Unsupervised domain adaptation methods for cross-species transfer of regulatory code signals

P. Latyshev

HSE University, Moscow, Russia,

pavel.latyshev@yandex.ru

F. Pavlov

HSE University, Moscow, Russia,

theodore.pavlove@gmail.com

A. Herbert

HSE University, Moscow, Russia,

InsideOutBio, Charlestown, USA,

alan.herbert@insideoutbio.com

M. Poptsova

HSE University, Moscow, Russia,

mPoptsova@hse.ru

Advances in next generation sequencing (NGS) technologies have made it possible to generate whole-genome maps for various functional genomic elements in a variety of species. However, the high cost and limited availability of experimental data pose challenges for many species of interest. Deep learning methods have emerged as the leading computational approaches to analyze the existing data, although their focus tends to be limited to the studied species. In this work, we leverage advances in transfer learning, specifically unsupervised domain adaptation (UDA), to address this limitation. We evaluate nine UDA methods for predicting regulatory code signals in the genomes of other species. Our approach involves training deep learning models on experimental data from one species, and then refining the models using the genome sequence of the target species for prediction purposes.

Out of all nine domain adaptation architectures tested, the non-adversarial methods Minimum Class Confusion (MCC) and Deep Adaptation Network (DAN) show superior performance compared to the others. We provide an empirical evaluation of each approach using real-world data, specifically ChIP-seq data for transcription factor binding sites and histone marks in multiple genome assemblies: human, mouse, fruit fly and worm. Although, these findings

are applicable to any cross-species transfer scenario.

To evaluate the efficiency of each method, we use species for which experimental data are available for both source and target organisms. This evaluation allows us to determine how well each implementation performs when limited experimental data are available, thus aiding the design of future experiments in understudied organisms. Overall, our results validate the effectiveness of UDA methods in generating missing experimental data for histone marks and transcription factor binding sites across various genomes. Furthermore, our study emphasizes the robustness of these different approaches in handling incomplete, noisy, and analytically biased data.