HEALTH AND BIOSECURITY
www.csiro.au

Bioinformatics Student Exchange Program

CSIRO – Germany

1 BSEP 2018

Australia was featured in a recent Nature article stating that "Scientists from across the world are attracted to the country, which competes internationally by focusing on its strengths". The Commonwealth Scientific and Industrial Research Organisation (CSIRO) is one of the largest and most diverse scientific organisations in the world. By igniting the creative spirit of our people, we deliver great science and innovative solutions that benefit industry, society and the environment.

In order to give overseas students the opportunity to contribute to world-class research and gain experience in an international research environment, the eHealth program is running the Bioinformatics Student Exchange Program (BSEP) with foreign Universities. The program is aimed at Master and Honours students and invites them to join CSIRO to conduct original research. This is an exciting opportunity to forge new collaboration with CSIRO as the hub for bioinformatics research in Australia.

Master and Honours students in Bioinformatics will have the opportunity to join CSIRO for 23 weeks (5 months) and undertake a research project that contributes towards their Thesis. The project will be proposed by CSIRO researchers who also agree to co-supervise the student and assist in writing the thesis.

University	Contact Person
Freie Universität Berlin	Prof. Dr. Annalisa Marsico RNA Bioinformatik Phone: +49 30 8413 1843 Fax: +49 30 8413 1960 Email: Annalisa.Marsico@fu-berlin.de
Eberhard Karls University Tübingen	Dr. Julian Heinrich Applied Bioinformatics Group Email: heinrich@informatik.uni-tuebingen.de
Justus-Liebig-University Giessen	Prof. Dr. Alexander Goesmann Bioinformatik und Systembiologie Tel. +49 (0)641 99-35801 Email: Gwyneth.schulz@computational.bio.uni-giessen.de
CSIRO	Dr. Denis Bauer Transformational Bioinformatics, eHealth, CSIRO Phone: +61 2 9325 3174 Email: denis.bauer@csiro.au

1.1 Key dates

Date	
June	CSIRO calls for project proposals
31st July	Program Booklet sent to the Universities
Early August to early November	Deadline for PROMOS or equivalent funding application
Dec	Thesis committee assesses suitability of projects and identifies appropriate co-supervisor amongst the faculty.
Jan	Students choose proposals and CSIRO starts recruitment process (interview, visa)
May	Students commence research in Australia
Oct	Students return home
Nov	Students finalise reports and write master thesis with input from CSIRO researchers

1.2 Funding

Students are encouraged to apply for funding. Unless stated otherwise, the projects will not provide funding.

1.3 Germany

PROMOS

German funding through PROMOS (Deadline Early October to early November), which will cover

- from 300 to 500 EUR per month and / or
- Traveling costs up to 1950 EUR

Note, PROMOS is not explicitly paying a health insurance, this hence needs to be covered by the student.

DAAD

The DAAD offers FIT- Internationale Forschungsaufenthalte in der Informationstechnologie für Masterstudierende, which can be applied for at any stage, with notification of success within 3 Months (recommended application date no later than October)

- 875 EUR per months
- contributions to travel costs
- contributions to insurances

There are also other funding sources available such as http://www.ranke-heinemann.de.

1.4 How to apply and other resources

Please choose the project you are interested in and get in touch with your contact person listed above. Your first step will be to organize funding by applying for PROMOS or equivalent sources (DAAD). After a successful interview in January, CSIRO will issue a contract with a visa sponsorship number. It is crucial to apply for the Australian Visa quickly as it can take up to 3 months to be approved.

VISA:
https://www.border.gov.au/Trav/Visa-1/407-

Information about the VISA subclass 407 Trainee and Research.

Address where to send the application:

https://www.border.gov.au/Lega/Lega/Help/Location/australia
Tasmania-Hobart office

Health insurance:
http://www.health.gov.au/internet/main/Publishing.nsf/Content/Overseas+Student+Health+Cover+FAQ-1#insurersofferoshc
e.g. the BUPA caters for VISA subclass 407.

German information on going to Australia:
http://www.reisebine.de/

Official government website with information about studying and living in Australia
www.studyinaustralia.gov.au

1.5 Experience Report from 2017

We are Amnon and Marc, students of the Bioinformatics M.Sc exchange program at FU-Berlin. We both joined Dr. Bauer's transformational bioinformatics group, located in Sydney, in which we spent 6 months between May and November. We want to share with you our experience of working at CSIRO and living in Sydney.

The student exchange program offers a unique opportunity to experience full-time research as part of a professional research group. Being a governmental institute, CSIRO constitutes a fine balance between research and industry and therefore the researchers working here come from diverse professional and educational backgrounds, a fact that helps broaden one's perspective.

Working at CSIRO gave us a taste of how research life is, having regular group meetings, brainstorming sessions, and seminars. Furthermore, the dedication of the group members to help one another (us included) expressed itself in divoted supervision that was given to us by Denis, the head of the group and an additional supervisor from the team. All in all, it was a great opportunity for us to learn not only about the subject of our research, but also broaden our view on other related subjects and the work in a group. In addition, depending on the project, you might receive a living allowance which will help you to focus on your work, and – equally important – have a good time in Australia!

Participating in the exchange program requires a few bureaucratic steps. First, you need to secure funding via one of the many different private and government organizations handing out stipends to students undertaking research activities aboard. We both acquired a scholarship via PROMOS. You should select one of the projects from the BSEP booklet beforehand, as it helps you to formulate the motivational letter required by most stipend organizations.

Next, you should choose a local supervisor at FU (who will also grade your final thesis) and get in touch with your project supervisors in Australia. After this, it's time to apply for a visa and start looking for accomodation. We suggest you arrive a week or two before starting your project and take your time choosing the right location, as Sydney is a very spread out city and commuting times are quite high. We both chose to live in the city center and take a train for about 45 minutes to the office. Another option is to live closer to work, which is cheaper both in rent and transportation.

For us, living in Sydney was an exciting opportunity to live in a city with different culture and atmosphere to Berlin. It is a city with great nature (which can conveniently be accessed during the weekends), great weather and beautiful beaches. Sydney is a multicultural city and a place that attracts many travelers and temporary residents, which all makes it rather easy to meet new people.

To conclude, this student exchange program was a great experience for us and we warmly recommend applying for it, both for professional and non-professional reasons. If you have any further question, feel free to contact us.

Cheers,
Marc and Amnon
(Marc.Horlacher@csiro.au, Amnon.Bleich@csiro.au)

1.6 Projects

BSEP01: Analysis of influenza A specific T-cell receptor sequencing data

Influenza A remains a high biological and economic threat to modern societies globally. Flu vaccines have been widely used to combat influenza A viruses (IAV), however, these vaccines have only showed up to ~40% efficiency due to the high mutation rate of IAV. In recent years, T-cell receptor (TCR) engineering based immuno-therapy has been proposed as an alternative vaccination approach for IAV. In collaboration with Prof. Guobing Chen, Jinan University, China, we are generating high-throughput TCR sequencing data specific to 26 subtypes of IAV epitopes. Deep analysis of these IAV epitope-specific TCR sequencing data will help to elucidate the best candidate TCR peptide(s) for developing the next-generation IAV vaccines.

The student will perform:

- High-throughput TCR sequencing data analysis including sequencing data pre-processing and mapping
- Quantitative analysis of epitope-specific TCR data
- HLA super-family analysis

Relevant fields of study:

- Machine learning
- Statistical analysis
- Scripting programming (bash, R, perl/python etc.)
- Immuno-genomics

Supervisor

Oscar Luo (CSIRO, Health & Biosecurity)

Email: Oscar.Luo@csiro.au

Location: Sydney

Funding:
CSIRO offers a top-up stipend for this project: $300/week contribution towards living expenses

BSEP02: Epistasis data simulation for GWAS

GWAS has been a widely used approach in genetic research over the last years to identify disease causing genomic variations. Traditional analysis methods for GWAS such as logistic or linear regression tend to underestimate associations of SNPs with a non-Mendelian phenotype effect (one that derives from a multi loci). This led to use of different algorithms for associating SNPs to phenotypes, in a way that will take SNPs interaction into account. One major setback is the lack of gold standard data with known causal SNPs, that would serve as a reliable benchmarking base. This project will focus on GWAS data simulation with underlying causal SNPs including SNP interaction. The different confounding factors should be studied and simulated in a realistic way, taking into account factors such as minor-allele frequency, SNPs correlation, missing data, linkage disequilibrium etc. The different factors should be flexible so that the user can adjust it to his/her needs (e.g how strong should correlation be? The extent of missing data etc.)
Side note - several simulation tools already exist, such as epiSIM (Shang et al., 2013a) but not all confounding factors are included in it.

The student will perform:

- Study of GWAS and the confounding factors in current analysis methods.
- Study of population genomics
- Write algorithm for realistic GWAS data simulation, including confounding factors

Relevant fields of study:

- Statistics
- Machine learning
- Population genetics
- Scripting programming (bash, R, perl/python etc.)

Supervisor

Denis Bauer (CSIRO, Health & Biosecurity)

Email: Denis.Bauer@csiro.au

Location: Sydney

Funding:
CSIRO offers a top-up stipend for this project: $300/week contribution towards living expenses

BSEP03: Identifying Dangerous DNA within a genome

Integration of foreign DNA into an organism’s genome can have serious implications. The insertion of specific retroviruses into human liver cells has been linked to the development of cancer, while integrations of plasmids can transform benign bacteria into pathogenic strains.
It is therefore critical that we develop methods for detecting when and where foreign DNA has been inserted into a genome. The aim of this project will be to develop such a model, taking advantage of a genomes unique “genetic signature” in order to identify regions that do not belong.

The student will perform

- Design and implementation of a model to identify foreign regions within a genome
- Validation of model using a custom built experimental dataset

Relevant field/s of study

- machine learning
- Statistics
- Experience in R/Python or another suitable language

Supervisor
Laurence Wilson (CSIRO, Health & Biosecurity)
Email: laurence.wilson@csiro.au
Location: Sydney

Funding:
CSIRO offers a top-up stipend for this project: $300/week contribution towards living expenses

BSEP04: Visual interpretation of VariantSpark results

VariantSpark is a machine learning library for real-time genomic data analysis designed to cater for ``big'' (many samples) and ``wide'' (many features) datasets. VariantSpark uses a custom machine learning random forest implementation to find the most important variants attributing to a phenotype of interest. VariantSpark is able to identify co-occuring variants which are interactively associated with phenotype. The current implementation of VariantSpark does not allow for easy visualisation of interacting variants and their level of importance relative to phenotype of interest.

The student will perform:

- Add functionality to existing implementation of VariantSpark to allow visualisation of variant interactions and importance relative to phenotype.

Relevant fields of study:

- Computer science
- Machine learning
- Statistical analysis
- Scripting programming (Python/R)

Supervisor
Natalie Twine
Email: natalie.twine@csiro.au
Location: Sydney

Funding:
CSIRO offers a top-up stipend for this project: $300/week contribution towards living expenses

BSEP05: A fast implementation of genomic sequence data phasing for ‘TRIBES’

Accurate identification of genetic relatedness between individuals is important for multiple fields of research, including disease gene discovery and genome-wide association studies. We have developed ‘TRIBES’, a software tool to identify distant relatives based on their genome sequence. An essential step during relatedness analysis using ‘TRIBES’ is phasing of genomic sequence data. Current phasing tools (such as BEAGLE or SHAPEIT) require substantial compute time with large sample sizes. To improve the compute time for running ‘TRIBES’, a faster implementation of genomic sequence data phasing is required.

The student will perform:

- Add functionality to existing implementation of TRIBES to allow for fast phasing of genomic sequence data
- Demonstrate that the implementation is faster than gold standard tools, BEAGLE/SHAPEIT

Relevant fields of study:

- Computer science
- Machine learning
- Statistical analysis
- Scripting programming (Python/R)

Supervisor
Natalie Twine
Email: natalie.twine@csiro.au
Location: Sydney

Funding:
CSIRO offers a top-up stipend for this project: $300/week contribution towards living expenses

BSEP06: Optimizing for variable importance rather than prediction

Random-Forest is an ensemble and supervised machine learning methods used in prediction application. Random-forest is built from Decision tree models which are particularly strong in considering the interactions between various features. In the health domain, where we expect interactions between different omics data, Random-Forest could be an ideal chose for processing data. While the model has been well used for prediction purpose, it is also possible to do an association test with the model to find out how different features are correlated with a specific phenotype when considered together. However, there has been less effort to tailor the algorithm for association studies. Also, genomics is one of the health omic data which grows fast in terms of the volume of data to be processed. In order to be feasibility, Random-Forest parameter requires to be tuned for such multi-omics association test when massive genomic data is presented.

The student performs:

- Study current improvement to RandomForest method.
- Process and analyse existing real-world datasets.
- Identify the algorithmic weakness and provides solutions and improvements.

Relevant field of study:

- Bioinformatics
- Statistics
- Biology
- Computer Science.

Supervisor
Arash Bayat
Email: arash.bayat@csiro.au
Location: Sydney

Funding:
CSIRO offers a top-up stipend for this project: $300/week contribution towards living expenses

BSEP07: Apply genome-scale machine-learning to “big” datasets in other disciplines

The automatic collection of information such as from internet of things (IoT) devices causes datesets to grow rapidly in all disciplines. “Wide” data, that is millions of datapoints per sample, was originally encountered in the genomic discipline. Here, millions of genomic variants describe a patient, and with cohort sizes ranging in the 10 thousands, analysis tasks would easily reach several trillion datapoints. The analysis tool, VariantSpark, was developed to perform machine learning on such large, high-dimensional datasets. Be part of the team that applies the smarts of genome research to other disciplines.

The student will perform:

- Identify non-life-science dataset to apply VariantSpark to
- Compete in Kaggle competitions and make a name in the Machine Learning community
- Contribute to changes in VariantSpark to make it application agnostic.

Relevant fields of study:

- Computer science
- Machine learning
- Statistical analysis
- Programming (Python/Scala)

Supervisor
Denis Bauer
Email: Denis.Bauer@csiro.au
Location: Sydney

Funding:
CSIRO offers a top-up stipend for this project: $300/week contribution towards living expenses

BSEP08: CRISPR-Cas9 applications in human health

The development of the CRISPR-Cas9 system has revolutionized genome engineering, making it possible to directly target almost anywhere in the genome for editing. The reliable application of the CRISPR-Cas9 technology requires the identification of the optimal target site, as activity can vary substantially between sites. This is particularly important if the technology is hoped to be applied in the human health space.
The goal of this project will be to use BigData and Machine Learning approaches to understand what factors contribute to CRISPR-Cas9 activity and how these can be leveraged to improve predictions of target activity.
The student will build upon the team's extensive CRISPR software portfolio, contributing to the next version of GT-Scan with the option of writing a first-author paper. Furthermore, the student will have the opportunity to develop cloud-based services, likely in close collaboration with AWS Solution Architects in Sydney.

The student will perform

- Research and review of potential applications of CRISPR-Cas9 in the area of human health
- Analysis of CRISPR-Cas9 activity *in vitro*
- Development of predictive models

Relevant field/s of study

- Machine Learning
- Genome Engineering
- BigData

Supervisor
Laurence Wilson, Denis Bauer (CSIRO, Health&Biosecurity)
email Laurence.wilson@csiro.au
Location: Sydney NSW

Funding
CSIRO offers a top-up stipend for this project: $300/week contribution towards living expenses.

BSEP09: Extending the CursedForest (a distributed random forest) framework for genomics

CursedForest is a tailored Hadoop/Spark-based implementation of random forests specifically designed to cater for ``big'' (many samples) and ``wide'' (many features) datasets, which was recently featured in the Databricks Engineering Blog [1].

The current implementation has been successfully used in detecting SNPs associated with a phenotype in application with up to 80M variables. Making CursedForest applicable to more genomic research areas, such as transcription analysis, this project aims to add some of these features:

1) handling of categorical variables (by subset selection);

2) a regression penalty term (currently uses Gini index);

3) proximity matrix;

4) conditional inference trees.

[1] https://databricks.com/blog/2017/07/26/breaking-the-curse-of-dimensionality-in-genomics-using-wide-random-forests.html

The student will perform

- design of implementation
- coding in Scala and perhaps other languages
- testing of implementation, documentation

Relevant field/s of study

- Computer science
- machine learning
- statistics

Supervisor
Denis Bauer (CSIRO, Health&Biosecurity), Robert Dunne, Piotr Szul (CSIRO, D61)
Email: Denis.Bauer@CSIRO.au
Location: Sydney

Funding
CSIRO offers a top-up stipend for this project: $300/week contribution towards living expenses.

BSEP10: Web-based framework for genetic risk prediction

The continuously decreasing cost of sequencing (NovaSeq US$100) and the landmark decision by the FDA to approve genetic testing as done by 23andMe, a commercial company, has paved the way for a second wave of direct-to-consumer genetic products. CSIRO hence aims to expand its portfolio in this space. Building on the team's LifeDNA framework, which predicts obesity risk from genomic profiles, the student will utilize Genome England's PanelApp [1] to build a generic testing framework that takes expert approved genetic loci associated with a specific trait and joins them in a genome wide risk score.

[1] https://panelapp.extge.co.uk/

The student will perform

- Build framework to read in a JSON object that holds the formula for combining loci
- Develops the API to interact with the PanelApp framework
- Design and implement the web service to display the result

Relevant field/s of study

- Python
- Web services, API
- Genomic/Genetic

Supervisor
Denis Bauer (CSIRO, Health&Biosecurity)
email Denis.Bauer@CSIRO.au
Location: Sydney, NSW

Funding
CSIRO offers a top-up stipend for this project: $300/week contribution towards living expenses.

CONTACT US

t 1300 363 400
+61 3 9545 2176
e enquiries@csiro.au
w www.csiro.au

AT CSIRO WE SHAPE THE FUTURE

We do this by using science to solve real issues. Our research makes a difference to industry, people and the planet.

As Australia's national science agency we've been pushing the edge of what's possible for over 85 years. Today we have more than 5,000 talented people working out of 50-plus centres in Australia and internationally. Our people work closely with industry and communities to leave a lasting legacy. Collectively, our innovation and excellence places us in the top ten applied research agencies in the world.

WE ASK, WE SEEK AND WE SOLVE

FOR FURTHER INFORMATION

Health and Biosecurity
Denis Bauer
t +61 2 9325 3174
e Denis.Bauer@csiro.au
w https://aehrc.com