A general approach for discriminative *de novo* motif discovery from high-throughput data

Jan Grau [1], Stefan Posch [1], Ivo Grosse [1], and Jens Keilwagen [2,3]

[1]Institute of Computer Science, Martin Luther University Halle–Wittenberg, Halle (Saale), Germany
[2]Julius Kühn-Institut (JKI) - Federal Research Centre for Cultivated Plants, Quedlinburg, Germany
[3]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Seeland OT Gatersleben, Germany
{grau|posch|grosse}@informatik.uni-halle.de
jens.keilwagen@jki.bund.de

Abstract: High-throughput techniques like ChIP-seq, ChIP-exo, and protein binding microarrays (PBMs) demand for novel *de novo* motif discovery approaches that focus on accuracy and runtime on large data sets. While specialized algorithms have been designed for discovering motifs in *in-vivo* ChIP-seq/ChIP-exo *or* in *in-vitro* PBM data, none of these works equally well for all these high-throughput techniques. Here, we present Dimont, a general approach for fast and accurate de-novo motif discovery from high-throughput data, which achieves a competitive performance on both ChIP-seq and PBM data compared to recent approaches specifically designed for either technique. Hence, Dimont allows for investigating differences between *in-vitro* and *in-vivo* binding in an unbiased manner using a unified approach. For most transcription factors, Dimont discovers similar motifs from *in-vivo* and *in-vitro* data, but we also find notable exceptions. Scrutinizing the benefit of modeling dependencies between binding site positions, we find that more complex motif models often increase prediction performance and, hence, are a worthwhile field of research.

Original paper: doi: 10.1093/nar/gkt831
http://nar.oxfordjournals.org/content/41/21/e197

1 Introduction

Transcription factors are a major component of gene regulation as they bind to specific binding sites in promoters of genes and subsequently activate or repress gene expression. The *de novo* discovery of transcription factor binding motifs and binding sites from data obtained by wet-lab experiments is still a challenging problem in bioinformatics, and has not been fully solved yet.

Today, two prevalent sources of experimental data are chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq and ChIP-exo) and protein binding microarrays (PBMs). Chromatin IP experiments yield approximate genomic regions bound

by transcription factors *in-vivo*, where each of the bound regions is associated with a measure of transcription factors abundance (often termed *peak statistics*) at that specific region. PBM experiments yield information about binding affinity (measured as *probe intensities*) of transcription factors to an unbiased collection of probe sequences *in-vitro*.

2 Approach

We present Dimont [GPGK13], a *de novo* motif discovery approach especially tailored to data from these high-throughput techniques. In contrast to previous approaches, Dimont uses a weighted discriminative learning principle for learning model parameters, which exploits the measures of confidence obtained experimentally, namely peak statistics or probe intensities. This learning principle relies on a common rank-based weighting schema that allows for integrating these different measures of confidence. Strategies for runtime optimization implemented in Dimont result in runtimes of approximately 10 minutes on an average-sized ChIP-seq data set.

3 Key findings

Dimont successfully discovers all motifs of the ChIP-seq data sets of Ma *et al.* [M$^+$12]. On the data sets of Weirauch *et al.* [W$^+$13], it predicts PBM intensities from probe sequence with higher accuracy than any of the approaches specifically designed for that purpose. Dimont also reports the expected motifs for several ChIP-exo data sets. Hence, Dimont is the first approach that yields a competitive performance on both *in-vitro* and *in-vivo* data using a unified approach.

This allows us to investigate differences between *in-vitro* and *in-vivo* binding in an unbiased manner. We find that for most transcription factors, the motifs discovered by Dimont are in good accordance between techniques. However, we also find notable exceptions, where the motifs obtained from ChIP and PBM experiments for the same transcription factor differ substantially, although both yield accurate predictions on data sets obtained by the corresponding technique.

We use the common framework of Dimont to additionally study the impact of motif models incorporating dependencies between adjacent positions (inhomogeneous Markov models of higher order) compared to standard position weight matrices on prediction accuracy. We find that modeling adjacent dependencies indeed improves prediction accuracy for several transcription factors for both, *in-vitro* and *in-vivo* data. Notably, this improvement often persists for predictions across techniques, i.e., for learning a model from *in-vivo* data and testing it on *in-vitro* data, and vice versa. The latter finding supports that the more complex motif models indeed capture relevant biological information instead of amplifying experimental biases due to the different techniques.

4 Availability

We provide a Dimont web-server at `galaxy.informatik.uni-halle.de` and a command line application at `www.jstacs.de/index.php/Dimont`. For installing Dimont into a local Galaxy, the web-application is also available from the Galaxy tool shed [B+14] at `toolshed.g2.bx.psu.edu/ view/grau/dimont_motif_discovery`. Dimont has been integrated into the open source platform Chipster [K+11] since Chipster v2.11.

5 Talk outline

After a brief introduction into the problem of *de novo* motif discovery and experimental techniques, the first part of the talk focuses on the specifics of the approach implemented in Dimont. While some of the methods presented in this part are specific to motif discovery, others are also applicable to other fields of bioinformatics as, for instance, the common weighting schema developed for ChIP and PBM data. In the second part, we will present results of the comparison of *in-vivo* and *in-vitro* data and of modeling intra-motif dependencies. Finally, we give a brief overview of the Dimont web and command line applications.

References

[B+14] Daniel Blankenberg et al. Dissemination of scientific software with Galaxy ToolShed. *Genome Biology*, 15(2):403, 2014.

[GPGK13] Jan Grau, Stefan Posch, Ivo Grosse, and Jens Keilwagen. A general approach for discriminative de novo motif discovery from high-throughput data. *Nucleic Acids Research*, 41(21):e197, 2013.

[K+11] M Aleksi Kallio et al. Chipster: user-friendly analysis software for microarray and other high-throughput data. *BMC Genomics*, 12(1):507, 2011.

[M+12] Xiaotu Ma et al. A highly efficient and effective motif discovery method for ChIP-seq/ChIP-chip data using positional information. *Nucleic Acids Res*, 40(7):e50, 2012.

[W+13] Matthew T. Weirauch et al. Evaluation of methods for modeling transcription factor sequence specificity. *Nat Biotech*, 31:126–134, 2013.