Timeline of Deliverables

Saket Choudhary

March 15, 2016

Proposed title: Locating Transcription Factor Binding Sites: Methods in search of *moralistic* motifs
Expected length: 8-10 pages
References: See References section. In places where required I have mentioned relevant review articles that I would be keeping as my main reference.
Key reference: Assessing computational tools for the discovery of transcription factor binding sites [1]
Expected duration: 6 weeks(14th March to 30th April)
Section Organization: See table 2
Progress tracking: `https://trello.com/b/8hLZBkKA/review-paper`

Tool	Original Findings	Principle	Reference	*de-novo*
AlignACE	Discovery of over represented motifs in unaligned sequences, typically in the upstream region of corregulated genes	Gibbs sampling	[2]	✓
ANN-Spec	finding *low-complexity* patterns present in high frequency. Suitable for locating TFBS given a background sequence	Artificial Neural Network for parameter fitting to maximize posterior probability and then Gibbs sampling for	[3]	✓
Consensus	Statistically significant alignments of DNA or protein sequences to determine evolutionary/functional perspective	Greedy algorithm that searches for motifs maximizing information iteratively(Fixed width)	[4]	✗(Check, constraint: Fixed width)
Consite	TFBS prediction using phylogenetic footprinting	Accounts for evolutionary constraints by aligning regulatory sequence from orthologous pairs of genes	[5]	✗
GLAM	Locating functional sites by MSA	Simulated Annealing for automatically determining the width(An improvement over Stormo's Gibbs sampling approach [4]	[6]	✓
The Improbizer	Identifying cis-regulatory elements that activate gene-expression within pharyngeal gene clusters	EM algorithm, zero, first or second order markov model for background sequences	[7]	✓
MEME	Over-represented motifs in DNA, protein sequences	EM algorithm to search for optimum motif	[8]	✓
MONKEY	Identifying conserved transcription factor binding sites in MSA	Probabilistic model of binding site-specificity accounting for evolutionary tree by modeling one as background	[9]	✗
CENTIPEDE	Used to predict genome wide map of 800K TFBS of 200+ TFs	Integrates histone modification, gene annotation, DNAse I cleavage pattern to predict TFBS	[10]	✓
SeqGL	TFBS prediction using ChIP, DNAse or ATAC-seq data	Group lasso regularization to extract most important *k-mer* groups distinguishing peaks from flanking sequences followed by motif finding across regions that have non zero weight	[11]	✓
YMF	Statistically significant motifs	Given regulatory regions of *related* genes, find motifs with greater Z-score	[12]	✗
Weeder	Predicting Regulatory motifs	Models the significant occurrence of motifs over a seventh order markov chain expected background	[13]	✓
TFEM	TFBS prediction	Position specific priors based on phylogenetic conservation, penalization based on deviation from conserved profiles	[14]	✓

Kellis et al.	Prediction of regulator target for drosophila	BLS: Branch Length score based cutoffs for finding most significant motifs	[15]	✓
PhyloGibbs	Regulatory motif finding in multiple local sequence alignment of orthologous sequences	MCMC, simulated annealing based approach that treats alignments as the sites for binding and intergenic DNA as 'background', taking into account evolutionary distances	[16]	✗
REDUCE	Discovering cis-regulatory sequences using expression data without the need of gene clustering	Models the log fold change expression as a linear with the number of occurrences(or Information content) of motif as covariates. Motif lengths are determined by user	[17]	✗
GMEP	Modeling Sequence to expression(S2E) profiles. Hierarchical clustering of GMEP identified clusters of motifs with known TFs	Enumerate motifs for different length to find the weight contribution to gene expression, similar to the REDUCE algorithm discussed above	[18]	✗
EMnEM/PhyME	Identify motifs in orthologous sequences	Phylogenetic EM based approach	[19, 20]	✗
RCADE	Motif discovery in C2H2-ZF ChIP-seq data	Use a previously established recognition code for C2H2 to identify motifs in target sequences, which are then tested for enrichment using sequences from endogenous retroelements(ERE) and non-ERE regions	[21]	✗
DME	Identifying tissue specific TFBS	Identifies motifs over-represented in one set of sequences over the background(promoters of deferentially expressed genes)	[22]	✓
INSIGHT	Not for motif discovery, but to gauge impact of mutations on TFBS	Probabilistic model to measure impact of natural selection on TFBS	[23]	✗
rVISTA	Finding cis-regulatory sequences	Clustering of TFBS and inter-species sequence conservation	[24]	✗
TRAWLER	Regulatory motif discovery pipeline	Uses a suffix-tree implementation and a Z-score approximation	[25]	✓
MEME-prior	TFBS prediction	Prior probabilities based on phylogenetic or other background information assigned to bases	[26]	✓
FootPrinter	Regulatory motif prediction in homologous sequences	Uses MSA and evolutionary conservation to determine motifs	[27]	✗

Table 1: Tools, purpose, methods

References

[1] Martin Tompa, Nan Li, Timothy L Bailey, George M Church, Bart De Moor, Eleazar Eskin, Alexander V Favorov, Martin C Frith, Yutao Fu, W James Kent, Vsevolod J Makeev, Andrei A Mironov, William Stafford Noble, Giulio Pavesi, Graziano

Table 2: Timeline of deliverables

14^{th} **March**	Deliverables proposed.
22^{nd} **March**	Page 1-2: Introduction: • Why TFBSs matter? : Biological motivation (Relevant review: [?] • TFBS prediction : Computational Challenges(Hints from review article [28]) .
1^{st} **April**	Page 3-4:TFBS discovery methods overview: A two step approach: 1. Representation of Motif(Consensus Based or Profile Based) [Comprehensive Review: [29]] 2. Identifying binding sites given a motif representation .
15^{th} **April**	Page: 6-10 TFBS discovery methods: • Comparative approaches which do not account for conservation information, See table 1 (Just an overview) • Motif discovery and Phylogenetic footprinting various approaches, the relevant math involved, their shortcomings and biological relevance/generalization(Major chunk of discussion): – INSIGHT – EMnEM/PhyME – CENTIPEDE – PhyloGibbs – MONKEY – Kellis et al. – TFEM – Consensus – Consite .

Discussion:

- Recommendations for what method to choose for what type of data that capture more information biologically
- Recommendation for best practices? (steps to follow to minimize false positive motifs)
- Overview of statistics used for assessing motif(tool performance) quality

22^{nd} **April** ·····

.

29^{th} **April** ····· Conclusion.

Pesole, Mireille Rgnier, Nicolas Simonis, Saurabh Sinha, Gert Thijs, Jacques van Helden, Mathias Vandenbogaert, Zhiping Weng, Christopher Workman, Chun Ye, and Zhou Zhu. Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology*, 23(1):137–144, January 2005.

[2] Frederick P. Roth, Jason D. Hughes, Preston W. Estep, and George M. Church. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature biotechnology*, 16(10):939–945, 1998.

[3] C. T. Workman and G. D. Stormo. ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. In *Pac Symp Biocomput*, volume 5, pages 464–475. Citeseer, 2000.

[4] G. Z. Hertz and G. D. Stormo. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15(7):563–577, July 1999.

[5] A. Sandelin, W. W. Wasserman, and B. Lenhard. ConSite: web-based prediction of regulatory elements using cross-species comparison. *Nucleic Acids Research*, 32(Web Server):W249–W252, July 2004.

[6] Martin C. Frith, Ulla Hansen, John L. Spouge, and Zhiping Weng. Finding functional sequence elements by multiple local alignment. *Nucleic Acids Research*, 32(1):189–200, 2004.

[7] Wanyuan Ao, Jeb Gaudet, W. James Kent, Srikanth Muttumu, and Susan E. Mango. Environmentally Induced Foregut Remodeling by PHA-4/FoxA and DAF-12/NHR. *Science*, 305(5691):1743–1746, September 2004.

[8] Timothy L. Bailey, Charles Elkan, and others. Fitting a mixture model by expectation maximization to discover motifs in bipolymers. 1994.

[9] Alan M. Moses, Derek Y. Chiang, Daniel A. Pollard, Venky N. Iyer, and Michael B. Eisen. MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biol*, 5(12):R98, 2004.

[10] R. Pique-Regi, J. F. Degner, A. A. Pai, D. J. Gaffney, Y. Gilad, and J. K. Pritchard. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Research*, 21(3):447–455, March 2011.

[11] Manu Setty and Christina S. Leslie. SeqGL Identifies Context-Dependent Binding Signals in Genome-Wide Regulatory Element Maps. *PLOS Computational Biology*, 11(5):e1004271, May 2015.

[12] Saurabh Sinha and Martin Tompa. YMF: a program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Research*, 31(13):3586–3588, July 2003.

[13] G. Pavesi, P. Mereghetti, G. Mauri, and G. Pesole. Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Research*, 32(Web Server):W199–W203, July 2004.

[14] Katherina J Kechris, Erik van Zwet, Peter J Bickel, and Michael B Eisen. Detecting DNA regulatory motifs by incorporating positional trends in information content. *Genome Biology*, 5(7):R50, 2004.

[15] Pouya Kheradpour, Alexander Stark, Sushmita Roy, and Manolis Kellis. Reliable prediction of regulator targets using 12 Drosophila genomes. *Genome Research*, 17(12):1919–1931, December 2007.

[16] Rahul Siddharthan, Eric D. Siggia, and Erik van Nimwegen. PhyloGibbs: A Gibbs Sampling Motif Finder That Incorporates Phylogeny. *PLoS Computational Biology*, 1(7):e67, 2005.

[17] Harmen J. Bussemaker, Hao Li, and Eric D. Siggia. Regulatory element detection using correlation with expression. *Nature genetics*, 27(2):167–174, 2001.

[18] D. Y. Chiang, P. O. Brown, and M. B. Eisen. Visualizing associations between genome sequences and gene expression data using genome-mean expression profiles. *Bioinformatics (Oxford, England)*, 17 Suppl 1:S49–55, 2001.

[19] Saurabh Sinha, Mathieu Blanchette, and Martin Tompa. PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC bioinformatics*, 5(1):1, 2004.

[20] Alan M. Moses, Derek Y. Chiang, and Michael B. Eisen. Phylogenetic motif detection by expectation-maximization on evolutionary mixtures. In *Pacific Symposium on Biocomputing*, volume 9, pages 324–335. World Scientific, 2004.

[21] Hamed S. Najafabadi, Mihai Albu, and Timothy R. Hughes. Identification of C2h2-ZF binding preferences from ChIP-seq data using RCADE: Fig. 1. *Bioinformatics*, 31(17):2879–2881, September 2015.

[22] Andrew D. Smith, Pavel Sumazin, and Michael Q. Zhang. Identifying tissue-selective transcription factor binding sites in vertebrate promoters. *Proceedings of the National Academy of Sciences of the United States of America*, 102(5):1560–1565, 2005.

[23] Adam Siepel and Leonardo Arbiza. Cis-regulatory elements and human evolution. *Current Opinion in Genetics & Development*, 29:81–89, December 2014.

[24] Gabriela G. Loots, Ivan Ovcharenko, Lior Pachter, Inna Dubchak, and Edward M. Rubin. rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Research*, 12(5):832–839, May 2002.

[25] Laurence Ettwiller, Benedict Paten, Mirana Ramialison, Ewan Birney, and Joachim Wittbrodt. Trawler: de novo regulatory motif discovery pipeline for chromatin immunoprecipitation. *Nature Methods*, 4(7):563–565, July 2007.

[26] G. Cuellar-Partida, F. A. Buske, R. C. McLeay, T. Whitington, W. S. Noble, and T. L. Bailey. Epigenetic priors for identifying active transcription factor binding sites. *Bioinformatics*, 28(1):56–62, January 2012.

[27] M. Blanchette. FootPrinter: a program designed for phylogenetic footprinting. *Nucleic Acids Research*, 31(13):3840–3842, July 2003.

[28] Matthew T Weirauch, Atina Cote, Raquel Norel, Matti Annala, Yue Zhao, Todd R Riley, Julio Saez-Rodriguez, Thomas Cokelaer, Anastasia Vedenko, Shaheynoor Talukder, Phaedra Agius, Aaron Arvey, Philipp Bucher, Curtis G Callan, Cheng Wei Chang, Chien-Yu Chen, Yong-Syuan Chen, Yu-Wei Chu, Jan Grau, Ivo Grosse, Vidhya Jagannathan, Jens Keilwagen, Szymon M Kiebasa, Justin B Kinney, Holger Klein, Miron B Kursa, Harri Lhdesmki, Kirsti Laurila, Chengwei Lei, Christina Leslie, Chaim Linhart, Anand Murugan, Alena Myikov, William Stafford Noble, Matti Nykter, Yaron Orenstein, Stefan Posch, Jianhua Ruan, Witold R Rudnicki, Christoph D Schmid, Ron Shamir, Wing-Kin Sung, Martin Vingron, Zhizhuo Zhang, Harmen J Bussemaker, Quaid D Morris, Martha L Bulyk, Gustavo Stolovitzky, and Timothy R Hughes. Evaluation of methods for modeling transcription factor sequence specificity. *Nature Biotechnology*, 31(2):126–134, January 2013.

[29] Gary D. Stormo. Modeling the specificity of protein-DNA interactions. *Quantitative Biology*, 1(2):115–130, June 2013.

[30] G. D. Stormo. DNA binding sites: representation and discovery. *Bioinformatics (Oxford, England)*, 16(1):16–23, January 2000.

[31] Gerald Z. Hertz, George W. Hartzell, and Gary D. Stormo. Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Bioinformatics*, 6(2):81–92, 1990.

[32] Rahul Siddharthan. Dinucleotide Weight Matrices for Predicting Transcription Factor Binding Sites: Generalizing the Position Weight Matrix. *PLoS ONE*, 5(3):e9722, March 2010.

[33] Albin Sandelin and Wyeth W. Wasserman. Prediction of Nuclear Hormone Receptor Response Elements. *Molecular Endocrinology*, 19(3):595–606, March 2005.

[34] I. G. Lyakhov, A. Krishnamachari, and T. D. Schneider. Discovery of novel tumor suppressor p53 response elements using information theory. *Nucleic Acids Research*, 36(11):3828–3833, May 2008.

[35] Todd Riley, Xin Yu, Eduardo Sontag, and Arnold Levine. The p53hmm algorithm: using profile hidden markov models to detect p53-responsive genes. *BMC Bioinformatics*, 10(1):111, 2009.

[36] Li-San Wang, Shane T. Jensen, and Sridhar Hannenhalli. An interaction-dependent model for transcription factor binding. In *Systems Biology and Regulatory Genomics*, pages 225–234. Springer, 2007.

[37] Matti Annala, Kirsti Laurila, Harri Lhdesmki, and Matti Nykter. A Linear Model for Transcription Factor Binding Affinity Prediction in Protein Binding Microarrays. *PLoS ONE*, 6(5):e20059, May 2011.

[38] Yue Zhao, David Granas, and Gary D. Stormo. Inferring binding energies from selected binding sites. *PLoS Comput Biol*, 5(12):e1000590, 2009.