Transcription Factor Binding Site Prediction & Phylogenetic Footprinting: A Review

Saket Choudhary

March 2016

Introduction

The completion of human genome sequencing[1] revealed that a major portion of this genetic material does not code for protein sequences. Although this portion once termed "junk" DNA, is now known to be functionally active and conserved over evolution[2]. It is often referred to as *'conserved non-genic sequence'*(CNS)

Vertebrate genomes have around twice the number of genes compared to the invertebrates. A lot of these are attributed to gene duplication[3]. There is substantial evidence to establish that the difference in organism complexity arises from elaborate regulation of gene expression. A small number of genes could be exploited to generate complex systems in higher order organisms(vertebrates). One such mechanism involves alternative splicing, which involves one gene coding for multiple proteins. Another mechanism involves DNA rearrangement. However, it has often been argued that the physiological and behavioral complexity must arise from the elaboration of the complexity present at *cis-* regulatory DNA sequences[3, 4]. The regulatory landscape consists of *trans-* acting proteins that bind at different sites along the *cis*-regulatory sequence. These modules are regulated by multiple transcription factors and each transcription factor can interact with multiple genes. Sequence specific DNA-binding proteins called transcription factors(TF) bind near the transcription start site in the promoter region but can also bind to the region being transcribed or distal regulatory elements such as enhancers and silencers[5]. These DNA binding proteins can themselves be regulated at the transcription level. DNA-TF binding is sequence specific and hence TFs can bind at locations that are *similar* but not necessarily identical. The combinatorial nature of transcriptional regulation leads to dramatic increase in regulatory complexity.

Systematic identification of regions where the TFs bind can help us understand transcription regulation in a better way, providing the means to understand and model cellular response to different stimuli.

The Biological Problem

To understand the cis-regulatory mechanism, it is important to identify the sequence specific patterns or 'motifs' associated with TFs. Having identified these motifs, the next question to ask is where do these occur in the non-coding region. Both these problems are analogous to finishing a partially complete jigsaw puzzle, where the first step will involve looking for possible pieces that can fill the available gaps and then having chosen a piece, looking for a place where it would *most likely* fit. Sophisticated methods developed over the last one and half decades have tried to tackle both these problems simultaneously. In the following discussion we refer to 'motifs' as the specificity or pattern representation associated with transcription factors and 'instances' , 'matches' , 'TF binding sites' or 'regulatory sequences' to those positions in the non coding DNA where the 'motifs' bind. In the next few sections, I describe the different approaches taken to solve this problem.

The Computational Problem

Given a set of DNA sequences typically around 100-200 base pairs long, the goal is to find a *shorter* sequence typically 5-25bp long which occur *frequently*, allowing for errors at individual sites(since TF-DNA binding can involve similar but not necessarily identical sequences). Also for the motif to be *significant*, it should not be as frequent in another set of DNA sequences, either pooled randomly or generated using some 'biologically feasible' model.

Futility Theorem

Wasserman and Sandelin in their review of tools for motif discovery[6] conjectured 'Futility Theorem' asserting that essentially all TF binding sites(TFBS) predicted using models for the individual binding of the TFs will have no functional role. One of the frequent assumptions underlying methods for TFBS prediction is that each TF binds individually. However it is well known that it is the combinatorial interaction of multiple factors binding to multiple cis-regulatory sites regulate gene transcription[7]. These assumptions lead to a limitation - inability to distinguish between sites that are functional and non-functional, in-vivo. Few of the methods discussed in the following sections, often fail to overcome this theorem's claim.

First generation: Pattern matchers

First generation of motif discovery tools relied on performing a multiple sequence alignment and then searching for a motif which maximizes some objective function. For example, CONSENSUS[8] implemented such a *de-novo* discovery method by maximizing the information criterion of patterns in multiple sequence alignment. However, such methods that do not use any extra information apart from the alignment are susceptible to a lot of false positive motifs arising firstly due to the inherent degeneracy in TF binding specificity and secondly because of the fact that an over-represented sequence might not necessarily be a true binding site. This is often the case if the aligned sequences have a lot of repeats. These methods often rely on likelihood ratios and $p - values$ to determine statistically significant 'motifs' which are likely to be significant even for nonfunctional repeats. A possible example where Futility theorem holds.

Second generation: Combination and Conservation of Motifs

Tagle *et al.* proposed the term *phylogenetic footprint* referring to the phylogenetic comparisons that reflect evolutionary conserved functional elements in homologous genes[9].

Gelfand *et al.* were one of the first to use comparative genomics to predict transcription factor binding sites[10]. Using the available experimental data, they wanted to predict sites of operator sequences in E.coli. However, the information contained in these datasets was limited. They hypothesized that the regulatory signal must be conserved across some of the evolutionary close species of E.coli and hence adding sequences from co-regulated genes and orthologous regions of other species could help in discovering patterns in TFBS. By combining profile-based search with phylogenetic analysis, they were able to overcome the limitations of missing experimental data. This idea was further used by McGuire *et al.* [11] to discover new motifs in the upstream regions of co-regulated genes. Detecting motifs that are overrepresented is easy, however, pooling orthologous regions from closely related organisms provides an extra layer of information that can help in discovery of these conserved motifs, assuming these sites are always under constraint due to purifying selection. They demonstrated how an experimentally verified E.coli motif(MetR) is not found if only E.coli sequences are used, since there are very few instances of this motif compared to the *background* non-functional sequence. However, when pooled with orthologous regions from B.subtilis, it was found, since it is over represented in the pooled sample. Hence they concluded that a conserved operon is likely to represent a functional coupling if it is also found to be conserved across large evolutionary distances.

Using evolutionary distant species can help overcome Simpson's paradox(or the Yule-Simpson effect)[12] where closely related species can be the confounding factor. In other words, if two species are closely related evolutionary, there TFBS can show high conservation, but it is difficult to determine if the conservation arises from the fact that they are selectively conserved being TFBS or is it simply because all other regions are conserved too. A discussion on controlling for such conservation in methylation studies is discussed in a recent paper[13].

Caveats of Gelfand and McGuire approach:

- Ignores phylogenetic relationships, often giving way to Yule-Simpson effects: phylogenetically close sequences are bound to have most portions of the DNA conserved
- In at least one of the earlier studies by Lane *et al.* [1], the instances of orthologous regulatory regions were found to be more conserved(since they arose from the same ancestral sequence) when compared to instances across the co-regulated genes of the same genome. Since both these studies relied on pooling sequences from co-regulated genes along with the orthologs, there is an inherent loss of distinction
- number of occurrences of regulatory elements might not be comparable when they are pooled with their orthologs. Some regulatory regions will have zero occurrences while others will have multiple.

This variance is expected to be lower across the orthologous genes since they evolved from the single ancestral sequence

Conservation: Biological implications

My focus in the following sections is on methods that employ phylogenetic footprinting to predict TFBS. The approach of finding functional segments in non-coding DNA has been a pretty old one, probably dating to 1971 when Pribnow *et al.* [14] determined the sequence of promoters in T7 bacteriophages which was reported to have similarities with the bacteriophage fd and a lac UV5 promoter.

Several studies have used phylogenetic footprinting to identify functional regulatory elements. For example, Loots *et al.* [15] used comparative genomics to find distal regulatory regions. In their study a sequence alignment from chromosomal region 5q31 and an orthologous region in mouse lead to the discovery of 90 conserved noncoding sequences(ungapped sequence alignment of at least 100bp with at least 70% identity), which on experimental validation(knockout experiments) validated that they were indeed regulatory.

In a separate experiment contrasting the experiment, Kim *et al.* [16] found very slow divergence rates of enhancers of HOX clusters when compared to the primitive horn shark. So they essentially went beyond Loots *et al.'s* approach of just using phylogenetically close species(mammals), to say that a comparison between mammals and fish will reveal regions that are under stronger selection and hence likely to be involved in more critical functions. Essentially, regions conserved between human and mouse might not be so in human and fish, but the *most* conserved regions can be trusted to be functionally relevant. It is also important to note that criteria used for finding CNS at one locus might be too stringent for other loci. One such example of this study, is Gottgens *et al.* [17] which found a lot of CNS when comparing human-mouse sequences, but one enhancer was discovered in chicken-human sequence, which was then experimentally shown to be a neural enhancer

The simple premise underlying these comparative methods is that selective pressure acts on the functional elements in CNS and causes them to evolve at a much slower rate compared to the the flanking non-functional elements and hence set of orthologous regions showing conservation can be good candidates for TF binding. However, there is at least one caveat with this hypothesis. Species close to each other will show high degree of conservation because while evolving, the time for substitution to occur has been insufficient. As the distance increases, non functional sites will undergo far more substitutions and hence conservation detection can happen more confidently until the point that the one fails to find sufficient orthology between the species being aligned. Hence, it is important to ensure that the predicted conserved regions do not show conservation merely because they were found using two species which are very close to each other.

The following methods try to integrate two levels of information for *de-novo* motif discovery: over representation and conservation in orthologous regions. The motivation to do so is quite intuitive, if the motif is insignificant using either level of information, it might turn out to be significant when they are combined.

MONKEY

In a 2003 paper on evolution of transcription factor binding sites, Moses *et al.* [18] concluded that these sites evolved slowly compared to the surrounding sequences, thus asserting their hypothesis that they were under purifying selection. A striking result was the positive correlation between rate of evolution and the degeneracy of the binding site, implying if a particular site in the motif is evolving, the degeneracy of TF-DNA binding at that position also increases allowing for other bases to be present. They also leveraged this observation to predict rate of evolution at individual bases of TFBS under the assumption that the position weight matrix(PWM) reflects the allowed sequence specificity for TF-DNA binding. For example, consider a site where the PWM is of this form $f_A, f_C, f_G, f_T = (1, 0, 0, 0)$ then a mutation at any site will most likely inhibit TF-DNA binding. However distributions like $f_A, f_C, f_G, f_T = (1/2, 0, 0, 1/2)$ would lead to changes at C, G to be removed over the course of evolution, whereas for a distribution of the form $f_A, f_C, f_G, f_T = (1/4, 1/4, 1/4, 1/4)$ all substitutions are permitted. They used this model to make predictions about rate of evolution for transcription factors with experimentally characterized binding sites and found them to be in-sync. Though the motifs appeared to be conserved on average, the individual binding sites need not be perfectly conserved. And hence it is important to note that simply searching for perfectly conserved segments need not reveal the true binding sites. They thus demonstrated that conservation relative to flanking sequences and correlation between position-specific rate of evolution and intragenomic degeneracy can be used as a proxy to identify the *bona fide* transcription factor binding sites from computational artifacts.

Using the results from the earlier study, Moses *et al.* developed MONKEY[19] which employs a probabilistic framework to assess factor specificity and binding site evolution to compute the likelihoods of motif hits in multiple sequence alignments. The goal here was to identify conserved binding sites while simultaneously accounting for sequence specificity, pattern of evolution, and phylogenetic relationship of the species being compared.

MONKEY is not a *de-novo* motif finder. Given a multiple sequence alignment, a model of transcription factor's binding specificity, and a model for background noncoding DNA, it returns the likelihood ratio of each position being conserved over the probability that it is background.

It uses a statistic commonly used for scoring the similarity of a single sequence to a frequency matrix: ratio of the logarithm of the probabilities of observing the sequence under the motif model to the probability of observing it under a background model. Since the hypothesis is that the aligned sequences arose from a common ancestor(which is a hidden variable) this score is calculated by averaging over all possible ancestral sequences. These calculations also involve a rate matrix, quantifying the probability of observing each base in the ancestral sequence given the time or distance of divergence. The background model involves use of Jukes-Cantor or Hasegawa-Kishino-Yano model of DNA substitution.

With MONKEY another problem being investigated was which species are optimal for defining conservation. However this problem mostly remained unanswered since different transcription factors exhibited different conservation patterns for the same set of species, although in general the $p - values$ corresponding to the hypothesis of the TFBS originating from the background decreased with increasing evolutionary distances.

Caveats:

- Not suitable for *de-novo* motif discovery
- assumes each site is evolving independently (this assumption is almost always not true)

PhyME and PhyloGibbs

PhyME[20] uses expectation maximization to search for relevant motifs. Motif evaluation involves accounting for its occurrence in orthologous regions, which are assumed to be related by a probabilistic model that takes into account the different phylogenetic distances between species.

This approach is mostly similar to PhyloGibbs[21] except that here expectation-maximization(EM) is used instead of Gibbs sampling and secondly PhyloGibbs assumes star topology for phylogeny while there is no such assumption in PhyME. Another approach, orthoMEME[22] performs EM based motif elucidation by accounting for homology simultaneously. It however, has this stronger assumption that each motif occurrence has an orthologous analog in the other species. This is often not true, especially when the cis-regulatory sequence is specie specific.

It is important to note that PhyME is not a *de-novo* motif-finder. It requires the length of motif to be inputted along with the sequences. It is assumed that there is also a master reference sequence such that its sequence data comprises all the promoter sequences inputted. Besides this, the input consists of a phylogenetic tree over different species specifying the neutral point mutation rate at each branch point. By performing multiple sequence alignment it selects contiguous regions that are conserved in the master reference. There will be aligned portions in these multiple sequences and then there will be some 'bracketed' sequences which do not align with the reference but have flanking regions aligned to it. A hidden markov model is then trained using essentially two states: a motif or background over the aligned regions accounting for which species in the phylogenetic tree the alignments arose from, to maximize an objective function.

Caveats: Not suitable for *de-novo* motif discovery; Assumes all positions in the binding site evolve independently(this can probably be accounted for by fitting a higher order markov chain, but was not discussed in the original paper); requires presence of a 'master reference' with all promoters present in that sequence

INSIGHT:

Inference of Natural Selection from Interspersed Genomically coHerent elemenTs (INSIGHT) is a probabilistic method that characterizes the effects of natural selection on collections of short transcription factor binding sequences[23].

Just like the previous methods, they also argue that even though transcription factor binding sites have experienced weaker selection than protein coding regions, there is enough evidence of evolutionary adaptation.

It can be argued that mutations since cis-regulatory mutations are often co-dominant, and hence natural selection might act more strongly on these as compared to protein coding mutations.

It assumes that nucleotides within TFBs evolve by a mixture of four selective models: 1) neutral drift 2) weak negative selection 3) strong negative selection 4) positive selection. A strong positive or negative selection will cause mutations to rapidly reach fixation or be lost. A weak negative selection allows polymorphism, allowing degeneracy in the TFBS consensus sequence. Positive selection allows for fixation of derived alleles while negative selection ensures elimination of deleterious ones. The flanking sequences are assumed to be evolving neutrally. Information about overall prevalence of selection comes from the rate of alleles, positive selection rate comes from the rate of divergence, and information about weak negative selection comes from the relative rates of low and high frequency derived alleles. By analyzing different transcription factors from the ENCODE datasets, they detected a strong signature of natural selection in TFBS compared to the flanking sequences. They also found that the information content is positively correlated with the fraction of sites under selection, thus indicating higher information content at a particular site will often have a high conservation too.

Third generation: Combining Information from multiple sources

Over the past half decade, the advent of technologies such as ChIP-seq, DNAse-seq, and other epigenetic-seq studies has been used to create models that along with phylogenetic footprinting produce promising results.

MEME and priors

MEME[24] was the first EM based *de-novo* motif finder. It uses EM to find over-represented motifs and hence as such does not take into account any biological feature. However, this extra biological information can be easily integrated in the form of priors.

Several molecular mechanisms limit TF-DNA binding. Local chromatin structure in its 'closed' form hinders TF access to DNA. A lot of studies have also established the association of epigenetic marks such as mono and tri methylation of H3K4[25] and hypersensitivity to DNAse I digestion[26]

With the *easy* availability of this epigenomic data in the past few years, MEME leverages this extra information in the form of Bayesian *priors*. This is an extension of the work done earlier in a tool called PRIORITY[27, 28]

These priors represent the probability of a TF binding a certain position i given that the tag count at position is y_i. This tag count can be from any assay: histone modification, DNAse I etc. The tag counts can be mapped to probability values using a linear mapping function at each position i. So the prior $P(B_i = 1|y_i)$ is the probability of binding ($B_i = 1$) given the tag count at i^{th} position is y_i Using DNAse I sensitivity and histone marks H3K4Me1, H3K4Me3, H3K9Ac,H3K27Ac (which are known to correlate with transcriptional activation) they create a prior and using H3K27Me3 as a negative control asses the log posterior scores compared to a degenerate prior. They concluded that using H3K4Me3 priors improved TFBS prediction in mouse ES cells while using histone modification and DNAse I hypersensitivity improved predictions in human K562 and GM12878 cell lines.

CENTIPEDE

Centipede[29] is a probabilistic framework that integrates cell/tissue-specific experimental data(histone marks, DNAse I hypersensitivity, gene annotation and phylogenetic footprinting) and uses mixture models to predict TFBS. CENTIPEDE can identify binding sites for many factors from a single experimental assay.

The underlying hypothesis here is the same as that was used for MEME using epigenetic priors, the sites that are bound are more likely to be associated with an open chromatin region and would often be associated with active histone marks, evolutionary conservation. For each candidate binding site CENTIPEDE separates the information into two components: G and D. G: Genomic information independent of cell type or experimental conditions. D: Cell-specific experimental data(DNAse I, histone marks). Here the information contained in G is treated as a prior and then the likelihood $P(D|Bound)$ is modeled depending on whether the site is bound or not. In addition to the histone and DNAse I marks, other informative priors are also incorporated: PWM match score, proximity to the nearest TSS and evolutionary conservation.

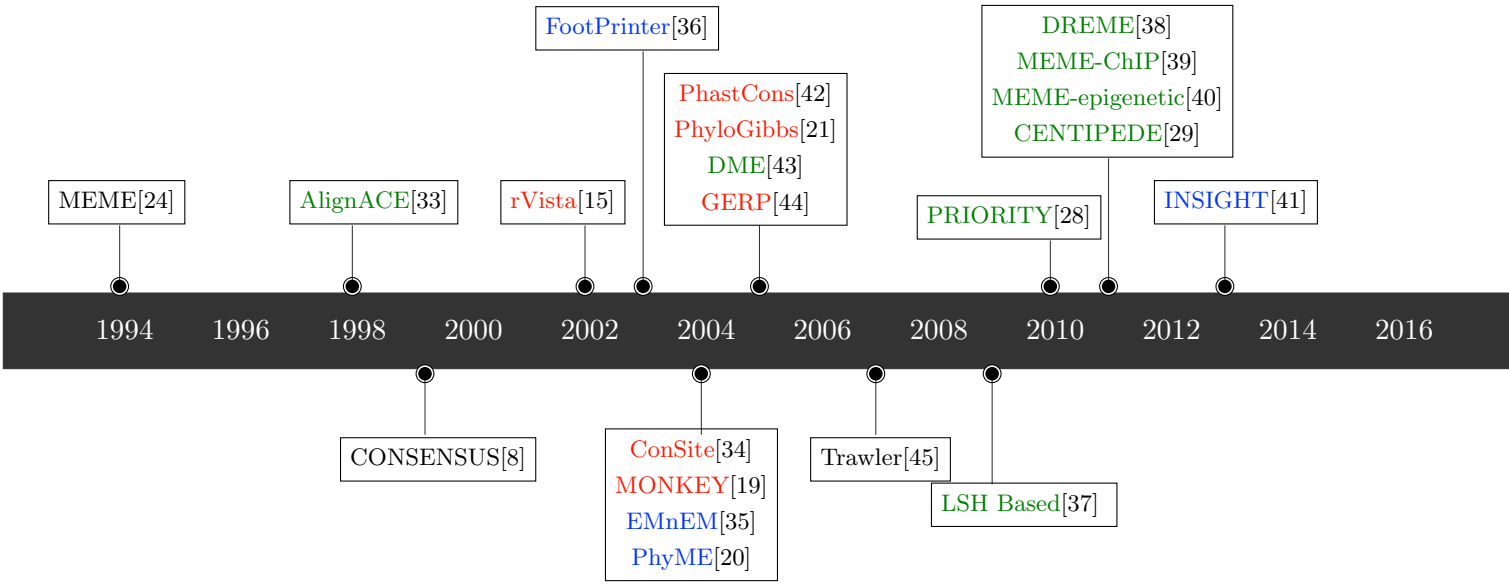

Figure 1: Timeline of different tools used for motif discovery or phylogenetic footprinting. The list is non-exhaustive. For a more detailed list refer to the following reviews:[46, 47, 48]

Conclusions

A lot of methods have leveraged phylogenetic footprinting to predict TFBS based on the hypothesis that these sequences evolve slowly compared to the neighboring sequences. However, leveraging just this piece of extra information has not been able to disprove the *Futility Theorem* . Use of mixture models and Bayesian priors leveraging information from the open chromatin region sounds promising. For example, a recent paper from Siepel *et al.* [30] explored using machine learning approach to reveal active transcriptional regulatory elements using GRO-seq[31].

These methods coupled with other *-seq technologies mapping open chromatin regions such as GRO-seq and ATAC-seq[32] can help in decoding the regulatory map of the genome at higher resolutions.

Thus improving these third generation methods that use cascaded information from different sources, can promise better characterized regulatory maps in the coming few years.

References

[1] Eric S. Lander, Lauren M. Linton, Bruce Birren, Chad Nusbaum, Michael C. Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, and others. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[2] Stylianos E. Antonarakis, Robert Lyle, Emmanouil T. Dermitzakis, Alexandre Reymond, and Samuel Deutsch. Chromosome 21 and down syndrome: from genomics to pathophysiology. *Nature Reviews Genetics*, 5(10):725–738, October 2004.

[3] Michael Levine and Robert Tjian. Transcription regulation and animal diversity. *Nature*, 424(6945):147–151, 2003.

[4] Michael Levine and Eric H. Davidson. Gene regulatory networks for development. *Proceedings of the National Academy of Sciences of the United States of America*, 102(14):4936–4942, 2005.

[5] Bryan Lemon and Robert Tjian. Orchestrated response: a symphony of transcription factors for gene control. *Genes & development*, 14(20):2551–2569, 2000.

[6] Wyeth W. Wasserman and Albin Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics*, 5(4):276–287, April 2004.

[7] Eric H. Davidson. *Genomic Regulatory Systems: In Development and Evolution*. Academic Press, January 2001.

[8] 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.

[9] Danilo A. Tagle, Ben F. Koop, Morris Goodman, Jerry L. Slightom, David L. Hess, and Richard T. Jones. Embryonic and globin genes of a prosimian primate (Galago crassicaudatus). *Journal of Molecular Biology*, 203(2):439–455, September 1988.

[10] M. S. Gelfand, E. V. Koonin, and A. A. Mironov. Prediction of transcription regulatory sites in Archaea by a comparative genomic approach. *Nucleic Acids Research*, 28(3):695–705, February 2000.

[11] Abigail Manson McGuire, Jason D. Hughes, and George M. Church. Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. *Genome research*, 10(6):744–757, 2000.

[12] E. H. Simpson. The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 13(2):238–241, 1951.

[13] Meromit Singer and Lior Pachter. Controlling for conservation in genome-wide DNA methylation studies. *BMC Genomics*, 16(1), December 2015.

[14] David Pribnow. Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. *Proceedings of the National Academy of Sciences*, 72(3):784–788, 1975.

[15] 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.

[16] Chang-Bae Kim, Chris Amemiya, Wendy Bailey, Kazuhiko Kawasaki, Jason Mezey, Webb Miller, Shinsei Minoshima, Nobuyoshi Shimizu, Gnter Wagner, and Frank Ruddle. Hox cluster genomics in the horn shark, Heterodontus francisci. *Proceedings of the National Academy of Sciences*, 97(4):1655–1660, 2000.

[17] Berthold Gttgens, Linda M. Barton, Michael A. Chapman, Angus M. Sinclair, Bjarne Knudsen, Darren Grafham, James GR Gilbert, Jane Rogers, David R. Bentley, and Anthony R. Green. Transcriptional regulation of the stem cell leukemia gene (SCL)Comparative analysis of five vertebrate SCL loci. *Genome research*, 12(5):749–759, 2002.

[18] Alan M. Moses, Derek Y. Chiang, Manolis Kellis, Eric S. Lander, and Michael B. Eisen. Position specific variation in the rate of evolution in transcription factor binding sites. *BMC evolutionary biology*, 3(1):1, 2003.

[19] 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.

[20] 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.

[21] 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.

[22] A. Prakash, M. Blanchette, S. Sinha, and M. Tompa. Motif discovery in heterogeneous sequence data. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 348–359, 2004.

[23] Leonardo Arbiza, Ilan Gronau, Bulent A. Aksoy, Melissa J. Hubisz, Brad Gulko, Alon Keinan, and Adam Siepel. Genome-wide inference of natural selection on human transcription factor binding sites. *Nature Genetics*, 45(7):723–9, July 2013.

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

[25] Artem Barski, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Dustin E. Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao. High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell*, 129(4):823–837, May 2007.

[26] J. A. Bernat. Distant conserved sequences flanking endothelial-specific promoters contain tissue-specific DNase-hypersensitive sites and over-represented motifs. *Human Molecular Genetics*, 15(13):2098–2105, May 2006.

[27] Raluca Gordn, Leelavati Narlikar, and Alexander J. Hartemink. A fast, alignment-free, conservation-based method for transcription factor binding site discovery. In *Research in Computational Molecular Biology*, pages 98–111. Springer, 2008.

[28] R. Gordan, L. Narlikar, and A. J. Hartemink. Finding regulatory DNA motifs using alignment-free evolutionary conservation information. *Nucleic Acids Research*, 38(6):e90–e90, April 2010.

[29] 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.

[30] Charles G Danko, Stephanie L Hyland, Leighton J Core, Andre L Martins, Colin T Waters, Hyung Won Lee, Vivian G Cheung, W Lee Kraus, John T Lis, and Adam Siepel. Identification of active transcriptional regulatory elements from GRO-seq data. *Nature Methods*, 12(5):433–438, March 2015.

[31] Leighton J. Core, Joshua J. Waterfall, and John T. Lis. Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. *Science*, 322(5909):1845–1848, December 2008.

[32] Jason D Buenrostro, Paul G Giresi, Lisa C Zaba, Howard Y Chang, and William J Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, 10(12):1213–1218, October 2013.

[33] 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.

[34] 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.

[35] 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.

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

[37] Zeeshan Syed, Piotr Indyk, and John Guttag. Learning approximate sequential patterns for classification. *The Journal of Machine Learning Research*, 10:1913–1936, 2009.

[38] T. L. Bailey. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, 27(12):1653–1659, June 2011.

[39] P. Machanick and T. L. Bailey. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics*, 27(12):1696–1697, June 2011.

[40] 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.

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

[42] Adam Siepel, Gill Bejerano, Jakob S. Pedersen, Angie S. Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, LaDeana W. Hillier, Stephen Richards, and others. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*, 15(8):1034–1050, 2005.

[43] 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.

[44] Gregory M. Cooper, Eric A. Stone, George Asimenos, Eric D. Green, Serafim Batzoglou, and Arend Sidow. Distribution and intensity of constraint in mammalian genomic sequence. *Genome research*, 15(7):901–913, 2005.

[45] 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.

[46] F. Zambelli, G. Pesole, and G. Pavesi. Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. *Briefings in Bioinformatics*, 14(2):225–237, March 2013.

[47] Alan Moses and Saurabh Sinha. Regulatory Motif Analysis. In David Edwards, Jason Stajich, and David Hansen, editors, *Bioinformatics*, pages 137–163. Springer New York, New York, NY, 2009.

[48] 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 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.