
# 1st Student Symposium on Computational Biology and Life Sciences

October 8, 2014, Pontypridd, Wales, United Kingdom

# Welcome to the 1st Student Symposium on Computational Biology and Life Sciences

The ISCB Regional Student Group in the UK is pleased to welcome you to the 1st edition of the student sysmposium in Wales, UK. With a motto of "For the Students, By the Students", this event marks the inauguration of the Regional Student Group (RSG) of the International Society for Computational Biology (ISCB) in the United Kingdom of Great Britain and Northern Ireland. The general format and style of the symposium follows the pattern of symposiums of ISCB and its student council.

Our aim of this meeting is to give young and early-career researchers in the United Kingdom a platform to share ideas and network with senior and leading academics in diverse areas of computational biology. We are glad to report that, alongside doctoral and post-doctoral researchers, today we have participation from students from undergraduates and taught masters courses. We strongly believe that, today's event and the upcoming ones would help us to shape the pipeline of next generations of computational biologists.

We are honoured to have Dr. Natasha De Vere, Dr. Alex bateman, Dr. Christopher Creevey and Prof. Tatiana Tatarinova, to deliver keynotes today. Each of these speakers have contributed to the student and research community at various levels. We are also deilighted to have Dr. Manuel Corpas, a founder of the Student Council of ISCB, who will deliver a note on the roles of student groups in computational biology research. This booklet contains abstracts of the works presented along with short biographies of the keynote speakers.

Everyone involved in the organization of this Symposium contributed significantly to make this event happen. We are grateful for the support we have recieved from industry and academia in making this event happen.

We encourage you to make the most out of this opportunity and to be very active in engaging other delegates, asking questions, discussing ideas and showcasing your own research. You can make this Symposium a starting point for fruitful future collaborations and another step towards a successful career in computational biology.
Enjoy your time in Pontypridd!

**Farzana Rahman**, Symposium Chair
**Mehedi Hassan**, Founding Chair, ISCB RSG UK

# CONTENTS

# Key people

## Organising Committee

| | |
|---|---|
| **Farzana Rahman** (Chair) | – University of South Wales, UK |
| **Sayoni Das** | – University College London, UK |
| **Fatima Vayani** | – Imperial College, UK |
| **Rohit Farmer** | – University of Birmingham, UK |
| **Mehedi Hassan** | – University of South Wales, UK |
| **Prof. Denis Murphy** (Academic Advisor) | – University of South Wales, UK |

## Acknowledgements

The organisers would like to acknowledge help and guidance from the following members of academia and industrial partners from across the globe at various phases to make this event happen.

| | |
|---|---|
| **Dr. Timothy Ebbels** | – Imperial College, London, UK |
| **Dr. Eran Elhaik** | – University of Sheffield |
| **Dr. Tomas Di Domenico** | – University of Cambridge, UK |
| **Dr. Owain Kerton** | – University of South Wales |
| **Umesh Nandal** | – Academic Medical Centre,<br>– Universiteit van Amsterdam, Netherlands |
| **Chinmay Kumar Dwibedi** | – Ume University, Sweden |
| **Mehnaz Tabassum Khan** | – Khulna University, Bangladesh |

## Sponsors

The organisers would like to thank **Fujitsu Services UK Limited** and **HPC Wales Constorium** for their financial and logistics support to the event. We would also like to extend our thanks to the **Graduate Research Office** of the **University of South Wales** , the **International Society for Computational Biology's Student Council** for their continuous support.

# Academic Advisors Message

I very much welcome the opportunity to say some words about this first Student Symposium on Computational Biology and Life Sciences to be organized in the UK. The meeting is an important occasion for several reasons.

Firstly, it marks the inauguration of a new network of UK-based student researchers as part of the global work of the International Society for Computational Biology. We very much hope that this symposium will catalyze productive interactions between computational biologists in the UK and overseas, and especially the student members of this growing community who provide many of its most innovative workers as well as its future leaders. The second reason for the timeliness of this meeting is that we are on the cusp of a massive increase in the rollout of bioinformatics tools and applications that have the potential to transform many aspects of bioscience and its applications for the welfare of humankind. Computational Biology and its particular application in bioinformatics involves the use of high performance computational systems and novel mathematical tools to make sense of big data in fields related to bioscience.

Over the past decade, DNA sequencing throughput and costs have fallen at a spectacular rate that has been far more rapid even that the celebrated Moores Law of computation. The first human genome cost over $3 billion to sequence and was announced in 2000. In 2014, it is possible to sequence a human genome for $1-3,000, which represents a one million-fold fall in sequencing costs. We have now sequenced many animals as well as dozens of plants and hundreds of bacterial genomes and are still generating petabytes of new sequence data in publicly accessible databases such as GenBank.

One of the major challenges of 21st century biology is to make sense of this plethora of big data and to convert it into more manageable sets of information that can in turn be used to derive useful knowledge about the biological systems being studied. The two most important areas where Computational Biology is already making its presence felt are in biomedicine and agriculture. More than a decade after the first human genome was sequenced, the sequencing of part or all of each patients genome is now being trialed in various clinical settings. This is set to usher in an era of personalized medicine that has great potential to improve the quality of human life. Earlier in 2014 the genome of a major crop staple, wheat, was sequenced. This genome is three times larger than the human genome and contains three almost complete constituent genomes from the ancestors of modern breadwheat. The ability to sequence and analyze crop genomes is already contributing to efforts to increase both the quantity and quality of our food supply. This will be essential in the coming years as we are faced with increasing population pressures, resource depletion, and climate change that are threatening food security for huge areas of the world.

These are just two examples of the vital role that Computational Biology is set to play in the future. There is a global shortage of graduates and trained postgraduates in Computational Biology, and this is especially true in the UK. It is therefore very important that our growing community of researchers is able to make contact with each other, inform each other about our work, and facilitate future productive interactions and this is the real aim of any scientific meeting. I very much hope that this process can be started at the meeting and that it is then sustained by future student-led events of a similar nature.

**Professor Denis Murphy**,
Head of Genomics and Computational Biology Research Group,
University of South Wales

# Programme of the day

| | | |
|---|---|---|
| 08:30-09:00 | Registration | |
| **Session One** | | |
| 09:00 | Welcome Address | |
| 09:05 | RSG Introduction | Developing the Next Generation of Computational Biologists through RSGs by Dr. Manuel Corpas, The Genome Analysis Centre |
| **09:20-10:00** | **Keynote 1** | DNA Barcoding: creating an open access resource for people, wildlife & the environment by Dr. Natasha De Vere - National Botanic Gardens of Wales |
| 10:00-10:15 | | Coffee Break |
| **Session Two** | | |
| 10:15-10:35 | PostDoc Talk | Evolutionary pattern of the phosphoproteome in 18 yeast species by Romain Studer of EBI, UK |
| 10:40-10:55 | Student Talk | Dynamical analysis of diluted associative networks: a minimal model for the adaptive immune system by Silvia Bartolucci of Kings College London |
| 11:00-11:15 | Student Talk | Goldilocks: Locating genomic regions that are 'just right' by Sam Nichols from Aberystwyth University |
| **11:20-12:00** | **Keynote 2** | Biological Databases, not just stamp collection by Dr. Alex Bateman - Bateman Lab, EBI-Cambridge |
| 12:00-12:05 | | High Performance Computing Wales |
| 12:05-12:15 | | Industry Talk |
| 12:15-13:00 | | Lunch and Poster Setup/Session |
| **Session Three** | | |
| **13:05-13:45** | **Keynote 3** | Finding home using DNA , by Dr. Tatiana Tatarinova, University of Southern California |
| 13:50-14:05 | Student Talk | Metagenomic analysis of the Rumen microbiome reveals functional isoforms drive niche differentiation for nutrient acquisition and use by Francesco Rubino of Aberystwyth University |
| **14:05-15:00** | **Interactive Session** | Networking and Poster Session (with refreshments) |
| 15:05-15:25 | PostDoc Talk | Evolution of miRNAs and their targets among hominoid primates by Dr. Ranajit Das from University of Sheffield |
| 15:30-15:45 | Student Talk | Predicting tumour grade across multiple adenocarcinomas using exome sequence data, by Russel Sutherland of Kings College London |
| **15:50-16:30** | **Keynote 4** | Systems-level approaches to understanding microbial community interactions by - Dr. Chris Creevey |
| 16:35 | | Awards and Closing Remarks |

# RSG Introduction

## Developing the Next Generation of Computational Biologists through Regional Student Groups(RSGs)

by **Dr. Manuel Corpas**

Manuel Corpas (MC) was the founder and inaugural chair of International Society for Computational Biology (ISCB) Student Council. He has been involved in the Junior PI initiative for mid-career scientists in ISCB and elected Board of Directors for the Society since 2014. He joined The Genome Analysis Centre (TGAC, Norwich, UK) in 2012, working on the genomic analysis of plants and coordinating the BioJavaScript (BioJS) project. Since 2014 he is the technical coordinator of the ELIXIR-UK bioinformatics pan-European infrastructure. Previous to TGAC, he was a postdoc at the Wellcome Trust Sanger Institute, after completing his PhD in Computer Science at the University of Manchester, UK. Among other current duties, Corpas is 1) Chair of the Technical Committee for the Global Organisation for Bioinformatics Learning, Education and Training (GOBLET), 2) Chair of ISCB Africa Conference ( 150 attendees, every other year) and 3) BioVisNet Steering Committee, the recently BBSRC-funded UK biological visualisation network. Manuel has 1,557 Twitter followers and a blog with 117 published posts and >2,600 visits/month. He currently has 29 publications.

# Keynotes

## **Keynote 1** : DNA Barcoding: creating an open access resource for people, wildlife and the environment.

by **Dr. Natasha De Vere**

Dr Natasha de Vere is Head of Conservation and Research at the National Botanic Garden of Wales and a senior lecturer in plant ecology and conservation at the Institute of Biological, Environmental and Rural Sciences at Aberystwyth University.

Her research focuses on using genetic information to answer conservation questions. She led the team that made Wales the first nation in the world to DNA barcode all of their native flowering plants and she currently works on a wide range of DNA barcoding applications. She is dedicated to the public engagement of science with particular emphasis on art-science projects.

## **Keynote 2 :** Biological Databases, not just stamp collection

by **Dr. Alex Bateman**

Dr. Alex Bateman leads the the Protein sequence resources cluster that includes world leading databases including UniProt, Pfam, Rfam, TreeFam and MEROPS. As Head of Protein Sequence Resources at EMBL-EBI, Dr. Bateman also has an important strategic leadership role in the InterPro and UniProtKB databases. He has a long held interest in classification of protein and ncRNA sequences. Over the years he has published a large number of novel protein domains and families of particularly high interest.

Dr. Bateman obtained his PhD from Cambridge University in 1997 under the supervision of Cyrus Chothia and collaborated extensively with Sean Eddy and Alexey Murzin. He then moved the the Wellcome Trust Sanger Institute to lead the Pfam database. In 2012 he moved to the EMBL-European Bioinformatics Institute. Dr. Bateman was the Editor for the Nucleic Acids Research Database issue between 2003 and 2008, Executive Editor for Bioinformatics between 2004 and 2012 and is currently the Chairman of the International Society for Biocuration.

## **Keynote 3 :** Revisiting our ancestral history

### by **Prof. Tatiana Tatarinova**

Tatiana V. Tatarinova, PhD is a computational biologist with over 15 years of experience. She received her undergraduate degree in Theoretical Physics from the Moscow Engineering Physics Institute, earned her MSc in Physics from University of Utah, Salt Lake City, Utah, and a PhD in Applied Mathematics from the University of Southern California. For eight years, Tatiana worked at Ceres, Inc., a local biotech company, where she became the inventor of 15 U.S. and European patents. After leaving Ceres, she established and led the Glamorgan Computational Biology Research group at the University of South Wales for four years. In 2013 Tatiana joined CHLA as an Associate Professor of Research.

Research currently in progress include: Development of novel methods for analysis of bio-medical data, such as genome annotations, computational ancestry prediction for personalized medicine, bacterial toxicity, and cancer biomarkers, Prediction of methylation levels from other genomic features and Development of algorithms to support Barcode-of-Life initiative dedicated to supporting the development of DNA barcoding as a global standard for species identification

## **Keynote 4 :** Systems-level approaches to understanding microbial community interactions

### by **Dr. Christopher Creevey**

Dr Chris Creevey is currently a Reader in Rumen Systems Biology at the Institute of Biological, Environmental and Rural Sciences (IBERS) in Aberystwyth University in Wales. His main research interests involve identifying the genomic factors influencing phenotypic changes in organisms from Bacteria to Eukaryotes, with an emphasis on the development of approaches for understanding microbial communities. He received his Ph.D. in 2002 from the National University of Ireland for his work in the area of phylogenetics and comparative genomics.

Following this he worked as a postdoctoral researcher in NUI Maynooth and the European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany. In 2009 he was awarded a Science Foundation Ireland Stokes lecturership in Teagasc Ireland. He took up his current position in Aberystwyth University in 2013.

# Oral Talk Abstracts

## Early Career Researcher Talks

### 1. Evolutionary pattern of the phosphoproteome in 18 yeast species

**Romain Studer**, EMBL-EBI
**Co-Authors:** P. Beltrao (EMBL-EBI) and J. Villen (University of Washington)

Phosphorylation plays a critical role in the regulation of diverse cellular functions by modulating protein activities via different mechanisms. Phosphosites have been mainly investigated in model species (i.e. human, baker's yeast), but few phylogenetic analysis of phosphorylation have been done to date. To address this, we estimated the evolutionary history of phosphosites in yeast species and analysed them in the light of functional categories and effect on the phenotype.
An average of 5'000 phosphosites per proteome has been identified by mass spectrometry (MS) in 18 yeast species. We then developed a novel approach to infer phosphosite ancestral states based on the species phylogeny, the MS data and sequence information. This data was used to study the evolutionary history of regulation of biological processes and protein complexes. Finally, using mutational data, we observe that the age of phosphosites correlates with the functional importance of the regulated positions.

### 2. Evolution of miRNAs and their targets among hominoid primates

**Ranajit Das**, Sheffield University
**Co-Authors:** M. Jensen-Seaman

miRNAs are short ($\sim$22bp) RNA molecule that regulates protein-coding genes gene expression at post transcription level. From an evolutionary point of view, changes in miRNA regulation may cause several species-specific adaptations. The unique gains and losses of miRNAs and its potential implication within hominoids have remained understudied. The overall goal of this project was to identify uniquely gained and lost miRNAs within hominoids. miRNAs from human, chimpanzee, gorilla, orangutan and rhesus were collected from miRBase. Additional miRNA information was gathered from microRNAviewer. Reciprocal BLAST approach was employed to finalize the unique miRNAs. I found 14 miRNAs uniquely gained in humans. Maximum uniquely gained and lost miRNAs were found to be brain specific. The targets of uniquely gained miRNAs in human are also associated with brain-associated functions. Older miRNAs and their target sites were found to be more conserved compared to the newer miRNAs gained $<$15 Mya.

# Oral Talk Abstracts

## Student Talks

### 1. Predicting tumour grade across multiple adenocarcinomas using exome sequence data.

**Russel Sutherland**, Kings College London
**Co-Authors:** S. Diaz-Cano, J. Moorhead, R. J.Dobson

There is a need for tumour grading tools that are applicable across multiple cancer types in order to make tumour grading a more objective process. Most tumour grading systems are only applicable to a single cancer type. The aim of this study was to develop a tumour grading prediction model using tumour/normal exome sequence data as a tumour grading decision support tool for Pathologists.
We used exome sequence tumour/normal variant data and clinical data from the Pan Cancer Analysis from 970 patients with Kidney Renal Cell Carcinoma, Ovarian Carcinoma and Endometrial Carcinoma. We created a sample (S) by protein coding gene (P) "binary mutation matrix" that indicated the presence of any protein coding mutation for a sample (s) at a protein (p). Using the "binary mutation matrix" together with age, gender and cancer stage clinical variables we used multivariate logistic regression and Akaike Information Criterion (AIC) based backwards model selection to create classification models to predict the high grade or low grade status of tumours across cancer types and within cancer types. It was important to include the gender clinical variable to control for any gender bias introduced by the Endometrial and Ovarian Carcinomas. Classification accuracy, sensitivity and specificity were measured in an independent test set along with the area under the receiver operator characteristic (ROC) curves. Results: Across cancer types an AIC refined multivariate logistic regression model including 13 protein coding genes; TP53, MUC4, ANK2, CCDC171, CDH7, CTCF, CTNNB1, MYOF, NALCN, PARD3, PKD1L1, PTPRK and RAPGEF2 along with the cancer stage and age clinical variables was able to assign tumours to the correct grade status with an area under the curve (AUC) of 0.821. By comparison, the model using gender and cancer stage clinical variables refined from a model including clinical variables achieved an AUC of 0.756. The multivariate logistic regression model in Endometrial Carcinoma refined from protein coding genes and clinical variables age and stage included TP53, PTEN and cancer stage and performed with an AUC of 0.879 in comparison to a model including age and stage clinical variables that achieved an AUC of 0.753.
There are protein coding genes across cancer types that are predictive of tumour grade when adjusting for cancer stage. The genes ANK2, CDH7, and CTNNB1 are involved in tumour and stromal cell interactions important in tumour progression. MUC4 is involved in glandular differentiation, important for adenocarcinomas. Mutations in TP53 can be indicative of aggressive tumour types. The 13 protein features capture predictive information in addition to that provided by age gender and cancer stage clinical variables.

**2. Dynamical analysis of diluted associative networks: a minimal model for the adaptive immune system**

**Silvia Bartolucci**, King's College London
**Co-Authors:** A. Annibale

In this work we adopt a statistical mechanics approach to investigate basic, systemic features exhibited by adaptive immune systems. The lymphocyte network made by B-cells and T-cells is modeled by a bipartite spin-glass, where, following biological prescriptions, links connecting B-cells and T-cells are sparse. Interestingly, the dilution performed on links is shown to make the system able to orchestrate parallel strategies to fight several pathogens at the same time; this multitasking capability constitutes a remarkable, key property of immune systems as multiple antigens are always present within the host. In this work we solve the dynamics of these networks, evolving via sequential Glauber update. We derive dynamical equations for the order parameters, that quantify the ability of the system to coordinate simultaneous immune responses, and analyse the nature and stability of the stationary solutions by means of linear stability analysis as well as Monte Carlo simulations. We investigate the system's behaviour in different regions of the phase space, tuning the number of agents and link in the network. Interestingly, we will show that the parallel processing can be performed in either a symmetric or a hierarchical way, depending on the dilution, storage and noise in the system. This may be relevant for applications in theoretical immunology where an open question is how the immune system prioritises immune responses against different pathogens, executed in parallel.

**3. Goldilocks: Locating genomic regions that are 'just right'**

**Sam Nicholls**, Aberystwyth University
**Co-Authors:** A. Clare (Aberystwyth University) and J. Randall (Wellcome Trust Sanger Institute)

We present Goldilocks, a Python package which provides users with functionality for locating suitable regions within a genome for analysis. Goldilocks was developed to support our work in the investigation of quality control for genetic sequencing. It was used to quickly locate regions on the human genome that expressed a desired level of variability, which were 'just right' for later variant calling and comparison. To enhance Goldilocks, the package has been made more flexible and can be used to find regions of interest based on other criteria such as GC-content, defined confidence metrics and missing nucleotides. We give examples of how Goldilocks may be used in a bioinformatics pipeline.
Goldilocks is freely available open-source software hosted at :
https://github.com/SamStudio8/goldilocks

**4. Metagenomic analysis of the Rumen microbiome reveals functional isoforms drive niche differentiation for nutrient acquisition and use**

**Francesco Rubino**, Aberystwyth University
**Co-Authors:** F. Rubino, S. Waters, D. Kenny, C. Carberry, M. McCabe and C. J. Creevey

Metagenomics has provided insights into the rumen microbial community and revealed that many species seem to share the same genes for acquiring and utilising nutrients. This questions whether niche specialisation between rumen microbes exists and if so, it may not be driven by gene presence or absence, but by the diversity of functional isoforms related to specialisation. This hypothesis was investigated using rumen fluid from 14 cows, DNA was extracted and sequenced.

Following assembly, gene prediction and taxonomic assignment, SNPs were identified and functional isoform diversity (pN/pS) was calculated. Estimates of pN/pS in genes involved in carbon and sugar metabolism showed significant differences between Prevotella and Clostridia and those genes with more functional isoforms in Prevotella were involved in different metabolic routes than those in Clostridia, supporting the hypothesis that rumen microbes use functional isoforms for niche specialisation in nutrient acquisition.
**Website:** http://www.creeveylab.org

# Poster Abstracts

## 1. BarcAlign: Improved Pipeline for the Alignment of Plant DNA Barcode region

**Hannah Garbett**, University of South Wales
**Co-Authors:** T. V. Tatarinova, N. de Vere, D. J. Murphy

Plant DNA barcoding uses short DNA sequences to identify species from a specimen (i.e. roots, pollen, leaf, etc). In 2009, the Consortium of the Barcode of Life (CBOL), established a standard for plant barcoding projects. The standard consists of the chloroplast genes, matK and rbcL. To analysis the data from this process, accurate and reliable alignment tools are required. Simply aligning the matK barcode regions using standard alignment tools often creates inaccurate results because the tools often break codons as they align DNA sequences, which leads to inaccurate attributions in the functionality of the gene. Hence, it is imperative to ensure that codons remain intact throughout the data analysis process. We selected a range of open source alignment tools based on their popularity in the scientific literature and reviewed their performance. We identified one tool, transAlign, which did not break any codons in our analysis. Some tools have the capability to align translated sequences but the sequences must be in the same reading frame, for the process to be successful. However, this is not always possible as the frame for each sequence maybe different in same sequence file due to editing of the sequence. We therefore modified transAlign to develop an algorithm for more efficient multiple alignment. Maximum likelihood models were used to detect the best possible Open Reading Frame (ORF). The pipeline includes several different open source alignment tools (i.e. ClustalW, MAFFT, MUSCLE) as default options. Other alignment

## 2. BioJS: an open source standard for biological visualisation

**Benjamen White**, The Genome Analysis Centre
**Co-Authors:** S. Wilzbach, A. Thanki, G. Yachdav, R. Jimenez, S. J. Carbon, A. Garcia, L. Garcia, T. Goldberg, J. Gomez, A. Kalderimis, S. E Lewis, I. Mulvany, A. Pawlik, F. Rowland, F. Schreiber, I. Sillitoe, W. H. Spooner, J. M. Villaveces, H. Hermjakob, G. Salazar, M. Corpas

BioJS is a library of biologically-driven components that are easy to reuse, maintain and deploy on the web. This poster presentation will showcase some of the current features and developments of the new BioJS 2.0; the current version most of the development is now done on. I will explain the rationale and updates of the new release, together with the steps of the BioJS 2.0 development workflow (discover, learn, combine, use, create, test, maintain). In addition to this, I will include details of how BioJS will be used in the visualisation and analysis of Dioscorea alata (DA) infected with Colletotricum gloeosporoides (CG), the main agent of yam anthracnose (yam dieback), to investigate the effectors CG used to infect and invade DA.
**Website:** www.biojs.net/

## 3. A recognition model of ACP-HCS interaction for programmed beta-branching in type I polyketide synthases

**Rohit Farmer**, School of Biosciences, University of Birmingham
**Co-Authors:** A. Haines, M. P. Crump, C. M. Thomas and P. J. Winn

Polyketide synthases (PKSs) are enzyme complexes that synthesise a wide range of natural products of medicinal interest, notably a large number of antibiotics. Type I polyketide

synthases can introduce beta-carbon branches into a growing polyketide chain via enzymes encoded by the "HMG-CoA synthase (HCS) cassette". One of the first polyketide biosynthesis cluster in which the HCS cassette was discovered is responsible for the synthesis of the antibiotic mupirocin by Pseudomonas fluorescens. MupH is the HMG-CoA synthase homologue responsible for -branching in the mupirocin synthesis pathway. To understand better what allows the HCS cassette to recognise -branch-associated acyl carrier proteins (ACPs) of the mupirocin synthesis pathway, we have computationally docked the modelled MupH with the NMR structure of ACPs. The docking results were also supported by the evolutionary trace data and the physical properties of the interface residues. Hidden Markov models (HMM) were used to classify ACPs as branching and non-branching. HMM analysis highlighted essential features for an ACP to behave like a branching ACP. Through modelling and mutagenesis we identified helix III of the ACP as a probable anchor point of the ACPHCS complex. The position of this helix is determined by the core of the ACP and substituting the interface residues modulates the interaction specificity. Our method for predicting -carbon branching lays a basis for determining the rules for ACP-HCS specificity and expands the potential for engineering new polyketides.

## 4. Towards the increase of the thermostability of a psychrophilic enzyme

**Stefani Dritsa**, Aberystwyth University

Lipases are enzymes with multiple industrial applications, from food processing, pulp, paper and oleochemistry to diagnostic tools, biosensors, cosmetics and biodiesel production. Their appeal to the industrial world can easily be comprehended, as lipases exhibit lucrative characteristics: activity, specificity, versatile and inexpensive catalysis, easy production. One of the qualities of lipases that interests industry in improving is the thermostability. High temperatures are used in most industrial processes, thus, the target of this project is to make one widely used psychrophilic lipase more thermostable through computational methods. 7 models were produced following an extensive study of prior art, a sequence and a structural analysis. Molecular dynamics (MD) simulations was the method chosen to study the enzyme's reaction to increasing temperature, and the 7 models have now been tested in 5 different temperatures (303K, 323K, 343K, 363K, 383K) using the GROMACS software. The radius of gyration (Rg), the root mean square deviation (rmsd) and root mean square fluctuation (rmsf) of C and backbone, the stability of the active site (the hydrophobic and total energy, rmsf of C) were measured and plotted. We are, now, performing the analysis of the data and looking into choosing the models with the best behaviour to be the input to a sequence optimization with Rosetta Design software.

## 5. Exploiting CATH-Gene3D functional families to suggest multifunctionality for moonlighting proteins

**Sayoni Das**, Institute of Structural and Molecular Biology, University College London
**Co-Authors:** D. Lee and C.A. Orengo

One of the main concerns faced by computational approaches for protein function prediction protocols is the functional diversity of moonlighting proteins which presents new challenges to existing function prediction methods.

We investigated the performance of a domain-based protein function prediction protocol in suggesting multifunctionality of proteins that exploits functional subclassification of protein domain superfamilies in CATH-Gene3D. Sub-classification of the protein domain superfamilies into functional families was done using a hierarchical agglomerative clustering algorithm supervised by a family identification protocol, FunFHMMer, that recognises highly conserved positions and specificity-determining positions in cluster alignments and uses this information to ensure functional coherence. The functional families were further associated with Gene Ontology terms proba-

bilistically, in order to predict functions for uncharacterised sequences.

The function annotations provided by FunFHMMer families for 144 moonlighting proteins from the moonlighting proteins database, MultitaskProtDB, are shown to be more precise than annotations provided by PSI-BLAST. Additionally, the highly conserved positions of these functional families can provide valuable information regarding specific sites or motifs which may be responsible for moonlighting activity of proteins.

### 6. A Molecular Dynamic Web Server for Analysis of DNA Structure

**Daniele Avancini**, Swansea University **Co-Authors:** G. Menzies , A. Brancale and P. Lewis

Cancer is driven by mutation and cataloguing spectra of mutations in different cancer types may help to develop new diagnostics and treatment regimes for personalized medicine. Molecular Dynamic (MD) approaches to analysis of chemical interactions with DNA are advanced and can provide one way to correlate mutagen exposure with mutation, DNA repair and disease. Using high performance computing (HPC Wales), our group has developed a pipeline to evaluate the mutagenic potential of a compound when it is bound to DNA by measuring 3D DNA structural parameters. The pipeline utilises GROMACS software to perform the MD simulations which are subsequently analyzed using CURVES+ a tool to calculate helicoidal parameters. Resulting data is then analysed using the R statistical environment. The entire process is time-consuming and limited to expert use. Therefore, the aim of this project is the development of a web server application housing a fully automated user-friendly version of this pipeline.

### 7. The possible use of Molecular Dynamics (MD) to under standing of the alterations of mucin structure in Cystic Fibrosis (CF)

**Georgina Menzies**, Swansea University
**Co-Authors:** P. Lewis

Cystic Fibrosis (CF) is a genetic disorder, where viscous mucus is hyper secreted. The production of this viscous mucus, impedes mucus clearance and promotes stasis, thus promoting infection. Mucins are a large component of mucus secreted from CF patients lungs and understanding how these are altered can lead to a greater understand of how to treat this disease. Mucins are large glycoproteins and in normal respiratory tract the predominant mucins secretions are MUC5AC and MUC5B. These structures are heavily glycosylated, which assists in the prevention of cellular desiccation, accumulation of foreign particles and prevent pathogen adhesion and invasion. Typical changes to these mucins in disease states involve addition of Lewis antigens to glycan cores and sialylation of the antigen. Modelling techniques were employed to both predict the consequences of and visualise the changes to MUC5AC when these sugar cores are altered and preliminary results show changes to the mucin structure when cores are altered.

### 8. MGkit: An Evolutionary Metagenomic Framework For The Study Of Microbial Communities

**Francesco Rubino**, Aberystwyth University **Co-Authors:** S. Waters, D. Kenny, C. Carberry, M. McCabe and C. J. Creevey

While metagenomics has been used extensively to study microbial communities from a taxonomic and functional perspective, little has been done to address how the species in a microbiome are adapted to and maintain specific roles in dynamic environments like the rumen. Identifying and assessing the level of this biological robustness for function is an important aspect that has not been addressed by any currently available metagenomic pipeline. To address this problem we have developed a

framework for the robust analysis of metagenomic data that includes automated analysis from assembly to, gene-predicition and taxonomic identification. Furthermore we implement approaches to estimate SNP diversity in metagenomic samples and carry out statistical tests to identify where differences exist. The framework allows easy customisation of any metagenomic workflow and platform independent. Automatically generated assembly and SNP-based statistics provide easy interpretation of the results.
**Website:** https://bitbucket.org/setsuna80/mgkit

### 9. Concatabominations: identifying unstable taxa in morphological and phylogenomic supertrees using Safe Taxonomic Reduction

**Karen Siu Ting**, University of Bristol **Co-Authors:** K. S. Ting, D. Pisani, C. J. Creevey, M. Wilkinson

Rogue taxa can be phylogenetically unstable because of limited and extensive missing data, leading to multiple trees, unresolved consensus topologies and an increase in run times. Concatabominations was developed as a heuristic extension to the Safe Taxonomic Reduction (STR) method that identifies unstable taxa. Our method is based on the experimental merging of data from pairs of "potential taxonomic equivalents" to create new "concatabominated" taxa (conceptually equivalent to forcing taxa together in a tree). Taxa that can be combined with many others without introducing additional homoplasy to the data (tested by means of compatibility) are candidate rogue taxa. Our pipeline implementation allows visualisation of taxonomic equivalence relations as a network. We assessed performance using gap-rich paleontological and genomic MRP datasets. The approach is potentially significant for identifying and thus for addressing cases of ineffective taxonomic overlap in phylogenomic datasets.
**Website:**
http://code.google.com/p/concatabominations/

### 10. High performance computing for comparative genomics

**Vasileios Panagiotis Lenis**, Aberystwyth, University

The advent of the Next Generation Sequencing technologies (NGS) and de novo assembly algorithms has enabled the reconstruction of several genomes at low cost. The comparison of these genomes remains a difficult problem necessitating a lot of computational resources. We aim to address this challenge by applying high performance computing technology on existing alignment algorithms. In order to automate the whole genome alignment process we have generated an alignment pipeline using several existing components in a parallel way. In order to optimize it farther, we are trying to identify Highly Conserved Elements (HCE) among the genomes which we then use as anchors. We predicted the HCE among twenty birds, and we found that there exist around 3 million of HCE bigger than 10bp. Finally, we aim to use the results of this work to produce an improved and refined version of the sheep genome and generate a Single Nucleotide Polymorphism (SNP) resource for Welsh native breeds.

### 11. Deterministic versus stochastic agent based models

**Elizabeth Donkin**, Aberystwyth University
**Co-Authors:** J. Warren

A major challenge of agent based modelling (ABM) of ecological systems is including enough complexity within the model to accurately simulate real-life organisms and their behaviour. Some ABMs are fully deterministic, the behaviour of agents and the environment within the model have no random processes. Others employ the use of stochastic processes, those that are based on random numbers and differ at every replication of the model. There is a balance to be met between the two. In natural populations stochasticity of the environment can influence behaviour and survival, however too many stochastic environmental inputs can make it difficult to analyse the more interesting variability that occurs from the underlying de-

terministic and adaptive behaviour of the modelled organisms. This project aims to develop agent based models to simulate invertebrate herbivore movement in plant populations varying in the level of genetic diversity of defensive traits. A significant part of this project will involve determining which, if any, stochastic variables to include to accurately recreate herbivore movement recorded in trials using live model plants and insects.

### 12. Micropropagation technique and HPLC bioactive chemical characterization of chamomile and yarrow species (Asteraceae)

**Banaz Mahmood**, Plymouth University

Asteraceae is the most commonly used plant species in medicine worldwide. Chamomile and Yarrow are common plant species belonging to this family. Phenol and flavonoid are the most numerous compounds found in these plants. The use of in vitro micropropagation technique is therefore essential to increase these compounds via plant tissue culture using different PGRs. Subsequently, the quantitative and qualitative testing is also an important step in the identification of these compounds in plant material using methanolic extraction and HPLC analysis. The maximum occurrence of chamomile callus (85.8%), shoots (93.3%) and roots (75.3%) were observed on MS medium with 2, 4-D (1mg/ml), IAA (0.5mg/ml) and NAA (0.5mg/ml), however, the maximum rate of yarrow shoots (77.8%) and roots (82.3%) were detected with GA3 (1mg/ml) and BAP (1mg/ml) respectively. The major phenols and flavonoids of in vitro vegetative Chamomile and Yarrow were also investigated.

### 13. Monitoring seed dispersal with tracking devices

**Csaba Sarosi**, University of South Wales

The seed dispersal is the transport of the seeds away from the parent plant by abiotic and/or biotic factors. To study seed dispersal, monitoring the transport of the seeds is necessary. The purpose of this presentation is to suggest ideas how this transport can be tracked in various plants. The ideas based on two different devices, which were created to monitor the movement of ants in order to track seed dispersal. Ants are small and fast moving; these characteristics make them very similar to seeds. The first device was the software: AntTracks, which was made to video analysis and can help to study seed release. The second one was the radio tracking devices. In the case of monitoring seed dispersal, this technology may be useful to fight against invasive plants; protect endangered species; discover mutualistic relationships; check water currents; support specific studies and model seeds.

# Awards

Following awards sponsored by the University of South Wales, will be given at the end of today's event.

(1) **Best Oral Presentation**
(2) **Best Poster Presentation**
(3) **Poster Presentation Runner-Up**

Sponsored by ***Faculty of 1000 Limited*** - F1000.com, there will be additional 3 awards each comprising of a certificate and article processing fees waiver for publication in F1000's peer reviewed journal.

Above awards will be selected based on review by academics and general attendees. You are welcome to contribute.

Further to this, the **HPC Wales constoritum** will select suitable collaboration project(s) between a Welsh entity and any other research group/institute. Each project will be awarded a fees waiver for use of upto 5000 core hours of free computing resources the HPCWales grid.

# Travel Fellowship

Two students are selected for the travel fellowship to attend this event. They are:

(1) **Silvia Bartolucci** from King's College London, UK
(2) **Lenis Vasilis** from Aberystwyth University, Wales, UK