Design of synthetic plant and mammal gene regulatory networks using nonparametric Bayesian approaches

Primary contact for the team

Christopher Andrew Penfold, Wellcome/CRUK Gurdon Institute, University of Cambridge: Christopher is a postdoctoral research associate in Prof. Azim Surani’s lab, developing statistical machine learning approaches to decipher the molecular mechanisms that underpin early mammalian development. He was previously a postdoctoral research fellow in the Systems Biology Centre at the University of Warwick, where he developed network inference algorithms to elucidate plant responses to stress. c.penfold@gurdon.cam.ac.uk

Team

Marc Jones, Computational and Systems Biology, John Innes Centre: Marc is a postdoctoral fellow at JIC working on Brassica genomics. He recently completed his PhD in the group of Prof. Richard Morris, working on computational methods for unraveling how different Brassica napus varieties integrate environmental cues to determine flowering time. This work involved inferring gene regulatory network models from transcriptomics time series in different tissues and using these networks to build dynamic models of the floral transition. He wrote software for Bayesian network inference and developed approaches for identifying regulatory modules. marc.jones@jic.ac.uk

Iulia Gherman, Department of Biology, University of York. Iulia is a 3rd year synthetic biology PhD student studying and manipulating the stress response in Arabidopsis in Prof. Denby’s group. This work involves inferring gene regulatory networks (GRNs) from transcriptome time series data. Abiotic and biotic stresses act like external perturbations to the system so she is modelling perturbation mitigation strategies for the GRNs and testing them in planta in protoplasts for medium throughput screening in order to generate more stress-tolerant plants. iulia.gherman@york.ac.uk

Anastasiya Sybirna, Wellcome/CRUK Gurdon Institute, University of Cambridge: Anastasiya is a 3rd year PhD student in Prof. Azim Surani’s group studying the molecular mechanisms of human germ cell development. She uses an in vitro model of human primordial germ cell (PGC) specification from human embryonic stem cells (hESCs) to investigate the signalling and genetic regulators of this cell fate decision. As a part of her project, she uses a synthetic auxin-inducible degron (AID) system to address the involvement of specific transcription factors in early human development. as2439@cam.ac.uk

The team will be supported by Prof. Katherine Denby (York), Prof. Richard Morris (Norwich), and Prof. Azim Surani (Cambridge), who will provide advice and guidance. The whole team and the advisors will meet for a kick-off meeting in Cambridge and then
via video-conference every second week to discuss progress and arising problems, with another face-to-face meeting approx. half way through the project.

Summary
Despite some impressive success stories, synthetic biology approaches still require a significant element of trial and error. Our ability to design and predict is still poor. This is largely due to the challenge of simulating large networks and using them effectively to inform experiments. Here, we aim to develop a suite of computational tools for the rational data-driven (re)design of gene regulatory networks (GRNs) that can be used to guide synthetic biology efforts. Our approaches will be validated in silico, and using data from well characterised plant and mammalian systems. As a proof of principle, GRNs associated with stress responses in Arabidopsis thaliana will be synthetically rewired and tested in protoplasts, whilst future work will aim to engineer synthetic switches in human embryonic stem cells to rationally direct their cell fate decisions.

Proposal
A key objective of synthetic biology is the design of GRNs to achieve a desired biological function (Liu and Stewart 2015). In plant systems, this might involve the synthetic rewiring of GRNs associated with stress responses to render the plant resistant to a variety of pathogens, or rewiring GRNs associated with development and growth to maximise crop yield. Likewise, in mammalian systems, the synthetic manipulation of GRNs can be used as a basis for investigating development programs (Hörner and Weber 2012; Ausländer and Fussenegger 2013; Gordley et al. 2016) and disease (Kis et al. 2015), as well as providing novel therapeutic outputs (Lienert et al. 2014; Kis et al. 2015).

Synthetic rewiring of GRNs has been achieved in microbial systems, including Escherichia coli (Isalan et al. 2008; Baumstark et al. 2015) and the yeast Pichia pastoris (Windram et al. 2017). However, this rewiring was achieved by creating large libraries of either random combinations of promoters and protein coding sequences, or testing all possible combinations. We propose a principled approach to the design of rewiring in order to achieve a desired phenotype, such as increased robustness or tolerance to stress, which would greatly reduce the number of constructs needed for testing.

This targeted approach to the synthetic design of GRNs involves preliminary computational and statistical studies to identify priority network structures likely to yield the desired traits, followed by a fabrication stage, wherein the networks are implemented in vivo or in vitro, and increasingly in cell free environments (see Figure 1(a)).
Figure 1: (a) Workflow for synthetic redesign of transcriptional networks in plants (reproduced from Liu and Stewart, 2015). This involved a two-step process: computational and statistical approaches are used to design novel network architectures; architectures are then implemented, usually in vitro. (b) Nonparametric approaches to nonlinear dynamical systems allow quantitative prediction of knockout behaviour in Drosophila segmentation networks. Here we indicate observed (left) and predicted (right) expression values along the Drosophila anterior-posterior axis for a range of genes.

Whilst the molecular toolkit for synthetic manipulation of plant and mammalian networks continues to grow, a major bottleneck exists during the initial design phase. Accurate prediction of the effect of network rewiring remains elusive. This is partially due to the underlying complexity of GRNs whose underlying structure may often be unknown with a large degree of inherent redundancy. Further to this, much of the early work in synthetic design of networks has focussed on predicting and quantifying the impact and behaviour of small synthetic networks, such as those based on toggle switches (Gardner, Cantor, and Collins 2000) and the three-component repressilator (Elowitz and Leibler 2000). Although more complex synthetic networks now exist (Cantone et al. 2009), their size remains significantly smaller than number of genes implicated in complex biological functions (Windram et al. 2012; Lewis et al. 2015). The use of ODE systems for large networks remains impractical. Scalable approaches that allow the (re)design of genetic circuitry for moderately large networks (tens to hundreds of genes) with uncertain topologies are clearly needed.

We will exploit recent advances in nonparametric Bayesian approximations to nonlinear dynamical systems (Klemm, 2008; Penfold and Wild 2011) for the synthetic redesign of gene networks. These approaches allow nonparametric representations of arbitrarily complex ODE systems (Klemm, 2008; Penfold and Wild 2011), whilst retaining a high degree of scalability (Penfold et al. 2015). Crucially, these methods allow quantitative predictions of the effect of novel perturbations, and have been successfully used to predict networks in Arabidopsis (Windram et al. 2012) and the effects of traditional gene knockouts in Drosophila melanogaster segmentation networks (Penfold, unpublished data; Figure 1(b)). We will use these approaches to predict the effect of a range of synthetic modifications, including transcriptional rewiring, where regulatory links are
modified, and select networks that maximise appropriate objective functions e.g., that maximally up-regulate a set of genes positively associated with pathogen resistance. Existing implementations are not yet able to predict the effect of more subtle synthetic modifications, such as the shuffling of individual cis-regulatory elements, however recent developments (Lloyd et al., 2014) will allow these nonparametric approaches to do so.

Within this project, we will develop a suite of computational tools, building on existing nonparametric approaches to nonlinear dynamical systems, for the synthetic design of networks for particular function. The fidelity of these approaches will be extensively quantified in silico to assess the feasibility of large scale biological experiments. Finally, once the viability of our methods has been firmly established in silico, they will be used to predict network rewiring likely to increase Arabidopsis resistance to the necrotrophic fungal pathogen, Botrytis cinerea. High-throughput platforms for testing network rewiring in protoplast have been developed by Gherman (York), using chitin to mimic the effect of Botrytis infection and reporter constructs to monitor network gene expression. RNAseq analysis has identified the limited set of genes perturbed by protoplasting (~1000), and the ability to use this protoplast system for rapid characterisation of network expression after rewiring.

Future studies will aim to apply similar approaches to infer changes during the floral transition in B. napus and the redesign of mammalian networks to direct cell fate decisions that can be tested in vitro models. As the networks associated with primordial germ cells have been partially elucidated in both mice and humans, we will focus on rewiring to increase the efficiency of the derivation of primordial germ cells in vitro.

Benefits and outcomes
Completion of the project will provide a number of tangible outcomes. The primary outcome will be the development of a suite of computational tools that allow the rational (re)design of synthetic GRNs in a data-driven manner. This is a necessary step towards being able to design GRNs for a particular function.

The ability to rationally redesign networks for specific purposes has a huge potential to contribute towards food security, as well as fighting human disease with new therapeutics. Therefore, fast and accurate synthetic design that sidesteps the need for prolonged testing of all possible synthetic permutations would have an important impact. The validation of these approaches within our pilot study in protoplasts will, itself, prove to be a useful outcome, that will be of interest to the synthetic biology community, as well as those that use network biology.

This project will see a new, interdisciplinary collaboration between the Surani lab (Cambridge), specialising in epigenetics and early mammalian development, the Morris lab (Norwich), with expertise in computational systems biology, and the Denby lab (York), working in plant systems and synthetic biology. The diverse range of experience and expertise across the three groups is underpinned by a common unifying interest in
Another key development from this preliminary study will be the eventual transfer of our framework to mammalian systems. Adapting natural or synthetic regulatory networks from other organisms has been widely used to control gene expression and achieve desired phenotypes. Recently, plant hormone-inducible degrons have been harnessed to selectively degrade proteins of interest in models ranging from yeast to mammalian cells (Nemhauser and Torii, 2016). Ultimately, synthetic networks designed using our approaches could be reconstituted in heterologous systems, such as human embryonic stem cells (hESCs). This would prove their robustness in a different regulatory setting and potentially provide novel tools to control gene expression in human cells, which might help rationally direct cell fate decisions in hESCs. The rapid kinetics of these approaches could allow more precise control of the phenotypic traits and provide new avenues for research.

The methods developed in this call will build on existing nonparametric Bayesian approaches pioneered by Penfold and colleagues (Cambridge, York) and Jones (Norwich). Previous software from Penfold has been incorporated into the CyVerseUK platform, an open access, cloud-computing platform for bioinformatics analysis (Polański et al., 2017), and were accompanied by two-day tutorial workshops funded by GARNet. Upon successful completion of this project, software will be made available via CyVerseUK and via a GitHub repository under an open MIT license. Due to the success of the earlier workshops (http://cyverseuk.org/events/cyverse-uk-workshop), we will run a one-day workshop to train postgraduates and early career researchers on using these tools and the biological insights they enable. We will run this workshop immediately after the 2018 GARNet meeting (Sept 17th-18th at York) to enable researchers to stay on from that meeting for the workshop. In the workshop we will cover the tools developed in this project as well as the RNAseq analysis tools on CyVerseUK. Tutorial packages developed for the workshop will be hosted on Github for use by others. We could run the workshop separately within the 6 month period of the project but feel linking it to GARNet meeting will make it easier for people to attend.

Finally, another output from this collaboration will be the dissemination of knowledge within the participating organisations (Cambridge, Norwich, and York), to the wider scientific community, and to the general public. Internal dissemination within host organisations will be achieved via seminars and training courses, whilst participation in conferences and publication in scientific journals will allow communication to the scientific community. Cambridge also runs an annual Festival of Science (http://www.sciencefestival.cam.ac.uk) whilst York holds an annual Festival of Ideas (http://yorkfestivalofideas.com) and Pint of Science events (https://pintofscience.co.uk/events/york), all excellent venues for dialogue with the general public. Additionally, Gherman is part of a science communication project, BiobyDesign, which reports the latest synthetic biology research and developments. As synthetic biology remains a topic of general public interest, key findings could be
communicated to the general public and scientific community at one or more of these events in the future.

**Timeline:**
Month 1: Generation of *in silico* datasets for benchmarking. Begin testing existing nonparametric models to optimise variety of objective functions.
Month 2: Extensions to existing model implemented. Systematic benchmarking of competing methods.
Month 3 -4: Predictions of *Arabidopsis* stress-response networks from existing data and design of synthetic networks to yield resistance to *Botrytis cinerea*.
Months 4 -6: Testing of priority network topologies in protoplast. Finalise documentation for codebase.

**Sponsor for the research and cost centre**

Professor Azim Surani,
Gurdon Institute,
University of Cambridge.
a.surani@gurdon.cam.ac.uk

*I confirm that I have the full support of the sponsor listed above and that they can be added to the OpenPlant Fund mailing list to receive project updates (to which they can unsubscribe at any time).*

**Budget**

1. Project meetings (travel, accommodation for two meetings): **£600**
   * Whilst the majority of the work in this project can be achieved virtually with follow up meetings via Skype, two face-to-face meetings are necessary to ensure an efficient and successful collaboration.
2. Generating re-wiring constructs (consumables for up to 8 constructs: PCR reagents, RNA extraction, cDNA synthesis, sequence verification, general bacterial growth/lab consumables): **£1600**
3. Protoplast test of network re-wiring predictions (plant growth, enzymes and consumables for protoplasting, luciferase detection): **£600**
4. One-day Workshop (catering, travel, accommodation expenses of workshop leaders): **£1200**

In kind support:
- York, Cambridge: access to labs.
- Cambridge: access to HPC clusters (Gurdon Institute cluster and Darwin cluster).

**References**


