PhD offer in applied mathematics – 2024-2027

Stochastic modeling and statistics for quantifying and predicting the evolution of tumor heterogeneity in chronic lymphocytic leukemia

Keywords: Applied probability, stochastic modeling, statistical modeling for medicine, variational Bayesian methods, clonal heterogeneity, chronic lymphocytic leukemia

Biological context

The development of targeted therapies has allowed considerable progress in the treatment of many cancers, but their efficacy is dependent on intra-tumor heterogeneity. In lymphomas and leukemias, the identification of gene alterations by high-throughput sequencing allows the characterization of this heterogeneity. In healthy B cells, the maturation process provides a unique sequence of DNA, called *VDJ genes*, encoding for the immune repertoire of the antigen receptor (BCR) by combining 3 immunoglobulin chains V, D and J (Fig. 1(A)). In contrast, in hemopathies, every B cell in the initial leukemic *clone* (i.e. population of tumor cells with the same genome) has the same antigen receptor encoded by a specific VDJ gene sequence. The occurrence of additional mutations in VDJ genes may be responsible for the emergence of subclones with increased antigen receptor reactivity further complicating the clonal heterogeneity of these hemopathies (Fig. 1(B)). Leukemic B cells therefore have two levels of heterogeneity: the heterogeneity of cancer genes (a feature shared by any cancer) and the heterogeneity of VDJ genes (a feature specific to leukemia). However, these two levels of clonal heterogeneity and their co-evolution remain poorly characterized and are not considered in the management of these cancers today.

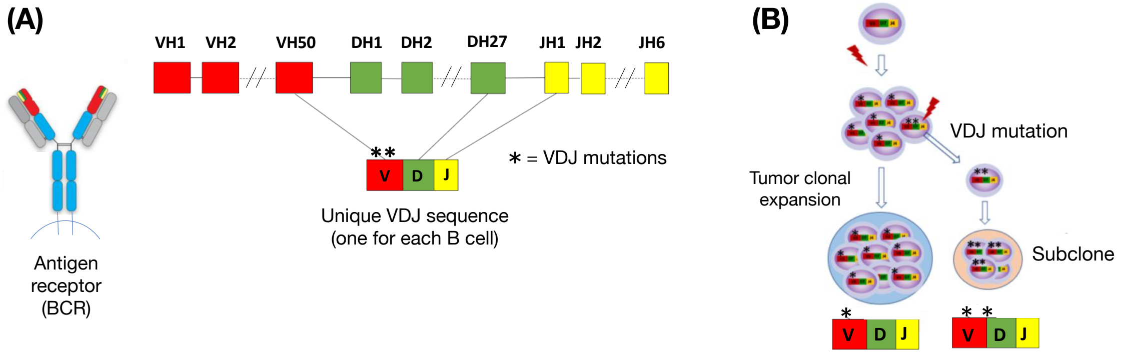

Figure 1: Biological context of B cell heterogeneity with respect to VDJ genes. (A) Physiological process of VDJ recombination: each mature B cell ends up with a unique VDJ sequence coding for its membrane receptor. (B) Tumor clonal expansion in the context of chronic lymphocytic leukemia produces a clonal population of B cells, which can be identified from their common VDJ sequence. Subsequent mutations can lead to subclones.

Project description

We propose to develop a mathematical model for the evolution of the two levels of clonal heterogeneity in leukemia, allowing to characterize their evolution from temporal bulk sequencing data of VDJ and cancer genes mutations using a Bayesian approach. We will test the predictive performance of clonal evolution from the inferred model.

Tasks

In this PhD project, we propose to tackle the problem of clonal reconstruction, first from data collected at a single time (already available), and second from longitudinal data. Data will be collected throughout the duration of the PhD thesis.

The main problem consists in reconstructing the phylogenetic tree of mutations and the dynamics of frequencies of each clone. The originality comes from the fact that data are heterogeneous: we will have the full profile of VDJ mutations of clones with frequencies and each cancer genes variants with allele frequencies. From the mathematical modeling perspective, VDJ data share common features with single-cell data since full sequences can be reconstructed using tools like MiXCR. Existing packages for clonal heterogeneity analysis are B-SCITE (Malikic et al., 2017) and ddClone (Salehi et al., 2017). They are able to deal with both types of data (bulk and single-cell) and could in principle be used here. However, there are specificities of CLL that do not fit into these methods.

The PhD student will first construct a probabilistic model accounting for all the data. This model will contain the phylogenetic tree as latent variable, where each node in the tree corresponds either to a VDJ mutation, a mutation of cancer genes, or a chromosomic alteration, where each mutation occurs only once in the tree. The observations will then be obtained, following the classical rules of the *infinitely many sites* model, as linear combinations of the frequency of every clone in the sample (which are other latent variables), possibly with some noise.

Treating latent variables as parameters, we could use the maximum likelihood method, but maximization is a difficult problem in practice due to the very large number of possible trees. We will test genetic algorithms (Metropolis-Hastings, MCMC...), but we expect better results using a Bayesian approach, combined with a variational method to maximize the a posteriori likelihood.

Second, the PhD student will validate the method from data simulated from our model, then using the benchmark simulation tool proposed by Foglierini et al. (2020), adapting them to our double heterogeneity context, and finally comparing with single-cell sequencing data of 3D *in vitro* cultures of proliferating cells that will be collected all along the project. Prediction performances will also be tested.

Finally, we will try to detect if groups of patients have similar mutational patterns (such as phylogenetic tree topology), which could correspond to a similar tumorigenesis, or a similar stage of progression, or a similar response to treatments. This is a clustering problem that can be addressed by model-free artificial intelligence tools (such as latent Dirichlet allocation: Pritchard et al., 2000; Falush et al., 2003), or using models like those developed by Beerenwinkel et al. (2004, 2005). This will allow us to build a predictive model of treatment efficiency given the clonal heterogeneity of a patient, that can be used by clinicians in a context of personalized medicine.

Thesis context

The thesis will take place in the Probability and Statistics team of the Institut Élie Cartan de Lorraine (IECL) in Nancy and in the SIMBA team (Statistical Inference and Modeling for Biological Applications) of Inria Nancy. The PhD student will be involved in discussions with staff at the Strasbourg University Hospital on medical and data aspects all along the

PhD project. During the thesis, the PhD student will have the opportunity to discover the world of mathematical research through the life of a dynamic mathematics laboratory, and to attend seminars and working groups in probability and statistics.

Skills

The candidate should have skills in statistics and/or stochastic modeling. R, Python or Matlab programming skills are also required. An affinity or experience with medical applications will be highly appreciated.

Remuneration

The thesis is funded by ITMO cancer (INSERM funding). Monthly gross salary amounting to 2082 euros.

Supervision

The thesis will be supervised by Nicolas Champagnat, Coralie Fritsch and Ulysse Herbach (IECL and INRIA Nancy - Grand Est) for the mathematical part and by Laurent Vallat (CHRU Strasbourg and University of Strasbourg) for the medical part.

How to apply?

On the Inria website.

Contacts

nicolas.champagnat@inria.fr, coralie.fritsch@inria.fr, ulysse.herbach@inria.fr

Bibliography

- Malikic, S. et al. (2019). Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data. Nature communications, 10(1), 1-12. https://doi.org/10.1038/s41467-019-10737-5
- Salehi, S. et al. (2017). ddClone: joint statistical inference of clonal populations from single cell and bulk tumour sequencing data. Genome biology, 18(1), 1-18. https://doi.org/10.1186/s13059-017-1169-3
- Foglierini, M. et al. (2020). AncesTree: An interactive immunoglobulin lineage tree visualizer. PLoS computational biology, 16(7), e1007731. https://doi.org/10.1371/journal.pcbi.1007731
- Pritchard, J. K. et al. (2000). Inference of population structure using multilocus genotype data. Genetics, 155(2), 945-959. https://doi.org/10.1093/genetics/155.2.945
- Falush, D. et al. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics, 164(4), 1567-1587. https://doi.org/10.1093/genetics/164.4.1567
- Beerenwinkel, N. et al. (2004). Learning multiple evolutionary pathways from cross-sectional data. In Proceedings of the eighth annual international conference on Research in computational molecular biology (pp. 36-44). https://doi.org/10.1145/974614.974620
- Beerenwinkel, N. et al. (2005). Mtreemix: a software package for learning and using mixture models of mutagenetic trees. Bioinformatics, 21(9), 2106-2107. https://doi.org/10.1093/bioinformatics/bti274