(for either slots 1 or 2.)

Project: *Causal inference for protein signalling networks.*

Supervisors:
Email c.oates@warwick.ac.uk.
s.e.f.spencer@warwick.ac.uk.

Dr Chris Oates (Statistics)

Dr Simon Spencer (Statistics)

Background: The modern "systems" view of cellular signalling can be conceptualised as a network of interacting components together with biochemical parameters that specify rates of reaction. In many scientifically important settings, including disease states such as cancer, both network topology and biochemical parameters are generally unknown. Reverse-phase protein arrays (RPPA [1]) and related technologies are emerging as important experimental tools for the systems-level investigation of cellular signalling. Such data provide a multiplex readout of protein concentrations in a biological sample. Most data obtained in this way are "observational", rather than "interventional", as controlled manipulation of cellular systems remains difficult. An important statistical challenge is therefore to estimate these unknowns from observational proteomic data. From a theoretical perspective, this project will be based on ideas from causal inference, where a directed acyclic graph (DAG) is used to describe causal relationships between random variables.

Objectives: This mini-project seeks to answer one simple question, namely, is it possible to infer a causal DAG that describes protein phosphorylation, using only observational proteomic data. In principle this is challenging [2], since (i) many other protein concentrations are highly correlated with the true causal proteins, and (ii) the causal DAG itself need not be uniquely identifiable from observational data. ([3] proves that one can only identify the "skeleton" of the causal DAG, along with any "v-structures", i.e. A -> B <- C.) The novel core of this project is to leverage joint modelling of both phosphorylated and total protein concentrations in order to induce additional v-structures into the causal DAG, rendering all causes (theoretically) identifiable from observational data. A careful statistical analysis of this approach will be timely and publication of the results would interest the research community.

"What the student will do": He/she will exploit recent advances in computational and graphical statistics [4] to undertake Bayesian model selection using causal models for RPPA data. He/she can expect to gain experience in applied Bayesian statistics, a working knowledge of causal inference and graphical statistics, and an understanding of high-throughput proteomics. These skills are currently highly sought-after in many areas of quantitative science and the student will be well prepared for their future research. Applications of causal inference within molecular biology are currently lacking; progression to a PhD would focus on establishing fundamental modelling frameworks that will allow for the data-driven investigation of cellular signalling systems.

Prerequisites: A basic competence in a suitable programming language (MATLAB, R, C, etc.).

[1] Hennessy, B.T. et al. (2010) A Technical Assessment of the Utility of Reverse Phase Protein Arrays for the Study of the Functional Proteome in Nonmicrodissected Human Breast Cancer. Clin. Proteomics 6:129-151.

[2] Sachs et al (2005) Causal protein-signaling networks derived from multiparameter single-cell data. Science 308:523-529.

[3] Pearl, J. (2009) Causality: models, reasoning and inference (Second Edition). Cambridge: MIT press.

[4] Barlett, M., Cussens, J. (2013) Advances in Bayesian Network Learning using Integer Programming. Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI 2013): 182-191.