Project: **Reconstruction of gene regulatory interactions using machine learning**

Cis-regulatory elements (also called enhancers) are regions of the genome that control expression of neighboring genes though binding of specific proteins and making a physical connection with regions proximal to gene transcription start sites (promoters). In any given cell type, locations of active cis-regulatory elements can be predicted using such techniques as ATAC-seq and DNAse-seq.

Description

Machine learning methods are currently revolutionizing the field of computational biology. Given enough training data, almost any prediction task can be learned to a very high degree of accuracy. In this project you will attempt to learn genes associated with particular cis-regulatory elements (enhancers). The major challenge comes from the fact that a cis-regulatory element may regulate genes located hundreds of thousands of nucleotides away in genomic coordinates, this regulation can be tissue-specific and one cis-regulatory element can regulate several genes.

For the learning and validation tasks, you will use several data sets of open chromatin regions (putative enhancers and promoters) in different cell types and hi-ChIP data showing experimental enhancer-promoter interactions in the corresponding cells.

The method you will create is likely to become an important part of state-of-the-art analysis pipelines of open chromatin data (ATAC-seq and DNAse-seq). Most importantly, it will help construct tissue-specific gene regulatory networks.

Techniques

This project will involve developing a stand-alone application (in your favorite programming language) to predict enhancer-promoter interactions from open chromatin data. You will build and validate a predictive model based on a set of biologically relevant features. You could choose between different approaches to solve the task: from constructing an empirical algorithm to applying deep learning.

Requirements

You should have a strong background in programming and data analysis, and at least basic knowledge of processes related transcriptional regulation and chromatin organization. The project will be divided roughly into 60% development, 30% theory and literature review, and 10% writing of the thesis.

You will work on this project in the laboratory of Computational Genetics and Epigenetics of Cancer (Department of Computer Science, ETH Zurich).

Goals

- Demonstrate the feasibility of learning approaches for prediction of enhancer-promoter interactions solely based on open chromatin data
- Assess the predictive power of the method on available data sets
- Thoroughly evaluate technique for thesis and potentially for a publication

Supervisor:

Prof. Valentina Boeva, CAB building, Department of Computer Science, ETH Zurich, valentina.boeva@inf.ethz.ch