Technical Report CoSBi 05/2006

Book of Abstracts

Scientific Opening
of
The Microsoft Reasearch Centre for Computational and Systems Biology
Trento, Italy

April 3-5, 2006

Jane Hillston – University of Edinburgh
Title: Systems Biology Activity at Edinburgh

ABSTRACT
In this talk I will give an overview of some of the Systems Biology activity currently being undertaken at Edinburgh. Within the Centre for Systems Biology at Edinburgh we are undertaking a programme of work focussed on issues surrounding modelling the dynamics of biological processes based on all available data. As part of this programme we seek to develop the first integrated, publicly-accessible informatics infrastructure for systems biology, covering all stages of systems-level research. We also aim to develop tools from computer science, based on process calculi, to represent, compare and check biological models.

Pietro Liò – University of Cambridge
Title: Modelling the immune system: quasi species dynamics in HIV infection

ABSTRACT
During the HIV infection several quasispecies of the virus arise, which are able to use different coreceptors, in particular the CCR5 and CXCR4 coreceptors (R5 and X4 phenotypes, respectively). The switch in coreceptor usage has been correlated with a faster progression of the disease to the AIDS phase. As several pharmaceutical companies are starting large phase III trials for R5 and X4 drugs, models are needed to predict the co-evolutionary and competitive dynamics of virus strains. I present a model of HIV early infection which describes the dynamics of R5 quasispecies and a model of HIV late infection which describes the R5 to X4 switch. I report the following findings: after superinfection or coinfection, quasispecies dynamics has time scales of several months and becomes even slower at low number of CD4+ T cells. The progressive loss of CD4+ T cells can be described taking into account the X4-related Tumor Necrosis Factor dynamics. I also hypothesise that virus mutational pathway may generate R5 variants able to interact with chemokine receptors different from CXCR4. This may explain the massive signalling disruptions in the immune system observed during AIDS late stages and may have relevance for vaccination and therapy.

F. Amato¤, M. Bansal *y*, C. Cosentino¤, W. Curatola¤, D. di Bernardo*y*
Title: Piecewise-Affine Dynamical Model of the Cell Cycle Regulatory Network in Fission Yeast

ABSTRACT
The intrinsically complex nature of biological systems derives from the coexistence of continuous dynamics and discrete events in the underlying regulation mechanisms. Several approaches can be found in literature to the problems of modelling and identification of biological systems, focusing either on the logical evolutionary rules (formal languages, Petri nets, membrane computing) or on the dynamical behavior of averaged state variables (differential/difference equation systems, stochastic models). The approach pursued in this work, aims at the integration of the two perspectives, addressing both of them by means of a class of hybrid model, namely the class of Piecewise-Affine (PWA) systems. With reference to the case-study of cell cycle regulatory network of fission yeast, a method is presented to identifying both continuous evolution of the expression levels of genes, and the switching points in the cycle, corresponding to the occurrence of logical conditions. The identified model is first assessed by comparison with an in silico model, then it is exploited for the network reconstruction starting by in vitro data available in literature.

¤F. Amato, C. Cosentino and W. Curatola are with the School of Computer and Biomedical Engineering, Universita' degli Studi Magna Grecia di Catanzaro, Via T. Campanella 115, 88100 Catanzaro, Italy (amato,carlo.cosentino,walter.curatola)@unicz.it
yM. Bansal and D. di Bernardo are with the Telethon Institute of Genetics and Medicine, via P. Castellino 111, 80131 Napoli, Italy (bansal,dibernardo)@tigem.it

Cesare Furlanello - ITC-irst Trento
Title: Predictive profiling for high-throughput functional genomics

ABSTRACT
Class prediction and feature selection are two learning tasks that are strictly paired in the search for molecular profiles. It is however easy to incur a selection bias effect when dealing with high-throughput molecular data. Complex validation setups and computing resources are thus required to avoid overly optimistic estimates of accuracy on novel data and the incorrect selection of biomarkers. In this talk, I will outline methodology and computational solutions implemented in the BioDCV system, a complete tool for high-throughput data analysis. With examples of recent tasks on microarray and proteomics data, I will discuss biomarker selection and stability, discovery of outliers and of potential subtypes, treatment of time-varying patterns. (Joint work with G. Jurman, S. Merler, A. Barla, S. Paoli, D. Albanese, B. Irler, R. Flor. Website: http://biodcv.itc.it)

Enrico Blanzieri – University of Trento
Title: Data quality and a priori information in microarray gene expression analysis

ABSTRACT
The study of gene expression with microarrays are a mainstream high throughput technique for the analysis of transcriptome. Data quality can be assessed exploiting information coming from different sources. We present examples of automatic or semiautomatic techniques for assessing data quality using information on the experimental design, continuity in time, consistency of the statistical model and Gene Ontology. We will argue that data quality of microarray should also take into account information coming from comprehensive biology system models.

Paola Quaglia – University of Trento
Title: Shaped interactions in Beta-binders

ABSTRACT
The talk presents Beta-binders, a formalism that adopts a typed (vs key-lock) interaction mechanism. Typed sites allow the modelling of interactions depending on the degree of 'affinity' of the interacting parties. We show how this feature can accommodate the representation of some aspects of the 'shape space' theory developed for immunology.

Ela Hunt – University of Glasgow
Title: Database support and visualisation for biological data types

ABSTRACT
The amounts of data that a biologist has to manage and understand are large, and data searching, integration and visualisation are necessary foundations for future discoveries.

We outline our findings in the area of indexed searching for peptides, and for micro array probe mapping. In this scenario we want to define exactly the search criteria (number mismatches and match length) and compute how many matches there are and where they are positioned. Such matches can later be visualised in the context of other information needed for data interpretation, using our SyntenyVista visualisation.
We then turn our attention to the issue of data integration. We are interested in automated merging of large XML data sets, including efficient joins on textual values. We hope to develop powerful tools that will allow us to merge arbitrary data sets with a greater degree of automation than currently possible. We are developing statistical and algorithmic approaches to support this process.

Giovanni Cuda and Mario Cannataro - University Magna Græcia of Catanzaro
Title: Management, Preprocessing and Data Mining Analysis of Mass Spectrometry Proteomics Data

ABSTRACT
Mass Spectrometry (MS) based proteomics produces a huge volume of data, said spectra, that contain large set of measures (intensity, m/Z), representing the abundance of biomolecules having certain mass to charge ratios. MS data hides a lot of information about cell functions and disease conditions and can be used for various analysis, e.g. biomarker discovery, peptide/protein identification, and sample classification. The discovering of such information needs the combined use of bioinformatics and data mining, and requires the efficient access to huge spectra datasets and various software tools for to the loading, management, preprocessing, and mining of spectra, as well as the interpretation and visualization of discovered knowledge models.
The increasing use of MS in clinical studies causes the collection of spectra data from large sample populations, e.g. to control the progression of a disease. In addition, the comparative study of a disease may require the analysis of spectra produced in different laboratories, so it is possible to envision that in few years biomedical researchers will need to collect and analyze more and more spectra data. Since spectra have a high dimensionality and are often affected by errors and noise, specialized spectra databases and preprocessing techniques are needed.
Finally, MS involves different technological platforms, such as sample treatments, MS techniques, spectra processing, data mining analysis, and results visualization. Choosing the right methods and tools requires multidisciplinary knowledge from MS specialists to biologists and computer scientists, thus, modelling the semantic of processes, tools, and data is a key issue to simplify application design.
Ontologies constitute a well established tool to model the steps of data mining applications and support the application design, while Grid technology may provide the broadband infrastructure and the computational power needed by preprocessing and mining algorithms.
To address the key issues of spectra data management and analysis we propose MS-Analyzer, a software platform for the design and execution of mass spectrometry-based experiments. It offers to the biologist a set of high level services, namely:
~ spectra management services, providing spectra format conversion and efficient spectra storage through a specialized spectra database;
~ pre-processing services, that implement common spectra pre-processing algorithms, such as base line subtraction, smoothing, normalization, binning, peaks extraction, peaks alignment;
~ data preparation services, that provide the spectra reorganization needed when applying data mining tools (e.g. Weka tools require spectra dataset formatted in a unique input file having a specific metadata header);
~ data mining services, obtained by wrapping Weka tools, a popular data mining suite; moreover, tools for knowledge models visualization are provided.

An Ontology-based Workflow Editor allows the concept-based browsing/searching of such services modelled through the MS-Analyzer ontologies. By using MS-Analyzer a user can easily design a data mining application with the help and the constraint checks provided by such ontologies, and without worrying of software details, having a suite of specialized spectra management services that simplify and automate the path to knowledge discovery.

[1] M. Cannataro, G. Cuda, P. Veltri (2005), Modeling and Designing a Proteomics Application on PROTEUS, (2005) *Methods of Information in Medicine*, 44(2):221-226, Schattauer Publishers.

Ralf Blossey - IRI Lille
Title: Coarse-grained models for gene regulation

ABSTRACT
Current modelling approaches in Systems Biology resemble an attempt to build a tunnel under a mountain, whereby the teams on both sides hope to meet somewhere in the middle but do not know how to ensure it. The teams on the one side of the mountain derive models from high throughput data, which relies heavily on statistical methods. The team! s on the other side try to build models 'one piece at a time', based on the detailed information available from molecular and cellular biology. I discuss some tentative ideas for mid-level approaches, combining coarse-grained models with experimental data.

Magali Roux-Rouquié - LIP6 – CNRS UPMC, Paris
Title: Linking data integration to system analysis

ABSTRACT
Most of data in Biology are produced out of the context of mathematical modeling and account for heterogeneous sets too much complex to be further integrated into models. To circumvent these limitations, biological observations can be viewed as systems requirements and approaches developed for complex software engineering could be adapted for the analysis of biological behavior.
Taking advantage of previous work on standardization of Omics sciences and graphical notation, we delineated a set of rules using the Unified Modeling Language (UML) that allow describing biological systems, from genes to organisms. To link data integration to system analysis, we are developing (collaboration: INIST, http://www.inist.fr/index_en.php) a platform dedicated to the design and the storage of models specifications. This is performed in the special context of sterols import/export and synthesis by handling data in *Saccharomyces cerevisiae* and *Caenorhabditis elegans* (collaboration: D. Pompon and R Legouis, CGM; M. Bolotin and Th Berges, LGM). Providing further, models translation will be performed into Bioambients to assess biological behavior (collaboration: D. Schuch da Rosa, UNITN). To achieve these goals, several problems have to be overcome; we will present progress and perspectives.

Oksana Tymchyshyn[1], Gethin Norman[1], Marta Z. Kwiatkowska[1,] and John K Heath[2,3] - School of Computer Science[1] and CRUK Growth Factor Group, School of Biosciences[2], University of Birmingham, Edgbaston, Birmingham B15 2TT, UK. [3] j.k.heath@bham.ac.uk
Title: Computer assisted biological reasoning: applying stochastic pi-calculus techniques to simulation and analysis of FGF signalling pathway dynamics.

ABSTRACT
Many biological systems exhibit the potential for substantial behavioural complexity which is not readily understood or analysed by traditional methods of biological intuition. Computational tools

which assist in reasoning about biological systems are valuable accelerators of the discovery process. Process calculi offer many attractions for simulating biological systems because computational primitives can be readily mapped onto the real worlds of molecules, interactions and state changes. The FGF signalling pathway is an interesting test case for computational modelling because alterations in pathway dynamics (ie signal duration and amplitude) underlie common pathologies and many components in the pathway are known. There is however debate in the biological literature regarding the influence of different pathway reactions on signalling dynamics. Here we describe implementation of a Stochastic Pi calculus model of a fragment of the FGF pathway explicitly designed to reason about known biological events: ligand activation, substrate phosphorylation/dephosphorylation and degradation; protein/protein interaction/competition and protein complex relocation. We show that the base model reproduces the known dynamics of FGF signalling. We then interrogate the model by "in silico genetics" and exploration of parameter space. The results of these experiments indicate that pathway dynamics are dominated by two variables: the rate of receptor kinase activation and the rate of signalling complex relocation. By removal of components from the model we test the role of different molecules in shaping signalling dynamics. Surprisingly this reveals that a known attenuator of FGF signalling Sprouty – explicitly modelled to function by protein/protein competition- fails to act as described in the biological literature. Further parameter exploration suggests that this may reflect the design of biological experiments. We conclude by outlining future challenges for process calculi approaches in simulating biological signalling pathways.

Alessandro Quattrone – University of Firenze
Title: An ontology-based information system for the dynamic integration of high throughput biological data with textual knowledge

ABSTRACT
Nanobiotecnologies provide high-throughput (HT) quantitative data about the whole organization of biological macromolecules in cells and tissues. These data are measures of the flux of genomic information, and are increasingly complemented by qualitative descriptions on all levels of cellular organization of biological macromolecules, like interactome maps and signalling and metabolic pathways.
Semantic systems of classification of biological objects (biological ontologies) have been recently built to allow the interpretation of HT data in terms of engagement of subcellular systems and subsequent phenotypic responses. Here we provide an example coming from the experimental characterization of a model of tumor progression to discuss the need of development of biological ontologies covering new important domains, such as post-transcriptional regulation of gene expression. To this aim, we propose a general classification scheme for biological and biomedical ontologies.
We also present a project for the integration of HT data and biological knowledge based on a system of weighted edges between ontology nodes. The weight of these edges is defined using resources of biomedical knowledge organized in public domain databases available online. In such architecture, the loading of nodes of some input ontologies with HT data and the evaluation of the statistical over-representation of these data should allow the translation of the ontological relations in biologically and biomedically relevant information.

Roberto Gorrieri (joint work with Nadia Busi and Cristian Versari) – University of Bologna
Title: Modelling Biological Systems with Polyadic Synchronization and Priority

ABSTRACT
We show some initial ideas about the study of a simple, conservative, yet tremendously more expressive, extension of the pi-calculus suited for the modelling of complex phenomena of

biological systems. The two key features of the proposed calculus, called pi@ (pronounced like the french "paiette"), are: compartments (modeled by means of polyadic synchronization) and atomicity of complex operations (modeled by means of priority). Some preliminary thoughts about expressiveness are discussed: in particular, some subcalculi (prio-pi, poly-pi, etc..) are introduced to single out the relative expressive power of the various mechanisms occurring in pi@. Finally, some natural and elegant encodings of the brane calculus and of bioambients into pi@ are sketched, hence justifying our claim that pi@ can be a good candidate for modelling biological systems.

Esther Graudens - CNRS, Paris
Title: Modelling cellular states of innate tumor drug response

ABSTRACT
Background: The molecular mechanisms underlying innate tumor drug resistance, a major obstacle to successful cancer therapy, remain poorly understood. In colorectal cancer (CRC), molecular studies have focused on drug-selected tumor cell lines or individual candidate genes using samples derived from patients already treated with drugs, so that very little data are available prior to drug treatment. Results: Transcriptional profiles of clinical samples collected from CRC patients prior to their exposure to a combined chemotherapy of folinic acid, 5-fluorouracil and irinote can were established using microarrays. Vigilant experimental design, power simulations and robust statistics were used to restrain the rates of false negative and false positive hybridizations, allowing successful discrimination between drug resistance and sensitivity states with restricted sampling. A list of 679genes was established that intrinsically differentiates, for the first time prior to drug exposure, subsequently diagnosed chemo-sensitive and resistant patients. Independent biological validation performed through quantitative PCR confirmed the expression pattern on two additional patients. Careful annotation of interconnected functional networks provided a unique representation of the cellular states underlying drug responses. Conclusion: Molecular interaction networks are described that provide a solid foundation on which to anchor working hypotheses about mechanisms underlying *in vivo* innate tumor drug responses. These broad-spectrum cellular signatures represent a starting point from which by-pass chemotherapy schemes, targeting simultaneously several of the molecular mechanisms involved, may be developed for critical therapeutic intervention in CRC patients. The demonstrated power of this research strategy makes it generally applicable to other physiological and pathological situations.

Matteo Cavaliere – The Microsoft Research-University of Trento Centre for Computational and Systems Biology
Title: Modeling (and Simulating) Biological Processes with Stochastic Multiset Rewriting.

ABSTRACT
We propose a prototype of biological simulator based on multisets rewriting executed in cell-like compartments (as introduced in the area of membrane systems). The software can simulate chemical reactions, enriched with stochastic rates. The syntax of the simulator is essentially based on rewriting rules and in this way chemical reactions can be easily modeled. A series of basic experiments is presented.

Graziano Pesole – University of Bari
Title: Computational identification of novel genes and regulatory elements in the human genome

ABSTRACT
Although the complete nucleotide sequences of human and other vertebrate genomes are already known we are far from the complete inventory of all encoded gene and gene expression isoforms, as well as of the regulatory elements modulating gene expression at the transcriptional and post-transcriptional level.
We recently developed several computational tools for the identification of known and novel regulatory elements acting as transcriptional or post-transcriptional regulators. Furthermore, we also developed specific tools for the identification of sequence tags (CSTs) conserved across multiple genomes and their functional characterization as protein coding or not. Indeed, clusters of protein coding CSTs, located in intergenic regions, may represent unannotated genes that could be eventually experimentally validated, whereas non coding CSTs may represent novel regulatory elements controlling gene expression.
We will discuss the general features and application areas of such tools that can be freely accessed on the web at www.pesolelab.it (Tools section).

Riccardo Velasco – IASMA Research Centre
Title: Beyond the genome: what's next in plant system biology

ABSTRACT
Genome analysis in plant needs strong support from bioinformatics to supply the scientific community of tools to assign function and possible interaction to putative proteins. Reading of a genome is not any longer more than a technical task. The real deal is to understand what a nucleotide sequence code for, what a putative protein does, which relationship two or more proteins have, how do a protein complex work, and finally how do entire pathways work and interact each others during the cell life. Progress in yeast and human science can be useful to understand partially such biological events, but specificity of plants needs further efforts of bioinformatics to sustain plant science.

Giancarlo Mauri – University of Milano-Bicocca
Title: Modelling biological processes with P-systems

ABSTRACT
P-systems, or membrane systems, have been recently introduced by Gheorghe Paun [1] as a computation model taking inspiration from the structure and the functioning of living cells. Paun proposed to abstract from the architecture of the cell and the way biological substances are both modified and moved among internal compartments, and to interpret the phenomena occurring inside the cell as computing processes.
In this formal model, a membrane structure is described by a finite string of well matching parentheses, and graphically represented as regions on the plane, hierarchically embedded in an external region. The chemical substances and reactions are represented by means of *objects* and *evolution rules*. Objects are described as symbols or strings over a given alphabet, evolution rules are given as rewriting rules. The rules act on objects, by modifying and moving them, and they can also affect the membrane structure, by dissolving the membranes.
A computation in P systems is obtained by starting from an initial configuration, identified by the membrane structure, the objects and the rules initially present inside it, and then letting the system evolve. The application of rules is performed in a nondeterministic and maximal parallel manner: all the applicable rules have to be used to modify all objects which can be the subject of a

rule, and this is done in parallel for all membranes (a universal clock is assumed to exist). Whenever no rule can be further applied, the computation halts and the output is defined in terms of the objects sent out the external membrane or, alternatively, collected inside a specified membrane. No output is obtained if the computation never halts (that is, whenever a rule can be continuously applied). It should be emphasized here that P-systems were not initially intended to be a model of the cell, but were defined with the purpose of investigating some computational features which can be abstracted from the cellular biology. Hence, they have been extensively studied in the area of Natural Computing from the point of view of their computational power, and compared with other models like DNA computing or splicing systems.
However, they can also be considered as a powerful tool to model and simulate cellular phenomena, hence returning meaningful and useful information to biologists. in the frame of systems biology. Some examples of use of P-systems to model cellular processes (i.e., Na-K pump or Mechanosensitive channels) will be given in the talk. The design of an appropriate software simulator, based on the models we will present, would then provide an easier way to check both the effectiveness and the correctness of the models, and hopefully become a tool for biologists for testing known data, predicting unknown scenarios and returning meaningful information.

1. G. Paun. *Membrane Computing. An Introduction.* Springer–Verlag, Berlin, 2002.