Genome analysis

CCSeq: Clusters of Colocalized Sequences

Stefan Golas

Abstract

0.1 Motivation

Potential transcription factor (TF) complexes may be identified by testing whether the binding sequences of individual TF proteins form clusters with each other. These clusters may also indicate TF inhibition due to competitive occupancy of enhancer regions. Genome annotation data containing the coordinates of enhancer sequences is highly accessible via position-weight matrix tools.

0.2 Results

An algorithm called CCSeq (Clusters of Colocalized Sequences) was developed for identifying clusters of sequences along a one-dimensional line, such as a chromosome, given genome annotation files and a cut-off distance as inputs. The algorithm was applied to the binding sequences of the constituent proteins of two known transcription factor complexes, the HSF1 homotrimer and one form of the NF- κ B complex, a dimer of NFKB2 and RELB. 28 clusters of HSF1 trimer binding sequences were identified on chromosome Y, and 16 clusters of the NFKB2 and RELB dimer were identified on chromosome 17, compared to 0 clusters identified in any of the five simulated random distributions for each of the two sets of TF proteins. Additionally, structural patterns of these binding sequence clusters are described.

0.3 Availability and Implementation

This algorithm is freely available as an R package on the open source R repository CRAN at the following link: <https://cran.r-project.org/package=colocalized>. Genome annotation files were obtained from the PWMScan tool at <https://cgg.epfl.ch/pwmtools/pwmtools.php> hosted by the Swiss Institute of Bioinformatics (2) (3).

1 Introduction

Transcription factor (TF) proteins are a key element in gene regulatory circuits and other mechanisms for controlling gene expression (7), and many diseases are related to mutations in TFs. Many TFs bind with each other to form complexes, such as the heat shock factor 1 (HSF1) homotrimer (9) or the NF- κ B heterodimer (6). These and other complexes are

critical regulators of cell responses to stimuli, and have been implicated in a wide range of diseases including cancer. It is reasonable to assume that the different subunits of a TF complex will have DNA-binding sequences in close physical proximity to each other along a chromosome, and we can exploit this tendency to identify potential complexes using only sequence data. An algorithm CCSeq was created in which the user provides an input selection of sequence sets, where each set corresponds to the binding sequences of a particular TF (these binding sequences are also known as enhancers). The algorithm identifies clusters of DNA-binding sequences colocalized to each other within a given cut-off distance. The first step of the algorithm identifies pairs of colocalized binding sequences across sequence sets, each set corresponding to the binding sequences of a transcription factor in the input selection, and the locations of these sequence pairs are stored in colocalization matrices. Next, pairs of colocalized sequences are compared across multiple colocalization matrices to identify chains of paired sequences. Very simply, if we find that a sequence a pairs with sequence b and b pairs with c , the cluster abc is recorded. A cluster will be defined as a set of sequences where every member is colocalized to at least one other member of the cluster, two sequences are colocalized if they are within a given cut-off distance of each other. A number of papers have studied combinatorial binding in TFs, but these either rely on overlapping binding regions from individual ChIP-Seq experiments or otherwise take into account other measurements such as chromatin accessibility (8), whereas this algorithm relies solely on genome annotation files generated via position-weight matrices. The terms sequence and interval are used somewhat interchangeably in this paper. For clarity, a sequence is a substring along a chromosome whose interval is its starting and ending coordinates. Since this paper is concerned with the spacing and distribution of sequences rather than their contents, the interval will suffice to refer to the sequence and vice versa.

2 Algorithm

The input selection given to the algorithm consists of multiple sets of DNA-binding sequences, where each sequence set corresponding to a particular TF protein. A cut-off distance for colocalized pairs is also provided in the input. The two main steps in the algorithm are the creation of colocalization matrices for each pair of sequence sets, and the multiplication of these matrices in dimensionally sound sequences to locate clusters. The details

of these steps will be elaborated in section 2.2, while an illustration of a simple example is given in figure 2.2.

2.1 Notation

Upper-case letters will in general refer to matrices or sequence sets, and lower-case letters to integers. The colon ":" will index all the elements of a particular row or column of a matrix, e.g. $K_{:,z}$ will refer to all elements in the z th column in K . Certain variables i, j, k , and z are used to represent different values at different points throughout the pseudocode where they are used to index arbitrary elements in a vector or matrix. The meaning of each usage of these variables should be clear within each context.

2.2 Pseudocode

Algorithm 1 CCSeq

```

1: procedure CCSeq( $S, n$ )
2:    $\beta^{2 \times m} = P(|S|, 2)$ 
3:   for  $i$  in  $m$  do
4:      $A^i = \text{colocalized}(S_{\beta_{1,i}}, S_{\beta_{2,i}}, n)$ 
5:   end for
6:    $K = P(m, |S| - 1)$ 
7:    $K^* \leftarrow \emptyset$ 
8:   for  $p$  in  $\text{columns}(K)$  do
9:     if  $\beta_{K_{i-1,p},2} = \beta_{K_{i,p},1} \forall i \in K_{:,p}$  then
10:       $K^* \leftarrow \text{cbind}(K^*, K_{:,p})$ 
11:    end if
12:  end for
13:  for  $j$  in  $\text{columns}(K^*)$  do
14:     $Q^j \leftarrow \prod_{m=1}^{|K^*_{:,j}|} A^{K^*_{m,j}}$ 
15:  end for
16:  return  $Q$ 
17: end procedure

```

For an input selection S and a cut-off distance n , the function $\text{CCSeq}(S, n)$ proceeds according to the pseudocode. Line 2 generates all 2-element permutations on the index of S and stores these in the matrix β . Lines 3-5 iterates the function $\text{colocalized}(X, Y, n)$ over columns of $\beta^{2 \times m}$. Very simply, $\text{colocalized}(X, Y, n)$ forms a colocalization matrix such as has been already described from two sequence sets X and Y and a cut-off distance. For a colocalization matrix $M = \text{colocalized}(X, Y, n)$, a matrix entry $M_{i,j} = 1$ where the i th element of X and the j th element of Y are less than n base pairs from each other and 0 otherwise. In lines 6-12, a matrix K is formed containing every permutation over the index of A containing $|S| - 1$ elements with permutations stored as columns. These columns are filtered to form the matrix K^* , which will index elements of $A = A_i, A_j, \dots$ such that the sequence set represented by the column of some $A^{K_{i,p}}$ is the same as that represented by the row of $A^{K_{i+1,p}}$. This will ensure that taking the product of some sequence of colocalization matrices in A indexed by a column in K^* will be logical according to the formula for forming a cluster by chaining colocalization pairs with shared partners. If the columns of a colocalization matrix A represents the same sequence set as that represented by the rows of a matrix B , then $C = AB$ will have entries $C_{i,k} = A_{i,1}B_{1,k} + A_{i,2}B_{2,k} \dots + A_{i,j}B_{j,k}$ where the shared sequence set has length j . The formula for entries of the composite matrix C formed from A and B attempts to form clusters by finding where colocalized pairs from A and B share a sequence from the matching sequence set, and an entry $C_{i,k} = 1$ where $A_{i,j} = 1$ and $B_{j,k} = 1$. This can be considered in terms of Boolean logic, where $C_{i,k}$

is *true* if there exists a j where $A_{i,j}$ and $B_{j,k}$ are both *true*, and the summation to find $C_{i,k}$ checks every possible value of j . In this sense Boolean logic can entirely replace the standard arithmetic formula and the same result will be arrived at. Indeed, this fact is exploited in the implementation of this algorithm for significantly improved run-time. Lines 13-15 perform this matrix multiplication in series on colocalization matrices in A . The loop which filters K to K^* uses the functions $\text{columns}(K)$ and $\text{cbind}(K^*, K_{:,p})$. The function $\text{columns}(K)$ extracts from K the index over its columns, such that an element j from $\text{columns}(K)$ can be used to extract a column from K via $K_{:,p}$ where the ":" symbol implies every entry in the column p . The function $\text{cbind}()$ creates a matrix by binding the columns of two input matrices. It is here used exactly as it is used in the language R. After performing the sequence matrix multiplication in lines 13-15, matrices $Q^j \in Q$ will contain a positive entry $Q^j_{y,z} = 1$ where colocalization pairs had shared partners at every matrix in the matrix multiplication sequence. For example, if $Q = ABC$, then $Q_{y,z} = 1$ where $(AB)_{y,y'}C_{y',z} = 1$ for some y' and $A_{y,y'}B_{y',y''} = 1$ for some y' and the same y'' . Therefore, positive values in Q^j and its factors will index the locations of individual sequences belonging to clusters within their respective sequence sets.

An example of the algorithm is shown for illustration. Three sequence sets are depicted corresponding to hypothetical proteins **green**, **purple**, and **red**, with one sequence each for simplicity.

AGGATTTCAGTTTCAGCCGCAGATTTCGAGAGATGGCGTAAGTGGGATGAAGGCATAGAGGCT

1. An input selection and cut-off distance are specified. Sequences from each set in the input selection are highlighted in **green**, **purple**, and **red**, respectively. The cut-off distance is 20bp.

AGGATTTCAGTTTCAGCCGCAGATTTCGAGAGATGGCGTAAGTGGGATGAAGGCATAGAGGCT

2. Sequence-to-sequence distances are examined and the locations of colocalized sequence pairs are stored separately in matrices, where each matrix corresponds to a pair of sequence sets e.g. **green** and **purple**.

AGGATTTCAGTTTCAGCCGCAGATTTCGAGAGATGGCGTAAGTGGGATGAAGGCATAGAGGCT

3. The series of colocalization matrices is examined to determine shared colocalization partners between pairings. If a set of sequences exists where each sequence is colocalized to at least one other sequence in the set, that set is said to be a cluster. If a cluster contains a sequence from every sequence set in the input selection, that cluster satisfies the input selection and will be reported by the algorithm.

3 Implementation

Two known TF complexes, the HSF1 homotrimer and the NF-kB heterodimer (composed of NFKB2 and RELB), were tested as inputs to the algorithm on chromosome Y and chromosome 17, respectively, each with cut-off distances of 150bp. 16 clusters of NF-kB dimer sequences were detected on chromosome 17, while 28 clusters of HSF1 trimer binding sequences were detected on chromosome Y. Further data on these clusters are given in the supplementary data. Not only are clusters of binding sequences associated with the formation of complexes by proteins, the structure of the clusters in the two example sets is notably specific. In the case of the NF-kB complex, NFKB2 and RELB bind to complementary sequences on opposite strands, so the presence of one binding sequence almost always implies the presence of the other. For the HSF1 homotrimer, the three binding sequences are always on the same strand; 16/28 clusters are spaced according to the pattern [115, 30] and 5/28 are spaced according to the pattern [99, 69], with the remaining 7/28 not conforming to any regular spacing pattern. Spacing patterns are denoted according to [x,y..] where x is the distance between the lowest-numbered sequence and the 2nd-lowest-numbered, and so forth.

3.1 Simulating a Null Hypothesis

It is necessary to consider that these clusters may exist by chance and that their existence or frequency of occurrence has no significance as to whether the proteins of their constituent sequences form a binding complex. For a null hypothesis, we assume that individual binding sequences are distributed randomly across the specified chromosome. If there is a sequence set x in the input selection, then to test the null hypothesis a sequence set called x' is generated with the same number of sequences as x and whose intervals are all the same size as that of the average interval size in x , but which are distributed randomly throughout the length of the corresponding chromosome. A randomized sequence set is generated for every member of the input selection, and the algorithm is executed on the randomized selection with the same parameters (i.e. cut-off distance and chromosome). In order to reject the null hypothesis, the algorithm must detect consistently more clusters in the input selection than in the randomized sequence sets. In the examples of HSF1 and NF- κ B given above, no clusters were detected in randomized sequence sets across five randomized samples each. In this case the null hypothesis is rejected.

4 Discussion

Combinatorial regulation of gene expression has been a focus of intense study for years, in part due to its role as a built-in platform for cellular logic circuits and computation (18) (19) (20) (21) (22) (23). It has long been understood that these logic circuits can be exploited for metabolic engineering of industrially relevant cell lines (25) (26) (27) (28). Additionally, networks of interacting TFs play a key role in defining cancer cell lines (29). Clustered enhancer sequences within the genome provide crucial evidence for characterizing transcriptional circuitry and networks, and when combined with patterns of gene co-expression and protein-protein interaction data, may be used to definitively identify candidates for TF complexes. Although this paper has primarily focused on TF complexes as the motivation for identifying clustered binding sequences, transcriptional repression of one TF by another via competitive occupancy of enhancers is another potential explanation for clustering with experimental evidence (30). The structure of binding-sequence clusters may also be used to investigate patterns in the physical interaction of TF proteins with each other and the chromatin itself. The strength of this algorithm in investigating elements of transcriptional circuitry is in the high accessibility of genome annotation data given the availability of web tools such as PWMScan and JASPAR.

4.1 Comparison with other algorithms

Although other algorithms and studies have examined combinatorial TF regulation involving proximal binding sites (10) (11) (12) (13), this algorithm diverges from those in several ways. In general other algorithms investigating clusters of enhancers for different TFs have relied on overlaps between ChIP-Seq peaks, which recover accessible binding sites on chromatin for specific cell types and experimental conditions. This is because ChIP (chromatin immunoprecipitation) studies require that a physical binding event take place in order to detect a signal, which depends highly upon chromatin accessibility due to epigenetic regulation, which itself depends on complex external and internal signals (14) (15) (16) (17). This makes ChIP-Seq highly appropriate for investigating the interaction of environment, epigenetics, and cell fate, but it will fail to identify enhancer clusters in regions that are effectively "switched off" under the conditions of a given experiment. CCSeq relies exclusively on genome annotation files generated by the application of PWMs to whole genomes, and will identify clusters in locations which may not be accessible to TFs under a given set of experimental conditions, or any conditions whatsoever. From

this perspective, ChIP-Seq based algorithms should be used as part of a targeted approach under specific experimental conditions, while CCSeq offers a more exploratory approach that can be applied to genome annotation files generated using widely available web programs and open-source data. The algorithm will also report the exact locations of clustered binding sequences rather than regions with overlapping peaks.

5 Acknowledgements

Thank you to Ed Reitman and Jack Tuszynski for supporting this project through its duration with your excellent feedback and ideas. Thank you Philip Winter, Jack Tuszynski, and the University of Alberta for providing access to the Compute Canada research cluster which was invaluable for testing this algorithm and developing its parallel computing capabilities. Thank you to Renato Spacek for helping with the math very early in the project.

References

- [1]Dolcet, X., Llobet, D., Pallares, J., & Matias-Guiu, X. (2005). NF- κ B in development and progression of human cancer. *Virchows archiv*, 446(5), 475-482.
- [2]Ambrosini G., Groux R., and Bucher P. (2019) *PWMScan: a fast tool for scanning entire genomes with a position-specific weight matrix* Bioinformatics, bty127
- [3][Ambrosini G. (2018)] Ambrosini G. (2018), PWMTools, <http://ccg.vital-it.ch/pwmttools>
- [4][Kahn A et al.(2018)] Kahn A et al.(2018), JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework *Nucleic Acids Res.* 2018; 46:D260-D266. DOI:10.1093/nar/gkx1126 PMID:PMC5753243
- [5]Ivan V. Kulakovskiy; Ilya E. Vorontsov; Ivan S. Yevshin; Ruslan N. Sharipov; Alla D. Fedorova; Eugene I. Rumynskiy; Yulia A. Medvedeva; Arturo Magana-Mora; Vladimir B. Bajic; Dmitry A. Papatsenko; Fedor A. Kolpakov; Vsevolod J. Makeev, HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis *Nucl. Acids Res.*, Database issue, gkx1106 (11 November 2017) doi: 10.1093/nar/gkx1106
- [6]Lawrence T. (2009), The nuclear factor NF- κ B pathway in inflammation. *Cold Spring Harbor perspectives in biology*, 1(6), a001651. doi:10.1101/cshperspect.a001651
- [7][Xu, L., Glass, C. K., & Rosenfeld, M. G. (1999)] Xu, L., Glass, C. K., & Rosenfeld, M. G. (1999), Coactivator and corepressor complexes in nuclear receptor function. *Current opinion in genetics and development*, 9(2), 140-147.
- [8]Majid Kazemian, Hannah Pham, Scot A. Wolfe, Michael H. Brodsky, Saurabh Sinha, Widespread evidence of cooperative DNA binding by transcription factors in *Drosophila* development, *Nucleic Acids Research*, Volume 41, Issue 17, 1 September 2013, Pages 8237–8252, <https://doi.org/10.1093/nar/gkt598>
- [9]Repression of heat shock transcription factor HSF1 activation by HSP90 (HSP90 complex) that forms a stress-sensitive complex with HSF1. *Cell*, 94(4), 471-480.
- [10]Tom Whittington, Martin C. Frith, James Johnson, Timothy L. Bailey (2011), Inferring transcription factor complexes from ChIP-seq data, *Nucleic Acids Research*, Volume 39, Issue 15, 1 August 2011, Page e98, <https://doi.org/10.1093/nar/gkr341>
- [11]Teng, L., He, B., Gao, P., Gao, L., & Tan, K. (2013), Discover context-specific combinatorial transcription factor interactions by integrating diverse ChIP-Seq data sets. *Nucleic acids research*, 42(4), e24-e24.

- [12]Ka-Chun Wong, Yue Li, Chengbin Peng, Zhaolei Zhang (2015), SignalSpider: probabilistic pattern discovery on multiple normalized ChIP-Seq signal profiles, *Bioinformatics*, Volume 31, Issue 1, 1 January 2015, Pages 17–24, <https://doi.org/10.1093/bioinformatics/btu604>
- [13]Zinzen, R. P., Girardot, C., Gagneur, J., Braun, M., & Furlong, E. E. (2009), Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature*, 462(7269), 65.
- [14]Epigenetics: Beyond Chromatin Modifications and Complex Genetic Regulation Steven R. Eichten, Robert J. Schmitz, Nathan M. Springer *Plant Physiology* Jul 2014, 165 (3) 933-947; DOI: 10.1104/pp.113.234211
- [15]Meaney, M. J., and Szyf, M. (2005). Environmental programming of stress responses through DNA methylation: life at the interface between a dynamic environment and a fixed genome. *Dialogues in clinical neuroscience*, 7(2), 103.
- [16]Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., ... and Garg, K. (2012). The accessible chromatin landscape of the human genome. *Nature*, 489(7414), 75.
- [17]John, S., Sabo, P. J., Thurman, R. E., Sung, M. H., Biddie, S. C., Johnson, T. A., ... and Stamatoyannopoulos, J. A. (2011). Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nature genetics*, 43(3), 264.
- [18]Benfey, P. N., & Chua, N. H. (1990). The cauliflower mosaic virus 35S promoter: combinatorial regulation of transcription in plants. *Science*, 250(4983), 959-966.
- [19]Balaji, S., Babu, M. M., Iyer, L. M., Luscombe, N. M., Aravind, L. (2006). Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. *Journal of molecular biology*, 360(1), 213-227.
- [20]Moon, T. S., Lou, C., Tamsir, A., Stanton, B. C., & Voigt, C. A. (2012). Genetic programs constructed from layered logic gates in single cells. *Nature*, 491(7423), 249.
- [21]Buchler, N. E., Gerland, U., & Hwa, T. (2003). On schemes of combinatorial transcription logic. *Proceedings of the National Academy of Sciences*, 100(9), 5136-5141.
- [22]Hermesen, R., Tans, S., & Ten Wolde, P. R. (2006). Transcriptional regulation by competing transcription factor modules. *PLoS computational biology*, 2(12), e164.
- [23]Kramer, B. P., Fischer, C., & Fussenegger, M. (2004). BioLogic gates enable logical transcription control in mammalian cells. *Biotechnology and bioengineering*, 87(4), 478-484.
- [24]Weiss, R., Basu, S., Hooshangi, S., Kalmbach, A., Karig, D., Mehreja, R., Netravali, I. (2003). Genetic circuit building blocks for cellular computation, communications, and signal processing. *Natural Computing*, 2(1), 47-84.
- [25]Hoynes-O'Connor, A., & Moon, T. S. (2015). Programmable genetic circuits for pathway engineering. *Current opinion in biotechnology*, 36, 115-121.
- [26]Jensen, M. K., & Keasling, J. D. (2015). Recent applications of synthetic biology tools for yeast metabolic engineering. *FEMS Yeast Res*, 15(1), 1-10.
- [27]Piscitelli, A., Giardina, P., Lettera, V., Pezzella, C., Sannia, G., & Faraco, V. (2011). Induction and transcriptional regulation of laccases in fungi. *Current Genomics*, 12(2), 104-112.
- [28]Mach, R., & Zeilinger, S. (2003). Regulation of gene expression in industrial fungi: *Trichoderma*. *Applied microbiology and biotechnology*, 60(5), 515-522.
- [29]Wang, Q., Li, W., Liu, X. S., Carroll, J. S., Jänne, O. A., Keeton, E. K., ... & Brown, M. (2007). A hierarchical network of transcription factors governs androgen receptor-dependent prostate cancer growth. *Molecular cell*, 27(3), 380-392.
- [30]Mailly, F., Berube, G., Harada, R., Mao, P. L., Phillips, S., & Nepveu, A. (1996). The human cut homeodomain protein can repress gene expression by two distinct mechanisms: active repression and competition for binding site occupancy. *Molecular and Cellular Biology*, 16(10), 5346-5357.