



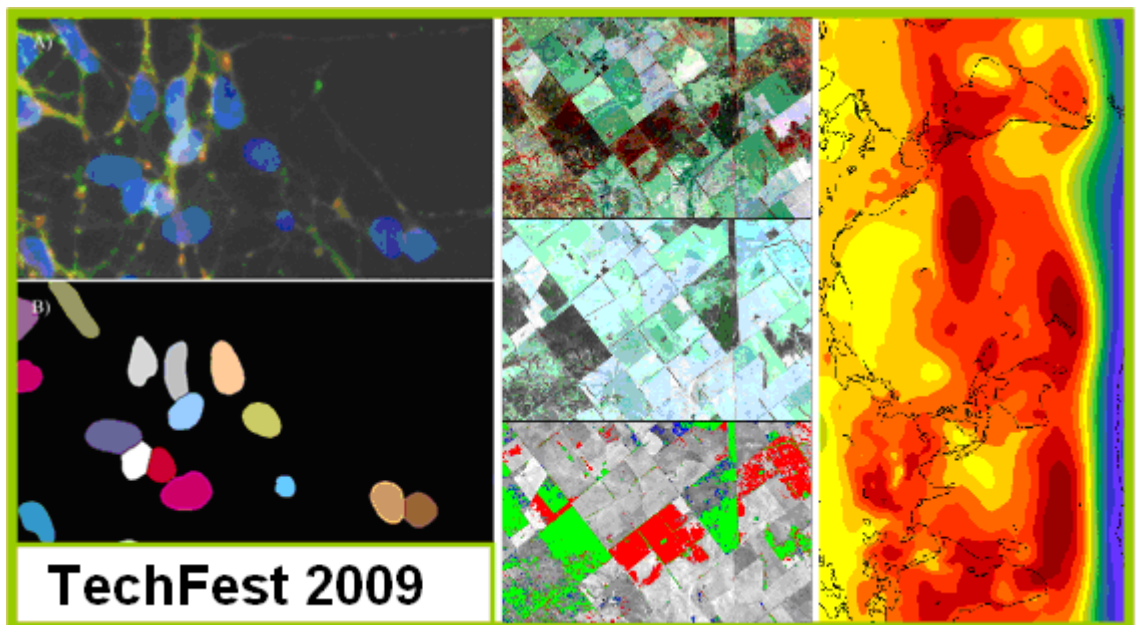
Analysis of High-Dimensional and Large Data

TechFest 09 Book of Abstracts

CSIRO Mathematical and Information Sciences

CMIS Report 09/66

Edited by Frank de Hoog, Emma Huang, Amy Nason and Erin Peterson



Enquiries should be addressed to:

Emma Huang

Emma.Huang@csiro.au

Copyright and Disclaimer

© 2009 CSIRO To the extent permitted by law, all rights are reserved and no part of this publication covered by copyright may be reproduced or copied in any form or by any means except with the written permission of CSIRO.

Important Disclaimer

CSIRO advises that the information contained in this publication comprises general statements based on scientific research. The reader is advised and needs to be aware that such information may be incomplete or unable to be used in any specific situation. No reliance or actions must therefore be made on that information without seeking prior expert professional, scientific and technical advice. To the extent permitted by law, CSIRO (including its employees and consultants) excludes all liability to any person for any consequences, including but not limited to all losses, damages, costs, expenses and any other compensation, arising directly or indirectly from using this publication (in part or in whole) and any information or material contained in it.

Contents

1. Preface	4
2. BIOLOGICAL APPLICATIONS Techfest Day One: 11 June 2009.....	5
3. ENVIRONMENTAL APPLICATIONS Techfest Day Two: 12 June 2009.....	7
4. Day 1 Abstracts.....	9
0910-0935: Embedded p-rep designs for use in QTL analysis.....	9
0935-1000: Unlocking higher order epistatic interactions in wheat quality traits.....	10
1000-1025: Variable penalty dynamic time warping for aligning GC-MS data.....	11
1100-1125: Data challenges in preventative health	12
1125-1150: A computationally efficient spatial modelling of disease count data	13
1150-1215: High-dimensional data analysis in bioinformatics	13
1215-1240: Automated classification of high throughput subcellular imaging and data visualisation	14
1350-1415: Quality control and analysis of 454 sequencing data.....	15
1415-1440: Assembly of the sheep genome and identification of SNPs using 454 and Illumina sequencing and design of the Illumina Ovine SNP50 BeadChip	16
1440-1505: Development of bioinformatics approaches to extract information from metagenomic data produced by next generation sequencing technologies.....	17
1505-1530: Climbing the k-mer Mountains	18
1610-1635: Studying transcriptomes at single nucleotide resolutions	19
1635-1700: Asking the right question of microarray data: the quest for causal mutations	20
1700-1725: A computational framework to promoter sequence analysis: Application to the regulation of differentially expressed genes in parasite infested sheep	21
5. Poster Abstracts.....	23
Combining Bayesian analyses of genetic linkage using SNP chip data.....	23
Managing CSIRO's Scientific Data.....	23
CSIRO IM&T Advanced Scientific Computing	24
From the ground up: Linkage map construction in integrated crosses	24
Regional water and energy flux estimation using geostationary satellite thermal data	25
Genome annotation and distributed annotation system.....	26
Visualisation tools for unsequenced genomes.....	27
After Dinner Speaker: Dr David Lovell	28
6. Day 2 Abstracts	29
0835-0905: Enterprise data management services	29
0905-0935: Katmandoo, a biosciences database.....	30
0935-1005: The flood of data: Large-scale sensor networks for water storage monitoring	31
1005-1035: The storage, transfer, analysis and archive of underwater video	32

1055-1125: Organising and searching high dimensional data sets	32
1130-1200: Terrestrial ecosystem data with high spatial and temporal variability	33
1200-1230: Challenges in analysing Seqwater intensive data	33
1230-1300: A statistical downscaling model for southern Australia winter rainfall.....	34
1345-1415: Operational large-area land-cover monitoring using medium spatial resolution satellite imagery	35
1415-1445: Spectroscopy and hyperspectral imaging for mineral and environmental applications	37
1445-1515: Fast Bayesian analysis of spatial dynamic factor models for large space time data sets	38
List of Participants.....	40

List of Tables

Table 1: TechFest Day 1.....	5
Table 2: TechFest Day 2.....	7

1. PREFACE

TechFests are series of workshops organised by CSIRO Mathematical and Information Sciences (CMIS), each of which targets an area of the mathematical sciences. The objectives of these workshops are:

- to share research achievements with CSIRO colleagues;
- to identify synergies and possible collaborations in CSIRO;
- to share emerging trends and developments in science.

The theme for TechFest '09 is Analysis of High-Dimensional and Large Data. We extend our warmest welcome to the Riverglenn Conference Centre, Indooroopilly for this conference.

TechFest itself is a two day conference. The first day of the conference is on Biological Applications and the second day covers Environmental Applications. A dinner and poster session will be held on the evening of the 11th of June to give participants the opportunity to network in an informal setting. This year marks the first that external participants and speakers have been invited, with representatives from the University of Queensland, Queensland University of Technology, the Queensland Institute of Medical Research, and the NSW Department of Primary Industries. Overall approximately 70 people will attend, with about half of these from CMIS, a quarter from other CSIRO divisions and the rest external.

This book includes abstracts of all the talks and posters from the two days of the TechFest. Contact details of all the participants of the TechFest are also included.

We would like to thank the current and previous members of the Science Development Group and Frank De Hoog within CMIS for their assistance in the organisation of the TechFest. We would also like to thank Amy Nason and Daphne Bruce for maintaining the TechFest website and registration list and Dr David Lovell for his address at the TechFest dinner.

Emma Huang and Erin Peterson

June 2009

2. BIOLOGICAL APPLICATIONS TECHFEST DAY ONE: 11 JUNE 2009

Daintree Room, Riverglenn Conference Centre

Table 1: TechFest Day 1

Time	Presentation	Presenter
0815-0845	Registration (Tea/Coffee on arrival)	
0845-0900	Welcome & Introductions	Frank De Hoog CMIS
0900-1030	Session 1: Agricultural Applications Chair: Ari Verbyla, CMIS	
0900-0910	Introduction	Chair
0910-0935	Embedded p-rep designs for use in QTL analysis	Brian Cullis NSW DPI
0935-1000	Unlocking higher order epistatic interactions in wheat quality traits	Julian Taylor CMIS
1000-1025	Variable penalty dynamic time warping for aligning GC-MS data	David Clifford CMIS
1030-1100	Morning Tea	
1100-1240	Session 2: Health Applications Chair: Ian Saunders, CMIS	
1100-1125	Data challenges in preventative health	Richard Head P-Health Flagship
1125-1150	A computationally efficient spatial modelling of disease count data	Louise Ryan CMIS
1150-1215	High-dimensional data analysis in bioinformatics	Harri Kiiveri CMIS
1215-1240	Automated classification of high throughput subcellular imaging and data visualisation	Nick Hamilton UQ
1240-1340	Lunch	

1340-1530	Session 3: Sequence Analysis Chair: Mark Morrison, CLI	
1340-1350	Introduction	Chair
1350-1415	Quality control and analysis of 454 sequencing data	Jen Taylor CPI
1415-1440	Assembly of the sheep genome and identification of SNPs using 454 and Illumina sequencing and design of the Illumina Ovine SNP50 BeadChip	Brian Dalrymple CLI
1440-1505	Development of bioinformatics approaches to extract information from metagenomic data produced by next generation sequencing technologies	Lauren Bragg CMIS
1505-1530	Climbing the k-mer mountains	Paul Greenfield CMIS
1530-1600	Afternoon Tea	
1600-1730	Session 4: Expression Analysis Chair: Brian Dalrymple, CLI	
1600-1610	Introduction	Chair
1610-1635	Studying transcriptomes at single nucleotide resolutions	Sean Grimmond IMB
1635-1700	Asking the right question of microarray data: the quest for causal mutations	Nick Hudson CLI
1700-1725	A computational framework to promoter sequence analysis: Application to the regulation of differentially expressed genes in parasite infested sheep	Shivashankar Hiriyyur-Nagaraj CLI
1800-1900	Poster Session & Pre Dinner Drinks	Mossman Room
1900-Late	Dinner Speaker: Dr David Lovell	Restaurant

3. ENVIRONMENTAL APPLICATIONS TECHFEST DAY TWO: 12 JUNE 2009

Daintree Room, Riverglenn Conference Centre

Table 2: TechFest Day 2

Time	Talk	Presenter
0815-0830	Registration (Tea/Coffee on arrival)	
0830-1125	Session 1: Large data storage and transfer Chair: Alf Uhlherr, CSIRO IM&T	
0830-0835	Introduction	Chair
0835-0905	Enterprise data management services	John Morrissey CSIRO IM&T
0905-0935	Katmandoo, a biosciences database	David Butler QDPI&F
0935-1005	The flood of data: Large-scale sensor networks for water storage monitoring	Matt Dunbabin ICT Centre
1005-1035	Analysis of the Moreton Bay BRUV data	Ian McLeod CMAR
1035-1055	Morning Tea	
1055-1125	Organising and searching high dimensional datasets	Dave Abel CMIS
1125-1300	Session 2: Environmental Applications Chair: Keith Hayes, CMIS	
1125-1130	Introduction	Chair
1130-1200	Terrestrial ecosystem data with high spatial and temporal variability	Peter Grace QUT
1200-1230	Challenges in analysing Seqwater intensive data	Sarah Lennox CMIS
1230-1300	A statistical downscaling model for southern Australia winter rainfall	Yun Li CMIS
1300-1340	Lunch	

Time	Talk	Presenter
1340-1515	Session 3: Remote Sensing Chair: Mark Berman, CMIS	
1340-1345	Introduction	Chair
1345-1415	Operational large-area land-cover monitoring using medium spatial resolution satellite imagery	Stuart Phinn, Tony Gill UQ Robert Denham DERM
1415-1445	The analysis of mineral and environmental spectroscopic and hyperspectral data	Mark Berman CMIS
1445-1515	Fast Bayesian analysis of spatial dynamic factor models for large space time data sets (Remote Sensing)	Chris Strickland QUT
1515-1530	Wrap-Up	Louise Ryan CMIS

4. DAY 1 ABSTRACTS

0910-0935: Embedded p-rep designs for use in QTL analysis

Brian Cullis

NSW Department of Primary Industries, Wagga Wagga Agricultural Institute, Pine Gully Road,
Wagga Wagga NSW 2650

Phone: +61 2 6938 1855

email: Brian.Cullis@dpi.nsw.gov.au

Brian Cullis is based at the Wagga Wagga Agricultural Institute, where he heads the Biometrics group and is an adjunct Professor at the University of Sydney. He is internationally recognised as an expert in the analysis of field experiments using spatial techniques. His other main research interest is the analysis of series of crop variety trials, commonly known as multi-environment trials (METs). Methods developed by him are now the basis of the protocols which are in widespread use for the analysis of plant improvement data in Australia.

Abstract

The literature on the design and analysis of cereal variety trials has focussed on the trait of grain yield. Such trials are also used to obtain information on grain quality traits but these are rarely subjected to the same level of statistical rigour. The data are often obtained using composite rather than individual replicate samples. This precludes the use of an efficient statistical analysis. In this paper we propose an approach in which a proportion of varieties is grain quality tested using individual replicate samples. This is achieved by embedding a partially replicated design (for measuring quality traits) within a fully replicated design (for measuring yield). This allows application of efficient mixed model analyses for both grain yield and grain quality traits.

0935-1000: Unlocking higher order epistatic interactions in wheat quality traits

Julian Taylor and Ari Verbyla

CSIRO Mathematical and Information Sciences, Private Bag No 2, Glen Osmond SA 5064

Phone: +61 8 8303 8792

email: Julian.Taylor@csiro.au

Julian Taylor is a postdoctoral fellow with CSIRO Mathematical and Information Sciences in Adelaide. He completed his Ph.D. in Biometrics in 2005 under the supervision of Dr. Ari Verbyla. He currently works on several agricultural genetics projects, including high-dimensional data analysis with the incorporation of mixed model components, computational whole-genome analysis, and large-scale gene x environment analysis of indicator wheat crops.

Abstract

There has been a recent focus on variable selection methods in the biosciences to help to understand the influence of underlying genetics on traits of interest. In the plant breeding context, this involves the analysis of Quantitative Trait Loci (QTLs) from traits measured in complex designed experiments. Due to the nature of these experiments, extra components of variation, such as spatial trends and extraneous environmental variation, needs to be accounted for and can be achieved using linear mixed models. However, with these models the inclusion of an additional high dimensional genetic component becomes problematic. This talk discusses the incorporation of high dimensional genetic variable selection in a mixed model framework. The methodology shows that this incorporation can be achieved in a natural way even when the number of genetic variables exceeds the number of observations. This method is then applied to wheat quality traits and a well established genetic wheat map of 411 markers obtained from the Future Grains group in the Food Futures Flagship in CSIRO. This example focusses on the simultaneous incorporation and selection from 75,000 genetic variables (main QTL effects and epistatic interactions) for some wheat quality traits of interest. The results show, possibly for the first time, that QTL epistatic interactions are obtainable for traits measured in a highly complex designed experiment.

1000-1025: Variable penalty dynamic time warping for aligning GC-MS data

David Clifford

CSIRO Mathematical and Information Sciences, Locked Bag 17, North Ryde, NSW 1670

Phone: +61 2 9325 3210

email: David.Clifford@csiro.au

David Clifford has been a research scientist with CSIRO Mathematical and Information Sciences in Sydney since 2006, after joining the division as a postdoctoral fellow two years previously. He has worked in a range of areas including spatial statistics, bioinformatics, precision agriculture and health risk models. One of his recent projects has focused on developing methodology for mass spectrometry data.

Abstract

Gas Chromatography Mass Spectrometry (GC-MS) is a technology used in environmental monitoring, criminal forensics, security, food/beverage/perfume analysis, astrochemistry and medicine. It is considered the “gold standard” for detecting and identifying trace quantities of compounds in test substances. The data collected by GC-MS is high dimensional and large as the technology divides the substance into and quantifies the amount of each compound that make up the test substance. Typically the first step involved in an analysis of data like this is the alignment of the data to correct for the often subtle drift of the gas chromatography part of the system that can occur over time. Once properly aligned, these high-dimensional data is used to find compounds that distinguish between test substances – e.g. different kinds of meat, wine of different quality, blood serum from healthy/non-healthy individuals etc.

In this talk I highlight a novel variation on dynamic time warping (DTW) for aligning chromatogram signals. We are interested in sets of signals that can be aligned well locally, but not globally, by shifting individual signals in time. This kind of alignment is often sufficient for aligning gas chromatography data. Regular DTW often “over-warps” signals and introduces artificial features into the aligned data. To overcome this one can introduce a variable penalty into the DTW process. The penalty is added to the distance metric whenever a nondiagonal step is taken. I will discuss penalty selection and showcase the method using three examples from agricultural and health research. The use of variable penalty DTW significantly reduces the number of nondiagonal moves. In the examples presented here, this reduction is by a factor of 30, with no cost to visual quality of the alignment.

1100-1125: Data challenges in preventative health

Richard Head

CSIRO Preventative Health, Level 3, Elizabeth House, 231 North Terrace, Adelaide SA 5000

Phone: +61 61 8 8303 8819

email: Richard.Head@csiro.au

Richard Head is the Director of the CSIRO Preventative Health Flagship, and has had 40 years of research experience nationally and internationally in investigating the aetiology of chronic diseases that affect societies. He oversees development of approaches to early detection and intervention in colorectal cancer, Alzheimer's disease and obesity. He also provides leadership to a large multidisciplinary team with skills in genomics and proteomics, biotechnology, bioinformatics and statistical data analysis, structural biology, information technology, psychology and food and nutritional sciences.

Abstract

Biology including biomedical research is by nature a descriptive science. It achieves great vitality when it is integrated with mathematics. The P-Health Flagship has by design brought together differing disciplines to tackle major national challenges. In doing so, the integration of mathematics with all of the disciplines required to achieve input is critical. The specific challenges of relevance to this Conference are:

1. How does one integrate and range across data derived from image (PET/MRI) with gene expression?
2. How does one integrate and range across expression data based upon case controls in the human and treatment in animal and cell equivalents?
3. How does one approach the modelling of the roles of adduct repair, DNA strand breakage, apoptosis and proliferation regulation?

While not having solutions to the above, the presentation is designed to highlight these challenges and encourage collaborative discussion.

1125-1150: A computationally efficient spatial modelling of disease count data

Louise Ryan

CSIRO Mathematical and Information Sciences, Locked Bag 17, North Ryde, NSW 1670

Phone: +61 2 9325 3203

email: Louise.Ryan@csiro.au

Louise Ryan took up her post as Chief of CSIRO Mathematical and Information Sciences in 2009 after serving as Chair of the Biostatistics Department at Harvard University. She is well known for her contributions to statistical methods for cancer and environmental health research. She has a distinguished career in biostatistics, having authored or co-authored over 200 papers in peer-reviewed journals.

Abstract

We illustrate the usefulness of a simple type of Gauss-Seidel algorithm for fitting spatially correlated generalized linear models. The methods are illustrated using data provided by NSW health and involving 33 million observations. Our method is able to handle this large dataset very easily, whereas standard methods fail. The method allows us to identify very interesting interactions of gender and age on the impact of socio-economic status on the incidence of ischemic heart disease.

1150-1215: High-dimensional data analysis in bioinformatics

Harri Kiiveri

CSIRO Mathematical and Information Sciences, 65 Brockway Rd, Floreat WA 6014

Phone: +61 8 9333 6315

email: Harri.Kiiveri@csiroau

Harri Kiiveri is a Research Scientist with CSIRO Mathematical and Information Sciences in Perth. He currently works in the area of bioinformatics, with a particular emphasis on the analysis of microarray and SNP data. He has developed methodology for fitting statistical models to data with many more variables than observations, and for the construction of local and global gene networks.

Abstract

In this talk I'll present some ideas for analysing data where the number of variables is several orders of magnitude greater than the number of samples. The ideas will be illustrated with microarray and SNP (Single Nucleotide Polymorphism) chip datasets from the human health area. I'll touch on response modelling useful in finding molecular diagnostics and prognostics for disease as well as the construction and simulation of local gene networks to try and identify influential genes in a disease process. The examples will span the range of sample sizes of the order of a hundred and variable numbers from tens of thousands to millions.

1215-1240: Automated classification of high throughput subcellular imaging and data visualisation

Nick Hamilton

Institute for Molecular Bioscience, University of Queensland, St Lucia QLD 4072

Phone: +61 7 3346 2033

email: n.hamilton@imb.uq.edu.au

Nick Hamilton is a researcher with the Institute for Molecular Biology at the University of Queensland. His group develops the methodologies, algorithms and tools to maximise the benefit of the newly available data sources in biological imaging. He collaborates closely with cell biology, bioinformatics and mathematics groups in creating these methodologies and utilises and develops techniques in areas such as machine learning, data clustering, graph algorithms, image segmentation, statistical testing and mathematical modelling.



Abstract

Obtaining the sequence of numerous genomes and subsequent identification of the encoded proteome has created the need for large-scale systematic approaches to understand the functions of the tens of thousands of proteins at the cellular level. The desire and the ability to carry out high-throughput screenings of protein localization and trafficking for applications such as drug discovery is leading to a rapid growth in cell images in need of analysis on a scale comparable to that of the genomic revolution. However, the bottleneck in the pipeline is the need for a researcher to examine the wealth of data that modern automated microscopes are capturing.

Towards this, my group has been developing the Automated Subcellular Phenotype Classification (ASPiC) system. Within ASPiC novel image statistics have been developed that in combination with machine learning techniques have are able to distinguish and classify microscopy images of 10 subcellular localisations with up to 98.2% accuracy. Other applications include similarity ranking, clustering, statistical differentiation under varying experimental conditions, and representative image selection.

While high-throughput and automated image classification have many benefits, they suffer by removing the best pattern recognition system, the human eye, from the viewing the images. To deal with this I have developed the iCluster subcellular image visualisation and clustering system. Within iCluster, large image sets (up to 1400 images simultaneously) are automatically arranged in 2 or 3 dimensions in such a way that similar images are spatially close. In this way large image sets may be quickly viewed and subtle patterns in the image data discerned and conclusions drawn. While designed initially designed for image visualisation, iCluster may be used to visualise the relationships in (high dimensional) vector based data sets, or datasets on which a distance relation is defined between objects. It is open source and available to download from <http://icluster.imb.uq.edu.au>.

1350-1415: Quality control and analysis of 454 sequencing data

Jen Taylor and Andrew Spriggs

CSIRO Plant Industry Bioinformatics Team, GPO Box 1600, Canberra ACT 2601

Phone: +61 2 6246 4929

email: Jen.Taylor@csiro.au

Jen Taylor is the Bioinformatics Team Leader in CSIRO Plant Industry in Canberra. She joined CSIRO in November 2008 from a position as the Head of Functional Analysis at the Wellcome Trust Centre for Human Genetics in Oxford, and is an Adjunct Fellow in the Department of Mathematics at the Australian National University. Her team researches computational and analytical methods to make the most of modern, high-throughput genomics analysis.

Abstract

Roche 454 next-generation sequencing facilitates the generation of high-throughput nucleic sequences of biological sample. Unlike other next-generation sequencing approaches, 454 technology utilises pyrosequencing chemistry to produce millions of relatively long sequences reads averaging 360 bases in length. The 454 sequencing approach is being quickly exploited in the sequencing of complex mixtures of RNA transcripts where the added read length better informs unambiguous identification and classification of sequences to DNA template and/or species. The work presented investigated the properties of 454 sequencing data with regard to error rates, sequencing bias and technical and biological variation.

1415-1440: Assembly of the sheep genome and identification of SNPs using 454 and Illumina sequencing and design of the Illumina Ovine SNP50 BeadChip

Brian Dalrymple on behalf of the International Sheep Genomics Consortium
CSIRO Livestock Industries, 306 Carmody Rd, St Lucia QLD 4067
Phone: +61 7 3214 2503 email: Brian.Dalrymple@csiro.au

Brian Dalrymple is a Senior Principal Research Scientist with CSIRO Livestock Industries in Brisbane. He heads bioinformatics research with particular interests in the assembly of genome sequences, primarily for cattle and sheep; the development of genomics and SNP resources; non-coding RNA and the comparative genomics of mammals. He is a participant in the cattle and sheep genome international sequencing consortia.

Abstract

A whistle stop tour through the assembly of the first mammalian genome to be put together in the southern hemisphere. What were the big challenges handling the large datasets and the

complex processes required to generate the assembly? I will describe the use of different sequencing methodologies and comparative genomics to generate a draft 3 X coverage of the sheep genome with potentially a more accurate order of contigs than the bovine genome which underpinned the assembly. To complement the sheep SNPs discovered from the 454 skim sequencing of the sheep genome deeper, reduced representational sequencing (RRS) was also undertaken using the Illumina Genome Analyzer. Three size ranges of HaeIII digested sheep DNA (pooled from sixty diverse animals) were sequenced to an average depth of ~20 fold. After extensive filtering of the Illumina sequencing reads and mapping to the sheep genome assembly 76,000 high quality SNPs were identified. The 1536 pilot chip SNPs (identified using Sanger resequencing) along with the 454 and RRS-derived SNPs were then remapped to the sheep genome assembly. All SNPs were scored for probability of assay conversion by Illumina and minor allele frequency was calculated for the 1536 and Illumina-derived SNPs. SNPs with an assay conversion score of less than 0.8, and/or a minor allele frequency of less than 0.2 were excluded. SNPs for the BeadChip were then selected from adjacent windows of the assembled genome sequence with as close to equal spacing as possible. A hierarchy of selection criteria was applied in each window, choosing Sanger before Illumina before 454, Infinium II (requires one bead on the BeadChip) over Infinium I (requires two beads on the BeadChip), highest assay conversion score and minor allele frequency closest to 0.5. In a final quality control SNPs with identical oligonucleotide sequences and SNPs with oligonucleotides with multiple hits to the sheep genome assembly were removed and replaced with the closest adjacent SNP that met the selection criteria. The final BeadChip carries 59,494 sheep SNPs. The chip has been in use for almost six months, how well did we do with our predictions and selections now that we have the real answers?

1440-1505: Development of bioinformatics approaches to extract information from metagenomic data produced by next generation sequencing technologies

Lauren Bragg, S. E. Denman, P. Evans, E. Ling, G. Stone, A-D. Wright, C.S. McSweeney, D. Edwards and M. Morrison

CSIRO Mathematical and Information Sciences, University of Queensland, 306 Carmody Rd, St Lucia QLD 4067

Phone: +61 7 3214 2954

email: Lauren.Bragg@csiro.au

Lauren Bragg is a bioinformatician with CSIRO Mathematical and Information Sciences who is currently pursuing a Ph.D. at the University of Queensland. She uses her computer programming, statistical and biological knowledge to develop analytical pipelines for data from a variety of biological sources. She has been involved in research to develop novel sequence space microarray platforms as well as current metagenomic projects.

Abstract

In the last 5 years, microbial genomics has progressed from the sequencing of single microbial genomes, to the production of metagenomic data from “real-world” samples, including those microbes not yet cultured in the laboratory. As the read length and accuracy of the “next generation” sequencing methods improve, so too does the quantity of data produced. However, these methods also impose new and complex bioinformatics challenges, most notably the correct assembly of raw metagenomic data into contiguous sequence that will support interpretative analysis of community structure-function relationships. As our first step in addressing these challenges we have embarked on the use of 454 GS-FLX and Solexa sequencing technologies to produce metagenomic data, and we are developing a suite of assembly and analysis approaches with the goal of producing accurately assembled genomic data for specific microbes. The microbes are a part of a “simplified” mixture of both well-known and “new” microbes, produced by selective enrichment from the microbiomes resident in the foregut compartment of the Tammar wallaby (*Macropus eugenii*) and the bovine (*Bos taurus*). The numerically predominant member of each enrichment is phylogenetically assigned to a novel group of archaeobacteria that has not yet been cultured axentially or characterized. We consider these enrichments will support a useful benchmark to assess and overcome the challenges that would be faced with more complex metagenomic datasets and I will explore some of the diagnostic and analytical methods required for this new and complex data.

1505-1530: Climbing the k-mer Mountains

Paul Greenfield

CSIRO Mathematical and Information Sciences, Locked Bag 17, North Ryde, NSW 1670

Phone: +61 2 9325 3250

email: Paul.Greenfield@csiro.au

Paul Greenfield is a research projects officer with CSIRO Mathematical and Information Sciences in Sydney. He has had positions as a Group Leader in CMIS and the ICT centre before his background in computer science led to his involvement in problems at the

intersection of computing and biology. He currently works on the Transformational Biology project of non-assembly methods for analysing and interpreting short-read sequence data.

Abstract

Genomic data, both sequence data and reference genomes, can be viewed as (very large) collections of overlapping k-mers that can be used to characterise and compare organisms. Most of the work discussed in this session is concerned with 25-mers. The 25-mer space is large (there are around 10^{15} possible 25-mers) and the set of 25-mers derived from a single organism's genome has few repeats, with few overlaps with the 25-mers derived from other organisms. One application of these properties was to develop a 10,000m view of the bacteria, using shared 25-mers as a similarity metric, by cross-comparing all 25-mers from all (728) available bacteria sequences. The results of this comparison both confirm the established taxonomy and highlight possible anomalies, and validate the use of 25-mers as identifying tags for individual bacterial species/families. The same analysis generates gene-level mappings between organisms that highlight conserved genes, the differences between closely related organisms and possible gene homologies. These types of questions are largely answered through the use of single SQL queries over a database that holds information about the gene-level sharing of 25-mers.

The uniqueness of short k-mers has also been used to 'heal' bacterial sequence data – recovering accurate sequencing data from reads with sequencing errors. Deriving overlapping 25-mers from short reads effectively oversamples the read data and 'good' 25-mers accumulate along a Poisson distribution – given enough depth of sequencing. 25-mer tiles containing errors are often singletons, and generating and matching variants of these possibly erroneous tiles can undo some sequencing errors and clean up sequence data without needing a reference genome.

1610-1635: Studying transcriptomes at single nucleotide resolutions

Sean Grimmond

Expression Genomics Laboratory, Institute for Molecular Bioscience, University of Queensland, St Lucia QLD 4072

Phone: +61 7 3346 2057

email: s.grimmond@imb.uq.edu.au

Sean Grimmond is an Associate Professor with the Institute for Molecular Bioscience at the University of Queensland. He is an NHMRC Senior Research Fellow who established the Expression Genomics Laboratory in 2001. The group focuses on globally surveying genomic, transcriptomic, and epigenomic information and using the data to define the underlying molecular networks controlling key biological processes and pathological states.

Abstract

While array-based transcriptomics has revolutionized our ability to globally survey gene activity and define transcriptional programs controlling in biological processes and pathological states, it is clear these approaches fail to capture the true complexity of mammalian gene expression. We are using multi-gigabase scale transcriptome sequencing to completely define the transcriptional output of mammalian systems. The development of pipelines for monitoring and quantifying the complexity of transcriptional output for each locus, exon splice junction and promoter usage will all be presented. In addition to surveying the known loci, strategies for studying of novel gene expression will be discussed.

1635-1700: Asking the right question of microarray data: the quest for causal mutations

Nicholas Hudson, Antonio Reverter and Brian Dalrymple

CSIRO Food Futures Flagship and Livestock Industries, 306 Carmody Rd, St Lucia QLD 4067

Phone: +61 7 3214 2212

email: Nick.Hudson@csiro.au

Nick Hudson is a muscle systems biologist with CSIRO Livestock Industries in Brisbane. He is involved in a Food Future Flagship project to use a systems approach to understand gene expression in cattle muscle by analysing bovine muscle microarray data. His work aims to determine the genetic regulation of such processes as muscle growth, fibre composition and mitochondrial biogenesis in order to allow livestock producers to tailor muscling, feed efficiency and meat quality in their animals.

Abstract

The assembly of biological tissues is governed by regulatory circuits where the nodes are transcriptional regulators. Phenotypic variation between species, breeds and individuals is a product of differences in regulatory circuit wiring. Consequently, there is great interest in system-wide high-throughput methods that can identify the causal mutation / perturbation

responsible for the wiring differences. Here, based solely on microarray expression data from just 27 animals, we show that we can identify myostatin as responsible for the network perturbation underpinning the difference in muscle structure between highly muscular Piedmontese cattle and less muscular Wagyu. This specific result is noteworthy as myostatin is not actually differentially expressed (DE) in Piedmontese, but exerts its effect through encoding a dysfunctional protein. Such independence of activity and expression level is not uncommon for perturbations influencing regulatory genes. While these remain completely invisible to conventional DE analysis we show they can be identified through contrasting the topology of co-expression networks. By identifying the question to which myostatin is the answer we present a systems-wide comparison of network connectivity that is potentially transferable to a diverse range of phenotypes.

1700-1725: A computational framework to promoter sequence analysis: Application to the regulation of differentially expressed genes in parasite infested sheep

Shivashankar Hiriyur-Nagaraj, Aaron Ingham and Antonio Reverter

CSIRO Food Futures Flagship and Livestock Industries, 306 Carmody Rd, St Lucia QLD 4067

Phone: +61 7 3214 2524

email: Shivashankar.Hiriyur-Nagaraj@csiro.au

Shiv Hiriyur-Nagaraj is a postdoctoral research fellow in the Systems Biology Group with CSIRO Livestock Industries in Brisbane. He received his Ph.D. in bioinformatics from Macquarie University in 2008, and has research interests in systems biology, transcriptomics, infectious disease and host-parasite interactions.

Abstract

Regulation of transcription is a pivotal mechanism for determining whether or not a gene is expressed and how much mRNA and protein is produced. Regulatory sequences along with transcription factors tightly control this process in every cell and tissue in eukaryotes. We have designed a comprehensive computational schema to analyse regulatory sequences in mammals that includes statistically over-represented sequence motifs, promoter models and inter-species comparisons. Using this schema, and a proof of concept, we have discovered several novel regulatory motifs in genes differentially expressed in sheep, following a nematode parasite challenge. Finally, we have experimentally validated our in-silico predicted findings using

mobility shift assays and mass-spectrometry. These studies are crucial, given the current, serious resistance problems in parasites against most treatments, and residue problems in meat, milk and the environment.

5. POSTER ABSTRACTS

Mossman Room, Riverglenn Conference Centre

Time: 1800-1900

Combining Bayesian analyses of genetic linkage using SNP chip data

Ian Saunders

CSIRO Mathematical and Information Sciences, Private Bag No 2, Glen Osmond SA 5064

Phone: +61 8 8303 8788

email: Ian.Saunders@csiro.au

Abstract

Single Nucleotide Polymorphisms or "SNPs" are genetic markers that occur at very high densities in the genomes of many organisms. "SNP Chips" are a cheap method of measuring large number of SNP markers simultaneously. A range of different chips are now available, with different marker densities and locations. Combining the results of genetic linkage studies using different chips and different experimental designs is a challenge. A Bayesian approach provides a natural way to integrate the results from multiple studies. Some results from a P-Health study of colon cancer genetics will be presented as an example of how this can be done.

Managing CSIRO's Scientific Data

Cynthia Love, Gerry Ryder and John Morrissey

CSIRO Information Management and Technology, Private Bag 10, Clayton VIC 3169

Phone: +61 3 9545 8176

email: Cynthia.Love@csiro.au

Abstract

Data Management is a growing issue for research in CSIRO with an increasing volume of data being generated daily. This poster outlines IMT's strategic and practical approach to addressing the issue. It covers an organisational approach within a collaborative context.

CSIRO IM&T Advanced Scientific Computing

Justin Baker

CSIRO Information Management and Technology, 700 Collins Street, Docklands VIC 3008

Phone: +61 3 8601 3801

email: justin.baker@csiro.au

Abstract

CSIRO IM&T Advanced Scientific Computing provides facilities and services needed to support computational science and eResearch. These include high-performance computing systems, large-scale data storage, scientific software and access to a team of ASC specialists. ASC shared facilities and support are freely and immediately available to all CSIRO researchers. Resources are continually upgraded to provide significant computational capability, notably through our partnerships under the Federal Government's Platforms for Collaboration (PfC) program.

From the ground up: Linkage map construction in integrated crosses

B. Emma Huang, Andrew W. George, Colin Cavanagh, Matthew Morell

CSIRO Mathematical and Information Sciences, St Lucia Bioscience Precinct, 306 Carmody Road, St. Lucia QLD 4067

Phone: +61 7 3214 2953

email: Emma.Huang@csiro.au

Abstract

The integrated cross is an exciting new experimental design which enables genomic regions housing genes of commercial significance to be detected with far greater precision than previously possible. By using four or eight parents rather than a traditional biparental design and breeding generations through to fixation, these crosses represent an abundance of genetic diversity with the potential for high mapping resolution. CSIRO is currently conducting the world's first integrated cross in wheat. This promises to be instrumental in unlocking the genetic secrets of agronomic, disease and quality traits of one of the world's most important domesticated crops.

While there are similarities between this project and the Collaborative Cross in mice, a key difference is in the lack of physical maps or genome sequence for wheat. Thus a crucial element for the success of the project is the production of highly accurate DNA marker maps. We have developed statistical methods and computational tools to address this challenge. We use two- and three-point haplotype probabilities to group and order loci within linkage groups. These algorithms and software have been tested through extensive simulations and will be applied to 4-way and 8-way cross data as it becomes available.

Regional water and energy flux estimation using geostationary satellite thermal data

Luigi Renzullo

CSIRO Land and Water, Black Mountain, Canberra ACT 2601

Phone: +61 2 6246 5758

email: Luigi.Renzullo@csiro.au

Abstract

The ability to accurately map the flux and stores of water in the landscape is an essential part of water resources assessment, with implications for water management and the forecasting of future availability. Land surface models (LSMs) can provide estimates of soil moisture and evaporative fluxes for large parts of the continent, but model constraints are limited to data from a small number of scattered in-situ monitoring sites, and therefore have high uncertainty in those ungauged areas of the country.

Another source of LSM constraint is provided by indirect measurements from remote sensing systems. Satellite-based remote sensing data have the advantage of continental coverage and sub-daily observation frequency. A difficulty however is that, with the exception of some inferred image products, rarely is the LSM output of interest matched exactly by an equivalent remotely-sensed observation. Therefore, if these data are to be of use, it is necessary to either develop retrieval schemes (i.e. inverse models) that derive as accurately as possible land surface variables from the satellite observations that correspond to model output, or to develop appropriate observational models that relate the LSM states/variables to the remotely-sensed observations.

This work focuses on the role of geostationary MTSAT-1R satellite's thermal observations as observational constraints on surface water and energy flux estimation. The retrieval of land surface temperature (LST) estimates from the MTSAT-1R thermal brightness temperature observations is described. Results of assimilating the derived LST images in a simple coupled water-energy balance model are presented for a number of study sites around Australia. Discussion is also given on the computational challenges of working with high frequency (hourly), moderate resolution (~ 4 km x 4 km) geostationary data, particularly when seeking to extend analysis to the entire continent.

Genome annotation and distributed annotation system

Andrew Spriggs

CSIRO Plant Industry Bioinformatics Team, GPO Box 1600, Canberra ACT 2601

Phone: +61 2 6246 5193

email: Andrew.Spriggs@csiro.au

Abstract

It has always been desirable in genomics research to integrate generated data with public data resources for the purposes of hypothesis generation, accumulation of support of biological interpretations and data dissemination. In particular, it is vital that high-throughput sequencing data can be visualised and browsed in a genome-wide context with maximal annotation. DAS (Distributed Annotation System) is an answer to this. Underlying genome browser clients, it consists of a central reference sequence server, distributed annotation-containing servers and standardised XML data formats allowing for easy sharing of annotation data around the world. This presentation will discuss the functionality and suitability of DAS implementations to typical modern genomic datasets.

Visualisation tools for unsequenced genomes

Jen Taylor and Andrew Spriggs

CSIRO Plant Industry, Black Mountain Laboratories, Clunies Ross Street, Black Mountain
ACT 2601

Phone: +61 2 6246 4929

email: jen.taylor@csiro.au

Abstract

Genome browsing or visualisation is an integral part of modern genomics research. It facilitates the presentation of various meta-data, such as the relative positions of gene and regulatory elements and the efficient comparison of different types of genomic data such as whole genome gene polymorphisms and gene expression data. Typically, genome browsers use a linear coordinate system based on a largely sequenced reference genome assembly. Unfortunately, this approach is unsuited for visualisation of genomic data in partially assembled or unsequenced genomes such as wheat, barley and sugarcane. This poster presents an overview of tools adaptable to visualising genomic data in unsequenced genomes.

After Dinner Speaker: Dr David Lovell

Dr David Lovell (BEng, PhD, PGDipMgmt)

CSIRO Mathematical & Information Sciences, CSIRO ICT Centre & CMIS, Computer Science and Information Technology, Building, North Road, Acton ACT 2601

Phone: 61 2 6216 7042

email: David.Lovell@csiro.au

Biography:

David Lovell is an electrical engineer by training, and has been involved in a variety of roles related to the analysis of high-dimensional and large data, and a variety of roles related to managing this, and other kinds of research.

After receiving his PhD from the University of Queensland, Dr. Lovell completed postdoctoral research in perinatal risk predication at Cambridge University before joining CSIRO's Analysis of Large and Complex Datasets Group in 1998. In addition to this research role, Dr Lovell has also been Executive Officer to the CEO, and a member of CSIRO's Corporate Information Technology Management team.

Dr Lovell returned to CSIRO Mathematical and Information Sciences in 2004 as leader of Statistical Bioinformatics for Agribusiness, and then leader of the Division's Bio Program. In 2008 he was appointed co-leader of Transformational Biology, an initiative to help transform CSIRO's approach to biological research, and to help CSIRO's research transform biology itself.

6. DAY 2 ABSTRACTS

0835-0905: Enterprise data management services

John Morrissey

CSIRO Information Management and Technology, Bld 1, Banks Street, Yarralumla, ACT 2600

Phone: +61 2 6124 1411

email: John.Morrissey@csiro.au

John Morrissey is currently the Technology Theme leader for the CSIRO IM&T eScience Information Management program. With over 25 years experience in a wide variety of ICT roles in CSIRO ranging from Advanced Network Design to Data Management, John uses his skills to work with various research groups to find solutions to meet their ICT needs. John has spent the last two years leading CSIRO's engagement with the National Collaborative Research Infrastructure Strategy Platforms for Collaboration capability (NCRIS PfC). Recently John has been heavily involved in the development of ePublish, a new tool that combines a workflow management tool with a data repository in order to track and manage the scientific publication process across CSIRO.

Abstract

Science at CSIRO today is defined by high dimensional and large data sets. There is a global trend toward eResearch capabilities in the way that science is performed in the 21st Century.

Features of the scientific process today include:

- the use of specialised data gathering equipment such as sensors, meters and remote gathering tools;
- massive-scale data sets generated by 24/7 automated tools and processors
- the need for much better data management and the description and curation of key data assets
- a need for upscale computational resources to process these data sets
- the ability to collaborate with research partners residing outside one's own organisation, geographical region and research discipline

Speakers from IM&T will share some of their recent experiences where we have partnered successfully with key CSIRO research projects who have been experiencing these challenges.

We will share the various solutions we worked together to implement and some of the key learnings from those experiences. We will also outline some of the services or solutions we might be able to offer to other research leaders seeking assistance with moving and storing large data sets in this new world of science.

0905-0935: Katmandoo, a biosciences database.

David Butler

Queensland Department of Primary Industries and Fisheries, Plant Science, Toowoomba QLD 4350

Phone: +61 7 4688 1210

email: David.Butler@dpi.qld.gov.au

David Butler is a consulting biometrician with Qld Primary Industries & Fisheries and has worked closely with the organization's crop improvement programs for over 20 years. David has had an equally long interest in database and automation systems for these programs as well as being responsible for their statistical components for much of that time. More recently, David has become a collaborator with the ASReml software project and is currently responsible for the R language implementation of the package.

Abstract

Genetic improvement programs give rise to (often) large volumes of data from phenotypic observations, ancestral records, molecular data from laboratory genotyping and environmental characterization. Katmandoo integrates these sources in a relational database where the core relations respect the statistical principles of experimental units and sampling units . Simple extensions to the data model allow flexible treatment structures and accommodate multivariate response and repeated measures applications. This talk focuses on the design principles underlying components of Katmandoo and considers some options for large applications.

0935-1005: The flood of data: Large-scale sensor networks for water storage monitoring

Matthew Dunbabin

CSIRO ICT Centre, Pullenvale QLD 4069

Phone: +61 7 3327 4754

email: matthew.dunbabin@csiro.au

Matthew Dunbabin graduated from the Royal Melbourne Institute of Technology with a Bachelors Degree in Aerospace Engineering in 1995. He then worked as a research engineer at Roaduser Research, a Melbourne based consultancy before commencing his PhD at the Queensland University of Technology. In 2002 he received his PhD and joined the CSIRO field robotics team, which later became the ICT Centre Autonomous Systems Laboratory. He currently has the roles of Senior Research Scientist as well as Research Stream Leader for the development of advanced solutions to monitor and understand coastal marine environments. His research interests are in the area of field robotics, in particular underwater and mining robots, with focus on dynamics and control, underwater vision-based navigation, as well as robot and sensor network interactions.

Abstract

The CSIRO has been developing and installing a number of large-scale sensor networks for environmental monitoring across Queensland. Wireless sensor networks can complement existing monitoring programs by collecting an order of magnitude greater spatial and temporal data than traditional manual sampling programs. Through utilization of mobile sensors (robots), data collection rates can be further enhanced not only providing scientific measurements but also supporting static network operation. However, as the scale of the networks increase a number of challenges arise in terms of data distribution, storage and retrieval. Using the Lake Wivenhoe Sensor Network as an example, this presentation will describe the role that fixed and mobile wireless sensor networks can play in water storage and catchment monitoring. An overview of the technical aspects of the network including its purpose, size, information flow and modes of operation will be given. Also our latest achievements and preliminary work towards event detection within the network and our future research and operational directions will be described.

1005-1035: The storage, transfer, analysis and archive of underwater video

Ian McLeod

CSIRO Marine and Atmospheric Research, PO Box 120, Cleveland QLD 4163

Phone: + 61 7 3826 7185

email: Ian.Mcleod@csiro.au

Ian McLeod is a Spatial Analyst with CSIRO Marine & Atmospheric Research. He has been involved in technical computing and the analysis of large datasets for nearly 24years.

Abstract

Modern video technology now delivers vast quantities of Data. Compression technologies can help but in a lot of cases valuable information content is lost. Storage and transfer of these large information poor datasets poses difficult and interesting problems.

1055-1125: Organising and searching high dimensional data sets

Dave Abel

CSIRO Mathematical and Information Sciences, Building 108, North Road, ANU Campus, Acton ACT 2601

Phone: +61 2 6216 7033

email: dave.abel@csiro.au

Dave Abel is a Post-Retirement Fellow with CMIS and was previously a Chief Research Scientist. His current interests are in fundamental search operations and data mining algorithms for large multi-dimensional data sets. His approaches draw on experience in Operations Research, geospatial databases and distributed systems. His work over the past 5 years has been aimed at very large scientific data sets. Dave has been working on astronomical data to identify clusters of galaxies from sky survey data.

Abstract

A key computational issue associated with the analysis of high dimensional data is that of sorting, an example being the nearest neighbour query problem. Nearly all approaches to design of access methods for the exact k Nearest Neighbours query problem are based on partitioning using heuristics that are essentially greedy algorithms using simplified cost models. An

alternative not yet considered in depth is treat partitioning as a global optimisation problem. Key aspects of the design are formulating the partitioning operation as an optimization problem and devising a solution technique for partitioning that is reliable while not being prohibitively expensive. These and related issues will be addressed in this talk.

1130-1200: Terrestrial ecosystem data with high spatial and temporal variability

Peter Grace

Institute for Sustainable Resources, Queensland University of Technology, Brisbane QLD 4001

Phone: +61 7 3138 9283

email: pr.grace@qut.edu.au

Peter Grace is Professor of Global Change at QUT and Director of the Institute for Sustainable Resources. He is an agroecologist with specific interest in biocomplexity, multi-variate trend analysis and the simulation of plant-soil-atmosphere interactions.

Abstract

The assessment of ecosystem health requires the acquisition and synthesis of multiple streams of biotic and abiotic data. Two examples are provided, greenhouse gas emissions, specifically nitrous oxide emissions from agricultural systems, and landscape and organism specific acoustic data.

1200-1230: Challenges in analysing Seqwater intensive data

Sarah Lennox

CSIRO Mathematical and Information Sciences, Building 29, Long Pocket Laboratories, 120 Meiers Rd, Indooroopilly QLD 4068 Australia

Phone: +61 7 3214 2717

email: Sarah.Lennox@csiro.au

Sarah Lennox was appointed as a research scientist within CSIRO Mathematical and Information Sciences in January 2008. Sarah's current research interests lie in the area of environmental statistics, statistical modelling and trend analysis. Prior to the appointment with CSIRO she worked for the Queensland Climate Change Centre of Excellence. Sarah graduated from the University of Southern Queensland in 2003 with a Bachelor of Science (Honours) majoring in Applied Mathematics Statistics.

Abstract

The Queensland Bulk Water Supply Authority (trading as Seqwater) manages catchments, storages and water treatment plants to ensure a sustainable and high quality water supply. There is a growing need to understand the variability in the system and impacts of events on storages. Data therefore needs to be collected on smaller timescales (daily/hourly/minutes) at a variety of sites and depths. Two such sampling programs are currently in operation on Lake Wivenhoe. The first sampling program measures 11 water quality indicators at 3 sites, sampling every metre from surface to bottom every two hours. Stratification causes and event impacts across depths can be investigated along with daily and longer-term variability. The more recent placement of 50 wireless sensors on Wivenhoe currently recording water temperature at 6 depths every minute provides opportunities to study the spatio-temporal variability and to monitor short and longer-term event impacts. To analyse these invaluable datasets requires appropriate model structures and efficient computational techniques to handle the many issues that arise.

1230-1300: A statistical downscaling model for southern Australia winter rainfall

Yun Li

CSIRO Mathematical and Information Sciences, Leeuwin Centre, 65 Brockway Rd, Floreat WA 6014

Phone: +61 8 9333 6388

email: Yun.Li@csiro.au

Dr Yun Li is a senior research scientist with CSIRO Mathematical and Information Sciences (CMIS). He uses his expertise in statistical modelling of climatology and oceanography problems under projects supported by Australian Government, Western Australian State Government and CSIRO Climate Adaptation Flagship. Dr Li has developed deep expertise in statistical modeling of extreme rainfall events, and focuses on developing a high degree of expertise in statistical downscaling models using a nonparametric and semi-parametric modelling approach. He is the CMIS supervisor of the PhD program "Nonparametric and Semi-parametric Modelling of Spatial Data", supported by a CSIRO Corporate Postgraduate Scholarship and an Australian Postgraduate Award from the School of Mathematics and Statistics at the University of Western Australia. Dr Li is currently a principal investigator of the project "Research on Rainfall and Climate Change in both China and Australia" funded

through the Australia-China Bilateral Climate Change Partnership Programme, run by Australian Department of Climate Change.

Abstract

A technique for obtaining downscaled rainfall projections from climate model simulations is described. This technique makes use of the close association between mean sea level pressure (MSLP) patterns and rainfall over southern Australia during winter. Principal components of seasonal mean MSLP anomalies are linked to observed rainfall anomalies at regional, grid point and point scales. A maximum of 4 components is sufficient to capture a relatively large fraction of the observed variance in rainfall at most locations. These are used to interpret the MSLP patterns from a single climate model which has been used to simulate the both present day and future climate. The resulting downscaled values provide (a) a closer representation of the observed present day rainfall than the raw climate model values and, (b) provide alternative estimates of future changes to rainfall that arise due to changes in mean MSLP. While decreases are simulated for later this century (under a single emissions scenario), the downscaled values, in percentage terms, tend to be less

This work was supported by a Western Australian State Government Project-Indian Ocean Climate Initiative, the CSIRO Climate Adaptation Flagship project and the CSIRO Water for a Healthy Country Flagship project and the Australian Department of Climate Change via the Australian-China Climate Change Partnership Program.

1345-1415: Operational large-area land-cover monitoring using medium spatial resolution satellite imagery

Stuart Phinn, Tony Gill & Robert Dehnam

The University of Queensland, St Lucia, QLD 4072

Phone: +61 7 334 6 7019

email: t.gill1@uq.edu.au

Phone: +61 7 3365 6526 / 3346 7020

email: s.phinn@uq.edu.au

Department of Environment & Resource Management, QCCA Building, 80 Meiers Rd, Indooroopilly QLD 4068

Phone: +61 7 3896 9899

email: robert.denham@derm.qld.gov.au

*Stuart Phinn runs the Centre for Remote Sensing and Spatial Information Science (**Error! Hyperlink reference not valid.**), coordinates the Joint remote Sensing Research Program and*

teaches undergraduate and postgraduate remote sensing courses at the University of Queensland. His research and student project work focus on developing remote sensing applications for monitoring environmental conditions in a range of environments. Stuart sits on local, state and national advisory panels for environmental monitoring and management projects, providing advice on how to use make best use of current and future remote sensing data. He also plays a similar role for private industry and non-government organisations worldwide. He also provides a science advisory role for projects run by NASA, ESA, US National Science Foundation, UK National Environment Research Council and the Australian Research Council.

Tony Gill received the B.S. (Honours) degree in applied mathematics in 2002 and was a 3D applications programmer with the CSIRO from 2002 to 2005. He received his PhD in remote sensing in 2008 from the University of Queensland. He is currently a Postdoctoral Research Fellow with the Centre for Remote Sensing and Spatial Information Science at the University of Queensland. His current interests are vicarious calibration and radiometric correction of medium spatial resolution (10–30 m) satellite imagery and large area land cover monitoring.

Dr Robert Denham is an environmental statistician with the Remote Sensing Centre, Queensland Department of Environment and Resource Management. He has an interest in Bayesian statistics and particularly how it can be used to combine information from different sources and of different qualities. A current area of research involves the analysis of image time series, and he works in collaboration with DERM scientists and Dr Chris Strickland from Queensland University of Technology to develop efficient algorithms to model temporally dense remotely sensed imagery.

Abstract

Governments globally are recognising the urgent need for regular collection of nationally standardised and validated biophysical spatial information to understand where, how and why our environments are changing. Such information is used to develop policy and legislation to support sustainable land management practices. For example, in Queensland and New South Wales medium spatial resolution (10 m – 100 m) satellite imagery is combined with field data collection to monitor vegetation cover and change. These programs support the Vegetation Management Act 1999 and the Delbessie Agreement (also known as the State Rural Leasehold Lands Strategy) 2008 in QLD, and the Native Vegetation Act 2003 and Native Vegetation Regulation 2005 in NSW. These large area land cover monitoring programs have both

operational and research requirements that require a collaborative approach to develop. To support these requirements the Joint Remote Sensing Research Program (JRSRP) was formed in 2008. The JRSRP is a formal collaboration between the QLD Department of Environment and Resource Management (DERM), The NSW Department of Environment and Climate Change (DECC), and the Centre for Remote Sensing and Spatial Information Science based at the University of Queensland.

Members from the JRSRP will present the following selection of operational and research activities that support large-area dynamic land-cover monitoring:

- the data storage, database management, scripting and processing systems that underpin the operational activities;
- examples of the vegetation cover products, derived from time-series of Landsat imagery, that support QLD government legislation;
- current research activities in the field of time series image analysis for automating the detection of clouds, mapping burnt areas, detecting forest disturbance, and crop monitoring using both medium and low spatial resolution imagery;
- how the image data management, products and technical expertise of the group, will contribute to the NCRIS Terrestrial Ecosystem Research Network (TERN) Auscover initiative.

1415-1445: Spectroscopy and hyperspectral imaging for mineral and environmental applications

Mark Berman

CSIRO Mathematical & Information Sciences (CMIS) Locked Bag 17, North Ryde NSW 1670

Phone: +61 2 9325 3205

Email: Mark.Berman@csiro.au

Mark Berman received his B.Sc.(Hons.) degree and University Medal in mathematical statistics from the University of New South Wales in 1974, and a Master of Statistics degree from the

same institution in 1976. In 1978, he was awarded Ph.D. and D.I.C. degrees in mathematical statistics by the Imperial College of Science and Technology, London.

He was a visiting lecturer in the Department of Statistics at the University of California, Berkeley during 1978-1979. Most of his time since then has been with the CSIRO Division of Mathematical and Information Sciences (CMIS), Sydney, where he is now a Chief Research Scientist. He led CMIS' Image Analysis Group from 1989 to 2000. He spent 1988 at the Melbourne Research Laboratories of Broken Hill Proprietary Ltd. where he established the Image Processing and Data Analysis Group. His research interests are in image analysis (especially hyperspectral), spectroscopy and spatial data analysis.

Abstract

Every material has a distinctive spectrum. The spectrum of a material tells us about its chemistry. Hyperspectral images produce a spectrum (represented as several hundred numbers) at each pixel in an image. So hyperspectral images enable us to map variations in chemistry.

A number of CSIRO Divisions have been involved in the collection and analysis of airborne hyperspectral images for more than a decade, primarily for exploration and environmental applications. However, there are now a range of terrestrial applications; there are even hyperspectral microscopes! In addition, a number of hyperspectral satellites are due to be launched in the next few years.

A significant issue in hyperspectral imaging is that the spectra at many pixels in an image are actually mixtures of the spectra of the pure ingredients. CMIS' main focus has been on developing fast and sophisticated algorithms and software for "unmixing" these spectra into their pure ingredients, both when the pure ingredients are known and when they are unknown.

In this presentation, I will discuss some application areas (especially in mining and the environment), how CMIS' algorithms and software are being or are likely to be used in these applications, and some outstanding research problems.

1445-1515: Fast Bayesian analysis of spatial dynamic factor models for large space time data sets

Christopher Strickland

Queensland University of Technology, Gardens Point, QLD 4001

Phone: +61 7 3138 8313

email: christopher.strickland@qut.edu.au

Chris Strickland completed a PhD in Econometrics at Monash University, focusing on the Bayesian estimation of non-Gaussian state space models, in 2007. He is currently working as a post doctoral research fellow for Professor Kerrie Mengersen at Queensland University of Technology (QUT). He is employed under a linkage grant between QUT and the Department of National Resources and Water (NRW). They are currently developing Bayesian space time methods to meet the needs of NRW. In particular, They are looking at methods that can be used for land use mapping, the identification of clearings and the identification of the effect of different land management practices.

Abstract

Remote sensing is one example where data sets that vary across space and time have become so large that 'standard' approaches employed by statistical modellers for applied analysis are no longer feasible. We present a Bayesian methodology, which makes use of recently developed algorithms in applied mathematics, for the analysis of large space time data sets. In particular, a Markov chain Monte Carlo algorithm is proposed for the efficient estimation of spatial dynamic factor models (DFMs). The spatial DFM is specified whereby spatial dependence is modelled through the columns of the factor loadings matrix using a Gaussian Markov random field. Krylov subspace methods are used to take advantage of the sparse matrix structures that are inherent in the model. The methodology is used to analyse remotely sensed data from the Moderate Imaging Spectroradiometer satellite. In particular, the proposed methodology is used in conjunction with high resolution imagery for the classification, in terms of land type, of two regions located in central Queensland, Australia.

LIST OF PARTICIPANTS

CSIRO Mathematical and Information Sciences (CMIS) attendees

Dr Dave Abel	Dr Simon Barry	Dr Mark Berman
Lauren Bragg	Dr Mike Buckley	Dr Sam Burnham
Dr Murray Cameron	David Clifford	Ross Darnell
Dr Frank de Hoog	Dr James Doecke	Dr Scott Foster
Liya Fu	Dr Andrew George	Norm Good
Paul Greenfield	Dr. Keith Hayes	Dr Bronwyn Harch
Dr Emma Huang	Paul Jackway	Dr Warren Jin
Dr Harri Kiiveri	Dr Petra Kuhnert	Rex Lau
Emma Lawrence	Sarah Lennox	Dr Yun Li
Dr Xunguo Lin	Dr David Lovell	Maree O'Sullivan
Mark Palmer	Dr Erin Peterson	Dr Louise Ryan
Dr Ian Saunders	Dr Quanxi Shao	Dr Glenn Stone
Dr Bruce Tabor	Dr John Taylor	Dr Julian Taylor
Prof Ari Verbyla	Dr You-Gan Wang	Dr Alec Zwart

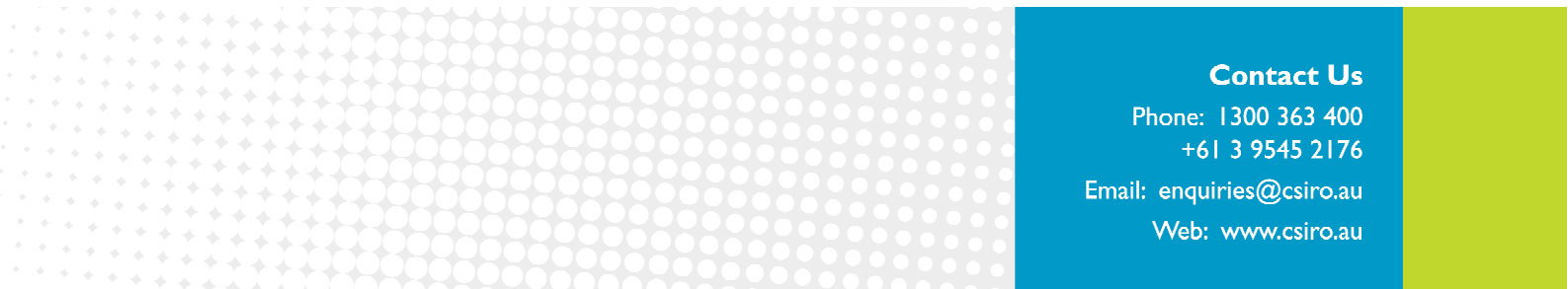
CSIRO attendees

Justin Baker	CSIRO IM&T
Dr Brian Dalrymple	CSIRO Livestock Industries
Dr Matthew Dunbabin	ICT Centre
Nicholas Hansen	CSIRO Plant Industry
Dr Richard Head	CSIRO Flagships and MXD Programs
Shivashankar Hiriyur-Nagaraj	CSIRO Livestock Industries
Dr Nick Hudson	CSIRO Livestock Industries
Dr Yutao Li	CSIRO Livestock Industries
Bryce Little	CSIRO Livestock Industries
Mr Ian McLeod	CMAR
Prof Mark Morrison	CSIRO Livestock Industries
John Morrissey	CSIRO IM&T
Yalchin Oytam	CSIRO MHT
Terry Rankine	CSIRO Exploration and Mining

Dr Luigi Renzullo	CSIRO Land and Water
Henry Scheele	CSIRO IM&T
Andrew Spriggs	CSIRO Plant Industry
Dr Jen Taylor	CSIRO Plant Industry
Dr Alf Uhlherr	CSIRO IM&T
Dr Gareth Williams	CSIRO HPSC

Non-CSIRO attendees

Jeremy Barker	Queensland Facility for Advanced Bioinformatics
Paul Berkman	University of Queensland
David Butler	Queensland Department of Primary Industries and Fisheries
Prof. Brian Cullis	NSW Department of Primary Industries
Robert Denham	DERM
Tony Gill	Centre for Remote Sensing and Spatial Information Science
Dr Dominique Gorse	Institute of Molecular Bioscience
Prof. Peter Grace	Queensland University of Technology
Dr Sean Grimmond	Institute of Molecular Bioscience
Dr Nick Hamilton	Institute of Molecular Bioscience
Edmund Ling	University of Queensland
Louise Marquart	Queensland Institute of Medical Research
Daniel Marshall	University of Queensland
Dr Geoff Morgan	Northern Rivers University Department of Rural Health
Dr Peter O'Rourke	Queensland Institute of Medical Research
Prof. Stuart Phinn	Centre for Remote Sensing and Spatial Information Science
Chris Strickland	Queensland University of Technology
Dr James Udy	SEQ Water
Dr Ian Wood	University of Queensland



Contact Us

Phone: 1300 363 400

+61 3 9545 2176

Email: enquiries@csiro.au

Web: www.csiro.au

Your CSIRO

Australia is founding its future on science and innovation. Its national science agency, CSIRO, is a powerhouse of ideas, technologies and skills for building prosperity, growth, health and sustainability. It serves governments, industries, business and communities across the nation.