Andreas Economou: Developing A New Artificial Intelligence-based Diagnostic Approach To Ulcerative Colitis And Crohn's Disease

Introduction

With the help of the Pathological Society Undergraduate Bursary I was able to undertake an 8 week summer research internship with the Elizabeth Soilleux Pathology Research Group based in Addenbrooke's Hospital Cambridge. Due to the coronavirus outbreak the project had to be undertaken online since all members of the group needed to return to their home countries. Regardless, since the project was already computer based and focused on bioinformatic analysis, research was carried out as normal with meetings carried out through online platforms such as "Microsoft Teams". When I joined the team, the group had been developing a new artificial intelligence based algorithm classifying patient intestinal T cell receptor repertoires using machine learning methods in order to identify patients with coeliac disease. This method intends to replace the current classical microscopy used in many areas of diagnosis. The main advantages would be that this method does not require patients to eat gluten prior to testing which increases the acceptability by patients and additionally such a method can potentially meet the currently unmet need for a more accurate and objective diagnostic test. I was immediately paired up with my supervisor, a bioengineering PhD student, who guided me throughout the internship. My main tasks involved assisting in analysing data generated at the lab from coeliac patients and developing and running improved algorithms in order to enhance the machine learning methods the group had already developed in order to cluster TCR repertoires and distinguish coeliac disease patients from healthy volunteers.

Coeliac Disease algorithm basis and mechanism of action

Adult coeliac disease testing strategies include serology for anti-tissue transglutaminase and anti-endomysial antibody, and histopathological examination of duodenal endoscopic biopsies. Biopsy examinations by pathologists are very subjective, with poor interobserver concordance, variable concordance with serology, and a high rate of "equivocal" biopsies. Both serology and endoscopic biopsy require patients to eat appreciable amounts of gluten for 6 weeks prior to testing to avoid false negative or equivocal results, meaning that many gluten-sensitive patients choose not to seek testing due to the unpleasant symptoms that follow gluten ingestion. These led to the need for a new method of diagnosis.

The TCR repertoire(TCRR) refers to the range of different TCRs expressed. By somatic recombination, random insertion, deletion and substitution, the small set of TCR genes can create an enormous number of unique TCR clonotypes. TCRs are heterodimers of TCR-αβ and TCR-γδ type. A randomly selected and recombined variable (V) and joining (J) segment encode the antigen binding region of TCR α-and γ-chains, while TCR β-and δ-chains are encoded by randomly selected and recombined V, J and diversity (D) regions. The TCRR is shaped by previously encountered antigens. The most variable part of the TCR, encoded by the V(D)J junction, known as the complementarity determining region 3 (CDR3) is critical in determining antigen specificity and can be used as a genetic 'barcode' to detect, track and analyse T-cells. Bulk sequencing of TCRRs is an important research tool, producing large datasets, but there are few machine learning algorithms for diagnosing immunological conditions from TCRRs, none in clinical use.

The group developed an algorithm capable of diagnosing coeliac disease, regardless of gluten consumption, on the basis of TCRR in duodenal biopsies. The approach is based on the hypothesis that there are multiple related TCR sequences with similar specificities (capable of binding gluten and possibly self-antigens), both within a single patient's TCRR and between patient TCRRs with coeliac disease. The final algorithm function includes: TCR sequences are translated into amino acids and broken into overlapping kmers (amino acid sequences of length k). These are positionally annotated by which third of CDR3 they derive from (start/ middle/ end). Then a matrix is compiled containing the frequency of each kmer in each patient sample. The matrix dimensionality is reduced by principal component analysis (PCA) and finally samples are clustered using the combination of principal components (PCs)1-10 which had the highest accuracy and gave the greatest separation between groups.

My specific involvement and progress

Since most of the algorithm discussed above was already written in python, I had to familiarize myself with this programming language and learn how to write code in order to contribute to the development and modification of the algorithm. Hence as a preparation for the internship, I taught myself(under the guidance of my supervisor) how to read and write basic code in python and also learned to interpret the huge dataset that had already been collected by the team and which was getting analysed by the algorithm. This dataset included the amino acid sequences of CDR3 regions of TCRs along with their frequencies of occurrence for a group of patients with coeliac disease and for a group of normal healthy individuals. Then, my first task was to write a code in python which would use the aforementioned dataset, extract the relevant information (the amino acid sequences of CDR3s along with their frequencies), break those down into all possible overlapping kmers whilst retaining the frequencies for each kmer and finally transform the calculated data into a matrix whereby each column would represent either a single patient sample or a single healthy sample, and each row would represent a particular kmer retrieved from the dataset. Hence, the final boxes within the matrix, had the frequency of occurrence of a particular kmer in a particular patient or healthy sample (these final frequencies could then be clustered and compared using principal component analysis in search for differentiation markers between patient and healthy samples). My next task involved modifying the algorithm so that it could distinguish between kmers found at the start, middle or end of a particular CDR3, and hence reproduce the matrix, but this time having positional sensitivity. This could help determine at later stages whether particular kmers with an epitope specificity eg specificity to gluten epitopes, where predominantly found at particular positions of the CDR3 sequence. Another part of my project involved classifying the kmers produced based on the most predominant amino acid features amongst the amino acids comprising them, eg being mainly hydrophobic, or charged or polar, and then performing ttests in search for differences amongst predominant patient kmer features and predominant healthy volunteer kmer features. However, this analysis did not yield significant results.

Covid impact

Around the half mark of my summer internship, the head of the research group managed to get access to a large dataset of TCR information from COVID patients, which meant that we could continue our research and try to apply our algorithm on the COVID dataset in search for a new, faster and reliable diagnostic method for COVID. Whilst some members of the group were getting promising results applying the aforementioned algorithm onto the COVID dataset, my supervisor's attention shifted to trying to implement other already published machine learning algorithms with comparable functions on this dataset in order to enhance the results the team was getting and potentially introduce new and improved clustering possibilities. My next task was to get familiar with the documentation for GLIPH(grouping of lymphocyte interactions by paratope hotspots) and GLIPH2 which are published algorithms for clustering TCRs that are predicted to bind to the same MHC-restricted peptide antigen. After lots of experimentation using these algorithms, I finally managed to analyse the COVID datasets using the code, and generated a set of interesting results including a table of enriched kmer motifs that were found to exist at particularly higher frequencies in COVID samples compared to reference datasets. These final tables were produced right at the end of my summer internship but my work is now being continued by another member of the research group in order to generate additional clusters using principal component analysis and contribute further results to the ones found using the aforementioned group algorithm.

Personal development and value of this elective

This research project has been an incredible experience that enriched both my scientific knowledge as well as my skillset regarding research. Firstly, it has been a great opportunity for me to learn coding and how that applies to bioinformatic analysis in the context of analysing and transforming clinical datasets in order to test hypotheses based on medical knowledge. This particular project served also as a revelation for me, since for the first time, I finally saw the real applications of medical knowledge I had accumulated in my first two preclinical years of medicine, eg the importance of understanding diversity in TCR formation. During the project, I learned how to problem solve on my own, interpret and use information from published papers and from the web(eg to solve a coding error), but also, when to ask for guidance and advice from my supervisors and how to benefit from constructive criticism. At a particular stage of the project, I needed to run a very complicated algorithm on a huge dataset which wouldn't normally run on a standard computer. Hence, the head of the group helped me gain remote access to the Cambridge cluster of supercomputers, and indeed after some help, I managed to run days worth of analysis remotely onto those servers. This was an amazing learning opportunity for me, which can be very useful in future bioinformatics projects. The frequent online meetings both with my supervisor as well as with the whole research group, gave me an invaluable insight into the way research is conducted through effective collaboration between researchers from a variety of scientific fields and helped me realise that I want to pursue a career that combines clinical practise with active research. Being an active member of the group, I learned how to present results in lab meetings with confidence and formulate coherent propositions to contribute ideas to group discussion. The opportunity to contribute to active research towards COVID diagnosis at this moment in time really made me appreciate the importance of pushing the boundaries of current medical knowledge. I would like to take this opportunity to thank the Pathological society for encouraging and financially supporting this project, which truly was a fascinating and enlightening experience.