Lab rotation at Computational Biomodeling Laboratory Vilhelm Suksi

Intent

During my biochemistry studies, I realized I missed something. Much of the material we had to learn we had to learn by heart. Even at the expense of the mind, it seems! Most pathways and cascades are so interconnected that everything affects everything. Thus, it almost became a joke how, during presentations or exam answers or whatever, if one's mind went blank one could always say that this or that can also contribute to cancer. And it's true! But it's not a satisfying answer. Instead, perhaps from a deterministic worldview, arises the notion that the relationships can and should be modelled. Maybe, one day, if there are enough atoms in the universe to carry out the computations, we can model life perfectly to the extent that randomness allows for.

Such thoughts were very intriguing, especially since modern biology is characterized by massive amounts of data and reductionistic research. Instead of chopping life into smaller and smaller pieces, maybe quantifying and modeling the phenomena could be worthwhile? Thanks to the internet and some academic freedom, I was able to start pursuing this path. Since the mathematics, statistics and related fields were unfamiliar to me, I needed a flexible context so that I could decide what I wanted to learn and in what order. Fittingly, there was some room in my curriculum for such meandering, in the form of a Lab Rotation course. If I only could find a suitable research group, of course.

And there it was, the Computational Biomodeling Laboratory! For two whole months, I was able to explore relevant domains and plan a learning path. The project I was given the "keys" to was rather steep for a first modelling project, but seminar participation gave some much needed context: there are other people doing this sort of thing. Even if I couldn't contribute anything useful, I came to understand the project very well and thought it was very interesting to follow along. In the following pages I will communicate my understanding of the project and the data science material I've learned during my time at CBL, along with reflection and some considerations to continue my learning.

Classification of diffuse glioma subtypes

Diffuse gliomas account for 80% of brain tumors[1]. Diagnosis of diffuse glioma and its subtypes has relied heavily on histological criteria. Histological classification, however, is riddled by the human factor. The difficulty of classifying diffuse gliomas is further confounded the grade of the tumor. Molecular characterization studies have given diffuse glioma classification a more secure footing. Promising classification schemes seek to identify low-grade diffuse gliomas before progression into glioblastoma multiforme. Classification irrespective of grade allows for earlier, more specific treatments, leading to improved clinical outcomes and less unnecessary suffering.

Both molecular and histological characterization of diffuse gliomas is invasive, requiring a biopsy and time. This makes diagnosis resource-intensive and complications more likely. The project at hand explores a putative solution for rapid, minimally invasive classification of diffuse gliomas by Raman spectroscopy. If the Raman spectra allows for accurate classification, a biopsy might not be needed at all. This could mean that subtype diagnosis and potential treatment by incision can be performed in a single session. The six-class classification problem is informed by molecular characterization studies, where mutations of the isocitrate dehydrogenase gene and the codeletion of chromosomes 1p and 19q are of central interest. These are strongly associated with other biochemical characteristics, which contribute to the clinical picture. For example, the subtypes were also recapitulated through analysis of methylation profiles and protein expression.

Since the novel six-class classification scheme is based on the molecular characterization of tumors, they could also be identifiable by methods that can detail the composition of materials. Indeed, the motivation for using Raman spectra comes from studies detailing the classification of materials using deep learning on Raman spectra. Deep learning is thought to be especially well suited for the task since it can detect very subtle patterns in the data. Moreover, depending on the exact architecture of the model, there may or may not be a need for preprocessing steps like baseline correction and principal component analysis. On the other hand, dimensionality reduction in the form of principal component analysis, for example, can be practical for other reasons.

Raman spectroscopy is based on the Raman scattering, where incident light is converted into a characteristic wavenumber spectrum for the material at hand depending on the composition[2]. The data at hand was Raman spectra consisting of intensity sampled 1738 regularly spaced wavenumbers. The data was collected by scanning a given biopsy at regular intervals. This amounted to 141582 samples collected from 59 patients. Since spectra from only 59 distinct tumors were collected, the possibility of not having enough data is real. This difficulty is pronounced for classes that are underrepresented in the data.

I was provided with a script based on a convolutional neural network, which I set out to understand step by step. This wasn’t limited to understanding what was implemented, but also why, what other options are available and how they might turn out. The speculation resulted in a rather unproductive state, paralysis by analysis, when I understood the multitude of options for dealing with the data. This, in turn, resulted in further speculation about how one could sample the space of all possible things that can be done with the data, in an attempt to find a global peak in the model landscape. Such considerations were hardly within the scope of my rotation, since I was really new to modelling, but it led to some interesting reflection. Eventually, I was convinced that the main problem the project faced was the lack of data. The need for synthetic data was discussed on one of the seminars, along with a solution, generative adversarial networks, which seemed quite promising indeed.

Coding toolkit

In order to start implementing my own modelling projects, I needed to become more comfortable with coding. In preparation for my rotation, I learned the very basics of coding in Python on an online course and applied some machine learning concepts in code in the Building AI course by the University of Helsinki. However, that course only required modification of existing code at the intermediate level that I completed it. The advanced level was too steep.

So, I set out to become a better coder. I'm not a fan of big textbooks, especially if one is assumed to work through it from beginning to end. Thus, I searched the internet for shorter, more specific courses, and found the data science platform Kaggle extremely helpful. On Kaggle, I completed the courses Intro to Machine Learning, Intermediate Machine Learning, Data Cleaning, Feature Engineering and Pandas. I didn't learn the material included in these courses by heart, but now I know what tools exist. A tangible result of completing these short courses was that I was then able to complete the Building AI course on the advanced level, which was too steep only weeks before.

On this coding front, the mission may well be completed in the sense that from here onwards, I can develop my coding skills in parallel with doing modelling projects from beginning to end. Perhaps this pattern will hold for other areas of my learning path as well. This could mean that there will be a point where different aspects of modelling will complement each other and constitute a rising tide of parallel learning.

Mathematics toolkit

Many modelling concepts can be understood without math, so the value of learning concepts through formalized mathematical reasoning was not immediately obvious to me. It doesn't help that there are so many libraries and easy-to-use tools for implementing the concepts: there is no need to apply the math in code.

Yet, at the beginning the Foundations of Machine Learning course, I glimpsed how understanding the mathematics can support the intuitive understanding and be much more precise. This precision then allows for communication of the concepts to persons who know the mathematical notation and logic. And even if there are libraries full of tools for modelling, they do not allow for the same freedom in tweaking the algorithms, if need be. Another benefit could be that the precision inherent to mathematics might inform the interpretation of the model, data or phenomenon at hand.

During the rotation, I learned just enough math follow the Foundations of Machine Learning course. However, to develop proficiency and fluency, one needs to navigate and apply the mathematical concepts. Thus, I will start filling in the theoretical background of what I've learned during the internship by studying linear algebra, calculus, statistics and probability. I'm hoping that learning the math in the context of this data science journey will make the math salient enough to encourage understanding the material and how one thing follows from the other.

Future directions

I’m at a branching point of sorts: do I want to focus on data science, as applied to biology, or should I keep exploring the interface between biology and mathematics more broadly? Given that my motivation for exploring the interface between biology and mathematics was to understand biological processes better, interpretability is important to consider. Unfortunately, predictive power and interpretability are typically at odds. Thus, there is room for reflection concerning future directions, especially regarding interpretability. At the most interpretable end of the spectrum are mechanistic models[3]. Developing mechanistic models also seems satisfying as it is theory-based. Speaking about the interpretability of mechanistic models might be redundant, as it is the interpretation of the phenomenon, or theory, that informs a mechanistic model. The reliance on domain knowledge could make the experience of developing mechanistic models more immersive and specialized, as one needs to know the literature inside out. This could also be a downside. For example, one may feel disconnected from the rest of the scientific community if one ends up developing mechanistic models in a very niche area of research. Another potential downside is that mechanistic models are less likely to benefit from the massive amounts of data. In a sense, one is missing out on possibilities granted by the modern high-throughput methods.

All in all, mechanistic modeling sounds like a more creative process in comparison to the empirical or phenomenological approach, where one has a set of tools available that can be thrown at the problem. In principle, one can end up at an optimal empirical model for the application at hand without any domain knowledge. Not relying on domain knowledge could mean less specialization. Less specialization in turn means that one’s project can be communicated more easily, and there is more material to fall back on if one is struggling. As far as job opportunities go, the predictive power of empirical modeling is hard to argue with. In a nutshell, perhaps one could say that building empirical models is more like a craft than an art.

I used to place modelling on a pedestal, thinking of it as some kind of creative wizardry that makes something of all the reductionistic biochemistry research, which seemed like a necessary evil. This was an unnecessarily harsh judgement. An unexpected result of this reflection was a newfound appreciation for creativity in biochemical research. Biochemical research can be very creative in that one has to dig very deep into the subject matter and combine very specific facts. Moreover, there is no toolkit available for these very specific situations, whereas in empirical modelling, there are distinct options as to what can be done with the data. Mechanistic models might strike the right balance for me, as it is quantitative but seems to allow for more creativity. On the other hand, empirical and mechanistic models can be combined, either in series or in parallel. In series, an empirical model may be used to estimate parameters for a mechanistic model. In parallel, an empirical model tries to predict the residuals that need to be added to optimize a mechanistic model. I will have to keep testing the waters and see where I end up.

References

1. Ceccarelli et al. (2016). Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. Cell, 164(3): 550-563
2. Liu et al. (2017). Deep convolutional neural networks for Raman spectrum recognition: a unified solution. Analyst, 142: 4067-4074
3. Baker et al. (2018). Mechanistic models versus machine learning, a fight worth fighting for the biological community? Biology Letters, 14(5)