MDC · Robert-Rössle-Straße 10 · 13125 Berlin

Prof. Dr. Uwe Ohler
Berlin Institute for Medical
Systems Biology (BIMSB)
Computational Regulatory Genomics

Robert-Rössle-Straße 10
13125 Berlin

Tel.: +49 (0)30 9406-1810
Fax: +49 (0)30 9406-1751

Uwe.Ohler@mdc-berlin.de
http://ohlerlab.mdc-berlin.de

Berlin, Jan 22, 2023

Evaluation of Anna Macioszek's PhD thesis "HMM-based method for identifying enrichment in signal form sequencing based experiments"

Deep sequencing approaches have transformed the field of biology and gene regulation, by enabling the genome-wide profiling of biological molecules or their interactions. In particular, chromatin immunoprecipitation followed by sequencing (ChIP-seq) has not only allowed for a better understanding of the target sites of transcription factors, but also of chromatin structure. Specifically, mapping the locations of modified histones has led to the identification of the histone code, i.e., which modifications co-/occur in which context of gene expression.

Many computational "peak calling" methods have been developed since these experimental data have become available, which take raw ChIP sequence reads to genomic annotation by identifying accumulations of read matches in a local neighborhood. While superficially appearing as a "solved" problem, this problem is still far from trivial, not solely because of varying quality and amount of data, but rather because of the conundrum that there is no one-size-fits-all solution: Depending on the entity that is profiled, the length of the genomic region in which a particular factor or modification is observed can vary over orders of magnitude, leading to vastly different statistical properties of the resulting data. Most peak callers have been focused on small regions, here defined as individual binding sites or a typical size of a regulatory region of a few hundred nucleotides such as a single promoter or enhancer. However, histone modifications that indicate larger active or repressed genomic domains, can stretch over tens of kilobases, and it is this type of phenomenon that is addressed in this work.

The thesis starts with a basic level introduction to genetics, genomics and transcriptional regulation and then turns to a description of the biological

experiments, resulting data, and a detailed overview of three related computational approaches: one that is likely the most widely used ChIP-seq analysis tools, and two that were also developed to identify larger domains.

This serves as motivation to develop a new approach based on hidden Markov models dubbed HERON, in which the input data and state structure is explicitly setup to define large domains without splitting them into arbitrary, smaller ones. The main conceptual novelty is to integrate negative binomial (NB) distributions as emission distributions, which are the current, most widely used standard as they reflect the read distributions of many deep sequencing assays. Mrs Macioszek carefully lays out the standard training equations for discrete and Gaussian distributions, before she addresses the issue of NB, which unfortunately has no closed form solution due to its coupled parameters.

The next chapters turn to three distinct application scenarios, which range from simulated data, to widely used datasets from a large consortium, to data from local collaborators, which are the main motivation of the project. Over the course of these experiments, it becomes clear that HERON typically works better on its specific task than the methods it is compared to, but that Gaussian distributions appear more successful than NB – most likely due to the closed form and more stable parameter estimation.

I found room for improvement largely in the relative lack of discussion and application of methods from recent years. There are multiple experimental protocols that have started to deliver related relevant data – from ChIP-exo/nexus to Cut&Tag and single cell approaches. Especially as Cut&Tag is becoming a de facto standard for histone modification mapping, it would have been helpful to discuss this approach and the potential to apply or modify HERON accordingly. Similarly, Mrs Macioszek's choices to compare the performance are well motivated, but limited to methods that are >10 years old by now. She briefly introduces several newer ones, but it does not become clear why none of these were included (MACS2 has been out for several years now as well). While well written, the biological discussion (i.e. what is the current state of knowledge underlying the histone modifications, esp. regarding the 3D architecture of these domain inside the nucleus) does not go exceedingly into depth. Finally, I was left wondering whether HERON might not be applicable to domains of positive histone modifications, specifically,

H3K27 acetylation, when it occurs over larger distances of so-called "super-enhancers".

That said, throughout the thesis, the description and results are very clearly laid out, the methods contain adequate detail, and the discussion of the results on multiple data sets is exemplary. Due to the lack of ground truth, evaluating performance is difficult for many biological questions, and the candidate makes a commendable effort to address this issue from multiple angles. Overall, Mrs Macioszek's thesis describes a distinctive achievement to develop a new and well-performing computational method to a specified problem.

From my assessment, both the written thesis as well as the work described in it are without any reservation sufficient to grant a PhD.

Uwe Ohler

Digitally signed by Uwe Ohler
Date: 2023.01.25 11:03:08 +01'00'

Dr.-Ing. Uwe Ohler

Professor of Systems Biology
Humboldt University Berlin

Senior Investigator
Max Delbruck Center for Molecular Medicine