

## PhD Programme in Simulation Science

The Structured PhD Programme in Simulation Science is a new multi-institutional collaborative Ph.D. programme involving University College Dublin, Trinity College Dublin, Queen's University Belfast, National University of Ireland Galway and is supported by the Irish Centre for High End Computing.

Funded PhD Fellowships are currently available at NUI Galway in this programme. These SimSci Fellowships are fully funded for 4 years and include a stipend of 16,000 Euro per year together with an allowance for research travel and expenses and cover fees for EU students.

The SimSci Fellowships are funded under the Programme for Research in Third Level Institutions (PRTL) Cycle 5 which is co-funded by the European Regional Development Fund (ERDF).

Applications are now being accepted for these Fellowships. Please send application, including full CV and names of referees by email to Ms Triona Quinn ([triona.quinn@nuigalway.ie](mailto:triona.quinn@nuigalway.ie)). Applications will be accepted up to 5 pm (GMT) on Friday September 16, 2011.

Please contact me with any queries.



**Triona Quinn | Administrative Assistant | NCBES National Centre for Biomedical Engineering Science | NUI Galway**

**E-mail: [Triona.quinn@nuigalway.ie](mailto:Triona.quinn@nuigalway.ie) | Web: [www.ncbes.ie](http://www.ncbes.ie) | Direct Line: 091495165**

# Project SSG -001: A Semantic Laboratory Information Management System for Stem Cell Research

*Supervisors: Prof. Frank Barry and Dr. Helena F. Deus (NUIG)*



One of the most pressing needs in understanding stem cell differentiation and accelerating its application to improving health care is the ability to reuse and integrate experimental results with existing biomedical data sources. This project will focus on the development of a semantic web based computational methodology for the integration of experimental results and public genomics, proteomics and metabolomics datasets.

## Description

Most modern laboratories engaged in stem cell research rely on several methodologies for data collection and correlation. More often than not, experimental results are stored in different, proprietary systems which complicate its integration for the design of comprehensive biological models. Even when there are attempts to create Laboratory Information Management Systems (LIMS) as integration tools, the high heterogeneity and frequent update of experimental data challenge automated integration. The state of the art in the development of LIMS has relied on relational or object databases with fixed schemas [1]. Such systems, however, rarely enable the changes to the data model which would be required to support capturing and recording novel experimental variables [2].

Efforts to standardize “omics” databases have resulted in document models such as e.g. the Minimum Information About a Microarray Experiment (MIAME). Such standards facilitate reusing genomics data across laboratories and experiments but they have become victims of their own success – the challenge of reusing experimental data has been complicated by the existence of too many standards to choose from [3]. Semantic web and Linked Data technologies can provide a solution for this problem. By relying on a high level abstraction for representing data, it becomes possible to cross-reference and disambiguate the multitude of standards and data models to represent, e.g. proteomics data [4].

This project will rely on devising and applying a solution to solve the data integration problem in Stem Cell research through the development of a Semantic Laboratory Information Management System (SLIMS). The methodologies used will rely on semantic web and linked data technologies to integrate and align proprietary data sources (where experimental results are collected) with public proteomics, genomics and metabolomics data sources available on the Web.

## References

- [1] Troshin,P.V. et al. (2008) Laboratory information management system for membrane protein structure initiative--from gene to crystal. *Molecular Membrane Biology*, 25, 639-652.
- [2] Almeida,J.S. et al. (2006) Data integration gets "Sloppy". *Nature biotechnology*, 24, 1070-1.
- [3] Quackenbush,J. (2006) Standardizing the standards. *Molecular Systems Biology*, 2, 2006.0010.
- [4] Wang,X. et al. (2005) From XML to RDF: how semantic web technologies will change the design of "omic" standards. *Nature biotechnology*, 23, 1099-103.

## Project SSG -002: The Autoimmune Potential of Transcriptional Noise

Supervisors: Prof. Cathal Seoighe and Prof. Rhodri Ceredig (NUIG)

### Description

The mammalian adaptive immune response is orchestrated by populations of T and B lymphocytes with surface receptors that bind foreign antigens. Lymphocytes are generated in large numbers in the primary lymphoid organs, namely the thymus for T cells) and the bone marrow for B cells. Each individual lymphocyte expresses a distinct receptor capable of recognising a particular antigenic determinant. The large pool of lymphocytes expressing a diverse receptor repertoire means that antigens from most pathogens will be recognised. Given that the generation of the lymphocyte receptor repertoire is essentially stochastic, the problem arises as to how to ensure that the receptors generated do not recognise peptides derived from the animal's own proteins. At an immature stage in their development in the primary lymphoid organs, engagement of a lymphocyte's receptor with a self antigen results in its elimination. It was unclear how exposure of developing lymphocytes to non-thymic self protein antigens was achieved but recent work has shown that genes encoding such proteins can be expressed by some cells in the thymus, thus ensuring that developing T cells are exposed to self antigens whose expression may ordinarily be restricted to specific tissue types. This is achieved, in part, through the expression in the thymus of the gene Autoimmune Regulator (AIRE), identified as the gene mutated in a group of patients with severe auto-immunity and which, when knocked out in mice, results in organ-specific autoimmunity. A recent search for proteins bound to Aire identified a set of proteins involved in the splicing of primary transcripts (Mathis and Benoist 2009) and these authors reported that AIRE influenced spliced forms of transcripts more than their un-spliced precursors. The fact that transcripts encoding tissue-restricted antigen (TRA) proteins can be differentially spliced increases the possible repertoire of molecules the thymus must express in order to efficiently purge the system of auto-reactive T cells. Thus, it is of considerable interest to analyze the transcriptome of thymus cells in greater detail. In particular, this project aims to determine the extent to which tissue-specific alternative splicing gives rise to TRAs and whether there are mechanisms to prevent autoimmune responses against these TRAs. We will investigate the expression of tissue-restricted splice isoforms in thymus using existing data in the public domain and by applying RNA-seq to mouse thymic epithelial cells. This will be investigated further by testing, experimentally, whether a high rate of mis-splicing can result in autoimmunity in the mouse.

This research project will examine the links between transcriptional noise, mis-splicing and autoimmunity. We will use simulation together with mathematical or probabilistic modelling to understand the relationship between the expression of cryptic epitopes resulting from errors in the regulation of transcription or mRNA processing and the likelihood of stimulating an autoimmune response. Given the high proportion of human chronic diseases associated with autoimmunity and recent research indicating the importance of stochastic noise, both at the level of RNA transcription and splicing, the results of this investigation will have the potential for significant basic scientific and medical impact.

# Project SSG -003: The Application of Graphical Processing Units in Simulating Biological Systems

Supervisor : Dr. Andy Shearer (NUIG)

## Description

Graphical Processing Units (GPUs) represent the next stage in the development of high-performance computing. Over the past twenty years the powerful machines have moved from being dominated by vector processors, through clusters of processors, to tightly coupled clusters and NUMA architectures and finally to dedicated architectures like IBM's BlueGene. In the biological sciences we have some of the most computationally intensive tasks available. (Meredith et al 2007, ). In many cases GPUs represent a solution to accessing high performance computers at a relatively low cost, particularly for computational biology (see e.g. Shkutri et al, 2010). In NUI Galway we have hybrid system currently consisting of 4 TESLA GPU cards on Bull Novascale 512 core computer. On this platform we want to investigate the applicability of a number of computational biology platforms ranging from molecular dynamics problems (see e.g. Hampton et al, 2010) to medical imaging problems. Our work is intended to design and implement specific GPU libraries to streamline the development of biological and medical simulations. Our own work in this area ranges from putting plasma simulations onto GPUs for Astrophysical simulations (Rochford, et al, 2011, in preparation) development of Monte-Carlo simulations for radio-therapy dose determination (Downes et al, 2009), developing hybrid Matlab/MPI code for finite element analysis (Browne et al, 2006).

As well as performing detailed numerical simulations GPUs will also play an important role in rendering and displaying the output of complex models. We have been working on systems to visualise N dimensional models both as an aid to understanding the computational science and as means to undertake computational steering.

## References

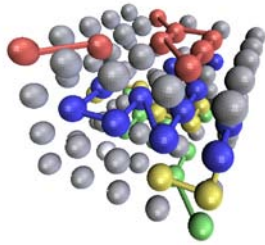
- Browne, M., Enmark, A., Andersen, T., and Shearer, A. "Optimised external computation for the Euro50 MATLAB based integrated model," *SPIE, Volume 6271, pp. 627103 (2006).*, 6271, (2006)
- Georgios Chalkidis, Masao Nagasaki, Satoru Miyano, "High Performance Hybrid Functional Petri Net Simulations of Biological Pathway Models on CUDA," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 24 Nov. 2010. IEEE computer Society Digital Library. IEEE Computer Society
- Downes P, Curran O, Cunniffe J & Shearer, A, "Distributed Radiotherapy Simulation with the Webcom Work Flow System ", *International Journal of High Performance Computing Applications*, 24 Issue: 2, 213-227 (2010)
- Hampton, Scott; Agarwal, Pratul K.; Alam, Sadaf R.; Crozier, Paul S.; , "Towards microsecond biological molecular dynamics simulations on hybrid processors," *High Performance Computing and Simulation (HPCS), 2010 International Conference on* , pp.98-107, June 28 2010-July 2 2010
- Meredith, J.S.; Alam, S.R.; Vetter, J.S.; , "Analysis of a Computational Biology Simulation Technique on Emerging Processing Architectures," *Parallel and Distributed Processing Symposium, 2007. IPDPS 2007. IEEE International* , vol., no., pp.1-8, 26-30 March 2007

O'Shea TP, Foley MJ, Rajasekar D, Downes, P. A., van der Putten, Moore, M. & Shearer, A, Electron beam therapy at extended source-to-surface distance: a Monte Carlo investigation", *Journal of Applied Clinical Medical Physics*, 9 Issue: 4, 57 (2008)

Shkurti, A.; Acquaviva, A.; Ficarra, E.; Macii, E.; Orsi, M.; Ruggiero, M.; , "GPU acceleration of simulation tool for lipid-bilayers," *Bioinformatics and Biomedicine Workshops (BIBMW), 2010 IEEE International Conference on* , vol., no., pp.849-850, 18-18 Dec. 2010

## Project SSG -004: Understating Cell Signalling through Linked Data

Supervisors: Prof. Frank Barry and Dr. Helena F. Deus (NUIG), Prof. Walter Kolch and Prof Boris Kholodenko (UCD)



Understanding and predicting protein-protein interactions (PPI) can have a direct effect on our ability to therapeutically target the signalling networks that ultimately regulate carcinogenic processes. This project will focus on making use of Linked Data technologies to create multiple, non-overlapping layers of information that can be used as inputs for devising encompassing PPI prediction models.

### Description

Many cell processes such as proliferation and differentiation are controlled by signalling cascades, i.e. chains of proteins responsible for communicating signals from the surface of the cell to the nucleus, effectively affecting protein transcription. One of the most important signalling cascades in carcinogenesis is the ERK/MAPK pathway - it is believed that mutations in the genes responsible for the proteins involved in this pathway may lead normal cells to become cancer cells. Recent research has revealed that such signal transduction pathways appear to be organised as communication networks where information is processed and integrated through relay stations formed by multi-protein complexes [1]. Identifying the proteins involved in these signalling cascades and understanding how they interact to produce a chain of events is therefore a crucial step towards our ability to devise drugs that restore normal activity in the cell.

Computational simulation methods have become a popular method for predicting potential protein-protein interactions based on 3D protein docking, domain-domain interactions or the co-evolution model. The accuracy and predictive power of such computer models relies heavily on the amount and quality of integrated information used as input [2]. The current state of the art in devising such models relies on *ad hoc* integration of the relevant information e.g. sequence and structure information, to build a useful predictive model. Every additional layer of information must be extracted, transformed and integrated separately before it can be used as input. Alternatively, Linked Data can be used as an integrative technology as it relies on the simple concept that existing relationships between entities, such as proteins, can be represented as a network where each individual entity is represented by a node and its relationships to other entities in the graph, e.g. drugs or other proteins, are represented by an arc. Moreover, both the entities and the links established between them can be deferred, i.e. their description and associated properties can be automatically retrieved from the Web to be used in the creation of new layers of integrated information. Multiple studies have shown that these technologies are suitable for integrating proteomics and genomics experimental results [3-5].

In this project, Linked Data technologies will be weaved to represent protein-protein interactions. The research focus will be on identifying the type of relationships that are best used to represent both the provenance of the interaction information (e.g. mass spectrometry, co-upregulation, etc) and its probabilistic value in order to create non-overlapping layers of information. Representing protein-protein interaction data in this format will enable the creation of mathematical constructs, e.g. adjacency matrixes that can be algebraically manipulated to identify the topology of the protein-protein interaction network. The advance beyond the state of the art will be the possibility to enrich

the predictive models with *ad hoc* layers of information such as drug interactions and its effect on the network topology.

## References

1. Kolch W: Coordinating ERK/MAPK signalling through scaffolds and inhibitors. *Nature Reviews Molecular Cell Biology* 2005, 6:827-837.
2. Wierling C, Herwig R, Lehrach H: Resources, standards and tools for systems biology. *Briefings in functional genomics proteomics* 2007, 6:240-251.
3. Anwar N, Hunt E: Francisella tularensis novicida proteomic and transcriptomic data integration and annotation based on semantic web technologies. *BMC Bioinformatics* 2009, 10:S3.
4. Deus HF, Prud E, Zhao J, Marshall MS, Samwald M: Provenance of Microarray Experiments for a Better Understanding of Experiment Results. In *ISWC 2010 SWPM*. 2010.
5. Deus HF, Veiga DF, Freire PR, et al. Exposing The Cancer Genome Atlas as a SPARQL endpoint. *Journal of Biomedical Informatics* 2010, 43:998-1008.

# Project SSG -005: Glycomic studies of Endothelial Progenitor Cells and Osteopontin

*Supervisors: Prof. Lokesh Joshi and Prof. Tim O'Brien*

Glycans cover every cell including progenitor cells and play crucial roles in modulating cellular response to the environment. This project will combine in-silico and experimental studies to investigate the role of glycosylation in EPC and associated cells.

## Description

Glycoconjugates (glycolipids, glycoproteins and proteoglycans) are found in abundance on all cell surfaces in nature, forming a dense array of glycans (known as the glycocalyx). Glycoconjugates mediate cell recognition and facilitate cellular interactions during numerous physiological and pathological processes, including cell adhesion, migration and differentiation, inflammation, autoimmune diseases and oncogenesis. The glycosylation pattern of a cell can change rapidly depending on the physiological states, stresses of cell-cell interaction, local environment and any disease state.

Endothelial Progenitor Cells (EPCs) contribute to angiogenesis and are able to differentiate into endothelial cells of the vasculature. In cardiovascular disease and diabetes EPCs are downregulated and often dysfunctional. Thus, understanding EPC function is key to understanding vascular diseases. It has also been shown that a heavily phosphorylated and glycosylated protein on EPCs, osteopontin (also called as bone-sialoprotein), plays a role in angiogenesis. However, the role of the osteopontin-EPC interaction in angiogenesis has yet to be established. Further, a population of cells related to EPCs called Circulating Angiogenic Cells (CACs) are also dysregulated in vascular disease. These cells confer an angiogenic effect through paracrine mechanisms where osteopontin is also believed to be involved.

There is considerable interest in understanding the mechanisms of EPC and CAC - mediated angiogenesis and the use of EPC/CAC to repair damaged blood vessels due to myocardial infarction and critical limb ischemia. However, little is known about the glycocalyx and its role in EPC/CAC biology. Cell surface is vital for intercellular and cell-environment communication and therefore, studying the glycocalyx and its components (specially osteopontin) may shed light on EPC/CAC biology.

The research project will investigate the glycomics of EPC/CAC using both in-silico and experimental tools. Transcriptomic data will be used to explore the glyco gene regulation and connect these

changes to other pathways using network biology methods. These studies will lead to candidate molecules that effect the remodeling of glycocalyx which will be validated by experimental studies. This project will also examine the effect of designed osteopontin on EPC/CAC biology to determine the molecular aspects of osteopontin involved in regulating the cellular behavior. The results of this project will open new vistas of EPC and CAC biology and their potential clinical application in cardiovascular and metabolic diseases.

