Cluster analysis on shape indices for ChIP-Seq data

Classificazione non supervisionata su indici di forma per dati ChIP-Seq

Marzia A. Cremona, Pier Giuseppe Pelicci, Laura Riva, Laura M. Sangalli, Piercesare Secchi, and Simone Vantini

Abstract In recent years many techniques have been developed to study genetic and epigenetic processes. Here we focus on a particular Next Generation Sequencing method called ChIP-Seq (Chromatin ImmunoPrecipitation Sequencing), that permits to investigate protein-DNA interactions. At present, in the relevant literature, the analysis of ChIP-Seq data is mainly restricted to the detection of enriched regions (peaks) in the genome, considering only signal intensity. The innovative approach that we propose takes into consideration also the shape of such peaks. We introduce some indices to summarize the shape and we use multivariate clustering techniques in order to detect statistically significant differences in peak shape, with the idea that it can be associated with a functional role and a biological meaning.
Abstract Negli ultimi anni sono state sviluppate molte tecniche per studiare i processi genetici ed epigenetici. In questo lavoro concentriamo la nostra attenzione su un particolare metodo di sequenziamento di nuova generazione chiamato ChIP-Seq (Chromatin ImmunoPrecipitation Sequencing), che permette di indagare le interazioni tra proteine e DNA. Al momento, in letteratura, l'analisi di dati ChIP-Seq è quasi totalmente ristretta alla localizzazione di zone arricchite (picchi) sul genoma e considera solo l'intensità del segnale. L'approccio innovativo che proponiamo tiene in considerazione anche la forma dei suddetti picchi. Introduciamo alcuni indici per sintetizzare la forma e usiamo tecniche di classificazione non supervisionata multi-

Marzia A. Cremona, Laura M. Sangalli, Piercesare Secchi and Simone Vantini
MOX - Dipartimento di Matematica, Politecnico di Milano, Milano (Italy)
e-mail: marziaangela.cremona@polimi.it; laura.sangalli@polimi.it; piercesare.secchi@polimi.it; simone.vantini@polimi.it

Pier Giuseppe Pelicci
Department of Experimental Oncology, European Institute of Oncology, Milano (Italy)
Dipartimento di Scienze della salute, Università degli Studi di Milano, Milano (Italy)
e-mail: piergiuseppe.pelicci@ieo.eu

Laura Riva
Center for Genomic Science of IIT@SEMM, Istituto Italiano di Tecnologia, Milano (Italy)
e-mail: laura.riva@iit.it

variata allo scopo di trovare differenze statisticamente significative nella forma dei picchi, con l'idea che questa possa essere associata a una funzione e a un significato biologico.

Key words: genomics, ChIP-Seq, cluster analysis, multivariate analysis

1 Chromatin ImmunoPrecipitation Sequencing (ChIP-Seq)

Next Generation Sequencing (NGS) techniques have been developed during the past few years, enabling scientists to directly study genetic and epigenetic processes. Here, we focus on a particular NGS method that allows to investigate protein-DNA interactions, namely the ChIP-Seq (Chromatin ImmunoPrecipitation Sequencing).

ChIP-Seq is widely used to detect regions in the genome in which DNA interacts with proteins, usually transcription factors and histones (see [6] for a review of the method). After cross-linking of protein-DNA in vivo and shearing of DNA molecules by sonication, a specific antibody permits to select and immunoprecipitate only the fragments with the protein of interest attached. Purified DNA is then sequenced, obtaining as a result a short read (single-end sequencing) or two short reads (paired-end sequencing) for each fragment.

Mapping reads back on a reference genome is the first pre-processing step that must be done in order to analyze ChIP-Seq data. Usually only high quality reads that maps uniquely to the genome are kept for further analysis. Then a peak caller is run, with the aim of finding regions in the genome with a high density of reads with respect to a control signal, defining the peaks, i.e. the areas where the protein of interest is attached to the genome. There exist several different peak callers, each using a specific technique to detect enriched regions, looking in a slightly different way at the intensity of signals. In this step the average length of the initial fragments must be estimated from reads positions in the two DNA strands.

Downstream computational analyses include comparison of different ChIP-Seq or study of different peaks in the same ChIP-Seq, essentially involving the location of peaks, the distance from annotated points (i.e. transcription start sites) or the total number of reads corresponding to each peak. We propose a novel technique that takes into consideration not only the intensity but also the shape of these peaks in ChIP-Seq analysis.

2 Multivariate analysis on shape indices

We start our analysis creating, for each peak found by the peak caller, the base by base count of the aligned fragments (elongated reads). We would like to use cluster analysis to evaluate whether peaks can be divided into groups, according to the shape of the step function that defines them. The simplest way to deal with

this problem is to select some shape indices, analyzing them in the framework of multivariate statistical analysis.

We use five indices to summarize the shape, the first two related to the intensity of the signal, and the others connected with the complexity of the peak:

1. The maximum height of the peak;
2. The area under the step function that defines the peak;
3. The full width at half maximum, that is the width of the peak at half of its maximum height;
4. The number of local peaks;
5. The shape index M, divided by the maximum height of the peak.

To compute the number of local peaks we need first at all to smooth the peak, in order to filter out noise. The index is then calculated as the local maxima of the resulting smoothed peak. We perform this step using a cubic B-splines basis on each peak (see, for example, [7]). The shape index M is a measure of the complexity of the peak that is robust to noise, computed as suggested in [1] and [3]: we build a tree starting from the peak, creating a new vertex in correspondence to an increase in the step function and moving toward the root when the step function goes down, and we calculate the index M as the number of edges in a maximal matching for the tree, using the blossom algorithm.

Once we have these indices, we can apply multivariate clustering techniques, such as k-means, k-medoids or hierarchical agglomerative clustering (see, for instance, [4] and [2]). Through total-within-sum-of-squares plot or dendrogram, we can estimate the "correct" number of clusters. Each cluster will be characterized by a specific distribution of the five indices, characterizing its shape.

3 Results: analysis of a ChIP-Seq for GATA1

In this section we show the application of the multivariate clustering technique presented in Section 2 to a ChIP-Seq for the transcription factor GATA1 in human cells (cell type K562) generated by the ENCODE Consortium (see [8]). Our analysis starts from reads already mapped on the genome. First, we run the peak caller MACS (Model-based Analysis of ChIP-Seq, see [9]) on the ChIP-Seq, with a more stringent threshold than the default one, in order to find the enriched regions of the genome and to estimate the average fragments length. Around 13 thousand peaks are found. For each peak, we calculate the base by base count, and hence we compute the five shape indices. We show here the results obtained by applying the k-mean algorithm on the standardized indices.

Looking at the plot of the total-within-clusters-sum-of-squares in Figure 1, $k = 3$ seems a reasonable compromise for the number of clusters. The three resulting clusters are shown in Figure 2. For visualization reasons, only 200 peaks are plotted, aligned according to their maximum height. Cluster 1 is the biggest of the three clusters found (about the 75% of the peaks are allocated to it), and contains uni-

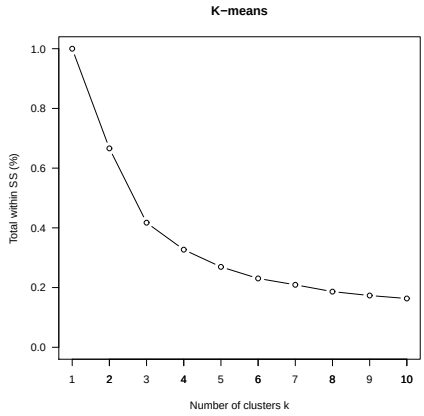

Fig. 1 Plot of the total within clusters sum of squares in multivariate k-means, for k=1,...,10

modal and not very high peaks. The second cluster is that of high, bell-shaped peaks; these are also the peaks selected with a high significance by the peak caller MACS (high score). Lastly, Cluster 3 contains multimodal and wider peaks. MACS search for highly localized peaks, so it should not find peaks of Cluster 3; interestingly, these peaks are not the ones at which MACS gives low score (less significant peaks according to MACS): if we reduce the threshold in peak detection, we keep on selecting these peaks.

To validate the clusters, we searched for recurrent motifs using the tool MEME-ChIP (see [5]). Motifs are short sequences of nucleotides conserved among peaks, that are specific to the proteins that bind the DNA. Figure 3 shows the results of such analysis on samples of size 100. With random peaks from the whole dataset, MEME-ChIP finds the motif GATA that is known to be associated to the protein GATA1. Selecting peaks from Cluster 1 or 2 the tool finds the same motif, but in the case of Cluster 2 we have a more significant E-value (the E-value is an estimate of the expected number of motifs having the same width and nucleotide frequency sequence, that one would find in a similarly sized set of random sequences; see [5]). Finally, using peaks from cluster 3 we do not get any specific motif. This analysis supports the existence of a possible relationship between peak shape and biological meaning.

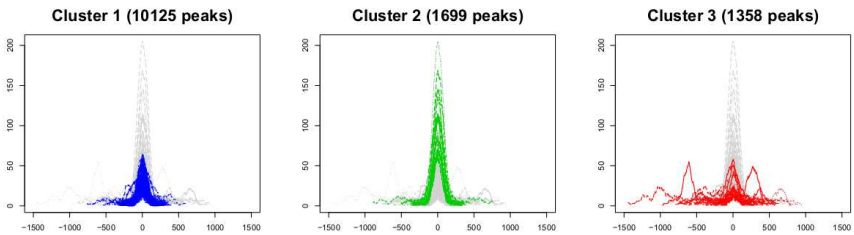

Fig. 2 Results of multivariate k-means with $k = 3$

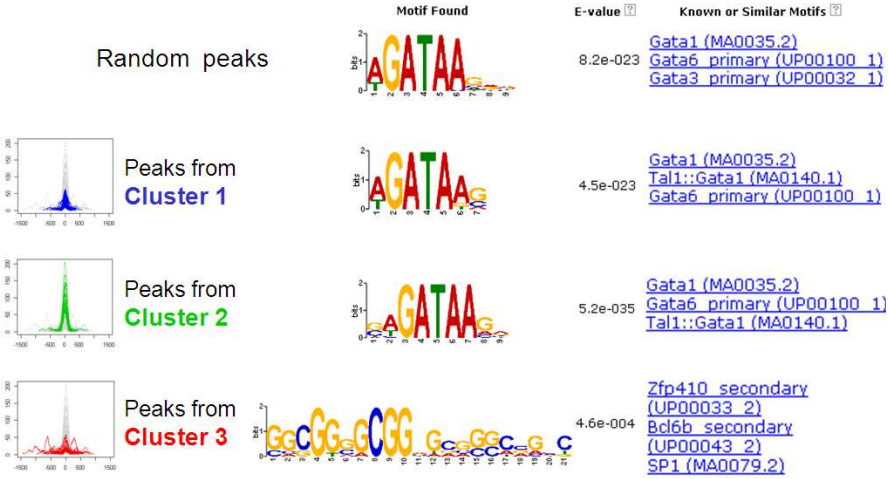

Fig. 3 Motif finding with MEME-ChIP

There are two replicates of ChIP-Seq for GATA1 produced by the ENCODE Consortium, so we can take advantage of this to examine the robustness of our method. Running the peak caller MACS on this second replica (with the same threshold) we find less than 5 thousand peaks, almost all overlapping some peak of the first replica. It is important to notice that these don't correspond only to regions selected with high significance in the first replicate. Moreover, the shape of the step functions in the same region of the genome looks more or less the same in both replicates. Despite the highly different efficiency of the two ChIP-Seq, our method is able to find the same three clusters in both datasets: more than 85% of peak pairs fall in the correspondent clusters in the two replicates. Motif finding with MEME-ChIP gives the same results too, suggesting that our technique is robust.

We also applied k-medoid algorithm and hierarchical agglomerative clustering in place of k-means, even if the latter is not very convenient when dealing with large datasets, due to its (at least) quadratic complexity. These different clustering techniques led to the same optimal number of clusters, $k = 3$. Moreover, the resulting clusters were rather comparable to the ones obtained with k-mean algorithm, with the difference that Cluster 1 identified by these techniques is smaller than Cluster 1 identified by the k-mean algorithm (part of these peaks being allocated instead to Clusters 2 and 3).

4 Acknowledgements

This work is part of the "Genomic Computing Project" in co-operation between European Institute of Oncology, Istituto Italiano di Tecnologia and Politecnico di Milano. We wish to thank Gaetano Ivan Dellino, Lucilla Luzi and Alice Parodi for

their collaboration to the project. We are grateful to Giancarlo Ferrari for his help in the implementation of these analysis.

References

1. Evans, S.N., Hower, V., Patcher, L.: Coverage statistics for sequence census methods. BMC Bioinformatics **11**(430) (2010)
2. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer (2009)
3. Hower, V., Evans, S.N., Patcher, L.: Shape-based peak identification for ChIP-Seq. BMC Bioinformatics **12**(15) (2011)
4. Johnson, D.W. Wichern, D.W.: Applied Multivariate Statistical Analysis. Pearson Prentice Hall (1992)
5. Machanick, P., Bailey, T.L.: MEME-ChIP: motif analysis of large DNA datasets. Bioinformatics **27**(12), 1696-1697 (2011)
6. Park, P.J.: ChIP-Seq: advantages and challenges of a maturing technology. Nat. Rev. Genet. **10**(10), 669-680 (2009)
7. Ramsay, J. O., Silverman, B. W.: Functional Data Analysis. Springer (2005)
8. The ENCODE Project Consortium: An integrated encyclopedia of DNA elements in the human genome. Nature **489**, 57-74 (2012)
9. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., Liu, X.S.: Model-based Analysis of ChIP-Seq (MACS). Genome Biol. **9**(9), R137 (2008)