Project topics for the course CS-E4880 Machine Learning in Bioinformatics

March 29, 2019

Determining epitope-specificity of TCRs with Gaussian processes

Instructor: Emmi Jokinen (emmi.jokinen@aalto.fi)

Background: T cell receptors (TCRs) of a T cell determine which epitopes (portions of foreign proteins, or antigens) that cell can recognize and thus what kind of immune responses it can initialize. TCRs can be sequenced from individuals and methods that can analyze the specificity of the TCRs can help us better understand the individual's immune status in different diseases.
In addition to getting high prediction accuracy for the epitope-specificity of a TCR, it's desirable to get biological insights to what properties are important in determining the specificity.

Each TCR has short CDR (complementarity determining region) sequences that consist of the 20 naturally occurring amino acids. CDR3 has most contact with the epitope presented to the TCR and is generally most used in modeling. However, CDR1, CDR2, and CDR2.5 may also provide additional useful information.

Aim: In this project, you will create a feature presentation for CDR sequences that utilizes k-mers. This feature presentation will then be used in Gaussian process (GP) framework to predict if TCRs recognize certain epitopes or not. You will also get familiar with GPflow, which is a package for building Gaussian process models in Python using TensorFlow. To better understand the underlying problem and the functioning of this model, you should infer which k-mers are important. For this, you can use either automatic relevance determination (ARD) e.g. with RBF-kernel and/or inducing points which are used with Sparse variational Gaussian processes (SVGP). With ARD, each dimension (corresponding to one k-mer) will be given its own optimized lengthscale, which also determines the importance of that dimension / k-mer. With SVGP and inducing points, the inducing point locations and values determine which k-mers are important in determining the epitope-specificity of TCRs.

In addition, choose 1-2 of the following:

- Experiment with different kernels. You may construct your own kernel, use known string kernels (e.g. GSkernel), or any other kernels of your choice
- Experiment with different feature representations. You can also use aligned sequences and utilize for example the different properties of the amino acids.
- Experiment with different GP techniques of your choice. For this, it is recommended that you have some prior knowledge of GPs

Materials and methods: This project should be implemented using Python with GPflow and TensorFlow. Use of Jupyter notebooks is recommended. You can use TCRGP implementation available at https://github.com/emmijokinen/TCRGP as a basis for your project. However, you may have to do some modifications and you will need to construct the feature presentation yourself. If necessary, you may limit your feature presentation to those k-mers that are present in the training set. The Github page also contains the data (epitope-specific and control TCRs) and some example code.

Prerequisites: Programming skills in Python, basic knowledge of machine learning. Some knowledge of kernel methods and Gaussian processes is beneficial.

References:

- Jokinen, E., et al. TCRGP: Determining epitope specificity of T cell receptors. Preprint at: https://www.biorxiv.org/content/10.1101/542332v1?rss=1
- Rasmussen, C. E. and Williams, C. K. I. (2006). Gaussian processes for machine learning. The MIT Press. http://gaussianprocess.org
- Matthews, A. G. d. G. et al. (2017). GPflow: A Gaussian process library using TensorFlow. Journal of Machine Learning Research, 18(40), 1–6. https://gpflow.readthedocs.io/
- Giguere, S. et al. (2013). Learning a peptide-protein binding affinity predictor with kernel ridge regression. BMC bioinformatics, 14(1), 82. https://github.com/GRAAL-Research/gs-kernel
- Kawashima, S. and Kanehisa, M. AAindex: amino acid index database. Nucleic Acids Res. 28, 374 (2000). https://www.genome.jp/aaindex/