MSc Bioinformatics
Birkbeck, University of London
Project proposals for full-time and first-year part-time students (2022-2023)
Version 1 - 10/03/2023

Guidelines

1. Read the booklet for the Project Module of the MSc Bioinformatics (uploaded on Moodle).

2. Read the list of projects carefully. You may be surprised by the variety of things on offer, even if you already think you know what you want to do for your project.

3. There are several projects from both internal supervisors (code starting with "I") and external supervisors (code starting with "E"). Many of these projects are available to both full- and part-time students. However, some are specifically for full-time students or for part-time students, and this is stated explicitly at the end of the project description.

4. Projects outside this list are possible, but you need to let me know as soon as possible, the title, description, and supervisor. Projects need to be agreed with me and the course director, Dr Adrian Shepherd. If the supervisor is external, we will assign a second internal supervisor for your project (you may also want to suggest an internal supervisor yourselves). The internal supervisor need not be an expert in the area, but if there is an obvious match in terms of research interests, we will try to involve that person.

5. Project selection:
a) Please talk to the supervisors whose projects you are considering. Many project descriptions are brief and supervisors expect that you contact them before making a decision. Supervisors will have the final say on whether they want you to take on a project.
b) You need to select 3 projects and provide me with the order of your preference. This applies to EVERYBODY, except for people who have made arrangements to do projects outside this list. Do not assume your first selection is "safe" as others may have picked it too.
c) If you select an external project, I reserve the right to let the external supervisor know what progress you have made so far on the MSc (i.e. any exam, coursework and assessed practical marks they may be interested in). This is to allow them to make an informed choice and to ensure that we will keep receiving projects from external supervisors in the future.
d) Internal supervisors will usually be first supervisors to only THREE new students each year (this limit may have to be dropped if external projects do not attract enough interest). External supervisors will normally be allowed to take only ONE student per year.

6. Timeline: Please return your top three choices to me (i.nobeli@bbk.ac.uk) via email no later than the end of FRIDAY 31st OF MARCH 2023. Please quote both the project code and title/supervisor in your emails to me and put "Project choices" in your email title.

6. No projects on this list will be assigned before I have received all choices and resolved all clashes.

7. Clashes will be resolved by a panel of academic staff following consultation with supervisors and students involved.

Any questions, please contact me.

Irilenia Nobeli
i.nobeli@bbk.ac.uk

Project Code: *I_2023_TC_1*

Cryo-EM fitting of flexible systems

First supervisor
Name: Tristan Cragnolini
Affiliation / Position: Birkbeck, Lecturer

Telephone:
Email: t.cragnolini@bbk.ac.uk

Project Description

Cryo-EM experiments of biological macromolecules are interpreted in term of 3-dimensinoal reconstructions and using structural models. In the ideal case, all particles (image of a single macromolecule) correspond to the same conformational state of the studied system. In practice, inherent flexibility means that reconstructed volumes correspond to average of a conformational ensemble.
In this project, we will look at improving the interpretation of the experimental data using conformational ensemble of models for systems of interest. By taking advantage of a recently developed probabilistic formulation of the relation between structural variations and the experimental cryo-EM density,[1] combined with sampling methods, [2] you will refine the interpretation and quality of fit between the underlying structural model and the cryo-EM data, by considering how each model relate and contribute to the reconstructed density.

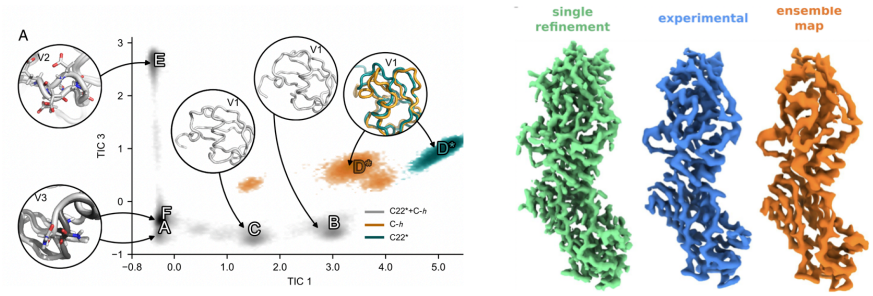


Figure 1: A.The clustering of molecular models can illustrate local conformational differences. B. An ensemble of models produces a better representation of the experimental density.

We will validate the methods developed using deposited data for benchmarking and test them on challenging datasets from intrinsically flexible systems.

[1] Cragnolini, Beton, and Topf, 'Cryo-EM Structure and B-Factor Refinement with Ensemble Representation'.
[2] Robustelli, Piana, and Shaw, 'Developing a Molecular Dynamics Force Field for Both Folded and Disordered Protein States'.

Any particular skills that are required for this project?

Basic Python programming is required, as well as familiarity with the command line. A background in statistics would be useful, but not necessary. All skills required are taught in the bioinformatics program.

Is the project suitable for full-time / part-time students?

Both.

Project Code: *I_2023_TC_2*

Ensemble sampling for structural predictions

First supervisor
Name: Tristan Cragnolini
Affiliation / Position: Birkbeck, Lecturer

Telephone:
Email: t.cragnolini@bbk.ac.uk

Project Description
Understanding the fold of proteins is crucial to understand their function. Next-generation sequencing methods and advance in proteomics and metagenomics have produced very large protein sequence datasets, but the number of solved protein structures has not grown commensurately. Structure prediction methods are a promising tool to bridge that gap: deep learning methods are now able to predict many protein conformations with very high accuracy.[1] Flexible domains, however, are often not predicted or incorrectly predicted. The relative orientation of domains separated by flexible linkers is also often incorrect. Altered pipelines have been proposed to remedy this issue,[2] but highly flexible regions are still a challenge. By combining the contact predictions from deep-learning models and using a constraint-propagation algorithm with molecular dynamics, new conformations can be sampled that have high likelihood, and a larger conformational ensemble can be proposed. The goal of this project is to use this method and assess its application to flexible multi-domain proteins, optimise the algorithm used, notably in term of conformation selection, and potentially incorporate it directly in the deep-learning model generation pipeline. We will validate the methods developed using deposited data for benchmarking and test them on challenging datasets from intrinsically flexible systems.

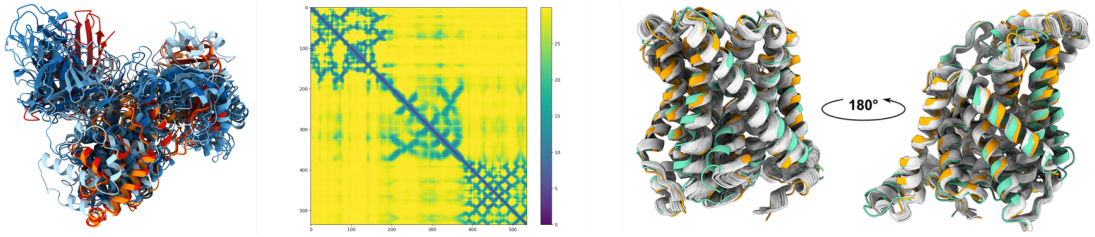


Figure 1: A. Predicted structures (blue) do not correctly predict the known conformers (orange, red) B. Contact map for the same protein. The contacts present in the experimental structures have high likelihood but are not present in the final prediction ensemble. C. Example of the conformational ensemble generated by current pipelines.[2]

[1] Jumper et al., 'Highly Accurate Protein Structure Prediction with AlphaFold'.
[2] del Alamo et al., 'Sampling Alternative Conformational States of Transporters and Receptors with AlphaFold2'.

Any particular skills that are required for this project?
Basic Python programming is required, as well as familiarity with the command line. Some understanding of machine learning packages will be needed. A background in statistics would be useful, but not necessary. All skills required are taught in the bioinformatics program.

Is the project suitable for full-time / part-time students?
Both.

Project Code: *I_2023_TC_3*

Hierarchical coarse-grain modelling

First supervisor
Name: Tristan Cragnolini
Affiliation / Position: Birkbeck, Lecturer

Telephone:
Email: t.cragnolini@bbk.ac.uk

Project Description

All-atom models can describe the dynamics and behaviours of complex systems with very high accuracy, but require large computational resources, and simulations on longer timescales are onerous. Coarse-grained models are excellent tools to study the dynamical behaviour of biomolecules on a larger scale,[1] but are limited in scope by the necessity of parametrising them, often with manual adjustments. Although progress has been made on automating this process,[2] the fundamental problem of defining an exact mapping between different level of description remain unanswered: as of now, there exists no practical alternative to hand-picking groups, based on one's knowledge of the system at hand, as it is often unclear how 'coarse' the grouping should be, and whether there is an optimal grouping. While some work has been done to try and evaluate the quality of a given mapping, those methods are expensive, usually requiring integration over the entire phase space for both the coarse-grained and full-atom system of interest. In this project, we will investigate how coarse-grain mappings at different resolutions relate to one another and run simulations on biomolecular systems using those simplified models.

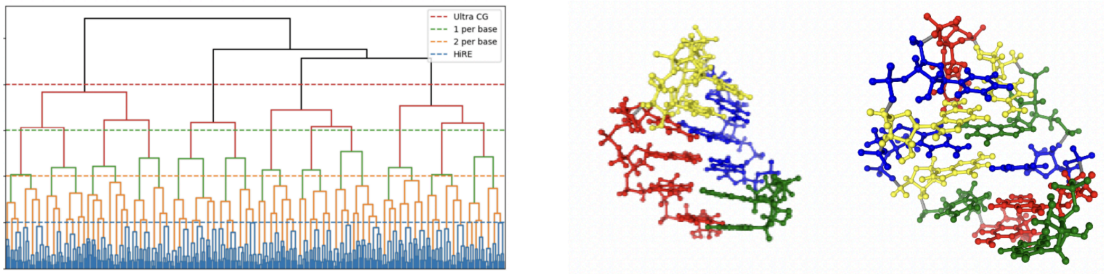


Figure 1: A. Hierarchy of coarse-grained models for RNA, and how they relate to each other when mapped on a small RNA structure. B. Mapping of the ultra-coarse-grain model on an all-atom RNA structure. C. Mapping of the HiRE-RNA model on the sameRNA structure.[1]

[1] Tristan Cragnolini, Yoann Laurin, Philippe Derreumaux, and Samuela Pasquali. Coarse-Grained HiRE-RNA Model for ab Initio RNA Folding beyond Simple Molecules, Including Noncanonical and Multiple Base Pair- ings. Journal of Chemical Theory and Computation, 11(7):3510–3522, 2015.
[2] Lee Ping Wang, Jiahao Chen, and Troy Van Voorhis. Systematic parametrization of oolarizable force fields from quantum chemistry data. Journal of Chemical Theory and Computation, 9(1):452–460, 2013.

Any particular skills that are required for this project?
Basic Python programming is required, as well as familiarity with the command line. Some understanding of machine learning packages will be needed. All skills required are taught in the bioinformatics program.

Is the project suitable for full-time / part-time students? Both.

Project Code: *I_2023_TC_4*

Contact prediction with deep learning in flexible proteins

First supervisor
Name: Tristan Cragnolini
Affiliation / Position: Birkbeck, Lecturer

Telephone:
Email: t.cragnolini@bbk.ac.uk

Project Description

About 30% of eukaryotic proteins are intrinsically disordered, and many more of the rest possess highly flexible loops, and nucleic acids are often highly flexible as well. Properly accounting for these regions is a challenge, both theoretically (due to the large conformational space available to them) and experimentally (where collecting data with a high signal-to-noise ratio is hampered by the heterogeneity within the sample). By using deep-learning predictors for contact prediction (such as transformer models) combined with molecular dynamics, we will build models of those flexible regions, and validate those results against experimental biophysical data.

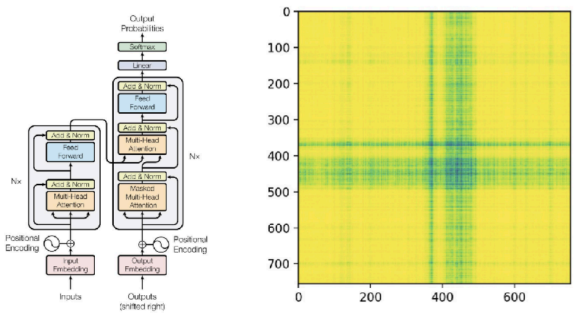


Figure 1: A. A transformer model[1] B. A predicted contact map generated with a BERT-like deep learning model.

[1] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. arXiv December 5, 2017. http://arxiv.org/abs/1706.03762
[2] Gomes, G.-N. W.; Krzeminski, M.; Namini, A.; Martin, E. W.; Mittag, T.; Head-Gordon, T.; Forman-Kay, J. D.; Gradinaru, C. C. Conformational Ensembles of an Intrinsically Disordered Protein Consistent with NMR, SAXS, and Single-Molecule FRET. J. Am. Chem. Soc. **2020**, 142 (37), 15697–15710. https://doi.org/10.1021/jacs.0c02088.

Any particular skills that are required for this project?
Basic Python programming is required, as well as familiarity with the command line. Some understanding of machine learning packages will be needed. All skills required are taught in the bioinformatics program.

Is the project suitable for full-time / part-time students?
Both.

Project Code: *I_2023_FF_1*

In silico characterization of a newly identified lipase in plants

First supervisor
Name: Franca Fraternali
Affiliation / Position: UCL/Birkbeck Professor in Integrative Computational Biology

Telephone:
Email: f.fraternali@ucl.ac.uk

Second supervisor
Name: Prof. Salvatore Adinolfi
Affiliation:University of Turin, Italy
Email: salvatore.adinolfi@unito.it

Project Description

Please insert your project description here (half a page to two pages are acceptable lengths; you may include figures and references)

Hazelnuts (*Corylus avellana* L.) are characterized by high fat content, with triacylglycerols as the main components. Quality retention during the marketing of raw and processed hazelnuts is still a common problem which needs to be solved on an industrial level. Indeed, while the relatively high abundance of mono- and unsaturated fatty acids gives this nut a high nutritional value, but it leads also to a considerable susceptibility to autoxidation and/or degradation via chemical and/or enzymatic reactions. Indeed, lipid rancidity in natural products is a combination of the activity of two main enzymes: lipase and lipoxygenase, which lead to hydrolytic rancidity and oxidation, respectively [1,2,3].
Lipases are present in reserve tissues of many oilseed plants and nuts where they mostly contribute to post-germination oil reserve mobilization; they are either not expressed or inactive in resting and intact seeds [4,5,6,7]. Lipase action results in the release of long-chain fatty acids, generated by hydrolysis of tri-, di- and monoacylglycerols [8]. These free fatty acids can be oxygenated by lipoxygenase or by autoxidation to form hydroperoxides that are further metabolised to yield off-flavour volatile compounds, many of which are associated with hydrolytic rancidity [9]. These enzymatic and/or chemical oxidations occur with faster kinetics on free fatty acids with respect to those esterified with triacylglycerols [10]. This issue is particularly relevant in the case of hazelnuts, due to the high amount of mono- and polyunsaturated fatty acids in the kernel. Indeed, an increase in free fatty acids has been associated with the increase of acidity in hazelnuts, the main contributor of rancidity. At the same time, storage, water activity (a_W), and temperature [11] have a clear and decisive impact on the fatty acids' chemical stability [11].
The hazelnut (*Corylus avellana)* seed's putative lipase has been purified in the lab of Prof. Adinolfi (University of Turin) and its biochemical and biophysical characterization has been carried out there as well. The characterisation of the protein for its structure-function and molecular interactions will be performed in the Fraternali laboratory at UCL/Birkbeck (London, U.K.).

The core of this project will be the in silico investigation of this putative lipase protein, in particular to elucidate a molecular enzyme-ligand complex to explain the lipase activity observed in vitro. The structure of the *Corylus avellana* hazelnut is

available as Alphafold model and by super-imposition with similar structures containing lipid substrates one can infer their biding properties to lipids and fatty acids. In order to characterise lipase activity, a computational analysis of possible catalytic triads forming the active site will be performed. Molecular dynamics simulations will be performed to optimise the geometry of such sites and to evaluate the overall complex stability.

1) Belitz, H.-D.; Grosch, W.; Schieberle, P. *Food Chemistry*; Springer: Berlin, Germany, 2013; ISBN 9783540699330.
2) Frega, N.; Mozzon, M.; Lercker, G. Effects of free fatty acids on oxidative stability of vegetable oil. *JAOCS J. Am. Oil Chem. Soc.* **1999**, *76*, 325–329.
3) Seyhan F., Tijskens L.M.M., Evranuz O. (2002). Modelling temperature and pH dependence of lipase and peroxidase activity in Turkish hazelnuts. Journal of Food Engineering, 52, 387-395.
4) Huang, A. H. C. (1984). In B. Borgstrom, & H. L. Brockman (Eds.) Lipase: Plant lipases (pp.419–422). Amsterdam: Elsevier Publishers.
5) Sanders, T. H., & Pattee, H. E. (1975). Lipids, 10(1), 50–54.
6) Abigor, R. D., Uadia, P. O., Foglia, T. A., Haas, M. J., Scott, K., & Savary, B. J. (2002). Journal of the American Oil Chemists Society, 79, 1123–1126.
7) Haas, M. J., Cichowicz, D. J., & Bailey, D. G. (1992). Lipids, 27, 571–576.
8) Casas-Godoy, L., Gasteazoro, F., Duquesne, S., Bordes, F., Marty, A., & Sandoval, G. (2018). Lipases: An overview. *Methods in Molecular Biology, 1835*, 181–238.
9) Doblado-Maldonado, A. F., Pike, O. A., Sweley, J. C., & Rose, D. J. (2012). Key issues and challenges in whole wheat flour milling and storage. *Journal of Cereal Science, 56*, 119–126.
10) Frega N., Mozzon M. and Lercker G. (1999) Effects of Free Fatty Acids on Oxidative Stability of Vegetable Oil. *JAOCS*, 76, 325-329.
11) Savage, G.P.; McNeil, D.L.; Dutta, P.C. Lipid composition and oxidative stability of oils in hazelnuts (Corylus avellana L.) grown in New Zealand. JAOCS, J. Am. Oil Chem. Soc. 1997, 74, 755-759.

Any particular skills that are required for this project?
The student should have strong inclination towards protein structure and their investigation via structural bioinformatics and evolutionary analyses.
The project would require scripting (phyton-based ideally but other programming languages are also possible). The student will be trained in the use of docking programs and molecular dynamics simulations.

Is the project suitable for full-time / part-time students?
Both

Project Code: *I_2023_CM_1*

Particle picking optimisation and evaluation for cryo-EM data using deep learning-based approaches

First supervisor
Name: Carolyn Moores
Affiliation / Position: Birkbeck/Professor of Structural Biology

Telephone: 020 3926 3516
Email: c.moores@bbk.ac.uk

Second supervisor (if applicable)
Name: Tianyang Liu
Affiliation: Birkbeck/Postdoctoral Researcher
Email: t.liu@bbk.ac.uk

Project Description

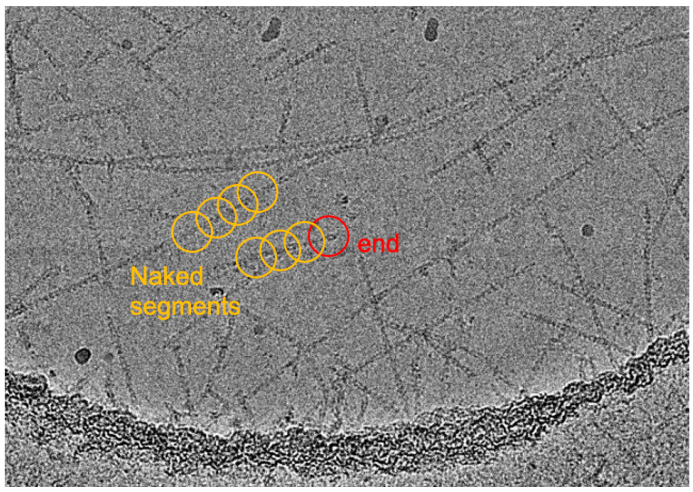
Actin filaments are key components of the cytoskeleton and play an important role in many cellular processes from cell motility to phagocytosis. Actin filaments can polymerise and depolymerise at two ends. Many actin end binding proteins regulate polymerisation dynamics, control the length of actin filaments and thus influence cellular processes. For example, Capping protein can tightly bind to the barbed ends of actin filaments to block their growth[1]. Spin90 and Arp2/3 together can nucleate an actin filament from its pointed end[2].

In order to understand the mechanisms by which these actin ending binding proteins interact with ends and regulate their dynamics, we use cryo-electron microscopy(cryo-EM) and single particle 3D reconstruction to determine the structures of actin filament ends with binding partners[3]. The workflow involves in vitro reconstitution of actin filaments in the presence of binding partners, plunge freezing of the sample to maintain its near-native state, data collection using an electron microscope and data processing of the noisy micrographs to reconstruct the 3D structure of filament ends.

One of the main bottlenecks in data processing is the accurate selection of actin filament ends from cryo-EM micrographs while ignoring various types of background noise, contaminants and non-end portions of actin filaments. Accurate target picking is crucial for the success of the 2D/3D data processing. Traditional particle picking includes manual picking and the use of templates, but manual picking is impractical given that the size of modern cryo-EM datasets are often > 10,000 micrographs. In addition, template picking based on cross correlation between template and micrograph is often not sufficiently accurate given the high rate of false positives in many samples.

Several automatic picking packages based on deep learning have been developed for cryo-EM data over the past few years such as TOPAZ[4] and crYOLO[5]. Although these deep learning-based packages have been reported to improve the accuracy and increase the speed of cryo-EM particle picking, they utilise different approaches. The focus of this project is 1) to compare the accuracy and efficiency of all picking

packages in terms of picking actin filament end structures from our existing datasets 2) to generate a particle picking workflow to maximise the accurate picking of actin filament end. 3) to integrate the deep learning- based particle picking into our data processing pipeline.



1 Funk, J. *et al.* A barbed end interference mechanism reveals how capping protein promotes nucleation in branched actin networks. *Nat Commun* **12**, 5329, doi:10.1038/s41467-021-25682-5 (2021).
2 Luan, Q., Liu, S. L., Helgeson, L. A. & Nolen, B. J. Structure of the nucleation-promoting factor SPIN90 bound to the actin filament nucleator Arp2/3 complex. *EMBO J* **37**, doi:10.15252/embj.2018100005 (2018).
3 Milne, J. L. *et al.* Cryo-electron microscopy--a primer for the non-microscopist. *FEBS J* **280**, 28-45, doi:10.1111/febs.12078 (2013).
4 Bepler, T. *et al.* Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs. *Nat Methods* **16**, 1153-1160, doi:10.1038/s41592-019-0575-8 (2019).
5 Wagner, T. *et al.* SPHIRE-crYOLO is a fast and accurate fully automated particle picker for cryo-EM. *Commun Biol* **2**, 218, doi:10.1038/s42003-019-0437-z (2019).

Any particular skills that are required for this project?

• Skills in using Linux commands
• Scripting

Is the project suitable for full-time / part-time students?

Both

Project Code: *I_2023_IN_1*

Towards the development of a simple fingerprint representation for the efficient comparison and clustering of large datasets of RNA sequences

First supervisor
Name: **Irilenia Nobeli**
Affiliation / Position: **Birkbeck, Biological Sciences (Lecturer)**

Email: **i.nobeli@bbk.ac.uk**

Project Description

Background & Aim

Interest in predicting RNA structure and classifying RNAs into families has increased considerably since high throughput methods showed that a large part of mammalian genomes is transcribed but does not code for proteins. Non-coding RNAs are likely to play a multitude of regulatory roles, especially in organisms of higher complexity, and their function will most certainly be tightly linked to their structure. Just as the protein world has benefited enormously from structure classification schemes, it is easy to imagine that having some method for clustering RNAs into structural families would help with computational function annotation of the thousands of sequences that are currently in databases and the many more that are likely to be added over the next few years. Indeed, the RFAM database [Kalvari et al. 2018], built on the principle that sequences of the same family can be represented with probabilistic models analogous to protein family Hidden Markov Models (covariance models), has contributed hugely to the difficult task of clustering the RNA universe. However, it is clear that many known RNA transcripts do not match any existing families and that for some, it is probably impossible to build multiple sequence alignments that are informative enough to result in useful models of the family. Perhaps more problematic is the expectation that a family has a characteristic secondary structure that reflects the "peak probability" of the covariance model. This assumption overlooks the fact that many RNAs will fold into mutually exclusive alternate structures, depending on the environment and interactions with other molecules. In this sense, they are fairly different to proteins, as the latter will generally keep their overall fold and only undergo local conformational changes in the presence of ligands or changes in the environment (this is an over-simplification of the conformational flexibility of proteins but RNA is certainly more susceptible to large changes affecting its overall fold). Thus, attempting to describe RNA families using single representative structures and models that do not explicitly encode mutual exclusivity is likely to have its limitations.

The aim of this project is to explore the feasibility of a novel approach for clustering/classificying RNA sequences that does not assume a single structure per family nor requires expensive calculations. The method borrows the basic idea of "path-based" (or hashed) fingerprints from chemoinformatics and adapts them to the description of RNA structure so that comparison and clustering of RNA sequences can be done efficiently for very large numbers of sequences, without the need for time-consuming energy minimisations or family-based covariance models.

Methods

In the method proposed here, we attempt to address two major issues in structure-based clustering of RNA:

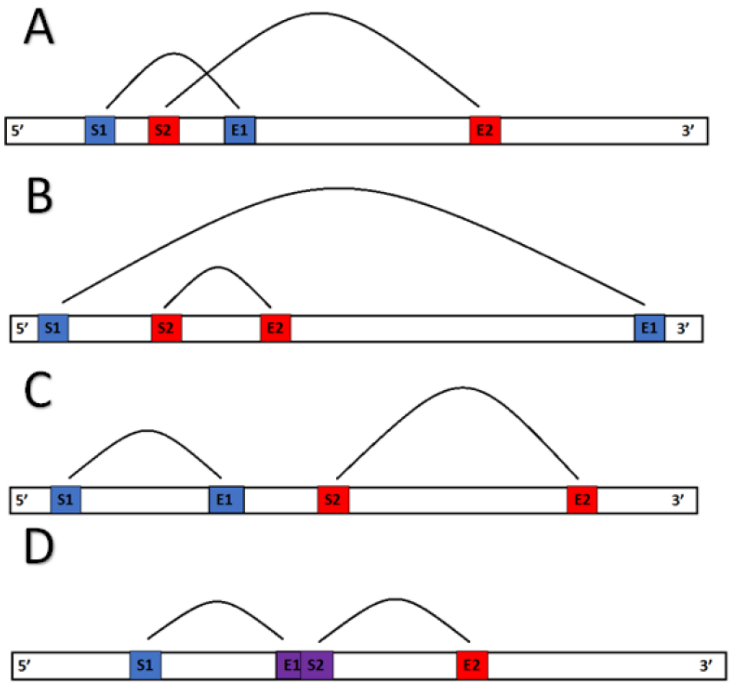
a) avoiding the RNA global folding problem (as results are anyway dubious for sequences larger than about 50 nucleotides) and

b) allowing multiple RNA conformations to be implicitly considered, even if some of these may not be significantly populated under normal conditions (but that could become important, for example, in the presence of a ligand, at a raised temperature etc).

We propose to achieve this by representing an RNA sequence with a binary array (fingerprint) encoding palindromes in the sequence and their relationships (essentially recording the most likely base-pairing possibilities within the sequence and how the base pairs relate to each other; see Figure 1). The number of possible palindromes in a random sequence increases very fast with the length of the sequence so the biggest challenge of this project is how to avoid finding all palindromes and how to prune the ones retrieved to generate a useful fingerprint. Here, we will adopt an empirical approach, whereby preliminary analysis of large sets of RNA sequences will be used to deduce appropriate filters for achieving efficiency in encoding with limited loss of information.

The project will take advantage (at least initially) of existing software for finding palindromes (Brazda et al. 2016) and stems (Kingsford et al. 2007) and will concentrate on a) setting rules for finding and encoding palindromes and b) building a prototype that can be tested on a few RNA families. The work involved is most likely beyond what can be achieved in the time of a full-time Masters but reaching some form of conclusion on whether the idea is worth implementing seriously would be considered a successful outcome. Someone with strong programming skills might be able to take this project to completion but I think this would be suitable for any student who has an interest in algorithms (regardless of programming expertise) and no aversion to risk.

Figure 1. The relationship of hairpins in an RNA sequence



Schematic diagram of an example RNA sequence comprising two hairpins (blue and red) and their possible arrangements. For each hairpin a start (S) and an end (E) point are shown. A) Overlapping hairpins: either the start of the end point of a hairpin falls within the other hairpin. B) Inclusive: one hairpin is totally encompassed by another. C) Separate: the hairpins do not

overlap. D) Exclusive: the hairpins overlap in a way that makes it impossible for both of them to form simultaneously.
Diagram drawn by Ajayi Dolapo (Birkbeck, MSc Bioinformatics, 2017).

References

Brazda, V., Kolomaznik, J., Lysek, J., Haronikova, L., Coufal, J. and Stastny, J. (2016) Palindrome analyser - A new web-based server for predicting and evaluating inverted repeats in nucleotide sequences. Biochem Bioph Res Co, 478, 1739-1745.

Kalvari, I. et al. (2018). Rfam 13.0 shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Res. 47: D335–D342.

Kingsford, C., Ayanbule, K. and Salzberg, S.L. (2007). Rapid, accurate computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. Genome Biol. 8:R22.

Any particular skills that are required for this project?

Programming in any language (but python is probably more suitable). An interest in RNA and algorithms would be helpful.

Is the project suitable for full-time / part-time students?

Both

Project Code: *I_2023_IN_2*

Towards improving the annotation of functional elements in species of the *Mycobacterium tuberculosis* complex: A web-based explorer of unannotated expressed regions

First supervisor
Name: **Dr Irilenia Nobeli**
Affiliation / Position: **Birkbeck**

Email: **i.nobeli@bbk.ac.uk**

Second supervisors (if applicable)
Name: **Jennifer Stiens**
Affiliation: **Birkbeck (PhD student in the Nobeli group)**

Name: **Dr Sharon Kendall**
Affiliation: **Royal Vet College**
Email:

Project Description

Although the human pathogen *Mycobacterium tuberculosis* and very closely-related species such as *Mycobacterium bovis* have been studied extensively, their genome annotation is still very poor for anything but coding regions of the genome (even those regions are often annotated simply as “hypothetical proteins”). The apparently non-coding parts that show clear expression in RNA-seq studies (such as short RNAs, untranslated regions of genes and antisense RNAs) are almost always absent from annotation files accompanying the genome assemblies (Stiens et al., 2022a).

Our group has a long-standing interest in using computational approaches to annotate all expressed elements in mycobacterial genomes and as part of this effort, we have created *baerhunter*, an R-based software for the prediction of potentially functional, non-coding elements in bacterial genomes from NGS transcriptomic data. More recently, we have used baerhunter to analyse a number of RNA-seq studies of *Mycobacterium tuberculosis* under a variety of conditions and have shown that several non-coding elements show differential expression under such conditions, often forming clusters with genes of known function, based on their co-expression across different studies (Stiens et al., 2022b).

In this project, we would like to make findings from our recent work easily accessible to other groups by producing a web-based resource for exploring the annotation (official and non-official from the literature) of both coding and non-coding regions in *M. tuberculosis* and related species. Some steps towards achieving this goal have been accomplished by previous MSc students. Thus, the project will concentrate on organising data we already have into a database and creating a web-based explorer that will include a genome browser, a search function and links to external resources of information. At the same time, we are developing a new program that identifies highly expressed regions that overlap (on the same strand) with known genes, something that baerhunter cannot do, and plan to use this program on nanopore data from *Mycobacterium smegmatis* (produced by Dr Kendall’s and Dr Shan Goh’s lab at the University of Hertfordshire) to identify potential short RNAs produced from longer transcripts, or expressed through different promoters in the same locus. Thus, additional data may become available during the time of the project to enrich the content of the resource. Finally, a very motivated student could do some additional

research into the possibility of some “non-coding” elements being instead short ORFs producing small peptides. The Riboseq data from the Cortes and Wade labs (Sawyer et al., 2021; Smith et al., 2021) could be used to check the evidence supporting the presence of short ORFs among these expressed elements.

References

Sawyer et al. (2021). A snapshot of translation in Mycobacterium tuberculosis during exponential growth and nutrient starvation revealed by ribosome profiling. Cell Reports 34 (5), 108695. Doi: https://doi.org/10.1016/j.celrep.2021.108695.

Smith et al. (2021). Pervasive translation in Mycobacterium tuberculosis. BioRxiv, doi: https://doi.org/10.1101/665208.

Stiens et al. (2022a). Challenges in defining the functional, non-coding, expressed genome of members of the Mycobacterium tuberculosis complex. Mol Microbiol. 2022 Jan;117(1):20-31. doi: 10.1111/mmi.14862. Epub 2021 Dec 27. PMID: 34894010.

Stiens et al. (2022b). Using whole-genome co-expression networks to inform functional groupings of hypothetical conserved and predicted genomic elements from Mycobacterium tuberculosis transcriptomic data. BioRxiv: https://doi.org/10.1101/2022.06.22.497203. *This paper has just been accepted for publication in Molecular Microbiology.*

Any particular skills that are required for this project?

Skills taught on the course. Anything extra will be supported by the supervisors.

Is the project suitable for full-time / part-time students?

Both

Project Code: *I_2023_IN_3*

Towards a web-based hub for the respectful reporting of scientific studies on autism

First supervisor
Name: **Dr Irilenia Nobeli**
Affiliation / Position: **Birkbeck**

Email: **i.nobeli@bbk.ac.uk**

Project Description

The big picture

The availability of "omics" technologies has allowed great steps to be taken towards our understanding of the genetic and molecular basis of autism and its comorbidities. However, scientific reports on studies of autism are often met with hostility by the autistic community, primarily because many scientists treat autism as a disease that must be cured – understandably, high-functioning autistic people that are engaged with such research find this approach offensive. In this long-term project, we aim to help bridge the two communities and share our understanding of the science with the wider public in a way that is acceptable to all stakeholders.

Specifics of this project

In this MSc project, we want to harness computational approaches for collecting data from public resources and creating a portal where scientific facts can be collated and linked. To narrow the focus of this particular project, we will be aiming to set up a resource for linking genes/transcripts/proteins to literature findings, with the goal of creating networks of knowledge that are specific to regulation of gene expression in autism and its comorbidities. I would expect the results of the project to be made available via a web-based hub that can be explored by both scientists and the lay public.

Who is this project suitable for?

This is an open-ended project and the specifics are not very clearly defined yet. Hence, the student taking this on would be expected to display initiative, independence, enthusiasm and competence in exploring approaches towards achieving the overall aim. There is no need to come up with a working website but an exploration of what can be achieved with current tools and some sort of prototype would be a great start for this project that is likely to run for a long time. Clearly, an interest in neurodiversity and the molecular mechanisms governing it would be key here. There may also be an opportunity to get involved in a side project that aims to disseminate autism research using the help of a comics illustrator and script writer.

Any particular skills that are required for this project?

Skills taught on the course - possible extras can be learned along the way.

Is the project suitable for full-time / part-time students?

This project is available to both FT and PT students but PT students would be more suitable because the details of the project are not yet well-defined.

Project Code: *I_2023_KOT_1*

Development of an Integrated Software Platform for the Analysis and Dissemination of Travelling Wave Ion Mobility Data

First supervisor
Name: Konstantinos Thalassinos
Affiliation / Position: ISMB at UCL / Birkbeck – Senior Lecturer

Telephone: 02076792197
Email: k.thalassinos@ucl.ac.uk

Project Description

Mass spectrometry (MS) and ion mobility coupled to mass spectrometry (IM-MS) are powerful analytical techniques that are increasingly used by biochemists to study protein conformation and dynamics, especially when these proteins are not amenable to analysis by more established methods such as X-ray or NMR. Using MS one can identify the stoichiometry of protein complexes, determine the topology of these complexes and, with the addition of ion mobility, probe conformational changes and monitor protein unfolding events.

IM-MS has only recently become widely available after the introduction of the first commercial IM-MS instrument in 2007 (Pringle et al.) and is increasingly used for the study of important biological problems such as protein misfolding diseases. Newer instruments have also been introduced to the market in 2011 and 2013. Despite the advances in both ion mobility instrumentation development and the growing applications, advances in the software to process such data has been limited. Our lab has developed the first open-source program (Sivalingam et al., 2013) to process ion mobility data called Amphitrite (http://www.homepages.ucl.ac.uk/~ucbtkth/resources.html and https://github.com/gnsiva/Amphitrite and recently the first genetic-algorithm (GA) based program used for the deconvolution of collision-induced unfolding ion mobility data (Sivalingam et al, 2018).
Recently, we were also one of the first labs to show how a novel cyclic ion mobility device can be used to study protein structure and dynamics (Eldrid et al, 2021 Eldrid 2019).

The current project will focus on extending the functionality of the programs we have already developed but also developing new programs to handle the new cyclic IMS data.

A large number of IM-MS data files obtained during the analysis of diverse protein systems such as alpha-1 antitrypsin, curved DNA binding protein A, HDACs etc. will be available for development and testing.

Relevant Publications

Deconvolution of ion mobility mass spectrometry arrival time distributions using a genetic algorithm approach: Application to α1-antitrypsin peptide binding,
Ganesh N. Sivalingam, Adam Cryar, Mark A. Williams, Bibek Gooptu, Konstantinos Thalassinos,

International Journal of Mass Spectrometry, Volume 426, March 2018, Pages 29-37,

Amphitrite: a program for processing travelling wave ion mobility data.
Sivalingam, G.N., Yan, J., Sahota, H., Thalassinos, K.
International Journal of Mass Spectrometry (2013) 345, 54–62.

An investigation of the mobility separation of some peptide and protein ions using a new hybrid quadrupole/travelling wave IMS/oa-ToF instrument.
Pringle, S., Giles, K., Wildgoose, J., Williams, J., Slade, S., Thalassinos, K, et al. (2007).
International Journal of Mass Spectrometry, 261(1), 1-12.

Cyclic Ion Mobility-Collision Activation Experiments Elucidate Protein Behavior in the Gas Phase
Eldrid, C., Ben-Younis, A., Ujma, J., Britt, H., Cragnolini, T., Kalfas, S., Cooper-Shepherd, D., Tomczyk, N., Giles, K., Morris, M., Akter, R., Raleigh, D., Thalassinos, K.
J Am Soc Mass Spectrom (2021) 32, 1545-1552

Gas Phase Stability of Protein Ions in a Cyclic Ion Mobility Spectrometry Traveling Wave Device
Eldrid, C., Ujma, J., Kalfas, S., Tomczyk, N., Giles, K., Morris, M., Thalassinos, K.
Analytical Chemistry (2019) 91, 7554-7561

Integration of Mass Spectrometry Data for Structural Biology
Britt, H. M., Cragnolini, T., Thalassinos, K.
Chem Rev (2021)

Any particular skills that are required for this project?

Any programming language ideally Python and/or Web based techonologies like HTML, CSS, JavaScript.

Is the project suitable for full-time / part-time students?

Both

Project Code: *I_2023_MAW_1*

Molecular dynamics studies of the recognition of Hsp90 chaperones by TPR domains

First supervisor
Name: Mark Williams
Affiliation / Position: ISMB Birkbeck / Lecturer in Biophysics
Email: ma.williams@bbk.ac.uk

Project Description

The Hsp90 chaperone proteins are an essential regulator of activity of a range of cell cycle proteins and homorone receptors [1]. Partly folded proteins (called clients) bind to Hsp70 and are transferred to Hsp90 and, together with about a dozen co-chaperone proteins, they act to protect their clients from aggregation as they are transported around the cell and to activate or inactivate them as required. Many of the co-chaperones have a two domain architecture consisting of an enzyme domain attached to a TPR (tetratricopeptide repeat) domain. The co-chaperone TPR domains all bind to the C-terminal peptide EEVD sequence motif of Hsp90 and Hsp70 and consequently localize the enzyme function to the chaperone and its client. It is known experimentally that the binding affinity varies by >10000 fold among the different co-chaperones and chaperones.

The aim of this project is to analyse the structural and energetic differences between human Hsp90-binding TPR domains alone and in complex with EEVD peptides and try to identify possible sources of their different binding behaviour.

References:

1. Maximillian M. Biebl and Johannes Buckner (2019). **Structure, function and regulation of the Hsp90 chaperone machinery.** CSH Perspectives in Biology doi: 10.1101/cshperspect.a034017

Any particular skills that are required for this project?

It would be helpful to have a reasonably good grasp of physical chemistry. The student will gain experience in: Comparative structural analysis of proteins. Structural analysis of protein-ligand interactions. Structural modelling of mutations. Molecular dynamics simulation.

Is the project suitable for full-time / part-time students?
Both

Project Code: *I_2023_MAW_2*

Evolutionary divergence and functional investigation of NAMLAA amidases – bacterial proteins essential to cell division.

First supervisor: Mark Williams
Affiliation / Position: ISMB Birkbeck / Lecturer in Biophysics
Email: ma.williams@bbk.ac.uk

Project description:

The bacterial cell wall is single mega-macromolecule comprised of linear glycan chains interlinked by short peptides. The cell wall is a dynamic structure that has to be modified to allow growth and division of bacteria. Among the most important enzymes in the remodelling process are the N-acetylmuramyl-L-alanine amidases (NAMLAAs), which are responsible for cleaving the peptide crosslinks from the gylcan chain. These metalloenzymes have a modular structure containing an amidase domain combined with a variety of additional enzymatic or cell wall binding domains [1].

The PFAM family Amidase_3 contains (or perhaps consists entirely of) NAMLAAs that are known to be essential to separation of the daughter cells during cell-division (during which they cleave peptidoglycan at the septum). The aim of this project is to collect and organise the present information on variations in the modular architecture and sequence of these Amidase_3 enzymes across species, and to then use sequence variation to investigate questions concerning their function. Specifically, experiments have suggested that some members of the family are auto-inhibited by an attached cell-wall binding domain and that they are then activated by another cell division protein (EnvC/NlpD) [2]. Large-scale multiple sequence alignment and covariation analysis should reveal whether or not there is any evidence in the evolving sequences for these suggested mechanisms.

References:

[1] Rocaboy et al. (2013) The crystal structure of the cell division amidase AmiC reveals the fold of the AMIN domain, a new peptidoglycan binding domain. Mol Microbiol 90:267-277. https://doi.org/10.1111/mmi.12361

[2] Peters, N.T., Dinh, T., *and* Bernhardt, T.G. *(*2011*)* A fail-safe mechanism in the septal ring assembly pathway generated by the sequential recruitment of cell separation amidases and their activators. J Bacteriol 193:4973-4983

Any particular skills that are required for this project (e.g., knowledge of a specific programming language, good knowledge of statistics, understanding of chemistry etc)?

A keen interest in protein structure and evolution. The project will provide experience in automating handling of protein sequence data and in comparative sequence and structure analysis.

Is the project suitable for full-time / part-time students? Both

Project Code: *I_2023_GZ_1*

Adapting AI molecule recognition in cryo-tomography software for coat complexes

First supervisor
Name: Giulia Zanetti
Affiliation / Position: Birkbeck College Biological Sciences, Reader in Structural Biology

Telephone:
Email: g.zanetti@bbk.ac.uk

Project Description
Cryo-tomography is a powerful technique to obtain 3D reconstructions of biological molecules in complex environments. Despite the recent improvements in hardware and software, mining cryo-electron tomograms to identify the coordinates of molecules of interest remains a challenge.
Lately, a number of neural network based software has been released for the task. However, all recently released software has been written and optimised in the developers' computational and biological systems, and it is challenging to adapt it to our lab needs.
In our lab, we study membrane assembled coat protein complexes in vitro and in vivo. We have datasets of various degrees of complexity where we need to automatically detect coat components.
The aim of this project is to take a library of recently released academic software which use AI to identify molecules in cryo-tomograms, and optimise the associated workflows (including training) to work on coat proteins.

Any particular skills that are required for this project?
Python programming, interest in image processing.

Is the project suitable for full-time / part-time students?
Either full time or part time.
Please note that due to space constraints the project can only be supervised remotely. There will be the possibility of planning regular in person progress report meetings but most of the interaction will be online.

Project Code: *I_2023_GZ_2*

Subtomogram averaging of the outer COPII complex

First supervisor
Name: Giulia Zanetti
Affiliation / Position: Birkbeck College Biological Sciences, Reader in Structural Biology

Telephone:
Email: g.zanetti@bbk.ac.uk

Project Description
We have collected a cryo-ET dataset of COPII-coated vesicles reconstituted in vitro from native microsomal membranes. We aim to obtain subtomogram averaging reconstructions of an inner and outer layer of the coat. In this project, you will focus on the outer coat layer.
You will use specialised software and learn the basics of image processing.

Any particular skills that are required for this project?
Python programming, interest in image processing.

Is the project suitable for full-time / part-time students?
Either full time or part time.
Please not that due to space constraints the project can only be supervised remotely. There will be the possibility of planning regular in person progress report meetings but most of the interaction will be online.

Projects with EXTERNAL supervisors

NOTE FOR 2023:
Projects classified as external are now all projects offered by supervisors who are not based within the Department of Biological Sciences and do not teach on the course, or if they are teaching on the course, they are not associated with Birkbeck/ISMB.

Project Code: *E_2023_GF_1*

Gene-to-gene Dynamic Methods for Time-Dependent Analysis for Biology Risk Assessments

First supervisor
Name: Dr George Fitton
Affiliation / Position: Unilever (SEAC) Computational Scientist
Telephone: N/A
Email: George.Fitton@Unilever.com

Second supervisor
Name: Dr Alistair Middleton
Affiliation: Unilever (SEAC) Computational Science Leader
Email: Alistair.Middleton@Unilever.com

Please note: An internal contact will be assigned to this project to act as a "tutor".

Project Description

In a fast-moving consumer goods environment, it is vital that safety assessments are conducted to ensure products are safe for humans and the environment either during the production or the use of products. Risk assessments historically have heavily relied on the use of in vivo animal testing to identify detrimental impacts of chemicals on organisms. This approach leads to clear ethical issues, is not compatible with sustained societal pressure to remove the use of animals testing for safety purposes and is not always relevant to answer questions regarding the risk to human health.

For more than 20 years, Unilever has been working towards developing Next Generation Risk Assessments (NGRAs) leveraging on advances in biology, genetics, computational sciences, mathematics, and statistics to develop novel in silico and in vitro based methods – New Approach Methodologies (NAMs) that robustly support safety decisions without testing on animals.

One of the *in vitro* assays that is readily used for safety assessment of chemicals is High Throughput Transcriptomics (HTTr). HTTr is used to detect changes in transcriptional activity in cells as a result of chemical treatment at different chemical concentrations. One question around this approach is which time after exposure to measure transcriptional changes. Time can also be an important factor when considering repeat dosing or the effect of chemicals on differentiating cells, for example during development. Time-course experimental design and dynamical modelling can therefore be particularly important for a range of questions in risk assessment. Despite their feasibility and high-demand, sophisticated dynamic methods aren't yet well validated in large-scale comparative studies and often lack statistical and computational rigor when compared to their static counterparts.

The project aims to implement and review several gene-to-gene dynamic methods including but not limited to: FunPat, Time-Course Gene Set Analysis (Tcgsaseq), and Dynamic Bayesian Neural Networks [1]. The student is expected to gain a good understanding of the models, e.g., their assumptions, modelling strategies, parameter, and sample size constraints, document their differences and applicability

domains before implementing and benchmarking the models on a range of suitable datasets.

A second important aspect of the project requires building and developing tools in R, using either existing scripts and packages provided by journal authors or writing custom pipelines for SEAC's experimental and bioinformatics teams. For example, taking advantage of current time-series analysis methods that aren't optimized for time-course gene set analysis, e.g., optimizing state-space models on longitudinal gene data. Ideally, the student will produce a tool using one of the reviewed methods.

Role of Unilever
During the project the student will interact with experts in computational science (Dr George Fitton) and mathematical modelling (Dr Alistair Middleton) in addition to the wider bioinformatics team. The project will give the student an insight into safety assessment of chemicals and how such assessments are performed at Unilever. The student will also learn how bioinformatics is applied in a business environment, as well as how to present and discuss their work.

About SEAC
Both supervisors are based at Unilever's Safety and Environmental Assurance Centre (SEAC) staffed by around 150 people, including scientists recognised externally as leaders in their areas of expertise. The core purpose of the SEAC team is to assess the safety and environmental sustainability of Unilever's products so that they are safe for the people who use them and better for the environment. In doing so, we use cutting-edge research, including exposure science, cell-based in vitro assays, computational chemistry, mathematical modelling, and bioinformatics. As well as conducting our own scientific research, we also work closely with leading scientific authorities around the world including regulators, government scientists, and academic experts. This collaboration ensures we are always using the most up-to-date scientific advances within our safety and environmental sustainability assessments. We regularly publish and present our research advances and make them available through our website. To learn more about our safety science take a look at our BBC StoryWorks film.

Literature

[1] Oh, Vera-Khlara S., and Robert W. Li. "Temporal dynamic methods for bulk RNA-Seq time series data." Genes 12.3 (2021): 352.

Any particular skills that are required for this project?

The student is expected to have some experience in programming in either R or Python. A background in statistics is preferred but not essential.

Is the project suitable for full-time / part-time students?

The project is suitable for both full-time and part-time students.

Is the student expected to work on site? If so, is the student expected to pay for travel/accommodation?

The student will work remotely but will have the opportunity to visit SEAC to meet the team. Travel expenses aren't provided.

Project Code: *E_2023_PK_1*

Toxicity pathways informed by protein-protein interactions

First supervisor
Name: Dr Predrag Kukic
Affiliation / Position: Science Leader
Telephone: 07446 209025
Email: predrag.kukic@unilever.com

Second supervisor
Name: Dr Claire Peart
Affiliation: Bioinformatician, Unilever
Email: claire.peart@unilever.com

Please note: An internal contact will be assigned to this project to act as a "tutor".

Project Description

Background

For almost a century, animal studies have been used to test the potential for adverse consequences of exposure to chemicals in the pharmacological, agrochemical and consumer goods industry. This approach is based on the use of laboratory animals and apical endpoints which are clinical signs or pathologic states that are indicative of a disease state in the animal that can result from exposure to a toxicant. Aside from the ethical issues, this approach is sometimes not relevant in answering questions regarding the risk to human health.

Recently, the landmark report *Toxicity Testing in the 21st Century: A Vision and a Strategy, 2007*, from the U.S. National Academy of Sciences, precipitated a major change in the way toxicity testing has been conducted. This approach is based on the use of *in vitro* assays on human cells or cell lines using robotic high throughput screening (RNAseq, proteomics, binding assays, cell painting, etc.) rather than whole animal models. The response measured in the *in vitro* assays is then combined with known human exposure (the dose of application) to make a safety decision on the proposed use of the chemical.

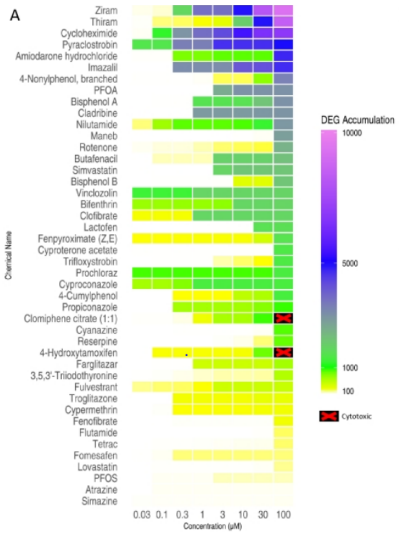


Figure 1. Adopted from Harrill et al. 2021. The figure shows accumulation of differentially expressed genes (DEG)s when a cell line is treated with an increasing concentration of the chemicals (x-axis). Names of the chemicals are provided on y-axis.

Objectives

One of the *in vitro* assays that is readily used for safety assessment of chemicals nowadays is High Throughput Transcriptomics (HTTr). As such, HTTr, is capable of detecting changes in transcriptional activity in a cell as a result of treatment with an increasing concentration of the chemical (see Figure 1). Given sufficient exposure to the chemical, cells will upregulate or downregulate the amount of mRNA that they produce in a gene-specific manner, which in turn alters the amount of proteins that are translated to perform their specific functions related to maintaining homeostasis (e.g., clearing or neutralizing foreign chemicals, repairing damage, increasing energy production, etc.). These alternations in mRNA production are indicators of perturbations in so called 'toxicity pathways' that can lead to adverse health outcomes under conditions of human exposure.

To implement the new toxicity-testing approach, identification of toxicity pathways is a key (see AOP Wiki for examples). However, most toxicity pathways are unknown since they do not always correspond to currently annotated cellular pathways that represent normal physiology (KEGG, Reactome, Wiki Pathways, etc.). Therefore, the aim of this project is to leverage the current knowledge of protein structures and protein-protein interactions to help identify cellular responses to toxicants.

Methodology

HepG2, MCF7, and HepaRG cells (5 biological replicates each) were treated for 24 h with 10 chemicals in a dose dependent manner (Middleton et al. 2022). The chemicals with well-studied mechanism of toxicity were carefully chosen as benchmark chemicals. For each chemical-cell combination, a list of differentially expressed genes (DEGs) was derived for each treatment dose.

- In the first part of the project, the student will analyse the lists of DEGs for each dose and each chemical using available information from protein-protein interaction databases (IntAct, String, etc.). The student will explore and identify presence/absence of protein-protein complexes in the list of identified DEGs and their statistical significance. This will be analysed per cell line per chemical.
- In the second part of the project, the identified protein-protein complexes and their function will be scrutinised against the known toxicity mechanism of the 10 chemicals.

Role of Unilever
During the project the student will interact with experts in structural bioinformatics (Dr. Predrag Kukic) and bioinformatics (Dr. Claire Peart). The project will give the student an insight into safety assessment of chemicals and how such assessments are performed at Unilever. The student will also learn how bioinformatics is applied in a business environment, as well as how to present and discuss their work.

About SEAC
Both supervisors are based at Unilever's Safety and Environmental Assurance Centre (SEAC) staffed by around 150 people, including scientists recognised externally as leaders in their areas of expertise. The core purpose of the SEAC team is to assess the safety and environmental sustainability of Unilever's products so that they are safe for the people who use them and better for the environment. In doing so, we use cutting-edge research, including exposure science, cell-based in vitro assays, computational chemistry, mathematical modelling, and bioinformatics. As well as conducting our own scientific research, we also work closely with leading scientific authorities around the world including regulators, government scientists, and academic experts. This collaboration ensures we are always using the most up-to-date scientific advances within our safety and environmental sustainability assessments. We regularly publish and present our research advances and make them available through our website SEAC web page. To learn more about our safety science take a look at our BBC StoryWorks film.

Literature
1. Harrill JA, Everett LJ, Haggard DE, Sheffield T, Bundy JL, Willis CM, Thomas RS, Shah I, Judson RS. High-Throughput Transcriptomics Platform for Screening Environmental Chemicals. Toxicol Sci. 2021 Apr 27;181(1):68-89. doi: 10.1093/toxsci/kfab009. PMID: 33538836.

2. Alistair M Middleton, Joe Reynolds, Sophie Cable, Maria Teresa Baltazar, Hequn Li, Samantha Bevan, Paul L Carmichael, Matthew Philip Dent, Sarah Hatherell, Jade Houghton, Predrag Kukic, Mark Liddell, Sophie Malcomber, Beate Nicol, Benjamin Park, Hiral Patel, Sharon Scott, Chris Sparham, Paul Walker, Andrew White, Are Non-animal Systemic Safety Assessments Protective? A Toolbox and Workflow, Toxicological Sciences, Volume 189, Issue 1, September 2022, Pages 124–147, https://doi.org/10.1093/toxsci/kfac068

Any particular skills that are required for this project?
The student should be familiar with essential computing skills (Python or any other relevant programming language), understand fundamental biological concepts, be familiar with the protein-protein interaction databases and possess basic statistical skills.

Is the project suitable for full-time / part-time students?
This project is suitable for both full-time and part-time students.

Is the student expected to work on site? If so, is the student expected to pay for travel/accommodation?

The student is not expected to work on site. For this work the student will need access to a laptop/desktop computer. No additional computational resources are required.