RSAT peak-motifs: Efficient prediction of transcription factor motifs and binding sites from genome-wide sequencing peak sets

Morgane Thomas-Chollier[1], Matthieu Defrance[2], Olivier Sand[3], Carl Herrman[4], Denis Thieffry[1], van Helden Jacques[3] ✉

[1]Institut de Biologie de l'Ecole Normale Supérieure, Paris, France
[2]Centro de Ciencias Genomicas, Universidad Nacional Autónoma de México, Cuernavaca, Mexico
[3]CNRS-UMR8199 Institut de Biologie de Lille, Lille, France
[4]INSERM U928 & Université de la Méditerranée, Marseille, France

Motivation and objectives

ChIP-seq is increasingly used to characterize transcription factor binding and chromatin marks at a genomic scale. Although various programs have been developed to perform read mapping and peak calling, the subsequent steps have not yet reached proper maturation: identifying relevant transcription factor binding motifs and the precise location of their binding sites remains a bottleneck. Most existing tools present limitations on sequence size, and they typically restrict motif discovery to a few hundred peaks, or to the central-most part of the peaks. To interpret genome-wide location data, there is a crucial need for time- and memory-efficient algorithms, interfaced as user-accessible tools to extract relevant information from high-throughput sequencing data.

For this purpose, we developed the software tool *peak-motifs* (Thomas-Chollier *et al.*, 2012a), which takes as input a set of peak sequences of interest, discovers key motifs, compares them with transcription factor binding motifs from various databases, predicts the location of binding sites within the peaks and exports them in a format suitable for visualization in the UCSC Genome Browser. Notably, all these steps, including motif discovery, are performed on the full-size sets of peak sequences, without restrictions on peak number or width.

Methods

The motif discovery step relies on a combination of algorithms integrated in the software suite regulatory sequence analysis tools (RSAT, http://rsat.ulb.ac.be/rsat/) (Thomas-Chollier *et al.*, 2011), which use complementary criteria to detect exceptional words (oligonucleotides and spaced motifs): global over-representation of oligonucleotides (*oligo-analysis*) or spaced pairs (*dyad-analysis*), heterogeneous positional distribution (*position-analysis*) and local over-representation (*local-word-analysis*).

The motif comparison step is performed by *compare-matrices* (Thomas-Chollier *et al.*, 2011), which supports a wide range of scoring metrics and displays the results as multiple alignments of logos, enabling to grasp the similarities between a discovered motif and several known motifs. This feature is particularly valuable to reveal adjacent fragments of the discovered motif showing similarities with two distinct known motifs, suggesting a bipartite motif for two factors.

Sequences are scanned with the discovered motifs to locate binding sites, and their positioning within peaks is analyzed (coverage, positional distribution along peaks).

Peak-motifs generates an HTML report summarizing the main results and giving access to each separate result file. The report page includes links, allowing users to upload input peaks and predicted sites to the UCSC Genome Browser in order to visualize them in their genomic context.

Results and discussion

We assessed peak-motifs performances on several published datasets. In all cases, relevant motifs are disclosed.

For example, we discovered individual Oct and Sox motifs in Sox2 and Oct4 peak collections, whereas the original study only found the composite Sox/Oct motif (Chen *et al.*, 2010; Thomas-Chollier *et al.*, 2012a).

Similarly, for ChIP-seq data targeting the generic transcriptional co-activator p300, peak-motifs identified motifs bound by tissue-specific transcription factors consistent with these two tissues (Visel *et al.*, 2009; Blow *et al.*, 2010; Thomas-Chollier *et al.*, 2012a).

We assessed the time efficiency of *peak-motifs* by analyzing data sets of increasing sizes (from 100 to 1 000 000 peaks of 100 bp each), with total sequence sizes ranging from 10 kb to 100 Mb. The computing time of the motif discovery algorithms integrated in peak-motifs increases linearly with sequence size and outperforms all the other existing motif discovery tools used in our comparison (Thomas-Chollier *et al.*, 2012a). Data sets of several tens of megabytes are processed in a few minutes on a personal computer (the most efficient tool, *oligo-analysis*, treats 100Mb in 3min). This linear time response enables peak-motifs to scale up efficiently with sequence size, and allows us to provide an easy access via a web interface, without any data size restriction. This moreover gives us the possibility to run four distinct algorithms in order to detect motifs of various types (oligonucleotides, spaced pairs) based on complementary criteria (over-representation, positional heterogeneity).

In conclusion, peak-motifs supports time-efficient and statistically reliable analysis of complete ChIP-seq datasets, while offering an online user-friendly and well-documented interface, as well as a detailed protocol (Thomas-Chollier *et al.*, 2012b)

Acknowledgements

M.T-C is supported by the Alexander von Humboldt foundation.

References

Blow MJ, McCulley DJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F et al. (2010) ChIP-Seq identification of weakly conserved heart enhancers. *Nat Genet* **42**, 806-10.

Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J et al. (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133**, 1106-17.

Thomas-Chollier M, Defrance M, Medina-Rivera A, Sand O, Herrmann C, Thieffry D, van Helden J (2011). RSAT 2011: Regulatory Sequence Analysis Tools. *Nucleic Acid Research* **39**: W86-91.

Thomas-Chollier M, Herrmann C, Defrance M, Sand O, Thieffry D, van Helden J (2012a). RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic Acid Research* **40**: e31.

Thomas-Chollier M, Darbo E, Herrmann C, Defrance M, Thieffry D, van Helden J (2012b). From peaks to motifs: a complete workflow for full-sized ChIP-seq (and similar) datasets. *Nature Protocols* **7**: 1551-68.

Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F et al. (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854-8.