

# Footprinting with additional knowledge

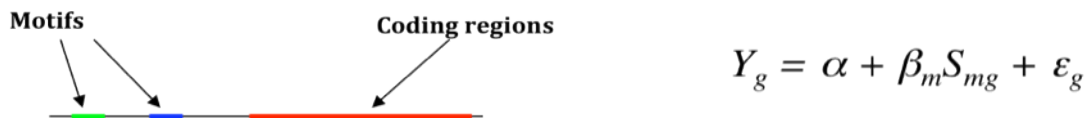
*Expression Levels, ChIPSEQ, Nucleosome Positioning and DNA Hypersensitive Sites*

16.2.10 Richard Mott and Jotun Hein

Finding regulatory motifs in front of genes can be approached in a variety of ways. If only sequences/genomes are observed, two complementary methods are: Observing independent genes with some commonality to their regulation and then search for common segments to these genes. An example of this approach is Lawrence et al. (1993). An alternative is footprinting that takes related genes and searches for slowly evolving segments or even more powerful, segments with a common mode of evolution (Satija et al. (2010), Moses et al., (2004)). We will mainly be concerned with footprinting, but the framework can be extended to incorporate independent sequences as well.

Alternatively or supplementary to the above pure sequence approach additional data can be available and this situation is increasingly predominant. Four major sources are expression levels, ChIPSEQ, nucleosome positioning, and DNA Hypersensitive Sites (DHS). A few comments on each:

1. Many genes and their expression levels are observed and it is then investigated, which signals in front of them are responsible for the expression level. Footprinting has been explored in many papers by now – recently in a series of papers by Rahul Satija, that also combines this with statistical alignment. Some of the first papers to use expression levels to define signals were Bussemaker et al. (2001) and Conlon et al. (2003). “The set of possible signals” is too large and Conlon reduced this considerably, by only investigating a pre-given subset of motifs. The effect of each motif was then determined by regression: the observed expression levels were fitted as a linear combination of effects from the possible signals as shown below. The assumptions of linear effects are often referred to as Omes Law.



$Y_g$  is the observed expression level of gene  $g$ .  $\alpha$  is basic common expression level,  $\beta_m$  is the effect of motif  $m$ ,  $S_{mg}$  is the 1/0 dependent of presence/absence of motif  $m$  in front of gene  $g$ .  $\epsilon_g$  is a gene specific error gaussian error term.

There are many natural extensions and issues to this simple setup. Three major are: motifs  $\rightarrow$  effect, set of motifs and finally inference in the many models thus defined. The first, motifs  $\rightarrow$  effect, is above assumed to be linear and independent, but could be considerably more complicated functions of a set of motifs and possibly involving concentration of regulatory molecules. Given the the size of the general solution space and the noisiness of the available data, it is reasonable to simplify in the first attempts at analyzing this problem and become more general, when data analysis, demonstrates the need. The first simplifying assumption is linearity and independence of effects, the second is to only consider small set of non-overlapping motifs. Inference can now be done by a sequential greedy standard goodness of fit where the best motif repeatedly added until no significant increase is obtained in fit.



ChIP SEQ experiments will give information on where a given protein (typically a regulatory molecule) interacts with DNA, but a considerable amount of error. Since the true protein-DNA interaction site is assumed to be the regulatory motif, this experiment is highly infomative.

2. ChIP-SEQ experiments can give information on the TF DNA interaction position, but originally only within 1-2 kb, but this clearly is valuable information considering the genome is about a million times longer - 3 Gb long. Several papers (Liu et al., 2002; Lun et al., 2009) have made models that used this information to predict motif

positions. Since many independent observations can be made, the final estimate of the TF-DNA interaction position can be quite narrow.

3. Several related experiments give information on the approximate locations of control elements in the genome. ChipSeq and its relatives (which, for example, detect DNase hypersensitive sites, and nucleosome positioning Segal et al.(2006), Narlikar et al.(2007a,b)) can be used to refine the search for motifs as they reduce the size of the search space several orders of magnitude. However the experiments also have technical limitations – the data are imperfect – and, more importantly, the regions identified vary by tissue and developmental stage. There may also be variation caused by differences in the genetic background: a region may be deleted in some individuals, or the presence of SNPs within the region may inhibit binding, or the activity of a site might be influenced by variation elsewhere.

By collecting data on variation of genome sequence, transcriptome and control elements in a defined set of conditions (tissues and developmental stage) in the same individuals it may be possible to understand precisely which DNA variants control levels of transcription and how this varies between tissues.

Combining the above additional sources of data with footprinting or motif overrepresentation has two major advantages: Firstly, it allows a rational selection in the set of possible signals to a much smaller set. Secondly, it allows simultaneous analysis of extra data from different species. Footprinting/motif overrepresentation can handle multiple pre-given signals, so incorporating an observed expression level to the likelihood function is in principle straightforward. A main issue is to make computations realistic.

**Project.** The aims of this project are to identify the binding sites that control transcription and sequence variants that modulate it. Optimal data analysis will use all sources of information simultaneously. It would be of major interest to develop a unified approach and quantify the value of different sources of data and apply this to simulated data and a very well defined data set.

### **Plan.**

- Week 1-2: Read key papers from the literature list and make preliminary contents of the report.
- Week 3-5: Use Chip-Seq data to identify approximate locations of regulatory regions (up to 200bp) and search for motifs within these regions
- Week 6-8: Use genome variation data to find sequence variants in the predicted regulatory regions
- Week 9-10: Correlate gene expression (RNA-Seq) data with DNA variation identified above.

### **References**

- Bussemaker, Li & Siggia (2001) *Regulatory element detection using correlation with expression nature genetics* 27.167-171
- Bussemaker et al. (2007) *Predictive Modeling of Genome-Wide mRNA Expression: From Modules to Molecules Annu. Rev. Biophys. Biomol. Struct.* 36:329–47
- Conlon et al. (2003) *Integrating Motif Discovery and Expression Analysis Proc.Natl.Acad.Sci.* 100.3339-44
- Eden et al. (2009) *Discovering Motifs in Ranked Lists of DNA Sequences PLOS compuBiol* 3.3.3e39
- Gao et al.(2004) *Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data BMC Bioinf.* 5.3140
- Halperin et al. (2009) *Allegro: Analyzing expression and sequence in concert to discover regulatory programs Nuc Ac. Rex* 37.5.1566-79
- Higgs, DR et al.(2007) *Using Genomics to Study How Chromatin Influences Gene Expression Annu Rev. Genom Hum Genet.* 8.299-325
- Holmes and Bruno (2000) *Finding Regulatory Elements Using Joint Likelihoods for Sequence and Expression Profile Data ISMB* 202-210
- Kim and Ren (2006) *Genome-Wide Analysis of Protein-DNA Interactions Annu Rev. Genom Hum Genet.* 7.81-102
- Liu, et al.(2002) *"An algorithm for finding protein-DNA binding sites with applications of chromatin-immunoprecipitation microarray experiments Nature biotechnology* 8.835-
- Lun et al. (2009) *A blind deconvolution approach to high-resolution mapping of transcription factor binding sites from ChIP-seq data Genome Biology* 10.R142
- Moses et al. (2004) *MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model Genome Biology* 5:R98
- Narlikar et al. (2007) *Nucleosome Occupancy Information Improves de novo Motif Discovery RECOMB107-121*
- Narlikar et al. (2007) *A Nucleosome-Guided Map of Transcription Factor Binding Sites in Yeast. PLOS CompuBiol.* 3.11.e215
- Pepke et al. (2009) *Computation for ChIP-seq and RNA-seq studies Nature Methods* 6.11.522-
- Satija et al.(2008) *Combining statistical alignment and phylogenetic footprinting to detect regulatory elements. Bioinformatics.* 2008 May 15;24(10):1236-4
- Satija et al.(2009) *BigFoot: Bayesian alignment and phylogenetic footprinting with MCMC. BMC Evol Biol.* 2009 Aug 28;9:217
- Segal et al. (2006) *A genomic code for nucleosome positioning Nature* 442.1772-
- Tavazoie et al.(1999) *Systematic determination of genetic network architecture Nature Genetics* 22.281-
- Won et al. (2008) *Prediction of regulatory elements in mammalian genomes using chromatin signatures BMC Bioinf.*9.547-
- Won et al. (2009) *An Integrated Approach to Identifying Cis-Regulatory Modules in the Human Genome PLOS ONE* 4.5.e5501
- Zhang et al.(2008) *Networks motif-based identification of transcription factor gene relationships by integrating multi-source biological data BMC Bioinformatics* 9.203-220