

**STUDY OF PROTEIN-DNA INTERACTION
USING NEW GENERATION DATA**

ZHIZHUO ZHANG

NATIONAL UNIVERSITY OF SINGAPORE

2013

STUDY OF PROTEIN-DNA INTERACTION USING NEW GENERATION DATA

ZHIZHUO ZHANG 2013

**STUDY OF PROTEIN-DNA INTERACTION
USING NEW GENERATION DATA**

ZHIZHUO ZHANG

(B. Tech. (Computer Engineering), SCUT China)

**A THESIS SUBMITTED
FOR THE DEGREE OF DOCTOR OF
PHILOSOPHY
DEPARTMENT OF COMPUTER SCIENCE
NATIONAL UNIVERSITY OF SINGAPORE**

2013

DECLARATION

I hereby declare that the thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

ZHIZHUO ZHANG

25 JAN 2013

ACKNOWLEDGEMENTS

I feel honoured to have been able to study at the NUS School of Computing, where I had the good fortune to meet many excellent professors like Prof. Lim-Soon Wong, Dr. Wing Kin Sung, Dr. Anthony K. H. TUNG and others. I learned so much from their insight, advice and tremendous knowledge of computer science and biology.

I wish to sincerely thank my supervisor Dr. Wing Kin Sung. Through all my PhD life, Dr. Sung encourages me to devote myself in my bioinformatics research by his passion and insightful regular discussions. Most of the ideas in this thesis were generated and refined through these discussions. It has been an amazing transformation for me, from a pure computer science student to a professional in computational biology during the past five years. He reminds me that bioinformatics is a data-driven research, and developing practical application to the real world data is much more important than solving theoretical toy problems. He also reminds me that the standard of a good research is whether that research can bring high impact to the field. Hence, throughout my research, although we have explored many different subjects in machine learning, statistics and algorithmics, the practical value of those works are most important thing that we have been concerned with. To achieve this goal, Dr. Sung has been very generous in giving his time whenever I want to discuss with him, even in the off-work hours. He has always been very supportive throughout my PhD and tolerant towards my shortcomings, which make him a valuable friend in my daily life and a role model leading me to become a better man.

I thank Dr. Hugo Willy for his generous support, invaluable trust, inspiring discussions and insightful suggestions on my research in computational biology, and for being a great friend and a source of encouragement and understanding. He always has a helping hand for everyone and teaches me how to achieve internal happiness and overcome personal difficulties.

I am also grateful to Prof. Toh Kim Chuan, whom I met in later part of my PhD study. He introduces me to semi-definite programming and helps me on the research of chromosome 3D modeling. As a mathematician, he is very generous to share with me the precise problem formulation and numerical techniques. Without his help, the publication from this work would have been impossible.

Thought-provoking discussions with my colleagues Dr. Edwin Cheung, Dr. Guoliang Li, Dr. Sucheendra Kumar and Dr. Huy Hoang Do have been valuable in deepening my understanding of my research subjects. Dr. Edwin Cheung not only shares with me his professional knowledge on the biology experiment protocol and gene regulation mechanism but also provides high quality wet-lab data for me to study. Dr. Guoliang Li is a great senior, and he impartially teaches me how to analyze ChIA-PET data with great patience. Dr. Huy Hoang Do is a heart-warming friend, and he is always patient to listen to me and inspire me on algorithmic problems.

I would like to thank my many friends and colleagues at NUS and GIS with whom I have had the pleasure of working over the years. These include Chandana Tennakoon, Zhang Haojun , Lim Jing Quan , Guan Peiyong, Benjamin Mate Gyori, Lim Jun Liang Kevin , Feng Mengling, Fan Mengyuan , Wei Xueliang, Yong Chern Han , Zhou Hufeng, Fabianus Hendriyan , Chang Chengwei, Cai ShaoJiang, Gaye Saginc, and all members of the School of Computing in NUS. Their encouragement, friendship, and help have brought me incredible joy during my NUS days.

Finally, I sincerely thank my parents and all members in my family for their sacrifices to support me and encouraging my pursuit of graduate studies. Finally, I thank my fiancée Liang JunQian for bearing with my countless weekends and late-night stays in the office, for listening to my wild ideas and endless details, and for her encouragement on my work with love, patience and understanding.

Zhizhuo Zhang

TABLE OF CONTENTS

CHAPTER - I INTRODUCTION	1
I-1 BACKGROUND	1
I-1.1 Gene Regulation	2
I-1.2 Nature of Protein-DNA Interaction	4
I-2 BIOTECHNOLOGY ADVANCES	7
I-2.1 Chromatin immunoprecipitation related Technology	9
I-2.2 Chromosome conformation capture	11
I-3 RESEARCH PROBLEMS	16
I-3.1 Binding Motif Enrichment Analysis	16
I-3.2 De Novo Motif Finding Analysis	16
I-3.3 3D Chromosome Structure Modeling	17
I-4 THESIS ORGANIZATION	17
CHAPTER - II LITERATURE REVIEW	19
II-1 MOTIF ENRICHMENT ANALYSIS	19
II-1.1 General Method	20
II-1.2 Method in ChIP-seq Era	22
II-2 DE NOVO MOTIF FINDING	24
II-2.1 General Method	25
II-2.2 Method in ChIP Era	27
II-3 3D CHROMOSOME STRUCTURE MODELING	28
II-3.1 Pair based Analysis	28
II-3.2 Heatmap based Analysis	29
II-3.3 3D Structure Modeling	30
II-4 REVIEW SUMMARY	34

**CHAPTER - III CENTDIST: MOTIF ENRICHMENT ANALYSIS FOR
CHIP-SEQ DATA..... 37**

III-1	INTRODUCTION	37
III-2	PROBLEM DEFINITION	38
III-3	CENTER DISTRIBUTION SCORE	39
III-3.1	<i>Generation of the frequency and velocity graphs</i>	40
III-3.2	<i>Z-score for frequency graph</i>	41
III-3.3	<i>Z-score for velocity graph</i>	42
III-3.4	<i>Center distribution score for a motif distribution</i>	43
III-4	IMPLEMENTATION OF CENTDIST	44
III-5	RESULTS	46
III-5.1	<i>Conclusion</i>	51

**CHAPTER - IV SIMULTANEOUSLY LEARNING DNA MOTIF
ALONG WITH ITS POSITION AND SEQUENCE RANK PREFERENCES
THROUGH EM ALGORITHM 54**

IV-1	INTRODUCTION	54
IV-2	SEME ALGORITHM	57
IV-2.1	<i>Review of Mixture Model for Motif Finding</i>	57
IV-2.2	<i>Mixture Model in SEME</i>	59
IV-2.3	<i>Identifying Over-represented l-mers</i>	62
IV-2.4	<i>Extending EM Procedure</i>	62
IV-2.5	<i>Re-sampling EM Procedure</i>	66
IV-2.6	<i>Sorting and Redundancy Filtering</i>	71
IV-3	RESULT	72
IV-3.1	<i>Profiling two novel EM procedures</i>	72
IV-3.2	<i>Comparing TF motif finding in large scale real datasets</i>	76
IV-4	CONCLUSION	85

CHAPTER - V INFERENCE OF SPATIAL ORGANIZATIONS OF CHROMOSOMES USING SEMI-DEFINITE EMBEDDING APPROACH AND HI-C DATA 86

V-1 INTRODUCTION86

V-2 METHOD88

V-2.1 From Distance Matrix To 3D Structure90

V-2.2 Formulation of SDP relaxation problems92

V-2.3 Obtaining 3D coordinates from the Kernel Matrix94

V-2.4 Searching for the Correct Conversion Factor95

V-3 RESULT99

V-3.1 Simulation Study99

V-3.2 Real Hi-C Data Study104

V-4 DISCUSSION111

CHAPTER - VI CONCLUSIONS AND FUTURE DIRECTIONS..... 112

VI-1 CONCLUSION112

VI-2 FUTURE WORKS115

Summary

Understanding how proteins interact with DNA is essential for decoding many biological processes and disease states. We can study protein-DNA interaction in two levels: sequence level and structure level. Recent improvement in biotechnology enables us to study protein-DNA interaction in a high-throughput manner. For sequence level, we have Protein Binding Microarray(PBM) and ChIP-seq. For structure level, we have Hi-C and ChIA-PET. Novel bioinformatics problems are generated when scientists analyze the data of these new technologies. The aim of this thesis is to propose novel computational methods to solve the new challenges brought by the new generation data.

For the sequence level, ChIP-chip or ChIP-seq experiments can identify the binding sites of a selected transcription factor. Given these binding sites, there are two bioinformatics problems. One is motif enrichment analysis, and the other one is *de novo* motif finding. For motif enrichment analysis, the performance of existing programs is heavily dependent on the proper background and other parameter settings. A novel parameter-free method called CENTDIST is developed in Chapter III, and it tunes its parameters automatically and assesses motif enrichment by utilizing center distribution property from the ChIP-seq data. For *de novo* motif finding, existing programs over-take the prior knowledge from the ChIP-seq data, which may be only suitable for ChIPed protein but not collaborating transcription factors (co-TF). A novel EM(expectation maximization)-based program called SEME

is developed in Chapter IV, and it learns different positional bias and sequence rank bias for different motifs by estimating the parameters in a mixture model with EM technique. Large-scale ChIP-seq and ChIP-chip experiments demonstrate that CENTDIST can obtain better result than existing programs without requiring expert knowledge in configuring the program. SEME not only reports more accurate co-TF motifs than other programs but also correctly estimates the position and sequence rank distribution of each co-TF's motif.

For the structure level, Hi-C or TCC experiments can identify the contact frequencies among different genome regions with a fine-grained resolution. This type of data leads to a novel bioinformatics problem, that is, to identify the underlying 3D structure of the genome. Although a few works have been proposed recently, they do not guarantee to reconstruct the correct structure even in the noise-free case. To fill in the gap, a semi-definite programming (SDP) based algorithm called ChromSDE is developed in Chapter V, which guarantees to recover the correct 3D structure when the structure is uniquely localizable. Furthermore, the parameter of conversion from contact frequency to spatial distance is proved to change under different resolutions theoretically and empirically. Comparing with existing methods, ChromSDE doesn't assume the conversion parameter is known or fixed, but searches for the correct value based on the input data. Experimental result indicates that ChromSDE is more accurate than existing methods and its predicted 3D structure can provide novel information of the chromosome spatial organization which is hidden from the linear view of the chromosome.

In conclusion, this study has achieved several important improvements on processing the new-generation protein-DNA interaction related data at both the sequence level and the structure level. Novel information, like co-TF binding features

and 3D interaction features, can be unveiled by our bioinformatics programs. While there is still much room to further explore these new-generation data, future works are given at the end of this thesis.

LIST OF TABLES

Table I-1: Comparison between different CHIP based techniques.	10
Table II-1 Summary of Different Motif Enrichment Analysis Programs. (*MT: Motif Matching Threshold; LEN: length of genomic region to be considered; CON: conservation information among species; ST: scoring type; PM: Primary Motif PWM;.....	23
Table III-1 The pseudo code of CENTDIST Algorithm.	44
Table III-2: Comparison of CENTDIST, CEAS, and CORE_TF for different ChIP-seq datasets. * The output result of CENTDIST* is ranked by the Z-score of frequency graph only. The columns 4th-6th are the results for CORE_TF using promoter background (promBG,default background for CORE_TF) with enriched region size 200-1000 respectively, and the column 7th-9th are the result of CORE_TF using random genome background(randBG) with enriched region size 200-1000 respectively. The last three columns are the results of CEAS with enriched region size 200-1000 respectively.	50
Table V-1: The conversion factors estimated by ChromSDP and BACH. Each table element is the mean value of the estimated conversion factor across all chromosomes, and the value in each bracket is the standard deviation of the corresponding mean.....	105

LIST OF FIGURES

Figure I-1: Transcription initiation process regulated by transcription factors. Red line presents enhancer sequence and green line present promoter sequence. (a) Sequence level of view. (b) Structure level of view.	4
Figure I-2: An example using consensus and PWM to represent the same set of binding sequences.	6
Figure I-3: Chromatin Immunoprecipitation (ChIP) experiment workflow. 8	
Figure I-4 Outline workflows for Hi-C and ChIA-PET. Two cross-linked chromatin complexes are shown at the top as the first step in both experiments. (Left Panel) For Hi-C, restrict enzyme digest the chromatins and ligation junctions with biotin(red circle) are attached to the end of DNA. Then, DNA with ligation junctions are sheared and followed up pair-end sequencing. Final, whole genome contact heatmap is generated by counting pair-end read falling into given two genomic region. (Right Panel) For ChIA-PET, chromatins are sonicated and filtered by specific antibody(black fork symbol) of the interested protein (orange color). Following two types of biotinylated linkers are ligated to the end of DNA, and the linkers are further ligated to one another. Then, the ligation products are digested by Mme1 enzyme and further sequenced. Finally, the read with hybrid types of linkers are chimeric, and filtered. The rest of reads are mapped to the reference genome for interaction calling and peak calling. 14	
Figure II-1: Workflow of Motif Enrichment Analysis.	22
Figure II-2. Demonstration Pol2 ChIA-PET interaction in MCF-7 cell on the linear chromosome view. In the first row, red color arcs present the intra-chromosome interaction of Pol2. In the second row, the green and blue color regions represent the gene bodies of different genes.....	29
Figure II-3: An example of Hi-C heatmap analysis on mouse chromosome 17 (derived from [43]). (a) Hi-C heatmap represents the contact frequencies for any pair of loci on chr17:10Mb-90Mb. Hotter color indicates higher contact frequency for given two loci. Topological domain can be defined as the hot sub-region in the heatmap (b) PC1(middle track) presents the first principle component of PCA analysis of the Hi-C heatmap above. It shows that the PC1 signal is highly correlated with histone modification (H3K4me2, red) and gene density(purple).....	31
Figure II-4: General workflow of chromosome 3D modeling. It starts from a contact frequency matrix (or Hi-C Heatmap), then the contact frequencies are converted to a set of spatial distance constraints among genomic locations. Further, the optimization problem is modeled to find the	

corresponding 3D coordinates for the genomic locations satisfying the spatial constraints. Final 3D structure is built by linking the resolved 3D coordinates based on their order in the linear genome.32

Figure III-1. **Frequency and velocity analyses of the AP4 motif.** (a) The frequency graph of the AP4 motif in an AR ChIP-seq dataset, (b) The velocity graph of the AR motif in an AR ChIP-seq dataset. In each graph, the dotted line partitions the distribution into the enriched region (left region) and the non-enriched region. The dotted line is determined by maximizing the frequency difference between the two regions.39

Figure III-2: **Demonstration of CENTDIST Capability.** (a) CENTDIST promotes the enrichment of AR motif in the AR ChIP-seq dataset (LNCaP cell line). The blue bar and red bar show the Z-scores of the AR motif computed using the traditional enrichment method under the default enriched region size of 200bp and the default PWM cut-off (1.32, FDR=0.0001) using random genome region background(blue) or promoter background(red) respectively. The green bars show the Z-score of the AR motif computed by CENTDIST after it optimized different parameters. (b) CENTDIST represses the enrichment of the false CG rich motif in the Pol2 ChIP-seq dataset. All Z-scores are computed exactly as in (a). Since CENTDIST considers the velocity graph of the false CG rich motif, the combined Z-score of CENTDIST finally drops and is significantly lower than that computed by the traditional enrichment based method. As a side note, this result also shows that random background (blue bar) can produce quite different results compared to promoter background(red bar), which highlights the difficulty of choosing a correct background in existing enrichment based methods.47

Figure III-3: **Co-TF motif analysis of 13 Embryonic Stem Cell TFs using CENTDIST, CEAS, and CORE_TF.** (a) A comparison of co-TF motif analysis results using CENTDIST, CORE_TF and CEAS on 13 different ChIP-seq datasets from ES cell. The best setting in each dataset for CORE_TF and CEAS were used for comparison. CENTDIST*=CENTDIST algorithm without the inclusion of velocity score. (b) Heat map representing the analysis of 11 ES cell core TFs motif enrichment in 13 ChIP-seq experiments. Every row corresponds to a PWM motif while every column corresponds to a ChIP-seq dataset. The color of each entry presents the center distribution score (in log scale) of the motifs with respect to the peaks of the ChIP-seq dataset. The figure showed that the enhancer motifs are enriched in the enhancer ChIP-seq datasets (top left gene rectangle) while the promoter motifs are enriched in the promoter ChIP-seq datasets (bottom right green rectangle).48

Figure III-4: **Frequency analysis of ES cell TFs.** Every row corresponds to a PWM motif while every column corresponds to a ChIP-seq dataset. Each entry shows the frequency graph of the motif with respect to the peaks of the ChIP-seq dataset. Each graph shows the center enrichment region in red color and flanking enrichment region in green color. We observed that the frequency graphs in the top left blue rectangle show center enrichment while the frequency graph in the bottom right rectangle

shows center enrichment. All motifs are extracted from TRANSFAC database except the ones with suffix “_ES” which are the de novo motifs from[14]. 52

Figure IV-1 **Algorithm description for SEME Pipeline.**62

Figure IV-2: **Pseudo code for Extending EM procedure.**.....63

Figure IV-3: **Pseudocode for Re-sampling EM procedure.**.....71

Figure IV-4: **Procedure for computing AUC score.** Given a set of positive sequences and negative sequences, and a PWM motif. We compute the best match score of the PWM motif in every sequence. Then using different PWM score cut-off, we can compute the "True Positive Rate" and "False Positive Rate" of the PWM and generate the receiver operating characteristic (ROC) curve. Finally, the AUC score of the given PWM can be calculated as the area under the ROC curve.72

Figure IV-5: **EEM Motif Estimation and REM Sampling Efficiency.** (a)Comparison of EEM estimated motif length and actual motif length: The estimation of PWM length by the EEM procedure closely matches the actual planted motif. When the planted PWM have degenerate positions on its flanking positions, EEM will predict a shorter PWM which excludes the latter. (b) Efficiency Ratio and Recall Rate across different sampling ratios: We apply different sampling ratios to run SEME on 75 simulation datasets. The left y axis is value of the average efficiency ratio of SEME biased sampling against the uniform sampling. The right y axis is the average recall rate of true planted sites. The error bar presents the region +/- one standard deviation of recall rate.....73

Figure IV-6: **The empirical performance of SEME on synthetic datasets.** (a) The accuracy of SEME’s PWM (both EEM step’s (unrefined) PWM and the REM’s (final) PWM are listed). We quantify accuracy using the commonly used Area-Under-ROC Curve (AUC) score and PWM divergence (PD). We showed that EEM’s predicted PWM is already significantly stronger than random; indicating the goodness of EEM’s PWM as starting point for the subsequent REM step. The scores also show that SEME’s PWMs are significantly better when compared to MEME’s. (b) Based on the performances of SEME and MEME on the Pax4 motif dataset, we observed that MEME has serious difficulties in mining PWMs with long gap region within them. (c) The running time of SEME is shown against increasing input size. We observed that CUDA-MEME, the GPU enabled version of MEME, still runs slower than SEME running on normal CPU (it takes 1 day to handle ≈ 6000 sequences while SEME takes around 1 hour for 10000 sequences).75

Figure IV-7: **Comparison of de-novo motif discovery tools on the metazoan compendium.** Each column of the table presents the results for one motif discovery tool, and each row corresponds to one data set of the metazoan compendium. The color of the checkmarks represents the accuracy of the motif discovered as measured by the normalized euclidean

distance, and we used the thresholds on the PWM divergence as proposed by Linhart et al[73]. The symbol ∞ marks long execution times (hour) that were aborted in[73]. In the last row of the table, we report the total number of motifs discovered by each of the tools.78

Figure IV-8: **SEME detected TF motifs with significant position preference to TSS Seven examples of SEME’s output of metazoan compendium dataset.** The result indicates these TF binding sites are enriched near the transcription start sites. The TSS position is located around 200bp from the rightmost position. The original 1200bp promoter sequences may be shortened after removing “N”-masked regions, so the TSS position may be shifted in those cases.79

Figure IV-9: **The performance of SEME compared to existing motif finding programs from large scale real data** (a) Comparison result on the metazoan compendium datasets. Four PWM motifs returned by each motif finding program are then compared to the known Transfac motifs using PWM divergence (PD) (as in [73]) and further classified into three matching categories (strong, medium, weak) corresponding to different PD cut-offs (0.12,0.18,0.24). (b) Comparison result on 164 ChIP-seq libraries over four different measurements: AUC, PPV (Positive Predictive Value), ASP (Average Site Performance) and SPC (Specificity). The result shows that most motif finders perform similarly well in detecting ChIPed TF (but SEME is consistently better than all of them). (c) Comparison result for Co-TF motif finding on 15 ChIP-seq libraries. The quality of reported PWMs is classified into three categories (strong, medium, weak) corresponding to different STAMP p-value cut-offs (0.0001, 0.01, 0.05). SEME reported the most number of co-TF’s motif which match the known PWM with STAMP p-value ≤ 0.0001 (strong match, blue bar). Overall, SEME also found the most number of co-TF motif (61) as compared to the second best program, Amadeus (48).82

Figure IV-10: **Automatic learning of the position and sequence rank preference from the input data.** Instead of requiring the user to input the expected co-TF motif preference distribution (position and/or sequence rank distribution), SEME learns such distributions directly from the input data. We show that most of the time, SEME can learn the correct distributions of each TF (as compared to real binding sites distribution in the rightmost column, defined by the ChIP-seq and the known PWM of the TF). For position distribution, the x-axis is +/-200bp from ChIP-seq peak summit (the black dash line), and the y-axis is the fraction of binding sites in a given position. For rank distribution, the x-axis is the rank of ChIP-seq peak (left : high ChIP intensity, right : low ChIP intensity), and the y-axis is the fraction of binding sites in a given rank. The ChIP-seq peak rank distributions (MCF7 ER ChIP, LNCaP AR ChIP) of FoxA1 and the position distribution of Myc are tested to be insignificant by SEME.84

Figure V-1: **Simulation result that compares the performance of different regularization parameters.** The results were generated using simulation on a Brownian Motion Curve as stated in Section V-3.1. In the

experiments, different regularization parameter lambda (0.1, 0.01, 0.001) and different types of SDPs (linear, quadratic) combination were tested. (a) Spearman correlation between the pair-wise distance matrices of the predicted structure and the true structure under different noise level. (b) The absolute error of the estimated value of conversion factor under different noise levels.92

Figure V-2: **The effect of different conversion factors on the 3D structures predicted by ChromSDE.** The correct structure is a helix and the true value of conversion factor is 1. Each sub-figure is the predicted structure by ChromSDE with the specific conversion factor indicated below it. 97

Figure V-3: **Absolution error of estimated frequency with different values of conversion factor.** (a) Simulation study: the data is generated by helix curve under different noise level and true conversion factor is 1. (b)Real Hi-C data: the data is from chromosome 16-18 of mESC Hind3 dataset. 98

Figure V-4: **Algorithm description for ChromSDE**.....99

Figure V-5: **Different types of structures used in the simulation study:** Helix curve(Left), Brownian motion simulation of a single particle(Middle) and Uniform random points in a cube(Right).99

Figure V-6. **Predicted 3D structures by different programs using simulated data.** The Red curve is the true structure and the green curve is the predicted structure. ChromSDE uses quadratic SDP here and the linear SDP has the same performance.....101

Figure V-7. **Performance of different methods on simulated data.** (a) Spearman correlation between the pair-wise distance matrices of the predicted structure and the true structure under different noise level. (b) Running times of tested programs given different number of pairs of observed frequency (test stop at 80000 pair-wise frequencies, ~1600 points). (c) The absolute error of the estimated value of conversion factor under different noise levels. (d) The Consensus Index predicted by ChromSDE (quadratic model) under different degree of mixture of helix curve(right) and Brownian motion curve(left).102

Figure V-8: **Simulation result that compares the performance of different methods.** (a)Root mean square deviation (RMSD) between the predicted structure and true structure (Brownian curve). Generally, RMSD increases as the noise level increases. The linear SDP and quadratic SDP of ChromSDE performs similarly below noise level 0.7 and better than other methods. (b) Consensus Index predicted by two SDP formulations under different noise levels. Generally, the consensus index decreases as the noise level increases. (c) The value of the average conversion factor of different mix factors in Figure 3(d) for given noise level and the error bar represents the standard deviation of estimated conversion factor. (d) Consensus index decreases when the proportion of the dominate structure in the mixture

decreases or noise level increases. For dominate structure ratio 1 to 0.5, the results were generated by simulating two Brownian structures with mix factor 1 to 0.5 correspondingly. For dominate structure ration 0.333, 0.2 and 0.1, the results were generated by simulating a mixture with 3, 5 and 10 Brownian structures with equal proportion correspondingly.....104

Figure V-9: Validate ChromSDE using mESC, GM Hi-C data with two different enzymes (Hind3, NcoI). (a) Average Spearman correlation across all chromosomes between inverse 3D distance and contact frequency from testing dataset. For each dataset, the best performer is highlighted. (b) Alignment between predicted structures of chromosome 1 of mESC Hind3(red) and mESC NcoI(green) by ChromSDE. (c) Alignment between predicted structures of chromosome 1 of GM Hind3(red) and GM NcoI(green) by ChromSDE.106

Figure V-10: 3D structures predicted by ChromSDE using different enzyme data (red: Hind3, green : NcoI). The 3D structures are built using 1Mbp resolution data, and quadratic SDP . (a) mouse ES cell. (b) human GM cell. 107

Figure V-11: The consensus indices for different Hi-C datasets. (a) The consensus indices estimated across different chromosomes using Hind3 enzyme (blue) and NcoI enzyme(red) in mESC Hi-C datasets. (b) The consensus indices estimated across different chromosomes using Hind3 enzyme (blue) and NcoI enzyme(red) in GM Hi-C datasets.108

Figure V-12. Predicted structure of chromosome 13 from mESC Hind3 data. (a) The predicted structure of chromosome 13 under 1Mbp,500kbp,200kbp resolutions. (b) The predicted structure of the region chr13:21Mb-25Mb under 40kbp resolution and the different signal tracks of mESC from UCSC genome browser [58].110

Figure V-13: The performance of existing methods under data of different resolutions. The conversion factor of MCMC5C is set to be 0.5 for different resolution, and its predicted structures under different resolutions are not so similar. The conversion factors predicted by BACH and BACH* increase when the resolution increases.110

IUPAC codes for degenerate nucleic acids

A - adenosine	M - A C (amino)
C - cytidine	S - G C (strong)
G - guanine	W - A T (weak)
T - thymidine	B - G T C
U - uridine	D - G A T
R - G A (purine)	H - A C T
Y - T C (pyrimidine)	V - G C A
K - G T (keto)	N - A G C T (any)

LIST OF ACRONYMS

DNA	Deoxyribonucleic acid
TF	Transcription Factor
3C	Chromosome conformation capture
3D	Three Dimensional
ASP	Average Site Performance
AUC	Area Under the Curve
co-TF	colocalized transcription factor
DBD	DNA-binding domain
EEM	extending Expectation Maximization
EM	Expectation Maximization
ES	embryonic stem
FDR	False Discovery Rate
IP	immunoprecipitation
MEA	motif enrichment analysis
MEME	Multiple EM for Motif Elicitation
PCA	Principle component analysis
PCR	polymerase chain reaction
Pol2	RNA-polymerase II
PPV	Positive Predictive Value
PWM	positional weight matrix
REM	re-sampling Expectation Maximization
RMSD	root mean square deviation
RNA	ribonucleic acid
ROC	receiver operating characteristic
SDP	semi-definite programming
SPC	Specificity
TSS	transcription start site

PUBLICATIONS

Zhang, Z.Z., Li, G.L., Toh, K.C., Sung, W.K., “Inference of Spatial Organizations of Chromosomes Using Semi-definite Embedding Approach and Hi-C Data.”, in *RECOMB*, 2013.

Zhang, Z.Z., Chang, C.W., Willy, H., Cheung, E., Sung, W.K., “Simultaneously learning DNA motif along with its position and sequence rank preferences through EM algorithm.” in *RECOMB*, 2012.

Zhang, Z.Z., Chang, C.W., Goh, W.L., Sung, W.K., Cheung, E. “CENTDIST: discovery of co-associated factors by motif distribution.” *Nucleic Acids Res* 39, Web-Server-Issue (2011): 391-399

Ghosh, A., Saginc, G., Leow, S.C., Khattar, E., Shin, E.M., Yan, T.D., Wong, M., **Zhang, Z.Z.**, Li, G., Sung, W.K., Zhou, J., Chng, W.J., Li, S., Liu, E., Tergaonkar, V. “Telomerase directly regulates NF- κ B-dependent transcription.” *Nature cell biology* 14, no. 12 (2012): 1270-1281

Li, G., Ruan, X., Auerbach, R. K., Sandhu, K. S., Zheng, M., Wang, P., Poh, H. M., Goh, Y., Lim, J., Zhang, J., Sim, H. S., Peh, S. Q., Mulawadi, F. H., Ong, C. T., Orlov, Y. L., Hong, S., **Zhang, Z.Z.**, Landt, S., Raha, D., Euskirchen, G., Wei, C. L., Ge, W., Wang, H., Davis, C., Fisher-Aylor, K. I., Mortazavi, A., Gerstein, M., Gingeras, T., Wold, B., Sun, Y., Fullwood, M. J., Cheung, E., Liu, E., Sung, W. K., Snyder, M., Ruan, Y. "Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation." *Cell* 148, no. 1 (2012): 84-98.

CHAPTER - I Introduction

I-1 Background

Understanding how proteins interact with DNA is essential for decoding biological processes and disease states. DNA-binding proteins are the main regulators of gene expression. For example, the protein RNA polymerase can bind to DNA and transcribe gene regions into mRNAs. There are also other DNA-binding proteins called transcription factors (TFs) that can recognize specific short stretches of DNA sequences in the genome and regulate the target genes' expression. Gene is usually not regulated by only a single protein, but by a group of collaborating proteins (co-factors) binding to chromatin in close proximity. Apart from controlling gene expression, DNA-binding proteins are also the main constructors of chromatin structure. Histone proteins[56] control DNA accessibility by wrapping the DNA around them, and CTCF protein[22] is believed to act as a chromatin barrier by preventing the spread of heterochromatin structures. Hence, protein-DNA interaction is a very important subject in genetics.

In genetics, a DNA molecule firstly is considered as a sequence of nucleotides, where each nucleotide is encoded by one of four nitrogenous bases A,C,G,T (viz. Adenine, Cytosine, Guanine, and Thymine). Two complementary strands pair up where Gs pair with Cs and As with Ts to form base pairs (bp). Further, a DNA chain has a double helical structure, and is tightly packed around histone proteins. Within cells, DNA is organized into long structures called chromosomes.

This thesis studies protein-DNA interaction at both the sequence and structure levels. At the sequence level, each chromosome is treated as a one-dimensional

sequence, and each element on the chromosome is encoded by its linear position on that sequence (genomic location). At the structure level, each chromosome has a three-dimensional structure in the nucleus, and each element on the chromosome is encoded by three-dimensional coordinates (spatial location). Generally, if two elements are close at the sequence level, they are also close at the structure level. However, the reverse statement is not necessarily true.

In the following sections, some basic concepts of molecular biology are provided, which establish a ground for introducing the new generation experimental data and the corresponding bioinformatics problems.

1-1.1 Gene Regulation

DNA encodes genetic information. But it does not perform most of the functional activities. These activities are carried out by a set of functional molecules called proteins, which are complex macromolecules of amino acids. The central dogma in biology[21] describes the flow of genetic information from DNA to its final product “Protein”. A set of short segments in the long DNA chain, called genes, provide the templates for synthesizing short ribonucleic acid (RNA) molecules in a process called transcription. Those RNA molecules encode the information needed to construct proteins.

Although a majority of the cells in the same organism contain the same genetic information (DNA), the cells of different tissues have different types of proteins or different amount of certain proteins in order to function differently. The difference is controlled by a set of transcription regulators, so that only a fraction of the genes in a cell are expressed at a time. In eukaryotes, each gene is transcribed by a RNA-polymerase, and the transcription is initiated at a specific genomic location,

called the transcription start site (TSS, the blue right arrow in Figure I-1). However, the RNA-polymerase enzyme is incapable of initiating transcription on its own. The initiation process is assisted by a number of DNA-specific binding proteins called transcription factors (TFs). This process can be explored at both the sequence level and the structure level.

For the sequence level, TFs bind to the DNA sequence and interact with RNA-polymerase as shown in Figure I-1 (a). The sequences bound by TFs are called regulatory sequences, which usually contain specific sequence pattern (motif). The regulatory sequence near the TSS is called *promoter sequence* (green line), and the regulatory sequence far away from the TSS is called *enhancer sequence* (red line).

For the structure level, both *enhancer sequence* (red line) and *promoter sequence* (green line) are spatially close to the TSS, as shown in Figure I-1 (b). Also, transcription initiation is associated with open chromatin state (loose DNA region), in which the DNA around the TSS is unpacked in order for RNA-polymerase to bind on it. Another interesting fact related to transcription initiation at the structure level is that people observed the TSS of different genes are gathering spatially during the transcription, and this observation points out that all genes are transcribed together but in a more efficient way by sharing the TFs and recycling the RNA-polymerases. This phenomenon is called transcription factory[91], which is hidden at the sequence level.

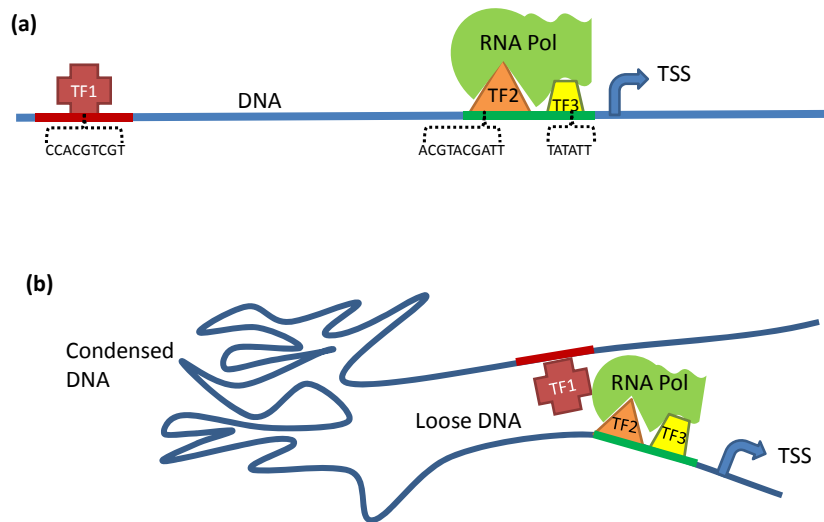


Figure I-1: **Transcription initiation process regulated by transcription factors.** Red line presents enhancer sequence and green line present promoter sequence. (a) Sequence level of view. (b) Structure level of view.

In short, a protein binding to the regulatory sequence can either directly interact with RNA polymerase or remodel the surrounding chromatin state, which promotes or inhibits RNA polymerase in the transcription process[19]. Thus the crucial point of the regulation mechanism is the binding of regulatory proteins.

I-1.2 Nature of Protein-DNA Interaction

There are two types of protein-DNA interaction based on how the protein binds on DNA. One type is sequence-specific binding. For example, a transcription factor (TF) contains one or more DNA-binding domains (DBDs), and has the affinity of binding to a specific DNA sequence. The other type is non-sequence-specific binding, with which the protein doesn't recognize a specific DNA sequence, but binds to DNA by forming complex with TFs or wrapping DNA around their surfaces.

I.1.2.1 Sequence-specific binding

The binding sequence of a TF is usually of length 5-30 bp and can be identified experimentally[12]. The quantitative modeling of TF binding specificity was firstly introduced by von Hippel and Berg [122]. Generally, the bases in a binding sequence are not equally important. Some bases can be substituted without affecting the affinity of the binding, but some bases are critical for binding and substitution at those bases can reduce binding affinity or completely inhibit binding. DNA motif is denoted as the conservation feature of binding sequence pattern for a TF, and there are two common ways for modeling DNA motif computationally.

One representation is called consensus pattern, which presents the motif of a set of binding sequences by the conserved nucleotide in each position. If the conserved pattern is significant, it can be changed to any binding sequence instance by a few substitutions. For example, Figure I-2 shows the consensus pattern of the binding sequences is “TTGACA”. Note that all the binding sequences can be formed from the consensus pattern by at most one substitution.

The other common representation is called *positional weight matrix* (PWM), which is numerically more precise than consensus pattern. The consensus pattern cannot tell the conservation of each base. Such information can be encoded in PWM. PWM models a motif of length m as a $4 \times l$ matrix Θ , where the entry $\Theta_{q,p}$ gives the probability that an occurrence of the motif contains a base q ($q \in \{A,T,C,G\}$) in its p -th position. Each column of the matrix therefore sums to one as illustrated in Figure I-2. Given a length- l sequence, let $s[i]$ denote the base at its i -th position, then the probability that Θ produces a particular sequence s is: $\Pr[s|\Theta] = \prod_{i=1}^l \Theta_{s[i],i}$. Given a set of motif occurrences S , the PWM can be easily computed by calculating the

frequency of each base in each position. Instead of reporting the probability, people commonly assess the goodness of a candidate binding sequence by PWM score, which is a log likelihood ratio of the probability of a given sequence under the PWM model, compared to a uniform 0-markov model[60]. Given a sequence s and a PWM Θ with length l , the PWM score is defined as:

$$\text{PWM Score} = \sum_{i=1}^l \log \left(\frac{\Theta_{s[i],i}}{0.25} \right) \quad (1.1)$$

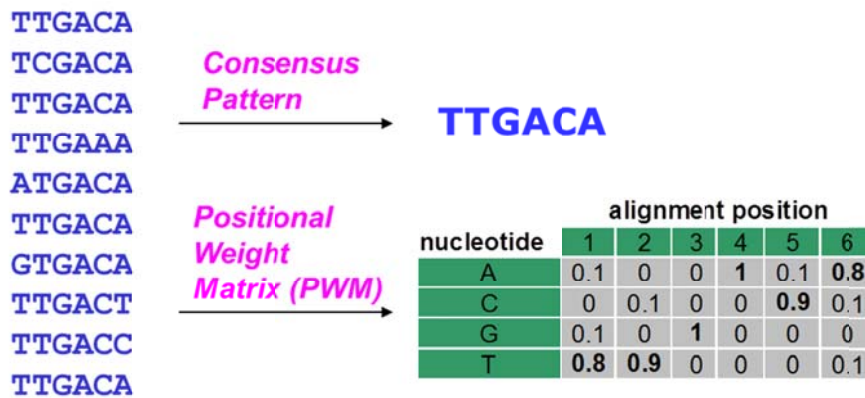


Figure I-2: An example using consensus and PWM to represent the same set of binding sequences.

I.1.2.2 Non-Sequence-specific binding

There are two well-understood types of non-sequence-specific protein-DNA interactions.

The first type is the histone protein. The histone proteins form an octamer complex, which contains two complete turns of DNA wrapped around its surface. These non-sequence-specific interactions are stabilized by the ionic bonds between basic residues in the histones and the acidic sugar-phosphate backbone of the DNA, which thus are independent of the nucleotide types on the DNA[83]. Chemical

modifications (i.e., methylation, phosphorylation and acetylation) on these interacting residues can change the strength of the interaction between the DNA and the histones. As a result, the wrapping DNA becomes more or less accessible to transcription factors, which affect the rate of transcription[13].

The second type is general transcription regulator protein, which can interact with the DNA indirectly through binding on the transcription factors or histone proteins. For example, P300 protein is a general activator, who binds to several different DNA-binding proteins [77]. P300 can bind on a TF called CREB, through its protein interaction domain KIX to enhance the transcription of target genes of CREB. Moreover, p300 also interacts with histone through protein interaction domain HAT, which acetylates conserved lysine amino acids on histone proteins and relaxes the chromatin structure.

I-2 Biotechnology Advances

The knowledge of protein-DNA interaction is enriched through the advance in biotechnologies, including breakthroughs in chromatin immunoprecipitation (ChIP) technology, parallel-sequencing technology and chromosome conformation capture (3C) technology. At the sequence level, we have technologies like PBM[12] and ChIP-seq[59]. At the structure level, we have Hi-C[76] and ChIA-PET[37]. The data produced by the new technologies leads to more novel bioinformatics problem. This thesis addresses some of these problems. First, this section gives a brief description of these biotechnologies.

I-2.1 Protein Binding Microarray

Protein binding microarray (PBM) is an in vitro technology to characterize TF's sequence specificities in a high-throughput manner. In PBM experiments, a DNA binding protein of interest is expressed with the epitope tag. A double-stranded DNA microarray with more than 2.3 million probes is designed based on *de Bruijn* sequences[3], in which all 8-mers can be represented in an overlapping manner in an unbiased fashion. The non-bound proteins are further removed by washing the protein-bound microarray. Then those bounded-proteins are labeled with a fluorophore-conjugated antibody specific to the epitope. The fluorescence signals of the microarray provide the quantitative measurement of the relative amounts of protein bound to each of the probe sequences. However, it is known that the raw probe signal contains the bias of the relative binding position in a probe [128]. Specific *de novo* motif finding algorithms [1, 44, 90, 139] are developed for PBM to identify the binding sequence pattern for the probe sequences sorted by the normalized probe signals.

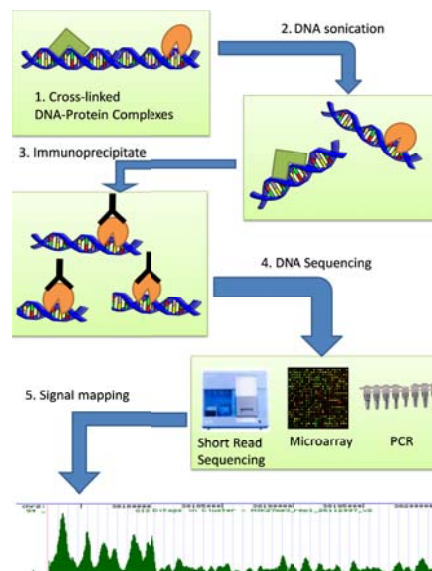


Figure I-3: Chromatin Immunoprecipitation (ChIP) experiment workflow.

I-2.2 Chromatin immunoprecipitation related Technology

Chromatin Immunoprecipitation (ChIP) is a type of immunoprecipitation experimental technique used to capture the interactions between specific proteins and DNA in the cell[113]. It identifies a set of protein-DNA complexes of interests using specific antibody. The workflow of ChIP experiment is shown in Figure I-3. First, a DNA-binding protein is cross-linked to its genomic DNA targets in vivo. Second, the protein-DNA complex is extracted from cells and the bounded DNA is further sheared by sonication into DNA fragments. Third, the cross-linked DNA fragments with the protein of interest are enriched by immunoprecipitation (IP) with an antibody that specifically binds that protein. Finally, the IP-enriched DNA fragments are examined using different techniques.

For example, ChIP-PCR [113] is used to test whether the pre-defined DNA sequences are enriched in the identified DNA fragments. Tiling array (ChIP-chip) or massive parallel sequencing (ChIP-seq) techniques can map the identified DNA fragments to the locations in the reference genome. Also, the enhanced version of ChIP called ChIP-exo[97] was developed recently, which applies a lambda exonuclease to further cut the unbound parts of the ChIP fragments and improve the resolution. Table I-1 shows the comparisons between different techniques. Note that the latest developed techniques are all high-throughput approaches, which target higher sensitivity (more genomic regions) and higher specificity (higher resolution). For these high-throughput approaches (ChIP-chip, ChIP-seq and ChIP-exo), they share similar dry lab protocol, which consists of three steps: mapping, peak calling and downstream analysis. The following sections briefly describe these three steps.

I.2.2.1 Mapping

The ChIP fragments can be processed using tiling array (ChIP-chip) or short read sequencing (ChIP-seq). For ChIP-chip, each probe in the array has its corresponding mapping location in the reference genome. The fluorescence signals of hybridized probes are mapped to the corresponding locations in the reference genome.

For ChIP-seq, millions of short reads are generated from the ChIP fragment and represent either a fragment start or end. The short reads can be mapped to a reference genome using different alignment program, such as BWA[75], Bowtie[69] and BatMis[115].

Table I-1: Comparison between different CHIP based techniques.

Technique	Sequencing type	Invention Time	Targeted region	Throughput	Resolution
ChIP-PCR [113]	PCR	1988	Selected region	Low	~100bp
ChIP-chip [2]	Microarray	2004	Promoters	High	~200bp
ChIP-seq [59]	Short read sequencing	2007	Whole genome	High	~100bp
ChIP-exo [97]	Short read sequencing	2010	Whole genome	High	~10bp

I.2.2.2 Peak Calling

Once the signals (either fluorescence signal or short read) are mapped to the reference genome, the peak calling procedure estimates the binding site location in the reference genome based on signal coverage. Usually the genomic background signal coverage is also required and prepared by performing the same ChIP experiment without immunoprecipitation with an antibody. Peak calling programs such as MACS[134] and CCAT[132], can identify a set of small regions with significant ChIP enrichment against background in the reference genome called ChIP peaks. Generally, each ChIP peak has two attributes: peak summit and ChIP intensity. Peak

summit indicates the most probable binding site location in the reference genome and ChIP intensity indicates the binding strength. They are important to the downstream analysis.

I.2.2.3 Downstream Analysis

Identified ChIP peaks can be used to analyze the binding profile of different DNA-interacting proteins including RNA polymerases, transcription factors, transcriptional co-factors, and histone proteins [106]. There are several common downstream analyses, such as peak-gene association, binding motif analysis, and peak annotation. For peak-gene association, the genes near the ChIP peak locations are treated as targeted genes, and the gene ontology analysis (or gene expression analysis) can be further performed to summarize the target genes function (or binding effect on gene expression). For binding motif analysis, the DNA sequences around ChIP peaks are extracted to identify whether any over-represented DNA motif enriched with the ChIP peaks, which can indicate the sequence-specific binding patterns of ChIPed proteins or their co-associate proteins. For peak annotation, the locations of ChIP peaks are overlapped with the annotation data in the reference genome, in order to check with whether the ChIP peaks significantly co-occurs with any type of annotation or not. In summary, these downstream analyses are very useful to understand the biology context of the ChIPed protein.

I-2.3 Chromosome conformation capture

Two genomic regions that are distal to one another on the linear view of genome can physically interact due to chromatin interaction. The ChIP technologies mentioned above cannot reflect this higher-order chromatin structure. To fill this gap, chromosome conformation capture or 3C [24], is a molecular biology technique used

to analyze the chromatin interactions in a population of cells. It measures the contact frequency between pairs of chromosomal loci, which can be used to further infer the structural properties and spatial organization of chromosomes.

Several enhanced techniques have been developed from 3C to increase the throughput of quantifying chromatin interactions in protein non-specific manner (4C [140], 5C[26], Hi-C[76] and TCC[61]) or in protein-specific manner (6C[118], ChIA-PET[37]). In this thesis, we focus on two methodologies: Hi-C and ChIA-PET, which have brought the assessment of chromatin interactions to the genome-wide scale.

I.2.3.1 Hi-C Experiment

Comparing to 3C, 4C or 5C, Hi-C introduces an unbiased way to measure the contact frequency of physical interaction between pairs of chromosomal loci on genome-wide scale. It solved the problem in previous 3C-related versions (including 4C, 5C), which require a set of pre-selected target loci, and are not designed for genome-wide studies.

Figure I-4(left panel) briefly shows the workflow of Hi-C, which contains seven steps. In Step 1, protein-DNA complexes are cross-linked with formaldehyde, such that interacting loci are bound to one another. In Step 2, the DNA is cut into a million pieces using a restriction enzyme. The specific restriction enzyme will recognize 6bp specific DNA sequences as cutting points (e.g., HindIII enzyme cuts “AAGCTT” sites). In Step 3, the ends of the overhang DNA fragments are filled in biotinylated residues. In Step 4, the overhang DNA fragments are further ligated with one another under dilute conditions. In Step 5, a set of ligation products with the biotin junctions are sheared, and pulled-down with streptavidin beads (not shown in the figure). In Step 6, the purified junction DNA fragments are subsequently

sequenced by high-throughput pair-end sequencing and the pair-end reads from two interacting DNA fragments can be read and mapped back to the reference genome. In Step 7, the contact frequency matrix is built by counting the number of pair-end read covered in any two-genome regions in resolution 40kbp-1Mbp.

Based on the contact frequency matrix (or contact heatmap), several common downstream analyses can be performed, such as chromosome topological domain study, 3D chromosome modeling, and interaction regions study.

I.2.3.2 ChIA-PET Experiment

Similar to Hi-C, ChIA-PET (i.e., Chromatin Interaction Analysis by Paired-End Tag Sequencing) allows the detection of long-range chromatin interactions on a genome-wide scale. The difference is that ChIA-PET integrates chromatin immunoprecipitation and chromatin proximity ligation. Hence, ChIP-seq identifies chromatin interactions mediated by specific protein only. Comparing to ChIP-seq, which is typically used for identification of the locations of TFBS[8, 125], and provides only one dimension information of those sites along the chromosomes (but not interactions between them), ChIA-PET further incorporates proximate ligation approach to link the free ends of the DNA fragments within the same protein-DNA complex, which captures the spatially contacting chromosome regions.

Figure I-4(right panel) briefly shows the workflow of ChIA-PET, which contains five steps. The first step is cross-linking protein-DNA, which is the same as Hi-C. In Step 2, the DNA is cut into a million pieces by sonication instead of restriction enzyme. In Step 3, protein of interest bound chromatin fragments are enriched by a specific antibody, which is the same as in the ChIP experiment. In Step 4, the ChIPed fragments are ligated with two types of biotinylated linkers(A and B)

containing an Mme1 restriction site, which allowed to ligate with one another under dilute condition. In Step 5, the purified ligation DNA fragments are digested by Mme1 and sequenced subsequently sequenced by high-throughput pair-end sequencing. In Step 6, the PET reads with AB linker composition are filtered, and the rest of PET reads are mapped to the reference genome. In Step 7, similar to peak calling in ChIP experiment, the “Self-ligation” peaks and “Inter-ligation” interactions are called based on the mapping coverage profile.

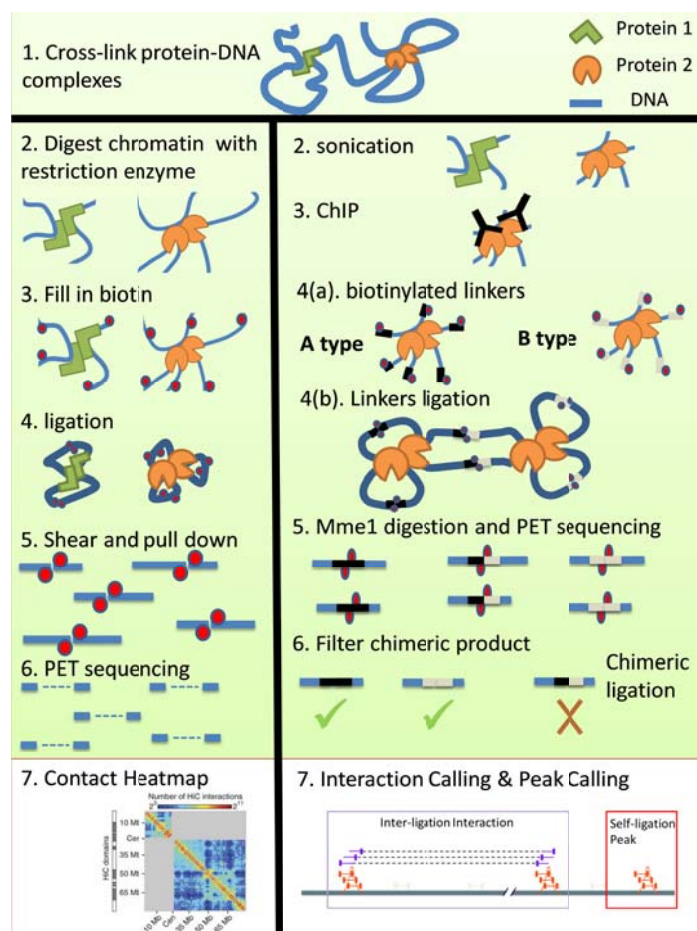


Figure I-4 **Outline workflows for Hi-C and ChIA-PET.** Two cross-linked chromatin complexes are shown at the top as the first step in both experiments. (Left Panel) For Hi-C, restrict enzyme digest the chromatins and ligation junctions with biotin (red circle) are attached to the end of DNA. Then, DNA with ligation junctions are sheared and followed up pair-end sequencing. Final, whole genome contact heatmap is generated by counting pair-end read falling into given

two genomic region. (Right Panel) For ChIA-PET, chromatin is sonicated and filtered by specific antibody (black fork symbol) of the interested protein (orange color). Following two types of biotinylated linkers are ligated to the end of DNA, and the linkers are further ligated to one another. Then, the ligation products are digested by MmeI enzyme and further sequenced. Finally, the read with hybrid types of linkers are chimeric, and filtered. The rest of reads are mapped to the reference genome for interaction calling and peak calling.

“Self-ligation” peaks can be used as ChIP-seq peaks in the downstream analysis. “Inter-ligation” interactions predict the list of genomic region pairs (1kb~5kb, from the same or different chromosomes) that have spatial proximity in the real cells. Similar to ChIP Intensity, PET count for each region pair can indicate the interaction strength between the two regions. Several common downstream analyses based on these interactions can be performed, such as promoter-enhancer association, chromatin loop annotation, and co-regulated genes analysis.

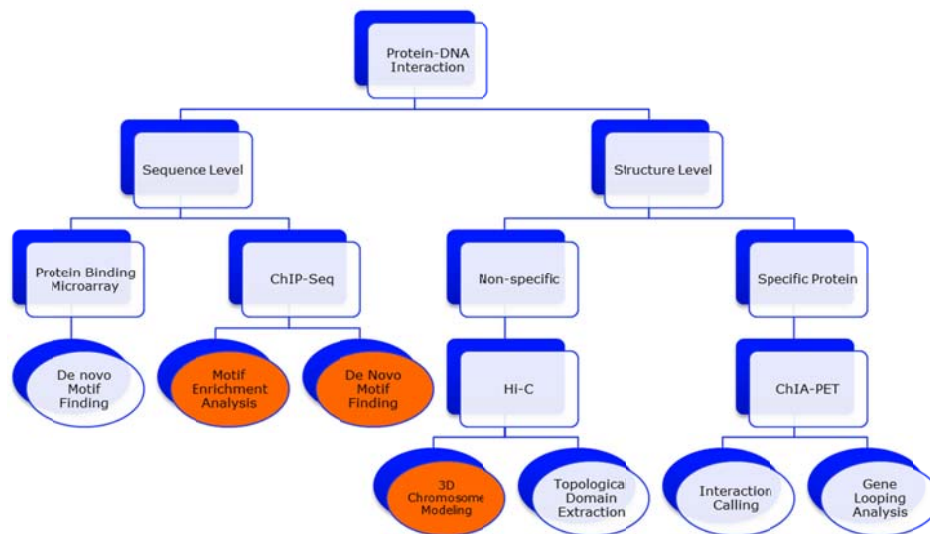


Figure I-5: New generation data of Protein-DNA interaction (rectangle shape) and new bioinformatics problems (eclipse shape). The three bioinformatics problems studied in this thesis are highlighted (orange color).

I-3 Research Problems

Despite of the breakthrough of these technologies, the extracted data lead to more novel bioinformatics problems (Figure I-5). This thesis addresses three of these problems.

I-3.1 Binding Motif Enrichment Analysis

The first problem is motif enrichment analysis using ChIP-seq data. The main application for solving this problem is to discover the co-TF using the known motif database. It assumes that the co-TF motifs will enrich around a binding sites of ChIPed TF. Recently, several studies[18, 16] showed that if two TFs are co-associated, their ChIP-seq peaks (or their binding sites) are not only in close proximity with each other, but the relative distance of each TF with respect to the other exhibits a peak-like distribution. We call this property the center distribution. In Chapter III, we examine whether the center distribution can be utilized for co-TF discovery.

I-3.2 De Novo Motif Finding Analysis

The second problem is de novo motif finding using ChIP data. Its main application is to recover the motifs for ChIPed TF and its co-TFs when the motif of the interested TF is not available in the known motif database. In the ChIP data, the ChIPed TF's motif (ChIPed TF is the TF pulled down in the ChIP experiment) prefers to occur in sequences with high ChIP intensity and also near the ChIP peak summits (thus having both position and rank preference). Hence, if we know the position preference and the sequence preference of the TF motifs in the input sequences, we can improve motif finding. However, it is an open question whether the position preference and the sequence preference can be treated as prior knowledge. In Chapter

IV, we explored an expectation-maximization method in de novo motif finding, which can automatically learn those preferences and provide novel finding.

I-3.3 3D Chromosome Structure Modeling

The third problem is building 3D chromosome structure using Hi-C data. The contact frequency matrix identified in Hi-C experiment gives a set of spatial distance constraints among different chromosome locations. Computationally, it is possible to embed each chromosome location in the 3D space and to satisfy all the spatial distance constraints indicated by Hi-C data. Many potential biological hypotheses are hidden when we assume that the chromosome is one dimension. However, they can be easily observed in the 3D space. This embedding problem can be solved naively using a non-convex constraint optimization method. However, such method cannot guarantee to find a feasible solution. In Chapter V, we proposed an elegant way to relax this problem as a semi-definite programming problem, which can be solved in polynomial time and guaranteed an optimal solution in the noise-free case and robust solution in the noisy case.

I-4 Thesis Organization

The thesis stands at the intersection of two areas, namely, computer science and molecular biology, and draws heavily on Bayesian statistics and optimization theory. Although the readers are not supposed to be the experts of these areas, general knowledge of basic concepts and techniques (e.g., DNA and Protein molecule, binomial statistics and convex optimization, etc.) is expected for the general audience in computational biology.

Three specific research problems, which are the main focus in thesis, have been briefly introduced in Section I-3, and each of them is presented in a separate chapter: motif enrichment analysis is addressed in Chapter III, followed by de novo motif finding problem in Chapter IV and finally chromosome 3D modeling problem in Chapter V. A review of the current literature within the scope of each research problem is given in Chapter II. Chapter III, IV and V are mostly self-contained and can be read separately from the rest. For the readers most interested in motif enrichment analysis, they are advised to read Section II-1 in Chapter II first and then Chapter III. For the readers interested in de novo motif finding, they are advised to read section II-2 in Chapter II first and then Chapter IV. For the readers interested in chromosome 3D modeling, they are advised to read section II-3 in Chapter 2 first and then Chapter V.

Chapter VI summarizes the contribution of this thesis and discusses some open questions and directions for future investigation.

Some of the material in this thesis has appeared before in my previous publications [135, 136, 138].

CHAPTER - II Literature Review

In this literature review, we will look at the following three fundamental genomic problems:(1) what are the TFs enriched in a set of regulatory sequences (Motif Enrichment Analysis); (2) What are the DNA binding motifs for a set of interested TFs (*De Novo* Motif Finding); (3) How do the chromosomes fold in 3D (chromosome 3D modeling). These problems are still unresolved even though many methods have been developed. Recently, novel experimental methodologies such as ChIP-chip ChIP-seq, ChIA-PET and Hi-C have been introduced. They provide unprecedented power for researchers to answer these fundamental problems.

II-1 Motif Enrichment Analysis

Transcription factors (TFs) will bind to specific DNA sequence pattern on the regulatory sequences of the targeted genes and regulate the expression of those genes. One basic question in bioinformatics is, given a set of regulatory sequences (e.g., promoters of a set of genes), to find TFs that bind on those sequences. If TF binding motifs are known, we can get the answer for this problem using motif enrichment analysis. Motif enrichment analysis is to determine whether the regulatory sequences have significantly higher than expected occurrences for a certain known DNA-binding motif. Such a motif is said to be "enriched" in that set of regulatory sequences. The TFs or microRNAs whose motifs are enriched in that set of regulatory sequences are candidate transcriptional regulators for some or all of the corresponding targeted genes.

II-1.1 General Method

Given a set of DNA sequences, motif enrichment analysis aims to identify over-represented known motifs in those sequences. The known motifs usually come from some public databases of TF motifs such as JASPAR [105] and Transfac [85].

When motif enrichment analysis is performed, we usually expect that input sequences enrich with binding sites of the same TF. Generally, such a set of input sequences can be a set of promoter sequences of co-regulated genes identified by expression microarray data or binding regions of certain TF identified using ChIP experiment data. Then, there are two stages in the analysis (see Figure *II-1*). The first stage is called "motif scanning". Each known motif in a database will be used to scan along the whole input sequences (and background sequences if provided) to determine their occurrences. Based on the motif model (usually PWM), the motif scanning program need to compute the matching scores for the tested motif and all substrings in the input sequences. If the matching score is higher than some user-defined threshold, the matched site is considered as one motif occurrence. The earliest developed motif scanning programs [20, 93] computed the matching score site by site and then filtered out the sites with the scores below given cut-off. However, this type of methods is considered to be slow since they need to scan the input sequences N times if we have N candidate motif models. To address the efficiency problem, some sophisticate data structures (e.g., Hash Table [108], suffix tree [68]) are used to index the original sequences. Instead of computing score for each site, they usually compute the matching scores for the k -mers appearing in the original sequences. Because the size of k -mers is usually much smaller than size of original sequences, the running time is improved.

In the second stage, different statistics can be applied to test whether the occurrences of the given known motif in the input sequences are significantly correlated with the signal labeled for the sequences or enriched under some null hypothesis (background model). The common types of statistics in motif enrichment analysis, are the Fisher Exact Test[54], the multi-hypergeometric test [29], the binomial test, the rank-sum test [109, 5], Clover[35] and Spearman's rank correlation[32]. Almost all these types of enrichment analysis require a background model or a set of background sequences. Hence, their power is limited to whether user can select the correct background.

Some programs are developed for motif enrichment analysis. ConTra[50], PASTAA[98] and oPOSSUM[48] can find the enriched known motifs in the promoters of the user input genes. Apart from considering statistical enrichment, they also check whether the matching sites are conserved in cross-species or not. This strategy is particularly useful when the number of input sequences is too small to make a statistically significant conclusion. As more and more functional inter-genic regions are identified, CEAS[57] and CORE_TF[47] are proposed and they allow user to input a list of genomic regions instead of only promoter regions. Although different programs compute different enrichment statistics, Robert C McLeay[86] gave a comprehensive examination for different enrichment statistics using a unified framework, and pointed out how to partition the data into positive set (regulatory sequences in interest) and negative set (background sequences), which would affect the result of motif enrichment analysis. Other than background, there are other factors affecting the motif enrichment analysis, such as motif matching threshold. A

summary for those factors corresponding to existing motif enrichment analysis programs is given in Table II-1.

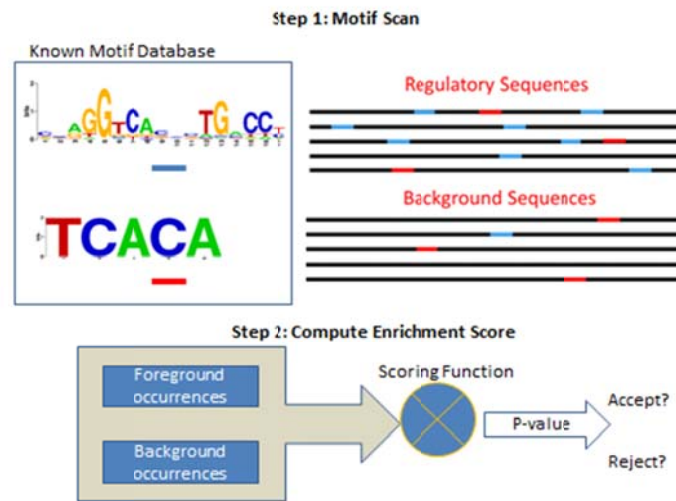


Figure II-1: **Workflow of Motif Enrichment Analysis.**

II-1.2 Method in ChIP-seq Era

The general motif enrichment analysis methods mentioned above are initially designed to analyze data from low throughput experiments such as promoter sequences of co-regulated genes. For the high-throughput experiments like ChIP-seq [125, 15, 31], which can identify thousands of binding sites of ChIPed TF in a high-resolution manner (~100bp). And most general motif enrichment analysis methods are able to correctly identify the motif of ChIPed TF with highest enrichment easily. Recently, people have shifted their focuses of motif enrichment analysis to further exploring the co-TF motif enrichment. The idea is to measure the enrichment of a set (pair) of motifs instead of single one. In order to interact with the ChIPed TF, people assume that the co-TFs sit close or with certain preferred distance to the ChIPed TF, that is, co-TFs' motifs co-occur with the ChIPed TF motifs in the input sequences. SpaMo[129] can identify a set of enriched co-TF motifs (from the known motif

database) with the given ChIPed TF motif by assuming they have preferred distance when forming TF complexes. However, this assumption may not be valid. On one hand, it is questionable whether there exists a fix distance between co-TF and ChIPed TF, since people believe most co-TFs can interact with the main TF when they are close enough. On the other hand, it failed to consider the case that ChIP-seq peaks may also present the indirect binding sites of ChIPed TF and there may be no ChIPed TF motif sites but the co-TF motif in those regions.

Table II-1 Summary of Different Motif Enrichment Analysis Programs.

(*)MT: Motif Matching Threshold; LEN: length of genomic region to be considered; CON: conservation information among species; ST: scoring type; PM: Primary Motif PWM;

Program Name	Genomic Regions	ChIP Peak Info	Background	Web	Additional Input*
ConTra	Promoter	NA	Need	yes	MT,LEN,CON
oPOSSUM	Promoter	NA	Need	yes	MT,LEN,CON
PASTAA	Promoter	NA	Need	yes	LEN
CEAS	Genome Wide	NA	Need	yes	LEN
CORE_TF	Genome Wide	NA	Need	yes	MT,LEN
AME	Genome Wide	ChIP Intensity	Optional	no	ST,MT,LEN
SpaMo	Genome Wide	Peak Location	No	yes	PM,LEN

To use the existing motif enrichment analysis in ChIP-seq data, apart from ChIP-seq peak locations, people need to input some additional information (the 6th column in Table II-1) before running the programs. Some information like the matching threshold, the length of interested regions and the choice of background regions can affect the final result of motif enrichment analysis, especially when the aim is to identify the potential co-TFs. Here, I summarize three limitations of existing methods using ChIP-seq data as follow:

The first limitation is to decide the distance between the ChIP peak summit and the motifs. The input sequences for motif enrichment analysis are usually extracted from regions around the ChIP peak summits. Different co-TF motifs may

have different distance distribution. If we scan motifs using a shorter or longer region around the ChIP peak summit, it may weaken the power of the enrichment statistics.

The second limitation is to compute the known motif occurrence. The DNA motif is usually modeled as a position weighted matrix (PWM) in the motif database. When we say a motif matches a sequence, it is usually referred to an approximate match (i.e., high PWM score). Hence, the cut-off for PWM score is usually needed and affects the follow-up enrichment analysis.

The last limitation may be the most serious problem. Almost all types of current enrichment analysis require a background model or a set of background sequences. Although AME[86] programs can compute the correlation between motif occurrences and the ChIP Intensity without any additional background, this method is only suitable to compute the ChIPed TF motif enrichment but not for computing the co-TF motif enrichment. The correct background model or background sequences are not easy to choose. For example, choosing all promoters sequences as background when analyzing mammal genome, the enrichment analysis may bias to AT-rich motif since most mammal promoters are CG-rich.

II-2 *De Novo* Motif Finding

Given a set of DNA sequences bound by the same TF, there are two ways to identify the binding motif for that TF. One is to search the know motif database to see which known motif is most enriched in the given set of sequences, which is stated in previous section. But this approach assumes the binding motif of the TF is known, which may not be true. The other approach is called *de novo* motif finding, which

tries to identify the recurring patterns (motifs) in the given set of sequences. This section reviews existing methods related to *de novo* motif finding.

II-2.1 General Method

De novo motif finding algorithm can generally be classified into three types based on their motif models: consensus, PWM and other forms.

For consensus based motif finding algorithms, the problem can be formalized as follows: Given a set of sequences, we aim to find a length l (6-12bp for TFBS) pattern so that the number of k -mismatch occurrences (where k is usually 1 or 2) of the motif is significantly over-represented in the input sequences. Exhaustive search for all 4^l candidate patterns is considered to be time-consuming. Different indexing structures of the input sequences have been proposed in this class of algorithm. Using indexing data structures (e.g. suffix tree[92], suffix array[68], and hash table[95]), it can efficiently identify short consensus motifs. Weeder[92], Trawler[30], YMF[111], DREME [4] are a few examples representing this line of approach.

Comparing to consensus, position weighted matrix (PWM,[110]) provides more powerful and flexible description of the binding specificity of a TF, and so it has been the most preferred way in motif modeling. The definition of PWM can be referred to Section I.1.1.2 in Chapter I. The *de novo* motif finding problem is to find a single or a set of PWMs, which can discriminate the input sequences and the background. Almost all the combinatorial optimization techniques (i.e. greedy[108], local search[6], stochastic search[99, 126] and so on) have been tried and applied over the years. Among all, Expectation-Maximization (EM)[6] and Gibbs sampling [99] are the two most common approaches to find a PWM but they usually require long running time. MEME algorithm [5, 81] models the motif finding problem as learning

the parameters for a mixture model. It assumes every length- l substring in the input sequences is generated from either a motif model Θ (i.e., PWM) or a background model \mathbf{B} (k^{th} -order markov model). To learn the parameters Θ and \mathbf{B} of a mixture model, EM algorithm is applied. The EM algorithm iterates two steps: E-step and M-step. In E-step, given the current best parameters Θ and \mathbf{B} , the likelihood for all l -mers in each input sequence are computed, and in M-step, MEME builds the new parameters Θ and \mathbf{B} using all l -mers weighted by the corresponding likelihood value. Different types of mixture model are provided in MEME, such as ZOO (zero or one occurrence per sequence), OOPS (only one occurrence per sequence) and TCM (general two component mixture model), in order to fit the prior expected occurrence per input sequence. Since this EM framework is very flexible, several MEME variants [81, 96, 43] are developed to speed up the EM process by parallel computing or data indexing, or to utilize additional information. Another related optimization approach is called Gibbs sampling, which is a stochastic counterpart of the EM. It has been implemented in several tools such as GibbsDNA[70], AlignACE[100], MotifSampler[117], BioProspector[79], ANN-spec[130], etc.

Other types of motif models are also proposed. Instead of assuming each position of a motif is independent like PWM, more complex motif models considering position dependency are developed. Bayesian network approach were used in[7, 11], and PWM was extended to di/tri nucleotide matrix to model the dependency of adjacent positions in a motif [141, 53]. Some other approaches use graph-based representation[34]. They represent each k -mer in the input sequences as a node in the graph and two k -mers are connected if they are similar. Such that a motif can be derived from the maximum density subgraphs. Although these complex

representations can capture inter-position dependency for the binding sites, they suffer seriously over-fitting and time-consuming problems.

II-2.2 Method in ChIP Era

For the general motif finding algorithms, they only examine the over-representation of sequence patterns, and often miss some real motifs and generate many false positives. Fortunately, additional information for the input sequences is found to be helpful to improve motif finding. As ChIP experiment is becoming a popular way to identify transcription factor binding sites, many new algorithms have been developed and optimized for ChIP data.

Similar to motif enrichment analysis, the input sequences for motif finding of ChIP data, are usually extracted from the regions around ChIP peak summit, and further sorted by the ChIP intensity (from high to low). Basically, the motif finding algorithm optimized for ChIP data is based on two assumptions: One is that the real ChIPed TF motif is more enriched in the regions (input sequences) with higher ChIP intensity. For example, MDscan [79] only considers high-ranking sequences to generate its initial candidate motifs. DRIM[28] searches the motif whose occurrences correlate with the ChIP intensity. The ChIPed TF motif is more enriched in positions close to the peak summits than the positions far away from the peak summits. Many programs allow users to specify the position prior distribution of motifs with respect to the peak summits [6, 92, 4, 67, 52]. Normal distribution or student-distribution is common used in modeling the motif position with respect to the ChIP peak summit.

However, the prior knowledge of ChIP data may only be true for the ChIPed TF motif, but may not for co-TF motifs. For example, co-TF motif may be a bit far away from the ChIP peak summits, which have already been occupied by ChIPed TF.

Also, it is possible that co-TF motif enriches in low ChIP intensity regions, which ChIPed TF may just indirectly bind to those regions through co-TF.

II-3 3D Chromosome Structure Modeling

In the previous sections, genome is usually assumed to be a set of linear chromosomes. This model, however, is oversimplified and it cannot explain the interactions among different genomic elements (e.g., enhancer, promoter, and gene). Chromosome actually forms a 3D structure within the nucleus and its spatial organization affects many chromosomal mechanisms such as gene regulation, DNA replication, epigenetic modification and maintenance of genome stability[23, 33, 42, 88, 87]. For example, the three-dimensional chromatin interactions have been shown to bring distal transcription factor binding sites into close spatial proximity to gene promoters [15]. ChIA-PET and Hi-C data give us the opportunity for global analysis of three-dimensional chromatin interactions in high-resolution and whole-genome manner. Generally, there are three popular ways to analyze interaction data: pair-based analysis, heatmap-based analysis and 3D structure modeling. The first two are briefly introduced in the following two sections, and this thesis focuses on 3D structure modeling, which is reviewed in the last section.

II-3.1 Pair based Analysis

Given a set of interacting loci pairs, people can easily overlap them with the known genome annotations. For example, we can annotate a set of pairs as "Promoter-Promoter" pairs, if the two ends of the pairs both overlap the promoter regions in the genome. Using this kind of annotation analysis, Li, et al. [74] shows that

promoters of genes with similar function will cluster together in 3D space and those genes will co-express.

Other than annotation, identifying sample-specific interaction pairs by comparing the data between different samples is also a common analysis. Usually, sample-specific interaction pairs give more biological insight for the regulation events under different experiment conditions [66]. Similarly, we can identify tissue specific interaction pairs by comparing the interaction pairs in particular tissue against the union of pairs in other tissues.

To visualize the distribution of interaction pairs, ChIA-PET tool developed by Li, et al. [73] shows interaction arcs over a linear chromosome(Figure II-2).

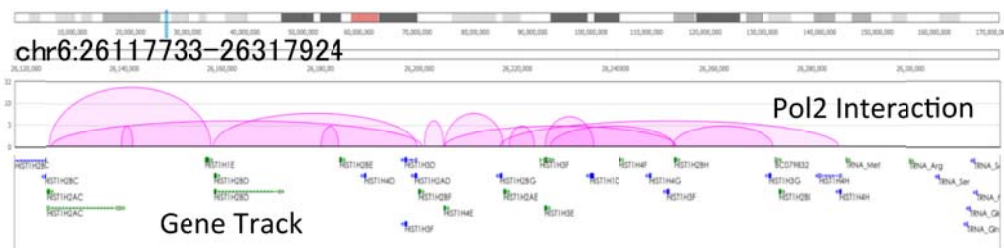


Figure II-2. **Demonstration of Pol2 ChIA-PET interaction in MCF-7 cell on the linear chromosome view.** In the first row, red color arcs present the intra-chromosome interaction of Pol2. In the second row, the green and blue color regions represent the gene bodies of different genes.

II-3.2 Heatmap based Analysis

Instead of showing interaction arcs over a linear chromosome, another way to present the interaction pairs is called contact frequency matrix, which is a popular representation in chromatin interaction data. The spatial contact matrix is constructed by dividing the genome into 1-Mb regions (“loci”) and defining the matrix entry M_{ij} to be the number of ligation products between locus i and locus j [76]. This matrix can reflect interaction frequency between any two genomic regions. As it is firstly used to

visualize the Hi-C data, it is also called a Hi-C heatmap (Figure II-3(a)). Figure II-3 shows an example for Hi-C heatmap analysis for mouse chromosome 17. A genomic region is defined as a topological domain if the interactions inside that region are much more frequent than the interaction between that region and other regions. Thus, based on the Hi-C heatmap, the topological domains can be computationally detected as the hot square regions on the heatmap (Figure II-3(a)). Moreover, Principle component analysis (PCA,[55]) is also a common analysis for the Hi-C heatmap. Figure II-3(b) demonstrates that the first principle component generated from the given contact frequency matrix are highly correlated with the H3K4me2 histone signal and the gene density.

II-3.3 3D Structure Modeling

The final goal of the interaction data is to understand the higher order architecture of genomic domains and entire genomes at unprecedented resolution. Therefore, given the interaction data, one interesting bioinformatics problem is to infer the 3D structure of the chromosomes.

Some progress has been made in reconstructing 3D structure of chromosomes using newly generated interaction data (4C, 5C, Hi-C and TCC), and most of them model this problem as a constraint optimization problem mathematically. Concretely, interaction data imposes a set of spatial distance constraints of the interacting points. The linear genomic distance and other prior knowledge provide additional spatial constraints, which enable us to model the problem as constraint optimization problem. The optimization problem resolves the three-dimension coordinates for each genomic location in the constraints, such that the linear genomic locations are mapped to the

three-dimension space. The general workflow of chromosome 3D modeling containing three components, and is shown in Figure II-4.

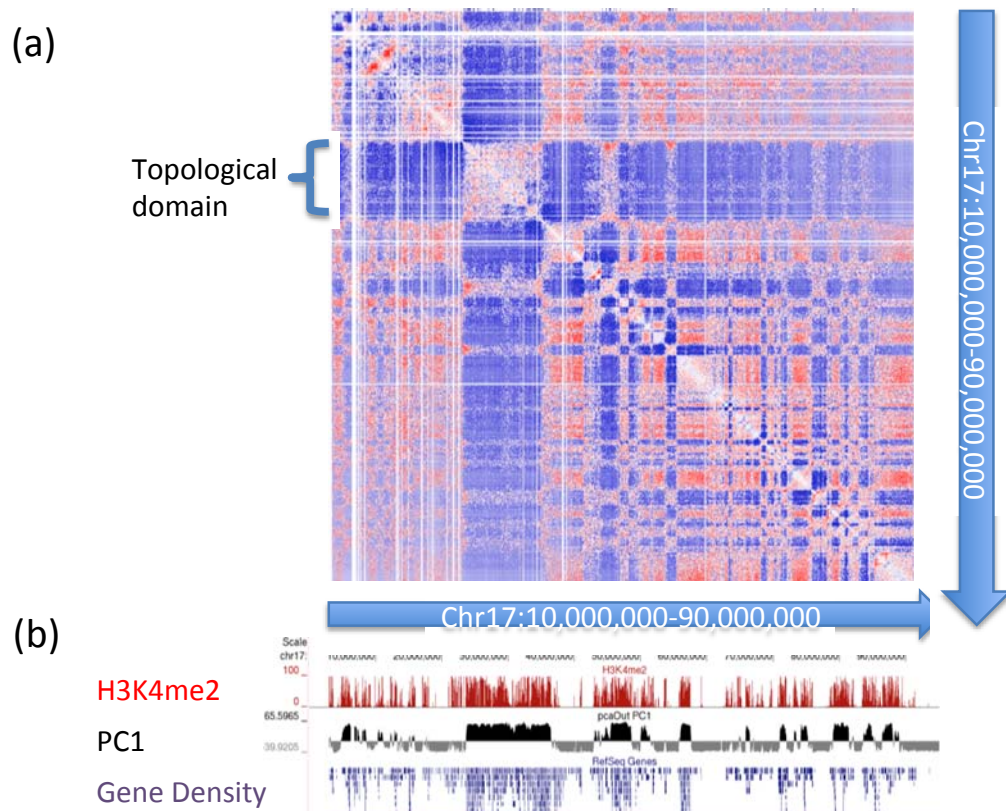


Figure II-3: **An example of Hi-C heatmap analysis on mouse chromosome 17** (derived from [46]). (a) Hi-C heatmap represents the contact frequencies for any pair of loci on chr17:10 Mb-90Mb. Hotter colors indicate higher contact frequency for given two loci. Topological domain can be defined as the hot sub-region in the heatmap (b) PC1 (middle track) presents the first principle component of PCA analysis of the Hi-C heatmap above. It shows that the PC1 signal is highly correlated with histone modification (H3K4me2, red) and gene density (purple).

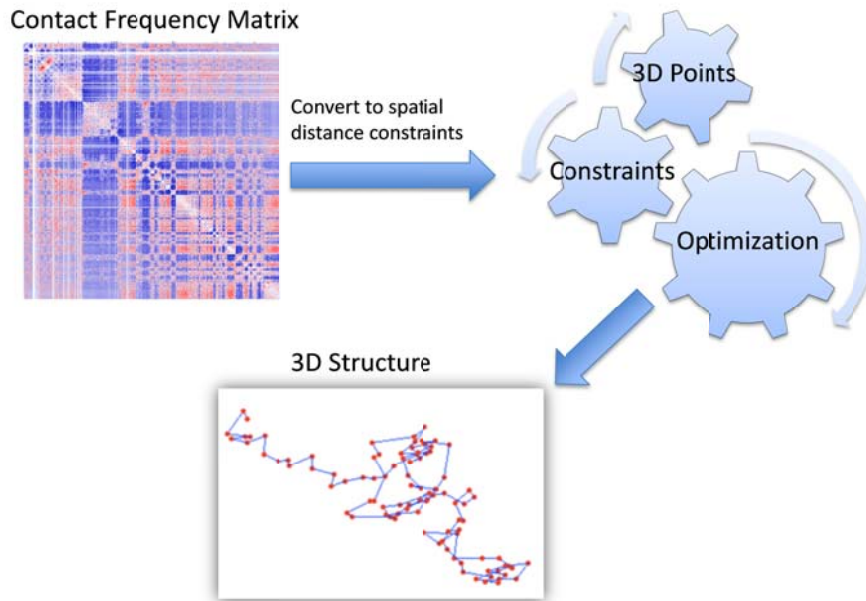


Figure II-4: **General workflow of chromosome 3D modeling.** It starts from a contact frequency matrix (or Hi-C Heatmap), and then the contact frequencies are converted to a set of spatial distance constraints among genomic locations. Further, the optimization problem is modeled to find the corresponding 3D coordinates for the genomic locations satisfying the spatial constraints. Final 3D structure is built by linking the resolved 3D coordinates based on their order in the linear genome.

According to this general workflow, a few works have been done. Duan, et al. [27] modeled this problem as a non-convex quadratic optimization problem. They presented each chromosome as a series of beads in 3D space, spaced 10kbp apart, and converted the contact frequencies extracted from the 4C experiment on yeast to spatial distances. Then they proposed to find a set of 3D coordinates for all the beads so that the differences between 3D distances and the prior expected distance are minimized, and meanwhile a set of distance constraints based on their prior knowledge is satisfied. IPOPT[123], a nonlinear constrained optimization program based on interior-point gradient-based methods, is applied to solve this problem.

Similar to [27], Baù and Marti-Renom [10] presented each chromosome as a series of particles in 3D space. They translated the contact frequencies extracted from

5C experiments to spatial distances by inverting the Z-score of contact frequencies and modeled the 3D structure modeling problem as finding an equilibrium state of a set of particles using Integrated Modeling Platform (IMP)[103]. With different optimization and searching strategy from [27], IMP simulates particles in many independent Monte Carlo rounds and local searching like simulated annealing is performed at each round to reduce the number of violated constraints. Finally, it ensembles the selected good particle simulations (with low number of violations) by rigid-body superposition and clustering.

More recently, two Markov-chain Monte Carlo (MCMC)[40] sampling-based methods, MCMC5C[101] and BACH[51], were proposed to infer the 3D structures by maximizing the likelihood of the observed Hi-C data. Both methods assume that the expected contact frequencies and spatial distances among loci follow the power law distribution. MCMC5C models the observed frequency with Gaussian distribution with respect to the expected frequency. BACH models the observed frequency with Poisson distribution with respect to the expected frequency and takes the enzyme cutting site bias (e.g., CG content, mappability, fragment length) into account.

Although some works have been done, these 3D modeling methods have common crucial defeats. (1) Existing methods infer the 3D chromosomal structure by heuristics and there is no guarantee that their final outputted 3D model satisfies all the imposed constraints and the result of local search heavily depends on the starting point. Even in the noise-free case, they cannot recover the 100 percent correct structure. (2) The conversion between the contact frequency and spatial distance has some parameters. Existing methods, except BACH, assume that those parameters are

fixed or known beforehand. However, it is not true. The parameters are actually different for different datasets and it is important to have a method to estimate them.

Besides, high-throughput sequencing data is derived from a population of cells instead of single cell, and people argue that the predicted chromosome structure based on the high-throughput sequencing data cannot represent the chromosome structure in the individual cell. Hence, population based approach is introduced by Kalhor, et al. [61], which claims that the Hi-C (or TCC) data can be better fitted by learning a set of 3D structures (since the sample has multiple cells where the chromatin structures in different cells are different) instead of one single structure. Only based on a small set of nuclear landmarks and molecular volume constraints, Tjong, et al. [119] further showed that the population simulation result can reproduce a contact frequency matrix highly correlated with the contact frequency matrix derived from 4C experiment in yeast genome. This result points out that the dominating factor of global chromosomes organization is the physical property of chromosome and nuclear instead of chromatin-bound proteins.

II-4 Review Summary

Based on the above literature review, different research gaps for three bioinformatics problems related protein-DNA interaction can be summarized below:

For motif enrichment analysis using ChIP-seq data, the extracted window size around ChIP peak summit and the cut-off of motif scanning will affect the final result. It is debatable whether the users can choose the correct parameters by themselves. Moreover, almost all types of current enrichment analysis require a background

model or a set of background sequences. The correct background model or background sequences are also not easy to choose.

For *de novo* motif finding using ChIP-seq data, the prior knowledge used by existing program may only be true for the ChIPed TF motif, but may not be true for co-TF motifs. Moreover, different TFs may have different position preference and sequence rank preference. Hence, it is impossible to ask the user to provide one prior distribution to satisfy different potential co-TFs the user is interested in.

For 3D chromosome structure modeling using Hi-C data, most existing methods assume that the conversion function from contact frequency to spatial distance is known. However, it is theoretically problematic for this assumption. Moreover, none of the existing methods can guarantee recovering the correct 3D structure even in the error-free case, since all of them are based on local search or random sampling.

Accordingly, the main objectives of this study were to propose three practical algorithms to fill the gaps listed above, and they were:

- To develop a novel motif enrichment analysis method for ChIP-seq called CENTDIST. CENTDIST does not require the input of any user-specific parameters and background information. Instead, CENTDIST automatically determines the best set of parameters and ranks co-TF motifs based on their distribution around ChIP-seq peaks.
- To propose a novel motif finding algorithm called SEME, which uses unsupervised mixture model learning to learn the motif pattern (PWM), position preference and sequence rank preference at the same time, instead of asking users to provide them as inputs. It does not assume

the presence of both preferences but automatically detects them during the motif refinement process by statistical significance testing.

- To propose an elegant semi-definite programming (SDP) formulation to solve the 3D chromosome modeling problem. It is also important that the algorithm can guarantee to recover correct structure in the noise-free case and automatically choose the correct way to convert contact frequency to the spatial distance.

The thesis may have significant impact on the study of protein-DNA interaction at both the sequence level and the structure level. They also may open a door for people to better understand the potential of the new generation data like ChIP-seq and Hi-C. Nevertheless, this study assumes the reference genome is generally correct and the structure variant is not considered here. It is a valid assumption when the data are from normal cells. For cancer cells, there may be genome rearrangements for them, and most of problems can only be solved approximately in this scenario by assuming that the reference genome is unchanged.

CHAPTER - III CENTDIST: Motif Enrichment Analysis for ChIP-seq data

ChIP-seq is one of most important technology advance to study Protein-DNA interaction in vivo. This chapter describes CENTDIST, a motif enrichment analysis method to identify co-TF motifs for ChIP-seq data. CENTDIST takes advantage of the ChIP-seq property to improve the accuracy. This is a joint work with Chang Cheng Wei. Parts of the material covered in this chapter were originally published in [135].

III-1 Introduction

As mentioned in the review in Chapter-II Section II-1, the success of motif enrichment analysis highly depends on several aspects: 1. The background model (which represents the non-binding sites), 2. The enriched region size (which models the range between the co-TF and the ChIP peak summit), and 3. The motif/PWM score cut-off (which determines if a site can be matched the given motif/PWM or not). Moreover, different TFs may satisfy different parameters, and existing methods can only assign one set of parameters, which reduces the accuracy and sensitivity of existing methods. Therefore, it would be ideal to have a method that automatically determines the background and estimates the enriched region size as well as the PWM score cut-off for every candidate motif. The motif enrichment analysis problem for ChIP-seq data is defined in Section III-2. A new scoring measure called center distribution score is introduced in Section III-3, which is based on two histograms of motif distribution around the ChIP-seq peaks. A user-friendly and accurate motif

enrichment analysis tool CENTDIST is developed in Section III-4 that utilizes the center distribution score to detect co-TF motifs associated with the given ChIP-seq data. The performance of CENTDIST against two enrichment-based programs on 13 ChIP-seq datasets generated from mouse embryonic stem cells [16] is reported in Section III-5, which showed that CENTDIST was the best performer among the three programs and also provides useful additional information that helps users select the best co-TF candidates.

III-2 Problem Definition

Recent advances in ChIP-seq allow researchers to identify binding sites of the selected TF (ChIPed TF) in genome-wide scale. One open challenge is to identify the co-TFs of the ChIPed TF given a list of ChIP-seq peaks. Assuming the binding motifs of candidate co-TFs are known, one approach to this challenge is motif enrichment analysis (MEA). The motif enrichment analysis problem can be defined as following:

Given a set of ChIP-seq peak summit locations $\mathbf{P}=\{P_1, P_2, \dots, P_n\}$, a list of motif patterns $\{\theta_1, \theta_2, \dots, \theta_m\}$ of the candidate co-TFs, and the reference genome sequence \mathbf{G} . The problem is to compute the enrichment score for each motif pattern θ given \mathbf{P} and \mathbf{G} , so that the motif patterns of true co-TFs get higher score than other motif patterns. Then, all the candidate co-TFs are sorted from high to low according to their motif enrichment scores. Finally, the top rank candidates are classified as potential co-TF candidates and subsequently validated experimentally.

From the problem definition, we know that the key issue is how to design a good motif enrichment score. In the next section, a parameter-free motif enrichment score called center distribution score will be introduced.

III-3 Center Distribution Score

For each motif, a center distribution score is designed for ChIP-seq data. It includes two goodness measures. The first measure is called the frequency score, which is computed from the distribution of the motif occurrences with respect to the peak summit (frequency graph, Figure III-1(a)) under different enriched region size and PWM score cut-offs. An optimal set of parameters (enriched region size, PWM score cut-off) is also found that maximizes the frequency score (Equation (3.1)) given the ChIP-seq peaks.

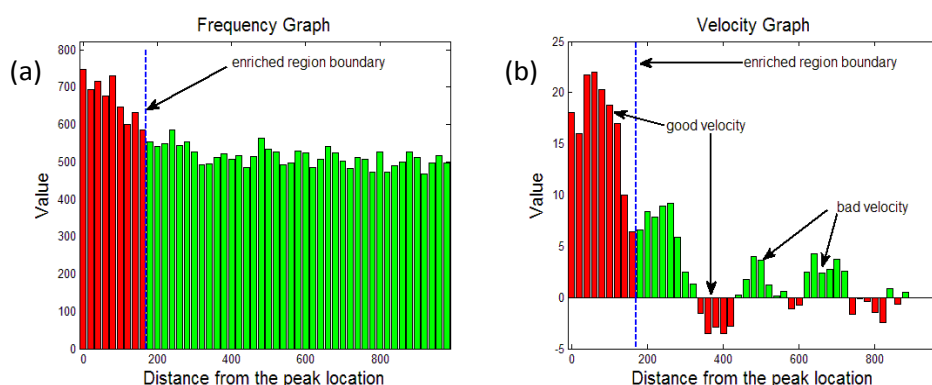


Figure III-1. **Frequency and velocity analyses of the AP4 motif.** (a) The frequency graph of the AP4 motif in an AR ChIP-seq dataset, (b) The velocity graph of the AR motif in an AR ChIP-seq dataset. In each graph, the dotted line partitions the distribution into the enriched region (left region) and the non-enriched region. The dotted line is determined by maximizing the frequency difference between the two regions.

The second measure is called the velocity score, which aims to correct the CG/AT bias in the peak regions. This score is derived from the velocity graph (Figure III-1(b)), i.e., the slope of the frequency graph. The final center distribution score for the given motif is the summation of the frequency score and the velocity score. In Section III-3.1, the details of generation for frequency graph and velocity graph are introduced. Next, the frequency score and the velocity score are described in Section

III-3.2 and III-3.3 respectively. Finally, Section III-3.4 gives the formal definition of center distribution score.

III-3.1 Generation of the frequency and velocity graphs

Firstly, the DNA sequences are extracted from the ± 1000 bp region of every peak in the ChIP-seq data. For each PWM motif Θ , we scanned all the extracted sequences and identified all occurrences whose PWM scores are above certain PWM score cutoff t . (PWM score is defined in Equation (1.1) in Chapter I)

From the list of occurrences of the motif Θ with matching scores higher than cut-off t , the frequency graph is constructed as follows. We partitioned every extracted sequence into 100 bins with respect to the distance to the peak summit, where each bin is of size 20bp. For $i=1, 2, \dots, 50$, let b_i be the number of occurrences of Θ in the range $[-20i, -20i+20]$ and $[20i-20, 20i]$ of all extracted sequences. The frequency graph presents the histogram of b_i for $i=1, 2, \dots, 50$ (Figure III-1(a)).

The velocity graph was obtained from the frequency graph. For every i , the velocity of enrichment is defined to be $b'_i = (b_i - b_{i+w})/w$, where w is a parameter to adjust the smoothness. In our implementation, we set $w=5$. The velocity graph presents the histogram of b'_i for $i=1, 2, \dots, 50-w$.

Given the motif enriched region (determined based on frequency graph, red region in Figure III-1(a)), the velocity is classified into two types as shown in Figure III-1(b). The type I velocity is colored red, including positive velocity inside the motif enriched region, and negative velocity outside the motif enriched region. The type II velocity is colored green, which includes negative velocity inside the motif enriched region, and positive velocity outside the motif enriched region. The motif enrichment

of velocity is computed by comparing the amount of these two types of velocity and will be introduced in Section III-3.3.

III-3.2 Z-score for frequency graph

Given a frequency histogram for a motif Θ under a PWM score cut-off t , the null hypothesis is that the histogram satisfies the uniform distribution, while the alternative hypothesis is that the first d bins are enriched. Let $|\Theta|_d$ be the total frequency of Θ in the first d bins in the histogram (i.e. $\sum_{i=1..d} b_i$) and $|\Theta|$ be the total frequency of Θ in all bins in the histogram. Under binomial distribution, the expected total frequency of Θ in the first d bins is $\frac{d \times |\Theta|}{50}$ and the standard deviation is $\sqrt{|\Theta| \times d/50 \times (1 - d/50)}$. Therefore, the frequency Z-score is

$$Z_{frequency,d}(\Theta, t) = \frac{(|\Theta|_d - \frac{d \times |\Theta|}{50})}{\sqrt{|\Theta| \times d/50 \times (1 - d/50)}} \quad (3.1)$$

The parameters include the enriched region size d and PWM score cut-off t will be chosen automatically for maximizing the frequency Z-score.

As shown Figure III-1(a), AP4 is a co-TF for AR, and the motif of AP4 (RNCAGCTG, IUPAC coding) occurs much more frequently near the center of the AR ChIP-seq peaks, when compared to the flanking regions. Thus, the AP4 motif would be considered as having a good frequency score and is a good candidate co-TF motif of AR.

III-3.3 Z-score for velocity graph

The frequency Z-score is a good measurement for the motif enrichment around the ChIP-seq peaks. However, there are occasions when noise (like CG/AT bias) could also be imbalancedly distributed around ChIP-seq peaks. Although such noise may be enriched, we expect it will not change dramatically near the center of ChIP-seq peaks compared to flanking regions. Therefore, to account for noise in the data, we include the velocity score. The velocity score is derived from a velocity graph of the co-TF motif (Figure III-1(b)), which is generated from the slope of the frequency graph (Figure III-1 (a)). If noise is assumed to change slowly (or linearly), the b'_i have similar values inside and outside the enriched region; otherwise, it will change dramatically near to the peaks as compared to the flanking regions. Specifically, the velocity score is a Z-score, which measures if the positive velocity increases dramatically.

Given a velocity histogram for a motif Θ under a PWM cutoff t , the null hypothesis is that the type I velocity and type II velocity uniformly distributed inside and outside the enriched region (assume enriched region is $[-d \times 20\text{bp}, d \times 20\text{bp}]$), while the alternative hypothesis is that the sum of type I velocity (represented by red bins) $A_{red,d}$ is larger than type II velocity (represented by green bins) $A_{green,d}$, where $A_{red,d}$ and $A_{green,d}$ are the area under red bins and green bins respectively. Under the null hypothesis, the expected area for each color is $(A_{red,d} + A_{green,d}) \times 0.5$, and the standard deviation is $\sqrt{(A_{red,d} + A_{green,d}) \times 0.5 \times (1 - 0.5)}$. Therefore, the Z-score of velocity graph with d enriched bins is

$$\begin{aligned}
Z_{velocity,d}(\theta, t) &= \frac{(A_{red,d} - (A_{red,d} + A_{green,d}) \times 0.5)}{\sqrt{(A_{red,d} + A_{green,d}) \times 0.5 \times (1 - 0.5)}} \\
&= \frac{(A_{red,d} - A_{green,d})}{\sqrt{(A_{red,d} + A_{green,d})}}
\end{aligned} \tag{3.2}$$

In short, by taking into consideration of the velocity distribution of motif occurrences (velocity graph), it will correct the frequency score biases due to CG (or AT) variation in the regions around the ChIP-seq peaks. For example, we observed a dramatic positive change in velocity (or slope) for the AP4 motif in the enriched region of the AR ChIP-seq peaks while the overall velocity remained small in the flanking region (Figure III-1(b)). In such instance, the AP4 motif would be classified as having a good velocity score.

III-3.4 Center distribution score for a motif distribution

The final scoring function used to assess motif distribution is called the center distribution score, which is the sum of two components: frequency score and velocity score. For a motif θ , the center distribution score $\tau(\theta)$ is defined as the sum of Z-scores for both the frequency distribution and the velocity distribution. Thus, we have:

$$\tau(\theta) = Z_{frequency,d}(\theta, t) + Z_{velocity,d}(\theta, t) \tag{3.3}$$

where the parameters d and t maximize $Z_{frequency,d}(\theta, t)$.

To assess the probability significance of the center distribution score, we compute the empirical p-value by assuming majority of the known motifs are not the co-TF motifs of the given ChIPed TF. The p-value of the center distribution score is computed as the tail probability of the given score assuming it comes from the normal distribution which is fitted using the scores of the lowest rank 80% of all motifs in TRANSFAC[85] database.

III-4 Implementation of CENTDIST

CENTDIST is a motif enrichment analysis (MEA) tool for ChIP-seq data, which requires minimal input from users. Unlike existing MEA methods, CENTDIST does not require any user-specific parameters. CENTDIST can automatically optimize the parameters like the enriched region size and the PWM score cut-off and computes the enrichment against the flanking regions. As a web-based MEA application, CENTDIST is fast, user-friendly, and capable of handling datasets with over a million ChIP-seq peaks.

Table III-1 **The pseudo code of CENTDIST Algorithm.**

Algorithm CENTDIST
INPUT: ChIP-seq peak locations P , Reference Genome G , Extracting Range L
Load a list of PWMs from TRANSFAC Database.
From the genome G , extract the list of sequences S from the regions $[-L/2, +L/2]$ of P .
FOR each PWM θ in the TRANSFAC database
Check each PWM threshold t from low to high and enriched region size d
Find $(t_{max}, d_{max}) = \max_t \max_{20 \leq d \leq L/4} Z_{frequency,d}(\theta, t)$
Compute the center distribution score $\tau(\theta) = Z_{frequency,d_{max}}(\theta, t_{max}) + Z_{velocity,d_{max}}(\theta, t_{max})$.
ENDFOR
Return the list of PWMs in decreasing order of center distribution scores.

The input of CENTDIST is a set of genomic locations representing ChIP-seq peaks (chromosome-peak summit position) and a list of candidate PWM motifs (provided by users or obtained from the TRANSFAC[85] or JASPAR[105] databases) representing co-TF binding sites. CENTDIST first extracts the sequences from the regions ± 1000 bp around the ChIP-seq peak locations. Next, CENTDIST scans the

sequences, obtains the initial occurrences of each PWM motif and searches the best enriched region size and PWM score cut-off to maximize the frequency Z -score for different frequency graphs. Then, CENTDIST computes the velocity Z -score using the velocity graph under the best enriched region size and PWM score cut-off. The center distribution score of each PWM motif is calculated as the sum of the two Z -scores. Finally, CENTDIST outputs the list of TF families ranked by the center distribution scores. The general algorithm of CENTDIST is shown in Table III-1.

Figure III-2 demonstrates how CENTDIST can promote true positive and repress false positive. To demonstrate the former, Figure III-2 (a) presents the motif occurrence of V\$AR_02 around AR ChIP-seq peaks. The enrichment score increased when we use flanking regions as background instead of promoter or random region. The V\$AR_02 motif enrichment progressively increases by identifying the optimal enriched region size, selecting the optimal PWM cut-off, and finally considering the velocity score. In contrast, to demonstrate the false positive repression, we examined the CG-rich yeast TF motif (F\$ADR1_01) in Pol2 (RNA polymerase II) ChIP-seq peaks in human K562 cells[94]. This CG-rich motif would have been determined incorrectly to be enriched around the Pol2 peak using traditional approaches due to CpG islands (i.e., regions known to contain many CG repeats) around Pol2 binding sites. As shown in Figure III-2(b), this motif has a modest center distribution score based on only the frequency score, but the final center distribution score was significantly lower after considering the velocity score.

III-5 RESULTS

To determine if CENTDIST can identify co-TFs better than existing enrichment based methods, we compared the performance of CENTDIST with two motif enrichment analysis (MEA) programs, CORE_TF[47] and CEAS[57]. We chose CORE_TF and CEAS for our comparisons because they were the only web-based programs we could find which can report enriched motifs from user-defined genomic regions at the time we were developing CENTDIST, while other web-based programs like Contra[49], oPOSSUM[48] and PASTAA[98], were limited to promoter regions only. For our comparisons, each program was optimized to their best performance.

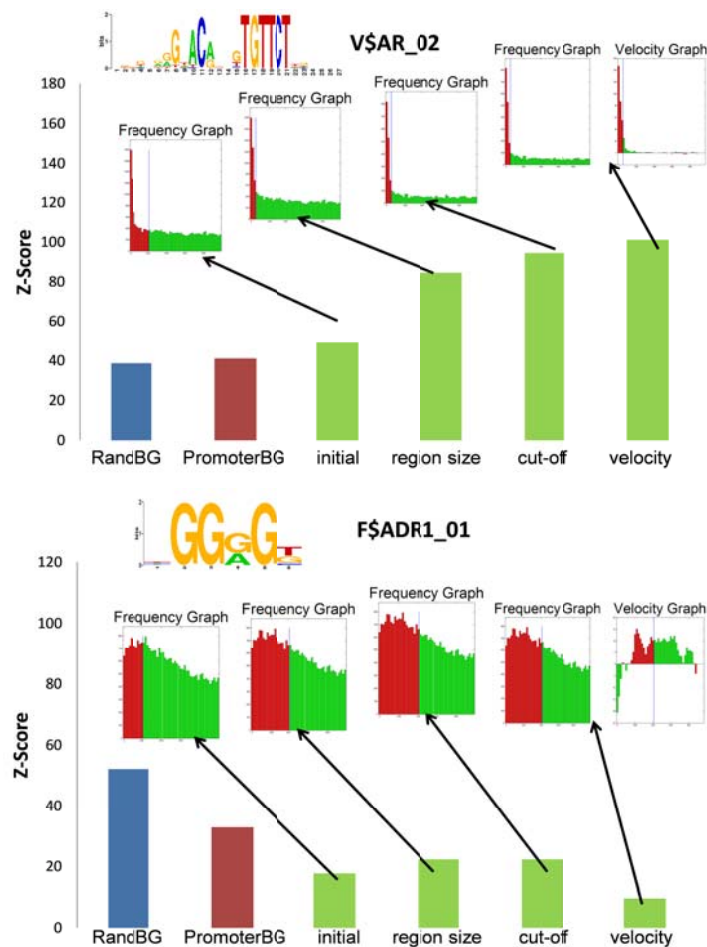


Figure III-2: **Demonstration of CENTDIST Capability.** (a) CENTDIST promotes the enrichment of AR motif in the AR ChIP-seq dataset (LNCaP cell line). The blue bar and red bar show the Z-scores of the AR motif computed using the traditional enrichment method under the default enriched region size of 200bp and the default PWM cut-off (1.32, FDR=0.0001) using random genome region background(blue) or promoter background(red) respectively. The green bars show the Z-score of the AR motif computed by CENTDIST after it optimized different parameters. (b) CENTDIST represses the enrichment of the false CG rich motif in the Pol2 ChIP-seq dataset. All Z-scores are computed exactly as in (a). Since CENTDIST considers the velocity graph of the false CG rich motif, the combined Z-score of CENTDIST finally drops and is significantly lower than that computed by the traditional enrichment based method. As a side note, this result also shows that random background (blue bar) can produce quite different results compared to promoter background(red bar), which highlights the difficulty of choosing a correct background in existing enrichment based methods.

Recently, 13 TF ChIP-seq maps were generated from mouse embryonic stem (ES) cells [16]. These 13 TFs were shown to cluster into two core transcriptional modules called MTLs (multiple transcription factor-binding loci), which can be highlighted (warm color) in Supp Table 1 by overlapping their ChIP-seq peaks. Because numerous co-TF relationships were discovered from the 13 factors, we decided to use these datasets for our comparisons of the three MEA programs. Only genomic locations of the ChIP-seq peaks and motifs from the TRANSFAC database were entered into CENTDIST. For CORE_TF and CEAS, input sequences with different enriched region size (± 100 bp, ± 200 bp and ± 500 bp) around the summit of the ChIP-seq peaks were extracted and different background settings were tested. The results from each program were compared against a table containing the co-TF motifs for each of the 13 ES TFs (Supp Table 2).

We assessed the performance of each program by the area under the receiver operating characteristic (ROC) curve (AUC)[45], which is computed as follow: For each ChIP-seq dataset, all tested programs ranked the same list of vertebrate

TRANSFAC motif families (Supp Table 3). With the pre-defined list of true co-TFs as a positive set (Supp Table 2) and the other motif families in the TRANSFAC as negative set, we then generate the ROC curve using the ranking list of the motif families reported by each program. Finally, the area under the ROC curve (AUC) was calculated using the trapezoid rule. The value of AUC ranges from 0 to 1 (a score of 0.5 is equivalent to random guessing).

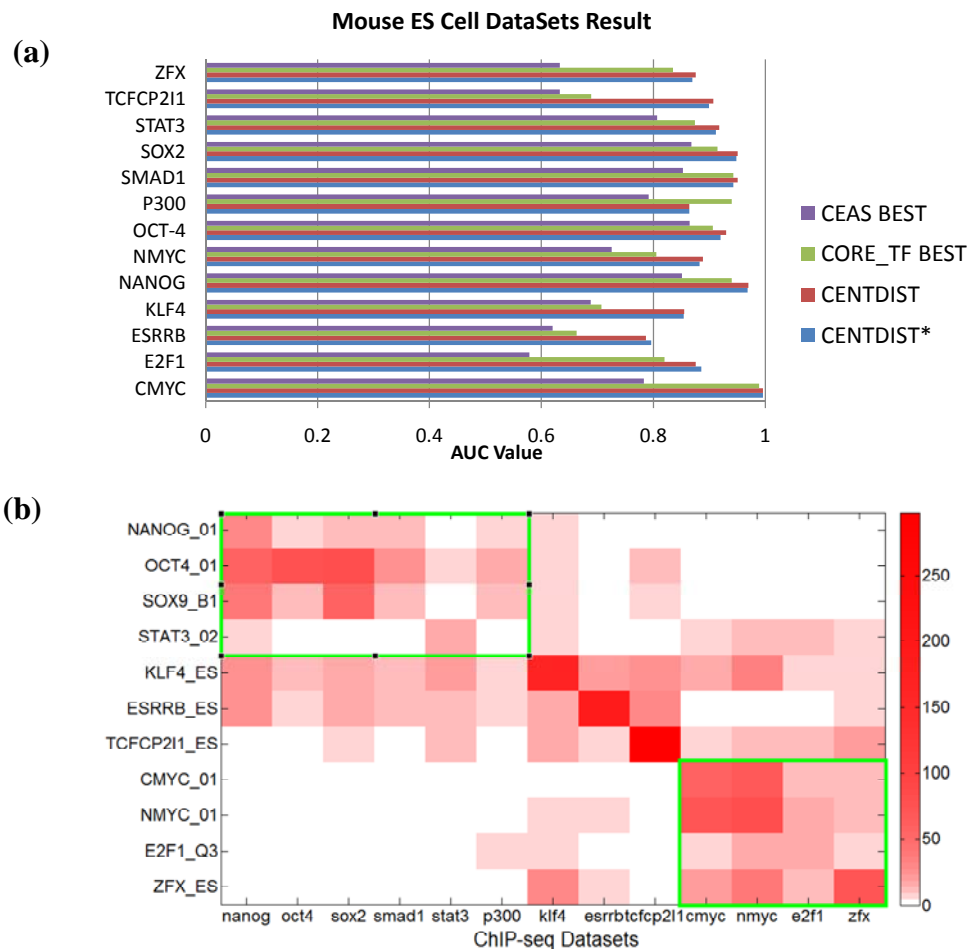


Figure III-3: **Co-TF motif analysis of 13 Embryonic Stem Cell TFs using CENTDIST, CEAS, and CORE_TF.** (a) A comparison of co-TF motif analysis results using CENTDIST, CORE_TF and CEAS on 13 different ChIP-seq datasets from ES cell. The best setting in each dataset for CORE_TF and CEAS were used for comparison. CENTDIST*=CENTDIST algorithm without the inclusion of velocity score. (b) Heat map representing the analysis of 11 ES cell core TFs motif enrichment in 13 ChIP-seq experiments. Every row

corresponds to a PWM motif while every column corresponds to a ChIP-seq dataset. The color of each entry presents the center distribution score (in log scale) of the motifs with respect to the peaks of the ChIP-seq dataset. The figure showed that the enhancer motifs are enriched in the enhancer ChIP-seq datasets (top left green rectangle) while the promoter motifs are enriched in the promoter ChIP-seq datasets (bottom right green rectangle).

Based on AUC scores, our results showed that CENTDIST significantly outperformed the best result from both CEAS and CORE_TF (Figure III-3(a) and Table III-2). We noticed that for CEAS and CORE_TF, different configurations led to different performances. This highlights the difficulty in selecting the appropriate parameters for co-TF motif analysis since no single set of parameters can be considered the best for each ChIP-seq dataset. CENTDIST, which requires neither background nor other parameter settings, performed significantly better (average AUC score=0.905) than the best configuration of CEAS (average AUC score=0.740) or CORE_TF (average AUC score=0.84084). Furthermore, we compared the results of CENTDIST with the results ranked by frequency score only (denoted as CENTDIST* in Figure III-3(a) and Table III-2). Although the AUC score changes may not very significant, we found that CENTDIST was consistently better than CENTDIST* in 11 out of 13 experiments and there is no parameter tuning cross these testing datasets, which indicates the velocity score can improve the motif ranking result.

Next, we examined the center distribution scores of 11 ES TF motifs (Smad1 and p300 do not have known motif) across 13 TF ChIP-seq datasets (Figure III-3(b)). From this analysis, we clearly saw two functional groups: the enhancer motifs (Oct4, Sox2, Nanog and Stat3) have good center distribution score in the enhancer TF ChIP-seq datasets (top left green box), while the promoter motifs (cMyc, nMyc, Zfx, and

E2f1) have good center distribution score in the promoter TF ChIP-seq datasets (bottom right green box) (Figure III-3(b)). Their frequency graphs are shown in Figure III-4. These results are in agreement with the previous findings [16]. Moreover, enhancer motifs did not show good center distribution in the promoter ChIP-seq datasets, and vice versa. The only exception was Stat3, which was classified as an enhancer TF but had good center distribution at the promoter. However, a recent report showed that Stat3 was also enriched in the promoter regions of ES cells, suggesting Stat3 can be located at both promoter and enhancer regions[64]. In short, the results from this large-scale comparison demonstrate that center distribution is a good statistical model for predicting the occurrences of co-TF motifs from ChIP-seq data.

Table III-2: **Comparison of CENTDIST, CEAS, and CORE_TF for different ChIP-seq datasets.** * The output result of CENTDIST* is ranked by the Z-score of frequency graph only. The columns 4th-6th are the results for CORE_TF using promoter background (promBG,default background for CORE_TF) with enriched region size 200-1000 respectively, and the column 7th-9th are the result of CORE_TF using random genome background(randBG) with enriched region size 200-1000 respectively. The last three columns are the results of CEAS with enriched region size 200-1000 respectively.

	CENTDIST*	CENTDIST	CORE_TF promBG 200	CORE_TF promBG 400	CORE_TF promBG 1000	CORE_TF randBG 200	CORE_TF randBG 400	CORE_TF randBG 1000	CEAS 200	CEAS 400	CEAS 1000
CMYC	0.9957	0.9957	0.9892	0.9742	0.9355	0.9742	0.9505	0.9097	0.7731	0.7828	0.5806
E2F1	0.8860	0.8761	0.8202	0.7966	0.7758	0.8076	0.7862	0.7303	0.5789	0.5625	0.5746
ESRRB	0.7961	0.7869	0.6373	0.6627	0.6065	0.5359	0.5451	0.6183	0.6203	0.6072	0.6111
KLF4	0.8542	0.8550	0.7075	0.7058	0.6908	0.7058	0.6950	0.6813	0.6708	0.6883	0.6021
NANOG	0.9686	0.9699	0.9320	0.9399	0.9020	0.9255	0.9046	0.8327	0.8386	0.8510	0.7268
NMYC	0.8824	0.8889	0.8052	0.7915	0.7627	0.7922	0.7719	0.7418	0.7255	0.6137	0.6039
OCT4	0.9200	0.9300	0.8767	0.8908	0.9067	0.8625	0.8342	0.7900	0.8650	0.8175	0.8017
P300	0.8646	0.8646	0.9397	0.9364	0.8657	0.8860	0.8169	0.7270	0.7917	0.7741	0.6184
SMAD1	0.9430	0.9507	0.9430	0.9287	0.8520	0.9364	0.9167	0.8191	0.7906	0.8531	0.7007
SOX2	0.9485	0.9507	0.9035	0.9068	0.8947	0.9145	0.8969	0.8235	0.8531	0.8448	0.8684
STAT3	0.9117	0.9175	0.8742	0.8525	0.7875	0.7892	0.7275	0.7300	0.8067	0.7513	0.7546
TCFCP2I1	0.8993	0.9072	0.6889	0.6719	0.5386	0.6627	0.6484	0.6641	0.6333	0.6144	0.6105
ZFX	0.8693	0.8758	0.8353	0.8248	0.7732	0.8288	0.8013	0.7190	0.6327	0.5137	0.5137
Average AUC	0.9030	0.9053	0.8425	0.8371	0.7917	0.8170	0.7919	0.7528	0.7369	0.7134	0.6590

III-5.1 Conclusion

This chapter introduces a parameter-free motif enrichment analysis (MEA) tool for ChIP-seq data called CENTDIST. Using CENTDIST, the biologists can easily identify the co-TFs with known motifs of the ChIPed TF. The existing MEA methods are heavily dependent on selecting the proper background and other parameter settings. In contrast, CENTDIST does not require an explicit background model and optimizes the parameters automatically based on the frequency information as well as slope information (velocity) of motif distribution. As a user-friendly web-based application, CENTDIST is capable of analyzing large-scale ChIP-seq datasets. It can test approximately seven hundred TRANSFAC motifs over 10,000 ChIP-seq peaks in only 10 minutes. The output of CENTDIST contains clean and rich information for users. Specifically, it groups the list of enriched motifs into TF families, and provides other information such as PWM logo, motif distribution graph, enrichment P-value, and the enriched region size of the enriched motifs. We examined CENTDIST on 13 ES cell ChIP-seq datasets and demonstrated that it is better than existing methods. We also showed that this could be achieved without requiring expert knowledge in configuring the program. More other biological interesting results found by CENTDIST can be referred in several other publications[66, 114, 142, 107, 39, 116].

CENTDIST does have certain limitations. CENTDIST assumes the co-TFs follow the proximity assumption (i.e. the occurrences of the co-TFs are over-represented near to the binding sites of the ChIPed TF). Although the proximity assumption is generally true, there is also possibility that the co-TF would keep a

certain distance to the ChIPed TF. For this case, CENTDIST may fail, and the alternative program like SpaMo[129] would be helpful, which can identify the co-TF motif with fix distance from the ChIPed TF motif. Also, the reader should be noted that the list co-TFs used in the study are derived from the literature and are not complete, and some of them were identified by traditional motif enrichment methods in the first place. All these limitation may affect the assessment of different motif enrichment methods.

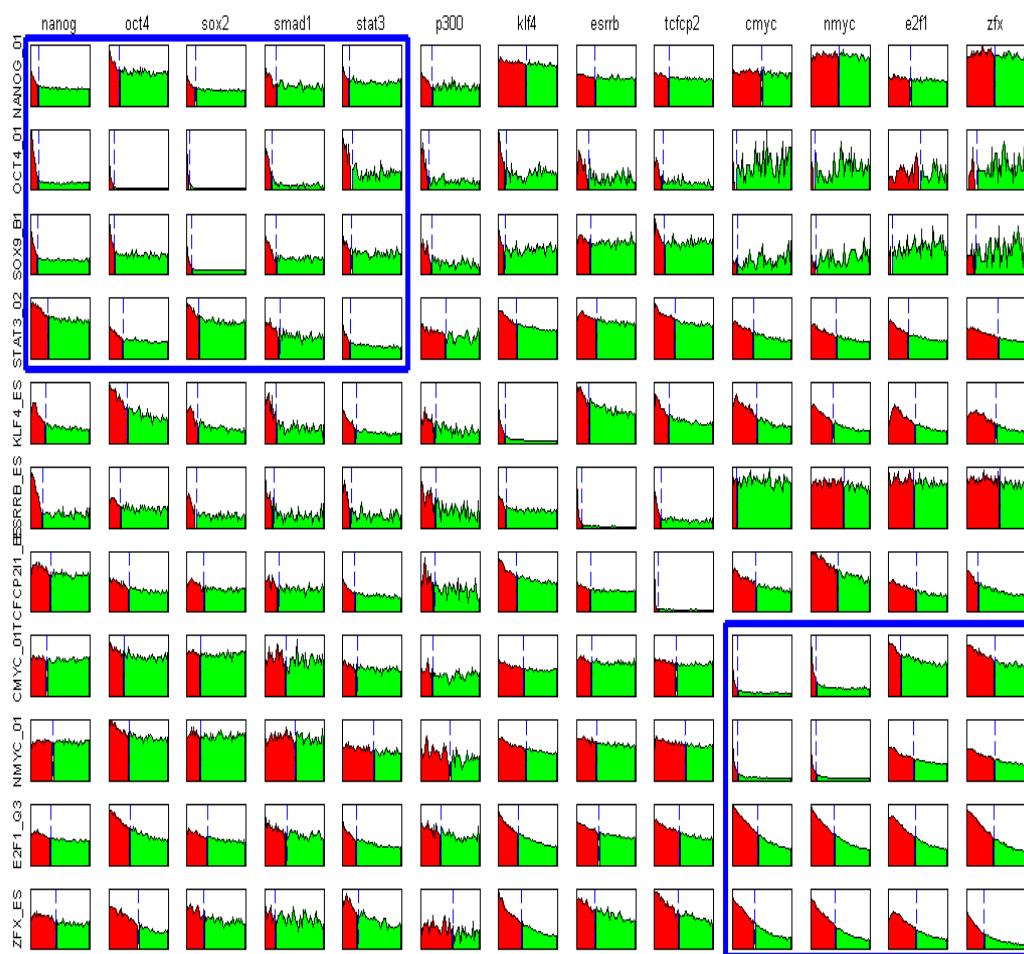


Figure III-4: **Frequency analysis of ES cell TFs.** Every row corresponds to a PWM motif while every column corresponds to a ChIP-seq dataset. Each entry shows the frequency graph of the motif with respect to the peaks of the ChIP-seq dataset. Each graph shows the center enrichment region in red color and flanking enrichment region in green color. We observed that the frequency graphs in the top left blue rectangle show center enrichment while the frequency graph in

the bottom right rectangle shows center enrichment. All motifs are extracted from TRANSFAC database except the ones with suffix “_ES” which are the de novo motifs from[16].

CHAPTER - IV Simultaneously Learning DNA Motif along with Its Position and Sequence Rank Preferences through EM Algorithm

In the last chapter, motif enrichment analysis for ChIP-seq data was introduced, which assuming the motifs of co-TFs is known. This chapter describes SEME, a *de novo* motif finding method to identify novel motifs for ChIP-seq data.

IV-1 Introduction

As mentioned in the review in Chapter-II Section II-2, *de novo* motif finding is an important classical bioinformatics problem. However, by only examining the over-representation of sequence patterns, the previous generation motif finders often miss some real motifs and generate many false positives. On the other hand, additional information for the input sequences is found to be helpful to improve motif finding. For example, some transcription factor (TF) binding motifs (e.g. TATA-box) are localized to certain intervals with respect to the transcription start site (TSS) of the gene. In this case, the position information can help to filter spurious sites. In protein binding microarray (PBM) [12] data, the *de Bruijn* sequences are ranked by their binding affinities and people expect the correct motif occurs in the high ranking sequences; such data has a rank preference. In the ChIP-seq data [121], the ChIPed TF's motif (ChIPed TF is the TF pulled down in the ChIP experiment) prefers to occur in sequences with high ChIP intensity and also near the ChIP peak summits (thus having both position and rank preference). Hence, if we know the position preference and the sequence rank preference of the TF motifs in the input sequences, we can improve motif finding. In fact, many existing motif finders already utilize such additional information. MDscan [80] only considers high ranking sequences to

generate its initial candidate motifs. Other programs allow users to specify the prior distribution of position preference or sequence rank preference [6, 52] by adding a prior distribution component in their scoring functions [89, 78, 36]. However, the users may not know the correct prior(s) to begin with. Even worse, different motifs may have different preferences. For example, in ChIP-seq experiments, some motifs prefer to occur in high ranking sequences and at the center of the ChIP peak summit while others do not.

To resolve such problem, we proposed a novel motif finding algorithm called SEME (Sampling with Expectation maximization for Motif Elicitation). SEME assumes the set of input sequences is a mixture of two models: a motif model and a background model. It uses EM-based algorithm to learn the motif pattern (PWM), position preference and sequence rank preference at the same time; instead of asking users to provide them as inputs. SEME does not assume the presence of both preferences but automatically detect them during the motif refinement process through statistical significance testing.

We also observed that EM algorithms are generally slow in analyzing large-scale high-throughput data. Speeding up EM using suffix tree was recently proposed [96] but the technique cannot be applied when one wants to also learn the position and sequence rank preferences. To improve the efficiency, SEME develops two EM procedures. The two EM procedures are based on the observations that the correct motifs usually have a short conserved pattern in it and majority of the sites in the input sequences are non-motif sites. The first EM procedure, called extending EM (EEM), starts by finding all over-represented short l -mers and then attempts to include and refine the flanking positions around the l -mers within the EM iterations. This way,

SEME recovers the proper motif length within a single run thus saving a substantial amount of time by avoiding multiple runs with different motif length (as done in other existing motif finders [78, 52]). The second EM procedure, called the re-sampling EM (REM), tries to further refine the motif produced by EEM. It is based on a theorem similar to importance sampling [41], which stated that the motif parameters can be learned unbiasedly using a biased subsampling. By this principle, we can sample more sites which are similar to the EEM's motif and fewer sites from the background. This way, REM is able to learn the correct motifs using significantly less background sites. In our implementation, REM is capable to produce the correct TF motifs using approximately 1% of the sites normally considered in a normal EM procedure.

Using 75 large scale synthetic datasets, we showed that SEME was better both in terms of accuracy and running time when compared to MEME, a popular EM-based motif finding program [6]. We found that MEME was unable to find motifs with gap regions while SEME's EEM procedure can successfully extend the motifs to include them. In the real experimental datasets, we performed comparison using 32 metazoan compendium datasets and 164 ChIP-seq libraries. SEME consistently outperformed seven existing motif finders. In general, we found that SEME not only found more TF motifs but also gave more accurate results (as evaluated using either PWM divergence, AUC score or STAMP's p-value [84]). When we compared the programs to find co-TF motifs from 15 ChIP-seq datasets, the superior performance of SEME was more pronounced. It indicates that SEME's ability to learn the underlying motif binding preference is crucial in its performance. We further confirmed the correctness of the position and sequence rank preference of the co-TF

motifs learned by SEME on three ChIP-seq datasets. The actual ChIP-seq data of the predicted co-TFs clearly shows that SEME managed to infer the correct preferences. We also showed that such preferences provided biological insights on the mechanism of the ChIPed TF—co-TF interactions.

IV-2 SEME Algorithm

IV-2.1 Review of Mixture Model for Motif Finding

Applying mixture model to learn motifs in a set of sequences is first proposed by MEME[6], it assumes the observed sequences are generated by two independent components: motif model and background model. Let the alphabet be $\{A, C, G, T\}$ for DNA sequence. The background model is a zero order markov model $\vec{\theta}_0 = (\theta_{0,A}, \theta_{0,C}, \theta_{0,G}, \theta_{0,T})$ where $\theta_{0,b}$ is the probability of observing nucleotide $b \in \{A, C, G, T\}$. The motif model describes a length- l sequence as l independent positions. It is represented as Θ , which is a $l \times 4$ matrix where $\Theta_{j,k}$ is the probability that the nucleotide a_k occurs at position j . For any length- l sequence X_i , the probability that X_i is generated from the motif model and the background model can be computed as follows.

$$Pr(X_i | Z_i = 1) = Pr(X_i | \Theta) = \prod_{j=1}^l \prod_b \Theta_{j,b}^{I(b, X_{i,j})} \quad (4.1)$$

$$Pr(X_i | Z_i = 0) = Pr(X_i | \vec{\theta}_0) = \prod_{j=1}^l \prod_b \theta_{0,b}^{I(b, X_{i,j})} \quad (4.2)$$

Where $X_{i,j}$ is the letter in the j -th position of sample X_i and $I(x,y)$ is an indicator function which is 1 if and only if $x=y$.

Any set of sequences can be conceptually splitted into a set $X = \{X_1, X_2, \dots, X_n\}$ of n overlapping subsequences of length l . MEME assumes those length- l subsequences in X are extracted from a mixture of motif model Θ and a background model $\vec{\theta}_0$, where λ ($0 < \lambda < 1$) is the parameter which defines the prior probability of X_i generated by motif model. The probability framework of the mixture model is defined as follows:

$$Pr(X) = \prod_{i=1}^n (\lambda Pr(X_i | Z_i = 1) + (1 - \lambda) Pr(X_i | Z_i = 0)) \quad (4.3)$$

Then MEME formulated the motif finding problem as an optimization problem, which finds a set of parameters $(\lambda, \vec{\theta}_0, \Theta)$ maximizing the likelihood of data $Pr(X)$. This optimization problem is NP-hard. EM algorithm is a state of the art method to solve this maximum likelihood problem. The EM algorithm makes use of the concept of missing data. In this case, the missing data Z_i is the knowledge of whether X_i is coming from a motif model. $Z_i = 1$ if X_i is from motif model; and $Z_i = 0$ otherwise. Also by definition, $Pr(Z_i = 1) = \lambda$. The objective function of EM can be revised as a “complete log likelihood function”:

$$\begin{aligned} \log Pr(X, Z | \lambda, \vec{\theta}_0, \Theta) &= \sum_{i=1}^n (Z_i \log(\lambda Pr(X_i | \Theta)) \\ &\quad + (1 - Z_i) \log((1 - \lambda) Pr(X_i | \vec{\theta}_0))) \end{aligned} \quad (4.4)$$

The EM algorithm iteratively maximizes the expected log likelihood over the conditional distribution of missing data Z_i given the current estimation of parameters $(\lambda, \vec{\theta}_0, \Theta)$. In the E-step, the expected value of Z_i in the iteration t can be computed as:

$$Z_i^{(t)} = \frac{\eta_i^{(t)}}{1 + \eta_i^{(t)}} \quad (4.5)$$

where η_i is likelihood ratio between the motif model and the background model.

$$\eta_i^{(t)} = \frac{\lambda^{(t-1)} Pr(X_i | \Theta^{(t-1)})}{(1 - \lambda^{(t-1)}) Pr(X_i | \bar{\theta}_0^{(t-1)})} \quad (4.6)$$

In the M-step, the parameters are estimated to maximize the expected log likelihood function given the expected value $\{Z_i\}$ in the last iteration:

$$\{\lambda^{(t)}, \bar{\theta}_0^{(t)}, \Theta^{(t)}\} = \arg \max_{\lambda, \bar{\theta}_0, \Theta} E_{Z^{(t)}} \left[\log Pr(X, Z^{(t)} | \lambda, \bar{\theta}_0, \Theta) \right] \quad (4.7)$$

and we can compute the explicit formulas for each parameter.

$$\lambda^{(t)} = \sum_{i=1}^n \frac{Z_i^{(t)}}{n} \quad (4.8)$$

For $b \in \{A, C, G, T\}$,

$$\theta_{0,b}^{(t)} = \frac{\sum_{i=1}^n \sum_{j=1}^l (1 - Z_i^{(t)}) I(b, X_{i,j})}{\sum_{i=1}^n \sum_{j=1}^l (1 - Z_i^{(t)})} \quad (4.9)$$

For $b \in \{A, C, G, T\}$ and $j = 1 \dots l$,

$$\Theta_{j,b}^{(t)} = \frac{\sum_{i=1}^n Z_i^{(t)} I(b, X_{i,j})}{\sum_{i=1}^n Z_i^{(t)}} \quad (4.10)$$

IV-2.2 Mixture Model in SEME

In SEME implementation, we consider two more binding preferences: position and sequence rank in addition to DNA sequence preference information in the traditional EM algorithm. The position preference tries to model if the binding site prefers certain positions. We discretize the positions into K bins. The probability that

a binding site occurs in the k -th position bin is denoted as α_k , for $k=1, \dots, K$, while the background distribution is assumed to be uniform. Precisely, for every X_i , we have :

$$Pr(X_i^{(pos)} = k | Z_i = 1) = \alpha_k; \quad Pr(X_i^{(pos)} = k | Z_i = 0) = \frac{1}{K} \quad (4.11)$$

Similarly, the sequence rank preference tries to model if the binding site prefers the sequences with certain range of ranks assuming input sequences are sorted by some measurement. We also discretize the ranks into K bins. The probability a binding site occurs in the k -th rank bin is denoted as β_k , $k=1, \dots, K$, while the background distribution is assumed to be uniform. Precisely, for every X_i , we have :

$$Pr(X_i^{(rank)} = k | Z_i = 1) = \beta_k; \quad Pr(X_i^{(rank)} = k | Z_i = 0) = \frac{1}{K} \quad (4.12)$$

We use naive bayesian approach to model three types of information (sequence, position, rank):

$$Pr(X_i | Z_i) = Pr(X_i^{(seq)} | Z_i) Pr(X_i^{(pos)} | Z_i) Pr(X_i^{(rank)} | Z_i) \quad (4.13)$$

where the probability of sequence information for bound state and unbound state $Pr(X_i^{(seq)} | Z_i)$ can be referred to Equations (4.1) and (4.2).

Similar to Equation (4.7), the ‘‘complete log-likelihood function’’ with additional binding preferences can be modified as follow:

$$\begin{aligned} \log Pr(X, Z | \Phi) = & \sum_{i=1}^n (Z_i \log(\lambda \prod_{j=1}^l (\prod_b \theta_{j,b}^{I(b, X_{i,j}^{(seq)})}) \alpha_{X_i^{(pos)}} \beta_{X_i^{(rank)}})) \\ & + (1 - Z_i) \log((1 - \lambda) (\prod_{j=1}^l \prod_b \theta_{0,b}^{I(b, X_{i,j}^{(seq)})}) / K^2)) \end{aligned} \quad (4.14)$$

Where $\Phi = (\lambda, \Theta, \vec{\theta}_0, \alpha_1, \dots, \alpha_K, \beta_1, \dots, \beta_K)$ are the parameters of mixture model in SEME.

Similarly, EM algorithm can be applied to optimize Equation (4.14). In the E-step, the likelihood ratio between the motif model and the background model is:

$$\eta_i^{(t)} = \frac{\lambda \left(\prod_{j=1}^l \prod_{b \in \{A,C,G,T\}} (\Theta_{j,b}^{(t-1)})^{I(b, X_{i,j}^{(seq)})} \right)}{1 - \lambda \left(\prod_{j=1}^l \prod_{b \in \{A,C,G,T\}} (\theta_{0,b}^{(t-1)})^{I(b, X_{i,j}^{(seq)})} \right)} \left(\frac{\alpha_{X_i^{(pos)}}^{(t-1)}}{K} \right) \left(\frac{\beta_{X_i^{(rank)}}^{(t-1)}}{K} \right) \quad (4.15)$$

and the expected value of $Z_i^{(t)}$ can be computed using Equation (4.5) .

In the M-step, the parameters $\Phi = (\lambda, \Theta, \vec{\theta}_0, \alpha_1, \dots, \alpha_K, \beta_1, \dots, \beta_2)$ can be estimated by maximizing the expected log likelihood function given the expected value Z_i in the last iteration.

$$\Phi^{(t)} = \arg \max_{\Phi} E_{Z^{(t)}} \left[\log Pr(X, Z^{(t)} | \Phi) \right] \quad (4.16)$$

The parameters $(\lambda, \Theta, \vec{\theta}_0)$ can be updated using Equations (4.8), (4.9) and (4.10), respectively. The additional parameters for position preference and sequence rank preference can be updated as follows:

$$\forall k \in \{1, \dots, K\}, \alpha_k^{(t)} = \frac{\sum_{X_i \in X} Z_i^{(t)} I(k, X_i^{(pos)})}{\sum_{X_i \in X} Z_i^{(t)}} \quad (4.17)$$

$$\forall k \in \{1, \dots, K\}, \beta_k^{(t)} = \frac{\sum_{X_i \in X} Z_i^{(t)} I(k, X_i^{(rank)})}{\sum_{X_i \in X} Z_i^{(t)}} \quad (4.18)$$

Above is the general probabilistic framework of SEME by applying classic EM algorithm. However, it cannot achieve good efficiency and accuracy for practical use if we directly apply classic EM algorithm to solve this problem. Hence, we developed four phases in the SEME pipeline (see Figure IV-1). To search for a good starting point, SEME first enumerates a set of over-represented short l -mers (phase 1) and extends each short l -mer to a proper length PWM motif by the extending EM (EEM) procedure (phase 2). The PWM reported by the extending EM procedure can

approximate the true motif when its starting l -mer captures the conserved region of the motif. To further refine EEM's PWM motif, SEME applies the re-sampling EM (REM) procedure (phase 3). It is an importance sampling version of the classical EM algorithm which greatly speeds up the EM iterations. Finally, the refined PWM motifs are scored and filtered for redundancies (phase 4).

SEME Pipeline
Require: A set of input DNA sequences (fasta format)
Ensure: Return a set of non-redundant motifs M

- 1: Identify a set of over-represented short l -mers Q in X ;
- 2: **for** every $q \in Q$ **do**
- 3: Extend q to full length PWM motif Θ^{EEM} using extending EM procedure;
- 4: Refine Θ^{EEM} to a more accurate PWM Θ^{REM} using re-sampling EM procedure;
- 5: Add Θ^{REM} to candidate motif set M ;
- 6: **end for**
- 7: Compute empirical scores (AUC or Z-score) for all the PWMs in M ;
- 8: Sort all the PWMs in M and filter lower scoring redundant PWMs in M ;

Figure IV-1 **Algorithm description for SEME Pipeline.**

IV-2.3 Identifying Over-represented l -mers

In the first phase, SEME computes the frequencies of all short l -mers ($l = 5$ by default) in the input sequences, and also their frequencies in background if control sequences or background model are provided. Then, all short l -mers with higher frequencies in the input sequences than the background are outputted to the next phase for further processing. If no background or control sequences are provided, 1st-order markov model is estimated from the input sequences as the background model.

IV-2.4 Extending EM Procedure

The classic EM algorithm does not allow varying the length of PWM within the EM iteration. Assume we know that the motif contains a conserved short l -mer seed q (obtained from the first phase), this section developed the extending EM (EEM)

which can extend the length- l seed while maximizes the likelihood of the observed data. We assume the maximum length of the motif is W_{max} . From the set of input sequences, SEME extract a set of length- $(2W_{max}-|q|)$ sequences $Y = \{X_i \in X | X_i^{(seq)} \text{ match } (N)^{W_{max}-|q|} q (N)^{W_{max}-|q|}\}$ (“N” is a wild char for A,C,G,T), whose middle part is q . For example, if the l -mer is “GGTCA” and the longest possible motif length is 10, Y are all the sites matching string pattern “NNNNNGGTCA>NNNNN”. By the definition of Y , we can consider all potential binding sites which contain the short conserved l -mer q with the length less than W_{max} .

Extending EM

Require: l -mer q , maximum allowed motif length w , input sequences X

Ensure: final extended PWM $\Theta^{(t)}$

- 1: $Y := \{X_i \in X | X_i^{(seq)} \text{ matches } (N)^{w-|q|} q (N)^{w-|q|}\};$
- 2: Initialize the parameter set $\Phi^{(0)}$ for the mixture model;
- 3: $t:=1;$
- 4: **repeat**
- 5: E-step: $\forall X_i \in Y$, compute the expectation of $Z_i^{(t)}$ using the parameter set $\Phi^{(t-1)}$ in the last iteration;
- 6: M-step: update the parameter set $\Phi^{(t)}$ by maximizing log likelihood $Pr(Y, Z^{(t)} | \Phi);$
- 7: **if** length of $\Theta^{(t)} < w$; **then**
- 8: Find a position j which maximizes the log likelihood increment in Equation 5 and denote J to be the corresponding nucleotide distribution of position j ;
- 9: **if** J is significantly different from the background distribution $\vec{\theta}_0^{(t)}$ using Chi-square test; **then**
- 10: Use J as the distribution in position j of PWM $\Theta^{(t)}$;
- 11: **end if**
- 12: **end if**
- 13: $t:=t+1;$
- 14: **until** PWM $\Theta^{(t)}$ converges;
- 15: The columns representing the l -mer q in $\Theta^{(t)}$ are diluted;

Figure IV-2: **Pseudo code for Extending EM procedure.**

Similar to the classic EM algorithm, we firstly define a wide PWM model Θ is a $(2W_{max}-|q|) \times 4$ matrix (may contain non-binding site positions, but is wide enough to cover all the potential binding site positions). Let a background model be $\vec{\theta}_0$, and

two variables l_1, l_2 index the real binding site start and end positions in Θ . In each EM iteration, a subset of columns in the wide PWM Θ will be used to compute the expectation, and a new column is included only if it can increase the likelihood in the M-step and show significant difference to background distribution. Let $\Theta_{[l_1, l_2]} = \{\Theta_{l_1}, \Theta_{l_1+1}, \dots, \Theta_{l_2}\}$, the computation for modeling sequence information will be carried on a subset of position in the sites, that is, the positions outside of $[l_1, l_2]$ will not be used and the positions for the given l -mer also will not be used because these l -mer positions are the same across Y . Here, we have,

$$\forall X_i \in Y$$

$$Pr(X_i^{(seq)} | \Theta_{[l_1, l_2]}) = \prod_{j=l_1, \dots, l_2 \wedge j \notin [W_{max}-|q|, W_{max}-1]} \prod_b (\Theta_{j,b})^{I(b, X_i \cdot j)} \quad (4.19)$$

$$Pr(X_i^{(seq)} | \bar{\theta}_0) = \prod_{j=l_1, \dots, l_2 \wedge j \notin [W_{max}-|q|, W_{max}-1]} \prod_b (\theta_{0,b})^{I(b, X_i \cdot j)} \quad (4.20)$$

Besides, the position model and sequence rank model remain the same as Equations (4.17) and (4.18). Then, we define each iteration of extending EM procedure as follow:

In the E-step, similar to Equation (4.15) and (4.5), for all $X_i \in Y$, we compute the likelihood ratio $\eta_i^{(t)}$ and $Z_i^{(t)}$ as,

$$\eta_i^{(t)} = \frac{\lambda^{(t-1)} Pr(X_i^{(seq)} | \Theta_{[l_1, l_2]}^{(t-1)}) \alpha_{X_i^{(pos)}}^{(t-1)} \beta_{X_i^{(rank)}}^{(t-1)}}{(1 - \lambda^{(t-1)}) Pr(X_i^{(seq)} | \bar{\theta}_0^{(t-1)}) / K^2} \quad (4.21)$$

$$Z_i^{(t)} = \frac{\eta_i^{(t)}}{1 + \eta_i^{(t)}} \quad (4.22)$$

In the M-step, the modeling parameters $(\lambda, \Theta, \vec{\theta}_0, \alpha_1, \dots, \alpha_K, \beta_1, \dots, \beta_2)$ are updated using Equations (4.8),(4.10),(4.9),(4.17) and (4.18), respectively, which are the exactly the same as the original EM algorithm except considering the selected sites Y instead of all the sites X . Moreover, the two indexing variables l_1, l_2 will also be updated in this step by trying to select a column outside $[l_1, l_2]$ to maximize the log likelihood objective function. Precisely, for each position $j = l_1, \dots, 2W_{max} - |q|$ not in $[l_1, l_2]$, we show that the maximum increment of the log likelihood before and after including the position j is $G(j)$ where

$$G(j) = \sup_J \sum_{X_i \in Y} Z_i^{(t)} \log\left(\frac{Pr(X_{i,j}^{(seq)} | J)}{Pr(X_{i,j}^{(seq)} | \theta_0^{(t)})}\right) \quad (4.23)$$

where J is any probability distribution over the nucleotides $\{A,C,G,T\}$.

In a greedy manner, the extending EM procedure chooses the column j with the largest $G(j)$, To avoid over-fitting, we require the selected column is (Chi-square) significantly different from the background frequency. Let $p = \arg \max_j G(j)$, the Chi-square statistics χ is defined as:

$$\chi = \sum_b \frac{\sum_{X_i \in Y} I(b, X_{i,j}^{(seq)}) Z_i^{(t)} (\Theta_{p,b}^{(t)} - \theta_{0,b}^{(t)})^2}{\sum_{X_i \in Y} I(b, X_{i,j}^{(seq)}) Z_i^{(t)} \theta_{0,b}^{(t)}} \quad (4.24)$$

then l_1, l_2 can be updated as $l_1^{(t)} = \min(l_1^{(t-1)}, p)$ and $l_2^{(t)} = \max(l_2^{(t-1)}, p)$, if and only if the Chi-square test is significant. The pseudo code of the extending EM procedure is described in Figure IV-2. The EEM procedure ends when PWM Θ converges. Finally, the columns in Θ representing the l -mer q will be further diluted (by setting all $[1.0, 0.0, 0.0, 0.0]$ columns representing "A" to $[0.52, 0.16, 0.16, 0.16]$)

while other nucleotides are handled similarly) before Θ is returned as the output of the EEM procedure.

IV-2.5 Re-sampling EM Procedure

In EEM, SEME finds a rough motif model with proper motif length. The motif can be further refined using classic EM algorithm to improve the accuracy. However, when the input data X is big, this step is slow. With the idea of importance sampling, we proposed the re-sampling EM (REM) procedure which reduces the running time by running EM algorithm on a subsample of the original data X .

Let $Q(\cdot)$ be the sampler function, where $Q(x) = 1$ if x is sampled; and 0 otherwise. $Q(\cdot)$ is a uniform random sampler, this approximation is trivial and we can directly use the classic EM algorithm and formulas in the sampled dataset in this case. Here, we generalize the formulas of EM to an arbitrary sampler $Q(\cdot)$, which satisfies $\Pr(Q(x) = 1) > 0, \forall x \in X$.

Theorem 1. Let $X_Q = \{X_i \in X | Q(X_i) = 1\}$ be the subset sampled from the original dataset X using the sampler function $Q(\cdot)$, then,

$$E_{X_Q} \left[\sum_{X_i \in X_Q} \frac{\log \Pr(X_i, Z_i | \Phi)}{\Pr(Q(X_i) = 1)} \right] = \sum_{X_i \in X} \log \Pr(X_i, Z_i | \Phi) \quad (4.25)$$

Where $E_{X_Q}[\cdot]$ is the expectation operator over all possible subset X_Q

Proof. According to sampling property:

$$E_{X_Q}[Q(x)] = 1 \cdot \Pr(Q(x) = 1) + 0 \cdot \Pr(Q(x) = 0) = \Pr(Q(x) = 1) \quad (4.26)$$

Then the proof is straightforward. For each site, the sampling process is independent. Hence, the expectation of the summation value of a subsampled set X_Q can be broken down to the expectation of contribution of each site $X_i \in X$ to the summation value.

$$\begin{aligned}
E_{X_Q} \left[\sum_{X_i \in X_Q} \frac{\log Pr(X_i, Z_i | \Phi)}{Pr(Q(X_i) = 1)} \right] &= \sum_{X_i \in X} E_{X_Q} [Q(X_i)] \cdot \frac{\log Pr(X_i, Z_i | \Phi)}{Pr(Q(X_i) = 1)} \\
&= \sum_{X_i \in X} \log Pr(X_i, Z_i | \Phi)
\end{aligned} \tag{4.27}$$

Where $\Phi = (\lambda, \theta, \vec{\theta}_0, \alpha_1, \dots, \alpha_K, \beta_1, \dots, \beta_2)$ are all the modeling parameters in our mixture model.

Q.E.D.

According to Theorem 1 and the large sample theory[71], we can expect to get the same log likelihood value as Equation (4.16), by weighting each subsequence X_i in the sampled dataset X_Q with $\frac{1}{Pr(Q(X_i) = 1)}$, when the sample size $|X_Q|$ is large enough. Therefore, Equation (4.16) can be approximated as:

$$\Phi^{(t)} = \arg \max_{\Phi} E_{Z^{(t)}} \left[\sum_{X_i \in X_Q} \frac{\log Pr(X_i, Z_i^{(t)} | \Phi)}{Pr(Q(X_i) = 1)} \right] \tag{4.28}$$

Interestingly, no matter how we choose the sampler function $Q(\cdot)$, the maximum likelihood estimation always converges to the original one, when the sample size is large enough. However, running EM using different $Q(\cdot)$ yields different sampling efficiencies. For example, we can use a uniform random sampler, i.e., $Pr(Q(X_i) = 1) = \mu$ for every $X_i \in X$, where $\mu \in [0, 1]$ is the sampling ratio. This function is expected to only cover $100 \cdot \mu\%$ of the correct motif sites from X which prohibits the use of small μ . In addition, the number of parameters in the motif model is much larger than that in the background model. In order to learn a motif model as good as the background model, it needs more samples from binding sites than from background sites. In the real dataset, the prior probability of binding site λ is usually very small (less than 0.01). This motivates us to perform biased sampling, i.e., we

want to use a sampler function, which tends to sample more binding sites than background sites. Here, we define our sampler to sample subsequences according to the PWM model outputted by extending EM $\Theta^{(EEM)}$, that is:

$$Pr(Q(x)=1) = \min(4^l \cdot \mu \cdot Pr(x | \Theta^{(EEM)}), 1), \forall x \in X \quad (4.29)$$

where μ is the sub-sampling ratio defined by the user, and l is the motif length.

Here is the rationale behind Equation (4.29). We want to control the final sample size to be roughly $\mu \cdot n$, where n is the total number of sites. For sequences of length l , there are 4^l possibilities, and if we use the $Q(\cdot)$ above to sample all these 4^l l -mers, the expected number of sampled sites is

$$\sum_{x \in \{A,C,G,T\}^l} 4^l \cdot \mu \cdot Pr(x | \Theta^{(EEM)}) = \mu \cdot 4^l \cdot \sum_{x \in \{A,C,G,T\}^l} Pr(x | \Theta^{(EEM)}) = \mu \cdot 4^l \quad (4.30)$$

Therefore, if the original dataset X of size n is formed by a uniform subset of those unique 4^l l -mers, we can expect the size of X_Q is $\mu \cdot n$.

This strategy is useful since we avoid most of the background sites in X . In fact, our simulation reveals that the REM procedure can achieve nearly 60% recall rate (of the correct motif sites) at the sampling ratio as small as 2^{-10} (≈ 0.001) and 90% recall rate at the sampling ratio of 2^{-5} (≈ 0.031) (see Figure IV-5(b)). We choose a default sampling ratio of 0.01 in all experiments in this chapter.

Below, we describe the implementation detail for re-sampling EM (REM) procedure. First, for the E-step, it is almost the same as the classic EM except that we add two Boolean parameters (τ_{pos} and τ_{rank}) to indicate whether the computation should consider the position model and the sequence model or not.

$$\eta_i^{(t)} = \frac{(\lambda \prod_{j=1}^l \prod_b (\Theta_{j,b}^{(t-1)})^{I(b, X_{i,j}^{(seq)})})}{(1-\lambda) (\prod_{j=1}^l \prod_b (\theta_{0,b}^{(t-1)})^{I(b, X_{i,j}^{(seq)})})} \left(\frac{\alpha_{X_i^{(pos)}}^{(t-1)}}{K} \right)^{\tau_{pos}^{(t-1)}} \left(\frac{\beta_{X_i^{(rank)}}^{(t-1)}}{K} \right)^{\tau_{rank}^{(t-1)}} \quad (4.31)$$

The motivation of introducing the indicator variables (τ_{pos} and τ_{rank}) is to avoid over-fitting the data in the final model by assuming the position preference and the sequence rank preference must exist. The position and sequence rank preferences are assumed to be non-existent at the beginning of the REM iterations (i.e., $Pr(X_i|Z_i) = Pr(X_i^{(seq)}|Z_i)$). The position and/or sequence rank preferences are considered only when the position and/or sequence rank distributions of $\{Z_i^{(t)}\}$ are significantly different from the uniform distribution (by Chi-square test). This strategy allows SEME to tell users which preference is really important for the predicted motif. These two indicator variables are updated in the M-step in each iteration and set to 1 only if the expected binding sites distribution is significantly different to uniform distribution (i.e., background distribution).

Next, we describe the M-step. Using the new objective function, $(\lambda, \Theta, \vec{\theta}_0)$ in the t -th iteration of the M-step can be estimated by Equations (4.32)-(4.34).

$$\lambda^{(t)} = \frac{\sum_{X_i \in X_Q} \frac{Z_i^{(t)}}{Pr(Q(X_i) = 1)}}{\sum_{X_i \in X_Q} \frac{1}{Pr(Q(X_i) = 1)}} \quad (4.32)$$

For $b \in \{A, C, G, T\}$ and $j=1, \dots, l$, we have

$$\theta_{0,b}^{(t)} = \frac{\sum_{X_i \in X_Q} \sum_{j=1}^l \frac{(1 - Z_i^{(t)}) I(b, X_{i,j})}{Pr(Q(X_i) = 1)}}{\sum_{X_i \in X_Q} \sum_{w=1}^l \frac{(1 - Z_i^{(t)})}{Pr(Q(X_i) = 1)}} \quad (4.33)$$

$$\Theta_{j,b}^{(t)} = \frac{\sum_{X_i \in X_Q} \frac{Z_i^{(t)} I(b, X_{i,j})}{Pr(Q(X_i) = 1)}}{\sum_{X_i \in X_Q} \frac{Z_i^{(t)}}{Pr(Q(X_i) = 1)}} \quad (4.34)$$

As the position and sequence rank modeling parameters are independent to our sampler function $Q(\cdot)$, so we do not have to re-weight each site in X_Q . $(\alpha_1, \dots, \alpha_K, \beta_1, \dots, \beta_K)$ are updated using Equations (4.17) and (4.18), except that we replace X with X_Q .

For the values τ_{pos} and τ_{rank} they are updated based on the result of two Chi-square tests. Precisely, $\tau_{pos} = 1$ if the positional distribution of binding sites $\{X_i^{(pos)} \cdot Pr(X_i^{(pos)} | Z_i = 1) | X_i \in X_Q\}$ is significantly different from the uniform distribution (Chi-square test); and $\tau_{pos} = 0$ otherwise. The Chi-square statistics χ is defined as: ($I(\cdot, \cdot)$ is the indicator function)

$$\chi_{pos} = \frac{\sum_{k=1}^K \frac{(\sum_{X_i \in X_Q} I(k, X_i^{(pos)}) Pr(X_i^{(pos)} | Z_i = 1) - \frac{1}{K} \sum_{X_i \in X_Q} Pr(X_i^{(pos)} | Z_i = 1))^2}{\frac{1}{K} \sum_{X_i \in X_Q} Pr(X_i^{(pos)} | Z_i = 1)}}}{\frac{1}{K} \sum_{X_i \in X_Q} Pr(X_i^{(pos)} | Z_i = 1)}} \quad (4.35)$$

Similar significance test can be applied for τ_{rank}

$$\chi_{rank} = \frac{\sum_{k=1}^K \frac{(\sum_{X_i \in X_Q} I(k, X_i^{(rank)}) Pr(X_i^{(rank)} | Z_i = 1) - \frac{1}{K} \sum_{X_i \in X_Q} Pr(X_i^{(rank)} | Z_i = 1))^2}{\frac{1}{K} \sum_{X_i \in X_Q} Pr(X_i^{(rank)} | Z_i = 1)}}}{\frac{1}{K} \sum_{X_i \in X_Q} Pr(X_i^{(rank)} | Z_i = 1)}} \quad (4.36)$$

As the default setting, if the p-value for the Chi-square test is less than 0.05, the indicator variable $\tau_{pos}^{(t)}$ or $\tau_{rank}^{(t)}$ will be updated to 1, respectively; and 0, otherwise.

Figure IV-3 is the pseudocode for this procedure.

Re-sampling EM

Require: the extended PWM $\Theta^{(EEM)}$, sampling rate μ , input sequences X

Ensure: Final refined PWM $\Theta^{(t)}$

- 1: Initialize the parameter set $\Phi^{(0)}$ for the mixture model;
- 2: $X_Q := \{X_i \in X \mid Q(X_i) = 1\}$ according to the probability $Pr(Q(X_i) = 1) = \min\{4^w \mu Pr(X_i \mid \Theta^{(EEM)}), 1\}$;
- 3: $t := 1$;
- 4: **repeat**
- 5: E-step: $\forall X_i \in X_Q$, compute $Z_i^{(t)}$ using the parameter set $\Phi^{(t-1)}$ in the last iteration;
- 6: M-step: update $\Phi^{(t)}$ by maximizing the weighted log likelihood $\sum_{X_i \in X_Q} \frac{\log Pr(X_i, Z_i \mid \Phi)}{Pr(Q(X_i) = 1)}$;
- 7: **if** the position distribution of $\{Z_i^{(t)}\}$ is significantly different from uniform distribution **then**
- 8: include position preference in the model;
- 9: **end if**
- 10: **if** sequence rank distribution of $\{Z_i^{(t)}\}$ is significantly different from uniform distribution **then**
- 11: include sequence rank preference in the model;
- 12: **end if**
- 13: $t := t + 1$;
- 14: **until** $\Theta^{(t)}$ converge;

Figure IV-3: Pseudocode for Re-sampling EM procedure.

IV-2.6 Sorting and Redundancy Filtering

The PWMs outputted by REM are further evaluated and sorted by empirical ROC-AUC (the area under the receiver-operator characteristic curve) or over-representation Z-score (representing the motif abundance) with the input data (details on each scoring are in the Figure IV-4 and Supp Figure 1). The first score is preferred for the case when the input sequences are short and most sequences contain at least one motif site (e.g., ChIPed TF motif finding); for the other cases, we suggest to use the Z-score (Supp Figure 1). We eliminate redundant PWMs from the sorted list as

follows. When the sites of a PWM motif overlap with those of another PWM motif by more than 10%, we will treat the PWM motif with the lower score as redundant and remove it.

Procedure 3 AUC

Input: PWM Θ , positive data X^{pos} , negative data X^{neg}

Output: AUC value

```

1: for all sequence  $X_i \in X^{pos} \cup X^{neg}$  do
2:    $Score_i :=$  compute maximum PWM score  $\log Pr(X_{i,j} | \Theta)$ , for all subsequence  $X_{i,j} \in X_i$ 
3: end for
4: Sort  $\{Score_i\}$  in descendent order, and  $\{i_1, \dots, i_n\}$  are the sorted indices
5: for  $k = 1, \dots, n$  do
6:    $TPR_k := \frac{|X_u \in X^{pos}, u \in \{i_1, \dots, i_k\}|}{|X^{pos}|}$  {true positive rate}
7:    $FPR_k := \frac{k - |X_u \in X^{pos}, u \in \{i_1, \dots, i_k\}|}{|X^{neg}|}$  {false positive rate}
8: end for
9: Compute ROC curve based on  $\{TPR_k, FPR_k\}$ 
10: return the area under the ROC curve

```

Figure IV-4: **Procedure for computing AUC score.** Given a set of positive sequences and negative sequences, and a PWM motif, we compute the best match score of the PWM motif in every sequence. Then using different PWM score cut-off, we can compute the "True Positive Rate" and "False Positive Rate" of the PWM and generate the receiver operating characteristic (ROC) curve. Finally, the AUC score of the given PWM can be calculated as the area under the ROC curve.

IV-3 Result

IV-3.1 Profiling two novel EM procedures

IV.3.1.1 EEM estimates the correct motif length

One of the strong points of SEME is that user need not provide any prior motif length (which is, in most cases, hard to estimate). As shown in Figure IV-5(a), for most cases, the EEM procedure estimated motif lengths that are very close to the planted motif length. We further observed that, when the estimated PWM length differs, EEM tends to underestimate the length. Sinice most motifs in JASPAR

contains very weak signal in their flanking positions, EEM excludes them to avoid over-fitting the training data.

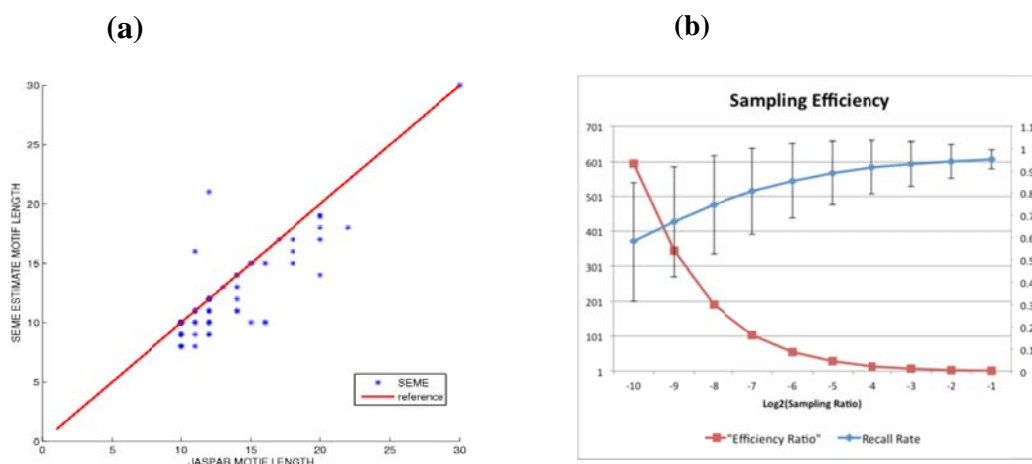


Figure IV-5: EEM Motif Estimation and REM Sampling Efficiency.

(a) Comparison of EEM estimated motif length and actual motif length: The estimation of PWM length by the EEM procedure closely matches the actual planted motif. When the planted PWM have degenerate positions on its flanking positions, EEM will predict a shorter PWM which excludes the latter. (b) Efficiency Ratio and Recall Rate across different sampling ratios: We apply different sampling ratios to run SEME on 75 simulation datasets. The left y axis is value of the average efficiency ratio of SEME biased sampling against the uniform sampling. The right y axis is the average recall rate of true planted sites. The error bar presents the region +/- one standard deviation of recall rate.

IV.3.1.2 REM recalls the majority of the actual motif sites

We want to study the sampling efficiency of re-sampling EM procedure of SEME. For each dataset and for each sampling ratio μ , we trace back the subsampled sites in the re-sampling EM procedure, and check how many true planted sites were included in the subsampled set. Under the same sampling ratio, we define a term “efficiency ratio” to be the ratio between the number of true planted sites in the biased subsampled set and that in the uniform subsampled set. In Figure IV-5(b), the red line shows the efficiency ratio changes from 600 to 2 when the sampling ratio μ changes

to 2^{-10} to 2^{-1} . It shows great efficiency ratio when the sampling ratio is low, because SEME performs subsampling using the output PWM from the extending EM procedure which has much higher chance to sample a true site than naive uniform sampling. Moreover, we observed that most true sites were sampled even in the low sampling ratio. After certain point, increasing the sampling ratio is almost the same as increasing the background ratio, which makes efficiency ratio drop dramatically. To illustrate this, we can check the average recall rate (blue line in Figure IV-5(b)) across different sampling ratio, and it shows near 60% recall rate at sampling ratio $2^{-10} \approx 0.001$ and 90% recall rate at sampling ratio $2^{-5} \approx 0.031$. The error bar in Figure IV-5(b) presents the interval \pm one standard deviation from the average recall rate, and we can see that higher sampling ratio can bring smaller variances of recall rate. To balance trade-off between the efficiency ratio and coverage, we fix default sampling ratio=0.01 in the later experiments of this chapter.

IV.3.1.3 SEME significantly outperforms MEME in recovering the planted PWM.

To analyze SEME's performance, we extract all seventy-five motifs with lengths longer than 9bp in JASPAR[124] vertebrate core database. For each such motif, we generated a training dataset of 1000 random sequences of length 400bp where 500 of them contain one motif site. These motif sites are planted uniformly across all positions and sequences.

For each dataset, we run SEME (EEM only), SEME (EEM + REM), and MEME (the classical EM-based motif finder) and obtain the top 5 predicted PWMs from each program. To test the goodness of the predicted PWMs, we compared the PWM divergence[78] between the predicted PWMs and the actual planted PWMs.

We also generate independent testing sequences with length 400bp (1000 positive sequences with one implanted motif site, 1000 negative sequences without motif site), and compute the ROC-AUC value for each predicted PWM. Figure IV-6(a) shows the comparison result. As expected, the random PWMs have the worst AUC values while the actual planted PWMs have the best AUC values. EEM’s predicted PWMs have significantly better discriminative capability (AUC) and similarity (less PWM divergence) to the actual planted PWM as compared to random PWMs. This indicates that EEM’s PWMs are good starting points for the subsequent REM procedure. REM’s predicted PWMs further improve the AUC score and are similar to the actual planted PWM (as indicated by the small PWM divergence).

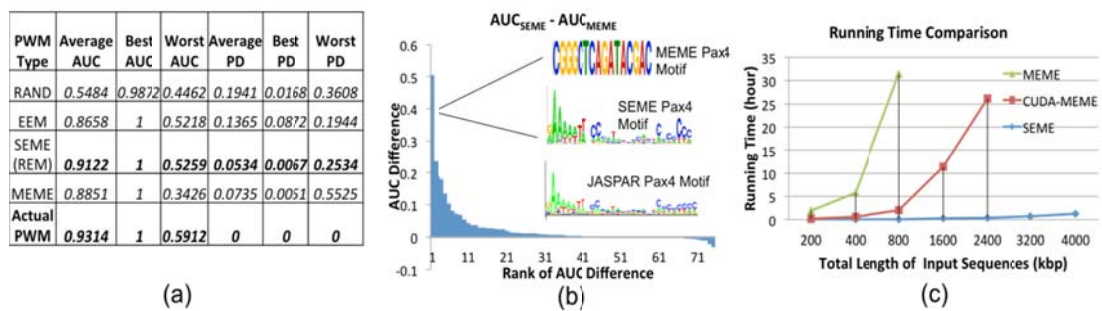


Figure IV-6: **The empirical performance of SEME on synthetic datasets.**

(a) The accuracy of SEME’s PWM (both EEM step’s (unrefined) PWM and the REM’s (final) PWM are listed). We quantify accuracy using the commonly used Area-Under-ROC Curve (AUC) score and PWM divergence (PD). We showed that EEM’s predicted PWM is already significantly stronger than random; indicating the goodness of EEM’s PWM as starting point for the subsequent REM step. The scores also show that SEME’s PWMs are significantly better when compared to MEME’s. (b) Based on the performances of SEME and MEME on the Pax4 motif dataset, we observed that MEME has serious difficulties in mining PWMs with long gap region within them. (c) The running time of SEME is shown against increasing input size. We observed that CUDA-MEME, the GPU enabled version of MEME, still runs slower than SEME running on normal CPU (it takes 1 day to handle ≈ 6000 sequences while SEME takes around 1 hour for 10000 sequences).

Figure IV-6(a) also shows that SEME outperformed MEME. In fact, SEME is better than MEME in 42 out of 75 experiments (the cases with positive AUC differences in Figure IV-6(b)). The cases where SEME performed worse have relatively small AUC score differences (less than 0.04). We examined the Pax4 dataset in which SEME gains the highest improvement against MEME. The implanted JASPAR Pax4 motif is a diverged PWM of length 30. SEME successfully extended and recovered the full Pax4 motif; thanks to the ability of its EEM procedure to handle long gaps in its extension step. In contrast, MEME failed to model the long gaps due to their starting point finding procedure which assumes that all of the PWM positions are equally important.

IV.3.1.4 SEME is more suitable in handling large-scale data.

We further generated 7 large datasets to observe the capability of SEME in handling large-scale data. Each dataset consists of different number of sequences (from 500 to 10000, each of length 400bp). Figure IV-6(c) shows that the original MEME program cannot process more than 2000 sequences within one day, hence we also used the GPU-accelerated version of MEME, CUDA-MEME [81](run on two Intel X5670 CPUs and two Fermi M2050 GPUs with 48GB RAM). SEME was run as a normal CPU program. SEME is still around 60 times faster than CUDA-MEME which runs on a highly parallelized GPU system. In addition, SEME can process up to 10000 sequences (a typical dataset size for ChIP-seq experiments) in 1 hour while the CUDA-MEME took more than one day to process 6000 sequences.

IV-3.2 Comparing TF motif finding in large scale real datasets

We compared the performance of SEME with other existing motif-finding programs on two large-scale TF binding data. We also studied the ability of SEME in

uncovering the hidden position and/or sequence rank preferences in the input dataset when they are present.

IV.3.2.1 The Metazoan Compendium datasets

The first benchmark is a metazoan compendium dataset published by Linhart et.al[78]; consisting of 32 datasets based on experimental data from microarray, ChIP-chip, ChIP-DSL, and DamID as well as Gene Ontology data. A list of the promoter sequences of many target genes (1000bp upstream and 200bp downstream the Transcription Start Site (TSS)) are used as the positive input for each motif-finding program and promoter sequences of other non-target genes are used as background sequences. The performance of six existing motif-finding programs, namely AlignACE[99], MEME [6], YMF [111], Trawler [30], Weeder [92], and Amadeus [78], were compared in the original benchmark study [78]. Each program's predicted PWMs are evaluated by the PWM divergence. Only PWMs with medium and strong matching with the known motifs (PWM divergence <0.18) are considered to be successfully detected[78].

The result of this comparison is shown in Figure IV-7. We found that SEME successfully detected the correct motifs in 21 datasets whereas the second best program, Amadeus, succeeded in 18. Weeder and Trawler found correct PWMs in 11 and 12 datasets, respectively. SEME also found more accurate motifs than the rest; it found 12 motifs with PWM divergence <0.12 . SEME further detected a significant position preference for the correct motifs for many datasets in this benchmark: most of them tend to bind nearer to the TSS position (see Figure IV-8 and Figure IV-9(a)).

Data set	AlignACE	MEME	YMF	Trawler	Weeder	Amadeus	SEME
Elegans.GATA_Pauli.1427	∞						
Fly.HSF_Machin.186.mapped							
Fly.MEF2_Sandmann.211.mapped							
Fly.MSL1_Legube.116.mapped							
Fly.Myc_Oryan.723.mapped							
Human.CREB_Zhang.2354	∞	∞					
Human.E2F_Ren.96							
Human.E2F4_Cam.203							
Human.ERa_Kwon.498							
Human.ETS1_Hollenhorst.1193	∞						
Human.HCC-G1S_Whitfield.268							
Human.HCC-G2M_Whitfield.350							
Human.HNF1a_Odom.207							
Human.HNF4a_Odom.1485	∞	∞					
Human.HNF6_Odom.214							
Human.HSF1_Page.333							
Human.ImmuneResponse_GO_Hs.619							
Human.Nanog_Boyer.720							
Human.NFkB_Schreiber.271							
Human.Nrf1_Cam.679							
Human.Oct4_Boyer.243							
Human.p53_Kannan.38							
Human.Sox2_Boyer.580							
Human.SRF_Cooper.174							
Human.YY1_XiRen.721							
Mouse.Foxp3_Marson.1071							
Mouse.ImmuneResponse_GO_Mm.335							
Mouse.MEF2_Blais.26							
Mouse.MyoD_Blais.105							
Mouse.MyoD_Cao.104							
Mouse.MyoG_Cao.78							
Mouse.Myogenin_Blais.110							
# successful detection (PD<0.18)	4	2	7	12	11	18	21
# Strong detection (PD<0.12)	2	2	2	6	8	9	12

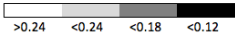
PWM Divergence (PD) Threshold: 

Figure IV-7: **Comparison of de-novo motif discovery tools on the metazoan compendium.** Each column of the table presents the results for one motif discovery tool, and each row corresponds to one data set of the metazoan compendium. The color of the checkmarks represents the accuracy of the motif discovered as measured by the normalized euclidean distance, and we used the thresholds on the PWM divergence as proposed by Linhart et al[78]. The symbol ∞ marks long execution times (hour) that were aborted in[78]. In the last row of the table, we report the total number of motifs discovered by each of the tools.

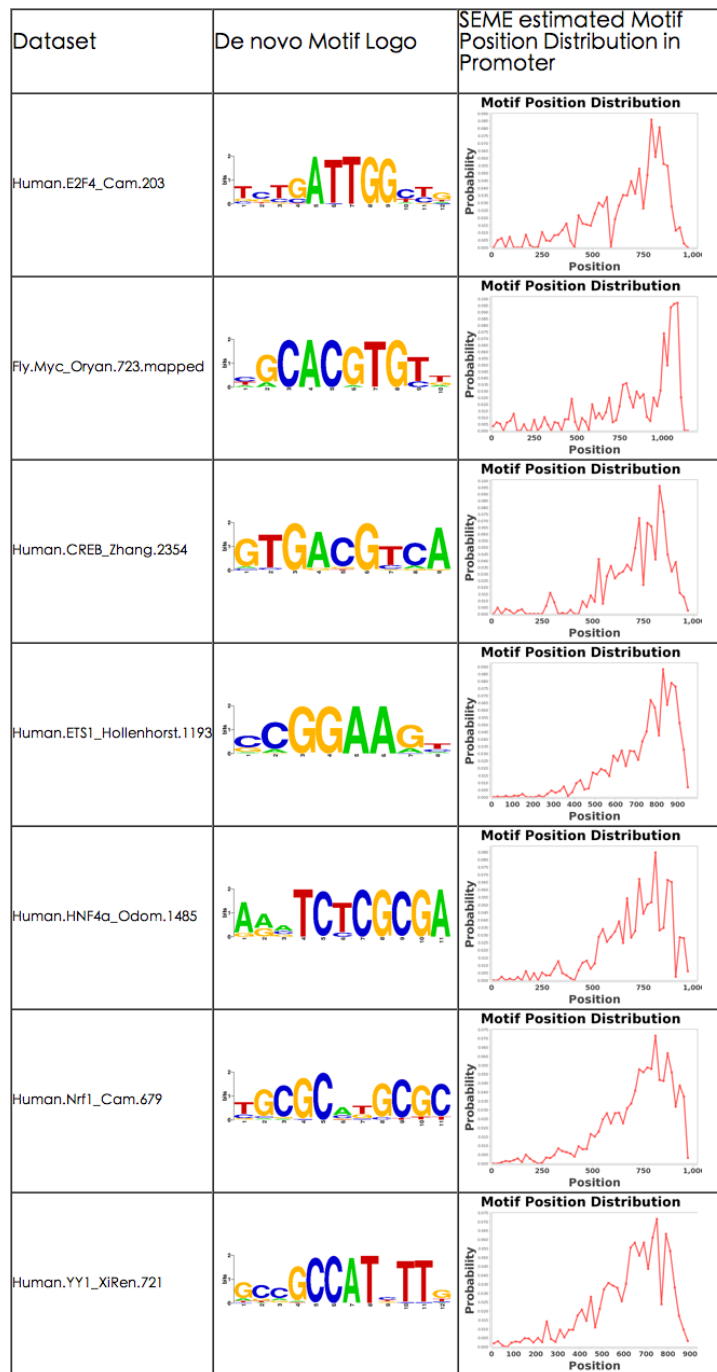


Figure IV-8: **SEME detected TF motifs with significant position preference to TSS** Seven examples of SEME’s output of **metazoan compendium dataset**. The result indicates these TF binding sites are enriched near the transcription start sites. The TSS position is located around 200bp from the rightmost position. The original 1200bp promoter sequences may be shortened after removing “N”-masked regions, so the TSS position may be shifted in those cases.

IV.3.2.2 ChIP-seq experimental datasets: Discovery of the ChIPed TF motif from ChIP-seq data.

The second benchmark is a collection of large scale ChIP-seq experimental data which consists of 164 published ChIP-seq libraries from the ENCODE project[31] and our lab over different cell-lines and TFs[17, 135, 65]. ChIP-seq usually reports more than 10000 target sequences with narrower target regions (100bp). We computed the Area Under ROC Curve, Positive Predictive Value, Average Site Performance and Specificity scores of each program's predicted PWM. The formula for the above scores are given in the Figure IV-4 and Equations(S. 1)-(S. 3). From each library, the 100bp sequences around the top 10000 ChIP-seq peaks were extracted (sorted by ChIP intensity) as our input data. For MEME and Weeder, we only used the top 2000 peaks due to their long running time. Peaks with odd numbered ranks were used for training while the even numbered peaks were used as positive testing data. The negative dataset is generated a 1st-order Markov model trained using the same number of 100bp random sequences extracted from the regions 1000bp away from the ChIP-seq peaks.

We compared SEME with 7 popular *de novo* motif finding programs for ChIP data: MEME, Weeder, Cisfinder, Trawler, Amadeus, CHIPMunk and HMS. Each program's top 5 motifs are evaluated using the four statistics measurements on the test data. For each scoring, the best of the 5 motifs were used to represent the performance of a program. Figure IV-9(b) shows the average performances of the motif finders. Again, we found that SEME is consistently better than all other programs (1st rank in Area under ROC Curve, Positive Predictive Value and Specificity, and 3rd rank in Average Site Performance).

IV.3.2.3 Discovery of co-TF motifs from ChIP-seq data

We noted that most motif finders showed good performance in finding the ChIPed TF motifs. This is expected since the ChIPed TFs are highly enriched in the extracted sequences[135]. Comparing to finding ChIPed TF motifs in ChIP-seq datasets, the problem of finding co-TF motifs in the ChIP-seq datasets is much more challenging. The co-TF motif sites are less abundant and most are not located exactly at the ChIP-seq peaks. Nevertheless, finding the co-TF(s) could potentially uncover previously unknown co-TFs interaction.

For co-TF motif comparison, we used 15 ChIP-seq libraries whose co-TFs have been characterized (the list of co-TFs for each ChIP-seq is in Supp Table 2). We extracted 400bp sequences around the ChIP-seq peaks and compared the top 20 *de novo* motifs of each program to the known co-TF motifs in the JASPAR[105] and TRANSFAC[85] database; we cannot use the previous statistical measurements since co-TFs may not occur in all ChIP-seq peaks. Furthermore, the ChIPed TF binding sites were masked before the co-TF motif finding. SEME and ChIPMunk can do this automatically and, for other programs without auto-masking mode, the input sequences were masked by the top 2 motifs reported from their ChIPed motif finding results.

STAMP program[84] was used to compute the p-value of the match between a predicted co-TF motif against the known co-TF motif. STAMP p-value provides a better match measurement compared to PWM divergence since it removes the motif length bias. We separated the p-value of the PWM matching into three significance levels: (1) weak match ($0.05 \geq \text{p-value} > 0.01$), (2) medium match ($0.01 \geq \text{p-value} > 0.0001$) and (3) strong match ($\text{p-value} \leq 0.0001$). Figure IV-9(c) shows the

performances for different motif-finding programs in finding the co-TF motifs from the 15 datasets. SEME recovered 61 known co-TF motifs; compared to Amadeus and MEME which find 48 and 44 co-TF motifs, respectively. 31 out of the 61 co-TF motifs of SEME belong to the strong match category (Amadeus only found 20) and another 27 are in the medium match category. This indicates that SEME’s predicted co-TF PWMs are highly accurate.

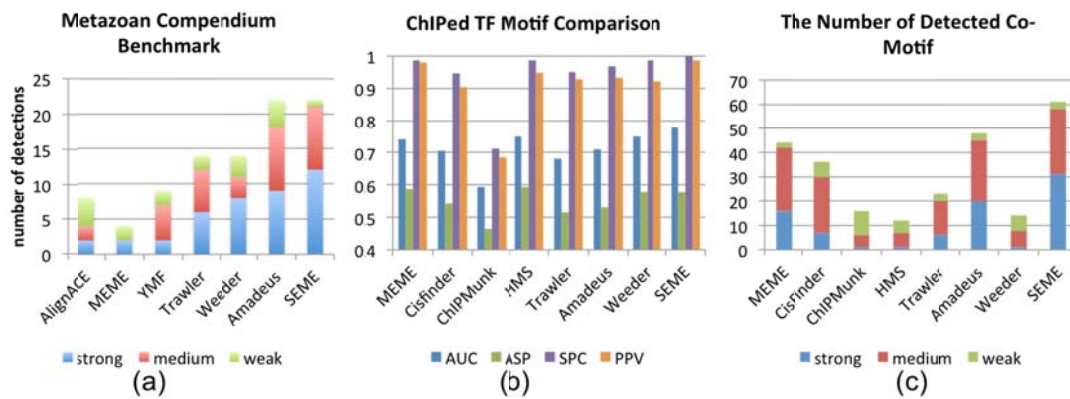


Figure IV-9: **The performance of SEME compared to existing motif finding programs from large scale real data** (a) Comparison result on the metazoan compendium datasets. Four PWM motifs returned by each motif finding program are then compared to the known Transfac motifs using PWM divergence (PD) (as in [78]) and further classified into three matching categories (strong, medium, weak) corresponding to different PD cut-offs (0.12,0.18,0.24). (b) Comparison result on 164 ChIP-seq libraries over four different measurements: AUC, PPV (Positive Predictive Value), ASP (Average Site Performance) and SPC (Specificity). The result shows that most motif finders perform similarly well in detecting ChIPed TF (but SEME is consistently better than all of them). (c) Comparison result for Co-TF motif finding on 15 ChIP-seq libraries. The quality of reported PWMs is classified into three categories (strong, medium, weak) corresponding to different STAMP p-value cut-offs (0.0001, 0.01, 0.05). SEME reported the most number of co-TF’s motif which match the known PWM with STAMP p-value ≤ 0.0001 (strong match, blue bar). Overall, SEME also found the most number of co-TF motif (61) as compared to the second best program, Amadeus (48).

To study the biological significance of the learnt preferences, we further study the output of three datasets, involving TFs like ER, AR, FoxA1, Oct4 and c-Myc, in details (see Figure IV-10). The real binding site of each TF is defined to be the site around +/-100bp around the TF's ChIP-seq peak whose known PWM score is better than a cutoff that yields FDR=0.01. If multiple matches occur, only the best scoring site is chosen. Comparison between SEME's learnt distributions (Figure IV-10, middle columns) and the real binding site distributions (Figure IV-10, rightmost columns) indicates that SEME is able to learn the correct position and sequence rank preferences of the tested TFs. We also found that the motif positions of FoxA1, a known co-TF of ER, is not enriched exactly at the ER ChIP-seq peak in the MCF7 data; instead it is found in the flanking regions near the ER peaks. Interestingly, in the LnCAP AR ChIP-seq dataset (FoxA1 is also a known co-TF of AR), we found that FoxA1 binds very closely to AR—it is enriched at the AR ChIP-seq peak summits. This observation is consistent with the previous report that FoxA1 can physically interact with AR[38]. It also indicates FoxA1 may play different roles when working with AR and ER[104]. In the ChIP-seq data of Oct4 from mouse's ES cell, SEME found the motif of c-Myc enriched within Oct4's low intensity peaks regions. We conjecture that, in these regions, Oct4 indirectly binds the DNA through c-Myc (hence explaining the ChIP-seq's low intensity). An earlier report showed that Oct4, along with Sox2, Nanog, and Stat3 form an enhancer module while c-Myc along with n-Myc, E2F1 and Zfx form a promoter module in the ES cell[17]. In fact, the interactions between these enhancer and promoter modules have also been reported previously[131].

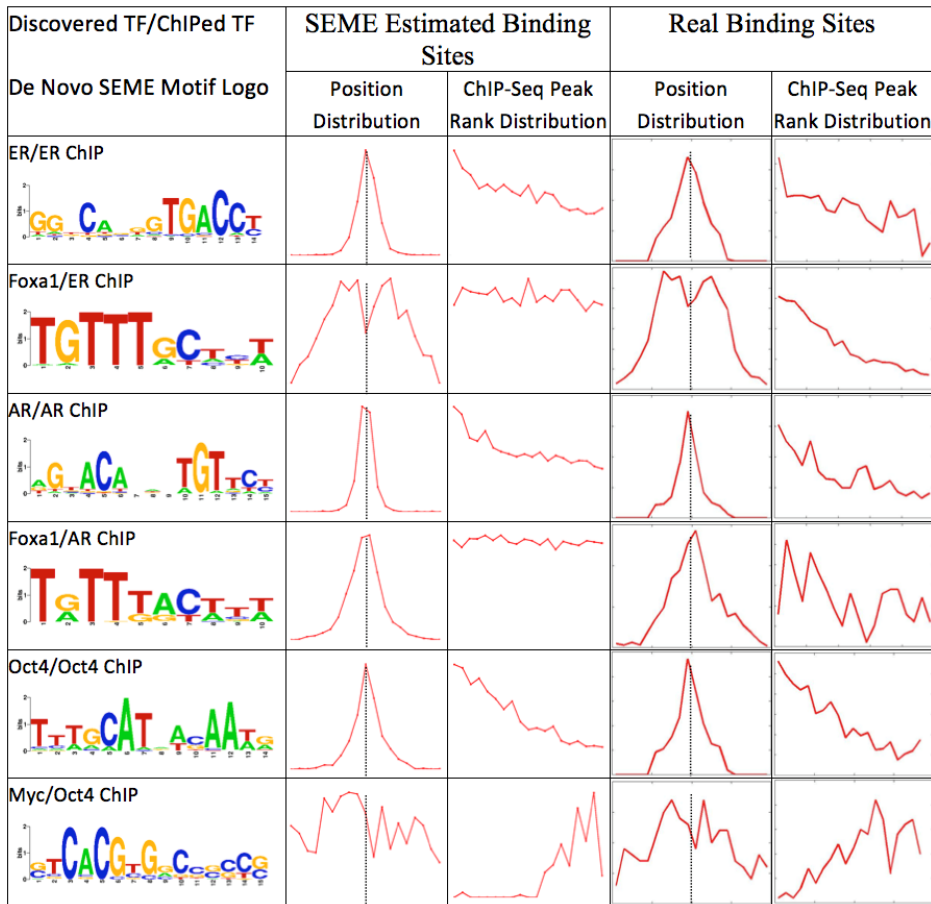


Figure IV-10: **Automatic learning of the position and sequence rank preference from the input data.** Instead of requiring the user to input the expected co-TF motif preference distribution (position and/or sequence rank distribution), SEME learns such distributions directly from the input data. We show that most of the time, SEME can learn the correct distributions of each TF (as compared to real binding sites distribution in the rightmost column, defined by the ChIP-seq and the known PWM of the TF). For position distribution, the x-axis is +/-200bp from ChIP-seq peak summit (the black dash line), and the y-axis is the fraction of binding sites in a given position. For rank distribution, the x-axis is the rank of ChIP-seq peak (left : high ChIP intensity, right : low ChIP intensity), and the y-axis is the fraction of binding sites in a given rank. The ChIP-seq peak rank distributions (MCF7 ER ChIP, LNCaP AR ChIP) of FoxA1 and the position distribution of Myc are tested to be insignificant by SEME.

These examples indicate that the position and sequence rank distribution learnt by SEME are reasonably accurate and users could use them to infer the nature of the interaction between the ChIPed TF and the co-TF(s). In this manner, SEME can be

used to generate biological hypothesis for further experimental validations. Moreover, the highly diverse preferences that we observe highlight the difficulty for users to provide the correct prior in the first place.

IV-4 Conclusion

This chapter developed a novel algorithm called SEME for mining motifs using mixture model and EM algorithm. We presented three important contributions: (1) automatic detection and learning of the position and sequence rank preferences of a candidate motif. (2) ability to estimate the correct TF motif length (with possible gaps within) and (3) using importance sampling for efficiency while still able to estimate the EM parameters unbiasedly. As a result, we showed that SEME is substantially better, both in terms of accuracy and efficiency, compared to the existing motif finding programs.

Moreover, in the task of finding co-TF motif in the ChIP-seq data, SEME not only reports more accurate co-TF motifs than other programs but also correctly estimates the position and sequence rank distribution of each co-TF's motif. We showed that such information provides useful insights on the interaction between the ChIPed TF and the predicted co-TFs. SEME does have a few limitations. Firstly, it assumes that the target motif contains a conserved 5-mer region. In cases without such 5-mer, SEME also allows user to provide custom seeds. Secondly, SEME is more suitable for large scale input (≥ 100 sequences) since it needs enough samples to determine whether it should do extension (EEM) or include additional binding preferences (REM).

CHAPTER - V Inference of Spatial Organizations of Chromosomes

Using Semi-definite Embedding Approach and Hi-C Data

The last two chapters focus on motif analysis for ChIP-seq data, which is sequence level study of Protein-DNA interaction. In this chapter, the focus has been shifted to the structure level study of Protein-DNA interaction using Hi-C data. ChromSDE, a novel chromosome 3D modeling method is introduced in this chapter.

V-1 Introduction

As mentioned in the review in Chapter-II Section II-3, the workflow chromosome 3D modeling contains two steps: (1) Converting the contact frequencies between loci to spatial distances and (2) Predicting the 3D chromosomal structure from the spatial distances.

Although some works [27, 10, 102, 62, 51] have been done, there are still unsolved issues in both steps 1 and 2. For step 1, the conversion between the contact frequency and spatial distance has one parameter. Existing methods, except BACH[51], assume that the parameter is fixed or is known beforehand. We found that the parameter is actually different for different datasets. Thus it is important to have a method to estimate the parameter. For step 2, existing methods infer the 3D chromosomal structure by heuristics. They are not guaranteed to reconstruct the correct structure even in the noise-free case.

To fill in these gaps, we propose a novel chromosome structure modeling algorithm called ChromSDE (Chromosome Semi-Definite Embedding). ChromSDE models the problem as two parts:

- Assuming that the parameter for the conversion from the contact frequency to the spatial distance is known, ChromSDE formulates the 3D structure modeling problem as a non-convex non-linear optimization problem similar to the previous works. Instead of directly solving the non-convex optimization which is NP-hard, ChromSDE relaxes it to a semi-definite programming (SDP) problem, whose global optimal solution can be computed in polynomial time. With this formulation, our approach is guaranteed to recover the correct 3D structure in the noise-free case when the structure is uniquely localizable[112].
- For the parameter in our conversion function from the contact frequency to the spatial distance, ChromSDE formulates it as a univariate optimization problem and estimate the correct parameter by a modified version of the golden section search method.

This chapter may have significant impact in three aspects. First, the SDP relaxation method in ChromSDE is a powerful relaxation technique, which is theoretically guaranteed to recover the correct structure in the uniquely localizable noise-free case[112]. The SDP approach has been successfully applied in other graph realization problems[14, 72, 127], but to our best knowledge, no one has introduced it in chromosome structure modeling. Second, we proved theoretically and empirically that the conversion parameter changes if we examine the data under different resolutions. Thus, it is inappropriate to assume that the conversion is known. We developed an efficient algorithm to estimate the correct conversion parameter from the input data. Third, we proposed a measure called *Consensus Index* that can quantify if the input frequency data comes from a consensus structure or a mixture of different structures. It is arguable if Hi-C data is appropriate for modeling 3D

structures, because the contact frequencies come from a population of cells instead of a single cell. Our simulation showed that if the data is from a consensus structure, the *Consensus Index* is high.

We evaluated our method with simulated data and real Hi-C data. Through simulation study, we showed that ChromSDE could perfectly recover different types of simulated structures in the noise-free setting while other tested programs fail in many cases. Even with noise, ChromSDE still significantly outperforms other tested programs. In addition, we also showed that ChromSDE could accurately estimate the conversion parameter and output the *Consensus Index* that can reflect the degree of mixture. Next, real Hi-C data replicates with different enzyme cutting sites are used to further validate the robustness and accuracy of ChromSDE comparing to other tested programs. The result indicates that ChromSDE can infer a more accurate and robust 3D model than existing methods. Finally, we showed that ChromSDE can robustly handle different resolution data and the predicted high-resolution 3D structure unveils interesting biological findings.

V-2 Method

The Hi-C and TCC technologies enable us to obtain paired-end reads from interacting loci in the genome. The interaction data can be summarized by a contact frequency matrix F , in which F_{ij} represents the number of contacts between loci i and j (loci i and j are genomic regions in a fixed bin size such as 1Mbp or 40kb). We expect two loci are close if and only if the contact frequency between them is high. A further note is that the raw Hi-C or TCC interaction frequencies are affected by

various biases (GC content, mappability and fragment length), and should be normalized [133].

The chromatin 3D modeling problem is defined as follows: Given a normalized interaction frequency matrix F , infer a 3D structure whose pairwise distances highly correlate with the interaction frequencies in F . This problem can be solved by two steps: 1) converting the frequency matrix F into a distance matrix D that describes the expected pairwise distance among the loci; 2) learning a 3D structure from the distance matrix D . Step 1 is based on the observation of Lieberman-Aiden, et al. [76] that the conversion between the frequency matrix F and the distance matrix D follows the power law distribution (Equation (5.1)) where α is a parameter called the conversion factor and D_{ij} and F_{ij} are the distance and frequency between loci i and j .

$$D_{ij} = \begin{cases} (1/F_{ij})^\alpha & \text{if } F_{ij} > 0 \\ \infty & \text{otherwise} \end{cases} \quad (5.1)$$

There are two main challenges in this approach: 1) estimate α ; and 2) convert the distance matrix D to the 3D model. In the following two sub-sections, we present ChromSDE that resolves these two challenges. Firstly, assuming that the conversion factor α is known, we describe a method that estimates the 3D structure from the expected distance matrix D . Then, the next section explains how ChromSDE estimates the correct value of the conversion factor α . To note that, the scale between the converted distance and the real physical distance is not considered here, since the relative distance (without the scale) does not affect the predicted structure for visualization and further study.

V-2.1 From Distance Matrix To 3D Structure

Assuming the conversion factor α (>0) is known, the interaction frequency matrix F can be converted to the expected distance matrix D by Equation (5.1). The 3D chromatin structure modeling problem aims to compute a set of 3-dimensional coordinates $\{\vec{x}_1, \dots, \vec{x}_n\}$ for the n loci, such that their distances can fit the distance matrix D well. In other words, we hope to ensure that $\|\vec{x}_i - \vec{x}_j\|$ (distance between loci i and j) is approximately the same as D_{ij} for all loci i and j . Mathematically, this problem can be formulated as three alternative optimization models in Equations (5.2)-(5.4), where $\|\cdot\|$ denotes the Euclidean norm. Each equation has two terms. The first term aims to minimize the errors between the embedding distances and the expected distances. These three alternatives apply three different commonly used error functions in the literatures: (a) sum of square errors of the distance differences [9, 27], (b) sum of absolute errors of the distance square differences [14, 72] and (c) sum of square errors of the distance square differences [14, 82]. The second term is the same for the three alternatives. It is a regularization term that maximizes the pairwise distances for the loci without any interaction frequency data. It is based on the assumption that the spatial distances of loci pairs not captured by the experiment cannot be too short.

$$\min_{x_1, \dots, x_n \in R^3} \sum_{\{i,j|D_{ij}<\infty\}} \omega_{ij} (\|\vec{x}_i - \vec{x}_j\| - D_{ij})^2 - \lambda \sum_{\{i,j|D_{ij}=\infty\}} \|\vec{x}_i - \vec{x}_j\|^2 \quad (5.2)$$

$$\min_{x_1, \dots, x_n \in R^3} \sum_{\{i,j|D_{ij}<\infty\}} \omega_{ij} \left| \|\vec{x}_i - \vec{x}_j\|^2 - D_{ij}^2 \right| - \lambda \sum_{\{i,j|D_{ij}=\infty\}} \|\vec{x}_i - \vec{x}_j\|^2 \quad (5.3)$$

$$\min_{x_1, \dots, x_n \in R^3} \sum_{\{i,j|D_{ij}<\infty\}} \omega_{ij} \left(\|\vec{x}_i - \vec{x}_j\|^2 - D_{ij}^2 \right)^2 - \lambda \sum_{\{i,j|D_{ij}=\infty\}} \|\vec{x}_i - \vec{x}_j\|^2 \quad (5.4)$$

In the formulas, ω_j represents the weight or confidence of the observed data D_{ij} . Since we expect the confidence of D_{ij} is higher when F_{ij} is large, this chapter simply set $\omega_j=1/D_{ij}$. The parameter $\lambda > 0$ in the second term is the regularization coefficient to balance the error term and the regularization term. In practice, we found that the results are stable for $0.001 < \lambda < 0.1$ (Figure V-1) and we fix it to 0.01 in this chapter. All three formulations (5.2)-(5.4) are non-convex non-linear optimization problems, which are NP-hard to solve for their global minimizers. Existing methods solved them by heuristics like MCMC sampling [101, 51], or local search[27, 61, 103]. Here, we show that, by relaxing the solution space of every \vec{x}_i from R^3 to R^n (n is the number of loci), formulations (5.3) and (5.4) become convex semidefinite programming (SDP) problems for which we can compute their global minimizers to any given degree of accuracy in polynomial time. Furthermore, if the expected distance matrix is indeed generated from a 3D object and is noise-free, the above relaxations can reconstruct the optimal R^3 solution by projecting the R^n points to certain R^3 subspace in theory [112]. In practice, even if the distance matrix is not noise-free, we still can find a good approximated solution in the R^3 subspace. The projecting technique to obtain a solution in R^3 will be introduced in Section V-2.3.

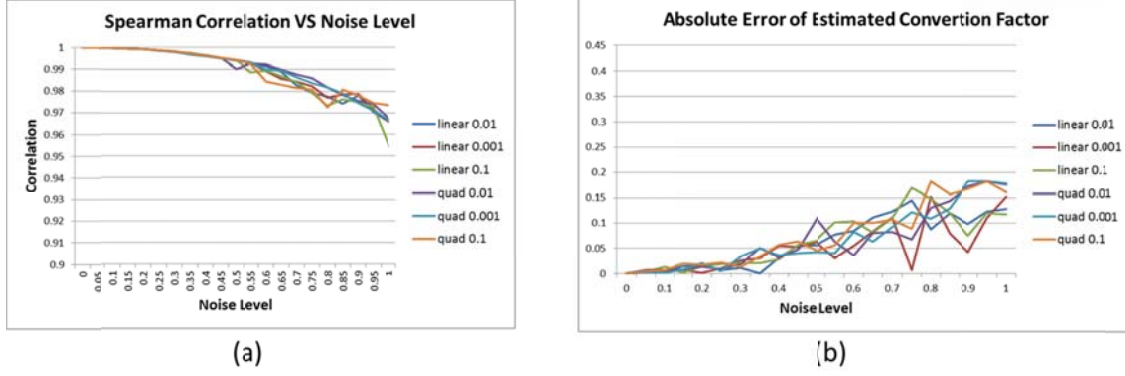


Figure V-1: **Simulation result that compares the performance of different regularization parameters.** The results were generated using simulation on a Brownian Motion Curve as stated in Section V-3.1. In the experiments, different regularization parameter lambda (0.1, 0.01, 0.001) and different types of SDPs (linear, quadratic) combination were tested. (a) Spearman correlation between the pairwise distance matrices of the predicted structure and the true structure under different noise level. (b) The absolute error of the estimated value of conversion factor under different noise levels.

V-2.2 Formulation of SDP relaxation problems

This section describes how to reformulate Equations (5.3) and (5.4) as linear and quadratic semidefinite programming (SDP) problems by relaxing the solution space of every \vec{x}_i from R^3 to R^n . Let K be the kernel matrix for $X = \{\vec{x}_1, \dots, \vec{x}_n\}$ (i.e., $K_{ij} = \vec{x}_i \cdot \vec{x}_j = K_{ji}$), then every square distance can be expressed in term of K . Precisely, we have: $\|\vec{x}_i - \vec{x}_j\|^2 = K_{ii} + K_{jj} - 2K_{ij}$. In addition, we set the center of the points to be the origin, that is:

$$\sum_{i=1}^n \vec{x}_i = 0 \Rightarrow \left\| \sum_{i=1}^n \vec{x}_i \right\|^2 = 0 \Rightarrow \sum_{i,j} K_{ij} = 0 \quad (5.5)$$

By our definition of the kernel matrix, K must be symmetric positive semidefinite (i.e., $K \succeq 0$). We first describe the quadratic relaxation (Equation (5.4)), which is stated as below:

$$\begin{aligned}
\min \quad & \sum_{\{i,j|D_{ij}<\infty\}} \omega_{ij}(K_{ii} + K_{jj} - 2K_{ij} - D_{ij}^2)^2 - \lambda \sum_{\{i,j|D_{ij}=\infty\}} (K_{ii} + K_{jj} - 2K_{ij}) \\
\text{s.t.} \quad & \sum_{i,j} K_{ij} = 0, \quad K \succeq 0.
\end{aligned} \tag{5.6}$$

For Formulation (5.3), the error term contains the absolute value operator $|\cdot|$, which cannot be handled directly by standard SDP solvers. Fortunately, without increasing the problem complexity, we can replace the absolute value operator $|\cdot|$ by adding two sets of slack variables. The linear SDP relaxation of Equation (5.3) is stated as below:

$$\begin{aligned}
\min \quad & \sum_{\{i,j|D_{ij}<\infty\}} \omega_{ij}(\varepsilon_{ij}^+ + \varepsilon_{ij}^-) - \lambda \sum_{\{i,j|D_{ij}=\infty\}} (K_{ii} + K_{jj} - 2K_{ij}) \\
\text{s.t.} \quad & K_{ii} + K_{jj} - 2K_{ij} + \varepsilon_{ij}^+ - \varepsilon_{ij}^- = D_{ij}^2 \\
& \sum_{ij} K_{ij} = 0, \quad K \succeq 0, \quad \varepsilon_{ij}^+, \varepsilon_{ij}^- \geq 0.
\end{aligned} \tag{5.7}$$

Note that ε_{ij}^+ (and ε_{ij}^- respectively) represents the penalty when the embedding distance is shorter (and longer respectively) than the expected distance. Moreover, at least one of them must be zero in the final solution since they are non-negative and their summation is minimized.

A general purpose SDP solver, such as SDPT3[120], can be used to solve the two SDP problems above. However, all the current general-purpose SDP solvers (which are all based on interior-point methods) cannot handle large-scale SDP problems. They can only comfortably handle distance matrix with around 40,000 expected distances (≈ 200 loci). Fortunately, for convex quadratic SDP such as the Formulation (5.6), recently developed advanced algorithm[58] based on partial proximal-point method (with semi-smooth Newton-CG method for solving the subproblems) can handle such a problem very efficiently even when the problem

scale is large. In particular, it can handle 10,000,000 expected distances (≈ 3000 loci). In the result section, we present the results for both SDP relaxations in the small-scale problems and the results for the quadratic SDP relaxation in the large-scale problems (if not specially mentioned, the result is generated by quadratic SDP).

V-2.3 Obtaining 3D coordinates from the Kernel Matrix

By solving the SDP Formulation (5.6) or (5.7), we obtain the solution as a positive semidefinite kernel matrix K . By computing the eigenvalue decomposition of K , the R^3 coordinates $X = \{\vec{x}_1, \dots, \vec{x}_n\}$ can be recovered from K (i.e., $K \approx X^T X$). A 3-dimensional representation that approximately satisfies $K_{ij} \approx \vec{x}_i \cdot \vec{x}_j$ can be obtained from the top 3 eigenvalues $(\gamma_1, \gamma_2, \gamma_3)$ and eigenvectors $(\vec{v}_1, \vec{v}_2, \vec{v}_3)$ of K . That is,

$$\vec{x}_i = [\sqrt{\gamma_1} \cdot v_{1,i} \quad \sqrt{\gamma_2} \cdot v_{2,i} \quad \sqrt{\gamma_3} \cdot v_{3,i}]^T \quad (5.8)$$

In the ideal case where the input expected distance matrix is noise-free and dense enough (i.e., it has sufficient constraints to uniquely present a 3D structure), it can be shown that the approximation (5.8) is the exact solution and all other eigenvalues (except top 3) are equal to zero. The property is called unique localizability [112].

When the input expected distance matrix is noisy, ChromSDE performs further local refinement to the 3D coordinates obtained from the SDP relaxation problems[14]. Specifically, our ChromSDE algorithm applies a local optimization method such as a quasi-Newton method or a gradient descent method to the original non-convex problem by using the 3D positions obtained from the SDP problems as the starting point. Because the 3D positions produced by the SDP problems are

generally close to a local minimizer, a local optimization method can generally converge to a good local minimizer for the original non-convex problems.

To measure if the input distance matrix can be represented as a single 3D structure, we propose a measure called *Consensus Index*, which includes two parts: the first part measures how the input distance matrix D satisfying the triangle inequality, and is presented as the ratio between the embedded distance in R^n and the input distance; the second part measures how good the R^3 approximation is, and is presented as the ratio between the sum of top 3 eigenvalues (i.e., $\sum_{i=1}^3 \gamma_i$) and the sum of all eigenvalues of K (i.e., $\sum_{i=1}^n \gamma_i$). Precisely, Let $D'_{ij} = \sqrt{K_{ii} - 2K_{ij} + K_{jj}}$ be the embedded distance in R^n , then we have:

$$Consensus\ Index = \frac{\sum_{\{i,j|D_{ij}<\infty\}} \min(D'_{ij}/D_{ij}, D_{ij}/D'_{ij})}{|\{i,j|D_{ij}<\infty\}|} \cdot \frac{\sum_{i=1}^3 \gamma_i}{\sum_{i=1}^n \gamma_i} \quad (5.9)$$

Note that the *Consensus Index* is between 0 and 1. When the *Consensus Index* trends to 1, this means that the input distance matrix fits a single 3D structure well. The result section showed that the *Consensus Index* is a good indicator on whether the input data corresponds to a single 3D structure or a mixture of 3D structures.

V-2.4 Searching for the Correct Conversion Factor

In Section V-2.2, the conversion factor $\alpha(> 0)$ is assumed to be known. However, the assumption is not valid in practice. Even worse, Lemma 1 shows that the conversion factor changes with different resolutions.

Lemma 1 Consider the frequency matrix F for loci x_1, \dots, x_{2n} . Let the conversion factor of F be $\alpha > 0$, i.e., distance between loci x_i and x_j is $d_{ij} = (1/F_{ij})^\alpha$. Now, we reduce the resolution by merging adjacent loci, i.e., we generate the frequency matrix F' for the low resolution loci y_1, \dots, y_n , where y_i is formed by merging adjacent loci x_{2i-1} and x_{2i} . Suppose $F'_{ij} = (F_{2i-1,2j-1} + F_{2i-1,2j} + F_{2i,2j-1} + F_{2i,2j})$ and d'_{ij} can be approximated as either arithmetic mean or geometry mean of $\{d_{2i-1,2j-1}, d_{2i-1,2j}, d_{2i,2j-1}, d_{2i,2j}\}$. Then the conversion factor α' of F' is less than or equal to α .

Proof. Note that $\log F_{p,q} > 0$ and $\log d_{p,q} < 0$ since $F_{p,q} \geq 1$. Let $d_{\min} = \min_{p \in \{2i, 2i-1\}, q \in \{2j, 2j-1\}} d_{p,q}$. Since $d'_{ij} \geq d_{\min}$, we have $\log d'_{ij} \geq \log d_{\min}$. We also have

$$F'_{ij} = \sum_{p \in \{2i, 2i-1\}, q \in \{2j, 2j-1\}} F_{p,q} = \sum_{p \in \{2i, 2i-1\}, q \in \{2j, 2j-1\}} \frac{1}{d_{p,q}^{1/\alpha}} \geq \frac{1}{d_{\min}^{1/\alpha}}$$

Hence $\log F'_{ij} \geq -\frac{1}{\alpha} \log d_{\min}$. As $d'_{ij} = (1/F'_{ij})^{\alpha'}$, we have

$$\alpha' = \frac{-\log d'_{ij}}{\log F'_{ij}} \leq \frac{-\log d_{\min}}{-\frac{1}{\alpha} \log d_{\min}} = \alpha.$$

Q. E. D

The Lemma 1 implies that the conversion factor of high-resolution Hi-C datasets is usually larger than that of low-resolution Hi-C datasets. Hence, we cannot assume that the conversion factor is a prior or is a fix value for different datasets. In fact, the predicted 3D structure is quite sensitive to the conversion factor. Given the

same frequency matrix, different conversion factor leads to different expected distances and finally implies very different 3D structures (Figure V-2). Therefore, estimating the correct conversion factor for a frequency matrix F is important.

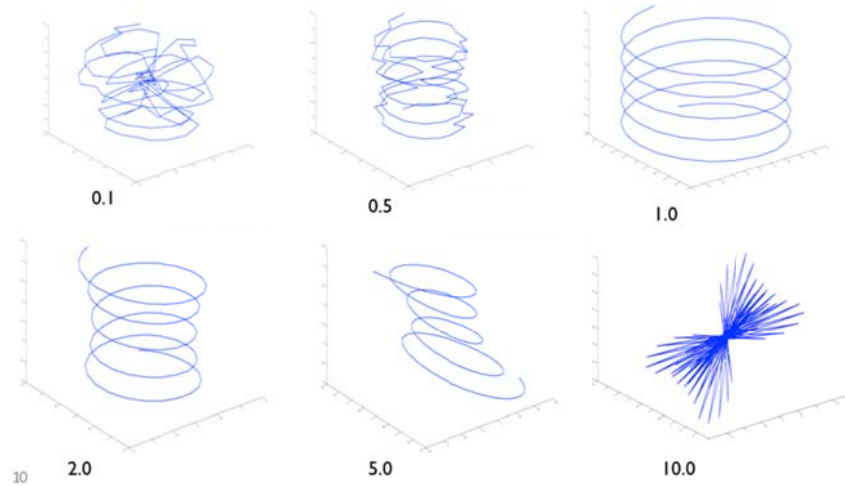


Figure V-2: **The effect of different conversion factors on the 3D structures predicted by ChromSDE.** The correct structure is a helix and the true value of conversion factor is 1. Each sub-figure is the predicted structure by ChromSDE with the specific conversion factor indicated below it.

A correct conversion factor enables us to convert a frequency matrix to a correct 3D model, and vice versa. Based on this principle, for a frequency matrix F , the goodness of a conversion factor α ($goodness(\alpha, F)$) can be determined by comparing the predicted frequency matrix \hat{F} and the input frequency matrix F . Figure V-4 details the function to compute $goodness(\alpha, F)$.

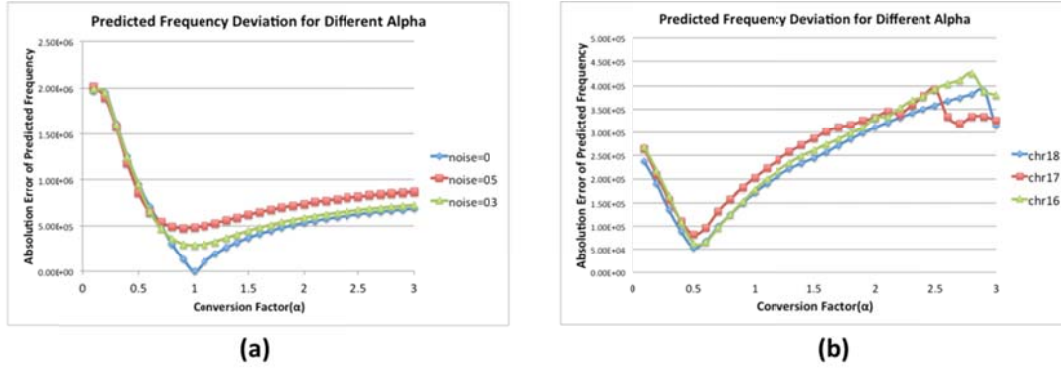


Figure V-3: **Absolution error of estimated frequency with different values of conversion factor.** (a) Simulation study: the data is generated by helix curve under different noise level and true conversion factor is 1. (b)Real Hi-C data: the data is from chromosome 16-18 of mESC Hind3 dataset.

Our aim is to compute α that maximizes the goodness function. As there is no obvious well defined gradient for the goodness function, we cannot use methods such as gradient descent or Newton's method to optimize α . Instead, we perform the golden section search method to optimal α , assuming that the goodness function is unimodal in the search interval. Since $d_{ij} = (1/F_{ij})^\alpha$, we deduce that α cannot be too small for otherwise the spatial distance will be independent of the frequency (when $\alpha \rightarrow 0$). Also, α cannot be too large, otherwise a small difference in frequencies will lead a very big difference in spatial distances, and small noise will seriously violate of the triangle inequality. In this chapter, we assume that $0.1 \leq \alpha \leq 3$. Moreover, we observed that applying the standard golden section search on the logarithm domain of the interval is more efficient (see Figure V-3). The algorithm detail is in Figure V-4.

Algorithm ChromSDE	
Require: normalized frequency matrix F	
Ensure: a set of 3D coordinates X , conversion factor α	
1: $\alpha_{min} = 0.1, \alpha_{max} = 3$	# set search boundary for α
2: $\varphi = \frac{\sqrt{5}-1}{2}$	# golden section ratio
3: repeat	
4: $\eta = (\frac{\alpha_{max}}{\alpha_{min}})^\varphi$	# step size for updating α
5: $x1 \leftarrow \alpha_{min} \cdot \eta, f1 \leftarrow goodness(x1, F)$	
6: $x2 \leftarrow \alpha_{max}/\eta, f2 \leftarrow goodness(x2, F)$	
7: if $f1 > f2$ then	
8: $\alpha_{min} \leftarrow x2$	# increase lower bound
9: else	
10: $\alpha_{max} \leftarrow x1$	# decrease upper bound
11: end if	
12: until $(\alpha_{max} - \alpha_{min}) < tolerance$	
13: $\alpha \leftarrow \alpha_{min}$	# final value of α
14: $D \leftarrow (1/F)^\alpha$	# expected distance matrix
15: $X \leftarrow$ compute 3D structure using SDP method based on D	
Function goodness (α, F)	
1: $D \leftarrow (1/F)^\alpha$	
2: $X \leftarrow$ compute 3D structure using SDP method based on D	
3: $D' \leftarrow$ compute pair-wise distances from X	
4: $F' \leftarrow (1/D')^{1/\alpha}$	
5: Return $\sum_{\{(i,j) F_{i,j}>0\}} - F'_{i,j} - F_{i,j} $	

Figure V-4: Algorithm description for ChromSDE.

V-3 Result

V-3.1 Simulation Study

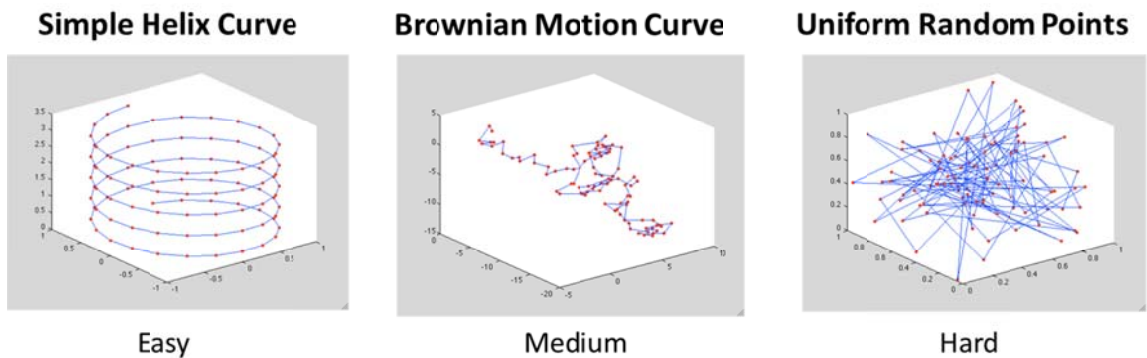


Figure V-5: Different types of structures used in the simulation study: Helix curve(Left), Brownian motion simulation of a single particle(Middle) and Uniform random points in a cube(Right).

To analyze the performance of ChromSDE, we generated three different types of 3D structures (Figure V-5): (1) Helix curve, (2) Brownian motion simulation of a single particle and (3) Uniform random points in a cube. Each structure is represented

by 100 points. We assume that the Hi-C technique is sensitive enough to capture interactions with at most 50 nearest neighbours and the conversion factor α is 1, i.e., the contact frequency f of two given points can be computed as $f = (1/d)^{1/\alpha} = 1/d$, where d is the spatial distance between given points. We compared our algorithm with the existing methods MCMC5C[101] and BACH[51], which are the only publicly available standalone programs that are suitable for general Hi-C data. For MCMC5C, it cannot estimate the conversion factor by itself, so we supplied it with the correct value. For BACH, it can estimate the conversion factor with the default starting point equal to 1 (i.e., the correct value in our simulation study). Since there is no enzyme bias in our simulation, we also modified BACH to suppress this feature (called BACH*). For ChromSDE, we just assume that the conversion factor is within the range (0.1, 3), so we give advantages to the existing programs, but not our ChromSDE.

V.3.1.1 ChromSDE guarantees optimality in noise-free case

Figure V-6 shows the true simulated structures and the predicted structures by different programs. For the helix curve, all three programs can recover the structure correctly. For the Brownian motion curve, both ChromSDE and MCMC5C can almost perfectly recover the true structure and BACH* can only reproduce a not-so-accurate but similar structure. For the third case, MCMC5C produced a not-so-accurate structure and BACH* completely failed in this case, while our ChromSDE still can perfectly recover the true structure. The result is not surprising since SDP method is the only one that can guarantee perfect recovery of the true structure when the input data is noise-free and the structure is uniquely localizable. Based on the

RMSD(root mean square deviation), ChromSDE also outperforms the other two methods in all the three simulated cases.

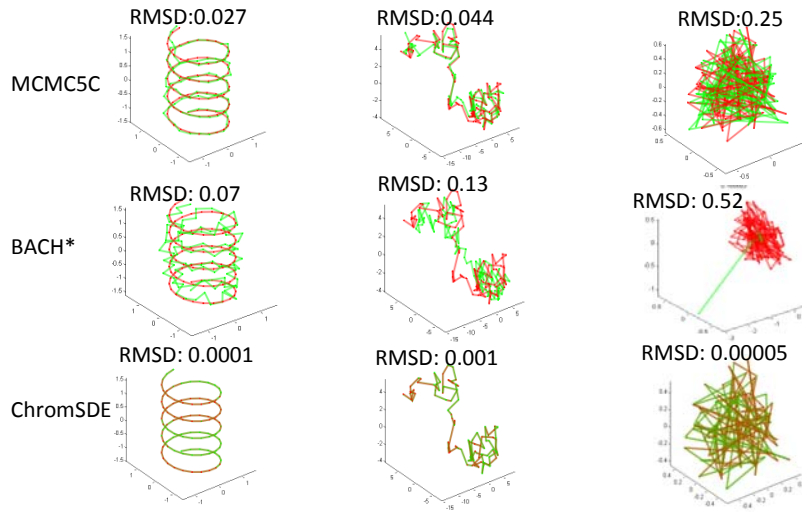


Figure V-6. **Predicted 3D structures by different programs using simulated data.** The Red curve is the true structure and the green curve is the predicted structure. ChromSDE uses quadratic SDP here and the linear SDP has the same performance.

V.3.1.2 ChromSDE outperforms the existing methods in noisy-data

The previous section showed that ChromSDE could recover the optimal chromatin structure in the noise-free case. Now, we test whether ChromSDE is robust in a noisy data setting. To study this, we simulated noisy contact frequency data in different noise level based on the Brownian curve structure. For any two loci i and j , the noisy frequency \tilde{F}_{ij} is deviated from the true frequency $F_{ij} = 1/D_{ij}$ (D_{ij} is the spatial distance between loci i and j) by adding a uniform random noise δ within a given noise level. Precisely, $\tilde{F}_{ij} = F_{ij}(1 + \delta)$ where $|\delta|$ is smaller than the noise level.

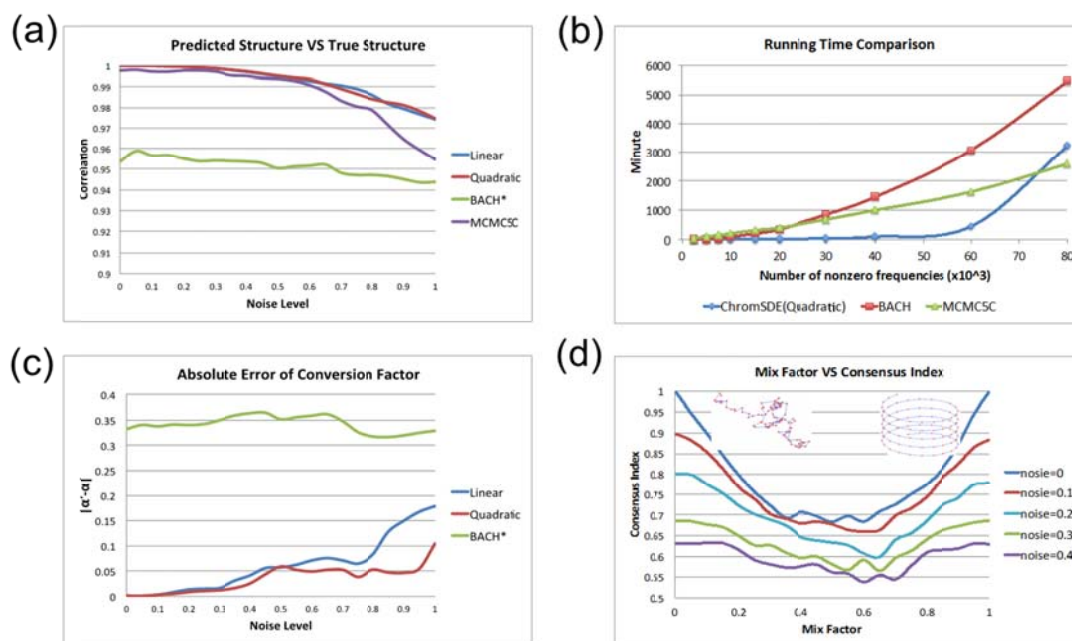


Figure V-7. **Performance of different methods on simulated data.** (a) Spearman correlation between the pair-wise distance matrices of the predicted structure and the true structure under different noise level. (b) Running times of tested programs given different number of pairs of observed frequency (test stop at 80000 pair-wise frequencies, ~ 1600 points). (c) The absolute error of the estimated value of conversion factor under different noise levels. (d) The Consensus Index predicted by ChromSDE (quadratic model) under different degree of mixture of helix curve(right) and Brownian motion curve(left).

Figure V-7 shows the performance of the programs with different noise levels under different measurements. Figure V-7(a) shows that, when the noise level increases, the Spearman correlation between the pairwise distances from the predicted structure and those from the true structure generally decreases. ChromSDE and MCMC5C perform similarly when the noise level < 0.6 and ChromSDE (both linear SDP and quadratic SDP) outperforms others when the noise level is higher than 0.6. Similar result is observed when we measure the RMSD between the predicted structure and the true structure (Figure V-8 (a)). In Figure V-7(c), we observed that ChromSDE estimated the conversion factor quite accurately (deviation < 0.1) when

noise level <0.7 . In contrast, the estimated conversion factor from BACH* tends to be incorrect (deviation around 0.35). This may be the reason why BACH* has worse performance comparing to others across different noise levels. Moreover, ChromSDE is faster than BACH and comparable to MCMC5C even though ChromSDE needs to search for the correct conversion factor but MCMC5C does not (Figure V-7(b)). In summary, the result shows that the linear SDP and quadratic SDP models perform quite consistently and ChromSDE is more robust and accurate than existing methods.

V.3.1.3 *Consensus Index* indicates the degree of mixture of 3D structures

In Hi-C and TCC experiments, the data is from a population of cells, and each potentially has different 3D chromosomal structure. The method section proposed to use the *Consensus Index* to determine if the data is from a consensus 3D structure. To show that the *Consensus Index* is a good indicator of the degree of mixture, we generated a frequency matrix F_{merge} by merging the frequency matrix from the helix curve F_1 and the Brownian motion curve F_2 under different mix factor γ (i.e., $F_{merge} = \gamma F_1 + (1-\gamma)F_2$). Figure V-7(d) shows that the *Consensus Index* is affected by both the noise level and mix factor. For the same noise level, the *Consensus Index* approaches the minimum when the mix factor is close to 0.5. This indicates that the *Consensus Index* is the lowest when the two structures are highly mixed. For different noise levels, the *Consensus Index* decreases as the noise level increases. From Figure V-7 (d), we can estimate a lower bound for the percentage of the dominant 3D structure by examining the curve of noise level=0. Also we note that the estimated conversion factors by ChromSDE are quite consistent with its true value even under different mix factors and noise levels (Figure V-8 (c)).

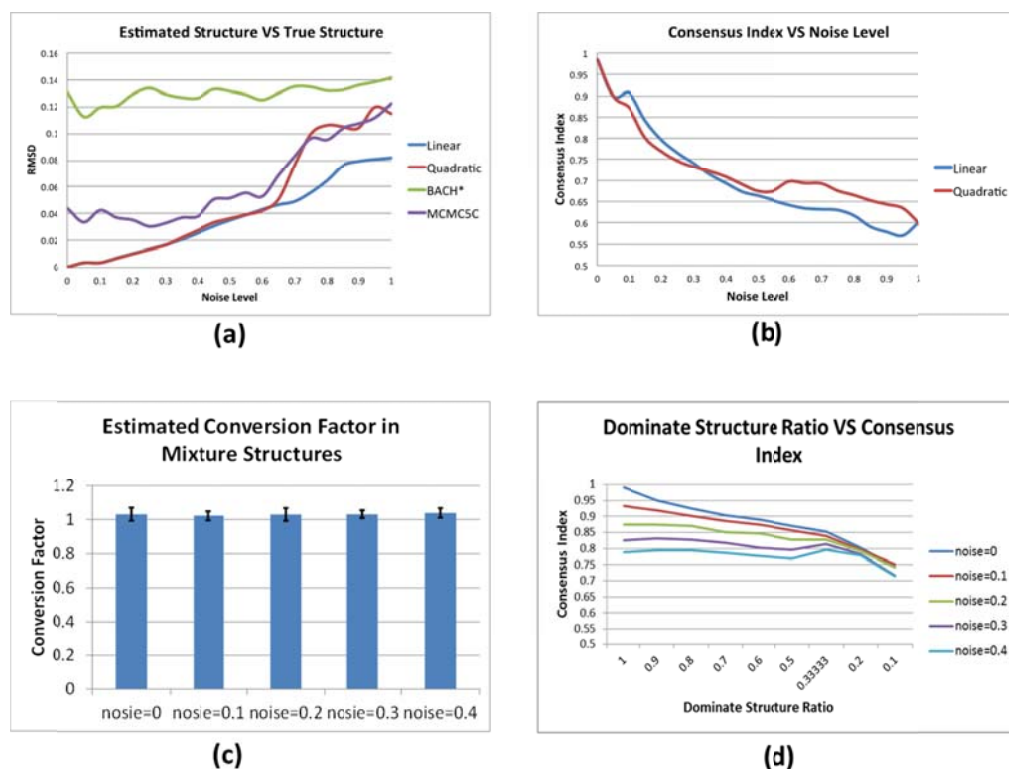


Figure V-8: **Simulation result that compares the performance of different methods.** (a) Root mean square deviation (RMSD) between the predicted structure and true structure (Brownian curve). Generally, RMSD increases as the noise level increases. The linear SDP and quadratic SDP of ChromSDE performs similarly below noise level 0.7 and better than other methods. (b) Consensus Index predicted by two SDP formulations under different noise levels. Generally, the consensus index decreases as the noise level increases. (c) The value of the average conversion factor of different mix factors in Figure 3(d) for given noise level and the error bar represents the standard deviation of estimated conversion factor. (d) Consensus index decreases when the proportion of the dominate structure in the mixture decreases or noise level increases. For dominate structure ratio 1 to 0.5, the results were generated by simulating two Brownian structures with mix factor 1 to 0.5 correspondingly. For dominate structure ratio 0.333, 0.2 and 0.1, the results were generated by simulating a mixture with 3, 5 and 10 Brownian structures with equal proportion correspondingly.

V-3.2 Real Hi-C Data Study

V.3.2.1 Validate ChromSDE using two enzyme replicates

From the literature, two different enzymes (Hind3, NcoI) were used to generate Hi-C replicate data from the mouse ES cell (mESC)[25] and the human

GM06990 cell(GM)[76]. Each enzyme replicate is an independent observation of the chromosome structure in the same cell type. Hence, we expect the result produced by a robust algorithm using one enzyme data can be validated using the other enzyme data.

We applied four different programs ChromSDE, BACH*, BACH and MCMC5C to predict the 3D structures of different chromosomes in the two cell lines using the Hi-C data from two replicates. For ChromSDE, BACH* and MCMC5C, the input is a normalized frequency matrix using the normalization pipeline by Yaffe and Tanay [133]. For BACH, we provide the raw Hi-C frequency and enzyme cutting point feature data.

Table V-1: **The conversion factors estimated by ChromSDP and BACH.**

Each table element is the mean value of the estimated conversion factor across all chromosomes, and the value in each bracket is the standard deviation of the corresponding mean.

Conversion Factor Estimation				
	Quadratic SDP	Linear SDP	BACH	BACH*
mESC_NcoI	0.5455(0.0167)	0.5437(0.0153)	0.4130(0.0129)	0.4285(0.0439)
mESC_Hind3	0.5354(0.0145)	0.5390(0.0183)	0.4182(0.0143)	0.4408(0.0902)
GM_NcoI	0.6284(0.0489)	0.5780(0.0382)	0.5942(0.0487)	0.7078(0.3104)
GM_Hind3	0.6381(0.0568)	0.6075(0.0490)	0.7342(0.3487)	0.8818(0.6990)

We compute Spearman correlation between the normalized frequency of one enzyme data and the estimated frequency ($frequency \sim 1 / distance$) of the predicted structure from the other enzyme data. (We use Spearman correlation instead of Pearson correlation since the Spearman correlation is independent to the conversion between frequency and distance; hence it is fair to every tested program.) Figure V-9(a) shows that ChromSDE (both Linear SDP and Quadratic SDP) outperforms the other programs by at least 5% across all four tested Hi-C datasets. Especially, in the

mESC dataset, ChromSDE obtains the average correlation of 0.9 across all chromosomes but other tested programs only obtain correlation at most 0.82. What's more, Figure V-9(b),(c) and Figure V-10 showed the 3D structures of different chromosomes predicted by ChromSDE are highly reproducible and the conversion factors estimated by ChromSDE are more consistent than the ones estimated by BACH and BACH* across different chromosomes and different enzymes (Table V-1).

(a)

Validate Correlation	Quadratic SDP	Linear SDP	BACH	BACH*	MCMC5C
mESC_NcoI	0.9096	0.9095	0.7918	0.7871	0.8187
mESC_Hind3	0.9080	0.9070	0.8061	0.7999	0.7676
GM_NcoI	0.7674	0.7899	0.6740	0.7244	0.6942
GM_Hind3	0.7832	0.7763	0.7066	0.7358	0.7293

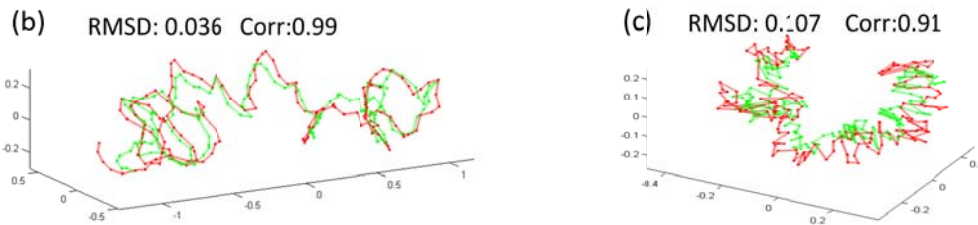


Figure V-9: **Validate ChromSDE using mESC, GM Hi-C data with two different enzymes (Hind3, NcoI).** (a) Average Spearman correlation across all chromosomes between inverse 3D distance and contact frequency from testing dataset. For each dataset, the best performer is highlighted. (b) Alignment between predicted structures of chromosome 1 of mESC Hind3(red) and mESC NcoI(green) by ChromSDE. (c) Alignment between predicted structures of chromosome 1 of GM Hind3(red) and GM NcoI(green) by ChromSDE.

Besides, we observed that all the tested programs perform worse in GM than in mESC and the *Consensus Index* is around 0.9 in mESC and is only 0.7 in GM (Figure V-11). It indicates that mESC has a consensus 3D structure for its genome and GM is relatively diverse or has higher noise level due to the low sequencing depth.

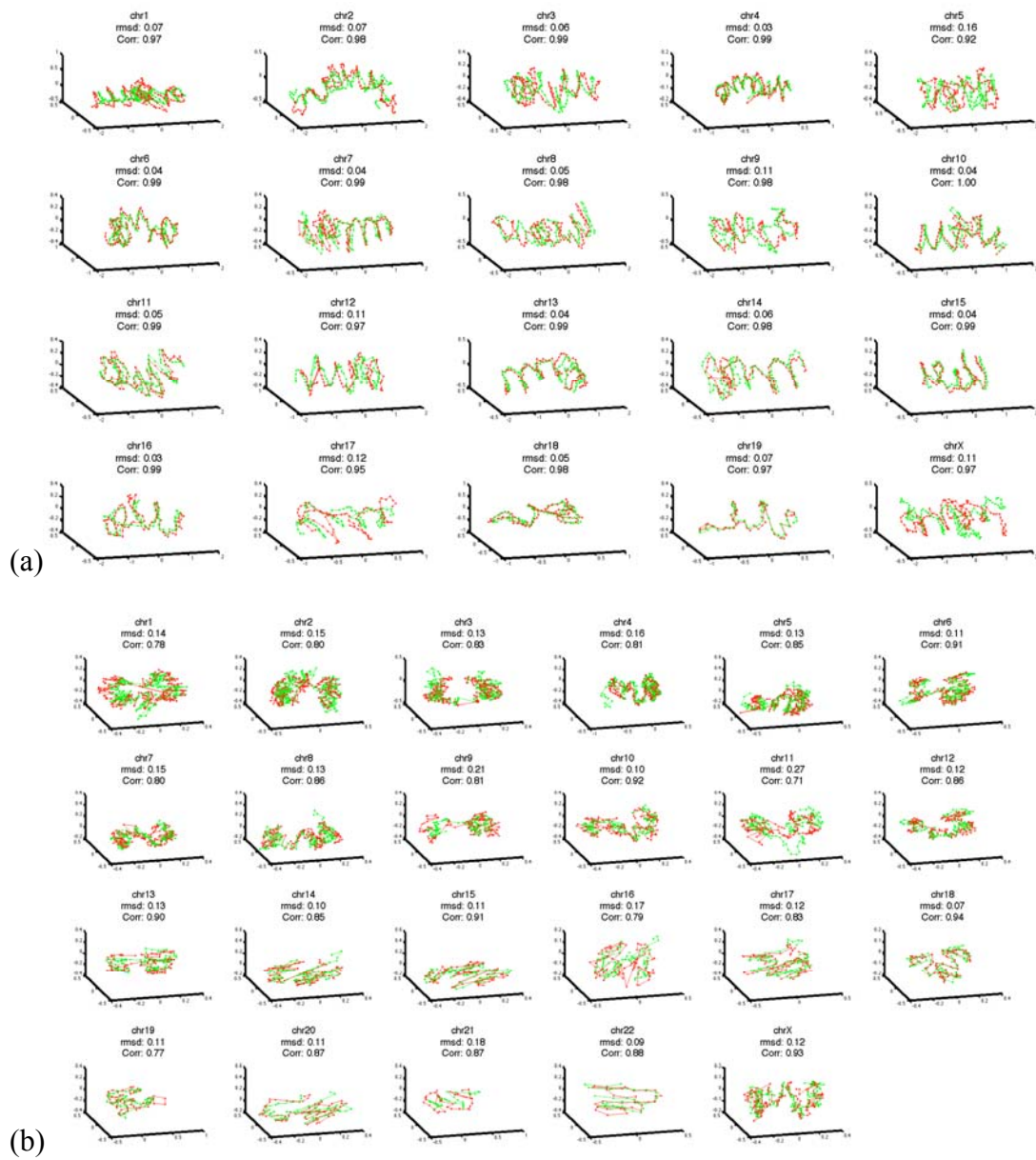


Figure V-10: **3D structures predicted by ChromSDE using different enzyme data** (red: Hind3, green : NcoI). The 3D structures are built using 1Mbp resolution data, and quadratic SDP . (a) mouse ES cell. (b) human GM cell.

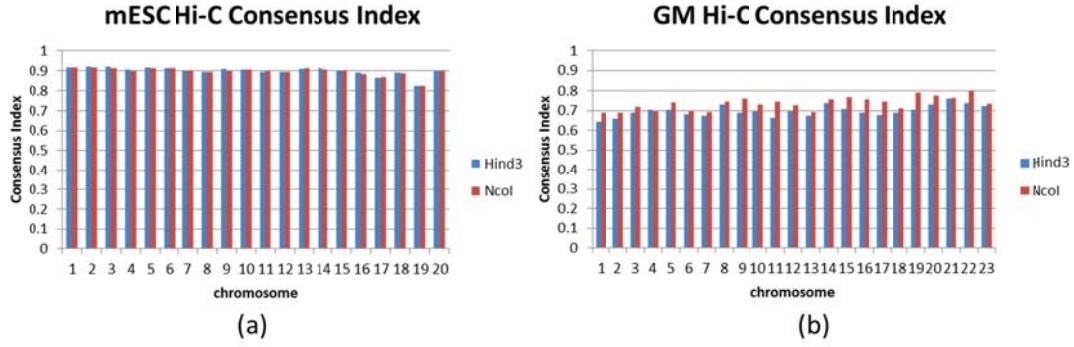


Figure V-11: **The consensus indices for different Hi-C datasets.** (a) The consensus indices estimated across different chromosomes using Hind3 enzyme (blue) and NcoI enzyme (red) in mESC Hi-C datasets. (b) The consensus indices estimated across different chromosomes using Hind3 enzyme (blue) and NcoI enzyme (red) in GM Hi-C datasets.

V.3.2.2 ChromSDE can generate consistent 3D structures from different genomic resolutions

We further tested ChromSDE on different genomic resolution data. Figure V-12(a) showed that ChromSDE can predict similar structures of chromosome 13 under different resolutions using mESC Hind3 data (average Spearman correlation is 0.97, average RMSD is 0.08). In contrast, other existing programs cannot reproduce similar structures with different resolution data, especially for MCMC5C which cannot estimate the correct conversion factor (Figure V-13). We also showed that the conversion factor for each predicted structure in Figure V-12(a). It demonstrated that the conversion factor increases as the resolution increases (also supported by BACH in Figure V-13). This further confirms the correctness of Lemma 1 even though the frequency has been normalized under different genomic resolutions.

To demonstrate the application of our predicted 3D structure, we generated a high resolution chromosome 3D structure for the region chr13:21Mb-25Mb (Figure V-12 (b)) using ChromSDE and mouse ES cell Hind3 data (40kbp resolution, estimated α is 0.83). Hist1h genes are highlighted with yellow color in the 3D

structure, and we find that two groups of Hist1h genes are separated quite far away (~1.5Mbp) in the linear genomic locations. In contrast, the promoters of two groups of Hist1h genes are spatially close to each other. To test if these two groups of genes interact each other for transcription, we checked the Pol2 ChIA-PET data available in our lab. We found that there are strong interactions (red dash line) between these two promoter regions mediated by Pol2, which indicates that the histone genes are co-regulated in the mouse ES cell.

Moreover, we found that the dense region and the loose region in the predicted 3D structure can be used to indicate the level of activity of those regions (from the snapshot of UCSC genome browser [63]). Dense regions (purple and blue color) correspond to repressive chromatin state in the cell, and there are few active histone modification and transcription factor-binding events occurring in those regions. In contrast, loose regions (green and yellow color) correspond to active chromatin state in the cell, and there are a lot of histone modification and transcription factor-binding events occurring in those regions. Also, we found that loose regions usually containing more genes and are associated with early replication timing than the dense regions. It is also noted that the purple region is associated with LaminB1 binding and late replication timing, which suggests that Lamin may plays a part in the histone genes regulation and DNA replication.

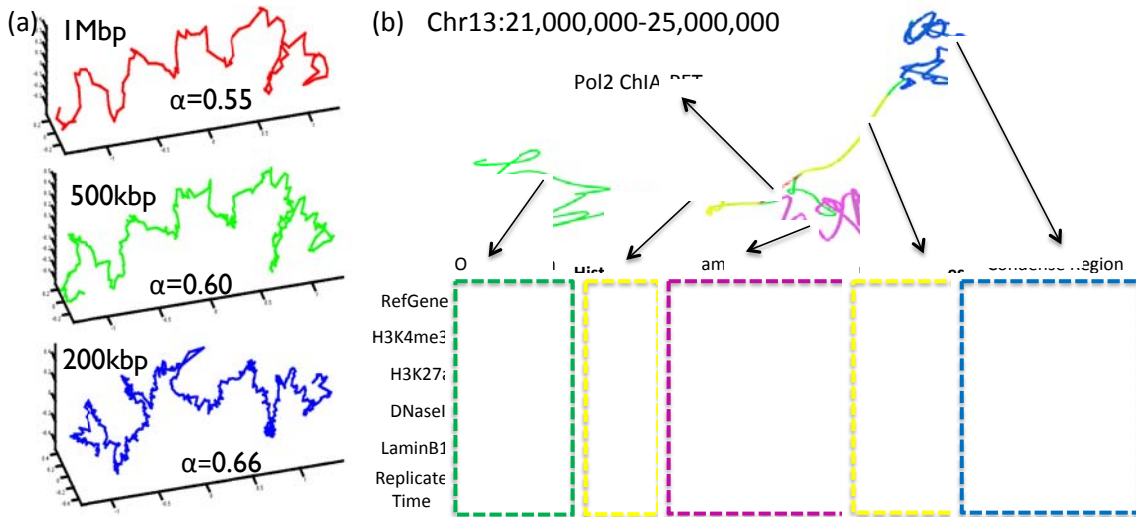


Figure V-12. **Predicted structure of chromosome 13 from mESC Hind3 data.** (a) The predicted structure of chromosome 13 under 1Mbp,500kbp,200kbp resolutions. (b) The predicted structure of the region chr13:21Mb-25Mb under 40kbp resolution and the different signal tracks of mESC from UCSC genome browser [63].

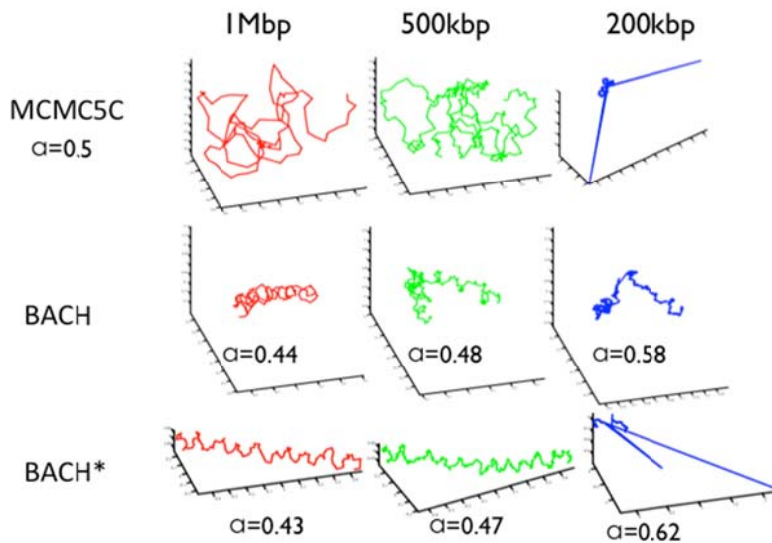


Figure V-13: **The performance of existing methods under data of different resolutions.** The conversion factor of MCMC5C is set to be 0.5 for different resolution, and its predicted structures under different resolutions are not so similar. The conversion factors predicted by BACH and BACH* increase when the resolution increases.

V-4 Discussion

In this chapter, we presented a method ChromSDE to reconstruct the consensus/dominate chromatin 3D structure of the given HiC data. To our best knowledge, ChromSDE is the only method, which can guarantee recovering the correct structure in the noise-free case. In the noisy case, ChromSDE is much more accurate and robust than existing methods in both simulation and real data study. In addition, ChromSDE can automatically estimate the conversion factor, which is proved to change under different resolutions theoretically and empirically. Furthermore, we demonstrate that interesting biological findings can be uncovered from our predicted 3D structure.

We also developed the *Consensus Index* to determine how good the data can be explained by a single 3D structure. However, *Consensus Index* may not be informative when the noise level of the data is high or the mixing structures are similar. When the mixing structures are similar to each other then ChromSDE will learn the average structure. One future research is to recover all the mixing structures using Hi-C data.

There are some possible limitations for this study. Due to the dynamics and heterogeneity, the predicted structure from Hi-C data may be quite different from the real chromosome structure [119], although we believe it retains some statistical spatial features. And in this study, we only consider the intra-chromosome contact frequency and single chromosome modeling. However, the inter-chromosome contact may also affect the prediction of single chromosome structure, especially considering the volume exclusion effect among different chromosomes[119].

CHAPTER - VI Conclusions and Future Directions

VI-1 Conclusion

This thesis explored a new set of wet lab experimental data for protein-DNA interaction including PBM, ChIP-seq, Hi-C and ChIA-PET. It also studied two levels of protein-DNA interaction, namely, sequence and structure. At the sequence level, each chromosome is treated as a one-dimensional sequence, and each element (the site on DNA bound by protein) is encoded by its one-dimensional position in the chromosome (genomic location). At the structure level, each chromosome has a three-dimensional structure in the nucleus, and each element on the chromosome is encoded by a three-dimensional coordinate (spatial location).

Two computational problems in sequence level were presented in this thesis: motif enrichment analysis and *de novo* motif finding. Although they are classic bioinformatics problems, the new-generation data (ChIP-seq) provides the statistical power to solve more challenging problem (i.e., finding collaborated binding protein motifs) because the data is much higher resolution and higher throughput than previous generation data.

In our work for motif enrichment analysis, a motif enrichment analysis program for ChIP-seq called CENTDIST was developed, which is described in Chapter III and published in [135]. The performance of motif enrichment analysis methods is heavily dependent on selecting the proper background and other parameter settings. Comparing with existing methods, CENTDIST is a background-free approach and utilizes frequency information as well as slope information (velocity) of the motif occurrence distribution around ChIP-seq peak to predict whether a motif is

enriched or not. We examined CENTDIST on 13 ES cell ChIP-seq datasets and demonstrated that it is better than existing methods, thus showing that good result can be obtained without requiring expert knowledge in configuring the program. This approach has been taken as the first step in integrating the automatic parameter tuning technique for solving the general enrichment analysis problem in the bioinformatics field. However, it should be noted that the proposed parameter tuning technique and scoring function proposed here may not be the best one. The parameter tuning technique ignores the bias in the multiple testing and the hybrid scoring function cannot directly associate with the common statistical measure like p-value. Thus, the framework can be improved by including multi-test correction and probabilistic modeling.

In our work for *de novo* motif finding, a novel motif finding program called SEME was developed, which is described in Chapter IV and published in Zhang, et al. [137]. SEME can automatically utilize positional bias and sequence rank bias in many experimental data (e.g., ChIP-seq, ChIP-chip and Promoter sequence) to improve the quality of the discovered motifs. In the task of finding co-TF motif in the ChIP-seq data, SEME not only reports more accurate co-TF motifs than other programs but also correctly estimates the position and sequence rank distribution of each co-TF's motif. Such information provides useful insights on the interaction between the ChIPed TF and the predicted co-TFs, like interaction distance or indirect binding. A most important feature of SEME is that it does not rely on prior knowledge and applies unsupervised learning to let the data to tell its own story. It should be noted that our method requires enough sequence data in order to make a robust parameter estimation,

and will over-fit the data if the number of sequences is small, in which case more prior information is needed.

The last research problem in this thesis is to reconstruct the three-dimension structure of chromosomes based on chromatin interaction (Hi-C) data. Recently, a few works have been proposed to build 3D model of genome using chromatin interaction, and all of them used hybrid heuristic to solve a non-convex optimization, which are not guaranteed to reconstruct the correct structure even in the noise-free case. To fill-in the gap, we proposed a novel chromosome structure modeling algorithm called ChromSDE, which is a semi-definite programming (SDP) relaxation for the original non-convex optimization problem, and is guaranteed to recover the correct 3D structure in the noise-free case when the structure is uniquely localizable. Further, we proved that the parameter of conversion from contact frequency to spatial distance will change under different resolutions theoretically and empirically. Comparing to existing methods, ChromSDE does not assume the conversion parameter is known or fixed, but search the correct value of it based on the input data. Our result indicates that 3D structure can provide novel information for the spatial organization such as co-expression of far-away genes, different histone mark in condense or sparse regions, which are hidden in the linear view of the chromosome. The novel algorithm presented here is of considerable importance since it is one of the most theoretically sound and practical methods, which can translate the chromatin interaction data directly into 3D structure and makes a worthwhile contribution towards understanding genetic mechanism in the 3D perspective. However, it should be pointed out that, the current chromatin interaction data cannot differentiate the sister chromosomes and different cell cycles. So the predicted structure in this study may not reflect the true

structure of an individual chromosome when the structures from two sister chromosomes or from different cell cycles are different.

In summary, the methods developed in this study have unlocked the potential provided by the new generation sequencing data of protein-DNA interaction and gave more in-depth understanding for biological mechanism compared to the existing methods.

VI-2 Future works

The proposed research problems in this thesis are current hot research topics. On one hand, with the new generation sequencing data, the focuses of some classic bioinformatics problem like motif enrichment and de novo motif finding have been shifted to the collaborative transcription factors instead of the main transcription factor. And there is still much room to improve the current algorithms to fulfill the new focuses. On the other hand, for the newborn bioinformatics problems like chromosome 3D modeling, there are a lot of unexplored (and not well-defined) directions for further investigation. Hence, I list some directions related to the research problems in this thesis, which I think are worth further study.

1. Identifying co-TF through motif enrichment analysis can be further improved by incorporating the peak rank distribution. CENTDIST and other ChIP-based algorithms such as SpaMo have utilized the position distribution of the motif, but not peak rank distribution. When ChIP peaks are sorted by ChIP intensity, the low ranking ChIP peaks (low ChIP intensity) usually indicates weak binding or indirect binding. If a motif specially enriches in the low ranking ChIP peaks, it is possible to belong to a co-TF bound by the ChIPed TF.

Hence, based on the position distribution and peak rank distribution, it is possible to define different types of interactions between the ChIPed TF and co-TF (i.e., co-binding or indirect binding).

2. Identifying co-TF motifs through *de novo* motif finding can be further improved by categorizing ChIP peaks. Since co-TF only occurs in the subsets of the ChIP peaks, so it is easier to identify co-TF motif if we can correctly partition the input ChIP peaks into different subsets. For example, the set of peaks can be partitioned into two sets: one with the ChIPed TF motif and the other without ChIPed TF motif. For the peaks with ChIPed TF motif, we can extract the DNA sequence around the ChIPed TF motif position, and apply SEME to identify co-TF motifs. For the peaks without ChIPed TF motif, they are usually less confident, and it will be helpful to incorporate other information like evolutionary conservation and performs motif finding only on the high confident regions.
3. There is one open question in chromosome 3D modeling, that is, whether there exists a consensus 3D structure for a chromosome. Although *consensus index* has been proposed in Chapter V, the solution for the mixture structures are still not well developed. Mathematically, it is a very challenging problem, even the problem is relaxed to high dimension as in ChromSDE. However, it is possible to solve this problem by giving some prior information on the interaction. That is, if each interaction can be annotated to belong to which cell state, then the 3D structure for each state can be constructed using ChromSDE on the corresponding subset of spatial distance constraints.

REFERENCES

- [1] M. ANNALA, K. LAURILA, H. LAHDESMAKI and M. NYKTER, *A linear model for transcription factor binding affinity prediction in protein binding microarrays*, PloS one, 6 (2011), pp. e20059.
- [2] O. APARICIO, J. V. GEISBERG and K. STRUHL, *Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo*, Curr Protoc Cell Biol, Chapter 17 (2004), pp. Unit 17 7.
- [3] M. ARITA and S. KOBAYASHI, *DNA sequence design using templates*, New Generation Computing, 20 (2002), pp. 263-277.
- [4] T. L. BAILEY, *DREME: Motif discovery in transcription factor ChIP-seq data*, Bioinformatics, 27 (2011), pp. 1653.
- [5] T. L. BAILEY, M. BODEN, F. A. BUSKE, M. FRITH, C. E. GRANT, L. CLEMENTI, J. REN, W. W. LI and W. S. NOBLE, *MEME SUITE: tools for motif discovery and searching*, Nucleic Acids Res, 37 (2009), pp. W202-8.
- [6] T. L. BAILEY and C. ELKAN, *Fitting a mixture model by expectation maximization to discover motifs in biopolymers*, Proc. Int. Conf. Intell. Syst. Mol. Biol, 1994, pp. 28-36.
- [7] Y. BARASH, G. ELIDAN, N. FRIEDMAN and T. KAPLAN, *Modeling dependencies in protein-DNA binding sites*, ACM, 2003, pp. 28-37.
- [8] A. BARSKI, S. CUDDAPAH, K. CUI, T. Y. ROH, D. E. SCHONES, Z. WANG, G. WEI, I. CHEPELEV and K. ZHAO, *High-resolution profiling of histone methylations in the human genome*, Cell, 129 (2007), pp. 823-37.
- [9] D. BAU and M. A. MARTI-RENO, *Genome structure determination via 3C-based data integration by the Integrative Modeling Platform*, Methods (2012).
- [10] D. BAU and M. A. MARTI-RENO, *Structure determination of genomic domains by satisfaction of spatial restraints*, Chromosome Research, 19 (2011), pp. 25-35.
- [11] I. BEN-GAL, A. SHANI, A. GOHR, J. GRAU, S. ARVIV, A. SHMILOVICI, S. POSCH and I. GROSSE, *Identification of transcription factor binding sites with variable-order Bayesian networks*, Bioinformatics, 21 (2005), pp. 2657-2666.
- [12] M. F. BERGER and M. L. BULYK, *Protein binding microarrays (PBMs) for rapid, high-throughput characterization of the sequence specificities of DNA binding proteins*, METHODS IN MOLECULAR BIOLOGY-CLIFTON THEN TOTOWA-, 338 (2006), pp. 245.
- [13] S. L. BERGER, *Histone modifications in transcriptional regulation*, Current opinion in genetics & development, 12 (2002), pp. 142-148.
- [14] P. BISWAS, T. C. LIANG, K. C. TOH, Y. YE and T. C. WANG, *Semidefinite programming approaches for sensor network localization with noisy distance measurements*, Automation Science and Engineering, IEEE Transactions on, 3 (2006), pp. 360-371.
- [15] J. S. CARROLL, X. S. LIU, A. S. BRODSKY, W. LI, C. A. MEYER, A. J. SZARY, J. ECKHOUTE, W. SHAO, E. V. HESTERMANN, T. R. GEISTLINGER, E. A. FOX, P. A. SILVER and M. BROWN, *Chromosome-*

- wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1*, Cell, 122 (2005), pp. 33-43.
- [16] X. CHEN, H. XU, P. YUAN, F. FANG, M. HUSS, V. B. VEGA, E. WONG, Y. L. ORLOV, W. ZHANG, J. JIANG, Y. H. LOH, H. C. YEO, Z. X. YEO, V. NARANG, K. R. GOVINDARAJAN, B. LEONG, A. SHAHAB, Y. RUAN, G. BOURQUE, W. K. SUNG, N. D. CLARKE, C. L. WEI and H. H. NG, *Integration of external signaling pathways with the core transcriptional network in embryonic stem cells*, Cell, 133 (2008), pp. 1106-17.
- [17] X. CHEN, H. XU, P. YUAN, F. FANG, M. HUSS, V. B. VEGA, E. WONG, Y. L. ORLOV, W. ZHANG, J. JIANG and OTHERS, *Integration of external signaling pathways with the core transcriptional network in embryonic stem cells*, Cell, 133 (2008), pp. 1106-1117.
- [18] E. CHEUNG and W. L. KRAUS, *Genomic analyses of hormone signaling and gene regulation*, Annual review of physiology, 72 (2010), pp. 191-218.
- [19] S. CHOUDHURI, *Gene regulation and molecular toxicology*, Toxicol Mech Methods, 15 (2004), pp. 1-23.
- [20] J. M. CLAVERIE and S. AUDIC, *The statistical significance of nucleotide position-weight matrix matches*, Computer applications in the biosciences : CABIOS, 12 (1996), pp. 431-9.
- [21] F. CRICK, *Central dogma of molecular biology*, Nature, 227 (1970), pp. 561-3.
- [22] S. CUDDAPAH, R. JOTHI, D. E. SCHONES, T. Y. ROH, K. CUI and K. ZHAO, *Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains*, Genome research, 19 (2009), pp. 24-32.
- [23] J. DEKKER, *Gene regulation in the third dimension*, Science, 319 (2008), pp. 1793-4.
- [24] J. DEKKER, K. RIPPE, M. DEKKER and N. KLECKNER, *Capturing chromosome conformation*, Science, 295 (2002), pp. 1306-11.
- [25] J. R. DIXON, S. SELVARAJ, F. YUE, A. KIM, Y. LI, Y. SHEN, M. HU, J. S. LIU and B. REN, *Topological domains in mammalian genomes identified by analysis of chromatin interactions*, Nature, 485 (2012), pp. 376-380.
- [26] J. DOSTIE and J. DEKKER, *Mapping networks of physical interactions between genomic elements using 5C technology*, Nature protocols, 2 (2007), pp. 988-1002.
- [27] Z. DUAN, M. ANDRONESCU, K. SCHUTZ, S. MCILWAIN, Y. J. KIM, C. LEE, J. SHENDURE, S. FIELDS, C. A. BLAU and W. S. NOBLE, *A three-dimensional model of the yeast genome*, Nature, 465 (2010), pp. 363-7.
- [28] E. EDEN, D. LIPSON, S. YOGEV and Z. YAKHINI, *Discovering motifs in ranked lists of DNA sequences*, PLoS Comput Biol, 3 (2007), pp. e39.
- [29] R. ELKON, C. LINHART, R. SHARAN, R. SHAMIR and Y. SHILOH, *Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells*, Genome Res, 13 (2003), pp. 773-80.
- [30] L. ETTWILLER, B. PATEN, M. RAMIALISON, E. BIRNEY and J. WITTBRODT, *Trawler: de novo regulatory motif discovery pipeline for chromatin immunoprecipitation*, Nature Methods, 4 (2007), pp. 563-565.
- [31] G. M. EUSKIRCHEN, J. S. ROZOWSKY, C. L. WEI, W. H. LEE, Z. D. ZHANG, S. HARTMAN, O. EMANUELSSON, V. STOLC, S. WEISSMAN,

- M. B. GERSTEIN and OTHERS, *Mapping of transcription factor binding regions in mammalian cells by ChIP: comparison of array-and sequencing-based technologies*, *Genome Res*, 17 (2007), pp. 898.
- [32] E. FIELLER, H. HARTLEY and E. PEARSON, *Tests for rank correlation coefficients. I*, *Biometrika*, 44 (1957), pp. 470-481.
- [33] P. FRASER and W. BICKMORE, *Nuclear organization of the genome and the potential for gene regulation*, *Nature*, 447 (2007), pp. 413-7.
- [34] E. FRATKIN, B. T. NAUGHTON, D. L. BRUTLAG and S. BATZOGLOU, *MotifCut: regulatory motifs finding with maximum density subgraphs*, *Bioinformatics*, 22 (2006), pp. e150-7.
- [35] M. C. FRITH, Y. FU, L. YU, J. F. CHEN, U. HANSEN and Z. WENG, *Detection of functional DNA motifs via statistical over-representation*, *Nucleic Acids Res*, 32 (2004), pp. 1372-81.
- [36] M. C. FRITH, U. HANSEN, J. L. SPOUGE and Z. WENG, *Finding functional sequence elements by multiple local alignment*, *Nucleic Acids Res*, 32 (2004), pp. 189.
- [37] M. J. FULLWOOD, M. H. LIU, Y. F. PAN, J. LIU, H. XU, Y. B. MOHAMED, Y. L. ORLOV, S. VELKOV, A. HO, P. H. MEI, E. G. CHEW, P. Y. HUANG, W. J. WELBORN, Y. HAN, H. S. OOI, P. N. ARIYARATNE, V. B. VEGA, Y. LUO, P. Y. TAN, P. Y. CHOY, K. D. WANSA, B. ZHAO, K. S. LIM, S. C. LEOW, J. S. YOW, R. JOSEPH, H. LI, K. V. DESAI, J. S. THOMSEN, Y. K. LEE, R. K. KARUTURI, T. HERVE, G. BOURQUE, H. G. STUNNENBERG, X. RUAN, V. CACHEUX-RATABOUL, W. K. SUNG, E. T. LIU, C. L. WEI, E. CHEUNG and Y. RUAN, *An oestrogen-receptor-alpha-bound human chromatin interactome*, *Nature*, 462 (2009), pp. 58-64.
- [38] N. GAO, J. ZHANG, M. A. RAO, T. C. CASE, J. MIROSEVICH, Y. WANG, R. JIN, A. GUPTA, P. S. RENNIE and R. J. MATUSIK, *The role of hepatocyte nuclear factor-3 α (Forkhead Box A1) and androgen receptor in transcriptional regulation of prostatic genes*, *Molecular Endocrinology*, 17 (2003), pp. 1484.
- [39] A. GHOSH, G. SAGINC, S. C. LEOW, E. KHATTAR, E. M. SHIN, T. D. YAN, M. WONG, Z. ZHANG, G. LI, W. K. SUNG, J. ZHOU, W. J. CHNG, S. LI, E. LIU and V. TERGAONKAR, *Telomerase directly regulates NF-kappaB-dependent transcription*, *Nat Cell Biol*, 14 (2012), pp. 1270-81.
- [40] W. R. GILKS, *Markov chain monte carlo*, *Encyclopedia of Biostatistics* (2005).
- [41] P. W. GLYNN and D. L. IGLEHART, *Importance sampling for stochastic simulations*, *Management Science* (1989), pp. 1367-1392.
- [42] F. GONG, L. SUN, Z. WANG, J. SHI, W. LI, S. WANG, X. HAN and Y. SUN, *The BCL2 gene is regulated by a special AT-rich sequence binding protein 1-mediated long range chromosomal interaction between the promoter and the distal element located within the 3'-UTR*, *Nucleic acids research*, 39 (2011), pp. 4640-52.
- [43] W. N. GRUNDY, T. L. BAILEY and C. P. ELKAN, *ParaMEME: A parallel implementation and a web interface for a DNA and protein motif discovery tool*, *Computer applications in the biosciences: CABIOS*, 12 (1996), pp. 303.

- [44] Y. HALPERIN, C. LINHART, I. ULITSKY and R. SHAMIR, *Allegro: analyzing expression and sequence in concert to discover regulatory programs*, Nucleic acids research, 37 (2009), pp. 1566-79.
- [45] J. A. HANLEY and B. J. MCNEIL, *The meaning and use of the area under a receiver operating characteristic (ROC) curve*, Radiology, 143 (1982), pp. 29-36.
- [46] S. HEINZ, C. BENNER, N. SPANN, E. BERTOLINO, Y. C. LIN, P. LASLO, J. X. CHENG, C. MURRE, H. SINGH and C. K. GLASS, *Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities*, Molecular cell, 38 (2010), pp. 576-89.
- [47] M. S. HESTAND, M. VAN GALEN, M. P. VILLERIUS, G. J. VAN OMMEN, J. T. DEN DUNNEN and P. A. T HOEN, *CORE_TF: a user-friendly interface to identify evolutionary conserved transcription factor binding sites in sets of co-regulated genes*, BMC Bioinformatics, 9 (2008), pp. 495.
- [48] S. J. HO SUI, D. L. FULTON, D. J. ARENILLAS, A. T. KWON and W. W. WASSERMAN, *oPOSSUM: integrated tools for analysis of regulatory motif over-representation*, Nucleic Acids Res, 35 (2007), pp. W245-52.
- [49] B. HOOGHE, P. HULPIAU, F. VAN ROY and P. DE BLESER, *ConTra: a promoter alignment analysis tool for identification of transcription factor binding sites across species*, Nucleic acids research, 36 (2008), pp. W128-32.
- [50] B. HOOGHE, P. HULPIAU, F. VAN ROY and P. DE BLESER, *ConTra: a promoter alignment analysis tool for identification of transcription factor binding sites across species*, Nucleic Acids Res, 36 (2008), pp. W128-32.
- [51] D. K. HU M, QIN ZS, DIXON J, SELVARAJ S, FANG J, REN B AND LIU JS., *Bayesian inference of spatial organizations of chromosomes*, PLoS Computational Biology, In press (2012).
- [52] M. HU, J. YU, J. M. G. TAYLOR, A. M. CHINNAIYAN and Z. S. QIN, *On the detection and refinement of transcription factor binding sites using ChIP-Seq data*, Nucleic Acids Res, 38 (2010), pp. 2154.
- [53] M. HU, J. YU, J. M. G. TAYLOR, A. M. CHINNAIYAN and Z. S. QIN, *On the detection and refinement of transcription factor binding sites using ChIP-Seq data.*, Nucleic Acids Res, 38 (2010), pp. 2154-2167.
- [54] J. D. HUGHES, P. W. ESTEP, S. TAVAZOIE and G. M. CHURCH, *Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae*, J Mol Biol, 296 (2000), pp. 1205-14.
- [55] M. IMAKAEV, G. FUDENBERG, R. P. MCCORD, N. NAUMOVA, A. GOLOBORODKO, B. R. LAJOIE, J. DEKKER and L. A. MIRNY, *Iterative correction of Hi-C data reveals hallmarks of chromosome organization*, Nat Methods, 9 (2012), pp. 999-1003.
- [56] T. JENUWEIN and C. D. ALLIS, *Translating the histone code*, Science's STKE, 293 (2001), pp. 1074.
- [57] X. JI, W. LI, J. SONG, L. WEI and X. S. LIU, *CEAS: cis-regulatory element annotation system*, Nucleic Acids Res, 34 (2006), pp. W551-4.
- [58] K. JIANG, D. SUN and K. C. TOH, *A partial proximal point algorithm for nuclear norm regularized matrix least squares problems*, (2012).

- [59] D. S. JOHNSON, A. MORTAZAVI, R. M. MYERS and B. WOLD, *Genome-wide mapping of in vivo protein-DNA interactions*, Science, 316 (2007), pp. 1497.
- [60] W. JU and F. C. BROSIUS, 3RD, *Understanding kidney disease: toward the integration of regulatory networks across species*, Semin Nephrol, 30 (2010), pp. 512-9.
- [61] R. KALHOR, H. TJONG, N. JAYATHILAKA, F. ALBER and L. CHEN, *Genome architectures revealed by tethered chromosome conformation capture and population-based modeling*, Nat Biotechnol, 30 (2012), pp. 90-8.
- [62] R. KALHOR, H. TJONG, N. JAYATHILAKA, F. ALBER and L. CHEN, *Genome architectures revealed by tethered chromosome conformation capture and population-based modeling*, Nature biotechnology, 30 (2012), pp. 90-8.
- [63] D. KAROLCHIK, A. S. HINRICHS and W. J. KENT, *The UCSC genome browser*, Current protocols in bioinformatics (2009), pp. 1.4. 1-1.4. 26.
- [64] B. L. KIDDER, J. YANG and S. PALMER, *Stat3 and c-Myc genome-wide promoter occupancy in embryonic stem cells*, PLoS One, 3 (2008), pp. e3932.
- [65] S. L. KONG, G. LI, S. L. LOH, W. K. SUNG and E. T. LIU, *Cellular reprogramming by the conjoint action of ER α , FOXA1, and GATA3 to a ligand-inducible growth state*, Molecular Systems Biology, 7 (2011).
- [66] S. L. KONG, G. LI, S. L. LOH, W. K. SUNG and E. T. LIU, *Cellular reprogramming by the conjoint action of ER α , FOXA1, and GATA3 to a ligand-inducible growth state*, Mol Syst Biol, 7 (2011), pp. 526.
- [67] I. KULAKOVSKIY, V. BOEVA, A. FAVOROV and V. MAKEEV, *Deep and wide digging for binding motifs in ChIP-Seq data*, Bioinformatics, 26 (2010), pp. 2622.
- [68] T. W. LAM, K. SADAKANE, W. K. SUNG and S. M. YIU, *A space and time efficient algorithm for constructing compressed suffix arrays*, Computing and Combinatorics (2002), pp. 21-26.
- [69] B. LANGMEAD and S. L. SALZBERG, *Fast gapped-read alignment with Bowtie 2*, Nature methods, 9 (2012), pp. 357-359.
- [70] C. E. LAWRENCE, S. F. ALTSCHUL, M. S. BOGUSKI, J. S. LIU, A. F. NEUWALD and J. C. WOOTTON, *Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment*, Science, 262 (1993), pp. 208-14.
- [71] E. L. LEHMANN, *Elements of large-sample theory*, Springer, 1999.
- [72] N. H. Z. LEUNG and K. C. TOH, *An SDP-Based Divide-and-Conquer Algorithm for Large-Scale Noisy Anchor-Free Graph Realization*, (2009).
- [73] G. LI, M. J. FULLWOOD, H. XU, F. H. MULAWADI, S. VELKOV, V. VEGA, P. N. ARIYARATNE, Y. B. MOHAMED, H. S. OOI, C. TENNAKOON, C. L. WEI, Y. RUAN and W. K. SUNG, *ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing*, Genome Biol, 11 (2010), pp. R22.
- [74] G. LI, X. RUAN, R. K. AUERBACH, K. S. SANDHU, M. ZHENG, P. WANG, H. M. POH, Y. GOH, J. LIM, J. ZHANG, H. S. SIM, S. Q. PEH, F. H. MULAWADI, C. T. ONG, Y. L. ORLOV, S. HONG, Z. Z. ZHANG, S. LANDT, D. RAHA, G. EUSKIRCHEN, C. L. WEI, W. GE, H. WANG, C. DAVIS, K. I. FISHER-AYLOR, A. MORTAZAVI, M. GERSTEIN, T. GINGERAS, B. WOLD, Y. SUN, M. J. FULLWOOD, E. CHEUNG, E. LIU,

- W. K. SUNG, M. SNYDER and Y. RUAN, *Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation*, *Cell*, 148 (2012), pp. 84-98.
- [75] H. LI and R. DURBIN, *Fast and accurate short read alignment with Burrows–Wheeler transform*, *Bioinformatics*, 25 (2009), pp. 1754-1760.
- [76] E. LIEBERMAN-AIDEN, N. L. VAN BERKUM, L. WILLIAMS, M. IMAKAEV, T. RAGOCZY, A. TELLING, I. AMIT, B. R. LAJOIE, P. J. SABO, M. O. DORSCHNER, R. SANDSTROM, B. BERNSTEIN, M. A. BENDER, M. GROUDINE, A. GNIRKE, J. STAMATOYANNOPOULOS, L. A. MIRNY, E. S. LANDER and J. DEKKER, *Comprehensive mapping of long-range interactions reveals folding principles of the human genome*, *Science*, 326 (2009), pp. 289-93.
- [77] C. H. LIN, B. J. HARE, G. WAGNER, S. C. HARRISON, T. MANIATIS and E. FRAENKEL, *A small domain of CBP/p300 binds diverse proteins: solution structure and functional studies*, *Mol Cell*, 8 (2001), pp. 581-90.
- [78] C. LINHART, Y. HALPERIN and R. SHAMIR, *Transcription factor and microRNA motif discovery: The Amadeus platform and a compendium of metazoan target sets*, *Genome Res*, 18 (2008), pp. 1180.
- [79] X. S. LIU, D. L. BRUTLAG and J. S. LIU, *An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments*, *Nat Biotechnol*, 20 (2002), pp. 835-9.
- [80] X. S. LIU, D. L. BRUTLAG and J. S. LIU, *An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments*, *Nature biotechnology*, 20 (2002), pp. 835-839.
- [81] Y. LIU, B. SCHMIDT, W. LIU and D. L. MASKELL, *CUDA-MEME: Accelerating motif discovery in biological sequences using CUDA-enabled graphics processing units*, *Pattern Recognition Letters* (2009).
- [82] Y. J. LIU, D. SUN and K. C. TOH, *An implementable proximal point algorithmic framework for nuclear norm minimization*, *Mathematical Programming* (2009), pp. 1-38.
- [83] K. LUGER, A. W. MADER, R. K. RICHMOND, D. F. SARGENT and T. J. RICHMOND, *Crystal structure of the nucleosome core particle at 2.8 Å resolution*, *Nature*, 389 (1997), pp. 251-260.
- [84] S. MAHONY, P. E. AURON and P. V. BENOS, *DNA familial binding profiles made easy: comparison of various motif alignment and clustering strategies*, *PLoS computational biology*, 3 (2007), pp. e61.
- [85] V. MATYS, E. FRICKE, R. GEFFERS, E. GOSSLING, M. HAUBROCK, R. HEHL, K. HORNISCHER, D. KARAS, A. E. KEL, O. V. KEL-MARGOULIS, D. U. KLOOS, S. LAND, B. LEWICKI-POTAPOV, H. MICHAEL, R. MUNCH, I. REUTER, S. ROTERT, H. SAXEL, M. SCHEER, S. THIELE and E. WINGENDER, *TRANSFAC: transcriptional regulation, from patterns to profiles*, *Nucleic Acids Res*, 31 (2003), pp. 374-8.
- [86] R. C. MCLEAY and T. L. BAILEY, *Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data*, *BMC Bioinformatics*, 11 (2010), pp. 165.
- [87] T. MISTELI, *Beyond the sequence: cellular organization of genome function*, *Cell*, 128 (2007), pp. 787-800.

- [88] T. MISTELI, *Spatial positioning; a new dimension in genome function*, Cell, 119 (2004), pp. 153-6.
- [89] V. NARANG, A. MITTAL and W. K. SUNG, *Localized motif discovery in gene regulatory sequences*, Bioinformatics, 26 (2010), pp. 1152.
- [90] Y. ORENSTEIN, E. MICK and R. SHAMIR, *RAP: accurate and fast motif finding based on protein-binding microarray data*, J Comput Biol, 20 (2013), pp. 375-82.
- [91] C. S. OSBORNE, L. CHAKALOVA, K. E. BROWN, D. CARTER, A. HORTON, E. DEBRAND, B. GOYENECHEA, J. A. MITCHELL, S. LOPES, W. REIK and P. FRASER, *Active genes dynamically colocalize to shared sites of ongoing transcription*, Nat Genet, 36 (2004), pp. 1065-71.
- [92] G. PAVESI, G. MAURI and G. PESOLE, *An algorithm for finding signals of unknown length in DNA sequences*, Bioinformatics, 17 (2001), pp. S207-S214.
- [93] G. PESOLE, S. LIUNI and M. D'SOUZA, *PatSearch: a pattern matcher software that finds functional elements in nucleotide and protein sequences and assesses their statistical significance*, Bioinformatics, 16 (2000), pp. 439-50.
- [94] D. RAHA, Z. WANG, Z. MOQTADERI, L. WU, G. ZHONG, M. GERSTEIN, K. STRUHL and M. SNYDER, *Close association of RNA polymerase II and many transcription factors with Pol III genes*, Proc Natl Acad Sci U S A, 107 (2010), pp. 3639-44.
- [95] B. RAPHAEL, L. T. LIU and G. VARGHESE, *A uniform projection method for motif discovery in DNA sequences*, IEEE Transactions on Computational biology and Bioinformatics (2004), pp. 91-94.
- [96] J. E. REID and L. WERNISCH, *STEME: efficient EM to find motifs in large data sets*, Nucleic Acids Res, 39 (2011), pp. e126-e126.
- [97] H. S. RHEE and B. F. PUGH, *Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution*, Cell, 147 (2011), pp. 1408-1419.
- [98] H. G. ROIDER, T. MANKE, S. O'KEEFFE, M. VINGRON and S. A. HAAS, *PASTAA: identifying transcription factors associated with sets of co-regulated genes*, Bioinformatics, 25 (2009), pp. 435-42.
- [99] F. P. ROTH1JT, J. D. HUGHES, P. W. ESTEP and G. M. CHURCH, *Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation*, Nature Biotechnology, 16 (1998), pp. 939.
- [100] F. P. ROTH, J. D. HUGHES, P. W. ESTEP and G. M. CHURCH, *Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation*, Nature Biotechnology, 16 (1998), pp. 939-945.
- [101] M. ROUSSEAU, J. FRASER, M. A. FERRAIUOLO, J. DOSTIE and M. BLANCHETTE, *Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling*, BMC Bioinformatics, 12 (2011), pp. 414.
- [102] M. ROUSSEAU, J. FRASER, M. A. FERRAIUOLO, J. DOSTIE and M. BLANCHETTE, *Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling*, BMC bioinformatics, 12 (2011), pp. 414.

- [103] D. RUSSEL, K. LASKER, B. WEBB, J. VELAZQUEZ-MURIEL, E. TJIOE, D. SCHNEIDMAN-DUHOVNY, B. PETERSON and A. SALI, *Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies*, PLoS Biol, 10 (2012), pp. e1001244.
- [104] B. SAHU, M. LAAKSO, K. OVASKA, T. MIRTTI, J. LUNDIN, A. RANNIKKO, A. SANKILA, J. P. TURUNEN, M. LUNDIN, J. KONSTI and OTHERS, *Dual role of FoxA1 in androgen receptor binding to chromatin, androgen signalling and prostate cancer*, The EMBO Journal, 30 (2011), pp. 3962-3976.
- [105] A. SANDELIN, W. ALKEMA, P. ENGSTROM, W. W. WASSERMAN and B. LENHARD, *JASPAR: an open-access database for eukaryotic transcription factor binding profiles*, Nucleic Acids Res, 32 (2004), pp. D91-4.
- [106] C. D. SCHMID and P. BUCHER, *ChIP-Seq data reveal nucleosome architecture of human promoters*, Cell, 131 (2007), pp. 831-2; author reply 832-3.
- [107] A. A. SERANDOUR, S. AVNER, F. OGER, M. BIZOT, F. PERCEVAULT, C. LUCCHETTI-MIGANEH, G. PALIERNE, C. GHEERAERT, F. BARLOY-HUBLER, C. L. PERON, T. MADIGOU, E. DURAND, P. FROGUEL, B. STAELS, P. LEFEBVRE, R. METIVIER, J. EECKHOUTE and G. SALBERT, *Dynamic hydroxymethylation of deoxyribonucleic acid marks differentiation-associated enhancers*, Nucleic Acids Res, 40 (2012), pp. 8255-65.
- [108] A. A. SHAROV and M. S. KO, *Exhaustive search for over-represented DNA sequence motifs with CisFinder*, DNA Res, 16 (2009), pp. 261-73.
- [109] S. SINHA, *Discriminative motifs*, J Comput Biol, 10 (2003), pp. 599-615.
- [110] S. SINHA, *On counting position weight matrix matches in a sequence, with application to discriminative motif finding*, Bioinformatics, 22 (2006).
- [111] S. SINHA and M. TOMPA, *A statistical method for finding transcription factor binding sites*, Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology, 2000, pp. 344-354.
- [112] A. M. C. SO and Y. YE, *Theory of semidefinite programming for sensor network localization*, Mathematical Programming, 109 (2007), pp. 367-384.
- [113] M. J. SOLOMON, P. L. LARSEN and A. VARSHAVSKY, *Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene*, Cell, 53 (1988), pp. 937.
- [114] P. Y. TAN, C. W. CHANG, K. R. CHNG, K. D. WANSA, W. K. SUNG and E. CHEUNG, *Integration of regulatory networks by NKX3-1 promotes androgen-dependent prostate cancer survival*, Mol Cell Biol, 32 (2012), pp. 399-414.
- [115] C. TENNAKOON, R. W. PURBOJATI and W. K. SUNG, *BatMis: a fast algorithm for k-mismatch mapping*, Bioinformatics, 28 (2012), pp. 2122-2128.
- [116] A. K. TEWARI, G. G. YARDIMCI, Y. SHIBATA, N. C. SHEFFIELD, L. SONG, B. S. TAYLOR, S. G. GEORGIEV, G. A. COETZEE, U. OHLER, T. S. FUREY, G. E. CRAWFORD and P. G. FEBBO, *Chromatin accessibility reveals insights into androgen receptor activation and transcriptional specificity*, Genome Biol, 13 (2012), pp. R88.

- [117] G. THIJS, M. LESCOT, K. MARCHAL, S. ROMBAUTS, B. DE MOOR, P. ROUZE and Y. MOREAU, *A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling*, *Bioinformatics*, 17 (2001), pp. 1113-22.
- [118] V. K. TIWARI and S. B. BAYLIN, *Mapping networks of protein-mediated physical interactions between chromatin elements*, *Curr Protoc Mol Biol*, Chapter 21 (2010), pp. Unit 21 16 1-13.
- [119] H. TJONG, K. GONG, L. CHEN and F. ALBER, *Physical tethering and volume exclusion determine higher-order genome organization in budding yeast*, *Genome research*, 22 (2012), pp. 1295-305.
- [120] K. C. TOH, M. J. TODD and R. H. TÖTÜNCÜ, *SDPT3—a MATLAB software package for semidefinite programming, version 1.3*, *Optimization Methods and Software*, 11 (1999), pp. 545-581.
- [121] A. VALOUEV, D. S. JOHNSON, A. SUNDQUIST, C. MEDINA, E. ANTON, S. BATZOGLOU, R. M. MYERS and A. SIDOW, *Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data*, *Nature methods*, 5 (2008), pp. 829.
- [122] P. H. VON HIPPEL and O. G. BERG, *On the specificity of DNA-protein interactions*, *Proceedings of the National Academy of Sciences of the United States of America*, 83 (1986), pp. 1608-12.
- [123] A. WÄCHTER and L. T. BIEGLER, *On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming*, *Mathematical Programming*, 106 (2006), pp. 25-57.
- [124] W. W. WASSERMAN and A. SANDELIN, *Applied bioinformatics for the identification of regulatory elements*, *Nature Reviews Genetics*, 5 (2004), pp. 276-287.
- [125] C. L. WEI, Q. WU, V. B. VEGA, K. P. CHIU, P. NG, T. ZHANG, A. SHAHAB, H. C. YONG, Y. FU, Z. WENG, J. LIU, X. D. ZHAO, J. L. CHEW, Y. L. LEE, V. A. KUZNETSOV, W. K. SUNG, L. D. MILLER, B. LIM, E. T. LIU, Q. YU, H. H. NG and Y. RUAN, *A global map of p53 transcription-factor binding sites in the human genome*, *Cell*, 124 (2006), pp. 207-19.
- [126] Z. WEI and S. T. JENSEN, *GAME: detecting cis-regulatory elements using a genetic algorithm*, *Bioinformatics*, 22 (2006), pp. 1577-84.
- [127] K. Q. WEINBERGER and L. K. SAUL, *Unsupervised learning of image manifolds by semidefinite programming*, *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, IEEE, 2004, pp. II-988-II-995 Vol. 2.
- [128] M. T. WEIRAUCH, A. COTE, R. NOREL, M. ANNALA, Y. ZHAO, T. R. RILEY, J. SAEZ-RODRIGUEZ, T. COKELAER, A. VEDENKO, S. TALUKDER, H. J. BUSSEMAKER, Q. D. MORRIS, M. L. BULYK, G. STOLOVITZKY and T. R. HUGHES, *Evaluation of methods for modeling transcription factor sequence specificity*, *Nat Biotechnol*, 31 (2013), pp. 126-34.
- [129] T. WHITINGTON, M. C. FRITH, J. JOHNSON and T. L. BAILEY, *Inferring transcription factor complexes from ChIP-seq data*, *Nucleic Acids Res*, 39 (2011), pp. e98.

- [130] C. T. WORKMAN and G. D. STORMO, *ANN-Spec: a method for discovering transcription factor binding sites with improved specificity*, Pac Symp Biocomput (2000), pp. 467-78.
- [131] Q. WU and H. H. NG, *Mark the transition: chromatin modifications and cell fate decision*, Cell Research (2011).
- [132] H. XU, L. HANDOKO, X. WEI, C. YE, J. SHENG, C. L. WEI, F. LIN and W. K. SUNG, *A signal-noise model for significance analysis of ChIP-seq with negative control*, Bioinformatics, 26 (2010), pp. 1199-204.
- [133] E. YAFFE and A. TANAY, *Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture*, Nature genetics, 43 (2011), pp. 1059-65.
- [134] Y. ZHANG, T. LIU, C. A. MEYER, J. EECKHOUTE, D. S. JOHNSON, B. E. BERNSTEIN, C. NUSBAUM, R. M. MYERS, M. BROWN, W. LI and X. S. LIU, *Model-based analysis of ChIP-Seq (MACS)*, Genome Biol, 9 (2008), pp. R137.
- [135] Z. ZHANG, C. W. CHANG, W. L. GOH, W. K. SUNG and E. CHEUNG, *CENTDIST: discovery of co-associated factors by motif distribution*, Nucleic Acids Res, 39 (2011), pp. W391.
- [136] Z. Z. ZHANG, C. CHANG, W. HUGO, E. CHEUNG and W. K. SUNG, *Simultaneously learning DNA motif along with its position and sequence rank preferences through EM algorithm*, Research in Computational Molecular Biology, Springer, 2012, pp. 355-370.
- [137] Z. Z. ZHANG, C. CHANG, W. HUGO, E. CHEUNG and W. K. SUNG, *Simultaneously learning DNA motif along with its position and sequence rank preferences through EM algorithm*, Springer, 2012, pp. 355-370.
- [138] Z. Z. ZHANG, G. L. LI, K. C. TOH and W. K. SUNG, *Inference of Spatial Organizations of Chromosomes Using Semi-definite Embedding Approach and Hi-C Data*, Research in Computational Molecular Biology, 2013.
- [139] Y. ZHAO and G. D. STORMO, *Quantitative analysis demonstrates most transcription factors require only simple models of specificity*, Nat Biotechnol, 29 (2011), pp. 480-483.
- [140] Z. ZHAO, G. TAVOOSIDANA, M. SJOLINDER, A. GONDOR, P. MARIANO, S. WANG, C. KANDURI, M. LEZCANO, K. S. SANDHU, U. SINGH, V. PANT, V. TIWARI, S. KURUKUTI and R. OHLSSON, *Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions*, Nature genetics, 38 (2006), pp. 1341-7.
- [141] Q. ZHOU and J. S. LIU, *Modeling within-motif dependence for transcription factor binding site predictions*, Bioinformatics, 20 (2004), pp. 909-16.
- [142] B. M. ZHU, K. KANG, J. H. YU, W. CHEN, H. E. SMITH, D. LEE, H. W. SUN, L. WEI and L. HENNIGHAUSEN, *Genome-wide analyses reveal the extent of opportunistic STAT5 binding that does not yield transcriptional activation of neighboring genes*, Nucleic Acids Res, 40 (2012), pp. 4461-72.

Appendix

Supp Table 1: **Overlapping peak percentage among ChIP-seq experiments.**

Each entry in this table shows the percentage of overlap (within 100 bp) of the ChIP-seq peak set. The overlap of TF1 and TF2's peak sets is defined as $|\text{TF1} \cap \text{TF2}| / \min(|\text{TF1}|, |\text{TF2}|)$.

	P300	NANOG	OCT4	SOX2	SMAD1	STAT3	KLF4	ESRRB	TCFCP211	CMYC	NMYC	ZFX	E2F1
P300	100%	44.47%	31.68%	37.60%	28.82%	22.14%	27.48%	24.43%	12.21%	3.44%	5.73%	4.77%	7.25%
NANOG	44.47%	100%	40.12%	57.58%	73.27%	23.49%	12.57%	10.42%	8.16%	3.59%	4.19%	2.34%	4.44%
OCT4	31.68%	40.12%	100%	37.76%	42.27%	15.04%	20.90%	12.02%	11.49%	7.54%	14.54%	9.65%	16.62%
SOX2	37.60%	57.58%	37.76%	100%	55.51%	16.34%	18.38%	14.05%	12.95%	2.60%	5.04%	4.18%	7.47%
SMAD1	28.82%	73.27%	42.27%	55.51%	100%	22.11%	34.46%	29.31%	18.21%	1.33%	5.68%	3.64%	7.55%
STAT3	22.14%	23.49%	15.04%	16.34%	22.11%	100%	25.69%	17.79%	14.49%	4.95%	11.63%	8.25%	13.08%
KLF4	27.48%	12.57%	20.90%	18.38%	34.46%	25.69%	100%	12.06%	9.98%	18.70%	21.35%	12.18%	18.10%
ESRRB	24.43%	10.42%	12.02%	14.05%	29.31%	17.79%	12.06%	100%	8.86%	5.08%	5.76%	5.70%	4.63%
TCFCP211	12.21%	8.16%	11.49%	12.95%	18.21%	14.49%	9.98%	8.86%	100%	6.52%	6.89%	5.12%	6.98%
CMYC	3.44%	3.59%	7.54%	2.60%	1.33%	4.95%	18.70%	5.08%	6.52%	100%	70.66%	30.77%	46.08%
NMYC	5.73%	4.19%	14.54%	5.04%	5.68%	11.63%	21.35%	5.76%	6.89%	70.66%	100%	26.54%	37.86%
ZFX	4.77%	2.34%	9.65%	4.18%	3.64%	8.25%	12.18%	5.70%	5.12%	30.77%	26.54%	100%	22.41%
E2F1	7.25%	4.44%	16.62%	7.47%	7.55%	13.08%	18.10%	4.63%	6.98%	46.08%	37.86%	22.41%	100%

Supp Table 2: **Co-TFs list table for each tested ChIP-seq dataset. The table contains a list of motif families for the ChIPed TF and co-TFs for each ChIP-seq experiment.**

DataSet	ChIPed TF motif and Co-TF motif list
Nanog	NANOG OCT SOX ERE STAT
Oct4	NANOG OCT SOX STAT ERE CP2 E2F EBOX
SOX2	NANOG OCT SOX STAT ERE CP2
Smad1	NANOG OCT SOX STAT ERE CP2
Stat3	NANOG OCT SOX STAT ERE CP2 E2F EBOX
P300	NANOG OCT SOX STAT ERE CP2
Klf4	NANOG ERE EBOX ZF5 E2F OCT STAT SOX
Esrrb	ERE NANOG OCT SOX STAT
Tcfcp12	CP2 E2F OCT SOX STAT
Cmyc	EBOX ZF5 E2F
Nmyc	EBOX ZF5 E2F OCT STAT
Zfx	CP2 EBOX ZF5 E2F OCT
E2f1	CP2 EBOX ZF5 E2F OCT STAT
AR	FoxA1 GATA NF1 OCT1 CEBP ETS1 NKX3
ER	AP1 FoxA1 PAX2 OCT1 GATA CEBP NF1 MYC NKX3 SP1 LEF1

Other Statistical Measurements

Given a set of positive sequences and negative sequences, and a PWM motif. We get all the sites matching the PWM motif higher than the PWM score cut-off (under FDR=0.001) in both positive sequences and negative sequences. Then, we define TP as the number of matched sites in the positive sequences and FP as the number of matched sites in the negative sequences. And TN and FN denote the number of unmatched sites in the positive sequences and the negative sequences, respectively. Then, some common measurements are defined as follow.

PPV (Positive Predictive Value)

$$PPV = \frac{TP}{TP + FP} \quad (S. 1)$$

SPC (Specificity)

$$SPC = \frac{TN}{TN + FP} \quad (S. 2)$$

ASP(Average site performance)

$$ASP = \frac{\frac{TP}{TP + FP} + \frac{TP}{TP + FN}}{2} \quad (S. 3)$$

Compute Zscore

Input: PWM Θ , positive data X^{pos} , negative data X^{neg}

Output: Zscore value

- 1: **for all** sequence $X_i \in X^{pos}$ **do**
- 2: $Score_{i,j}^+ :=$ compute PWM score $\log Pr(X_{i,j}|\Theta)$, for all subsequence $X_{i,j} \in X_i$
- 3: **end for**
- 4: **for all** sequence $X_i \in X^{neg}$ **do**
- 5: $Score_{i,j}^- :=$ compute PWM score $\log Pr(X_{i,j}|\Theta)$, for all subsequence $X_{i,j} \in X_i$
- 6: **end for**
- 7: compute PWM score cut-off γ based FDR 0.1% from $Score^-$
- 8: $T := |Score^+|$ {true count}
- 9: $F := |Score^-|$ {false count}
- 10: $TP := |Score^+ > \gamma|$ {true positive count}
- 11: $FP := |Score^- > \gamma|$ {false positive count}
- 12: $Zscore := \frac{TP - FP \cdot \frac{T}{F}}{\sqrt{T(1 - \frac{FP}{F}) \frac{FP}{F}}}$
- 13: **return** $Zscore$

Supp Figure 1: **Procedure for computing Zscore.** Consider a set of positive sequences and negative sequences, and a PWM motif. Under the PWM score cutoff with FDR=0.001, we get all the sites matching the PWM motif in both positive sequences and negative sequences. Then, we define TP be the number of matched sites in the positive sequences and FP be the number of matched sites in the negative sequences. And T and F denote the total number of sites in the positive and negative sequences, respectively.

Supp Table 3: **The table of TRANSFAC TF Families and their corresponding members (vertebrate only).**

TF Family	Family Members: {TRANSFAC PWM Identifier Motif Name}
HELIOS	M01003 V\$HELIOSA_01 M01004 V\$HELIOSA_02 M00935 V\$NFAT_Q4_01 M00302 V\$NFAT_Q6
MOVO	M01104 V\$MOVOB_01
GLI	M01037 V\$GLI_Q2 M00448 V\$ZIC1_01 M00449 V\$ZIC2_01 M00450 V\$ZIC3_01
AP2	M00800 V\$AP2_Q3 M00189 V\$AP2_Q6 M00915 V\$AP2_Q6_01 M00469 V\$AP2ALPHA_01 M01045 V\$AP2ALPHA_02 M01047 V\$AP2ALPHA_03 M00470 V\$AP2GAMMA_01 M00468 V\$AP2REP_01
AP3	M00690 V\$AP3_Q6
AP1	M00517 V\$AP1_01 M00199 V\$AP1_C M00173 V\$AP1_Q2 M00924 V\$AP1_Q2_01 M00188 V\$AP1_Q4 M00926 V\$AP1_Q4_01 M00174 V\$AP1_Q6 M00925 V\$AP1_Q6_01 M00172 V\$AP1FJ_Q2 M00037 V\$NFE2_01 M00285 V\$TCF11_01 M00284 V\$TCF11MAFG_01
AP4	M00005 V\$AP4_01 M00175 V\$AP4_Q5 M00176 V\$AP4_Q6 M00927 V\$AP4_Q6_01
MEF2	M00403 V\$AMEF2_Q6 M00406 V\$HMEF2_Q6 M00006 V\$MEF2_01 M00231 V\$MEF2_02 M00232 V\$MEF2_03 M00233 V\$MEF2_04 M00941 V\$MEF2_Q6_01 M00405 V\$MMEF2_Q6 M00026 V\$RSRFC4_01

	M00407 V\$RSRFC4_Q2
MEF3	M00319 V\$MEF3_B
ROAZ	M00467 V\$ROAZ_01
CACCC	M00720 V\$CACBINDINGPROTEIN_Q6 M00721 V\$CACCCBINDINGFACTOR_Q6
SPZ	M00446 V\$SPZ1_01
SEF1	M00214 V\$SEF1_C
HSF	M00641 V\$HSF_Q6 M00146 V\$HSF1_01 M01023 V\$HSF1_Q6 M00147 V\$HSF2_01
DEAF1	M01001 V\$DEAF1_01 M01002 V\$DEAF1_02
PAX	M01069 V\$GZF1_01 M00808 V\$PAX_Q6 M00326 V\$PAX1_B M00098 V\$PAX2_01 M00486 V\$PAX2_02 M00360 V\$PAX3_01 M00327 V\$PAX3_B M00373 V\$PAX4_01 M00377 V\$PAX4_02 M00378 V\$PAX4_03 M00380 V\$PAX4_04 M00143 V\$PAX5_01 M00144 V\$PAX5_02 M00097 V\$PAX6_01 M00979 V\$PAX6_Q2 M00717 V\$PAX8_01 M00328 V\$PAX8_B M00329 V\$PAX9_B
EGR	M00807 V\$EGR_Q6 M00243 V\$EGR1_01 M00246 V\$EGR2_01 M00245 V\$EGR3_01 M00982 V\$KROX_Q6 M00244 V\$NGFIC_01
HNF1	M00132 V\$HNF1_01 M00206 V\$HNF1_C M00790 V\$HNF1_Q6 M01011 V\$HNF1_Q6_01
MYB	M00004 V\$CMYB_01 M00773 V\$MYB_Q3 M00913 V\$MYB_Q5_01 M00183 V\$MYB_Q6
HNF6	M00639 V\$HNF6_Q6
WT1	M01118 V\$WT1_Q6
PIT1	M00802 V\$PIT1_Q6
SP1	M01113 V\$CACD_01 M00695 V\$ETF_Q6 M00255 V\$GC_01 M00649 V\$MAZ_Q6 M00491 V\$MAZR_01 M00008 V\$SP1_01 M00933 V\$SP1_Q2_01 M00932 V\$SP1_Q4_01 M00196 V\$SP1_Q6 M00931 V\$SP1_Q6_01 M01068 V\$UF1H3BETA_Q6
SP3	M00665 V\$SP3_Q3
ZBRK1	M01105 V\$ZBRK1_01
ZID	M00085 V\$ZID_01
BACH	M00495 V\$BACH1_01 M00490 V\$BACH2_01
SOX	M01014 V\$SOX_Q6 M01131 V\$SOX10_Q6 M01016 V\$SOX17_01 M00042 V\$SOX5_01 M00410 V\$SOX9_B1
ARP1	M00155 V\$ARP1_01
XPF1	M00684 V\$XPF1_Q6
IPF	M00436 V\$IPF1_Q4 M01013 V\$IPF1_Q4_01
AFP1	M00616 V\$AFP1_Q6
CART1	M00416 V\$CART1_01
NFKB	M00053 V\$CREL_01 M00054 V\$NFKAPPAB_01 M00051 V\$NFKAPPAB50_01 M00052 V\$NFKAPPAB65_01 M00208 V\$NFKB_C M00194 V\$NFKB_Q6 M00774 V\$NFKB_Q6_01 M00651 V\$NFMUE1_Q6
ZF5	M00716 V\$ZF5_01 M00333 V\$ZF5_B
TGTGGT	M00769 V\$AML_Q6 M00271 V\$AML1_01 M00751 V\$AML1_Q6 M01079 V\$CBF_01 M01080 V\$CBF_02 M00722 V\$COREBINDINGFACTOR_Q6 M00731 V\$OSF2_Q6 M00211 V\$PADS_C M00984 V\$PEBP_Q6
AHR	M00139 V\$AHR_01 M00778 V\$AHR_Q5 M00235 V\$AHRARNT_01 M00237 V\$AHRARNT_02 M00976 V\$AHRHIF_Q6
MAF	M00648 V\$MAF_Q6 M00983 V\$MAF_Q6_01

BRCA	M01082 V\$BRCA_01
DBP	M00624 V\$DBP_Q6
CP2	M00072 V\$CP2_01 M00947 V\$CP2_02
STAT	M00223 V\$STAT_01 M00777 V\$STAT_Q6 M00224 V\$STAT1_01 M00492 V\$STAT1_02 M00496 V\$STAT1_03 M00225 V\$STAT3_01 M00497 V\$STAT3_02 M00498 V\$STAT4_01 M00457 V\$STAT5A_01 M00460 V\$STAT5A_02 M00493 V\$STAT5A_03 M00499 V\$STAT5A_04 M00459 V\$STAT5B_01 M00494 V\$STAT6_01 M00500 V\$STAT6_02
CRX	M00623 V\$CRX_Q4
ZNF219	M01122 V\$ZNF219_01
STAF	M00262 V\$STAF_01 M00264 V\$STAF_02
BRACH	M00150 V\$BRACH_01 M01019 V\$TBX5_01 M01020 V\$TBX5_02 M01044 V\$TBX5_Q5
LDSPOLYA	M00317 V\$LDSPOLYA_B
OLF1	M00261 V\$OLF1_01
DEC	M00997 V\$DEC_Q1
OCT	M00210 V\$OCT_C M00795 V\$OCT_Q6 M00135 V\$OCT1_01 M00136 V\$OCT1_02 M00137 V\$OCT1_03 M00138 V\$OCT1_04 M00161 V\$OCT1_05 M00162 V\$OCT1_06 M00248 V\$OCT1_07 M00342 V\$OCT1_B M00930 V\$OCT1_Q5_01 M00195 V\$OCT1_Q6 M01125 V\$OCT4_01 M01124 V\$OCT4_02
XVENT1	M00445 V\$XVENT1_01
LUN1	M00480 V\$LUN1_01
CMAF	M01070 V\$CMAF_01
LMAF	M01139 V\$LMAF_Q2
MEIS1	M00419 V\$MEIS1_01 M00420 V\$MEIS1AHOXA9_01 M00421 V\$MEIS1BHOXA9_02
NF1	M00193 V\$NF1_Q6 M00806 V\$NF1_Q6_01
GATA	M00203 V\$GATA_C M00789 V\$GATA_Q6 M00075 V\$GATA1_01 M00126 V\$GATA1_02 M00127 V\$GATA1_03 M00128 V\$GATA1_04 M00346 V\$GATA1_05 M00347 V\$GATA1_06 M00076 V\$GATA2_01 M00348 V\$GATA2_02 M00349 V\$GATA2_03 M00077 V\$GATA3_01 M00350 V\$GATA3_02 M00351 V\$GATA3_03 M00632 V\$GATA4_Q3 M00462 V\$GATA6_01 M00278 V\$LMO2COM_02
E2	M00107 V\$E2_01 M00181 V\$E2_Q6 M00928 V\$E2_Q6_01
SZF11	M01109 V\$SZF11_01
EN	M00396 V\$EN1_01
FOX	M00809 V\$FOX_Q2 M00130 V\$FOXD3_01 M00422 V\$FOXJ2_01 M00423 V\$FOXJ2_02 M00630 V\$FOXM1_01 M00473 V\$FOXO1_01 M00474 V\$FOXO1_02 M00477 V\$FOXO3_01 M01137 V\$FOXO3A_Q1 M00472 V\$FOXO4_01 M00476 V\$FOXO4_02 M00987 V\$FOXP1_01 M00992 V\$FOXP3_Q4 M00290 V\$FREAC2_01 M00291 V\$FREAC3_01 M00292 V\$FREAC4_01 M00293 V\$FREAC7_01 M00129 V\$HFB1_01 M00289 V\$HFB3_01 M00742 V\$HFB4_01 M00294 V\$HFB8_01 M00791 V\$HNF3_Q6 M01012 V\$HNF3_Q6_01 M00724 V\$HNF3ALPHA_Q6 M00131 V\$HNF3B_01 M00148 V\$SRY_01 M00160 V\$SRY_02 M00267 V\$XFD1_01 M00268 V\$XFD2_01 M00269 V\$XFD3_01
MYOGNF1	M00056 V\$MYOGNF1_01
RFX	M00975 V\$RFX_Q6 M00280 V\$RFX1_01 M00281 V\$RFX1_02
GCM	M00634 V\$GCM_Q2

PBX	M00998 V\$PBX_Q3 M01017 V\$PBX1_Q3	M00096 V\$PBX1_01	M00124 V\$PBX1_02
RUSH1A	M01107 V\$RUSH1A_Q2		
SMAD	M00792 V\$SMAD_Q6 M00733 V\$SMAD4_Q6	M00974 V\$SMAD_Q6_01	M00701 V\$SMAD3_Q6
EFC	M00626 V\$EFC_Q6		
TAACC	M00331 V\$TAACC_B		
BRN2	M00145 V\$BRN2_01		
LYF1	M00141 V\$LYF1_01		
LRF	M01100 V\$LRF_Q2		
BLIMP1	M01066 V\$BLIMP1_Q6		
TFIIA	M00707 V\$TFIIA_Q6		
TFIII	M00706 V\$TFIII_Q6		
CREB	M00017 V\$ATF_01 M00514 V\$ATF4_Q2 M00113 V\$CREB_Q2 M00801 V\$CREB_Q3 M00981 V\$CREBATF_Q6 M00041 V\$CREBP1CJUN_01 M00325 V\$NRSE_B M00114 V\$TAXCREB_01	M00338 V\$ATF_B M00483 V\$ATF6_01 M00177 V\$CREB_Q2 M00178 V\$CREB_Q4 M00040 V\$CREBP1_01 M00694 V\$E4F1_Q6 M00256 V\$NRSF_01 M00115 V\$TAXCREB_02	M00691 V\$ATF1_Q6 M00513 V\$ATF3_Q6 M00039 V\$CREB_Q1 M00916 V\$CREB_Q2_01 M00917 V\$CREB_Q4_01 M00179 V\$CREBP1_Q2 M00260 V\$HLF_01 M01028 V\$NRSF_Q4 M00036 V\$VJUN_01
ETS	M00743 V\$CETS168_Q6 M01078 V\$CETS1P54_Q3 M00025 V\$ELK1_Q2 M00340 V\$ETS2_B M00655 V\$PEA3_Q6	M00032 V\$CETS1P54_01 M00746 V\$ELF1_Q6 M00771 V\$ETS_Q4 M00341 V\$GABP_B M00658 V\$PU1_Q6	M00074 V\$CETS1P54_Q2 M00007 V\$ELK1_Q1 M00971 V\$ETS_Q6 M00339 V\$ETS1_B M00531 V\$NERF_Q2 M00678 V\$TEL2_Q6
S8	M00099 V\$S8_01		
VMAF	M00035 V\$VMAF_01		
BCD	M01117 V\$OTX_Q1		
LRH1	M01142 V\$LRH1_Q5		
SUH	M01112 V\$RBPJK_Q1 M01111 V\$RBPJK_Q4		
MIF1	M00279 V\$MIF1_01		
COMP1	M00057 V\$COMP1_01		
VMYB	M00003 V\$VMYB_01 M00227 V\$VMYB_Q2		
CDC5	M00478 V\$CDC5_01		
NANOG	M01123 V\$NANOG_01		
HIC1	M01072 V\$HIC1_Q2 M01073 V\$HIC1_Q3		
CEBP	M00159 V\$CEBP_Q1 M00912 V\$CEBP_Q2_01 M00109 V\$CEBPB_Q1 M00622 V\$CEBPGAMMA_Q6	M00201 V\$CEBP_C M00770 V\$CEBP_Q3 M00117 V\$CEBPB_Q2	M00190 V\$CEBP_Q2 M00116 V\$CEBPA_Q1 M00621 V\$CEBPDELTA_Q6
ATCGAT	M00095 V\$CDP_Q1 M00105 V\$CDPCR3_Q1	M00102 V\$CDP_Q2 M00106 V\$CDPCR3HD_Q1	M00104 V\$CDPCR1_Q1
RP58	M00532 V\$RP58_01		
WHN	M00332 V\$WHN_B		
E2F	M00024 V\$E2F_Q1 M00425 V\$E2F_Q3 M00919 V\$E2F_Q4_01 M00428 V\$E2F1_Q3	M00050 V\$E2F_Q2 M00918 V\$E2F_Q3_01 M00427 V\$E2F_Q6 M00938 V\$E2F1_Q3_01	M00516 V\$E2F_Q3 M00426 V\$E2F_Q4 M00920 V\$E2F_Q6_01 M00430 V\$E2F1_Q4

	M00939 V\$E2F1_Q4_01 M00431 V\$E2F1_Q6 M00940 V\$E2F1_Q6_01 M00736 V\$E2F1DP1_01 M00740 V\$E2F1DP1RB_01 M00737 V\$E2F1DP2_01 M00738 V\$E2F4DP1_01 M00739 V\$E2F4DP2_01
TEF	M00672 V\$TEF_Q6 M00704 V\$TEF1_Q6
GGG	M00986 V\$CHCH_01
HOX	M00023 V\$HOX13_01 M00395 V\$HOXA3_01 M00640 V\$HOXA4_Q2 M01108 V\$HOXA7_01
SREB	M00776 V\$SREBP_Q3 M00221 V\$SREBP1_02 M00749 V\$SREBP1_Q6
HMGYIY	M01010 V\$HMGYIY_Q3 M00750 V\$HMGYIY_Q6
MINI	M00323 V\$MINI19_B M00324 V\$MINI20_B M00321 V\$MUSCLE_INI_B
SRF	M00152 V\$SRF_01 M00215 V\$SRF_C M00810 V\$SRF_Q4 M00922 V\$SRF_Q5_01 M01007 V\$SRF_Q5_02 M00186 V\$SRF_Q6
HES	M01009 V\$HES1_Q2
ATATA	M00311 V\$ATATA_B
R	M00273 V\$R_01
HP1SITE	M00725 V\$HP1SITEFACTOR_Q6
HEN	M00068 V\$HEN1_01 M00058 V\$HEN1_02 M00644 V\$LBP1_Q6
CAAT	M00687 V\$ALPHACP1_01 M00254 V\$CAAT_01 M00200 V\$CAAT_C M00334 V\$DTYPEPA_B M00287 V\$NFY_01 M00209 V\$NFY_C M00185 V\$NFY_Q6 M00775 V\$NFY_Q6_01 M00059 V\$YY1_01 M00069 V\$YY1_02 M00793 V\$YY1_Q6 M01035 V\$YY1_Q6_02
ZEC	M01081 V\$ZEC_01
LEF	M00805 V\$LEF1_Q2 M01022 V\$LEF1_Q2_01 M00978 V\$LEF1TCF1_Q4
MTF1	M00650 V\$MTF1_Q4
TGIF	M00418 V\$TGIF_01
TATA	M00100 V\$CDXA_01 M00101 V\$CDXA_02 M00320 V\$MTATA_B M00252 V\$TATA_01 M00216 V\$TATA_C M00471 V\$TBP_01 M00980 V\$TBP_Q6
PTF1BETA	M00657 V\$PTF1BETA_Q6
TST1	M00133 V\$TST1_01
AAAAA	M00734 V\$CIZ_01
CDX	M00991 V\$CDX_Q5 M00729 V\$CDX2_Q5
MRF2	M00454 V\$MRF2_01
HAND1E47	M00222 V\$HAND1E47_01
HMX1	M00433 V\$HMX1_01
TITF1	M00432 V\$TITF1_Q3
PLZF	M01075 V\$PLZF_02
CAP	M00253 V\$CAP_01
TTF1	M00794 V\$TTF1_Q6
ERE	M00158 V\$COUP_01 M00765 V\$COUP_DR1_Q6 M01036 V\$COUPTF_Q6 M00762 V\$DR1_Q3 M00966 V\$DR3_Q4 M00965 V\$DR4_Q2 M00191 V\$ER_Q6 M00959 V\$ER_Q6_02 M00511 V\$ERR1_Q2 M00526 V\$GCNF_01 M00134 V\$HNF4_01 M00411 V\$HNF4_01_B M00764 V\$HNF4_DR1_Q3 M00967 V\$HNF4_Q6 M01031 V\$HNF4_Q6_01 M01032 V\$HNF4_Q6_02 M01033 V\$HNF4_Q6_03 M00638 V\$HNF4ALPHA_Q6 M00646 V\$LFA1_Q6 M00766 V\$LXR_DR4_Q3 M00647 V\$LXR_Q3 M00763 V\$PPAR_DR1_Q2 M00242 V\$PPARA_01 M00518 V\$PPARA_02 M00512 V\$PPARG_01 M00515 V\$PPARG_02 M00528 V\$PPARG_03 M01152 V\$PXRXR_01 M01153 V\$PXRXR_02 M01138 V\$RORA_Q4 M00156 V\$RORA1_01

	M00157 V\$RORA2_01 M00727 V\$SF1_Q6 M01132 V\$SF1_Q6_01 M00239 V\$T3R_01 M00963 V\$T3R_Q6 M00671 V\$TCF4_Q5 M00444 V\$VDR_Q3 M00961 V\$VDR_Q6 M00711 V\$ZTA_Q2
GATA_DIMER	M00078 V\$EV11_01 M00079 V\$EV11_02 M00080 V\$EV11_03 M00081 V\$EV11_04 M00082 V\$EV11_05 M00011 V\$EV11_06
PITX2	M00482 V\$PITX2_Q2
FAC1	M00456 V\$FAC1_01
CLOX	M00103 V\$CLOX_01
DMRT	M01146 V\$DMRT1_01 M01147 V\$DMRT2_01 M01148 V\$DMRT3_01 M01149 V\$DMRT4_01 M01150 V\$DMRT5_01 M01151 V\$DMRT7_01
MZF1	M00083 V\$MZF1_01 M00084 V\$MZF1_02
BARBIE	M00238 V\$BARBIE_01
POU	M00744 V\$POU1F1_Q6 M00463 V\$POU3F2_01 M00464 V\$POU3F2_02 M00465 V\$POU6F1_01
ALX4	M00619 V\$ALX4_01
CHOP	M00249 V\$CHOP_01
AT_RICH	M00437 V\$CHX10_01 M00510 V\$LHX3_01
MSX1	M00394 V\$MSX1_01
KAISO	M01119 V\$KAISO_01
P53	M00034 V\$P53_01 M00272 V\$P53_02 M00761 V\$P53_DECAMER_Q2
Initiator	M00315 V\$GEN_INI_B M00313 V\$GEN_INI2_B M00314 V\$GEN_INI3_B
FXR	M00767 V\$FXR_IR1_Q6 M00631 V\$FXR_Q3 M00964 V\$PXR_Q2
IK	M00086 V\$IK1_01 M00087 V\$IK2_01 M00088 V\$IK3_01
AR	M00481 V\$AR_01 M00953 V\$AR_02 M00956 V\$AR_03 M00447 V\$AR_Q2 M00962 V\$AR_Q6 M00955 V\$GR_01 M00192 V\$GR_Q6 M00921 V\$GR_Q6_01 M00205 V\$GRE_C M00954 V\$PR_01 M00957 V\$PR_02 M00960 V\$PR_Q2
SREB/EBOX	M00220 V\$SREBP1_01
ACAAT	M00309 V\$ACAAT_B
POLYC	M00212 V\$POLY_C
POLYA	M00310 V\$APOLYA_B M00318 V\$LPOLYA_B
BEL1	M00312 V\$BEL1_B
IRF	M00699 V\$ICSBP_Q6 M00772 V\$IRF_Q6 M00972 V\$IRF_Q6_01 M00062 V\$IRF1_01 M00747 V\$IRF1_Q6 M00063 V\$IRF2_01 M00453 V\$IRF7_01 M00258 V\$ISRE_01
P300	M00033 V\$P300_01
NCX	M00484 V\$NCX_01
EBF	M00977 V\$EBF_Q6
EBOX	M00236 V\$ARNT_01 M00539 V\$ARNT_02 M01116 V\$CLOCKBMAL_Q6 M01145 V\$CMYC_01 M01154 V\$CMYC_02 M00693 V\$E12_Q6 M00804 V\$E2A_Q2 M00973 V\$E2A_Q6 M00002 V\$E47_01 M00071 V\$E47_02 M01034 V\$EBOX_Q6_01 M00698 V\$HEB_Q6 M00797 V\$HIF1_Q3 M00466 V\$HIF1_Q5 M00538 V\$HTF_01 M00277 V\$LMO2COM_01 M00119 V\$MAX_01 M00799 V\$MYC_Q2 M00118 V\$MYCMAX_01 M00123 V\$MYCMAX_02 M00615 V\$MYCMAX_03 M00322 V\$MYCMAX_B M00001 V\$MYOD_01 M00184 V\$MYOD_Q6 M00929 V\$MYOD_Q6_01 M00712 V\$MYOGENIN_Q6 M00055 V\$NMYC_01 M00985 V\$STRA13_01 M00993 V\$TAL1_Q6 M00066 V\$TAL1ALPHA47_01 M00065 V\$TAL1BETA47_01 M00070 V\$TAL1BETAITF2_01 M01029 V\$TFE_Q6 M00121 V\$USF_01 M00122 V\$USF_02 M00217 V\$USF_C M00187 V\$USF_Q6

	M00796 V\$USF_Q6_01	M00726 V\$USF2_Q6	M00228 V\$VBP_01
	M00251 V\$XBP1_01		
AIRE	M00999 V\$AIRE_01	M01000 V\$AIRE_02	
E4BP4	M00045 V\$E4BP4_01		
CACCT	M00412 V\$AREB6_01	M00413 V\$AREB6_02	M00414 V\$AREB6_03
	M00415 V\$AREB6_04	M00073 V\$DELTAEF1_01	
NRF	M00652 V\$NRF1_Q6	M00108 V\$NRF2_01	M00821 V\$NRF2_Q4
GFI	M00250 V\$GFI1_01	M01067 V\$GFI1_Q6	M01058 V\$GFI1B_01
NKX	M00485 V\$NKX22_01	M00240 V\$NKX25_01	M00241 V\$NKX25_02
	M01043 V\$NKX25_Q5	M00451 V\$NKX3A_01	M00424 V\$NKX61_01
	M00489 V\$NKX62_Q2		
RREB	M00257 V\$RREB1_01		

