Leeds Annual Statistical Research workshops
DEPARTMENT OF STATISTICS

MIMOmics

"Biostatistics and Machine Learning Methods in Omics Research"
34th LASR and MIMOmics Workshop

Room 1.33, Maurice Keyworth Building, Leeds University Business School

DAY 1: MONDAY 26 JUNE

10:30-11:00 Registration and Coffee

Welcome: Chair John Kent

11:00-11:05 Welcome and Introduction

11:05-11.25 Sir Alan Langlands, Vice Chancellor, University of Leeds

11.25-12.10 Some latest cutting age applications of shape analysis, and the centenary of D'Arcy Thompson's "On Growth and Form"
Kanti V. Mardia

12:10-12.30 MIMOmics
Jeanine Houwing-Duistermaat

12:30-13:30 Lunch

Session 1: Networks – Chair Felix Agakov

13:30-14.00 Multilayer networks: new methods and techniques
Ginestra Bianconi

14.00-14.30 Network integration of multi-omics data for the development of novel tumour targeting strategies
Daniel Remondini

14.30-15.00 Network diffusion-based analysis of genomic data
Ettore Mosca

15.00-15.30 A tale of two networks: two GGMs and their differences
Wessel van Wieringen

15.30-16.00 Coffee/tea break

Session 2: Bioinformatics and Statistics – Chair Jeanine Houwing-Duistermaat

16.00-16.30 Directional mixed effects models for compositional data
Janice Scealy

16.30-17.00 Correlative approaches to intepretation of genetic contributions to disease: Moving beyond the paradigm of functional enrichment
Winston Hide

17.00-18.30 Reception, in the School of Mathematics

DAY 2: TUESDAY 27 JUNE

Session 3: Keynote Lecture 1 – Chair Kanti Mardia

9:30-10.30	The role of formal statistical theory *Heather Battey*
10:30-11.00	Speed poster session I
11:00-11:30	Coffee/tea break

Session 4: Dimension Reduction – Chair John Kent

11:30-12:00	Methods for integrated analysis of omics datasets: prediction and biology *Ruth Pfeiffer*
12:00-12:30	Latent variable modelling for multiple omics data *Hae Won Uh*
12:30-13:00	Cox proportional hazard model with genomic data *Arief Gusnanto*
13:00-13:30	Speed poster session II
13:30-13:45	Workshop photo
13:45-15:30	Lunch break + posters + coffee

Session 5: Machine and Deep Learning – Chair Charles Taylor

15:30-16:00	Comparative analysis of deep learning architectures for multiomic inference *Pietro Lio*
16:00-16:30	Machine learning for precision medicine: some lessons learned *Felix Agakov*
16:30-17:00	Methods for integrated analysis of omics datasets: prediction and biology *Nathan Intrator*
19:00-	Workshop dinner (for MIMOmics members and invited speakers)

DAY 3: WEDNESDAY 28 JUNE

Session 6: Keynote Lecture 2 – Chair Jeanine Houwing-Duistermaat

9:30-10:30	The STATegra road map for the design and analysis of multiomics perturation experiments *Ana Conesa*
10:30-11:00	Coffee/tea break

Session 7: Mediation and Causality – Chair Arief Gusnanto

11:00-11:30	Analysis of metabolite-mediated gene-module co-expression using multivariate linear models *Tomasz Burzykowski*
11:30-12:00	Identification of causal pathways: a Bayesian approach *Huo Guo*
12:00-12:15	Closure
	Lunch

Posters

SPEED SESSION 1:

- Characterization of data features obtained from Nanostring technology
  Alessandra Merlotti
- Statistical modelling of CG interdistance across multiple organisms
  Alessandra Merlotti
- Efficiency and accuracy for Bayesian filtering using directional statistics in the orbital tracking problem
  Shambo Bhattacharjee
- Looking for kinks in protein helices
  Mai Alfahad
- Estimating metabolite networks subject to dietary preferences and lifestyle
  Georgios Bartzis
- Network-based approach for building predition models using multiple omics datasets
  Renaud Tissier
- Network Diffusion Detection of PPI Network Sub-modules Enriched in Altered Gene in Acute Myeloid Leukemia Datasets
  Gastone Castellani

SPEED SESSION 2:

- Lifting on clustering
  Nebahat Bozkus
- Particle Monte Carlo methods to integrate multiple datasets
  Nathan Cunningham
- Probabilistic integrative analysis of two datasets with PO2PLS
  Said el Bouhaddani
- Integrating multiple datasets with the OmicsPLS R package
  Said el Bouhaddani
- Investigation in Filtering Structure of Partial Least Squares for High-Dimensional Datasets
  Mohammed Abdullah Alshahrani
- Random Forest for Omics Data: classification and feature selection
  Umashanger Thayasivam
- Rheumatoid Arthritis Driven Alteration in T-cell Epigenetic Programing
  Rujiraporn Pitaksalee

Welcome

Welcome to the joint meeting of the EU funded MIMOmics consortium, which develops methods for the integrated analysis of multiple omics datasets, and the 34th Leeds Annual Statistical Research Workshop. We are happy to offer you an interesting programme with contributions in omics research from Biostatistics to Deep Learning presented by excellent speakers. We are delighted that the Vice Chancellor of the University of Leeds, Sir Alan Langlands, will be opening this Workshop. This workshop is sponsored by the MIMOmics project.

The Leeds Workshop is a well-established annual statistical meeting and was founded more than four decades ago by Kanti Mardia. We celebrated this milestone in the 2015 LASR, where it all started in 1975 when Leeds hosted a research symposium where distinguished and prominent figures in statistics attended. Although begun as an internal event, it now enjoys the support of a number of visitors each year from other universities and research bodies around the world. Over recent years, the theme of the workshop has reflected the growing, but not exclusive, departmental interest and expertise in bioinformatics, image and shape analysis, wavelets to data analytics and statistics for big data.

This year is the centenary of D'Arcy Thompson's book "On Growth and Form". This book kick-started the subject of Shape Analysis so Kanti Mardia will pay tribute to DArcy Thompson in his talk and give a glimpse of how influential this pioneering book has been in the development of the subject. Indeed, many of the LASR Workshops have focussed on this area. One of the most memorable workshop was in 1995 where almost all world experts in shape analysis participated. Also, this is the first time we prepared a LASR Proceeding and it was dedicated to the pioneers David Kendall and Fred Bookstein of modern shape analysis.

The MIMOmics project, funded by EU in 2012, is a consortium of experts in data science (machine learning, bioinformatics, biostatistics), epidemiology and high throughput omics measurements. The input of expertise beyond data analytics is essential: our methods will acknowledge the measurement error structure and will also answer relevant epidemiological questions.

Analysis of novel omics datasets is challenging. Within MIMOmics we have access to unique metabolomics and Glycomics datasets. For Glycomics technical improvements are ongoing and preprocessing of the data needs to be defined. Borrowing methods from other omics fields, MIMOmics developed standard operation procedures and guidelines for batch correction and normalization. Overlap between technical platforms has been assessed based on biological knowledge and by estimation of common latent factors. Appropriate handling of omics data will increase efficiency and interpretability when analyzing the data with regard to outcome variables.

During this workshop we will share the result with regard to methods for data analysis which have been developed over the last five years of research. Our results in epidemiology will be presented in a workshop in Cambridge, August 21st to 23rd, 2017.

It is really pleasing that we succeeded in creating a very attractive line-up and we hope that you will enjoy our programme.

Jeanine Houwing-Duistermaat and Kanti Mardia

The MIMOmics project has received funding from the European Unions Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 305280

Participants at the LASR Workshop 2015

Members of the MIMOmics network

Information Sheet

Locations

- All talks will take place in room 1.33 in the Maurice Keyworth Building, Leeds University Business School. Breaks will take place in the room next door (1.32).
- A map showing the main locations for the conference can be found on the next page. A Google map is also available at `https://drive.google.com/open?id=1vBOoyVvC8OBKlmbNaNyEzirkDzU&usp=sharing`

  All the locations in Leeds City Centre are within walking distance.
- The Reception on Monday will take place in the School of Mathematics.
- The workshop dinner on Tuesday evening is for invited speakers and MIMOmics members, and will be in University House.

Talks and Posters

- Presenters of talks should leave at least 5 minutes for discussion. Session chairs are asked to keep speakers to the timetable.
- We remind presenters that we have a wide ranging, interdisciplinary audience and talks should be given at an appropriate level.
- Room 1.33 is equipped with a PC projector and a computer. We encourage presenters to load their talks on to the computer as soon as possible by contacting Hae-Won Uh.
- We encourage presenters to display their posters for the entire duration of the conference (if possible). On Tuesday 27 June, there will be a dedicated poster session in room 1.32 – presenters are asked to be available by their posters during this session. There will also be two speed poster sessions beforehand.

Local Information

- **WiFi Access at the University:** If your home institution is a member of eduroam, you will be able to use that service on the University Campus. Temporary University of Leeds usernames will also be available throughout the conference please see Jessica Brennan for more information.
- **Banks and Shops:** There are cash points located within the Student Union building situated adjacent to the Refectory on the University campus. The Student Union building also houses a number of shops, bars and coffee bars, and Essentials mini-supermarket selling newspapers, magazines, stationery, drinks, sandwiches, snacks and confectionery.
- **Smoking:** All meeting rooms, lecture theatres, foyers, public areas, bars, doorways, entrances and bedrooms within the University operate a no smoking policy.
- **Safety and Security:** In the unlikely event of a fire alarm, please follow the guidance of the fire wardens and leave the building by the nearest exit.

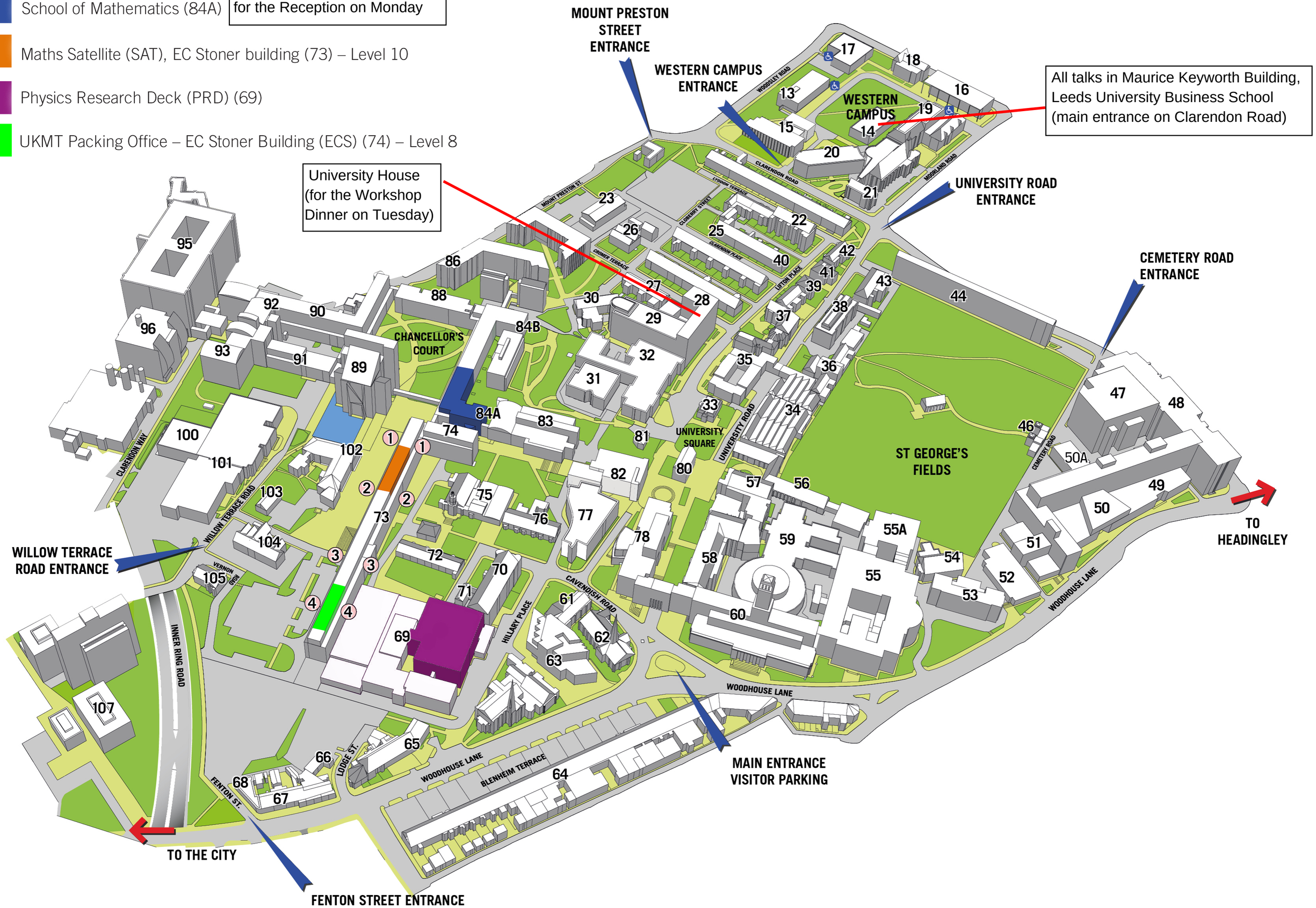
  Please do not leave your belongings unattended as we cannot guarantee their safety.
- **Health and First Aid:** If first aid is required on campus, please contact a member of staff in the building or for emergencies, call Security via an internal telephone on x32222 or externally on +44(0)113 343 2222 - available 24-hours.
- **Useful Telephone Numbers:**
  Jessica Brennan, +44 (0)113 343 5116
  School of Mathematics Reception, +44 (0)113 343 35130

School of Mathematics (84A) for the Reception on Monday

Maths Satellite (SAT), EC Stoner building (73) – Level 10

Physics Research Deck (PRD) (69)

UKMT Packing Office – EC Stoner Building (ECS) (74) – Level 8



ABSTRACTS

Some latest cutting age applications of shape analysis, and the centenary of D'Arcy Thompson's "On Growth and Form"

Kanti V. Mardia

There have been ongoing modern developments in statistical shape analysis in particular and statistics on manifolds in general, motivated by some challenging applications. The aim of this talk is to provide three case studies demonstrating the effective use of statistics of shape analysis in health and medical science which have come through our international, industrial and interdisciplinary collaborations.

My first case study is concerned with saving lives by analysing the shape of the brain. Specifically, our methods have been used on brain images to assess the extent of brain damage in people suffering from fetal alcohol spectrum disorders (FASD) and in turn our assessments have been used in court cases related to the death penalty for murderers. This work continue to have a very high societal impact.

Secondly, we are currently working with on analysing 4-D images related to cleft-lip reconstructive surgery and I will describe its potential to have a very high societal impact. One measure of success for the surgery in this case is the ability of the patient to make "normal-looking" facial expressions, such as smiling. At the moment this judgment is made subjectively. The aim of the research is to develop and assess more objective measures. I will describe from a pilot study that the smile can be characterised by two separate movements which are constant for normal people.

Thirdly, it is well known in Bioinformatics that the existence of a kink in an $\alpha-$helix can play a central role in the function of membrane proteins and thus in turn for drug discovery. We have formulated a parametric statistical model which incorporates the cylindrical nature of the helix and, I will describe a change point technique "Kink-Detector" to find the presence of a kink and its location along the helix. Our biological building block is crowdsourcing data, which consists of straight helices and kinked helices; this data has set a gold standard.

D'Arcy Thompson in his 1917's book "On Growth and Form" (in particular, chapter XVII On the Comparison of Related Forms) foreshadowed a range of biometric and bioinformatic approaches to shape data that have come to fruition only now, a century later, when data processing could finally catch up to his imagination. We will highlight his contributions through the case studies described above.

Department of Statistics, University of Leeds, Leeds; and
Department of Statistics, University of Oxford, UK
k.v.mardia@leeds.ac.uk

Multilayer networks methods and techniques

Ginestra Bianconi

The spectacular advancement made in these last decades on high-throughput molecular biology has resulted in an increasing number of biological network data. This includes co-expression networks, protein interaction networks, gene interaction networks and transcription networks. Currently, the computational methods used to analyse biological data apply extensively tools and methods coming from network science. However most approaches are focused in analysing single biological networks while the integration of multiple dataset is necessary to fully take advantage of the data deluge in molecular biology. Recently significant progress has been made on multilayer networks. These are networks formed by several interacting networks and constitute a powerful tool to integrate information coming from different datasets.

In this talk I will give an overview of the newly developed tools in multilayer networks and I will show evidence that with a multilayer analysis it is possible to extract from datasets more information than the one present on single networks taken in isolation.

School of Mathematical Sciences, Queen Mary University of London, UK
ginestra.bianconi@gmail.com

Network integration of multi-omics data for the development of novel tumor targeting strategies

Giulia Menichetti[1],
Gastone Castellani[2],
Daniel Remondini[2]

High-throughput molecular pro ling has changed the approach to study cancer: the genomic pro ling of tumour samples has revealed differences and similarities that go beyond the histopathological classification. Notwithstanding the enormous increase of knowledge on tumour processes, actually, a practical application of this knowledge to new treatment strategies has not advanced with the same pace.

The huge amount of heterogeneous types of data for a large number of tumours requires novel approaches capable to integrate such information into a uni ed framework: for this aim, we propose a study of gene networks based on expression pro ling and mutational data extracted from The Cancer Genome Atlas (TCGA) [1], in combination with cancer-specific functional annotation described in the Ontocancro database [2], and mapped onto the BioPlex protein-protein interaction network [3]. A structural analysis of the obtained networks, based on node centrality, allowed us to rank their relevance and to obtain specific signatures, that may provide multi-tumour drug targets, prognostic markers, and a molecular taxonomy for effective cancer categorization.

Our approach allowed to define and to prioritize gene targets: we proposed new pharmacological strategies that were validated by in vitro experiments, that showed inhibition of cell growth in two tumour cell lines with significant synergistic effects. This study thus provides a list of genes and pathways with the potential to be used, singularly or in combination, for the design of novel treatment strategies.

References

[1] HTTPS://CANCERGENOME.NIH.GOV,

[2] HTTP://ONTOCANCRO.INF.UFSM.BR,

[3] HTTP://BIOPLEX.HMS.HARVARD.EDU

[1]Northeastern University, Center for Complex Network Research, Boston, Massachusetts, USA
menicgiulia@gmail.com

[2]Department of Physics and Astronomy, Bologna University, IT
gastone.castellani@unibo.it, daniel.remondini@unibo.it

Network diffusion-based analysis of genomic data

Ettore Mosca

The information on direct and indirect molecular interactions can be modelled defining genome-scale networks. The integrated analysis of "-omics" measurements and molecular interactions, referred to as network-based analysis, provides several opportunities for a better interpretation of omics data, which is often hindered by biological complexity and experimental biases. In comparison to the independent analysis of every statistical unit, network-based quantities interpret omics data taking into account the modular and functional architecture of cells. Network-based approaches have been proposed in relation to several problems, including gene module identification, pathway analysis and patient stratification, just to mention a few.

Among network-based approaches, the principle of spreading information throughout a network – namely network diffusion – has been recently used in several applications, mainly related to the "smoothing" of sparse input quantities and the prioritization of molecular entities in network proximity. Here, we describe a network-diffusion based framework for -omics data analysis aimed at identifying gene modules [1]. This framework is based on indices that jointly quantify molecular measurements and network location. The resulting ranked gene list is then analysed to assess the presence of significant subnetworks. After having introduced the method – implemented as an R package named dmfind – and its performance in a controlled scenario, we present the results obtained on prostate cancer molecular profiles (somatic mutations and gene expression) and multiple gene lists associated with autism spectrum disorders.

References

[1] M. Bersanelli*, E. Mosca*, D. Remondini, G. Castellani, L. Milanesi, Network diffusion-based analysis of high-throughput data for the detection of differentially enriched modules, *Scientific Reports* **6** (2016).

Institute of Biomedical Technologies, National Research Council of Italy, Segrate (MI), Italy
ettore.mosca@itb.cnr.it

A tale of two networks: two GGMs and their differences

Wessel van Wieringen

The two-sample problem is addressed from the perspective of Gaussian graphical models (GGMs), in exploratory and confirmatory fashion. The former amounts to the estimation of a precision matrix for each group. This may be done group-wise by means of penalized maximum likelihood with an algebraically proper l2-penalty. But to link the groups the ridge penalty is augmented with a fused term, which penalizes the difference between the group precisions.

The confirmatory part concentrates on the situation in which partial correlations are systematically smaller/larger (in an absolute sense) in one of the groups. Data in both groups again are assumed to follow a GGM but now their partial correlations are proportional, differing by a multiplier (common to all partial correlations). The multiplier reflects the overall strength of the conditional dependencies. As before model parameters are estimated by means of penalized maximum likelihood, now using a ridge-like penalty. A permutation scheme to test for the multiplier differing from zero is proposed.

A re-analysis of publicly available data on the Hedgehog pathway in normal and cancer prostate tissue combines both strategies to show its activation in the disease group.

Department of Mathematics, Vrije Universiteit, Amsterdam, The Netherlands
w.n.van.wieringen@vu.nl

Directional mixed effects models for compositional data

Janice Scealy[1],
Alan Welsh[2]

Compositional data are vectors of proportions defined on the unit simplex and this type of constrained data occur frequently in applications. It is also possible for the compositional data to be correlated due to the clustering or grouping of the observations, for example longitudinal microbiome relative abundance data. We propose a new class of mixed model for compositional data based on the Kent distribution for directional data, where the random effects also have Kent distributions. The advantage of this approach is that it handles zero components directly and the new model has a fully flexible underlying covariance structure. One useful property of the new directional mixed model is that the marginal mean direction has a closed form and is interpretable. The random effects enter the model in a multiplicative way via the product of a set of rotation matrices and the conditional mean direction is a random rotation of the marginal mean direction. For estimation we apply a quasi-likelihood method which results in solving a new set of generalised estimating equations and these are shown to have low bias in typical situations. For inference we use a non-parametric bootstrap method for clustered data which does not rely on estimates of the shape parameters (shape parameters are difficult to estimate in Kent models). The new approach is shown to be more tractable than the traditional approach based on the logratio transformation.

[1]Research School of Finance, Actuarial Studies and Statistics, Australian National University, Canberra, Australia
`janice.scealy@anu.edu.au`

[2]Mathematical Sciences Institute, Australian National University, Canberra, Australia
`alan.welsh@anu.edu.au`

Correlative approaches to interpretation of genetic contributions to disease: Moving beyond the paradigm of functional enrichment

Yered Pita-Juarez[1],
Gabriel Altschuler[2],
Katjusa Kholer[2],
Wenbin Wei[2],
Winston Hide[1,2]

There is a need to move beyond the identification of individual genes associated with a disease. With the goal of using integrative genetic and functional models as a powerful approach to defining therapeutic targets for interventions; Pathway enrichment of identified variants yields a bridge between the data driven statistical association of a variant with a disease and the known function of a pathway defined by a set of genes within it.

Current approaches for understanding directed interaction between pathways rely on shared genes, combining information from databases and interaction networks, or using direct physical interaction between genes and gene products to determine likely interaction.

We have extended this concept to define the relationship between pathways by their co-activity. Systematic quantification of the relationship between pathways provides a high-level map of related cellular functions to reveal the relationship between biological functions by their interactions. We interrogate disease variants in combination with disease gene expression signatures to reveal key interacting pathways enriched with disease variants. We extend genome variant associations in specific pathways enabling analysis of influence of previously unknown pathway relationships. A pathway correlation network (PCxN) reveals co-activity between gene sets for integrative genetic and functional models for experiment or genome association study. PCxN reveals other co-activated pathways seeded by Gene Set Enrichment Analyses and genomic signatures to provide new insight and discovery of pathways influenced by genetic lesions.

[1]Department of Biostatistics, Harvard Chan School of Public Health, Boston, USA

[2]Sheffield Institute for Translational Neuroscience Centre for Genome Translation, University of Sheffield, Sheffield, UK
winhide@sheffield.ac.uk

The role of formal statistical theory

Heather Battey

The need for an eclectic approach to statistical theory is emphasized. Such theory has several roles, the synthesis of previous ideas on design and analysis and the provision of a basis for tackling new issues being the most obvious. Also there may be some tension between the roles of guiding the work of individual investigators and the need to communicate conclusions and the evidence for them in a public way, these calling for different emphases. These rather vague ideas will be illustrated with specific examples.

Imperial College London, London, UK
h.battey@imperial.ac.uk

Methods for integrated analysis of omics datasets: prediction and biology

Ruth Pfeiffer

Only a few procedures have been proposed so far that address how to combine information from multiple correlated markers that are also left and/or right censored due to lower or upper limits of detection. We extend dimension reduction approaches, specifically likelihood-based sufficient dimension reduction (LDR) to regression or classification with censored predictors. These methods apply generally to any type of outcome, including continuous and categorical outcomes. Using an expectation maximization (EM) algorithm, we find linear combinations that contain all the information contained in correlated markers for modeling and prediction of an outcome variable, while accounting for left and right censoring due to detection limits. We also allow for selection of important variables through penalization. We assess the performance of our methods extensively in simulations and apply them to data from a study conducted to assess associations of 47 inflammatory markers and lung cancer risk and build prediction models.

This is joint work with Diego Tomassi, Liliana Forzani and Efstathia Bura E

Senior Investigator/Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, HHS
pfeiffer@mail.nih.gov

Latent variable modelling for multiple omics data

Hae-Won Uh[1],
Said el Bouhaddani[1],
Jeanine Houwing-Duistermaat[2]

Recent advances in life science and technology have enabled the collection of a virtually unlimited quantity of data from multiple sources. The 'systems biology' approach often involves parallel investigation of, for example, transcriptomic, proteomic, or metabolomic datasets. It generally consists of two steps: each of the data sources are analysed in parallel, and the results are summarized to outline novel findings or hypotheses based on prior knowledge of existing pathways. Instead of analysing each dataset in parallel, we consider simultaneous approaches to integrate multiple omics datasets to unravel underlying latent relationships.

Omics datasets are noisy, heterogeneous, highly correlated within each dataset and among the datasets, and often high-dimensional ($p \gg n$). To address high-dimensionality and high correlation dimension reduction techniques can be applied. To unravel underlying biologically relevant information, latent variable modelling such as common factor analysis can be considered. This well-established but computationally intensive method, however, is not equipped to handle high-dimensional data. Dimension reduction as well as data integration can be achieved using a Partial Least Squares regression (PLS) variant. A drawback of PLS is the lack of a probabilistic model that leads to statistical inference and hypothesis testing.

A unified approach of latent variable modelling that has merits of PLS is developed for integration of multiple omics datasets. Various real world applications will be shown.

[1]Department of Medical Statistics
h.uh@lumc.nl, S.el_Bouhaddani@lumc.nl

[2]Department of Statistics, University of Leeds, Leeds, UK
j.duistermaat@leeds.ac.uk

Cox proportional hazard model with genomic data

Khaled Alqahtani[1],
Henry Wood[2],
Charles Taylor[3],
Arief Gusnanto[3]

Copy number alterations (CNA) are chromosomal changes in the genome, where some regions exhibit more or less copy number than the normal two copies. The genome-wide CNA patterns ('profiles') differ from one individual to the next and this profile provides critical information in tumour progression. Hence, it is informative for patients' survival. It is currently a statistical challenge to model patients' survival using their genomic CNA profiles while at the same time identify regions in the genome that are associated with patients survival. Current standard models, for example Cox proportional hazard (PH) model with Lasso penalty, is not appropriate because it totally ignores dependencies between genomic regions and does not produce sensible interpretation of the results. In this talk, we will describe how this challenge is addessed by modifying the assumption on random effects distribution to respect the dependencies between regions. The usual normal penalty in the joint partial likelihood is extended by including the penalty from the assumption on Cauchy and Laplace distribution. The results indicate that a sparse solution is obtained with sensible interpretation. Many genomic regions have parameter estimates at zero, while the other regions have non-zero estimates but with smooth transition across neighbouring adjacent regions. This enables us to identify interesting genomic regions that are related to survival within a single modelling framework.

[1]Department of Mathematics, College of Science and Humanitarian Studies, Prince Sattam Bin Abdulaziz University, Riyadh, Saudi Arabia
`k.alqahtani@psau.edu.sa`

[2]Leeds Institute of Cancer and Pathology, University of Leeds, Leeds LS9 7TF, United Kingdom
`h.m.wood@leeds.ac.uk`

[3]Department of Statistics, School of Mathematics, University of Leeds, Leeds LS2 9JT, United Kingdom
`charles@maths.leeds.ac.uk`, `A.Gusnanto@leeds.ac.uk`

Comparative analysis of deep learning architectures for multiomic inference

Pietro Lio

The aims of the talk is to describe the design and implementation of different types of neural network architectures to perform inference on gene expression and methylation data. The models were chosen such that each of them explores different properties of the epigenetic data. All of these models were subsequently compared in order to assess their pattern recognition abilities. The talk will contain elements of tutorial so also beginners willing to practice on deep learning could benefit by bringing their laptops.

Computer Laboratory, University of Cambridge, Cambridge, UK
pl219@cam.ac.uk

Machine learning for precision medicine: some lessons learned

Felix Agakov

We will present several machine learning methods we developed for addressing disparate tasks of molecular biology and precision medicine, highlighting both successful and unsuccessful cases. We will start by showing applications in data preprocessing and signal extraction, where we were able to achieve the state-of-the-art performance at a fraction of the time and costs of the current industry standards. We will then discuss several approaches to predictions of phenotypes from molecular data, covering discriminative mixtures and network-based classifiers. We will conclude by showing how an approach combining machine learning and epidemiology methods can advance our understanding of causal inference.

We will summarize some challenges and lessons learned while developing machine learning methods for precision medicine, and suggest several possible ways forward.

Pharmatics Limited, Edinburgh, UK
felix@pharmaticsltd.com

Methods for integrated analysis of omics datasets: prediction and biology

Nathan Intrator

The talk will discuss current and future ways in which the recent advances in AI will impact proteomics.

Few examples of "exotic" architectures will be discussed.

AcousticView Ltd, Tel Aviv, Israel
nathan@acousticview.com

The STATegra road map for the design and analysis of multiomics perturbation experiments

Ana Conesa, on behalf of the STATegra Consortium

The omics technologies have evolved over the last two decades to offer an array of platforms that are able to measure different types of molecular features on a cell-wide scale. The combination of several of these methods has been instrumental in profiling mechanisms that link gene expression with other regulatory layers and we are currently witnessing a dramatic increase in the number of studies where multiple omics technologies are combined. The multiomics approach aims at providing a more complete description of cellular components and reveal systems interactions that cannot be fully understood using only one type of measurement. Moreover, the integration of biomarkers from multiple molecular layers may potentially increase our capacity for developing tools for precision medicine.

However, while the multiomics approach is becoming a common strategy in genomic research, there is still a lack of clear guidelines to address the particular problems of multiomic projects. Experimental design, integrative analysis and visualization of multiomic data pose a number of challenges that were not present in studies where only one type of measurement was considered. For example, are sample size requirements the same across omics platforms? how do we deal with many-to-many possible relationships among omics features? how does the coverage of the feature space in each technology compare and how does this impact analysis? how do we represent data where chromatin and pathway level information needs to be combined? Motivated for these questions and the lack of software and guidelines for the analysis of multiomics data the STATegra project was develop. In my talk I would present STATegra efforts to propose a road map for multiomic projects. We highlight six different steps involving experimental design, preprocessing and explorative analysis, matching omics features, statistical methods for data integration to define gene expression regulatory models, visualization and validation. At each step, I will present specific solutions developed by the STATegra Multiomic Data Analysis Project. We show case these methods using a controlled experimental system mimicking a defined stage in the differentiation of the B-cell lineage in mouse and created a time course multiomic dataset comprising up to 8 different platforms. I hope this presentation will provide lights on the complexity, the possibilities and the existing challenges for applying multiomics approaches to solve basic and practical biological questions

Genomics of Gene Expression Lab., Principe Felipe Research Center, Spain
aconesa@cipf.es

Analysis of metabolite-mediated gene-module co-expression using multivariate linear models

Tomasz Burzykowski[1],
Trishanta Padayachee[2]

The current widespread availability of multiple "omics" technologies is particularly beneficial for uncovering the complex interplay between genes and their cellular environments. One way to enhance our understanding of the development and progression of complex diseases is to investigate the regulatory mechanisms behind gene co-expression (correlation). Often, changes in gene co-expression are investigated across levels of a covariate like, e.g., metabolite levels. We present an approach to study the relationships between gene-module co-expression and covariate(s) by using general linear models (GLMs) for correlated data. The use of GLMs allows studying the effect of covariates of different nature (categorical, continuous). In case of continuous covariates, different patterns (linear and non-linear) of co-expression can be studied. Furthermore, the modelling approach offers a formal framework for testing the significance of the observed pattern(s) of co-expression. In our paper, we discuss the theoretical issues related to the construction of the test statistics and the computational challenges related to fitting the GLMs. The versatility of the approach is illustrated by using a real-life data on the co-expression of genes from the core LL module and several metabolites [1].

References

[1] T. Padayachee, T. Khamiakova, Z. Shkedy, M. Perola, P. Salo, T. Burzykowski, The detection of metabolite-mediated gene module co-expression using multivariate linear models, *PLOS ONE* **11**, e0150257.

[1]I-BioStat, Hasselt University, Diepenbeek, Belgium
`tomasz.burzykowski@uhasselt.be`

[2]I-BioStat, Hasselt University, Diepenbeek, Belgium
`trishanta.padayachee@uhasselt.be`

Identification of causal pathways: a Bayesian approach

Hui Guo[1],
Carlo Berzuini[1],
Luisa Bernardinelli[2],
Jeanine Houwing-Duistermaat[3]

Genome-wide association studies have found many genetic variants, in particular, single nucleotide polymorphisms (SNPs), associated with certain clinical outcomes. Some of these SNPs are also associated with modifiable exposures. This overlap can be leveraged to learn about the biological mechanisms underlying these outcomes. Interventions can then be tailored on the basis of well-targeted causal pathways. Existing state-of-the-art methods including Mendelian randomization (MR) and statistical colocalization make important steps towards genetic causal inference.

Both MR and colocalization exploit results from two association studies: SNP-exposure and SNP-outcome. However, they are different in several ways. MR is tailored to identify a causal effect of the exposure on the outcome. Thus, SNPs used as instruments need not be causal for either. Colocalization is designed to identify common SNPs which are causal for both. MR uses exposure associated SNPs as instruments, while colocalization often starts with outcome associated regions. Both approaches can be applied to the scenario which comprises three essential components: SNPs, exposure and outcome. Colocalization can also be used to test for common causal signals of multiple variables that occur in no particular temporal order.

Taken together, there is a need to take forward strengths of the two methods in a unifying approach. We will focus on applying Bayesian colocalization [1] and Bayesian MR [2] to identification of genetic causal pathways.

References

[1] H. Guo, M. D. Fortune, O. S. Burren, E. Schofield, J. A. Todd and C. Wallace, Integration of disease association and eQTL data using a Bayesian colocalisation approach highlights six candidate causal genes in immune-mediated diseases, *Human Molecular Genetics* **24** (2015), 3305–3313.

[2] C. Berzuini, H. Guo, S. Burgess and L. Bernardinelli, A Bayesian Framework for Multi-Instrument Mendelian Randomisation in the Presence of Pleiotropy, submitted to *Biostatistics* (2017).

[1]Centre for Biostatistics, The University of Manchester, UK
`hui.guo@manchester.ac.uk`, `carlo.berzuini@manchester.ac.uk`

[2]Department of Brain and Behavioural Sciences, University of Pavia, Italy
`L.Bernardinelli@statslab.cam.ac.uk`

[3]School of Mathematics, University of Leeds, UK
`j.duistermaat@leeds.ac.uk`

POSTERS

Characterization of data features obtained from Nanostring technology

Alessandra Merlotti,
Daniel Remondini,
Gastone Castellani

We describe the main features and distribution underlying the data generated by Nanostring technology, that can be related to RNA-seq or SAGE data. In particular, we consider the role of noise as a function of transcript level in both signals and control data, consisting of no signal, housekeeping genes and known transcripts with progressive diluition. The data seem to show clear markings related to a specific statistical distribution.

In particular we found that the logarithm of transcript levels is normally distributed, as proved by the $y = x$ trend of the data in quantile-quantile plots. We also analyze the relationship between the variance and the mean, showing that they are independent. Furthermore, we show that the logarithm of transcript levels with progressive diluition follows a linear trend, except for low concentration levels, as confirmed by linear regression results obtained using the least squares method.

Acknowledgments We thank Dr. Alessia Ciarrocchi from IRCCS for providing the data.

Department of Physics and Astronomy, University of Bologna,
Viale B. Pichat 6/2, 40127 Bologna, Italy
daniel.remondini@unibo.it

Statistical modelling of CG interdistance across multiple organisms

Alessandra Merlotti,
Ítalo Faria do Valle,
Daniel Remondini,
Gastone Castellani

CG dinucleotides play an important role inside human genome since almost the 80% of them undergoes methylation, a fundamental epigenetic mechanisms involved in gene regulation and structural conformation of chromatin. This peculiar functional and structural role might be reflected in their interdistances distribution, which shows very different features from non-CG distributions, especially among mammals.

We therefore selected a set of 18 different organisms in order to characterize CG interdistances distributions, using an empirical approach. We considered four different types of models: exponential, double exponential, stretched exponential and gamma distribution. We thus fitted the data in semilogarithmic scale using non-linear least squares method and we estimated the goodness of fit analyzing residuals plots and calculating p-value through a test based on resampling technique and Kuiper's statistic.

In order to understand which process could give rise to such a difference between CG and non-CGs distributions in mammals, we elaborated a null model based on the following hypothesis: originally CG distribution followed the same trend of non-CGs and only because of random mutations it differentiated from the others.

Finally, we extended the analysis of CG interdistances distribution to 4425 genomes of different organisms downloaded from GenBank database and belonging to 7 different categories: invertebrates, vertebrates non-mammals, vertebrates mammals, bacteria, fungi, protozoa and plants.

Best fit results show that mammals CG interdistances distributions are well described by a gamma distribution; furthermore, the application of the null model shows that, given a non-CG distribution, it is always possible to obtain a new distribution similar to that of CG, by randomly changing a specific number of non-CG dinucleotides into another. Finally, fitting CG interdistances distributions of all 4425 organisms to a gamma distribution and plotting its 2 parameters, we find 7 clusters corresponding to the 7 considered categories; we therefore believe that this method can provide new tools for organisms characterization and classification, such as taxonomy.

Department of Physics and Astronomy, University of Bologna,
Viale B. Pichat 6/2, 40127 Bologna, Italy
daniel.remondini@unibo.it

Efficiency and accuracy for Bayesian filtering using directional statistics in the orbital tracking problem

John T. Kent,
Shambo Bhattacharjee

The problem of space debris tracking from a sequence of observations can be viewed as an example of Bayesian filtering. The state vector describes the position and velocity of the space object, and can be represented as a 6-dimensional vector in Cartesian coordinates. The observation vector consists of an angles-only measurement, possibly plus range, a 2- or 3-dimensional vector. Filtering is simplest if the joint state and observation vectors follow a multivariate normal distribution. However, even if the initial uncertainty of the state is normally distributed, the propagated orbital state several periods into the future quickly becomes non-normal, with the distribution of the position vector having a distinctive "banana" shape in $\mathbb{R}^3$.

One solution to the filtering problem is to use a particle filter, but this approach can suffer from the curse of particle depletion. Another solution is to switch to an "adapted structural" coordinate system, where the joint distribution is much closer to normal. The phrase "adapted" means that the coordinate system depends on the data. The key step is to note that the state vector can also be described in terms of an ellipse (5 degrees of freedom) plus the position of the space object along the elliptical orbit (a one-dimensional angular or directional variable, namely, the mean anomaly). In particular, under idealized Keplerian dynamics, the ellipse parameters are constant in time and the true anomaly is an angle travelling around a unit circle. In this paper we demonstrate the usefulness, efficiency and accuracy of this approach to the filtering problem.

In addition, this approach will be extremely useful for IOD methods, the backward orbit determination problem, the data association problem and the light curve tracking problem.

Department of Statistics, University of Leeds, Leeds, UK
`j.t.kent@leeds.ac.uk`, `mmsb@leeds.ac.uk`

Looking for kinks in protein helices

Mai Alfahad,
John Kent,
Kanti Mardia

n this study we establish a method to determine if a given protein α-helix is kinked or not; and if it is kinked, then a way to find the kink position.

If the helix in known to be kinked, then the kink position can be found, so that we can fit each unkinked half of the helix individually.

Moreover, we specify a test statistic by simulation the null distribution of the unkinked helices, where a threshold value can be selected as the value of test statistics at significance level $\alpha = 0.05$. This allows to decide if the given helix is kinked or not if we were unsure to begin with.

Furthermore, if we assume the kink position is known, then the test statistic under the null hypothesis approximately follows an F-distribution.

Department of Statistics, University of Leeds, Leeds, UK
mmmfa@leeds.ac.uk, J.T.Kent@leeds.ac.uk, K.V.Mardia@leeds.ac.uk

Estimating metabolite networks subject to dietary preferences and lifestyle

Georgios Bartzis[1,2],
Kate Xu[1,3],
Hae Won Uh[1],
Fred van Eeuwijk[2]
Jeanine J. Houwing-Duistermaat[1,4]

In systems biology, when the metabolome is modeled, often the interest needs to be focused on variation coming from a certain source. In DILGOM study, metabolite concentration levels have been measured in 2 time points allowing the estimation of subject specific effects by capturing the correlation coming from repeated measurements by using a linear mixed model. These subject specific effects, represent dietary choices, lifestyle, genetics etc.

Although, several studies have examined the interplay between diet and metabolism, to date, no studies have investigated how metabolic patterns are influenced by dietary variation ($\boldsymbol{F}$) and lifestyle ($\boldsymbol{u}$) while genetic contribution ($\boldsymbol{G}$) and time ($\boldsymbol{T}$) have been modeled in the context of network analysis. By fitting a linear mixed model for each metabolite the total metabolic variation is decomposed into parts relevant to $\boldsymbol{F}$, $\boldsymbol{G}$, lifestyle $\boldsymbol{u}$, and $\boldsymbol{T}$; the part relevant to $\boldsymbol{F}$ and $\boldsymbol{u}$ instead of the original values is further used for network estimation. Dietary information resulted from Food Frequency Questionnaires and Genetic information is quantified by using Exploratory Factor Analysis and Polygenic Risk Scores, respectively

For estimating, describing and visualizing networks, a correlation-based network estimation method (WGCNA) is used. In the resulting networks, several groups of biologically associated metabolites (VLDL, HDL, AA/BCAA, *omega*-3 FA) were clustered together based on their association to 6 known diets. The novelty of our method is on taking into account all sources of metabolic variation and resulting in networks with higher interconnectedness and interpretability, meaning identifying meaningful metabolite groups sharing similar association to $\boldsymbol{F}$ and $\boldsymbol{u}$.

[1]Department of Medical Statistics, Leiden University Medical Center, Leiden, the Netherlands
G.Bartzis@lumc.nl, H.Uh@lumc.nl

[2]Biometris, Wageningen University and Research Center, Wageningen, the Netherlands
fred.vaneeuwijk@wur.nl

[3]Department of Pedagogical and Educational Sciences, Erasmus University Rotterdam
man.kate.xu@fsw.eur.nl

[4]Department of Statistics, University of Leeds, Leeds, UK
j.duistermaat@leeds.ac.uk

Network-based approach for building predition models using multiple omics datasets.

Renaud Tissier[1],
Mar Rodríguez Girondo[1],
Jeanine Houwing-Duistermaat[2]

Nowadays, in many studies several omics datasets are available for analysis. Because of the presence of correlation between the omics variables and the large number of variables, regularized regression techniques are typically used for this purpose. If theses techniques can perform quite well to build prediction models on each omics datasets separately, this is not the case when stacking omics datasets with different sizes, scales, structures and measurement errors. Another drawback of these methods is that the results might be hard to interpret by biologists and epidemiologists especially when the variables are correlated as it is the case in most omics datasets. Our goal is to build models which have a good predictive ability and are also biological interpretable.

We propose a three-step approach: namely 1) network construction per omics dataset, 2) clustering to empirically derive modules within omics and across omics datasets, and 3) building a prediction model, where we use the information on the modules. For the first step we use a commonly approach based on weighted correlation. Identification of modules (groups) of features is performed by hierarchical clustering. To incorporate the grouping information in a prediction model we conduct group-based variable selection with group-specific penalization parameters.

We compare the performance of our new approaches with standard regularized regression (LASSO and Ridge) approaches via simulations. To qualify performance we use cross-validated calibration measures and variable selection properties. Finally our approaches are applied to two different studies with omics datasets. Namely metabolomics and gene expression available in an epidemiological study with Body Mass Index as outcome variable and CNV and gene expression available in cell lines with drug response as outcome variable. The prediction tasks differ because the correlation between metabolomics and gene expression is relatively small, while it is large between CNV and gene expression.

[1]Department of Medical Statistics
r.l.m.tissier@lumc.nl, M.Rodriguez_Girondo@lumc.nl

[2]Department of Statistics, University of Leeds, Leeds, UK
j.duistermaat@leeds.ac.uk

Network Diffusion Detection of PPI Network Sub-modules Enriched in Altered Gene in Acute Myeloid Leukemia Datasets

Matteuzzi Tommaso[1]
Italo Faria Do Valle[1]
Enrico Giampieri[1]
Daniel Remondini[1]
Matteo Bersanelli[2]
Ettore Mosca[2]
Luciano Milanesi[2]
Gastone Castellani[1]

Within the cell, functional similarity between molecular entities is associated to network proximity and entities involved in the same disease are more likely to interact with one another. Starting from a set of query genes, such as the set of altered genes in cancer, made available by high-throughput experiments, our aim is to find a set of other genes related to these, ranked according to their proximity to the query set.

Network diffusion methods exploits the global structure of a network by simulating the behavior of a random walker. We exploit a novel method, recently proposed in literature, for detection of sub-modules of the BioPlex PPI network, enriched in altered genes in Acute Myeloid Leukemia.

The method is organized in two part. For the first, we select a set of source genes and we associate to each of them an initial information which reflects its *"degree" of alteration*, *i.e.* gene expression fold change or frequency of mutation, compared to a set of controls. The diffusion process propagates such information within the network and allow us to define a network-based ranking of the BioPlex genes, according to their proximity to the sources.

The enriched sub-modules identification is carried out by a network resampling procedure, based on the minimization of an objective function, starting from the network-based ranking.

The method allow us to retrieve some genes that in literature are already associated to Acute Myeloid Leukemia and new genes that are likely to have a role in this pathology because of their proximity to the sources. Moreover, we assess, by a pathway enrichment procedure, the association of the detected enriched sub-modules to cancer related pathways.

[1]Department of Physics, University of Bologna, Bologna, Italy
`gastone.castellani@unibo.it`
`tommaso.matteuzzi@studio.unibo.it`

[2]Institute of Biomedical Technologies, CNR Via F.lli Cervi 93, 20090, Segrate (MI), Italy

Lifting on Clustering

Nebahat Bozkus

One of the popular questions in hierarchical clustering is how many clusters exist in a data set, or where to 'cut the tree'. Even though many methods have been proposed, this topic still attracts the interest of researchers. Previous indices capture the number of clusters quite well if clusters are well separated, but when the clusters overlap or have unusual shapes, their performance deteriorates. I propose a new method based on a multiscale technique called lifting which has recently been developed to extend the 'denoising' abilities of wavelets to data on irregular structures.

In my method, lifting is applied to the structure of a dendrogram. I then assume that the distances between data points and cluster centroids are affected by noise. Denoising the mean distances of data points from each cluster centroid can help me decide where to cut the tree. This method will be illustrated with both real and simulated examples.

Department of Statistics, University of Leeds, Leeds, UK
mmnb@leeds.ac.uk

Particle Monte Carlo methods to integrate multiple datasets

Nathan Cunningham[1],
David Wild[1],
Jim Griffin[2]

Cluster analysis is a popular technique for discovering group structure in a dataset. While it has been applied in a number of fields, it is often applied in clustering genomics data where identifying groups of genes can help uncover gene function and discover regulatory networks. While modern, high-throughput technologies produce a vast array of data from disparate sources, classic clustering algorithms, such as k-means and hierarchical cluster analysis, are typically geared towards continuous data and are not suitable for mixed data types.

Previous work on integrating datasets, multiple dataset integration (MDI), employed a Bayesian non-parametric model, where observations were assigned to clusters using a one-at-a-time approach. This was found to be able to uncover information which would be difficult or impossible to uncover from approaches considering a single dataset. However it was found to have slow mixing properties. We present an extension to the original MDI algorithm where cluster allocations are determined using a particle Gibbs approach, which has been proposed as a means of escaping this problem.

[1]Department of Statistics, University of Warwick
n.cunningham@warwick.ac.uk

[2]Department of Statistics, University of Kent

Probabilistic integrative analysis of two datasets with PO2PLS

Said el Bouhaddani[1],
Hae Won Uh[1],
Geurt Jongbloed[2],
Jeanine Houwing-Duistermaat[3],

In biomedical research it is becoming standard to have access to multiple datasets on the same set of subjects. These data are characterized by high correlations between measurements and high dimensionality. Due to these characteristics traditional methods fail and parallel analyses of each dataset separately miss out important relationships between datasets. Therefore to understand more complex relationships between the underlying variables integrative analyses of these data are needed. This need has motivated development of several data integration methods, in particular latent variable regression models such as Partial Least Squares (PLS). However the majority of these methods focus on estimating joint parts across datasets, ignoring the presence of data-specific parts. Taking into account specific parts is vital, as these parts might contain important information about technical or biological variation.

Few approaches have been developed to model and estimate both joint and specific parts in the data, among them Two-way Orthogonal PLS (O2PLS). Common issues underlying these approaches are unidentifiability of the model parameters and a lack of a probabilistic model. This hampers generalizing the results to other studies and accounting for typical epidemiological issues such as missing data and heterogeneity across subjects.

We propose Probabilistic O2PLS (PO2PLS), a latent variable approach modelling data-joint and data-specific variation in two datasets. The likelihood formulation provides opportunities to address issues such as missing data and heterogeneity between subjects. The PO2PLS model is identifiable under standard constraints on the parameters. These constraints are incorporated in a constrained EM algorithm to obtain maximum likelihood estimates.

Results of a simulation study will be presented comparing PO2PLS with state-of-the-art methods. Data from a Croatian cohort will be analysed illustrating the PO2PLS model.

[1]Department of Medical Statistics and Bioinformatics, LUMC, Leiden, Netherlands
s.el_bouhaddani@lumc.nl, h.uh@lumc.nl

[2]Institute of Applied Mathematics, TU Delft, Delft, Netherlands
g.jongbloed@tudelft.nl

[3]Department of Statistics, University of Leeds, Leeds, UK
j.duistermaat@leeds.ac.uk

Integrating multiple datasets with the OmicsPLS R package

Said el Bouhaddani[1],
Hae Won Uh[1],
Geurt Jongbloed[2],
Jeanine Houwing-Duistermaat[3],

An increasingly common goal in biomedical research is the integrative analysis of multiple data sources on the same set of subjects. As a consequence several data integration methods have been developed, focusing on the overlapping part across datasets. However taking into account data-specific parts is equally desirable, as these parts might contain important information, for example on the presence of batch effects.

Approaches that have been developed to model and estimate both joint and specific parts in the data include Two-way Orthogonal Partial Least Squares (O2PLS), DISCO-SCA and JIVE. Both DISCO-SCA and JIVE have unrealistic assumptions and restrictions on the parameters. For example the data sources share the same underlying latent variables and require joint and specific components to be exactly orthogonal. As these assumptions and restrictions are not present in the O2PLS method, we consider O2PLS to decompose each dataset in a joint, specific and residual part. The joint components represent the statistical integration between two data sets, while the specific parts contain variation specific to each dataset and unrelated to the joint components. However applying O2PLS is hampered due to lack of freely available software.

Therefore we introduce OmicsPLS, an easily accessible open source R package (freely available on CRAN) to perform O2PLS and inspect the results. The main functionalities include cross-validating, fitting and plotting O2PLS models. We will discuss several improvements in computational efficiency for high dimensional data as well as various ways to create publication-quality figures of the results. The OmicsPLS package will be demonstrated using transcriptomic and metabolomic data from a freely available population cohort and compared with competing software packages.

[1]Department of Medical Statistics and Bioinformatics, LUMC, Leiden, Netherlands
`s.el_bouhaddani@lumc.nl`, `h.uh@lumc.nl`

[2]Institute of Applied Mathematics, Delft University of Technology, Delft, Netherlands
`g.jongbloed@tudelft.nl`

[3]Department of Statistics, University of Leeds, Leeds, UK
`j.duistermaat@leeds.ac.uk`

Investigation in Filtering Structure of Partial Least Squares for High-Dimensional Datasets

Mohammed Abdullah Alshahrani,
Arief Gusnanto,
Charles Taylor

Recently, high-dimensional datasets have been used particularly in spectroscopy and genomics where the number of variables (p) greatly exceeds the number of observations (n). In near infrared (NIR) spectroscopy, we measure the absorbances of n samples at p different wavelengths and in genomics we estimate the copy number alteration (CNA) at thousands of genomic regions from each individual. Partial least squares regression (PLS) is one of the methods used when $p \gg n$ besides principle component regression (PCR) and ridge regression (RR).

In the case where $n > p$, using shrinkage methods such as RR and the dimension reduction methods such as PLS and PCR, the estimation of coefficients are shrunk from ordinary least squares regression (OLS) solution by an amount called shrinkage factor. However, in the case where ($p > n$), the shrinkage factors might be called filter factors since the OLS solution does not exist.

Moreover, filter factors are similar to the shrinkage factors, and they are between zero and one as in RR and PCR, but in PLS they have strange behaviours where they are isolating around one, and they might be very large or even negative. In this poster, we will discuss the filter factors of all three methods PLS, PCR and RR in high-dimensional data, and more details about the filter factors in PLS especially when they become negative.

Department of Statistics, University of Leeds, Leeds, UK
mmmaa@leeds.ac.uk, A.Gusnanto@leeds.ac.uk, C.C.Taylor@leeds.ac.uk

Random Forest for Omics Data : classification and feature selection

Umashanger Thayasivam

Analyzing biomarkers is a rapidly emerging study that helps answering a wide range of biological questions. Modern biology has experienced an increasing use of data mining techniques for large scale and complex biomarker analysis. One of its greatest potentials is the development of criteria that allow us to diagnose a disease or to classify patients according to their risk for a clinical outcome of interest. Metabolite profiling data and omics data in general, pose several statistical challenges for classification and prediction. Metabolomics, like most omics technologies, suffers from the problem of having huge quantities of data without the ability to efficiently process it and gain valuable knowledge. Random Forest (RF) technique, which includes an ensemble of decision trees and incorporates feature selection and interactions naturally in the learning process, is a popular choice. RF is a popular nonparametric algorithm, which is interpretable, efficient with high prediction accuracy. Recent work in computational biology has shown an increased use of random forest, owing to its unique advantages in dealing with small sample. Furthermore, RF can perform feature selection (Variable importance) which is essential to discover new robust biomarkers being used for patient identification and/or stratification.

The focus of this presentation is twofold. First, to provide an overview of Random Forest (RF) technique in Biomarker discovery, including feature selection. Second, to briefly introduce application of RF by presenting results from different biomarker studies.

Department of Mathematics, Rowan University, New Jersey, USA
thayasivam@rowan.edu

Rheumatoid Arthritis Driven Alteration in T-cell Epigenetic Programing.

Rujiraporn Pitaksalee[1],
Agata N Burska[1],
Joshua Rogers[1],
Paul Emery[1],
Richard Hodgett[2],
Frederique Ponchel[1]

Background Alteration in epigenetic patterns have been related to several diseases including Rheumatoid Arthritis (RA). The goal of our project is to identifying the early change in DNA methylation pattern of naive and memory CD4+T-cells, and monocytes to help understanding early disease pathology and find potential biomarker. Methods The methylation patterns of 480,000 CpGs were measured in 3 cell types (memory T-cell, naive T-cell and monocytes) in 6 healthy control (HC) and 10 RA patients using an Illumina methylation genome-wide array. Standard t-test were performed to associate p-value to individual CpG-probe. Manhattan plot, Hierarchical clustering and Heatmaps and venn diagram were generated using R. Gene annotation and function analysis was performed using Panther analysis. Results Manhattan plot highlighted different levels of significance for CpG probes between HC and RA for 3 cell types. The number of probes differentially methylated varied between cells types with 7,209 probes for monocytes, 15 849 and 20 631 for memory and naive T-cells. Distinct clustering between HC and RA was observed for the 3 cell types, with major hyper-methylation in memory T-cells suggesting gene silencing, while under-methylation was prominent in naive T-cells suggestive of gene activation but more equally distributed in monocytes. We designed rules to prioritise CpG for further analysis, based on high significant and likeliness to have a 3-D effect on DNA accessibility by clustering of several differentially methylated CpG in a small DNA region. 1112, 881 and 106 probes from naive memory and monocytes were selected, respectively. Venn diagrams of gene associated with these probe showed 4 genes (ABAT, TYH3, EPS8, and ZBTB45) to be common to all 3 subsets, while 117 genes were only common to the T-cell subsets. The top most important genes appear to be DAXX, HIC1 and HMX2. A Panther analysis of 117 genes between T-cell showed genes to be related to HLA (4.2Conclusion This preliminary analysis demonstrate over and under methylation in several CpG island in early disease. The shared differentially methylated genes between T-cell (but not monocyte) may offer potential for the development of a methylation sensitive specific qPCR biomarker for the early diagnosis of RA.

[1]Leeds Institute of Rheumatic and Musculoskeletal Medicine and NIHR Leeds Musculoskeletal Biomedical Research Unit, University of Leeds, and the Leeds Trust Teaching Hospital, Leeds, UK
umrpi@leeds.ac.uk

[2]Business Analytics and Decision Science, Leeds University Business School, University of Leeds, Leeds, UK

List of Participants

Samira Abushilah, University of Leeds, samirafaisal@yahoo.com

Felix Agakov, Pharmatics, felix@pharmaticsltd.com

Yovaninna Alarcon, Universitat Politecnica de Catalunya, yovi.alarcon@gmail.com

Mai Alfahad, University of Leeds, mmmfa@leeds.ac.uk

Hassan Aljohani, University of Leeds, ml09h3a@leeds.ac.uk

Fatimah Almulhim, University of Leeds, Mmfa@leeds.ac.uk

Mohammed Alshahrani, University of Leeds, mmmaa@leeds.ac.uk

Huda Alshanbari, University of Leeds, ml11h3m@leeds.ac.uk

Stuart Barber, University of Leeds, stuart@maths.leeds.ac.uk

Jenny Barrett, University of Leeds, j.h.barrett@leeds.ac.uk

Georgios Bartzis, Leiden University Medical Center, G.Bartzis@lumc.nl

Heather Battey, Imperial College London, h.battey@imperial.ac.uk

Shambo Bhattacharjee, University of Leeds, mmsb@leeds.ac.uk

Ginestra Bianconi, Queen Mary University of London, ginestra.bianconi@gmail.com

Leonid Bogachev, University of Leeds, L.V.Bogachev@leeds.ac.uk

Nebahat Bozkus, University of Leeds, mmnb@leeds.ac.uk

Tomasz Burzykowski, Hasselt University, tomasz.burzykowski@uhasselt.be

Jacob Cancino-Romero, University of Leeds, mmjcr@leeds.ac.uk

Gastone Castellani, University of Bologna, Gastone.Castellani@unibo.it

Damianos Christophides, University of Leeds, D.Christophides@leeds.ac.uk

Ana Conesa, Centro de Investigacion Principe Felipe, aconesa@cipf.es

Nathan Cunningham, University of Warwick, n.cunningham@warwick.ac.uk

Vinny Davies, LICAP, University of Leeds, v.davies@leeds.ac.uk

Joey Mark Diaz, LICAP, umjmsd@leeds.ac.uk

Cecilia Diaz, University of Manchester, cecilia.diaz@manchester.ac.uk

Elizabeth Duncan, University of Leeds, e.j.duncan@leeds.ac.uk

Said el Bouhaddani, LUMC, s.el_bouhaddani@lumc.nl

Benjamin Eltzner, University of Goettingen, beltzne@uni-goettingen.de

Claire Green, University of Sheffield, cmgreen1@sheffield.ac.uk

Hui Guo, University of Manchester, hui.guo@manchester.ac.uk

Arief Gusnanto, University of Leeds, A.Gusnanto@leeds.ac.uk

Caroline Hayward, University of Edinburgh, Caroline.Hayward@igmm.ed.ac.uk

Winston Hide, University of Sheffield, j.chand@sheffield.ac.uk

Jeanine Houwing-Duistermaat, University of Leeds, J.Duistermaat@leeds.ac.uk

Mark Iles, University of Leeds, m.m.iles@leeds.ac.uk

Nathan Intrator, AcousticView Ltd, nathan@acousticview.com

Lennart Karssen, PolyOmica, l.c.karssen@polyomica.com

John Kent, University of Leeds, j.t.kent@leeds.ac.uk

Lucija Klaric, University of Edinburgh, lklaric@genos.hr

Ruwanthi Kolamunnage-Dona, University of Liverpool, kdrr@liverpool.ac.uk

Bethany Kuszlewicz, University of Leeds, umblk@leeds.ac.uk

Kersty Laing, Leeds University, ed15kml@leeds.ac.uk

Anne Le Maître, University of Vienna, anne.le.maitre@univie.ac.at

SangYu Lee, University of Leeds, mmsl@leeds.ac.uk

Pietro Lio, University of Cambridge, pl219@cam.ac.uk

Haiyan Liu, Department of Statistics, H.Liu1@leeds.ac.uk

Kanti Mardia, Universities of Leeds and Oxford, k.v.mardia@leeds.ac.uk

Ivonne Martin, Leiden University Medical Centre, I.Martin@lumc.nl

Bart Mertens, LUMC, b.mertens@lumc.nl

Luciano Milanesi, CNR, luciano.milanesi@itb.cnr.it

Sarah Morgan, Sheffield University, sarah.morgan@sheffield.ac.uk

Ettore Mosca, CNR-ITB, ettore.mosca@itb.cnr.it

Daniela Nickel, University of Leeds, D.V.Nickel@leeds.ac.uk

Jo Panni, NA, pannijo@hotmail.com

Ruth Pfeiffer, National Cancer Institute, pfeiffer@mail.nih.gov

Rujiraporn Pitaksalee, University of leeds, umrpi@leeds.ac.uk

Frederique Ponchel, LIRMM, mmefp@leeds.ac.uk

Smarti Reel, The Open University, smarti.reel@open.ac.uk

Parminder Singh Reel, University of Dundee, p.s.reel@dundee.ac.uk

Daniel Remondini, Bologna University, daniel.remondini@unibo.it

Janice Scealy, Australian National University, janice.scealy@anu.edu.au

Farag Shuweidhi, Bengahzi University, fstat2005@gmail.com

Georgette Tanner, Leeds Institute of Cancer and Pathology, medgnt@leeds.ac.uk

Charles Taylor, University of Leeds, c.c.taylor@leeds.ac.uk

Rohit Thakur, LICAP, umrth@leeds.ac.uk

Umashanger Thayasivam, Rowan University, thayasivam@rowan.edu

Renaud Tissier, Leiden University Medical Centre, r.l.m.tissier@fsw.leidenuniv.nl

Rahman Uddin, Faculty of Biological Sciences, bsru@leeds.ac.uk

Hae-Won Uh, LUMC, h.uh@lumc.nl

Emeka Uzochukwu, University of Leeds, mmecu@leeds.ac.uk

Wessel van Wieringen, Free University, Amsterdam, w.n.van.wieringen@vu.nl

Dodi Vionanda, University of Leeds, mmdv@leeds.ac.uk

Jochen Voss, University of Leeds, J.Voss@leeds.ac.uk

Fatin Nurzahirah Zainul Abidin, Faculty of Biological Sciences, bs12fnza@leeds.ac.uk

Xin Zhao, Leeds University, mmxzha@leeds.ac.uk