Deciphering genome-wide cis-regulation with RSAT: application to the glucocorticoid receptor

Morgane THOMAS-CHOLLIER[1], Matthieu DEFRANCE[2], Alejandra MEDINA-RIVERA[3], Olivier SAND[4], Pierre VINCENS[1], Carl HERRMAN[5], Sebastiaan H. MEIJSING[6], Denis THIEFFRY[1] and Jacques VAN HELDEN[5]

[1] Computational systems biology, Institute of Biology of ENS (IBENS), Paris, France
mthomas@biologie.ens.fr , thieffry@ens.fr , Pierre.Vincens@ens.fr

[2] Laboratory of Cancer Epigenetics, Faculty of Medicine, Université Libre de Bruxelles, Belgium
defrance@bigre.ulb.ac.be

[3] SickKids Research Institute, 101 College St. East Tower | Suite 15-306, Toronto, Ontario, Canada
alejandra.medina@sickkids.ca

[4] Génomique et maladies métaboliques, CNRS-UMR8199, Institut de Biologie de Lille, France
olivier.sand@good.ibl.fr

[5] Technological Advances for Genomics and Clinics (TAGC), INSERM U928 & Aix-Marseilles University, France
jacques.VAN-HELDEN@univ-amu.fr , carl.herrmann@univ-amu.fr

[6] Max Planck Institute for Molecular Genetics, Berlin, Germany
meijsing@molgen.mpg.de

Overview of RSAT

The regulatory sequence analysis tools (RSAT, http://rsat.ulb.ac.be/rsat/ and mirrors) is a software suite that integrates a large collection of modular tools for the detection of cis-regulatory elements in genomic sequences [1-3]. The web site has been running without interruption since 1998, and the suite has been continuously developed to accommodate novel types of data and experimental approaches over the years [2-4]. The suite includes programs for sequence retrieval, motif discovery, phylogenetic footprint detection, sequence scanning with regular expressions or position-specific scoring matrices, motif quality assessment and comparison, visualization and conversion utilities, along with a series of tools for random model generation and statistical evaluation. Genomes are regularly updated from various genome repositories (NCBI, Ensembl, UCSC browser) and the website currently supports 2517 genomes (March 2013).

RSAT enables genome-wide analysis of cis-regulatory elements with different types of input data: (i) groups of co-expressed genes produced by transcriptomic experiments, (ii) phylogenetically conserved regions and (iii) high-throughput binding data such as ChIP-seq.

In addition to motif discovery and pattern-matching approaches, the suite already provides various tools to analyse transcription factor binding motifs represented as matrices, including motif comparisons [3] and evaluation of matrix quality [5]. We are currently developing a motif clustering algorithm to ease the analysis of overlapping motifs (newly discovered or reported in databases) and assess potential motif diversity for a given transcription factor, in the context of motifs for cofactors.

The RSAT web server offers an intuitive interface, where each program can be accessed either separately or connected to the other tools. In addition, many tools are available as SOAP/WSDL web services, enabling their integration in programmatic workflows. Programs are documented with manual pages, while 'demo' buttons propose typical test cases. In addition, web tutorials and a series of published protocols help the users to master the different functionalities of RSAT [6-9], providing step-by-step guidelines about alternative options, as well as regarding the interpretation of results.

Cis-regulation analysis from high-throughput binding data

Several efficient and complementary motif discovery algorithms can predict transcription factor binding motifs from groups of co-expressed genes. Although these methods yield good results in yeast and bacteria

genomes, they are not suitable for vertebrates, due to the larger size and heterogeneity of non coding genomic sequences. The same algorithms nevertheless proved very efficient to analyse high-throughput transcription factor binding data, where the signal to noise ratio is higher. The workflow *peak-motifs* [10] was therefore developed to process large collections of peak sequences obtained from ChIP-seq or related technologies, to predict transcription factor binding motifs.

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.

Our workflow *peak-motifs* 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.

The motif discovery step relies on a combination of algorithms that 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* [3], 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.

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 [10]. 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.

Current developments aim at extending *peak-motifs* to support other high-throughput data, including chromatin marks or DNaseI.

Condition-specific binding of GR

We extensively used RSAT to study the binding of the glucocorticoid receptor (GR), with the aim to better understand its specificity of action in various cell types. Glucocorticoids are steroid hormones that bind to its nuclear receptor (GR) that in turn recognizes specific DNA sequences to regulate the expression of target genes. The target genes of GR are highly cell-type specific and they mediate the various physiological effects of GR on processes including glucose metabolism (hepatocytes), anti-inflamatory effects (leucocytes) and increase of bone resorption, [11]. In particular, the anti-inflammatory effect of glucocorticoids are of enormous therapeutic importance and are widely used to treat a broad range of immune-related diseases e.g. allergies, asthma and arthritis. Unfortunately however, the beneficial therapeutical effects of glucocorticoids are accompanied by several side effects including muscle wasting,

metabolic changes and osteoporosis. The desired as well as the side effects of glucocorticoids are mediated by specific target genes that are regulated in a tissue specific manner. To better understand how glucocorticoids elicit highly tissue specific effects we used ChIP-seq to compare the genomic binding profile of GR in cell lines derived from B cells, bone and lung.

We made several observations when we compared the genomic binding profile of GR in these cell lines: first, there is little overlap of the bound regions, suggesting that genomic binding is highly cell-type specific. Moreover, the fraction of promoter-proximal GR binding differs greatly between cell-types. Our motif analysis suggests that GR recognizes similar binding sites (imperfect palindromes), but that the presence of sequence motifs for potential cofactors varies between cell lines. Preliminary experimental validation showed that genomic regions harbouring such cell-type specific cofactors could recapitulate cell-type specific regulation by GR arguing for their importance in directing cell-type specific GR actions.

Furthermore, we compared the binding specificity of two isoforms of GR that differ by a single amino acid which is a consequence of an alternative splice event. We found that the binding motif of these isoforms, GRalpha and GRgamma, differed slightly and functional luciferase reporter studies confirmed the importance of these differences in mediating isoform-specific transcriptional regulation. We do not know the exact reason for the difference between the binding motifs for GR alpha and gamma, but biochemical and structural studies indicated a role for isoform specific DNA binding affinity and conformation.

Current work on the GR involves the analysis of ChIP-exo datasets, with the goal to obtain more precise information on the mechanisms by which GR recognizes its binding sites.

References

[1] van Helden J. Regulatory sequence analysis tools. *Nucleic Acids Research* 31: 3593-6, 2003.

[2] Thomas-Chollier M, Sand O, Turatsinze JV, Janky R, Defrance M, Vervisch E., Brohee S & van Helden J. RSAT: regulatory sequence analysis tools. Nucleic Acids Res*earch* 36: W119-27, 2008.

[3] Thomas-Chollier M, Defrance M, Medina-Rivera A, Sand O, Herrmann C, Thieffry D, van Helden J. RSAT 2011: regulatory sequence analysis tools. *Nucleic Acids Research* 39: W86-91, 2011.

[4] Defrance M, & van Helden J. Info-gibbs: a motif discovery algorithm that directly optimizes information content during sampling. *Bioinformatics* 25: 2715-22, 2009.

[5] Medina-Rivera A, Abreu-Goodger C, Salgado-Osorio H, Collado-Vides J & van Helden J. Empirical and theoretical evaluation of transcription factor binding motifs. *Nucleic Acids Research* 39: 808–24, 2011.

[6] Turatsinze JV, Thomas-Chollier M, Defrance M & van Helden J. Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. *Nature Protocols* 3 : 1578-88, 2008.

[7] Defrance M, Janky R, Sand O.& van Helden J. Using RSAT oligo-analysis and dyad-analysis tools to discover regulatory signals in nucleic sequences. *Nature Protocols* 3 : 1589-603, 2008.

[8] Sand O, Thomas-Chollier M, Vervisch E & van Helden J. Analyzing multiple data sets by interconnecting RSAT programs via SOAP Web services-an example with ChIP-chip data. *Nature Protocols* 3 : 1604-15, 2008.

[9] Thomas-Chollier M, Darbo E, Herrmann C, Defrance M, Thieffry D & van Helden J. A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs. *Nature Protocols* 7: 1551-68, 2012.

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

[11] Gross KL, Cidlowski JA: Tissue-specific glucocorticoid action: a family affair. Trends in endocrinology and metabolism: TEM 2008, 19(9):331-339.