Mathematical inference on diversity generation of B-cell receptors from sequence repertoires : pathological vs physiological cases

Vuk Milisic [*] Ian Morilla [*]

12 octobre 2017

The immune system recognizes infection and induces protective responses. Germinal centres (GCs) are transient structures that form within peripheral lymphoid organs in response to T cell-dependent antigen. Within GCs, B-cells expressing high-affinity antibodies develop and differentiate into antibody-secreting plasma cells and memory B cells that mediate and sustain protection against invading pathogens. The importance of the GC reaction is best shown by the immunodeficiency syndromes that are observed in patients who are unable to form GCs. The B-cell receptor (BCR), a protein located on the surface of the B-cell recognizes the antigen, if the BCR is affine enough, the B-cell is rescued and survives. Non-affine BCR makes B-cell die. This mechanism, in the origin of immune system diversity, is called antibody affinity maturation. The part of the BCR responsible for affinity is concentrated on 3 "hot spots" called complementary determining regions (CDR1, CDR2, CDR3).

In [6], the authors applied statistical inference methods in order to predict the generation probability of any specific CDR3 sequence for healthy human T-cells and they extended this physiological analysis later on to B-cell in further works [4, 5]. However, more methods approaching these questions have been developed as well, as for instance, machine learning methods (SVM, Random forests, Genetic Programming, ...). In [7], the authors use some of these methods to identify the relationship to describe the shape and progress of the affinity maturation to prevent error propagation rates derived from the number of sequences measured by next-generation sequencing (NGS) [1]. In the same way, the techniques of deep-learning have been also used to predict the sequence diversity in RNA or DNA-binding proteins [2].

We are mainly interested in applying these techniques on chronic lymphocytic leukemia (CLL), which is the commonest form of leukemia in Europe and North America. And primarily affecting, though not exclusively, elder individuals. This pathology can be summarized as an invasion of a B-cell specific clone that takes the place of a normal B-cell and prevent the proper activation of blood cells machinery facing the infection. The presence of these fake B-cell is lowly latent in the human body and frequently is detected in a quite advanced stage [3]. Though major progress has been made in identification of molecular and cellular markers that eventually predict disease progression in CLL patients, most of these mechanisms are still poorly characterized.

The selected candidate should become familiar with the literature presented above, write or use some related code and have a strong interest in bio-medical applications.

[1]Laboratoire Analyse, Géométrie et Applications, CNRS UMR 7539, Université Paris 13, Villetaneuse, France

This research fits in the frame of Labex *Inflamex* whose main topic is chronic inflammatory diseases. *LAGA*'s bio-math team is involved as a full partner inside this Labex. This master thesis is part of the collaboration with the research unit *Signalling adaptors in Haematology* headed by Nadine Varin-Blank. Huge amount of data are already available. This project shall provide a significant output in the biological community. Moreover, the candidate has the opportunity to participate to the summer-camp *CEMRACS 2018* in Luminy (Marseille). At last, this master theses could lead to a PhD position.

Références

[1] J. Benichou, R. Ben-Hamo, Y. Louzoun, and S. Efroni. Rep-Seq : uncovering the immunological repertoire through next-generation sequencing. *Immunology*, March 2012.

[2] J. Benichou, R. Ben-Hamo, Y. Louzoun, and S. Efroni. Predicting the sequence specificities of DnA- and RnA-binding proteins by deep learning. *Nature Biotechnology*, June 2015.

[3] G. Dighiero and T. J. Hamblin. Chronic lymphocytic leukaemia. *Lancet*, 371(9617) :1017–1029, Mar 2008.

[4] Y. Elhanati, Q. Marcou, T. Mora, and A. M. Walczak. repgenHMM : a dynamic programming tool to infer the rules of immune receptor generation from sequence data. *Bioinformatics*, 32(13) :1943–1951, Jul 2016.

[5] Y. Elhanati, Z. Sethna, Q. Marcou, C. G. Callan, T. Mora, and A. M. Walczak. Inferring processes underlying B-cell repertoire diversity. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, 370(1676), Sep 2015.

[6] A. Murugan, T. Mora, A. M. Walczak, and C. G. Callan. Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proc. Natl. Acad. Sci. U.S.A.*, 109(40) :16161–16166, Oct 2012.

[7] S. Schaller, J. Weinberger, M. Danzer, C. Gabriel, R. Oberbauer, and S.M. Winkler. Mathematical modeling of the diversity in human B and T cell receptors using machine learning. *The 26th European Modeling & Simulation Symposium*, Sep 2014.