# Industrial CASE Studentship

## Analysis and Prediction of antibody stability

**Keywords**: Antibodies; Protein stability; Machine learning; Data mining; High-throughput biophysics

## Proposed Project

50% of the top 10 grossing drugs in 2014 were antibodies as were ~1/3rd of drugs in development, making them one of the most important classes of pharmaceuticals. However, there are a number of problems in using biologic drugs including cost of development and manufacturing, shelf-life/storage and immungenicity. For example, some molecules are more prone to aggregation than others, some will express more poorly and some are likely to have a greater tendency to degrade or unfold. Prediction of these properties from primary sequence is currently not possible. Ensuring that biologics have high stability is important for improving shelf-life, avoiding cold-chain storage and crucially for reducing immunogenicity.

Working with antibodies has the added complexity of having variable light and heavy chains which must be paired with one another. UCB have recently developed high throughput methods for generating paired heavy and light chain antibody sequences from natural sources, cloning these into Fab fragment expression vectors and expressing at small scale in transient mammalian cell culture. In addition to generating information on any bias in light/heavy chain pairing and in expression yield, the Fab can be rapidly purified and then analysed through a number of small-scale biophysical methods to measure properties including thermal stability, hydrophobicity and surface charge. This linkage between sequence information and biophysical properties will provide a unique opportunity to probe the sequence features which dictate these properties. By the time this project starts, it should be possible to generate data for thousands of antibodies within a 3-month visit by the student. The aim of this inter-disciplinary project is to be able to predict which sequences are likely to have stability issues and to suggest mutations that could enhance stability.

The work will proceed in 4 phases: (1) The student will spend 3 months at UCB to generate paired antibody sequences together with stability data for at least 1000 antibodies; (2) The student will then spend the bulk of the project analyzing these data using machine-learning and data-mining approaches based around the Weka software package. From this we expect to be able to extract specific residues or motifs that are correlated with stability; (3) There are numerous computational methods for predicting the effects of mutations on protein stability (see reviews by Kan & Wihinen (2010) Hum Mutat 31:675-84 and Potapov et al (2009) PEDS 22:553-560). It is our hypothesis that a meta-learner (based on a number of these methods) together with a predictor trained exclusively on antibody data will provide high-confidence stability predictions for antibody sequences and the student will develop and test these ideas; (4) Examples of mutations expected to modify stability will be selected and the student will revisit UCB to generate mutants and evaluate the performance of the predictions.

## Project Impact

The project is potentially of high impact outside the academic environment. As stated above, antibodies have become one of the most important and highest grossing classes of drugs and a huge range of companies (from small biotechs to global pharma) are developing antibody-based

drugs. Being biologics, their development and production is generally more expensive than that of conventional small-molecule drugs; consequently there is huge value in cutting short the development of unsuitable leads early in the pipeline and of being able to improve their biophysical characteristics. The methods outlined in this proposal should enable the prediction of biophysical properties from sequence data, facilitating the early selection of the most stable lead antibodies and the identification of sequences with stability liabilities. Furthermore, the potential will exist to engineer enhanced stability into the sequence of unstable antibodies, enabling the progression and potentially the marketing of antibodies which would previously have not been possible.

We will investigate routes to commercializing any resulting software. The Martin group has an established history of making software available to the academic and non-academic communities and currently makes the abYsis database available over the internet (www.abysis.org) and for in-house use by companies via UCL Business.

Other than proprietary UCB data, all datasets created and used in the project will be made available via the web. Appropriate data management strategies (including on-site and off-site backup) will be in place throughout the research project. Where established standards are available for sharing similar data (e.g. ProTherm, www.abren.net/protherm/), the applicability of these formats to our data will be investigated. Any available current guidance on data formats and metadata and information on best practice will be employed.

Methods developed will be published in relevant journals and made available as servers via the web. This will be done in a timely fashion. All data and servers will be maintained on the web for a period of at least 10 years after the completion of the research project. In addition, where appropriate, the new UCL Research Data Services Data Archive will be used for long-term archiving.

## Academic Research Environment

The Institute of Structural and Molecular Biology (ISMB) is a joint institute between UCL and Birkbeck. Together with UCL's Centre for Mathematics, Physics and Engineering in the Life Sciences and Experimental Biology (CoMPLEX) and the UCL Research Department of Genetics, Evolution and Environment (GEE) it has one of the largest and strongest concentrations of computational biologists and bioinformaticians in the UK. The Martin lab shares research space with three other computational biology / Bioinformatics groups. In the 2014 REF, UCL had the greatest amount of 4* ('world leading') research in medicine and biological sciences.

The ISMB runs weekly seminars and informal presentations by PhD students and post docs. Students also attend CoMPLEX seminars, 'London BioGeeks' meetings and seminars at Imperial and Kings. The ISMB organizes alternating annual 2-day retreats (with PhD student and post doc talks and posters) and symposia (with international scientists and members of the ISMB). Students are encouraged to meet members of other labs via a "Postgraduate Student Society" and the departmental common room. The combination of research excellence, the sense of cohort, interaction with other programmes and the intensity of living in London generates a very vibrant postgraduate student community.

UCL is extremely well equipped for computational biology research. Together with Computer Science, we have dedicated local IT staff who maintain Linux servers and desktop machines and a compute farm of >6000 cores. The Martin group has a local ~30 core processor farm, ~40TB RAID disk space, and a number of disk/database/web servers. We mirror data every night including the PDB and UniProt. UCL also provides a central computing service which includes two HPC compute farms (Legion and Grace) each consisting of ~6000 cores for serial/small-parallel and large-parallel jobs respectively.

## Training and Support at UCL

Dr. Martin will give training related to programming and machine learning and the student can attend sessions on programming that are part of the Birkbeck Bioinformatics MSc. Best practice in programming (revision control using GitHub, well-written and commented code, and test-driven design) are strongly encouraged by Dr. Martin. All students are assigned a 'thesis committee' consisting of the academic and industry supervisors, a secondary academic supervisor and an independent committee chair. These staff are available to support the student at any time as well

as to monitor progress. UCL offers numerous training courses through faculties and the Doctoral Training School. These range from subject-specific training through to general and transferable skills (e.g. Statistics, presentation skills, LaTeX, thesis writing). Students are required to acquire 'Roberts Points' through attending training courses, and seminars and through demonstrating for practicals.

## Training and Support at UCB

As industrial partner, UCB will provide a training program, which reinforces aspects of the UCL training and extends training to additional areas of basic science, as well as skills relevant for the workplace. Guidance on experimental strategy, design and interpretation will be provided to ensure:

- science is hypothesis driven

- the key experiments are performed which drive understanding

- data are robust and reproducible; both from a statistical and biological perspective

- data are effectively integrated and placed into context with the literature

Importantly, the student will also learn how to integrate their data, in a project team environment, with those of other scientists working on different aspects of antibody analysis.

The UCB PhD program will allow them to network annually with the ~30 other PhD students on UCB/academic collaborations. Exposure to other PhD students from top universities across the country provides a broader base to their scientific development than can be offered by their project alone.

## The collaboration

UCB is one of the leading companies working with antibody therapeutics and will provide an excellent opportunity for the student to work in such an environment. They will offer a placement at the start of the project that will allow the student to learn and apply wet-lab experimental high-throughput approaches to generating paired light and heavy chains of antibodies and to obtaining necessary biophysical data. They will also offer a placement towards the end of the project to assess predictions made by the computational analysis of these data. These are clear added value placements as the necessary expertise is not available in Dr. Martin's lab at UCL and the high throughput generation of paired antibodies is a new technology developed recently at UCB.

UCB will offer business-related training. During the 3 month placement the student will experience first-hand the industrial biopharmaceutical workplace. Full training will be given in all safety and compliance aspects necessary to work within the laboratory, and the student will be exposed to all company seminars, group meetings and events as they get a flavour of the industrial research environment. In addition, regular meetings with industrial supervisors will be scheduled after the completion of the placement. UCB has a very good track record of supporting industrial placement studentships and has a large number of concurrent studentships. Annual PhD days are held at which students are expected to present posters of their original research, culminating in an oral presentation in their final year.

UCB offers novel experimental techniques that will allow the generation of a novel dataset that cannot be generated by Dr. Martin's group or obtained in sufficient quantities from other sources to enable machine learning and data mining. UCB will also provide the opportunity to evaluate the performance of predictive methods experimentally – again something that cannot be done in Dr. Martin's group. Dr. Martin brings a huge amount of experience in computational analysis of antibodies, machine learning and analysis of mutations as well as best practice in software development. This expertise will enable the computational aspects of the project.