## Graph based methods to identify drugs and metabolites in stroke patient urine

**Supervisors:**  
Prof Mike Barrett - [michael.barrett@glasgow.ac.uk](mailto:michael.barrett@glasgow.ac.uk)  
Dr Simon Rogers - [simon.rogers@glasgow.ac.uk](mailto:simon.rogers@glasgow.ac.uk)  
Dr Jesse Dawson - [jesse.dawson@glasgow.ac.uk](mailto:jesse.dawson@glasgow.ac.uk)  
Prof Sandosh Padmanabhan - [Sandosh.Padmanabhan@glasgow.ac.uk](mailto:Sandosh.Padmanabhan@glasgow.ac.uk)

**Abstract:**  
This project will be fundamentally related to developing novel Bioinformatic approaches, through machine learning, to gain new insight into drug metabolism and its impact in medicine.  
Understanding how patients metabolise drugs is increasingly recognized as fundamental to stratified medicine (1). Pathways of drug metabolism can be complicated and the genetic basis of drug metabolism is yet to be elucidated. Quantification of parent drug and metabolite levels can demonstrate directly how a patient alters drugs and we can then assess whether this relates to treatment response. Mass spectrometry based untargeted metabolomics offers unparalleled capability to identify drugs and their metabolites (2). The supervisory team have been working on Informatics based methodologies to probe such untargeted metabolomics data to detect drugs and their metabolites (2,3).  
Stroke is the second most common cause of death globally and there is substantial variability in outcome to drug treatment across individual patients.

**Aims**  
We propose to use untargeted mass spectrometry based metabolomics, with full sample fragmentation, to look at banks of urine samples taken from stroke patients during post-stroke therapy. The approach will determine levels of drug in urine and also the full array of metabolites associated with each drug.  
Central to the work will be refining the methods used in clustering related metabolites within datasets, substantial analysis of fragmentation patterns to ascertain the metabolic transitions under investigation and developing algorithms to stratify patients based on the drug metabolism ability. Novel algorithms enabling correlation between drug metabolism pathways and treatment outcomes will be developed providing novel methods enabling prediction of treatment outcome in stroke patients to be determined based on readily quantifiable drug metabolism capability

**Training outcomes**  
(1) Understanding of patient stratification in stroke Many patients are admitted with stroke despite taking preventative treatment such as anti-platelet medications, lipid lowering drugs and anti-hypertensive drugs. No drug will be             100% effective but this may reflect resistance to drugs or poor compliance. Detailed assessment of recent drug ingestion will help make this distinction, stratifying patients into those with and without genuine breakthrough                 events. This could guide choices regarding on going therapy  
(2)  Development of novel algorithmns to determine drug metabolism This will include training in advanced Machine Learning methods, as well as Bayesian Statistical Modelling. The candidate will get experience in implementing data          pre-processing steps and state-of-the art machine learning algorithms in Python and embedding them into a pre-existing pipeline. Many of these skills (handling large noisy datasets, implementing, using and evaluating Machine            Learning algorithms, Bayesian statistics) are transferable to other applications within metabolomics and other –omics domains.  
(3)  Development of algorithms linking drug metabolism to treatment outcome

**References**  
(1) Close, S.L. Clopidogrel pharmacogenetics: metabolism and drug interactions. Drug Metabol Drug Interact. 26, 45-51.(2011)  
(2) Creek DJ, Barrett MP. Determination of antiprotozoal drug mechanisms by metabolomics approaches. Parasitology. 141,83-92. (2014)  
(3) Daly R, Rogers S, Wandy J, Jankevics A, Burgess KE, Breitling R. MetAssign: probabilistic annotation of metabolites from LC-MS data using a Bayesian clustering approach. Bioinformatics. 30, 2764-71. (2014)

ear Ms Young,

Thank you for your recent application for a studentship position on the MRC DTP in Precision Medicine, University of Glasgow. We would like to invite you for interview on 16 May 2016 at 2.30pm.

Please report to the MVLS Graduate School, Wolfson Link Building, University of Glasgow reception – B10 on campus map: [http://www.gla.ac.uk/media/media\_335384\_en.pdf](https://mail.student.gla.ac.uk/owa/redir.aspx?REF=XatShM_nXqhHFsftT-pgiHCWkgoEST2Q_q_ql8H50EUdyHN7l3PTCAFodHRwOi8vd3d3LmdsYS5hYy51ay9tZWRpYS9tZWRpYV8zMzUzODRfZW4ucGRm). We would be grateful if you could arrive 10 minutes before your interview time. Your interview should last approximately 30 minutes, and please plan to leave the University two hours before/after your interview which will take around 30 minutes, this will give you an opportunity to meet with your proposed supervisors or their representatives.  Your proposed supervisors or their representatives will be in contact to organise this meeting opportunity.

You are asked to prepare a 10 minute (max) presentation on what you can bring to the project to which you have applied. The presentation should include your understanding of the project area. Please note that the University of Glasgow IT Services do not support Apple software, so please ensure that your presentation is saved onto a USB compatible with a PC.

Reasonable travel expenses will be reimbursed (standard rail/economy airfare etc.). Prior to confirming any travel bookings, please send a note of anticipated costs to me. When recouping incurred expenses the attached Sundry payee form should be completed, and you must include all receipts. Similarly, the college will arrange 1 x overnight accommodation for you should it be required, please let me know in your response to this email. Completed forms and original receipts should be sent to me at the address below. If you require further clarification regarding expenses, please let me know as soon as possible.

As there are considerable interests in the above studentship, please inform me by return no later than **12 noon on Friday 6 May 2016** if you will be attending for interview.

If you require further information, please do not hesitate to contact me.

Look forward to hearing from you.

Yours sincerely,

Alexis Merry.

# Graph based methods to identify drugs and metabolites in stroke patient urine

## Project Description

Background   
This project will be fundamentally related to developing novel Bioinformatic approaches, through machine learning, to gain new insight into drug metabolism and its impact in medicine.   
  
Understanding how patients metabolise drugs is increasingly recognized as fundamental to stratified medicine (1). Pathways of drug metabolism can be complicated and the genetic basis of drug metabolism is yet to be elucidated. Quantification of parent drug and metabolite levels can demonstrate directly how a patient alters drugs and we can then assess whether this relates to treatment response. Mass spectrometry based untargeted metabolomics offers unparalleled capability to identify drugs and their metabolites (2). The supervisory team have been working on Informatics based methodologies to probe such untargeted metabolomics data to detect drugs and their metabolites (2,3).   
  
Stroke is the second most common cause of death globally and there is substantial variability in outcome to drug treatment across individual patients.   
  
Aims   
We propose to use untargeted mass spectrometry based metabolomics, with full sample fragmentation, to look at banks of urine samples taken from stroke patients during post-stroke therapy. The approach will determine levels of drug in urine and also the full array of metabolites associated with each drug.   
  
Central to the work will be refining the methods used in clustering related metabolites within datasets, substantial analysis of fragmentation patterns to ascertain the metabolic transitions under investigation and developing algorithms to stratify patients based on the drug metabolism ability. Novel algorithms enabling correlation between drug metabolism pathways and treatment outcomes will be developed providing novel methods enabling prediction of treatment outcome in stroke patients to be determined based on readily quantifiable drug metabolism capability   
  
Training outcomes   
(1) Understanding of patient stratification in stroke Many patients are admitted with stroke despite taking preventative treatment such as anti-platelet medications, lipid lowering drugs and anti-hypertensive drugs. No drug will be 100% effective but this may reflect resistance to drugs or poor compliance. Detailed assessment of recent drug ingestion will help make this distinction, stratifying patients into those with and without genuine breakthrough events. This could guide choices regarding on going therapy   
  
(2) Development of novel algorithmns to determine drug metabolism This will include training in advanced Machine Learning methods, as well as Bayesian Statistical Modelling. The candidate will get experience in implementing data pre-processing steps and state-of-the art machine learning algorithms in Python and embedding them into a pre-existing pipeline. Many of these skills (handling large noisy datasets, implementing, using and evaluating Machine Learning algorithms, Bayesian statistics) are transferable to other applications within metabolomics and other –omics domains.   
  
(3) Development of algorithms linking drug metabolism to treatment outcome   
  
Application   
You can apply here: [http://www.gla.ac.uk/research/opportunities/howtoapplyforaresearchdegree/](https://www.findaphd.com/common/clickCount.aspx?theid=73708&type=184&DID=462&url=http%3a%2f%2fwww.gla.ac.uk%2fresearch%2fopportunities%2fhowtoapplyforaresearchdegree%2f)   
Within the application, at the programme of study search field option, please select ‘MRC DTP in Precision Medicine’.   
  
Please note that, in step 6 within the online application process, you are asked to detail supervisor/project title information. Please ensure that you clearly detail this information from the information provided within this abstract advert. Within the research area text box area, you can also add further details if necessary.   
  
Please ensure that all of the following supporting documents are uploaded at point of application:   
• CV/Resume   
• Degree certificate (if you have graduated prior to 1 July 2016)   
• Language test (if relevant)   
• Passport   
• Personal statement   
• Reference 1 (should be from an academic who has a knowledge of your academic ability from your most recent study/programme)   
• Reference 2 (should be from an academic who has a knowledge of your academic ability)   
• Transcript   
  
This is a joint programme between the Universities of Edinburgh and Glasgow. You will be registered at the host institution of the primary supervisor detailed in your project selection.   
  
For more information visit: [http://www.gla.ac.uk/colleges/mvls/graduateschool/precisionmedicine/](https://www.findaphd.com/common/clickCount.aspx?theid=73708&type=184&DID=462&url=http%3a%2f%2fwww.gla.ac.uk%2fcolleges%2fmvls%2fgraduateschool%2fprecisionmedicine%2f)

Hi Fran,

Sorry - got to this a bit late for a meeting. Here’s some context about the PhD project:

 - Fragmentation data is a bit under-used because decent computational methods don’t exist

 - We think there are useful parallels between frag data and text (weird…)

 - There are lots of advanced algorithms for text

 - So, we took one text algorithm (LDA - [https://en.wikipedia.org/wiki/Latent\_Dirichlet\_allocation](https://mail.student.gla.ac.uk/owa/redir.aspx?REF=WTigOADzDr93hA0g2FVtdT_8yrpFjc92einx5hQmptEDkw2pV3PTCAFodHRwczovL2VuLndpa2lwZWRpYS5vcmcvd2lraS9MYXRlbnRfRGlyaWNobGV0X2FsbG9jYXRpb24.)) and applied it to frag data and got some interesting results:

- e.g. we can find ‘topics’ that correspond to molecular substructures allowing us to annotate molecules that couldn’t normally be identified

- we can find topics that seem to correspond to differentially expressed metabolites so we can hypothesise that the substructure implied by the topic is ‘causing’ the change in expression

 - This has lots of potential applications in clinical projects as we think we will be able to see the substructures that are indicative of particular drugs, and how they change across individuals

 - The PhD is about doing that…we have frag data from urine samples of people who have had strokes and they have all been given cocktails of drugs and we want to use the text approaches to find out what’s going on

In terms of skills….well, I don’t think anyone exists who would have all of the skills necessary to be expert in text processing, fragmentation data, and strokes…so any candidate would have to learn some of those things. Programming is important, because we will need to implement new models (probably in python) and a willingness to be immersed in the stats of text modelling (as stats go, it’s quite a nice bit…everything is discrete so it comes down to counting). Aside from that, successful PhD candidates tend to be people willing to work hard and can communicate with Comp Scientists, clinicians, and life scientists.

No idea how competitive it is likely to be - this is the first year of that particular PhD scheme. My gut feeling is that they will have lots of applicants across the advertised projects without computational skills so (although you may not feel that your programming is at expert level), you should use it to your advantage!

Hope that helps - sorry we didn’t have time to meet.

As for projects - if the above seems interesting to you, I’m happy to supervise a project in this area (we have lots we want to do). I will be on email a tiny bit tomorrow but not much. Feel free to email Mark B (and cc me) saying that we will sort something out.

Simon

### Plan what to bring.

* Extra copies of your resume on quality paper
* A notepad or professional binder and pen
* A list of references
* Information you might need to complete an application
* A portfolio with samples of your work, if relevant

On 30 Mar 2016, at 09:37, Francesca Young <[2226768Y@student.gla.ac.uk](https://mail.student.gla.ac.uk/owa/redir.aspx?REF=GDmQNzNrmIlN6pzZimgxWzT5IlWnI0gh774PujLtIggDkw2pV3PTCAFtYWlsdG86MjIyNjc2OFlAc3R1ZGVudC5nbGEuYWMudWs.)> wrote:

Dear Fran and Ana-Maria,  
You’ve both been in touch regarding MSc projects so I thought it most efficient for me to reply both at the same time (apologies Fran for taking so long - hectic time of year).  
  
My research interest at the moment is in developing algorithms for aiding the analysis of LC/MS metabolomics data (particularly fragmentation data). This has two strands:  
- Developing new techniques: statistics & machine learning, playing with data, running in-silico experiments to assess if the methods are any good and better than the currently available techniques. This can be done in any language.  
- Implementing validated techniques into a form that can be used in, for example, the Glasgow Polyomics software pipeline: more software engineering - making sure code (Python) is efficient and can be trusted in deployed software. We’re trying to move code away from R - at the moment there are too many different languages being used in one pipeline and python is more convenient.  
  
If that sounds of interest to either of you, let me know….it will often be possible to create a project that is useful to both me (and Polyomics) and you. I’m off on paternity leave starting on Thurs for a couple of weeks but will be on email throughout. When do you need to fix up projects by?  
  
Simon

# How important is machine learning for bioinformatics?

B Machine learning plays an important role in a lot of bioinformatics problems. To list a few -

* **Gene Finding Algorithms**: Hidden Markov Models (HMM)
* **Gene Expression**: Clustering Algorithms  like k-means
* **Genome Alignment**: HMM
* **Population Stratification :** PCA, MDS, manifold learning
* **GWAS :** Linear and logistic regression (Mixed linear models)
* **Genomic Selection :** Classification Algorithms like Random Forest, Decision Trees, Naive Bayes,  SVM, Logistic Regression

In last one and half year that I have worked in a bioinformatics lab, I have worked on a bunch of projects involving machine learning. I have built classification models for predicting structural variants, used Bayesian Networks for incorporating prior information in a genomic selection model, used Bayesian optimization to estimate parameters for a model etc.   
Machine learning is immensely helpful in doing bioinformatics, in fact I see a new idea along the interface of ML in bioinformatics almost every week.ioinformatics is really data intensive. As bionformaticians we viably  survive as data parasites and essential to our survival is the ability  to grow the right set of "teeth" in order to munch and munge through our  data rich diet. Gene annotation, motif finding, pathway analysis are  all but routine applications for mining, classification, functional  prediction and further downstream analysis.  
  
Few  years ago we thought that next generation sequencing was the game  changer, it made data generation cheaper and revolutionized  bioinformatics in many ways, today we sit at the top of untapped  insights yet to be released from all the data we've accumulated and it's  clear that better ways to intelligently integrate and query this data  are needed so we make sense of what is hidden underneath it. Machine  learning allows just that through an arsenal of algorithms you'd be able  to classify, aggregate,  project insights and have fun while at it.   
  
So the short answer is machine learning is very important to the big biology data that drives bioinformatics.  
  
This paper here provides examples for machine learning applications in biology  
[Page on nature.com](http://www.nature.com/nrg/journal/v16/n6/full/nrg3920.html)  
  
Another relevant read that puts things in perspective is the recent Nature article on outlooks of  data titled "[Big data: The power of petabytes](http://www.nature.com/nature/journal/v527/n7576_supp/full/527S2a.html)" by Michale Eisenstein.

[Written Nov 11](https://www.quora.com/How-important-is-machine-learning-for-bioinformatics/answer/Hisham-Eldai) • [View Upvotes](https://www.quora.com/api/mobile_expanded_voter_list?type=answer&key=dyPZFrxspfi) • Answer requested by [David Tran](https://www.quora.com/profile/David-Tran-109)

You can only go so far by analyzing experimental data; whether your own or publicly available. The next step in becoming a good computational biologist is to be able to develop models and classifiers learning differences between different conditions and being able to predict what class a new data point will be assigned to. Machine learning has been a transformational subfield in computer science and traditionally support vector machines (SVM's) in particular lead the charge in classifying objects. This is a problem routinely encountered in biology. Imagine a scenario in which you need to learn from existing experimental data features that distinguish DNA sequence bound by different proteins and next classify novel DNA sequence based on which protein it will interact with. Machine learning techniques are very good at solving such problems. Especially now with the advent of deep learning machine learning in biology is more important and hotter than ever!

Ability to get things done with no structure, no clear objectives, and very little if any supervision.

earning how to learn (not just skills); techniques/technologies/skills become outdated soon- how do you pick up new ones when you need?

(Close, 2012)

Pharmocogenetics future of personalised medicine –genetic biomarkers will guide therapeutic approach.

Glossary

Pharmacokinetic

Pharmacodynamic

ACS acute choronary syndrome

Baysian Statistics

* Probabilities as beliefs
* Updates your beleifs as data is observred
* Requires a model that describes the data generation

LDA

* Appealing Baysian unsupervised learning model
* Training is difficult
* Validation is difficult
* Presentation is difficult – visualisation for the Baysian model is hard.