a satellite meeting of the
international summer
school /workshop on
"Crops, Chips &
Computers"

2015年"生物大数据"国际暑期研习班及学术研讨会

2015 International Workshop and Summer School on "Biological Big Bytes"

Programme

September 6th - 11th, 2015

Nanning, China

Guangxi Teacher Education University

WELCOME TO C3-B3’2015

Dear Colleagues and Friends:

Welcome to C3-B3’2015 at the Guangxi Teacher Education University, Nanning, China, September6th-11th,2015!

This workshop and summer school on Biological Big Bytes (B3) is a satellite meeting of our traditional annual international workshop and summer school on "Crops, Chips and Computers" (C3), aiming to teach best practice in the application of state-of-the-art information technology and computational techniques to big data research in life sciences. The workshop and summer school will cover interdisciplinary science, in a rapidly evolving field, bringing together ideas from the fields of mathematics, bioinformatics, genomics and computer sciences. Because of these attractions, we expect the school to be of interest to a range of students from all academic levels, from post-graduate students to postdoctoral workers to junior academics.

Chang-An Yuan (Guangxi Teachers Education University)
Qingfeng Chen (Guangxi University)
Andrew Harrison (University of Essex)
Ming Chen (Zhejiang University)

Organizers

Dr. Andrew Harrison (University of Essex)
Dr. Chang-An Yuan (Guangxi Teachers Education University)
Dr. Christian Klukas (Leibniz Institute of Plant Genetics and Crop Plant Research)
Dr. Dechang Xu (Harbin Institute of Technology)
Dr. Huaqin He (Fujian Agriculture and Forestry University University)
Dr. Hugh Shanahan (Royal Holloway, Unversity of London)
Dr. Ling-ling Chen (Huazhong Agricultural University)
Dr. Ming Chen (Zhejiang University)
Dr. Qingfeng Chen (Guangxi University)
Dr. Sean May (The University of Nottingham)
Dr. Yingzhou Bi (Guangxi Teachers Education University)
Dr. Yongsheng Chen (Inner Mongolian University for the Nationalities)
Dr. Ziding Zhang (China Agricultural University)

Funding and Partners

The 2015 International Workshop and Summer School on Biological Big Bytes is sponsored by Guangxi Teacher Education University. It is jointly organized by the Department of Bioinformatics, College of Life Sciences, Zhejiang University, School of Computer, Electronic and Information, Guangxi University, and the Departments of Mathematical Sciences & Biological Sciences, University of Essex. It is hosted by Guangxi Teacher Education University, Nanning, China.

Local organizer

Dr. Chang-An Yuan, (Guangxi Teachers Education University), Chair
Dr. Yingzhou Bi (Guangxi Teachers Education University)
Dr. Zhi Zhong (Guangxi Teachers Education University)
Ms. Min Huang (Guangxi Teachers Education University)
Dr. Qingfeng Chen (Guangxi University), Co-Chair

Tel: +86 15994360606
Address: School of Computer and Information Engineering, Guangxi Teachers Education University, Nanning, 530000 P. R. China

Summer School/Workshop programme

Sunday 6th Sep 2015

Registration

Guangxi Teachers Education University / Litai International Hotel

17:30~ Dinner

Monday 7th Sep 2015

- Registration
  Guangxi Teachers Education University, College of Computer and Information Engineering, Ms. Min Huang: + 86 15994360606

- **Section 1 (Lectures)**
  - **08:30~09:45** Introduction & Lecture 1
    An overview of what needs to be done at the school and the expected work load on each day.
    Team grouping
    history of data science and technology
    Lecturer: Andrew Harrison, Hugh Shanahan, Adam Carter
  - **10:00~12:00** Lecture 2
    Bioinformatics – from sequence to function
    Lecturer: Shoba Ranganathan
  - **12:00~13:30** Lunch

- **Section 2 (Practical)**
  - **14:30~17:30** Practice 1
    Automating tasks with the Unix shell.
    Lecturer: Mitchell Stanton-Cook, Yuxi Luo
  - **18:00~** Dinner

Tuesday 8th Sep 2015

- **Section 3 (Lectures)**
  - **08:30~9:45 Lecture 3**
    Models for Processing Big Data - Moving Compute to the Data - Workflows - Data Intensive Compute Architectures Models for Processing Big Data - MapReduce (includes practical session)
    Lecturer: Adam Carter

- **10:00~12:00** Lecture 4
  Advanced data mining algorithms for bioinformatics problems
  Lecturer: Jinyan Li
- **12:00~13:30** Lunch

➢ **Section 4 (Lecture & Practical)**

- **14:30~17:30** Practice 2
  Building programs with Python 1
  Lecturer: **Mitchell Stanton-Cook, Yuxi Luo**
- **18:00~** Dinner

Wednesday 9th Sep 2015

Workshop Day

Conference Hall, Libarary of Guangxi Teachers Education University

➢ **Section 5 (Chair: Chang-An Yuan, Andrew Harrison)**

- **08:30~09:10**
  Opening Ceremony
  Speaker: President
  Group photo
- **09:10~9:40**
  Speaker 1: Andrew Harrison
  Title: The history and design of the CODATA-RDA summer schools in Research Data Science
- **09:40~10:10**
  Speaker 2: De-shuang Huang
  Title: Title: Prediction, De-noising and Modeling of PPI Networks Based on Machine Learning Techniques
- **10:10~10:30**
  Coffee/Tea break

➢ **Section 6 (Chair: Qingfeng Chen)**

- **10:30~10:55**
  Speaker 3: Jinyan Li
  Title: Structure bioinformatics and graph theory
- **10:55~11:20**
  Speaker 4: Hugh Shanahan

Title: Annotation of sequencing protocols for NGS data in the SRA - a cause for concern

- **11:20~11:45**
  Speaker 5: Mitchell Stanton-Cook
  Title: Doing bioinformatics pipelines better
- **11:45~12:10**
  Speaker 6: Hong Li / Ziding Zhang
  Title: Competition-cooperation relationship networks (CCRNs) characterize the competition and cooperation between proteins
- **12:30~13:30** Lunch

➢ **Section 7 (Chair: Huaqin He)**

- **14:30~14:55**
  Speaker 7: Yuxi Luo
  Title:
- **14:55~15:20**
  Speaker 8: Qiang Jiang
  Title:
- **15:20~15:45**
  Speaker 9: Youhuang Bai
  Title: Genomics analysis for aflatoxin biosynthesis in Aspergillus flavus
- **15:45~16:10**
  Speaker 10: Jian Jin
  Title: Genomic analysis of Arabidopsis flower development gene regulatory network
- **16:10~16:30**
  Coffee/Tea break

➢ **Section 8 (Chair:)**

- **16:30~16:55**
  Speaker 11: Adam Carter
  Title: Model Driven Preservation - The PERICLES project
- **16:55~17:20**
  Speaker 12: Zeeshan Gillani/ Ming Chen
  Title: large scale image high throughput phenotyping in Plants
- **17:20~17:45**
  Speaker 13: Liang Zhao
  Title: MapReduce-based approach for next-generation sequencing data
- **18:00~20:00** Dinner & Discussion

Thursday 10th Sep 2015

- **Section 7 (Practical)**
  - **09:00~11:30** Practice 3
    Building programs with Python 2
    Lecturer: **Mitchell Stanton-Cook, Yuxi Luo**
  - **12:00~13:30** Lunch

- **Section 8 (Practical)**
  - **14:30~17:30** Practice 4
    Version control with Git
    Lecturer: **Mitchell Stanton-Cook, Yuxi Luo**
  - **18:00~** Dinner

Friday 11th Sep 2015

- **9:30~11:40** Close session & Awards ceremony
  Science in the UK, Australia and China, and moving among countries; career structures in these countries. We expect to recruit foreign students who may be considering spending part of their careers in China and we hope some of the Chinese students similarly look to spend part of their careers overseas.
- **12:00~** Lunch & Farewell

Workshop abstracts

Title：The history and design of the CODATA-RDA schools on Research Data Science

Andrew Harrison *University of Essex*

Abstract: All research fields are generating data at a copious rate. The need to make best use of this data is opening up new opportunities as well as identifying clear challenges. The organisations CODATA and Research Data Alliance are leading efforts to develop a coordinated global approach across all research disciplines to enable all scientists from all nations to benefit from the "Data Revolution". At the heart of this initiative is the development of a global educational programme in Research Data Science. The design of the programme is leading to a delivery model we describe as Education as a Service. I will outline the design and composition of the programme.

Title：Prediction, De-noising and Modeling of PPI Networks Based on Machine Learning Techniques

De-shuang Huang Tongji *University*

Abstract: In this talk I firstly over-review the general concepts for biological data analyses based on machine learning techniques. Then I will discuss how to predict the protein-protein interactions (PPI) from protein sequences. After that, I will present how to perform the denoising of PPIs based on network topology information from the viewpoints of manifold learning of ISOMAP. Further, how to model the PPI networks based on novel t-Logical Semantic Embedding geometric approach will be also presented in details. Particularly, some simulating experiments will be demonstrated to verify the performances for our approaches. Finally, some future research directions will be discussed

Title：Bioinformatics – from sequence to function

Shoba Ranganathan *Macquarie University*

Abstract: Bioinformatics is a relatively new scientific discipline, which seeks new biological insights, using computational approaches. It signifies the unification of biology, computer science and information technology. With the advent of high throughput analytical technologies in molecular biology and biochemistry, bioinformatics holds the key to understanding the information encapsulated by genes and proteins within genomes and proteomes and their interaction with metabolites and drug molecules. An individual's personal genetic makeup determines their state of health as well as their response to prescribed medications, for personalized health care. At the same time, the threat of emerging diseases places a huge burden of public health costs, requiring a detailed molecular level analysis of pathogens for vaccine development.

With the advances in sequencing technologies and high-throughput analytical instrumentation, individual biological sequences are generated at an ever-increasing pace. Where this data is stored and how to intelligently search these databases are the first topics to be presented. We will then look at how to compare these sequences and groups them into families. At the sequence level, patterns or motifs are quick tools to rapidly assign function to a new sequence. At the extreme end of sequencing and analytical efforts are the availability of entire genomes and proteomes. When we have lists of complete sets of genes and proteins, we are able to better analyse how they function as teams in pathways and how these teams interact as networks.

An example of how to put these ideas together, and apply it to a biological problem will be presented.

Title：Structure bioinformatics and graph theory

Jinyan Li *University of Technology Sydney*

Abstract: I will introduce three graph-related bioinformatics problems. One is protein complex prediction, the second one is protein binding hotspot prediction, the third one is conformational B-cell epitope prediction. I will illustrate why graph theories and algorithms are important to solve these problems.

Title：Annotation of sequencing protocols for NGS data in the SRA - a cause for concern

Hugh Shanahan *Unversity of London*

Abstract: The workflow for the production of high-throughput sequencing data from nucleic acid samples is complex. There are a series of protocol steps to be followed in the preparation of samples for next-generation sequencing. The quantification of bias in a number of protocol steps remains to be determined.

We examined the experimental metadata of the public repository Sequence Read Archive (SRA) in order to ascertain the level of annotation of important sequencing steps in submissions to the database. Using SQL relational database queries (using the SRAdb SQLite database generated by the Bioconductor consortium) to search for keywords commonly occurring in key preparatory protocol steps partitioned over studies, we found that 7.10%, 5.84% and 7.57% of all records (fragmentation, ligation and enrichment, respectively), had at least one keyword corresponding to one of the three protocol steps. Only 4.06% of all records, partitioned over studies, had keywords for all three steps in the protocol (5.58% of all SRA records).

The current level of annotation in the SRA inhibits systematic studies of bias due to these protocol steps. Downstream from this, meta-analyses and comparative studies based on these data will have a source of bias that cannot be quantified at present.

Title：Doing bioinformatics pipelines better

Mitchell Stanton-Cook *Unversity of Queensland*

Abstract: We have developed the Banzai Microbial Genomics Pipeline (https://github.com/mscook/Banzai-MicrobialGenomics-Pipeline). Banzai simplifies the analysis of microbial next-gen sequencing (NGS) datasets. Banzai was specifically designed to distribute workload over internal and external High Performance Computing (HPC) resources. Banzai (in most cases) does not provide new NGS algorithms, but harnesses the power of tried and tested NGS tools. Banzai simplifies, automates and distributes computational workloads, which is the typical bottleneck in analysis of large NGS datasets. Here, I will discuss how we future proofed and migrated our pipeline from an aging HPC resource using DevOPs approaches. I will particularly focus on the use of container technologies, which have allowed us to build, ship and run reproducible bioinformatics analyses. This is a first hand lesson how we significantly changed our development practices to consider long-term sustainability of our analysis pipeline. In less than three years, the Beatson Laboratory has had to scale from the simultaneous analysis of 10 to 100 to 1000 genomes. By understanding and employing DevOps based approaches the genomics community will be able to build and scale reproducible analysis pipelines with minimal modification to their existing processes.

Title： Competition-cooperation relationship networks (CCRNs) characterize the competition and cooperation between proteins

Hong Li / Ziding Zhang *China Agricultural University*

Abstract: By analyzing protein-protein interaction (PPI) networks, one can find that a protein may have multiple binding partners. However, it is difficult to determine whether the interactions with these partners occur simultaneously from binary PPIs alone. In this report, I will introduce the yeast and human competition-cooperation relationship networks (CCRNs) which are constructed based on protein structural interactomes to clearly exhibit the relationship (competition or cooperation) between two partners of the same protein. If two partners compete for the same interaction interface, they would be connected by a competitive edge; otherwise, they would be connected by a cooperative edge. The properties of three kinds of hubs (i.e., competitive, modest, and cooperative hubs) are analyzed in the CCRNs. Results show that competitive hubs have higher clustering coefficients and form clusters in the human CCRN, but these tendencies are not observed in the yeast CCRN. We find that the human-specific proteins contribute significantly to these differences. Subsequently, we conduct a series of computational experiments to investigate the regulatory mechanisms that avoid competition between proteins. Our comprehensive analyses reveal that for most yeast and human protein competitors, transcriptional regulation plays an important role. Moreover, the human-specific proteins have a particular preference for other regulatory mechanisms, such as alternative splicing.

Reference: Hong Li, Yuan Zhou, and Ziding Zhang. (2015). Competition-cooperation relationship networks characterize the competition and cooperation between proteins. Scientific reports 5:11619.

Title： Experimental study of speed up techniques for sparse autoencoder and my workflow

Yuxi Luo Institute of Modern Physics, CAS

Abstract: The success of machine learning algorithms generally depends on data representation. So far there has been a great deal of literature on unsupervised feature learning and joint training of deep learning. The training process of traditional sparse auto-encoder is very time-consuming. We propose a novel method for training sparse auto-encoders quick. And we hope this simple method would enlighten the other researchers to pay attention to the feature operations on their own field. This talk would also introduce the tools, tips and workflow which are refined in the research process.

Title：On Network

Qiang Jiang *City University of HongKong*

Abstract:

Title：Genomics analysis for aflatoxin biosynthesis in Aspergillus flavus

Youhuang Bai *Fujian Agriculture and Forestry University*

Abstract: *Aspergillus flavus* is a saprophytic filamentous fungus that is also able to contaminate economically important crops such as peanuts, cotton, maize and other oils seed crops during pre-harvest or storage. This opportunistic pathogen produces a wide range of secondary metabolites, including aflatoxins. Ingestion of food products contaminated with aflatoxins has been associated with hepatotoxicity, teratogenicity, immuno-suppression and liver cancer. The major determinant of aflatoxin production in *A. flavus* are temperature and water activity. Deep sequencing was performed to characterize the transcriptome and proteome level changes in response to temperature and/or water activity. We demonstrated that there was a low correlation between the transcriptome and proteome data, suggesting that post-transcriptional gene regulation influences different biological pathways and secondary metabolite gene clusters. In addition, miRNA-like (milRNA) genes identified imply that they might play important roles in the mycotoxin biosynthesis and mycelium growth in *A. flavus*.

Title：Genomic analysis of Arabidopsis flower development gene regulatory network

Jian Jin *Guangxi University*

Abstract: Flowering is a critical developmental stage for angiosperms to reach the next generation. The transition from vegetative growth to flower formation requires a large-scale alteration of transcriptional programs(Moyroud *et al,* 2010). Though many of them have been genetically characterized, the gene network structure still remains incomplete.

A synchronized floral induction system has been used for the study of *APETALA1*, which plays a key role in floral development. Genome-wide approaches such as microarray and Chip-seq (chromatin imunoprecipitation followed by next generation sequencing) are used to characterize the gene network controlled by *AP1*. (Kaufmann *et al*, 2010; Wellmer *et al,* 2006)

CAULIFLOWER, the closest homolog of *AP1*, acts in a partially redundant manner：
it participates in floral meristem identity determination (which is an *AP1* "early" function), but it is not involved in floral organ identity specification and development (a "late" function of *AP1*).

The aim of our study was to apply genome approaches to identify the gene network that is controlled by *CAL*. Furthermore, because of the partial redundancy between AP1 and CAL, *CAL* is used as a tool to dissect the early and late function of *AP1*.

Title：Model Driven Preservation - The PERICLES project

Adam Carter *University of Edinburgh*

Abstract: Digital Preservation is about more than preserving bits. The PERICLES project is a 4-year EU-funded project which is looking at novel ways to solve the problems of digital preservation. One of the approaches being explored in the project is model-driven preservation (where the models are expressed, e.g., in RDF and BMPN). In this overview talk, I’ll discuss how using graph-based models could facilitate re-use of digital objects in a constantly changing ecosystem.

Title：large scale image high throughput phenotyping in Plants

Zeeshan Gillani/ Ming Chen *Zhejiang University*

Abstract: There has been increase in the consumption of food and fuel due to rapidly increasing population and to ensure food security for rapidly increasing population, there is need for high yielding crops that can adapt to future climate. To solve these issues across globe, novel techniques and methodologies are needed to provide quantitative phenotypes to elucidate the genetic basis of agriculturally important traits and to screen germplasm with super performance in function under resource-limited environment. Currently, plant phenomics has offered and the focus had only been on integrated suite technologies for understanding the complete set of phenotypes of plants, towards the progression of the full characteristics of plants with whole sequenced genomes. Therefore high-throughput phenotyping platforms have been developed to capture extensive and intensive phenotyping data from non-destructive imaging over time. These advance has enabled us to study plant growth and performance with respect to changing climate and environment. In this presentation, we briefly discuss currently developed imaging techniques and challenges posed by them.

Title：MapReduce-based approach for next-generation sequencing data

Liang Zhao *Guangxi University*

Abstract: Next-generation sequencing platforms such as Illumina, Solexa and Sanger have produced huge amount of sequence data which are revolutionizing every aspect of genetic and genomic research. However, these sequence data sets contain many machine-induced errors, including substitutions, insertions and deletions. The error rate is high. For example, the errors due to substitution can be as high as 2.5%. Existing methods are not perfect in error correction, and they also have poor parallelization issues. The key steps such as k-mer frequency cut off, distance calculation between k-mers, or the iterative correction of k-mers cannot be parallelized. Their applicability is severely limited to large data sets. We propose a parallelizable error correction approach, a two-layered MapReduce algorithm, to gain both accuracy and scalability. At the first layer, the set of input sequences are mapped into subgroups so that the candidate erroneous bases can be identified in parallel; and at the second layer, the erroneous bases at the same position are mapped together from all the subgroups to improve the correction rates. This two-layered MapReduce has a well-controlled time and space complexity. Our experimental results on real and simulated data sets show that the newly designed algorithm outperforms remarkably over the existing approaches on the substitution error corrections which are the main type of errors to be corrected in the market.

Contact

Email: All questions related to the C3'2015 can be sent to
Dr. Chang-An Yuan (yuanchangan@126.com, Guangxi Teachers Education University)
Dr. Ming Chen (mchen@zju.edu.cn, Zhejiang University) or
Dr. Andrew Harrison (harry@essex.ac.uk) (University of Essex)
Ms. Min Huang (hm@gxtc.edu.cn, Guangxi Teachers Education University)

Travel

By Air-plan:

From Airport to Guangxi Teachers Education University:
You may come to Nanning via Beijing / Shanghai/Guangzhou.
Taxi: Taxi from airport to the Mingxiu Road campus of Guangxi Teachers Education University costs about 100 RMB
Airport Bus: Chaoyang Square Line: costs 20 RMB, stop at terminal station, and transfer to Mingxiu Road campus Guangxi Teachers Education University: cost about 30RMB

By Train:

There are two railway stations in Harbin.
From Nanning Station:
Taxi: Taxi from Harbin Station to the Mingxiu Road campus costs about 30 RMB.
From Nanning East Station:
Taxi: Taxi from Nanning East Station to the Mingxiu Road campus costs about 60 RMB.

Note. If you have any difficulty in airport/railway station, please contact Ms. Min Huang by mobile + 86 15994360606.

Hotel

Litai International Hotel
Mingxiu Road 157, Xixiang Tang District, Nanning

Local attractions include:

Nanning skyline (2008)

Nanning International Convention and Exhibition Center

Nanning Mosque

Jinhu Square

View from Diwang International Commerce Center

Workshop is organized by:

廣西師範學院
GUANGXI TEACHERS EDUCATION UNIVERSITY

计算機与信息工程學院
Computer and Information Engineering

Co-organized by:

亚热带农业生物资源保护与利用国家重点实验室
(SKL for Conservation and Utilization of Subtropical Agro-bioresources)

浙江省生物信息学学会
(Bioinformatics Society of Zhejiang Province)

廣西大學
Guangxi University

浙江大學
Zhejiang University

University of Essex