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*]

[1]Institute of Computer Science, Martin Luther University Halle-Wittenberg, Germany

[2]Julius Kühn-Institut, Institute for Biosafety in Plant Biotechnology, Germany

[*]Both authors contributed equally

Email of corresponding author: jens.keilwagen@jki.bund.de

De-novo motif discovery has been an important challenge of bioinformatics for the last two decades. Since the emergence of high-throughput techniques like ChIP-seq, ChIP-exo, and protein binding microarrays (PBMs), the focus of de-novo motif discovery has shifted to runtime and accuracy on large data sets. For this purpose, specialized algorithms have been designed for discovering motifs in ChIP-seq *or* PBM data. However, none of the existing approaches works perfectly for all three high-throughput techniques.

Here, we propose *Dimont*, a general approach for fast and accurate de-novo motif discovery from high-throughput data.

We demonstrate that Dimont yields a higher number of correct motifs from ChIP-seq data than any of the specialized approaches, and achieves a higher accuracy for predicting PBM intensities from probe sequence than any of the approaches specifically designed for that purpose. Dimont also reports the expected motifs for several ChIP-exo data sets.

Investigating differences between *in-vitro* and *in-vivo* binding, we find that for most transcription factors, the motifs discovered by Dimont are in good accordance between techniques, but we also find notable exceptions. We also observe that modeling intra-motif dependencies may increase accuracy, which indicates that more complex motif models are a worthwhile field of research.