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.

Two funded PhD Fellowships are currently available at NUI Galway in this programme for collaborative projects between the Digital Enterprise Research Instutite (DERI), Systems Biology Ireland (SBI) Institute and the Regenerative Medicine Institute (REMEDI) . 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 (PRTLI) 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 Dr. Helena F. Deus (helena.deus@deri.org) . Applications will be accepted up to **5 pm (GMT) on Monday October 31, 2011.** Please also indicate to which of the projects you are applying to.

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.

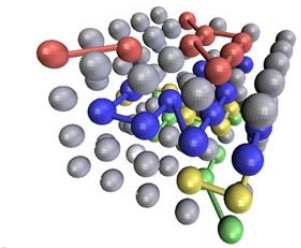
[2] Almeida,J.S. et al. (2006) Data integration gets "Sloppy". *Nature biotechnology*, 24, 1070-1.

[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 -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 deferenced, 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.

Ireland's EU Structural Funds Programmes 2007 - 2013

Co-funded by the Irish Government and the European Union

An Roinn Fiontar, Trádála agus Nuálaíochta

Department of Enterprise, Trade and Innovation

HEA

Higher Education Authority

An tÚdarás um Ard-Oideachas