Halla-aho, Viivi; Lähdesmäki, Harri

Efficient statistical methods for detecting differential methylation

Published: 22/07/2017

Please cite the original version:
Background and motivation

- Addition of the methyl group to the 5-position of a cytosine (5mC) is the most commonly studied epigenetic modification on DNA, and its effects on different diseases and cancer have been widely studied.
- We have previously developed a hierarchical generative model, LuxGLM [1], for analysing 5mC and oxidized methylcytosine species (oxi-mC).
- LuxGLM can take into account the different experimental parameters and confounding factors along with complex experimental design.
- To enhance the computational efficiency we propose the usage of variational inference (VI) instead of Hamiltonian Monte Carlo (HMC) sampling. VI is typically faster than MCMC sampling methods.

LuxGLM

- Read-out probabilities for single cytosine and for a population

The general linear model is used for calculating $\theta_i = (p(C), p(5mC), p(5hmC))$ for each sample $i = 1, \ldots, N$.

- General linear model

The linear part of the model with $p$ covariates has the following form

$$ Y = DB \cdot E, \quad (1) $$

where $Y \in \mathbb{R}^{N \times M}$ gives the parameters $\theta_i$ through Softmax transformation $\theta_i = \text{Softmax}(\text{row}(Y))$, $D \in \mathbb{R}^{N \times p}$ is the design matrix, $B \in \mathbb{R}^{p \times M}$ is the parameter matrix and $E \in \mathbb{R}^{1 \times M}$ represents normally distributed, zero-centered noise term.

- Bayes factors

To assess the difference in methylation between two conditions $i$ and $j$ the null hypothesis (no differential methylation) is

$$ H_0: \text{row}(B) - \text{row}(B) \succeq C_1 - C_2 = 0, \quad (2) $$

and alternative hypothesis (differential methylation) is

$$ H_1: \text{row}(B) - \text{row}(B) \not\succeq C_1 - C_2 \neq 0. \quad (3) $$

The Savage-Dickey density ratio approximates the Bayes factor between the models representing these hypotheses

$$ BF \approx \frac{p(C_1 - C_2 = 0 | H_1)}{p(C_1 - C_2 = 0 | H_0)}, \quad (4) $$

- Model hierarchy

HMC sampling from the posterior is done with Stan.

Variational inference for computation of the Bayes factors

- Variational inference approximates the posterior with a simpler distribution and to find the optimal approximative distribution, the expectation lower bound (ELBO) is maximized, which corresponds to minimizing the Kullback-Leibler distance.
- In the probabilistic programming language Stan, Automatic Differentiation Variational Inference (ADVI) algorithm has been implemented [2] and so the HMC sampling used by default in Stan can be easily switched to VI. ADVI algorithm parameters which can be tuned are number of gradient samples $N_{F}$ and number of ELBO samples $N_{E}$.
- The ELBO values for the approximations can be used to calculate another BF approximation

$$ BF \approx \exp(\text{ELBO}_{H_0} - \text{ELBO}_{H_1}). $$

Comparison of LuxGLM and state-of-the-art methods

Comparison table of LuxGLM, RADMeth [3] and MACAU [4] from [1]. In the comparison the area under receiver operating characteristic curve (AUROC) was calculated using simulated data sets. Perfect experimental steps and only BS-seq data were considered in the simulations, as experimental parameters and oxi-mC are not supported by the other methods.

<table>
<thead>
<tr>
<th>Number of reads</th>
<th>AUROC</th>
<th>Time</th>
<th>AUROC</th>
<th>Time</th>
<th>AUROC</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>0.958</td>
<td>0.42</td>
<td>0.895</td>
<td>0.75</td>
<td>0.976</td>
<td>0.67</td>
</tr>
<tr>
<td>6</td>
<td>0.976</td>
<td>0.67</td>
<td>0.985</td>
<td>0.84</td>
<td>0.760</td>
<td>0.67</td>
</tr>
<tr>
<td>8</td>
<td>0.976</td>
<td>0.67</td>
<td>0.985</td>
<td>0.84</td>
<td>0.760</td>
<td>0.67</td>
</tr>
<tr>
<td>10</td>
<td>0.976</td>
<td>0.67</td>
<td>0.985</td>
<td>0.84</td>
<td>0.760</td>
<td>0.67</td>
</tr>
<tr>
<td>12</td>
<td>0.976</td>
<td>0.67</td>
<td>0.985</td>
<td>0.84</td>
<td>0.760</td>
<td>0.67</td>
</tr>
<tr>
<td>16</td>
<td>0.976</td>
<td>0.67</td>
<td>0.985</td>
<td>0.84</td>
<td>0.760</td>
<td>0.67</td>
</tr>
</tbody>
</table>

Computation times for variational inference and comparison with HMC

- Computation times using Stan’s variational inference feature with different parameter values to compute the Savage-Dickey and ELBO approximations of the Bayes factor. The number of reads was 12 and number of replicates was 10.

- Comparison table of the AUROC values and mean computation times in seconds of the original Savage-Dickey estimate and Savage-Dickey and ELBO estimates calculated using variational inference for simulated data. The algorithm parameters were $N_{F} = 10$ and $N_{E} = 1000$ for ADVI.

<table>
<thead>
<tr>
<th>Number of replicates</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>20</td>
</tr>
</tbody>
</table>

- Scatterplot of the mean computation times and differences in AUROC with Savage-Dickey approximation calculated using HMC using different parameter values for ADVI. Number of reads was 12 and number of replicates was 10.

References