Machine learning for super resolution microscopy

PhD proposal

Nataliya Sokolovska
Sorbonne University
LCQB, Machine learning team
nataliya.sokolovska@sorbonne-universite.fr

Judith Miné-Hattab
Sorbonne University
LCQB, team FIONA
judith.mine-hattab@sorbonne-universite.fr

Context. Development of imaging techniques, especially, microscopic imaging techniques, has made it possible to visualize various biological phenomena. Biologists often investigate images manually, it requires a lot of effort, time, and concentration. Note that it is hardly possible to analyse a big amount of data reliably manually. At the same time, research on automated image processing is a very active domain of artificial intelligence and machine learning, and a number of efficient image processing and pattern recognition approaches to analyse biological images appeared recently. In recent years, different techniques called super resolution microscopy emerged allowing the observation of details inaccessible so far. Among them, single molecule microscopy offers the best resolution: it allows us to visualize individual molecules inside cells at 20 nanometers resolution [Betzig et al., 2006]. Single molecule microscopy generates very large amount of data, unfortunately, powerful tools to analyze and quantify automatically such images are still lacking. This project is highly interdisciplinary: it is a collaboration between two new teams of the LCQB, SU, namely, FIONA (Functional Imaging of Nuclear Architecture), led by Judith Miné-Hattab, and Computational medicine, led by Nataliya Sokolovska. The current PhD project will reinforce both newly created research units. Note that the collaboration between J. Miné-Hattab and N. Sokolovska has de facto already started in the context of a Master 2 BIM (Biology, Informatics, Modelling) internship; an intern student joined the LCQB in February 2023. He is supervised by J. Miné-Hattab and N. Sokolovska, and his internship is focused on the machine learning methods applied to the analysis of single molecule microscopy images.

Objectives. Our ultimate goal is to propose a novel and specific method to stratify single molecule microscopy images, so that the proposed approach ensures rapidity and reproducibility of bioimage analysis. Two types of images will be quantified automatically (see Figure 1): i) images containing the distribution of proteins (Photo-Activable Localization Microscopy - PALM) in which specific proteins patterns need to be recognized and quantified automatically; ii) images containing proteins dynamics (Single Particle Tracking - SPT) in which the motion of proteins will be quantified and classified. We will explore the state-of-the-art methods, among them random forests, XGBoost, and various deep learning architectures. Our images come from rich and original data sets produced by the FIONA team researchers. Remark, that several data sets including public data, are already available for the project. The projects aim is on the intersection between two scientific domains in which each team is already recognized at the international level: fundamental biology (cutting edge microscopy, DNA repair) and computer science (numerical optimisation, statistical machine learning). Thus, the results of the PhD project will have impact in both domains, namely, introducing novel mathematically sound methodology with theoretical guarantees, and revealing new knowledge about DNA repair proteins mobility in human cells.

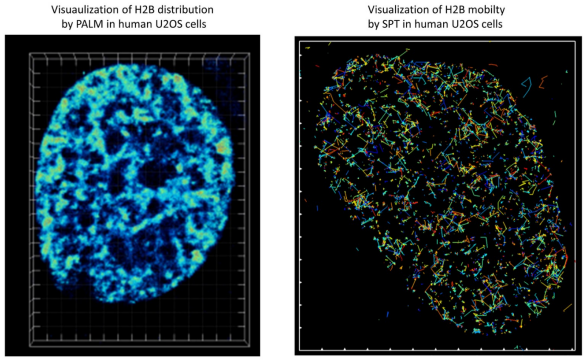

Figure 1: Left: Visualization of histones H2B fused to dendra2 in human U2OS cell by PALM (FIONA team). The color indicates the density of histones for each detected histone. Right: Visualization of histone H2B mobility by SPT in human U2OS cell (FIONA team). Each line represents a trajectory of a single histone molecule. The trajectories exhibit different kinds of mobility (slow, fast, more complex behaviour).

Models, methods, and goals. One of the important problems in live-cell experiments is that they are not quite repeatable. Even a small variation in cell culture, age of the cell, etc. changes the cells behaviour. Generating ground truth for such heterogeneous data is practically impossible. So, it is problematic to apply the modern machine learning methods, since they often rely on large, controlled, purified, and often extensively annotated sets. In our case, the data describe the precise distribution and motion trajectories of histones and DNA repair proteins. From the machine learning viewpoint, in this data we have a relatively small number of raw features (time and corresponding histones positions), and potentially a very big number of time points for each observation. Note that the number of time points varies from one histone to another, and is, therefore, proper to each observation. We have identified the following research directions which are interrelated to achieve our objectives taking into consideration the particularities of microscopy data:

- *Pattern recognition and quantification.* PALM allows to measure the distribution of proteins inside cells (local density, specific patterns, condensates) at unprecedented resolution. These images consist of a cloud of points, each point representing the position of a single molecule inside the observed cell. The first step of the project is to develop tools to quantify a cloud of points. Algorithms such as Non Linear Principal Component Analysis will be tested for example. We expect to determine shapes, quantify length and width, local density, curvature, lacunarity, etc. In particular, the FIONA can provide experimental data of i) histones distribution, ii) repair proteins distribution (PARP1 and in the future 53BP1) which form different patterns. The size and shape of these patterns vary in mutant cells, thus developing powerful algorithms to detect and quantify them rapidly will be an important breakthrough in the field of DNA repair.

- *Feature engineering for classification and motion analysis.* In real biological and medical studies, data are often huge dimensional and noisy, sometimes with some underlying unknown structure. The data come usually without any additional information on what methods should be preferred, and it is not obvious which characteristics should be extracted to construct an efficient predictive model. Single Particle Tracking allows us to measure the diffusion of individual proteins and to have a sense of where this protein is moving fast or slowly in the cell. This information is precious, and the captured dynamics can be used to compute the displacements, type of motion that are, e.g., clear indicators of cells health. More specifically, the FIONA team will provide data of histones and repair mobility in the absence and in the presence of DNA damage. We recently showed that several anti-cancer drugs such as PARP inhibitor affect histones mobility. Thus, we will apply the analysis developed here to better characterizing the effect of these drugs on histones mobility.

- *Weakly supervised learning.* Manual labelling costs time of human experts, and can also be erroneous. We aim to propose approaches to replace manual and fully supervised methods for living cells stratification. A research direction is self-supervised learning, see, e.g., [Robitaille et al., 2022] where it is discussed that there is an acute need to develop methods that do not ask for extensive input from users, i.e., do not need pre-processing steps such as full data labeling and training process supervision. Such a method would significantly increase the reproducibility of cell imagery results. Note that the state-of-the-art methods such as transfer learning – where pre-trained models are applied to similar data sets – often lead to poor performance on validation sets.

- *An online anytime approach to learn motion patterns directly on the microscope.* Our ambitious goal is to develop a new efficient, both computationally and in terms of performance, pipeline that allows biologists to verify their hypotheses fast and with some performance guarantees. The pipeline would include the automated feature engineering and will not ask human experts to perform manual data labelling. A research avenue that we are going to explore relies on that a complex motion pattern of an individual cell can be decomposed into piece-wise simple patterns, that can be analysed separately and together.

Deep learning (DL) methods were reported to achieve the state-of-the art results in image processing tasks, in various domains, including biological and medical applications. DL was developed for many ML scenarios such as supervised, unsupervised, reinforcement, structured learning. DL was also explored for microscopy data [von Chamier, 2021]. For example, in respect to our task of motion patterns analysis of individual proteins, a

shallow network with an attention mechanism can be used, where the type of motion pattern could be encoded by the networks layers. The main criticism of DL concerns its black-box nature, especially in high-stakes domains such as fundamental biology and clinical predictions. To overcome this drawback of DL, some steps to increase the interpretability of the deep models were proposed. In this project, we are particularly interested in deep decision trees and deep random forests, the recently proposed models that take the best of the both worlds: accurate prediction and explainability by human experts. To our knowledge, the most recent related results were published by [Maris et al., 2022] where the trajectory classification and motion analysis of molecules were considered and decision trees were applied, however, in [Maris et al., 2022] the decision rules were constructed manually.

Real applications: link with cancer research. Single molecule microscopy is developing rapidly and it became possible to observe the localization and diffusion of individual molecules in living cells. In particular, here we focus on the mechanisms of DNA repair and the effect of anti-cancer drugs (PARP1 inhibitor) on histones distribution and mobility. Recently, a strong interest in PARP inhibitor (PARPi) has grown for the treatment of cancers, originating from the observation that the PARPi Olaparib induces synthetic lethality in BRCA1/2 deficient tumour cells. The FIONA team will provide experimental data on the effects of PARP inhibitor on histones distribution and mobility using super resolution microscopy. The tools will be an important breakthrough to extract automatically more information from PALM and SPT images. Our work will be an initial step in the understanding of the impact of histones marks on cell sensitivity to PARPi. Therefore, it may contribute to identify new biomarkers helping to predict the efficiency of PARPi treatment and also pave the way for new therapies potentiating PARPi or reducing the risk of acquired resistance.

Requirements for the potential candidates. The candidate is expected to have a Master 2 in Computer Science or Applied Mathematics, or an equivalent engineering degree. A background in statistics, optimization, decision theory or any related field will be appreciated. An ideal candidate will propose, develop, and numerically test the developed methods. It is expected that the candidate provides some theoretical foundations for the methods and also implements them in Python, so that the final product could be publicly available for research purposes. Some interest in the biological applications is appreciated.

The role of supervisors. To provide expertise and supervision in fundamental biology, bioinformatics, and machine learning, the PhD will be co-supervised by Nataliya Sokolovska and Judith Miné-Hattab, LCQB, Sorbonne University. N. Sokolovska will lead the project on methodological and computational issues, i.e., development and testing of novel methods. J. Miné-Hattab being an internationally recognised specialist in DNA repair, will give insights into the biological challenge of the project, provide with real data, and evaluate the results. The PhD candidate will tightly collaborate with both supervisors.

Teams of supervisors. N. Sokolovska leads a machine learning team in the LCQB, SU; J. Miné-Hattab leads FIONA team, LCQB, SU. Both teams were created in 2022, therefore, the PhD student will reinforce both research units and develop a collaboration between them.

PhD school. The PhD student will be affiliated with the EDITE (*Ecole Doctorale Informatique, Télécommunications et Electronique*) doctoral school whose member is N. Sokolovska, and since the main goal of the project is to develop novel highly competitive statistical and probabilistic machine learning methods.

References

[1] M.C. Robitaille, J.M. Byers, J.A. Christodoulides, M. P. Raphael. Self-supervised machine learning for live cell imagery segmentation. *Commun Biol*, 5, 1162, 2022.

[2] P. Kontschieder, M. Fiterau, A. Criminisi, S. R. Bulò. Deep neural decision forests. *ICCV*, 2015.

[3] J. J. E. Maris, F. T. Rabouw, B. M. Weckhuysen, F. Meirer. Classification-based motion analysis of single-molecule trajectories using DiffusionLab. *Scientific Reports*, 12, 9595, 2022.

[4] L. von Chamier et al. Democratising deep learning for microscopy with ZeroCostDL4Mic. *Nature Comm.*, 12, 2276, 2021.

[5] E. Betzig et al. Imaging intracellular fluorescent proteins at nanometer resolution. *Science*, 313(5793), 2006.