Integrating Systems Biology and Network Science with the Rich Data of the CSIRO Biological Collections

OCE Cutting Edge Symposium

Dec 10-12, 2012

Symposium webiste

Symposium aim and objectives:
The aim of the symposium is to bring together researchers from the disciplines of Network Science and Systems Biology together with collection based researchers with genetic, phylogenetic, genomic, morphological, spatial and ecological data to investigate new opportunities of integrated data analyses. The major objectives are: ● Update researchers on data and analytical advances ● Develop new collaborations among workers in these fields ● Draft a review article on emerging topics in this research area

Format

Three days: Monday - Wednesday

- Morning seminars open to public
  - Rotation of data and analytical talks
- Afternoon working groups
- Monday evening Plenary speaker
- Tuesday evening symposium dinner
- Wednesday afternoon group discussion syntheses

Monday morning

- Introduction 9-9:15
  - **Bronwyn Harch (CMIS Chief)**: Opening Address
- Data 1: 9:15-10:00
  - **Joe Miller (CANBR Systematics and Evolution Group Leader)**: *An Overview of the Data in the Australian National Collections*
- Break 10:00-10:30
- Analytical 1: 10:30-11:15
  - **Grace Chiu (CMIS)**: *Modelling and Uncertainty: An Overview*
- Data 2: 11:15-12:00
  - **Warren Jin (CMIS)**: *Data mining: discovering potential useful knowledge from large data sets*

Monday Afternoon

- Group sessions 2:00-4:30

Monday Evening

- **Plenary: Mark Handcock (UCLA)**: *Statistical Modeling of Networks: Identifying structure from incomplete data*
    - Biography appears before list of abstracts

Tuesday morning

- Analytical 2: 9:00-9:45
    - **Ayesha Ali (U of Guelph)**: *The Econometrics of Ecology*
- Data 3: 9:45-10:30
    - **Justin Borevitz (ANU):** *Landscape Genomes for Climate Adaptation in Plants*
- Break 10:30-11:00
- Analytical 3: 11:00-11:45
    - **Anton Westveld (U of Arizona)**: *Latent space modelling for trophic food webs and other biological networks*
- Data 4**:** 11:45-12:30
    - **Nigel Andrew (UNE):** *Insect responses to climate change: integrating insect ecology, physiology and behavioural responses across populations, species and communities*

Tuesday Afternoon

- Group sessions 2:00-4:30

Tuesday Evening-Dinner (invited)

Wednesday morning

- Analytical 4: 9:00-9:45
    - **Simon Barry (CMIS):** *When did the dodo become extinct: inferring extinction from sighting records*
- Data 5: 9:45-10:30
    - **Mike McLeisch (Xishuangbanna Tropical Botanical Gardens):** *Plants, insects, and space: "With a Little Help from My Friends"*
- Break 10:30-11:00
- Analytical 5: 11:00-11:45
    - **Cang Hui (Stellenbosch University)**: *Building a causal model for the adaptive radiation of Australian trees*
- Data 6: 11:45-12:30
    - **Toni Reverter-Gomez (CAFHS):** *Gene network inference applied to three disparate scenarios: Cow puberty, wool pigmentation and colon cancer*

Wednesday Afternoon

- Group sessions 2:00-3:30
- Closing session 4:00-5:00

Topics of Afternoon Discussion Sessions

1. What are the modelling needs of the Australian National Collections? What other datasets would be critical to have when modelling biological collection data?
2. The challenge of presence only and missing data. What gaps can be filled by modelling and how far can we go without absence data?
3. Hierarchy of data structure, spatial dependence, correlation versus causation. Network linkages among datasets may indicate correlations but how can we model the causes of interactions, and what are the implications?
4. Data mining. What phenomena can we investigate using large amounts of data without a hypothesis?

Additional points to ponder:

- ATLAS of Living Australia - integrate it with more data mining and visualization tools?
- macroecological methods missing
- uncertainty and assumptions
- massiveness of spatial data
- practical concerns of extracting and collating from separate databases

People for Panels (to be confirmed)

Postdocs	***Statisticians***	***Collection scientists***
Luke Barrett (?)	Shuvo Bakar (Grp 3 **moderator**)	Nigel Andrew
Alice Hughes (Grp 2 **moderator**)	Grace Chiu	Justin Borevitz
Andrew Thornhill (?)	Simon Barry	Cang Hui
Holly Vuong (Grp 1 **moderator**)	Mark Handcock	Linda Karssies (Grp 4)
Russell Dinnage (Grp 4 moderator?)	Carolyn Huston (Grp 2)	Jars Lermiin (Grp 4)
	Warren Jin (Grp 4)	Mike McLeish
	Bill Venables	Joe Miller
	Anton Westveld	Alexie Papanicolaou (Grp 4)
	Alec Zwart (Grp 1)	Toni Reverter-Gomez

Local Arrangements

map	http://goo.gl/maps/0Unlq
limited shuttle service	register with Grace Chiu (grace.chiu@csiro.au)
Pavilion Hotel	Point A on map
bike rental (delivers to hotel)	http://www.realfun.com.au/canberra_bike_hire.html
Dickson (restaurants and drinking hangouts)	Point B on map
Braddon (excellent restaurants, pubs, cocktail bar)	Point C on map

Biography
of
Prof. Mark Handcock

Mark S. Handcock is Professor of Statistics at the University of California - Los Angeles. He received his B.Sc. from the University of Western Australia and his Ph.D. from the University of Chicago.

Dr. Handcock's research involves methodological development, and is based largely on motivation from questions in the social sciences, demography and epidemiology.

He has published extensively on network models and inference as well as network sampling methods. This includes the development of methods for analyzing social network data based on recent advances in the statistical modeling of random graphs. The new models, based on the statistical exponential family and more commonly referred to as "exponential random graph models" (ERGM).

Recent work focuses on the development of statistical methodology for the collection and analysis of social network data, combining population-level and individual-level information, spatial processes and longitudinal data arising in labor economics. Recent application has been to social relations networks with the objective of understanding the social determinants of HIV spread. Dr. Handcock has a particular interest in methodology for partially observed network data, including that from respondent driven sampling.

He has also developed models for stream networks that combine information from multiple environmental surveys, models of stochastic curves with application to job instability and age-earnings profiles, and developed graphical tools to allow distributional comparison.

He believes that it is important that advances in statistical methodology be translated into easy-to-use software tools if they are to be accessible to, and assessed by, others. He has collaborated with others to make this so (see, e.g., statnet.org).

He is the Director of the Center for Social Statistics at UCLA. He is a Fellow of the American Statistical Association.

Details of his work are available at his personal website.

Abstracts

The Econometrics of Ecology

Ayesha Ali

In this talk, we discuss how taking an econometric perspective on ecological problems can help us gain a deeper understanding of patterns in nature. For instance, in pollination ecology, researchers are trying to understand whether the plants visited by pollinators are totally random choices, or are the result of linkage rules that facilitate pollination. When pollinators are viewed as customers surveying a set of products (plants) with varying perceived assets, one can borrow meth from econometrics and collaborative filtering to model the interaction probabilities between plant and pollinator species. We will discuss two methods: latent Dirichlet allocation and Dirichlet-multinomial regression, which can be combined into a parsimonious interaction probability model for any mutualistic network and has its roots in random utility theory. The resulting generative model can also be interpreted from a causal modelling perspective.

Insect responses to climate change: integrating insect ecology, physiology and behavioural responses across populations, species and communities

Nigel Andrew

One of the biggest challenges we face today is how to predict how organisms will adapt and respond to a rapidly changing climate. Will individuals respond idiosyncratically or is there a general response across groups of populations, species and communities? How will these responses impact on other organisms within the landscape and the interactions between taxa? Currently there is a major emphasis on research into changes in species distribution, range size, abundance and community structure, but a substantially lower effort going into the biological reasons for these changes – such as physiological, behavioural, genetic and life-history traits. This makes developing predictions of future changes difficult. I will go through some of these issues, and the big ideas that 'bug' me when trying to develop predictive responses of insects to climate change.

When did the dodo become extinct: inferring extinction from sighting records

Simon <u>Barry</u>

Determining when a species is extinct is of interest in a range of contexts. There is the usual scientific curiosity (the Dodo), there is decision making in pest eradication, and there is decision making in species conservation. There are a range of approaches that consider how to make either inference or optimal management decisions from sighting records. Sighting records consist of historical information about the detection of species within a particular region.

This talk reviews previous approaches to this problem, and presents new work which explores how to make inference from this data and the extra information required alongside the sightings in order to make reasonable inference. We explore the adequacy of current models and consider the information content in sighting data and the implications of this for management actions. We present an analysis of data on the incursion of Foxes into Tasmania to illustrate the issues.

Landscape Genomes for Climate Adaptation in Plants

Justin <u>Borevitz</u>

I will present work from Genome Wide Association Studies (GWAS) in *Arabidopsis thaliana*. We show that major quantitative trait loci (QTL) control flowering time and depend on local seasonal conditions. Fall temperature acts as a threshold to convert select overwintering genotypes into rapid cycling under permissive conditions. Across the continental range, flowering time is correlated with population structure, however major QTL show the strongest clinal associations with latitude implicating selection by the environment. At the local landscape scale, genetically diverse populations can be considered as Natural Nested Association Mapping (NNAM) lines where genes for local adaptation can be mapped directly. Finally, simple time lapse imaging was used as high throughput phenotyping to map loci controlling fine growth and development under seasonal conditions.

These tools and methods can be applied in other model plants e.g., Brachypodium, crops, and foundation species that perform critical ecosystem services. To this end, methods of Genotyping By Sequencing (GBS) have been established in switchgrass (*Panicum virgatum*) and are now being applied in Eucalypts and Pelargonia.

Statistical Modelling of Networks: Identifying structure from incomplete data

Mark Handcock

Network models are widely used to represent relational information among interacting units and the implications of these relations. In studies of networks recent emphasis has been placed on random graph models where the nodes usually represent individual social actors and the edges represent a specified relationship between the actors.

In this talk we give an overview of social network analysis from the perspective of a statistician. The networks field is, and has been, broadly multidisciplinary with significant contributions from the social, natural and mathematical sciences. This has lead to a plethora of terminology, and network conceptualizations commensurate with the varied objectives of network analysis. As the primary focus of the social sciences has been the representation of social relations with the objective of understanding social structure, social scientists have been central to this development.

We illustrate these ideas with Exponential-family random graph models (ERGM) which attempt to represent the complex dependencies in networks in a parsimonious, tractable and interpretable way. A major barrier to the application of such models has been lack of understanding of model behavior and a sound statistical theory to evaluate model fit. This problem has at least three aspects: the specification of realistic models; the algorithmic difficulties of the inferential methods; and the assessment of the degree to which the network structure produced by the models matches that of the data.

Building a causal model for the adaptive radiation of Australian trees

Cang Hui

Adaptive radiation in a clad can be explored by correlations and, more appropriately, causal interactions between four elements: range overlaps, niche differentiation, trait similarity and phylogenetic distance. Early research is limited by data availability and mainly focuses on relationships between pairwise elements for sister species. Australia provides an excellent domain for exploring patterns of diversification as the continent has a rich biota including several speciose genera of trees with representatives in many types of ecosystems. Two groups of trees occurring throughout Australia are obvious choices for such a study: Australian acacias (previously grouped in *Acacia* subgenus Phyllodineae) and eucalypts (taxa previously grouped in *Eucalyptus* but now considered to form three genera: *Angophora*, *Eucalyptus* and *Corymbia*), with each including more than 1000 species native to Australia. We here provide a rare example of simultaneously testing multiple interrelated hypotheses for the diversification in Australian acacias by examining correlations and causal relationships between all four elements of multiple related species in a clad. We do so by compiling large dissimilarity/distance matrices from different databases and then building a causal model to unravel the drivers and consequences during the biogeographic diversification. In addition, I am mainly using Mantel and Partial Mantel Tests here.

Data mining: discovering potential useful knowledge from large data sets

Warren Jin

With rapid growing volumes of digital data, data mining has been widely used in areas like biology, medicine, business, engineering, health informatics etc. Data mining is the analysis step of the Knowledge Discovery in Databases (KDD) process. KDD, a field at the intersection of computer science and statistics, attempts to discover potential useful patterns or knowledge from large data sets. It consists of data cleaning, data integration, preprocessing, transformation, data mining, knowledge interpretation and validation steps. The analysis step, data mining, utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as predictive models (regression models, decision tree), groups of data records (cluster analysis), unusual records (outlier detection), dependency (association rule mining), or other understandable structures (like principle components, topics/themes). The talk will present some recent data mining techniques we developed. It will illustrate generating adverse drug reaction hypotheses from large health databases and discovering understandable topics from large documents. Our experience shows that domain experts play a key role in the interpretation and validation step in the whole procedure as well as data mining technique development. Actually, they determine whether pattern/information discovered is useful or not, or requires further investigation. Domain knowledge is often incorporated into data mining techniques too.

Plants, insects, and space: “With a Little Help from My Friends”

Mike McLeish, Carlos Gonzalez-Orozco, Joe Miller

The recent trend to incorporate ecological and evolutionary conceptual frameworks has opened opportunities to combine data in novel ways. New approaches often demand the integration of several fields of expertise commonly outside the reach of individual researchers. Like Ringo Starr whom apparently required more than “a little help” from his friends to hit the last note in the song, integrative approaches can demand expertise from several fields. For instance, if you are not a spacial modelling guru in this field of ever-increasing sophistication, it’s much better to have a colleague that is. Host plants represent a considerable part of the ecological and environmental niche of phytophagous insects. Testing hypotheses of host-plant niche requirements is a useful proxy for the niche requirements of the insect. In this presentation I share some of the results and lessons we have learned by combining new and/or preexisting locality and DNA sequence data in the development of insect-plant interaction hypotheses.

Gene network inference applied to three disparate scenarios: Cow puberty, wool pigmentation and colon cancer

Antonio (Toni) Reverter

The advent of cheaply available high-throughput genetic and genomic techniques has equipped geneticists with an unprecedented ability to generate massive amounts of data. As a result, large lists of genes differentially expressed in many experimental conditions of interests have been reported. Similarly, the association of an ever growing number of DNA variants with phenotypes of importance is now a routine endeavour. Inspired by this wealth of information, systems biology aims at the formal integration of seemingly disparate datasets allowing for a holistic view of the system as a whole and where the key properties emerge in a natural fashion. This talk will present three examples of rigorous ways of integrating molecular data anchored in the power of gene network inference. The first example in concerned with the onset of puberty in cattle raised in tropical regions of Australia (PNAS 2010, 107:13642-13647). The second example deals with piebald, a pigmentation phenotype in Merino sheep (PLoS ONE 2011, 6:e21158). Two networks were developed: a regulatory network and an epistatic one. The former is inferred based on promoter sequence analysis of differentially expressed genes. The epistatic network is built from two-locus models among all pair wise associated polymorphisms. At the intersection between these two networks, we revealed a set of genes and gene-gene interactions of validated and de novo predicted relevance to the piebald phenotype. The final example uses a Boolean approach to exploits non-numerical data in the identification of novel genes associated with colon cancer (BMC Systems Biology 2011, 5:35). Taken together, these approaches render attractive investigating systems biology mechanisms underlying complex phenotypes.

Latent space modelling for trophic food webs and other biological networks

Anton Westveld

A food web is a network of trophic species. The role of each trophic species as predator or prey determines the feeding relations that weave the web. Insight into the nature and degree of dependence among trophic species has ecological implications, and can facilitate discoveries about trophic levels, stability of the ecosystem under species extinction, etc. In a recent paper, we adapted a novel statistical methodology called latent space modelling - originally developed to study social networks - to yield comprehensive understanding of food web structure and its contributing factors. The understanding is achieved through rigorous quantitative inference and effective visualizations due to the unified modeling framework. Specifically, phylogeny is shown to have nontrivial influence on trophic relations in many webs, and for each web trophic clustering based on feeding activity and on feeding preference can differ substantially. In addition to food webs, this talk explores the potential of applying latent space modelling to other types of biological networks.