

EMBnet.journal

Volume 18
Supplement B
November 2012

NETTAB 2012

Workshop on "Integrated Bio-Search"

14-16 November 2012, Como, Italy

<http://www.nettab.org/2012/>

Editorial

The Network Tools and Applications in Biology (NETTAB) series of workshops in Bioinformatics held its "NETTAB 2012 workshop focused on integrated Bio-Search, 14-16 November 2012, Como, Italy.

The NETTAB 2012 workshop was held under the patronage of Bioinformatics Italian Society (BITS) and the European Molecular Biology network (EMBnet).

The event featured a range of lively and stimulating scientific sessions focused on innovative Information and Communication Technologies (ICT), tools, systems, applications and perspectives applied to the biomedical domain.

The EMBnet.journal is very pleased to publish the contributions to the workshop presented at this important event.

For this special issue (18.B), the selection of articles was overseen by the Conference Scientific Committee, while the layout and logistics were organised by the EMBnet.Journal Editorial Team.

For future conferences, our Online Journal System (OJS) can also be used for receiving, archiving and managing the full review process – note that we recently also published the proceedings of the Bioinformatics Italian Society (BITS) (Issue 18.A) 9th Annual Meeting, May 2-4, 2012, and the First Scientific Meeting of the COST Action, SeqAhead (issue 17.B).

We therefore welcome contributions from other Societies and Networks, and encourage interested parties to contact members of the Editorial Board.

EMBnet.journal Editorial Board

Contents

Editorial	2
The Bioinformatics Italian Society	3
Preface - NETTAB 2012	
Workshop on "Integrated Bio-Search"	4
Scientific Programme.....	8
Keynote Lectures.....	12
Tutorials	16
Oral Communications	19
Technological Industrial Communications	55
Short Oral Communications	61
Posters.....	80

EMBnet.journal Executive Editorial Board

Erik Bongcam-Rudloff, Department of Animal Breeding and Genetics, SLU, SE,
erik.bongcam@slu.se

Teresa K. Attwood, Faculty of Life Sciences and School of Computer Sciences, University of Manchester, UK,
teresa.k.attwood@manchester.ac.uk

Domenica D'Elia, Institute for Biomedical Technologies, CNR, Bari, IT,
domenica.delia@ba.itb.cnr.it

Andreas Gisel, Institute for Biomedical Technologies, CNR, Bari, IT,
andreas.gisel@ba.itb.cnr.it

Laurent Falquet, Swiss Institute of Bioinformatics, Génopode, Lausanne, CH,
Laurent.Falquet@isb-sib.ch

Pedro Fernandes, Instituto Gulbenkian, PT,
pfern@igc.gulbenkian.pt

Lubos Klucar, Institute of Molecular Biology, SAS Bratislava, SK,
klucar@EMBnet.sk

Martin Norling, Swedish University of Agriculture, SLU, Uppsala, SE,
martin.norling@slu.se

Bioinformatics Italian Society



www.bioinformatics.it

BITS is a scientific society founded in 2003 with the aim of studying and disseminating Bioinformatics in the academic and research context as well as in the technological and applicative world. In Italy, the foundation of BITS started in 1999 by a SIBBM working group for the cooperation in bioinformatics ("Gruppo di Cooperazione Bioinformatica") and then also by ABCD, has greatly promoted the spreading of the bioinformatics discipline in the Italian scientific community. The joining of researchers from different scientific areas, not only biology, is the consequence of the multidisciplinary nature of bioinformatics, and BITS has now about 250 members which constitute a young community with expertise and interest in bioinformatics and the different "classical" disciplines converging on it (molecular biology, biochemistry, genetics, medicine, biophysics, informatics, mathematics, statistics, physics).

BITS is recognized by the International Society for Computational Biology (ISCB) as an affiliated Regional Group. BITS activities include the organization of the annual scientific meeting of the Society, the support to numerous scientific events proposed by BITS members, distribution of news of interest for the involved community of researchers by means of its web site and mailing list, the coordination of educational initiatives in Italy, and the support for the participation of Italian researchers to international events and projects of relevance.

BITS is glad to sponsor the NETTAB2012 workshop and wishes the Organizers and participants a successful meeting.

Preface

NETTAB 2012

Workshop on "Integrated Bio-Search"



Marco Masseroli¹, Paolo Romano², Frédérique Lisacek³

¹Politecnico di Milano, Milan, Italy

²IRCCS San Martino IST, Genova, Italy

³SIB Swiss Institute of Bioinformatics, Geneva, Switzerland

NETTAB Workshops are a series of International meetings on "Network Tools and Applications in Biology" held annually in Italy. They are aimed at introducing participants to the most promising among those innovative Information and Communication Technologies (ICTs) that are being applied to the biomedical application domain. Workshops include many focused sessions which are devoted to tools, systems, applications, and perspectives. Keynote lectures introduce the sessions' topics, and are followed by presentations selected from among the submitted contributions after peer review by members of the Scientific Committee. Discussion is a key factor, both within sessions and in a special Panel Discussion. Tutorials and poster sessions complete the agenda of the NETTAB workshops. Each year, the workshop is focused on a different technology or domain. Since 2001, many different topics, often related to data integration issues, were discussed. These included, e.g., Standardization for data integration (Genoa, 2001), Multi agent systems (Bologna, 2002), Scientific workflows (Naples, 2005), Grid and Web Services (Santa Margherita di Pula, 2006), The Semantic Web (Pisa, 2007), Collaborative research and development (Catania, 2009), Biological wikis (Naples, 2010), and Clinical Bioinformatics (Pavia, 2011).

The NETTAB 2012 workshop, the twelfth in the series, was held in Como, Italy, on November 14-16 2012.

Its rationale is based on the consideration that the data deluge of the current post-genom-

ic era is providing scientists with potentially very valuable information, but it makes difficult to find and extract from the increasing available high-throughput omics data those information that are most reliable, specific and most related to the biomedical questions to be answered. Such questions are increasingly complex and they often simultaneously regard many heterogeneous aspects of an organism, tissue, cell and the role of all biomolecular entities. Several of these questions can be addressed only by comprehensively searching different types of data, which generally are distributed in many heterogeneous sources. Usually, scientists explore these data by using the individual search services and tools available in Internet and they then struggle in combining the essential information in order to answer their global questions. In this context, moreover, quality and consistency checking is a central issue that should be brought up

Searching and combining all open and linked data and algorithmic sources has the potential of reshaping the scenario of current bioinformatics applications, going beyond the capabilities of conventional tools, Web services and existing search engines. Yet, it also presents new technological challenges.

Solving data integration and automatic extraction problems requires new solutions, including the use of universal URIs, efficient indexing, partial or approximate value matching, rank aggregation, continuous or push-based search, exploratory methods and context-aware paradigms, collaborative and social search, and building new efficient information retrieval approaches, based on automation of workflows too that may contribute to new "good practices" in data searching, retrieval, and integration, with the specific goal of ensuring quality of procedures, as well as their reproducibility coupled with efficiency and efficacy.

On these premises, then, the NETTAB 2012 workshop has been focused on "Integrated Bio-Search", which includes all aspects that relate to technologies, methods, architectures, systems and applications for searching, retrieving, integrating and analyzing data, information, knowledge, infrastructures, services and tools that are required to answer complex bio-medical-molecular questions.

Workshop topics included four main areas. The first area relates to data integration. It in-

cludes syntactic and semantic methods and algorithms for biological and clinical data and knowledge integration, information and knowledge retrieval, data and knowledge query, data, information and knowledge extraction, and data and knowledge mining.

The second area refers to new and optimized technologies for data management. It includes federated databases, data warehouses, and triple stores. It also includes topics as biomedical terminologies and ontologies, systems' interoperability, natural language processing, and scientific workflow processing.

Tools and platforms for molecular data management and storage, deep sequencing analysis, omics data computing, search computing, decision support, and clinical bioinformatics are the third topic area, while the fourth area includes examples of applications of these methods, technologies and tools in different biomedical domains, such as knowledge assessment, integration, discovery, and validation, drug design, diagnosis and prognosis support, and personalized medicine.

The Call for abstracts was able to attract 34 submissions for oral communications. From

these submissions, the Scientific Committee of the workshop was able to select 12 oral communications, seven short oral communications, and three technological communications from industry. All submissions underwent peer review by at least two members of the Scientific Committee. At the workshop, 29 posters were also presented. Submissions for posters were also peer reviewed by one or two members of the Scientific Committee. This Supplement therefore includes about 50 abstracts, all revised according to reviews, which are grouped by submission type and ordered by first author name.

The NETTAB 2012 workshop has been a great meeting for all researchers involved in data search and integration in biology and medicine. It was possible to discuss ideas, and doubts, with such scientists as Erik Bongcam-Rudloff, Barend Mons, Eric Neumann, Alexander Kel, Katy Wolstencroft, who accepted to give invited lectures and tutorials, and many others who enthusiastically joined the workshop.

And, of course, the workshop has been a great occasion to enjoy Italian lifestyle....

Speakers

Keynote Speakers

Erik Bongcam-Rudloff

Swedish University of Agricultural Sciences, and Uppsala University, Sweden

Barend Mons

Leiden University Medical Center, and Netherlands Bioinformatics Center, The Netherlands

Eric Neumann

PanGenX, and Clinical Semantics Technologies, USA

Tutorials

Alexander Kel

GeneXplain GmbH, Wolfenbüttel, Germany, and Biosoft.ru, Skolkovo Center of Bioinformatics, Novosibirsk, Russian Federation

Katy Wolstencroft

School of Computer Science, University of Manchester, United Kingdom

Oral presentations

Claudio Angione

Bachir Balech

Esra Erdem

Giovanni Felici

Francesca Finotello

Matteo Gabetta

Francisco Gómez-Vela

Alejandra Gonzalez-Beltran

Claudia Gugenmus

Víctor Martínez

Marco Masseroli

Marco Muselli

Carmen Navarro

Piergiorgio Palla

Uberto Pozzoli

François Rechenmann

Fabio Rinaldi

Patrick Ruch

Bahar Sateli

Saif Ur-Rehman

Editors

Marco Masseroli
Paolo Romano
Frédérique Lisacek

Politecnico di Milano, Milan, Italy
 IRCCS San Martino IST, Genoa, Italy
 SIB Swiss Institute of Bioinformatics, Geneva, Switzerland

Editorial Staff

Davide Chicco
Arif Canakoglu

Politecnico di Milano, Milan, Italy
 Politecnico di Milano, Milan, Italy

Chairs and Conference Committees

Chairs

Marco Masseroli
Paolo Romano
Frédérique Lisacek

Politecnico di Milano, Milan, Italy
 IRCCS San Martino IST, Genoa, Italy
 SIB Swiss Institute of Bioinformatics, Geneva, Switzerland

Scientific committee

Francisco Azuaje
Riccardo Bellazzi
Olivier Bodenreider
Mario Cannataro
Bastien Chopard
Marie-Dominique Devignes
Christine Froidevaux
Carole Goble
Nicolas le Novère
Ulf Leser
Frédérique Lisacek
Paolo Magni
Roberto Marangoni
Marco Masseroli
Luciano Milanese
Paolo Missier
Heiko Muller
Norman Paton
Horacio Pérez-Sánchez
Paolo Romano
Patrick Ruch
Indra Neil Sarkar

Centre de Recherche Public de la Santé, Luxembourg
 University of Pavia, Italy
 National Institutes of Health, USA
 University "Magna Græcia" of Catanzaro, Italy
 Swiss Institute of Bioinformatics, University of Geneva, Switzerland
 CNRS LORIA, France
 University Paris-Sud, France
 The University of Manchester, United Kingdom
 European Bioinformatics Institute, United Kingdom
 Humboldt University, Germany
 SIB Swiss Institute of Bioinformatics, Geneva, Switzerland
 University of Pavia, Italy
 University of Pisa, Italy
 Politecnico di Milano, Milan, Italy
 Biomedical Technologies Institute, CNR, Italy
 Newcastle University, United Kingdom
 Istituto Italiano di Tecnologia, Italy
 University of Manchester, United Kingdom
 University of Murcia, Spain
 IRCCS San Martino IST, Genoa, Italy
 University of Applied Sciences, Geneva, Switzerland
 University of Vermont, USA

Organising Committee

Marco Masseroli
Arif Canakoglu
Davide Chicco

Politecnico di Milano, Milan, Italy
 Politecnico di Milano, Milan, Italy
 Politecnico di Milano, Milan, Italy

Organization Staff

Centro Studi Scientifici "Alessandro Volta", Como, Italy



Centro di Cultura Scientifica
 "Alessandro Volta"

Supporting Institutes, Scientific Societies and Projects

The workshop is held under the patronage of



Bioinformatics Italian Society
<http://www.bioinformatics.it/>



EMBnet: the Global Bioinformatics Network
<http://www.embnet.org/>

and with support from



**POLITECNICO
DI MILANO**

Politecnico di Milano, Italy
<http://www.polimi.it/>



San Martino IST, Genoa, Italy
<http://www.hsanmartino.it/>



**Swiss Institute of
Bioinformatics**

SIB Swiss Institute of Bioinformatics
<http://www.isb-sib.ch/>



Flagship INTEROMICS Project



Search Computing (SeCo) Project
<http://www.search-computing.it/>



CNR BIOINFORMATCS Project
<http://www.cnr.it/sitocnr/>

Sponsors



CRC Press
<http://www.crcpress.com/>



Camera di Commercio di Como
<http://www.co.camcom.gov.it/>

Scientific Programme

NETTAB 2012 **Workshop on “Integrated Bio-Search”**

14-16 November 2012, Como, Italy

<http://www.nettab.org/2012/>

Scientific Programme

Wednesday November 14

9.00 - 10.50 **Tutorial 1**

Multi-scale data integration and virtual exploration from promoters, through networks to drug targets

Alexander Kel, GeneXplain GmbH, Wolfenbüttel, Germany, and Biosoft.ru, Skolkovo Center of Bioinformatics, Novosibirsk, Russian Federation

10.50 - 11.10 *Break*

11.10 - 13.00 **Tutorial 2**

The Taverna Workbench: Integrating and analysing biological and clinical data with computerised workflows

Katy Wolstencroft, University of Manchester, United Kingdom

13.30 - 14.20 *Registration and poster hang-up*

14.20 - 14.30 *Welcome and Introduction*

14.30 - 15.10 **Invited Lecture**

Integration and analysis of multi-type high-throughput data for biomolecular knowledge discovery

Erik Bongcam-Rudloff, Swedish University of Agricultural Sciences, and Uppsala University, Sweden

15.10 - 15.50 **Scientific Session 1**

Using graph theory to analyze gene network coherence

Francisco Gómez-Vela, Norberto Díaz-Díaz, Jose Antonio Lagares, Jose Antonio Sánchez and Jesús S. Aguilar-Ruiz

Network-based gene-disease prioritization using PROPHNET

Víctor Martínez, Carlos Cano and Armando Blanco

15.50 - 16.20 **Coffee Break**

16.20 - 18.15 **Scientific Session 2**

Rational design of organelle compartments in cells

Claudio Angione, Giovanni Carapezza, Jole Costanza, Pietro Lio' and Giuseppe Nicosia

Filtering with alignment free distances for high throughput DNA reads assembly

Maria Cristina De Cola, Giovanni Felici, Daniele Santoni and Emanuel Weitschek

A strategy to reduce technical variability and bias in RNA sequencing data

Francesca Finotello, Enrico Lavezzo, Luisa Barzon, Paolo Mazzon, Paolo Fontana, Stefano Toppo, Barbara Di Camillo

Applications of a generic model of genomic variations functional analysis

Sarah N. Mapelli, Uberto Pozzoli

The Biovel project: robust phylogenetic workflows running on the Grid

Saverio Vicario, Bachir Balech, Giacinto Donvito, Pasquale Notarangelo, Graziano Pesole

Ranking-aware integration and explorative search of distributed bio-data

Marco Masseroli, Matteo Picozzi and Giorgio Ghisalberti

Development of a text search engine for medicinal chemistry patents

Emilie Pasche, Julien Gobeill, Fatma Oezdemir-Zaeche, Therese Vachon, Christian Lovis and Patrick Ruch

Thursday	November 15
8.30 - 9.00	<i>Registration and poster hang-up</i>
9.00 - 9.40	Invited Lecture <i>Semantics based biomedical knowledge search, integration and discovery</i> <i>Barend Mons, Leiden University Medical Center, and Netherlands Bioinformatics Center, The Netherlands</i>
9.40 - 10.20	Scientific Session 3 <i>Answering Gene Ontology terms to proteomics questions by supervised macro reading in Medline</i> <i>Julien Gobeill, Emilie Pasche, Douglas Teodoro, Anne-Lise Veuthey and Patrick Ruch</i> <i>IntelliGenWiki: An Intelligent Semantic Wiki for Life Sciences</i> <i>Bahar Sateli, Marie-Jean Meurs, Gregory Butler, Justin Powlowski, Adrian Tsang and René Witte</i>
10.20 - 12.00	Poster and Software Demonstration Session with Coffee Break
12.00 - 13.00	Technological - Industrial Session <i>Extracting knowledge from biomedical data through Logic Learning Machines and RuleX</i> <i>Marco Muselli</i> <i>Data modeling: the key to biological data integration</i> <i>François Rechenmann</i> <i>GeneGrid: finding disease causing variants in NGS data</i> <i>Jochen Supper, Claudia Gugenmus, Korbinian Grote and Frederic Eyber</i>
13.00 - 14.00	Lunch Break
14.00 - 15.30	Panel Discussion <i>Technological and methodological challenges for Integrated Bio-Search</i> <i>Erik Bongcam-Rudloff, Barend Mons, Eric Neumann, Alexander Kel, François Rechenmann, and Stefano Ceri introduce the topic, then open discussion follows</i>
15.30 - 19.00	Guided tour of Como and of the Educational Silk Museum of Como
20.00 - 23.00	Social Dinner

Friday	November 16
9.00 - 9.40	Invited Lecture <i>Clinical and genomic data integration in support of biomedical research and clinical practice</i> <i>Eric Neumann, PanGenX and Clinical Semantics Technologies, USA</i>
9.40 - 10.40	Scientific Session 4 <i>ROCK: a resource for integrative breast cancer data analysis</i> <i>Marketa Zvelebil, Costas Mitsopoulos and Saif Ur-Rehman</i> <i>QTreds: a flexible LIMS for omics laboratories</i> <i>Piergiorgio Palla, Gianfranco Frau, Laura Vargiu and Patricia Rodriguez-Tomé</i> <i>The open source ISA software suite and its international user community: knowledge management of experimental data</i> <i>Alejandra Gonzalez-Beltran, Eamonn Maguire, Philippe Rocca-Serra and Susanna-Assunta Sansone</i>
10.40 - 11.10	Coffee Break
11.10 - 12.30	Scientific Session 5 <i>The ontogene system: an advanced information extraction application for biological literature</i> <i>Fabio Rinaldi</i> <i>A semantic collaborative system for the management of translational research projects</i> <i>Matteo Gabetta, Giuseppe Milani, Cristiana Larizza, Valentina Favalli, Eloisa Arbustini and Riccardo Bellazzi</i> <i>BioQuery-ASP: querying biomedical databases and ontologies using Answer Set Programming</i> <i>Esra Erdem, Umut Oztok</i> <i>DiGSNP: a web tool for Disease-Gene-SNP hierarchical prioritization</i> <i>Carmen Navarro, Carlos Cano, Armando Blanco, Fernando García</i>
12.30 - 13.00	Announcement of NETTAB 2013 and Farewell

Keynote Lectures

Integration and analysis of multi-type high-throughput data for biomolecular knowledge discovery



Erik Bongcam-Rudloff

Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, and Department of Immunology, Genetics and Pathology, Uppsala University, Sweden

Erik Bongcam-Rudloff received his doctorate in medical sciences from Uppsala University, Sweden. He is now the Director of the SLU Global Bioinformatics Centre (SGBC) at the Swedish University of Agricultural Sciences.

Erik was the chairman of EMBnet (2003-2010), a science-based group of worldwide collaborating bioinformatics nodes. He also coordinated the "Test Cases" work package 4 in the FP6 EMBRACE project (2005-2010). The goal of WP4 was to collect test cases from the scientific community to identify real research problems and to provide solutions for content and tool integration.

Bongcam-Rudloff is today Chair of SeqAhead, a Biomedical European COST Action: "Next Generation Sequencing Data Analysis Network" and Coordinator of ALLBIO, a FP7 project:

"Broadening the Bioinformatics Infrastructure to unicellular, animal, and plant science". He is also the founder of eBioinformatics.org the creators of eBiotools, eBioKit, eBioX and eBioKit.

His main research deals with development of bioinformatics solutions for the Life Sciences community.

In this talk he will discuss the new techniques that are now driving the generation of knowledge (especially in biomedicine and molecular life sciences) to new dimensions eg. NGS. He will also discuss the new opportunities in human and non-human research that these techniques create but also the new challenges in the design of ontologies for data and methods, and choosing common interoperability standards.

Semantics based biomedical knowledge search, integration and discovery



Barend Mons

Leiden University Medical Center, Leiden, The Netherlands, and Netherlands Bioinformatics Center

Barend Mons holds a chair in Biosemantics at the LUMC and is one of the scientific directors of NBIC. In addition he acts as a Life Sciences 'eScience integrator' in the Netherlands eScience centre. Currently, he coordinates the creation of the Data Integration and Stewardship Centre (DISC-ELIXIR) and in that capacity he is also the scientific representative of The Netherlands in the interim board of the ELIXIR ESFRI project.

Barend Mons is a molecular biologist by training and received his PhD on genetic differentiation of malaria parasites from Leiden University (1986). He performed over a decade of research on malaria genetics and vaccine development, also serving for 3 years the research department of the European Commission in this field. He did gain further experience in science management at the Research council of The Netherlands (NWO).

Barend is the co founder of three spin-off companies in biotechnological and semantic technologies. In 2000, he switched back to academia, focusing on the development of semantic technologies to manage big data and he founded the Biosemantics group.

His research is currently focused on nanopublications as a substrate for in silico knowledge

discovery. Barend is also one of the founders of the Concept Web Alliance, with "nanopublications" as its first brainchild. Nanopublications are currently implemented in the semantic project of the Innovative Medicines Initiative (IMI) called Open PHACTS.

Barend Mons will talk about the role of semantic technologies and related standards applied to biomedical-molecular data integration and biomedical knowledge search and discovery.

He will challenge several established views in the field of the Semantic Web for Life Sciences, by also taking into account "data publishing" in a broad sense, including, e.g., biomedical communication, intellectual networking, and nanopublications, with an emphasis on the barriers to brake down in order to allow effective data exposure, sharing, searching, and integration, and to "in silico" discovery of new biomedical knowledge in the Big Data era.

This talk will introduce the need for a semantics based eScience approach for "in silico" knowledge discovery. It will also show how such approach can indeed already support search, integration, and discovery.

Clinical and genomic data integration in support of biomedical research and clinical practice



Eric Neumann

PanGenX, and Clinical Semantics Technologies, United States

Eric Neumann is a graduate from MIT and holds a PhD in neurobiology, developmental genetics, and pharmacology from Case Western Reserve University. He is a recognized expert in semantic information, and has worked on many information initiatives for the pharmaceutical and life sciences, including the W3C Semantic Web Healthcare and Life Science Interest Group (HCLSIG).

Eric Neumann was the Global Head of Knowledge Management for Scientific and Medical Affairs within Sanofi-Aventis and the VP of Informatics at BG Medicine. He founded Genstruct (now Sleventa) and has also worked at Bolt, Beranek, and Newman (now BBN Technologies - Raytheon) on several advanced scientific computation projects over the span of several years.

He is Founder and CTO of PanGenX, a personalized medicine company, whose mission is to optimize therapeutic care by facilitating the discovery and application of medical knowledge towards patient segmentation.

Starting from the current definition of pharmacogenomics, in his talk Eric Neumann will show how it drives the personalized medicine vision, and will discuss what new forms of clinical and genomic information will be required for making clinical decision in personalized medicine.

He will illustrate some available public data sources, showing what they still lack and the value of deep focus curation. He will contrast data vs. "Actionable Knowledge" and discuss how leveraging semantically linked data to help progress personalized medicine. He will also address large-scale analytics and argue about who will benefit from linked knowledge.

Tutorials

Clinical and genomic data integration in support of biomedical research and clinical practice



Alexander Kel

GeneXplain GmbH, Wolfenbüttel, Germany, and
Biosoft.ru, Skolkovo Center of Bioinformatics, Novosibirsk, Russian Federation

Alexander Kel studied biology and mathematics at Novosibirsk State University and obtained his MS in 1985. He worked for 15 years at the Institute of Cytology and Genetics, Russia (ICG) finally holding the position of Vice-Head of the Theoretical Molecular Genetics Lab. In 1990 he received his PhD in Bioinformatics, Molecular Biology and Genetics. In 1999 he organized a Bioinformatics group at ICG.

From 2000 to 2010, he has been the Senior Vice President Research & Development of BIOBASE GmbH.

During his career, he has worked in many branches of current bioinformatics. He is a prolific author of scientific publications, as well as of tutorials and education materials.

In the tutorial, he will approach the analysis and modeling of biological systems from several practical angles. First, he will introduce into systems biology and modeling from a network-

based perspective. He will introduce several pathway databases and describe how to use them for pathway analysis. Next, he will describe computational methods for analysis of pathway information and for reconstruction of signal transduction and gene regulatory pathways using gene expression data and knowledge from the pathway databases. This will be followed by methods of analysis of topological properties of biochemical and regulatory networks. This will lead to the application of such methods for revealing key nodes in networks as potential biomarkers or drug targets. He will then show examples of application of these methods for identification of disease related biomarkers and drug discovery.

The attendees of the tutorial will get demo of the online system geneXplain with the aim to enable them to use it in their lab.

The Taverna Workbench: Integrating and analysing biological and clinical data with computerised workflows



Katy Wolstencroft

School of Computer Science, University of Manchester, United Kingdom

Katy Wolstencroft is a Research Fellow in the School of Computer Science, University of Manchester and a visiting researcher in the Molecular Cell Physiology group at the Vrije Universiteit, Amsterdam. She has a PhD and MSc in Bioinformatics from the University of Manchester, and a BSc in Biochemistry from the University of Leeds.

Katy's work is primarily in the area of data and knowledge integration, where she leads the bioinformatics research activities in the myGrid consortium. myGrid is a UK e-Science initiative that has produced, amongst other things, the Taverna workflows workbench (<http://www.taverna.org.uk/>), the myExperiment workflow repository (<http://www.myexperiment.org>) and the BioCatalogue service catalogue (<http://www.biocatalogue.org>). Currently, her main focus is on the BBSRC funded SysMO SEEK project, to develop a data exchange and modelling environment for Systems Biology consortia in Europe

(<http://www.sysmo-db.org/>). It was designed for the SysMO consortium, (Systems Biology of Micro-Organisms), but it has now been adopted by many other consortia, providing a common platform for hundreds of research labs in Europe.

Katy also coordinates the training and outreach activities in myGrid. As such, she has been involved in teaching scientific workflows and related technologies in over 50 workshops, summer schools and conferences throughout the world. In this tutorial, she will provide an introduction to designing and reusing workflows for high-throughput bioinformatics data analysis, using Taverna and myExperiment. Scientific workflows enable the chaining together of distributed analysis resources and databases to construct complex analysis pipelines that are ideal for high throughput omics data analysis. These workflows are reusable experimental methods that can be shared and rerun for other data, or for experimental validation.

Oral Communications

Rational design of organelle compartments in cells

Claudio Angione^{1✉}, Giovanni Carapezza², Jole Costanza², Pietro Lió¹, Giuseppe Nicosia²

¹ Computer Laboratory, University of Cambridge, Cambridge, United Kingdom

² Department of Mathematics and Computer Science, University of Catania, Catania, Italy

Motivation and Objectives

In recent years there is a growing interest in researching on mitochondria, chloroplasts and other mitochondrion-like organelles (e.g. hydrogenosomes, mitosomes and apicoplasts) because of the integrate bio-search for comorbidities-related genes, pathway dysfunctions, the energy balance in aging, inflammation and disease, and the discovery of novel factors involved in organelle division, movement, signaling and adaptation to varying environmental and pathogenic conditions.

Furthermore, there is an impressive amount of mitochondria and chloroplasts sequence data (thousands of mitochondrial sequences from many species have been sequenced) that have been used in the last ten years to derive the history of species. Notably, there are no examples of examined eukaryotes without a mitochondrion-related organelle (Shiflett and Johnson, 2010). Despite these research efforts, there is a lack of knowledge about the relationships between the organelles in a cell and its metabolism.

We aim at investigating and comparing the complexity of these organelles through a common framework that includes single- and multi-objective optimization, robustness analysis and sensitivity analysis. The possibility of multi-objective-optimization in organelles such as the mitochondrion may be related to the different tasks of maximizing the ATP or the heat, or intermediate compounds of the Krebs cycle in order to provide input for biosynthetic pathways (e.g. the amino acids synthesis).

Furthermore, rather than focusing only on networks of molecules, we think of the cell as an integrated system (Yoneda et al, 2009). Indeed, the systems biology approach, i.e. taking into account only molecular networks, misses the analysis of the organization provided by organelles. An organelle can be viewed as a functional organization of macromolecules working to accomplish essential cellular functions. Many conditions depend on a variety of environmental and other factors, and therefore cannot be fully

investigated by conventional molecular-level approaches.

Methods

The cell contains many membrane-bound organelles, each specialized in one or more functions. External reactions can be thought of as links between organelles, since they involve metabolites found both in the cytoplasm and in the organelles. In our framework (Figure 1), we take into account complete models of the organelles in the cell, whose state space reflects its metabolism. The genetic algorithm underlying the optimization allows us to reach the optimal Pareto-front, i.e. to move the front towards the optimal point (e.g., maximum ATP and NADH), which is unfeasible if the two objective are negatively correlated with one another. The framework is implemented in MATLAB.

Results and Discussion

The genetic and the energy-converting networks of mitochondria and chloroplasts are descended, with little modification, from those of their ancestor bacteria. In this regard, we explore how the optimization and the Pareto front analysis can provide interesting insights into the evolutionary dynamics leading to the formation of organelle compartmentalization in the single- and multi-celled life. Furthermore, the sensitivity and robustness analyses can detect clusters of parameters corresponding to clusters of chemical reactions, which are often found in the cell and reflect the presence of different pathways or membrane-bound organelles. The interplay between optimization, sensitivity and robustness is useful not merely to reach the optimal configuration for the organelles, but also to conduct tentative analyses on their parameters.

We have applied our framework to investigate models of organelles, e.g. mitochondria and chloroplasts. In the mitochondrial model (Bazil et al, 2010), we found that the most sensitive parameters are the Hexokinase max rate and the F_1F_0 ATP synthase activity. In the multi-objective optimization stage we analyzed the ATP-NADH

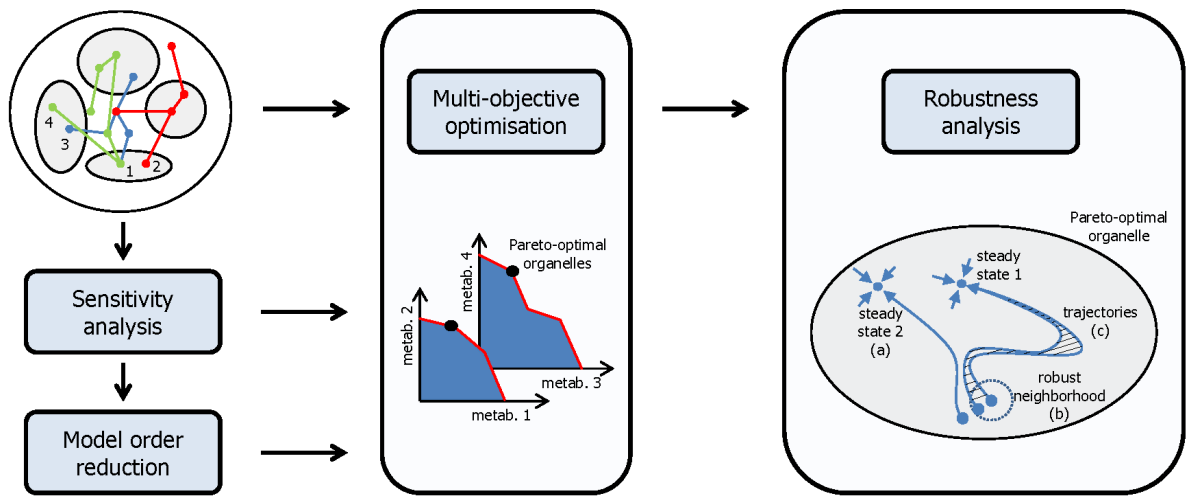


Figure 1. Framework for the rational design of the organelle network in the cell. [Left] First, we analyze the model in its high-dimensional parameter space and evaluate the sensitivity of all its parameters by perturbing them in a neighborhood of the original values. The sensitivity analysis gives both insights into the role of each parameter and hints for a possible model order reduction a technique widely used to reduce the complexity of a given model. [Center] Then, we perform a multi-objective optimization on the organelle metabolism, in order to find the Pareto-optimal front involving two or more metabolites of interest (e.g., ATP and NADH). [Right] Finally, we evaluate the robustness of the Pareto-optimal solutions. Kitano has remarked on the need for a general theory of biological robustness. According to him, a system is robust if it maintains its functionality, even if it transits through a new steady state (a) or if it is unstable. According to (Stracquadanio and Nicosia, 2011), the robustness of a system is the number of robust trials over the total number of trials; a perturbation trial is said to be robust when the perturbation is in the robust neighborhood (b) such that the output remains in a given interval. According to Gunawardena, the robustness to change of initial conditions is called dynamical stability. For instance one can evaluate the differences in the dynamics of the system (c); indeed, as highlighted by (Stelling et al, 2004) robustness can also apply to dynamic processes in development.

Pareto front obtained with several calcium concentrations: if Ca^{2+} increases, we obtain an increase in NADH formation, while ATP remains constant; if Ca^{2+} drastically decreases, there is a lower ATP synthesis. Unexpectedly, with a slow decrease of Ca^{2+} , both objectives are maximized. Our results highlight also that the natural mitochondrion is more robust than the optimized one, as it features a global robustness value of 26.94% and a local value of 9.00%.

In the chloroplast model (Zhu et al, 2007), our framework detected RuBisCO and GAP dehydrogenase as the most sensitive enzymes of the C3 cycle. The Pareto fronts allowed us to find a trade-off between the maximization of the CO_2 uptake rate and the minimization of the nitrogen consumption, with the aim of absorbing more CO_2 while consuming less "leaf-fuel". RuBisCO and PGA kinase are the most robust enzymes.

The functional optimization is not partitioned or delegated to the organelles. The selection is on the phenotype and acts on the whole cell's compartmentalized genomes. Indeed, it is the whole protozoan cell that competes with other

protozoa. The fact that an organelle is kept during the evolution means that its contribution to the overall protozoan's fitness is not marginal, i.e., the presence of the organelle ensures some advantages over losing it. The contribution of each organelle is both to maximize the energy production and to coevolve with the other cell structures, so as to ensure the maximum fitness of the cell. The organelles in a cell play also a key role in the neuronal degeneration. In this regard, neurotransmitters are found in vesicles, i.e. tiny organelles that allow to respond to packets ("units") of neuronal chemical signaling.

Following our framework and extending it, in the near future we plan to design, analyze and optimize the metabolism of systems composed of different species living and interacting in the same organism. In this project we have performed an *in silico* design that can explore the reaction network and seek in the search space the solutions that optimize two or more objectives. Therefore, our approach lies in the field of computational metabolic engineering. This kind of analysis could easily highlight the com-

plementarity of different metabolic networks. For instance, mitochondria and chloroplasts are (usually) both found in plants, and are part of the same functional pipeline: starting from CO₂, the photosynthesis in the chloroplast creates glucose that enters the mitochondria to create ATP.

The model of the whole cell of a human pathogen (Karr et al, 2012) has opened new frontiers in this research field. The whole-cell model refers to the *Mycoplasma genitalium* and consists of 28 submodels accounting for all the biological functions of the cell. We have already applied our methodology to Flux Balance Analysis, Gene-Protein-Reaction associations, Ordinary Differential Equations (ODEs) and Differential Algebraic Equations (DAEs). Hence, our framework is suitable for general purpose or black-box analysis, enabling us to investigate not only the model of metabolism, but also the whole-cell model. Our final goal is to improve our methodology in order to tackle any BioCAD problem.

In each Pareto front we have considered the single organelle, while in the cell there are usually many organelles that could differ for activity depending on their location in the cell. In a network of organelles, most of the reactions involve more than one organelle; an appropriate approach would be to build a Pareto front where each metabolite belongs to a different organelle in the network, linking with a set of Delay Differential Equations (DDEs). DDEs differ from ODEs in that they allow rates of change to depend on the state

of the system at an earlier time. In organelle systems, DDEs could account for diffusion processes and maturation events.

The integrated bio-search for a diseased, perturbed, misfunctional pathway often needs an accurate understanding of the relationships and interactions of that pathway with the organelle network system. The methodology proposed in our work could address most questions emerging in neurodegenerative and cancer disease investigation, which are focused on the interaction between organelle networks and cellular metabolism.

References

1. Bazil J N, Buzzard G T, Rundell A E (2010) Modeling mitochondrial bioenergetics with integrated volume dynamics. *PLoS computational biology*, 6(1):e1000632
2. Karr J R, Sanghvi J C, et al (2012) A Whole-Cell Computational Model Predicts Phenotype from Genotype. *Cell*, 150(2):389-401
3. Shiflett A and Johnson P J (2010) Mitochondrion-related organelles in parasitic eukaryotes. *Annual review of microbiology*, 64:409
4. Stelling J, et al (2004) Robustness of Cellular Functions. *Cell*, 118(6):675-685
5. Stracquadano G and Nicosia G (2011) Computational energy-based redesign of robust proteins. *Computers & chemical engineering*, 35(3):464-473
6. Yoneda Y, et al (2009) Frontier biomedical science underlying organelle network biology, <http://www.fbs.osaka-u.ac.jp/organelle-network/eng/greeting/greeting/>
7. Zhu X G, de Sturler E, Long S P (2007) Optimizing the distribution of resources between enzymes of carbon metabolism can dramatically increase photosynthetic rate. *Plant Physiology*, 145:513-526

Filtering with alignment free distances for high throughput DNA reads assembly

Maria C De Cola^{1,2✉}, Giovanni Felici², Daniele Santoni², Emanuel Weitschek^{2,3}

¹Department of Statistics, University La Sapienza, Rome, Italy

²Institute of Systems Analysis and Computer Science, National Research Council, Rome, Italy

³Department of Informatics and Automation, Università degli Studi Roma Tre, Rome, Italy

Motivation and Objectives

The output of a high throughput next generation sequencing (NGS) machine is a collection of short reads, which have to be properly assembled in order to reconstruct the original DNA sequence of the analyzed organism (Metzker, 2010; Earl, 2011). The DNA sequence assembly process is based on aligning and merging these reads for effectively reconstructing the real primary structure of the DNA sample sequence or reference genome. The use of NGS machines results in much larger sets of reads to be assembled, posing new problems for computer scientists and bioinformaticians. In particular, a relevant issue is related with the trade-off between precision of the assembly process and its computational time, stating the need for faster methods that can keep pace with the speed and volume of reads that are generated with NGS. An important step in DNA assembly is the identification of a subset of read pairs that have a high probability of being aligned sequentially in the reconstruction. Such step is often referred to as filtering, and amounts in selecting a significantly smaller subset of the initial set of read pairs (whose dimension is quadratic in the number of initial reads) that can be then processed by an alignment algorithm, usually quite time consuming. The desired effect of filtering is then to quickly filter out from the candidate set of read pairs those that would not provide a good alignment in the following phase. The computation cost of filtering should then be balanced by the speed-up obtained when a smaller set of read pairs is considered for alignment.

In this work we propose and test the use of alignment free distances to evaluate the similarity between two short reads as a technique for filtering good read pairs to be assembled.

The method operates in constant time in the string length and is tested in its ability to emulate, with a proper level of precision, much more

time consuming methods to evaluate the similarity between short DNA sequences, such as the established Needleman-Wunsch edit distance (Needleman, Wunsch, and Christian, 1970), often used in the final step of the assembly procedure. These preliminary experiments show the efficacy of this approach for filtering the promising read pairs - eligible candidates to successfully assemble the entire genome of a given organism. Therefore, the alignment free reads filtering may significantly accelerate the assembly process without a substantial loss in accuracy of the DNA sample sequence reconstruction.

Methods

ODNA sequence assembly

The DNA sequence assembly process is based on the alignment and merging of reads (stretch of sequences) in order to reconstruct the original primary structure of the DNA sample sequences. Given a set of sequences $S = \{s_1, s_2, \dots, s_n\}$, where $s \in S$ is a fragment of the primary structure of DNA (read) (e.g. $s = \{\text{ATTCGA...CTGACT}\}$), assembly is in charge of building the longest sequence from the set S where each pair of consequent reads obey certain similarity conditions.

DNA read pairs filtering and Alignment Free Distance

This step identifies the promising read pairs in order to reduce the amount of input data given to the real assembly algorithm. We adopted a very quick measure of the similarity between two reads, Alignment Free (AF) based distance (Vinga and Almeida, 2003). AF computes the similarity of two strings based only on the dictionary of their substrings, irrespective of their relative position. As a dictionary we considered the set of 4-mer (sequences composed of 4 different nucleotides) and then built a profile for each read composed by the relative frequencies of each 4-mer in the read. The Euclidean distance between the profiles of two reads was taken as

an inverse measure of the similarity of the two reads and thus as an indication that the two reads formed a promising pair to be considered in the assembly phase. AF filtering was then used defining a proper threshold on the AF distance and discarding all the pairs that exhibited a AF distance above the threshold. Computational complexity of AF distance is a constant linearly bounded by the number of k-mers adopted and the length of the strings to be compared.

Comparing with other distances: Needleman-Wunsch and "Bowtie" distance

Along with AF we considered the well-established Needleman-Wunsch edit distance (NW) and compared them in their ability to identify significant pairs. This comparison was based on the computation of a sort of perfect distance computed after an alignment over an already known sequence has been performed. Such distance, referred to as Bowtie distance (BT), was obtained as follows:

- a large number of reads coming from a known sequence were considered;
- these reads were aligned over the known sequence using the standard Bowtie algorithm (Langmead et al, 2009);
- any two reads received a maximum BT distance if their alignment did not intersect over the reference sequence, else they received a distance inversely proportional to their intersection over the sequence (e.g., they would have BT distance equal to 0 if they were

aligned one on top (or inside) of the other by the Bowtie algorithm). By construction we assumed BT distance to be the reference distance, e.g., the distance that expressed the best possible alignments - being based on the knowledge of the reference sequence - and tested the correlation of AF and NW with BT; moreover, we verified the ability of AF and NW to predict that a given read pair had BT distance above or below a given threshold.

Results and Discussion

For our test we considered the E.Coli genome and a set of reads from this genome reads obtained by Roche 454 sequencing machine. Reads have average length of ~235 nucleotides and standard deviation of approx. 10 (the large majority of them having length in the interval 225-245). Reads were aligned with the reference sequence with Bowtie and then 100,000 were sampled at random according to their alignment along the sequence. Reads were considered both forward and reversed, giving rise to a total of 200,000² read pairs. All 620,798 read pairs with BT distance < 1 were considered for the experiments; then, out of the remaining pairs, 233,099 were sampled at random. A total of 853,897 read pairs composed the working data set. For all these reads, NW distance and AF distance over the 4-mer were computed. AF, NW and BT distances were all normalized between 0 (maximal similarity) and 1 (maximal dissimilarity). The first interesting results was that the correla-

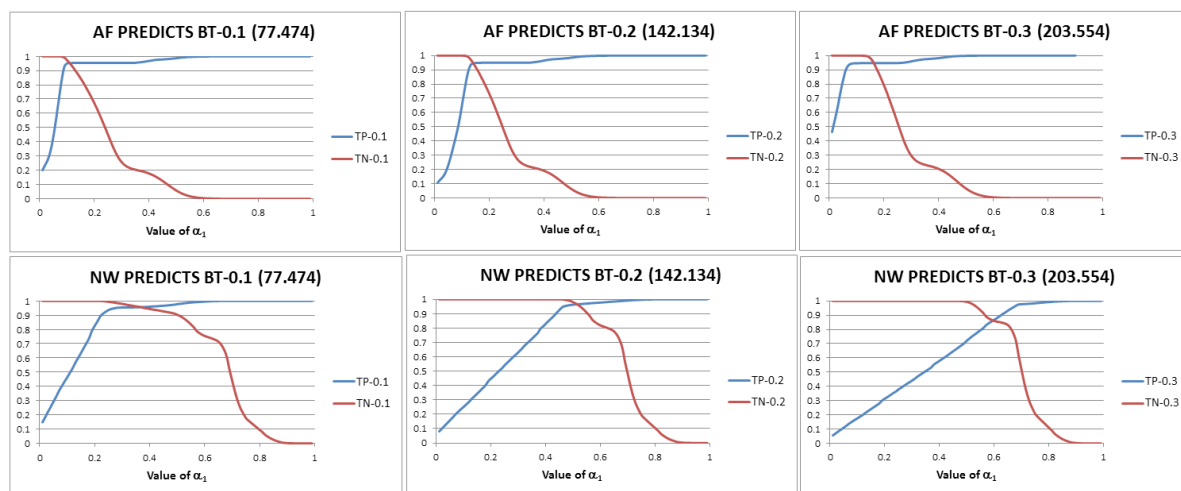


Figure 1. Error curves for predictors of BT. Error rates of threshold predictors for BT based on AF are plotted in the charts of the first row; predictors for BT based on AF are in the second row; blue lines represent True positive rates, red lines represent true negative rates. robustness can also apply to dynamic processes in development.

tion between distances showed that AF approximates BT somehow better than NW: we obtained a correlation coefficient of 0.761 for AF and BT, compared with a smaller 0.706 when NW and BT were considered (coherently, correlation between AF and NW is 0.721). The second interesting results was obtained when we compared the ability of AF and NW to predict whether BT was above or below a given threshold. We defined a threshold predictor for a given function F_2 based on function F_1 and on a given pair α_1, α_2 as follows: if $(F_1 < \alpha_1)$ then predict $(F_2 < \alpha_2)$, else predict $(F_2 \geq \alpha_2)$. To a given pair (α_1, α_2) , we associated the measure of True Positive rate (TP) (percentage of cases where $(F_1 < \alpha_1)$ and $(F_2 < \alpha_2)$) and of True Negative rate (TN) (percentage of cases where $(F_1 \geq \alpha_1)$ and $(F_2 \geq \alpha_2)$); analogously we defined False Positive rate (FP) and False Negative rate (FN).

For each (α_1, α_2) with both values ranging from 0 to 1, we then computed, with step 0.05, the positive and negative error rates taking AF as a predictor of BT and NW as a predictor of BT. Part of the results are summarized in the charts of Figure 1, that show for 3 different levels of α_2 (0.1, 0.2, and 0.3) the precision of the predictors (y-axis) when the value of α_1 is changed (x-axis), both when AF is used as a predictor of BT (charts in the first row) and when NW is used as a predictor of BT (charts in second row). Similar results are obtained also

for other levels of α_2 , here omitted for brevity. The curves bring to light very clearly how AF is a very good threshold predictor for BT for the considered data; despite its light computational complexity, it appears to perform significantly better than the more complex NW edit distance when its ability to support a threshold predictor is considered.

Acknowledgements

The authors are partially supported by the FLAGSHIP "InterOmics" project (PB.P05) funded by the Italian MIUR and CNR institutions, and by the cooperative programme 2010–2012 between the National Research Council of Italy (CNR) and the Polish Academy of Sciences (PAN).

References

1. Earl D et al (2011); Assemblathon 1: A competitive assessment of de novo short read assembly methods; *Genome Research*, 21
2. Langmead B, Trapnell C, et al (2009); Ultrafast and memory-efficient alignment of short DNA sequences to the human genome; *Genome Biology*, 10:R25
3. Metzker ML (2010); Sequencing technologies — the next generation; *Nat Rev Genet.*, 11(1)
4. Needleman SB, Wunsch CD. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* 48(3): 443–53
5. Vinga S, Almeida J (2003); Alignment-free sequence comparison—a review, *Bioinformatics* 19 (4), 513-523

A semantic collaborative system for the management of translational research projects

Matteo Gabetta^{1✉}, Giuseppe Milani¹, Cristiana Larizza¹, Valentina Favalli², Eloisa Arbustini², Riccardo Bellazzi¹

¹Dipartimento di Ingegneria Industriale e dell'Informazione, University of Pavia, Pavia, Italy

²IRCCS Fondazione Policlinico S. Matteo, Pavia Italy

Motivation and Objectives

Translational research projects aim at combining -omics, structural and functional studies with clinical investigation results to translate basic knowledge of genetic diseases into routine clinical practice. Biomedical informatics can fruitfully support this kind of research by implementing information technology solutions to support the multidisciplinary project team in the different phases of its investigation.

In this paper we present a semantic wiki-based system purposely implemented for supporting the consortium members of the EU project Inheritance in sharing and disseminating data and knowledge about genetic dilated cardiomyopathies (DCM) [Ahamad et al, 2005]. It consists of a collaborative system that is used to track project activities, share ideas and data, foster exchange of information between the investigators to support several activities of the INHERITANCE translational research project. Moreover, it can be used to easily manage the scientific research products by adding semantic tags on the basis of the underlying knowledge model. A Natural Language Processing (NLP) based module has been developed to this aim; it extracts the relevant molecular and medical concepts from the scientific material shared by the project team and store them as RDF form by enabling the semantic querying of data

Methods

The INHERITANCE Project's Semantic Wiki has been designed and implemented for two purposes: to manage in a collaborative and fast shareable way information and documents related to the organizational aspects of the project and to allow users to share scientific documents automatically analysed and annotated thanks to an integrated NLP based tool.

To build such a Wiki we choose to extend the standard MediaWiki [web site: <http://www.mediawiki.org/wiki/MediaWiki>], last accessed

on July 27, 2012) platform with its most popular semantic extension, called Semantic MediaWiki [Krotzsch et al, 2006].

The first step of the environment setup consisted of defining the Categories necessary to model the information managed inside the Wiki, and the Templates and Forms, which are required to define the content of each category.

In the first release of the Wiki we have implemented the "Person", "Organization", "Meeting" and "Work Package" Categories to represent the organizational aspects of the project, and the "Protein", "Gene" and "Dilated Cardiomyopathy Documents" Categories to model the scientific aspects.

In the typical system use case the authorized users manually insert the organizational data using the proper Templates and Forms; these information will be available for any further interrogation with the smart querying tools available in the Wiki. The main reason for not implementing an automatic import process of these data from the project material is their actual nature: indeed they are spread among many different documents, but their relatively small number doesn't justify the presence of an automatic extraction tool.

Differently, the scientific knowledge management section of the Wiki is designed to deal with an arbitrary large number of documents; therefore we implemented, on top of the Wiki, a concept extraction system able to: a) let the user upload a document (in plain text, pdf or MS Word format) and choose the name of the Wiki page where the document will be stored; b) extract genes and proteins cited inside the document, recursively checking if the gene/protein is already present in the Wiki (otherwise a page for the new gene/protein is created) and link these pages to the one containing the document; c) add the page representing the document to the Wiki.

To realize such a solution we designed a servlet directly accessible from a special page of the Wiki called "NLP"; the concept extraction module

of the servlet is based on Gate [H. Cunningham, 2002], an open-source library for natural language processing. This tool combines a standard (and already implemented) text analysis pipeline with some modules purposely developed in order to extract the cited genes (exploiting the Entrez Gene NCBI's database [Maglott et al, 2005]) and proteins (exploiting Uniprot [The UniProt Consortium, 2012]).

In addition, when a new page representing a gene or a protein is created, the system, thanks to the NCBI Entrez Programming Utilities tools [web site: <http://www.ncbi.nlm.nih.gov/books/NBK25500/>] (last accessed on July 27, 2012), automatically associates to the page the five most recent articles from Pubmed that have that gene/protein as topic.

Once the Wiki has been populated with the project's data, it is possible to perform, beyond all the standard tasks of a traditional Wiki (update, content modification, old pages restore, discussion, etc.), also some smart querying operations that exploit the semantic nature of the

data. The semantic query tools available in the Wiki use two distinct languages: a simple query language, to perform queries within the Wiki's data, and SPARQL [Herman, 2008] that is the standard query language for the semantic web, opening the Wiki to the possibility of a future integration with many other available repositories of linked data [web site: <http://linkeddata.org/>] (last accessed on July 27, 2012).

Results and Discussion

Actually, the INHERITANCE semantic wiki is up and running at the URL http://www.labmedinfo.org:8123/mediawiki/index.php/Main_Page and is made available to all the consortium members to track the project activities (meetings, partners, work packages) and manage every product of the project (deliverables, scientific papers). A Summary page has been defined to synthesize all the project activities and participants information. Moreover, the RelFinder browser [<http://www.visualdataweb.org/relfinder.php>] (last accessed on July 30, 2012), useful to look for rela-

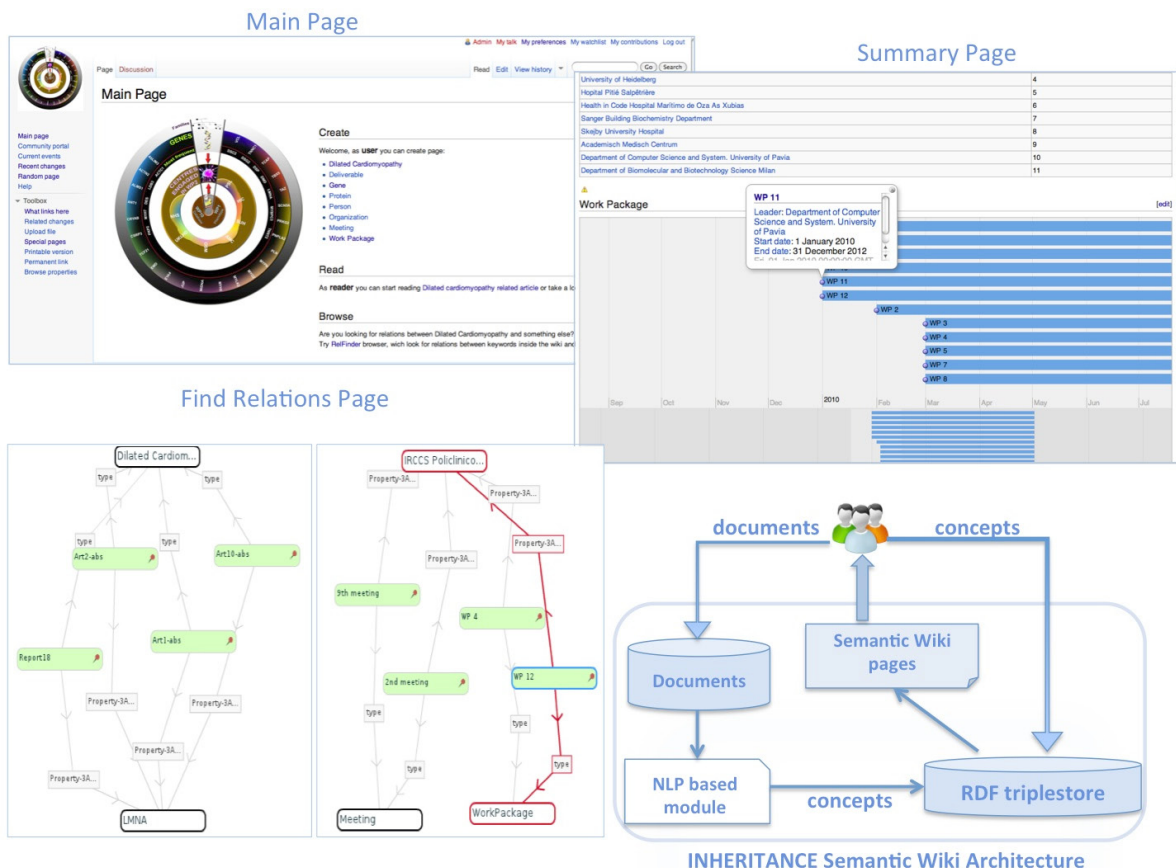


Figure 1 – The Semantic Wiki architecture and the Main, Summary and Find Relations project pages screenshots

tions between keywords inside the wiki and show the relations graph (eg. Person- Organization or Meeting-Organization relations), has been made available (Figure 1).

Currently the main goal of the semantic wiki is to support the INHERITANCE research group from two distinct points of view: the organizational and the scientific data management and sharing. While all the features related to the organizational aspects have been developed and tested by the users, the scientific knowledge management section of the wiki is still under development. The current prototype provides some basic features such as the scientific documents storage and mapping to custom categories, the NLP facilities for data extraction and the automatic linkage to relevant scientific literature. Nonetheless the upgrade of the system with new tools (e.g. link to specific DCM resources and integration with biological databases) doesn't entail relevant technical problem, and its actual implementation, although planned, depends on the future developments of the INHERITANCE project and on the users' feedback after the system evaluation.

At this moment the NLP based module has been used to annotate 10 documents and extract 13 genes and 10 proteins. In future we plan

to link the data to external resources from across the Linked Data community.

Acknowledgements

This work is part of the INHERITANCE Project, funded by the European Commission.

References

1. Ahamad F, Seidman JG, Seidman CE. (2005) The genetic basis of cardiac remodelling. *Annu Rev Genomics. Hum Genet* 6, 185. doi: 10.1146/annurev.genom.6.080604.162132
2. H. Cunningham, D. Maynard, et al (2002) GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia.
3. Maglott D, Ostell J, Pruitt KD, Tatusova T. (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*, 33 (Database Issue):D54-8
4. Herman, W3C Semantic Web Activity News - SPARQL is a Recommendation, http://www.w3.org/blog/SW/2008/01/15/sparql_is_a_recommendation/W3.org. 2008-01-15. (Last accessed on July 27, 2012)
5. The UniProt Consortium. (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, 40, D71-D75.
6. Krotzsch, M., Vrandečić, D. and Volkel, M. 2006. Semantic MediaWiki. Proceedings of the Fifth International Semantic Web Conference, pp 935-942, Springer, November 2006.

Answering Gene Ontology terms to proteomics questions by supervised macro reading in Medline

Julien Gobeill¹, Emilie Pasche², Douglas Teodoro², Anne-Lise Veuthey³, Patrick Ruch²

¹University of Applied Sciences, Information Sciences, Geneva

²Hospitals and University of Geneva, Geneva

³Swiss-Prot group, Swiss Institute of Bioinformatics, Geneva

Motivation and Objectives

Biomedical professionals have at their disposal a huge amount of literature. But when they have a precise question, they often have to deal with too many documents to efficiently find the appropriate answers in a reasonable time. Faced to this literature overload, the need for automatic assistance has been largely pointed out, and PubMed is argued to be only the beginning on how scientists use the biomedical literature (Hunter and Cohen, 2006).

Ontology-based search engines began to introduce semantics in search results. These systems still display documents, but the user visualizes clusters of PubMed results according to concepts which were extracted from the abstracts. GoPubMed (Doms and Schroeder, 2005) and EBIMed (Rebholz-Schuhmann et al, 2007) are popular examples of such ontology-based search engines in the biomedical domain. Question Answering (QA) systems are argued to be the next generation of semantic search engines (Wren, 2011). QA systems no more display documents but directly concepts which were extracted from the search results; these concepts are supposed to answer the user's question formulated in natural language. EAGLi (Gobeill et al, 2009), our locally developed system, is an example of such QA search engines.

Thus, both ontology-based and QA search engines, share the crucial task of efficiently extracting concepts from the result set, i.e. a set of documents. This task is sometimes called macro reading, in contrast with micro reading – or classification, categorization – which is a traditional Natural Language Processing task that aims at extracting concepts from a single document (Mitchell et al, 2009).

This paper focuses on macro reading of MEDLINE abstracts. Several experiments have been reported to find the best way to extract ontology terms out of a single MEDLINE abstract, i.e. micro reading. In particular, (Trieschnigg et al,

2009) compared the performances of six classification systems for reproducing the manual Medical Subject Headings (MeSH) annotation of a MEDLINE abstract. The evaluated systems included two morphosyntactic classifiers (sometimes also called thesaurus-based), which aim at literally finding ontology terms in the abstract by alignment of words, and a machine learning (or supervised) classifier, which aims at inferring the annotation from a knowledge base containing already annotated abstracts. The authors concluded that the machine learning approach outperformed the morphosyntactic ones. But the macro reading task is fundamentally different, as we look for the best way to extract then combine ontology terms from a set of MEDLINE abstracts.

The issue investigated in this paper is: to what extent the differences observed between two classifiers for a micro reading task are observed for a macro reading one? In particular, the redundancy hypothesis claims that the redundancy in large textual collections such as the Web or MEDLINE tends to smoothen performance differences across classifiers (Lin, 2007). To address this question, we compared a morphosyntactic and a machine learning classifiers for both tasks, focusing on the extraction of Gene Ontology (GO) terms, a controlled vocabulary for the characterization of proteins functions. The micro reading task consisted in extracting GO terms from a single MEDLINE abstract, as in the Trieschnigg et al's work; the macro reading task consisted in extracting GO terms from a set of MEDLINE abstracts in order to answer to proteomics questions asked to the EAGLi QA system.

Methods

We evaluated two statistical classifiers which were both studied in the Trieschnigg et al's work. The morphosyntactic classifier was EAGL. It is described comprehensively in (Ruch, 2006). It showed very competitive results when it was compared to other state-of-the-art morphosyntactic

classifiers, as during the official BioCreative I evaluation (Blaschke et al, 2005) or in the Trieschnigg et al's work against Metamap (Aronson and Lang, 2010). The machine learning classifier was a k-NN. The k-NN is a remarkably simple and scalable algorithm which assigns to a new abstract the GO terms that are the most prevalent among the k most similar abstracts contained in a knowledge base (Manning and Schütze, 1999). The knowledge base was designed from the GOA database, which contains 85'000 manually curated abstracts and is available at <http://www.ebi.ac.uk/GOA/>. Last accessed on August 1st, 2012). These abstracts were indexed with a classical Information Retrieval engine (Ounis et al, 2006) and, for each input text, the k=100 most lexically similar ones were retrieved in order to infer the GO terms.

For the micro reading task, we designed a so called GOA benchmark of one thousand MEDLINE abstracts sampled from the GOA database; the classifiers were evaluated on their ability to extract the GO terms that were manually associated with these abstracts by the GOA experts. For the macro reading task, we designed two benchmarks of fifty questions by exploiting two biological databases: the Comparative Toxicogenomics Database (CTD) contains more than 2'800 chemicals annotated with GO terms, and is available at <http://ctdbase.org/> (Last accessed on August 1st, 2012); the UniProt database contains millions of proteins annotated with GO terms, and is available at <http://www.uniprot.org/> (Last accessed on August 1st, 2012). Questions were sampled from these databases and dealt with molecular functions and a given chemical compound, such as "what molecular functions are affected by Aminophenols ?", or cellular components and a given protein, such as "what cellular component is the location of NPHP1?". The classifiers were successively embedded in the EAGLi's QA engine for extracting GO terms from a set of one hundred MEDLINE abstracts retrieved by EAGLi for each question. The most prevalent GO terms extracted from these abstracts were then proposed as answers by the QA engine. Please refer to (Gobeill et al, 2009) for a deeper description of EAGLi. Thus, their evaluation was extrinsic and was based on their ability to extract GO terms from a set of abstracts and then provide to EAGLi the answers contained in the databases.

There were on average 2.8 GO terms per abstract to return in the GOA benchmark, and

30/1.3 GO terms per question to find (literally to answer) for respectively the CTD/UniProt benchmark. As both categorizers output a ranked list of candidate GO term, we chose metrics from the Information Retrieval domain that were well-established during the TREC campaigns (Voorhees et al, 2001). For precision considerations, we computed the Mean Reciprocal Rank (MRR) which is the multiplicative inverse of the rank of the first correct outputted GO term.

Results and Discussion

For the micro reading task (i.e. extracting GO terms from a single abstract), as in the Trieschnigg et al's work with MeSH classification, the machine learning classifier (k-NN) outperforms the morphosyntactic one (EAGL). For the macro reading task (i.e. extracting GO terms from a set of abstracts), for both benchmarks, the k-NN also outperforms EAGL, and the observed differences in top-precision are similar and consistent with the micro-reading task. These results weaken the redundancy hypothesis, as the performance of classifiers for micro reading tasks appears to be of importance for macro reading tasks.

It is worth observing that, unlike other text mining tasks, Information Retrieval and Question Answering have been largely resisting to machine learning advances (Athenikosa and Hanb, 2009). Ontology-based search engines powered with morphosyntactic classifiers could benefit from such a new component, as it allows to inject knowledge contained in curated databases in the result set. This could provide promising research pathways for the biomedical data mining community.

Beyond comparisons, our QA engine with supervised macro reading in MEDLINE achieved a top-precision ranging from 0.58 to 0.69 to answer

Table1: top-precision for both GO classifiers observed in micro reading then macro reading tasks, along with the percentage of improvement with the k-NN.

	Micro reading task	Macro reading task	
	GOA benchmark	CTD benchmark	UniProt benchmark
EAGL	0,23	0,34	0,33
k-NN	.48 +109%	.69 +103%	.58 +76%

proteomics questions. This performance allows its users to save time on consulting the literature, as well as to automatically produce function predictions for massive proteomics datasets, such as in (Anton et al, 2012). EAGLi is available at <http://eagl.unige.ch/EAGLi/> (Last accessed on August 1st, 2012).

Acknowledgements

Work supported by the Swiss National Fund for Scientific Research [BiND project 3252B0-105755].

References

1. Anton BP, Chang YC, et al (2012) COMBEX: Design, Methodology, and Initial Results. Manuscript submitted for publication.
2. Aronson AR and Lang FM (2010) An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc.* 17(3), 229. doi:10.1136/jamia.2009.002733
3. Athenikosa S and Hanb H (2009) Biomedical question answering: A survey. *Comput Methods Programs Biomed.* 99(1), 1. doi:10.1016/j.cmpb.2009.10.003
4. Blaschke C, Leon EA, et al (2005) Evaluation of BioCreAtive assessment of task 2. *BMC Bioinformatics* 6(Suppl 1):S16. doi:10.1186/1471-2105-6-S1-S16
5. Doms A and Schroeder M (2005) GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Res.* 1(33), 783. doi:10.1093/nar/gki470
6. Gobeill J, Pasche E, et al (2009) Question answering for biology and medicine. *Information Technology and Application in Biomedicine, Larnaca, Cyprus.*
7. Hunter L and Cohen KB (2006) Biomedical language processing: what's beyond PubMed? *Mol Cell.* 21(5), 589. doi: 10.1016/j.molcel.2006.02.012
8. Lin J (2007) An exploration of the principles underlying redundancy-based factoid question answering. *ACM Trans. Inf. Syst.* 25(2). doi: 10.1145/1229179.1229180
9. Manning CD and Schütze H (1999) *Foundations of Statistical Natural Language Processing.* Cambridge, MA, MIT Press. doi:10.1023/A:1011424425034
10. Mitchell TM, Betteridge J, et al (2009) Populating the Semantic Web by Macro-reading Internet Text. *Proceedings of the 8th Intern. Semantic Web Conf.* doi: 10.1007/978-3-642-04930-9_66
11. Ounis I, Amati G, et al (2006) Terrier: A High Performance and Scalable Information Retrieval Platform. *Proceedings of ACM SIGIR'06 Workshop.*
12. Rebholz-Schuhmann D, Kirsch H, et al (2007) EBIMed--text crunching to gather facts for proteins from Medline. *Bioinformatics* 23(2), 237. doi:10.1093/bioinformatics/btl302
13. Ruch P (2006) Automatic assignment of biomedical categories: toward a generic approach. *Bioinformatics* 22(6), 658. doi:10.1093/bioinformatics/btl783
14. Trieschnigg D, Pezik P, et al (2009) MeSH Up: effective MeSH text classification for improved document retrieval. *Bioinformatics* 25(11), 1412. doi:10.1093/bioinformatics/btp249
15. Voorhees E (2001) Overview of the QA Track. In *Proceedings of the TREC-10 Conference.* NIST, Gaithersburg. 2001:157-165.
16. Wren JD (2011) Question answering systems in biology and medicine--the time is now. *Bioinformatics* 27(14). doi:10.1093/bioinformatics/btr327

Using graph theory to analyze gene network coherence

Francisco Gómez-Vela[✉], Norberto Díaz-Díaz, José A Lagares, José A Sánchez, Jesús S Aguilar-Ruiz

School of Engineering, Pablo de Olavide University, Seville

Motivation and Objectives

Gene networks (GNs) have become one of the most important approaches for modelling gene-gene relationships in Bioinformatics (Hecker et al, 2009). These networks allow us to carry out studies of different biological processes in a visual way.

Many GN inference algorithms have been developed as techniques for extracting biological knowledge (Ponzoni et al, 2007; Gallo et al, 2011). Once the network has been generated, it is very important to assure network reliability in order to illustrate the quality of the generated model. The quality of a GN can be measured by a direct comparison between the obtained GN and prior biological knowledge (Wei and Li, 2007; Zhou and Wong, 2011). However, these both approaches are not entirely accurate as they only take direct gene-gene interactions into account for the validation task, leaving aside the weak (indirect) relationships (Poyatos, 2011).

In this work the authors present a new methodology to assess the biological coherence of a GN. This coherence is obtained according to different biological gene-gene relationships sources. Our proposal is able to perform a complete functional analysis of the input GN. With this aim, graph theory is used to consider not only direct relationships but indirect ones as well.

Methods

The aim of our proposal is to evaluate the functional coherence of an input GN. The coherence is calculated according to current gene-gene interaction knowledge which is stored in public biological databases (DB). Thus, graph theory is applied with the aim of considering all gene-gene relationships (i.e. direct and indirect relationships) presented in the Input Network (IN).

Our approach works in various steps. First, the IN and the DB are converted into distance matrices (DM) using Floyd-Warshall algorithm (Asghar et al, 2012). This approach is a graph analysis method that solves the shortest path problem. This algorithm uses an adjacency matrix to com-

pute the minimum path for every pair of genes. In this sense, the shortest path between two vertices is computed by incrementally improving an estimate on the shortest path between those vertices, until the estimate is optimal. Hence, the minimum distance of all gene pair combinations are computed and stored in DMin and DMdb, respectively. Furthermore, a distance threshold (δ) is used to exclude relationships that lack biological meaning. This threshold denotes the maximum distance to be considered as relevant in the DM generation process. Thus, if the minimum distance between two genes is greater than δ , then no path between the genes will be assumed.

Once the distance matrices have been obtained, they are combined to generate a new one. The new matrix, hereafter called Coherence Matrix (CM), contains the existing gap between the common genes in either the DMin and the DMdb.

$$CM = |DM_{IN} - DM_{DB}|$$

Where $CM(i,j) = |DMin(i,j) - DMdb(i,j)|$ denotes the coherence of relationship between gene g_i and gene g_j with regard to the information stored in DB. Note that, relationships between genes within IN and DB will be only considered to generate CM. It is not possible to establish the quality of the rest of the relationships. DB contains no information to ascertain whether the relationships are biologically relevant or not.

According to the coherence values stored in CM and to an accuracy coherence level (θ), the differences and similarities between the GN and DB could be obtained. The differences are classified as false positives and false negatives, while the similarities as true positives and true negatives. Therefore, if $CM(i,j)$ is greater than θ it will be considered as a false positive, while if it is less than or equal to θ , it will be computed as true positive. In case there is no path between g_i and g_j in the IN, neither in DB ($IN(i,j)=DB(i,j)=infinite$), it will be considered as a true negative. Nevertheless, if there is no path in IN but there is in DB, it will be computed as a false negative.

Table 1: F-Measure and Accuracy values obtained by different input GN according to prior biological knowledge and chronologically sorted. The best results in each dataset are emphasized.

	Soinov		Bulashevskaja		Ponzoni (GRNCOP)		Gallo (GRNCOP2)	
	F-Measure	Accuracy	F-Measure	Accuracy	F-Measure	Accuracy	F-Measure	Accuracy
BioGrid	0,42	0,27	0,79	0,65	0,9	0,82	0,86	0,75
KEGG	0,48	0,58	0,5	0,34	0,43	0,28	0,61	0,47
SGD	0,47	0,31	0,69	0,53	1	1	0,73	0,58
YeastNet	0,45	0,29	0,66	0,5	1	1	0,77	0,62

Results and Discussion

In order to assess the robustness of our proposal, we present a set of analysis of different yeast cell cycle networks using four prior biological knowledge data sets.

Input networks were produced applying four inference network techniques (Soinov et al, 2003; Bulashevskaja and Eils 2005; Ponzoni et al, 2007; Gallo et al, 2011) on the well-known yeast cell cycle expression data set (Spellman et al, 1998). Finally, the functional coherence of GNs generated is measured using our proposal according to the gene-gene interaction knowledge stored in BioGRID (Stark et al, 2010), KEGG (Kanehisa et al, 2012), SGD (Cherry et al, 2012) and YeastNet (Lee et al, 2007).

Multiple studies were carried out using different threshold value combinations. δ and θ have been modified from one to five, generating 25 different combinations. The results show that the higher δ values, the greater is the noise introduced. Coherence level threshold (θ) shows similar behavior; the lower θ , the smaller is the noise. The most representative result, summarized in Table 1, was obtained for $\delta=4$ and $\theta=1$. This combination has a biological meaning. For each gene, only the interactions in a radius of four should be considered as relevant. Moreover, they ought to have a difference no greater than 1 to be considered as valid.

Table 1 shows that inference method proposed by Gallo (GRNCOP2) generates the most reliable result, although Ponzoni technique (GRNCOP) provides the best result in three of the four data sets. Soinov approach obtains the worst values.

These results are consistent with the experiment carried out in (Ponzoni et al, 2007) and (Gallo et al, 2011). GRNCOP was successfully

compared with Soinov and Bulashevskaja approaches, while Gallo et al presented a detailed analysis of the performance of GRNCOP and GRNCOP2, where the last one shows the most stable result. These behaviors are also found in the obtained results. GRNCOP presents better coherence values than Soinov and Bulashevskaja in BioGrid, SGD and YeastNet. Similarly, GRNCOP2 obtains more stable values than GRNCOP, especially for F-measure.

References

1. Asghar A, et al (2012) Speeding up the Floyd-Warshall algorithm for the cycled shortest path problem. *Applied Mathematics Letters* 25(1): 1
2. Bulashevskaja S and Eils R (2005) Inferring genetic regulatory logic from expression data. *Bioinformatics* 21(11):2706.
3. Cherry JM, et al (2012) Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Research* 40: D700-705. doi:10.1093/nar/gkr1029.
4. Gallo C, et al (2011) Discovering time-lagged rules from microarray data using gene profile classifiers. *BMC Bioinformatics* 12:123.
5. Hecker M, et al (2009) Gene regulatory network inference: Data integration in dynamic models – a review. *Biosystems* 96:86.
6. Kanehisa M, et al (2012) KEGG for integration and interpretation of large-scale molecular datasets. *Nucleic Acids Research* 40:D109-D114
7. Lee I, et al (2007) An improved, bias-reduced probabilistic-functional gene network of baker's yeast, *Saccharomyces cerevisiae*. *PLoS ONE* 2(10):e988.
8. Ponzoni I, et al (2007) Inferring adaptive regulation thresholds and association rules from gene expression data through combinatorial optimization learning. *IEEE/ACM Transaction on Computation Biology and Bioinformatics* 4(4):624.
9. Poyatos JF (2011). The balance of weak and strong interactions in genetic networks. *PLoS One* 6(2):e14598.
10. Soinov L, et al (2003) Toward reconstruction of gene networks from expression data by supervised learning. *Genome Biology* 4:R6.

11. Spellman PT, et al (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell* 9(12):3273.
12. Stark C, et al (2010) The BioGRID Interaction Database:2011 update. *Nucleic Acids Research* 39 (Database issue):D698
13. Wei Z and Li H (2007). A Markov random field model for network-based analysis of genomic data. *Bioinformatics* 23(12):153
- Zhou H and Wong L (2011). Comparative analysis and assessment of *M.tuberculosis* H37Rv protein-protein interaction datasets. *BMC genomics*, 12 (Suppl 3):S20
14. Zhou H and Wong L (2011). Comparative analysis and assessment of *M. tuberculosis* H37Rv protein-protein interaction datasets. *BMC genomics*, 12 (Suppl 3):S20.

The open source ISA software suite and its international user community: knowledge management of experimental data

Alejandra González-Beltrán[✉], Eamonn Maguire, Philippe Rocca-Serra, Susanna-Assunta Sansone

¹University of Oxford, Oxford e-Research Centre, Oxford, United Kingdom

Motivation and Objectives

Both in academia and industry, data generation is currently in the order of petabytes in the biomedical domain. The availability of this massive amount of data brings with it many challenges, especially when considering data sharing and integration aiming at a later re-use. In this context, the adoption of standard formats, minimum information guidelines and terminologies/ontologies for the rich annotation of experimental data is crucial. Annotation is a time-consuming task that must be supported by software tools, which should also enable querying, linking, integrating, reasoning and analysing the data as well as the information about it.

The Investigation/Study/Assay (ISA) infrastructure (Rocca-Serra et al 2010) aims at facilitating this rich description of heterogeneous experimental data and supporting the different steps of the data management workflow. The infrastructure revolves around a general-purpose file format (ISA-Tab) and includes an open source software suite supporting compliance with community standards and dealing with the harmonization of the experimental metadata. The ultimate goal is to allow for the gradual progression from unstructured, usually non-digital metadata

kept in lab notebooks to structure data that can be interpreted by machines (see Figure 1). The success of the ISA infrastructure is evidenced by the growing ISA Commons community (Sansone et al 2012), which encompasses increasingly diverse domains varying from metabolomics, (meta)genomics, proteomics, system biology to environmental health, environmental genomics and stem cell discovery (Ho Sui et al 2012).

We will present the components of the ISA infrastructure, the rationale behind them and their evolution. In particular, we will introduce our efforts to expand the infrastructure into three important directions: collaboration in a cloud environment, support for analysis with R, and the semantic web world. We will show use cases to exemplify the usage of the ISA infrastructure.

Methods

The ISA infrastructure software suite is the first one to support both experimentalists and curators in the description of multi-assay experiments (Rocca-Serra et al 2010). Studies using high-throughput (post)genomic technologies may involve multiple assays. For example, a system biology study in yeast (Castrillo et al 2007) includes transcription, metabolite and protein profiling using DNA microarray, NMR spectroscopy and mass spec-

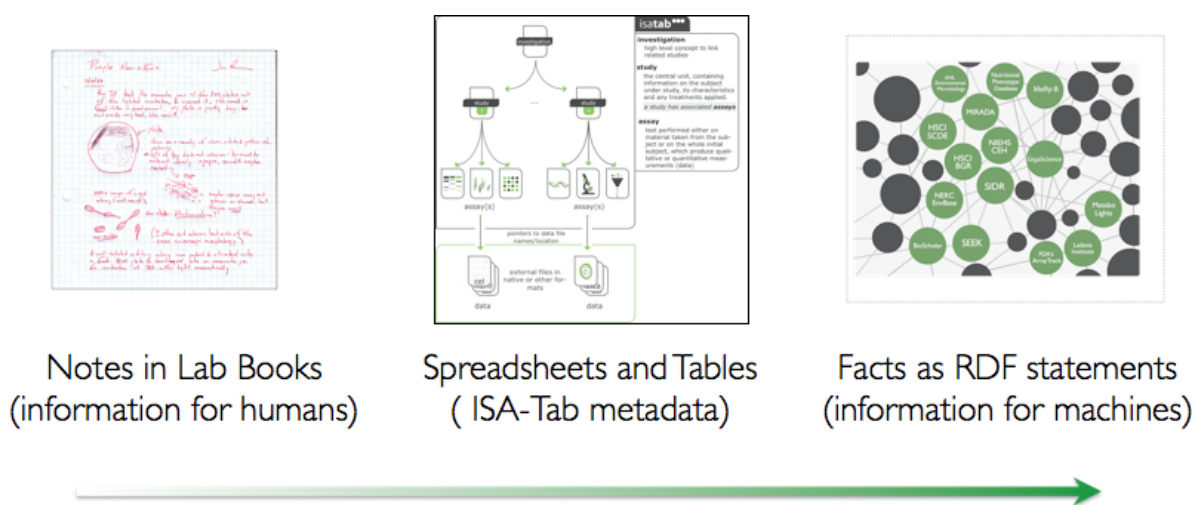


Figure 1. Experimental information with increasing level of structure.

troscopy, respectively.(BII-S-1: <http://www.ebi.ac.uk/bioinvindex/study.seam?studyId=BII-S-1>, (last accessed on 23rd July 2012).

The ISA-Tab format was designed to be domain-agnostic and includes the description of the experiments' contextual information such as the samples characteristics, the technology and measurement types, the instruments' parameters used, among other things. The availability of this metadata is crucial for the reproducibility of the experiments and posterior data re-use", i.e. add 'the' in 'for the reproducibility.

The ISA tools are open source (<http://isa-tools.org/>, GitHub: <https://github.com/ISA-tools>, (last accessed on 23rd July 2012) and follow a modular architecture, with standalone components providing functionality for each step of the data management workflow: *ISAcreeator* offers spreadsheet-based data acquisition and curation relying on BioPortal services (<http://bioportal.bioontology.org>); *ISAconfigurator* enables conformance to reporting standards (data formats, minimum information checklists, terminologies/ontologies); the *Bioinvestigation Index* provides for storage in a searchable repository with configurable access; *ISAconverter* facilitates reformatting for a growing number of acceptable formats; the tools enable submission of data to public repositories, validation and visualization.

Community engagement through case studies has been fundamental in the development of the infrastructure, and this community is grouped into the ISA commons (Sansone et al, 2012), ISA commons: <http://isacommons.org/> (last accessed on 23rd July 2012). The website attempts to list ISA users who have adopted or extended the format and/or tools for both public and private resources. An example resource using the ISA infrastructure is Metabolights (<http://www.ebi.ac.uk/metabolights> (last accessed on 27th September 2012) for metabolomics experiments.

The latest additions in the ISA software suite, described next, are: *OntoMaton*, *Risa* and *isa2owl*.

OntoMaton provides support for online collaborative data curation (Maguire et al 2012). It is implemented in Javascript using Google Apps Script API and offers functionality for searching bio-ontologies and for tagging free text with terms from ontologies. These functionalities rely on the NCBO BioPortal REST Services (<http://www.bioontology.org/wiki/index.php/BioPortal>

REST_services (last accessed on 23rd July 2012) and can be used for general semantic data annotation. Additionally, the *ISAConfigurator* 1.6 tool, which allows curators to create standards-compliant templates for ISA-Tab, has been extended to build templates to be included in the Google cloud environment and combined with *OntoMaton*. The *OntoMaton* Google Template can be found at: <https://drive.google.com/templates?type=spreadsheets&q=ontomaton> (last accessed on July 23rd 2012).

The *Risa* package, available in BioConductor 2.11 (<http://www.bioconductor.org/packages/2.11/bioc/html/Risa.html>) (last accessed on 27th September 2012), offers methods for parsing ISA-Tab datasets and building R objects that can be used for analysis using domain specific packages. *Risa* also provides interfaces to some of these domain specific R packages, such as the *xcms* R package (Tautenhahn et al 2008) if there are mass spectrometry assays within the ISA-Tab data-set. Also, it is possible to augment the ISA-Tab metadata after analysis, and save the new files from R.

Last but not least, the *isa2owl* Java package follows our approach to expose ISA-Tab datasets to the Linked Data cloud. Our methodology relies on the definition of mapping files, aligning the ISA terminology with existing domain ontologies. A noteworthy mapping is that between ISA and the Ontology of Biomedical Investigations (OBI) (Brinkman et al 2010), which in turn is built in the Basic Formal Ontology (BFO) framework (Simon et al, 2006). Given such mapping, ISA-Tab datasets are parsed to populate the ontology, following Linked Data best practices such as the five star scheme whereby data is made available on the web in a structured non-proprietary format using URIs for identifying elements and linking to other data to provide context (Heath and Bizer 2011). This approach allows for semantic querying, discovery of links to other resources and reasoning over the ISA-Tab metadata.

Results and Discussion

The ISA infrastructure provides a comprehensive solution to the knowledge management challenges for experimental data in the biomedical domain. The modular architecture of the open source ISA tools enables users to adopt, and if necessary extend, one or more of the tools according to their specific needs. The underlying

ing ISA-Tab cross-domain format has proven to be generic enough and simple enough to be adopted by a large and growing community: the ISA commons.

The latest components of the ISA infrastructure offer support for collaborative semantic annotation in the cloud-computing environment of Google spreadsheets (OntoMaton), interface to popular data analysis packages (Risa), and a large number of new opportunities for querying, linking and reasoning about the data through transformation to RDF/OWL (isa2owl), making ISA-Tab datasets available in the linked data world. We are confident these tools will continue to facilitate important data management tasks in the context of massive amounts of data being generated in the biomedical domain.

Acknowledgements

This work was supported by the Biotechnology and Biological Sciences Research Council [grant BB/I025840/1 to SAS, BB/I000771/1, BB/I000917/1 to SAS]

References

1. Brinkman et al (2010), Modeling biomedical experimental processes with OBI. *J Biomed Semantics* 1(Suppl 1): S7. doi:10.1186/2041-1480-1-S1-S7
2. Castrillo et al (2007), Growth control of the eukaryote cell: a systems biology study in yeast, *J. Biol.* 6 (2):4. doi:10.1186/jbiol54
3. Heath and Bizer (2011), *Linked Data: Evolving the Web into a Global Data Space* (1st edition). *Synthesis Lectures on the Semantic Web: Theory and Technology*, 1:1, 1-136. Morgan & Claypool. doi:10.2200/S00334ED1V01Y2011WB001
4. Ho Sui et al (2012), The Stem Cell Discovery Engine, *Nucleic Acids Research*. doi: 10.1093/nar/gkr1051
5. Maguire et al (2012). *OntoMaton: bringing semantic annotation to Google spreadsheets for collaborative data management*. Manuscript submitted.
6. Rocca-Serra et al (2010), ISA software suite. *Bioinformatics*, 26. doi:10.1093/bioinformatics/btq415
7. Sansone et al (2012), Toward interoperable bioscience data, *Nature Genetics*, 27. doi:10.1038/ng.1054
8. Simon et al (2006), Formal ontology for natural language processing and the integration of biomedical databases. *Int J Med Inform.* 2006 Mar-Apr;75(3-4):224-31. doi: 10.1016/j.ijmedinf.2005.07.015
9. Tautenhahn et al (2008), Highly sensitive feature detection for high resolution LC/MS *BMC Bioinformatics*, 9:504, doi:10.1186/1471-2105-9-504

Network-based gene-disease prioritization using PROPHNET

Víctor Martínez[✉], Carlos Cano, Armando Blanco

Department of Computer Science and A.I., University of Granada, Granada, Spain

Motivation and Objectives

A major goal in biomedicine is to determine the underlying genetic causes of human diseases in order to better understand them and support their prevention and treatment. However, the genetic bases of many multifactorial diseases are still unclear, and high-throughput technologies typically report hundred or thousands of genes associated to a disease of interest. In this context is where gene-disease prioritization methods are of use. These computational methods make use of available data to obtain prioritized lists of genes (diseases) associated to a query set of diseases (genes). Prioritization is based on “guilt-by-association” which states that biological entities that are associated or interacting are more likely to share function. This allows to infer new relationships from already known interactions.

Many network-based prioritization methods have been proposed in the literature, performing well across different validation tests (Wang et al., 2011; Barabasi et al., 2011; Navlakha et al., 2010). We focus our study on two recent methods: rcNet (Hwang et al., 2011) and domainRBF (Zhang et al., 2011) since they outperform previous methods. Despite their good performance, these methods have clear limitations. First, they are strongly tailored to a specific domain of interest (gene-disease prioritization for rcNet and protein domain-disease prioritization for domainRBF, respectively). Hence, they cannot be applied to the prioritization of other biological entities of interest. Second, they do not allow to consider more than two types of networks for performing the prioritization (gene and disease networks in rcNet and domain and disease networks in domainRBF). However, we hypothesise that simultaneously integrating data from more than two complementary sources may improve the obtained results. For example, a gene-disease prioritization may benefit from known relationships between genes and diseases, but also from known interactions between drugs targeting certain genes to prevent or treat a specific disease.

We present ProphNet, a generic method of prioritization that achieves a better performance

by integrating and propagating information in an arbitrary number of heterogeneous data networks. Our method is generic since it allows to prioritize any biological entity of any kind with respect to some biological entities of another kind. Therefore, the user can customize the goal of the prioritization task (disease-gene, domain-disease, etc.) and the networks that are being taking into account for prophNet to achieve this goal. ProphNet is available as a web application at <http://genome2.ugr.es/prophnet/>. MATLAB source code, datasets and detailed experiments can also be downloaded at <http://genome2.ugr.es/prophnet/prophnet.zip>. In this talk we present prophNet and compare its results to those obtained by rcNet and domainRBF in two cases of study associated to gene-disease and domain-disease prioritization, respectively.

Methods

To perform the prioritization task, our method measures the influence of a query set of biological entities of a certain type (e.g. genes or diseases) in a target set of entities of another type (e.g. diseases or genes, respectively). To this end, the algorithm uses a graph representation of data sources where each node corresponds to a biological entity of a type of interest (gene/protein, disease, protein domain, etc.), and the arcs between two nodes are labeled with a weight (from 0 to 1) representing the strength of the relationship between the connected entities. These weights are derived from different biological sources and their interpretation varies depending on the type of the connected entities and the final goal of the study. The nodes of the graph may also be labeled with a value, representing the degree of association of each entity to the query or target set.

Our method integrates a set of networks, each one connecting entities of one type, into a global network in which entities of different types are interconnected. In this work we focus on the prioritization of entities of different types, i.e. the query and target sets belong to two different networks. A simplified version of the method

described below can be used to prioritize biological entities with respect to other entities of the same type (basically limiting the propagation to other networks).

To measure the degree of relationship between the query set and the target set, we first assign an initial value to the nodes of these sets. This initial value is set to $1/|X|$ by default, where $|X|$ is the cardinal of the query set and target set, respectively. However, the initial values assigned to the entities of the query set may also be different in case we want to assign a different relative importance to the elements in this set, always satisfying that the sum of the assigned values equals one both in the target and query sets. Nodes not in the query or target sets are initially set to 0.

After the initial values have been set, these node values are propagated within each network and between networks. The propagation within a network is performed using the Flow Propagation algorithm (Vanunu et al., 2008) that iteratively propagates node values until convergence, taking into account the weight of the arc connecting two entities to perform the propagation of node values between these entities. The propagation of values from one network A to a neighbor network B is performed by assigning each node in B directly connected to nodes in A the average of the values of the neighbour nodes in A. Neighbour nodes which are connected with an arc labeled with a weight below a defined threshold are not considered in order to reduce the propagation noise.

This propagation within and between networks is performed through all the networks in the path connecting the query network to the target network. This process causes the nodes of the networks adjacent to the target network to take a value based on their degree of relationship to the query set.

Finally, to calculate the degree of relationship between the query set and the target set, we compute the correlation between the values of the target nodes and the values of the nodes from adjacent networks directly connected to the target nodes. To obtain a prioritized list of genes (target) associated to a particular disease (query), this prioritization algorithm is applied iteratively using each node (gene) from the target network as the target set and computing the degree of relationship with the query disease. The

prioritized list is obtained by ordering the resultant correlation values in decreasing order. Since our method requires to iteratively compute correlation values for each query node and each target node, ProphNet is computationally expensive. However, it can be highly optimized by pre-calculating propagation scores in target networks and using parallelization techniques for fast response times. This way, a typical run for a gene-disease prioritization task takes a few seconds in our servers.

Results and Discussion

To compare the results obtained by prophNet with those obtained by state-of-the-art methods such as rcNet and domainRBF, we applied prophNet to the prioritization of genes-diseases and domain-diseases, respectively.

To perform a fair comparison of the results, we used the same data sources and methodology applied by rcNet and domainRBF to build the global network. The phenotype network was extracted from OMIM using text-mining techniques (van Driel et al., 2006) yielding a phenotype network with 5080 diseases. The phenotype-gene connections were extracted from OMIM using BioMart. The gene network was obtained from the Human Protein Reference Database (HPRD) and the protein domain network was derived from DOMINE and InterDom, with the domain-gene and domain-phenotype relationships extracted from Pfam.

We ran rcNet, domainRBF and ProphNet and tested their performance on different leave-one-out (LOO) cross-validation experiments. These LOO experiments were created by iteratively removing one gene-disease or one domain-disease relation from the available global network and using the corresponding gene or domain as query to check whether the prioritization method was able to predict the removed relationship. The accuracy of the result was measured as the rank assigned to the disease associated to the removed relation.

Apart from the LOO cross-validation experiments, we also performed experiments to test whether the different methods were able to predict new associations recently added to OMIM.

To measure the performance of the different prioritization methods, we computed Receiver Operating Characteristic (ROC) curves (data not shown due to format restriction) by plotting the

Table1: comparison of results for cross-validation experiments and new predictions for gene-disease and domain-disease prioritization tasks. Our method clearly outperforms rcNet and domainRBF in terms of AUC and mean ranking values. All ranking values are computed for a list of 5080 diseases. ROC curves associated to the AUC values were not included due to format restrictions.

Test	Method	AUC	Mean rank (Std. Dev)	Norm. Mean rank
LOO gene-disease	ProphNet	0,94	309.28 (811.51)	0.0609 (0.1597)
prioritization	rcNet	0,81	987.77 (1243.59)	0.1944 (0.2448)
New gene	ProphNet	0,81	980.37 (1329.85)	0.1930 (0.2618)
prioritization	rcNet	0,72	1441.59 (1476.84)	0.2835 (0.2907)
LOO domain-disease	ProphNet	0,93	346.87 (779.09)	0.0683 (0.1537)
prioritization	domain-RBF	0,87	671.58 (1199.2)	0.1322 (0.2361)

fraction of true positives out of the positives vs. the fraction of false positives out of the negatives, at various threshold settings. Areas under the ROC curves (AUC) and the average position in which the correct entity was ranked (see Table 1) are also computed.

The obtained results show that prophNet outperforms the other methods in all the proposed experiments, achieving the best avg. ranking position with the lowest standard deviation. A t-test has been performed on the mean ranking values obtained by the two algorithms compared in each experiment. The obtained p-values are less than 0.0001 for all the tests.

Although ProphNet results are significantly better than those obtained by other methods, a high mean ranking value was obtained due to the high variability of the results. Detailed results (not shown) reveal that although in most LOO runs the correct disease is prioritized at the top positions, for a small fraction of cases the disease is much

worse ranked, increasing the mean ranking value. Further studies are needed to analyze these cases to improve the results of the algorithm.

ProphNet was also applied to obtain prioritized lists of genes associated to Alzheimer, Diabetes Mellitus Type II and Breast Cancer (results not shown). The resultant top-ranked genes were related to these diseases according to recent publications in the literature.

Acknowledgements

This project was supported by grants from the Plan Propio de Investigación of the University of Granada, Spain, and it is part of projects P08-TIC-4299 of J. A., Sevilla and TIN2009-13489 of DGICT, Madrid.

References

1. Wang X, Gulbahce N and Yu H (2011) Network-based methods for human disease gene prediction. *Briefings in Functional Genomics* 10(5):280-293. doi: [10.1093/bfpg/eli024](https://doi.org/10.1093/bfpg/eli024).
2. Barabasi A, Gulbahce N and Loscalzo J (2011) Network medicine: a network-based approach to human disease. *Nature Reviews Genetics* 12:56-68. doi: [10.1038/nrg2918](https://doi.org/10.1038/nrg2918).
3. Navlakha S and Kingsford C (2010) The power of protein interaction networks for associating genes with diseases. *Bioinformatics* 26(8):1057-1063. doi: [10.1093/bioinformatics/btq076](https://doi.org/10.1093/bioinformatics/btq076).
4. Hwang T, Zhang W, et al. (2011) Inferring disease and gene set associations with rank coherence in networks. *Bioinformatics* 27(19): 2692-2699. doi: [10.1093/bioinformatics/btr463](https://doi.org/10.1093/bioinformatics/btr463).
5. Zhang W, et al. (2011) DomainRBF: a Bayesian regression approach to the prioritization of candidate domains for complex diseases. *BMC Systems Biology* 5:55. doi: [10.1186/1752-0509-5-55](https://doi.org/10.1186/1752-0509-5-55).
6. Vanunu O and Sharan R (2008) A propagation based algorithm for inferring gene-disease associations. *Proceedings of the German Conference on Bioinformatics*. German Conference on Bioinformatics: 54-62.
7. van Driel MA, et al. (2006) A text-mining analysis of the human phenome. *European Journal of Human Genetics* 14: 535-542. doi: [10.1038/sj.ejhg.5201585](https://doi.org/10.1038/sj.ejhg.5201585).

QTreds: a flexible LIMS for omics laboratories

Piergiorgio Palla^{1,2✉}, Gianfranco Frau², Laura Vargiu², Patricia Rodriguez-Tomé²

¹Department of Electrical and Electronic Engineering (DIEE), University of Cagliari, Cagliari, Italy

²Center for Advanced Studies, Research and Development in Sardinia (CRS4),Pula, Italy

Motivation and Objectives

Advances in DNA/RNA sequencing technologies and the contemporary introduction of the so-called “next-generation” sequencing instruments during the last decade, made it possible to automate several steps of the laboratory processes, leading to an increased throughput.

As a direct consequence, there has been an exponential growth of data being generated and the development of more efficient and complex laboratory procedures.

To efficiently handle the large amounts of data produced and to implement a quality control plan, a Laboratory Information Management System (LIMS) is highly-recommended. A LIMS is a complex software platform aimed to the management of the laboratory data and processes. Several LIMS are currently available, but most of them have prohibitive costs and have proprietary code, lacking often flexibility and scalability.

We have designed and implemented a QUALITY and TRacEability Data System (QTreds), a software platform born to address the specific needs of the CRS4 sequencing laboratory.

The main purpose of our in-house software solution was to provide a system that could help researchers to have a complete knowledge of the laboratory processes at each step, managing and verifying the:

- workflow creation
- samples traceability;
- diverse experimental protocols definition;
- inventory of reagents;
- users' roles and privileges;
- customized report generation.

Tracking and monitoring all the phases of the laboratory workflow can help to identify and troubleshoot problems more quickly, reducing the risk of process failures and their related costs

Methods

To develop QTreds, we followed a software approach commonly referred to as Agile. Starting from a general description of the needs of the

laboratory and the main functional requirements that the system was expected to have, we created a working but incomplete prototype, refining it constantly through a continuous interaction with the researchers and the personnel of the laboratory until obtaining the desired results.

QTreds is a web application with a client-server architecture developed in the Ruby programming language - Ruby web site: <http://www.ruby-lang.org> (Last accessed on July 2nd, 2012), using the framework Rails - Rails web site: <http://www.rubyonrails.org> (Last accessed on July 3rd, 2012).

QTreds has been developed following a design pattern known as Model-View-Controller (MVC) which assigns to objects of our system one of these three roles (Model, View or Controller) and defines the way objects communicate with each other. Model objects encapsulate the data and define the logic and computational methods to manipulate them. The View is responsible for generating a user interface, usually based on data in the Model. The Controller acts as an intermediary between one or more application's Views and one or more of its Models.

The persistence layer of the platform was developed using Active Record, a Rails implementation of the object-relational mapping (ORM) pattern introduced by Martin Fowler – Active Record web site: <http://ar.rubyonrails.org> (Last accessed on July 2nd, 2012). The latter enables the communication between QTreds and the MySQL relational database used to store data, thus reducing the need to use the Structured Query Language (SQL), but allowing us to use it whenever we needed it – MySQL website: <http://www.mysql.com> (Last accessed on July 4th, 2012).

The implementation of QTreds also relies on the use of different open-source programming libraries. The web user interface integrates the Views generated through the Rails Action View module with the Prototype Javascript Framework which enabled us to deal with the Asynchronous JavaScript and XML (AJAX) technology in a very easy and efficient way - web site: <http://www.prototypejs.org> (Last accessed on July 4th, 2012).

Furthermore the use of the script.aculo.us set of Javascript libraries - web site: <http://script.aculo.us> (Last accessed on July 4th, 2012) provides us with a visual effects engine, that we used to enhance the interactive user experience with the application.

All the activities and operations allowed by the QTreds platform can be assigned to four functional areas shown in the graphical representation below:



Figure 1. QTreds functional areas

The protocol definition is a crucial phase performed by the lab supervisor.

A protocol is a sort of template that describes the workflow, that is the sequence of steps of a particular class of experiments (for instance, the Exome Enrichment Protocol, will list all the steps of that kind of experiments: First Hybridization, First Wash, Purification, PCR, etc). Each step, that we call "activity", provides a detailed description of the instruments involved, the reagents used and the items consumed at a given point of an experiment. The lab supervisor can perform this protocol definition task using a graphical user interface or writing an eXtensible Markup Language Schema (XML) document, that must be compiled following a set rules that we defined and collected in an XML Schema Definition (XSD) document. The resulting XML protocol will be processed and exploited by the system to graphically represent the experiment workflow as a "state diagram" that will guide the user step by step, enabling him to manage and monitor the progression of his experiment.

QTreds is also provided with a system for privilege management and authorization. The Authorization module defines different user roles,

each with a different access profile; each role includes a set of features and privileges to which the assigned user will have access.

At the moment we have implemented six main roles: administrator, supervisor, user, inventory manager, analyzer and viewer.

Depending on the role assigned, each user will be able to perform different levels of operations and to access different kinds of information. For example a simple user will be allowed to see only data related to his experiments or to the projects in which he is involved, while the administrator will have a complete view of all the activities and data processing operations. A user can have different roles in different projects.

The Inventory Management module tracks all the reagents and items used by the researchers for their experiments. It includes different components: 1) a Catalog where all the typologies or categories of items used in the laboratory have to be inserted; 2) a Stock to register the reagents and other items physically present in the lab; 3) a Topology, which defines a virtual representation of the laboratory to keep track of the locations of the instruments in which the items are stocked; 4) a Personal Stock, that is a sort of "shopping cart" in which each researcher must insert all the reagents and items needed to perform his experiments.

QTreds integrates all aspects of the DNA sequencing process which includes sample submission, handling and tracking and also allows to design the workflow of each experiment providing the data needed for subsequent analyses.

Results and Discussion

QTreds has been developed, starting from the needs of the CRS4 Sequencing and Genotyping Platform (CSGP), where it has been used since June 2011 to make almost 100 DNA library preparation and sequencing experiments, processing over 1500 samples.

A new version of QTreds is currently under test and will be released soon. The new release will be provided with an efficient Application Programming Interface (API) in order to allow smart and automated access to information. The API has been implemented according to the REpresentational State Transfer (REST) architecture (Fielding, 2000). Using this API, any authorized user or system will be able to retrieve resources and information via a standard Hypertext Transfer

Protocol (HTTP) request, providing the opportune parameters.

The API will also enable to insert data into the QTreds database, creating a bi-directional communication channel between our system and any other external application or tool. The upcoming release will also provide a complete reporting system to visualize and export data in different file formats.

A trial version of QTreds is available on demand for academic users. We are setting up a web site to give public access to the system. We intend to put it online in time for the conference. QTreds will be distributed as an open source software: the exact license model (copyleft-style or "permissive" open source) is currently under discussion. Also in this case we will disclose our choice very soon. Thanks to its flexibility our system can be easily adapted to address the issues and the needs of other kinds of laboratories; therefore we are currently developing pro-

totypes for some research groups in the fields of Metabolomics and Proteomics, with whom we are actively collaborating.

Acknowledgements

The authors are grateful to the personnel of the CSGP for their precious suggestions and the help provided in each phase of the development of the platform.

References

1. Melo A, Faria-Campos A, et al (2010) SIGLa: an adaptable LIMS for multiple laboratories, BMC Genomics 11(Suppl 5):S8. doi:10.1186/1471-2164-11-S5-S8
2. Stocker G, Fisher M, et al (2009) iLAP: a workflow-driven software for experimental protocol development, data acquisition and analysis, BMC Bioinformatics 10, 390. doi:10.1186/1471-2105-10-390
3. Triplet T, Butler G (2012) The EnzymeTracker: an open-source laboratory information management system for sample tracking, BMC Bioinformatics 13, 15. doi:10.1186/1471-2105-13-15

Development of a text search engine for medicinal chemistry patents

Emilie Pasche¹✉, Julien Gobeil², Fatma Oezdemir-Zaech³, Therese Vachon³, Christian Lovis¹, Patrick Ruch²

¹Division of Medical Information Sciences (SIMED), University Hospitals of Geneva and University of Geneva, Geneva, Switzerland

²Bibliomics and Text-Mining Group (BiTeM), Information Science Department, University of Applied Sciences, Geneva, Switzerland

³Novartis Institute for BioMedical Research – Text Mining Services (NIBR-IT/TMS), Novartis Pharma AG, Basel, Switzerland

Motivation and Objectives

Over the last decades, the size of patent collections has strongly increased. Thus, in 2009, it was estimated that there are globally about 50 millions patents (Bonino et al., 2010) with about 15-20 millions related to medicinal chemistry, which represent a corpus of knowledge comparable to the content of MEDLINE. These collections represent an important and high-quality source of knowledge. However, while the past years have seen the development of a wealth of search engines and text mining instruments to navigate the bibliome – a term coined by (Grivell, 2002) to refer to the post-omics biomedical literature – with applications such as EBIMed, EAGLi, GoPubMed and Twease to cite only a few of them, text-mining applications dedicated to patents of the biomedical field remain rare.

Recently, the development of such specialized patent retrieval engines has benefited from the effort of dynamic communities of researchers, encouraged by the emergence of several evaluation campaigns, such as the Text REtrieval Conferences (TREC), one of the most popular competitions to evaluate and compare search engines. Prestigious universities, but also corporate research centres have regularly participated to these competitions. Lately, a task of information retrieval dedicated to patent search for chemistry, called TREC-Chem, has been set up. The objective was to model a prior art search task on a sizeable patent collection (two millions patents).

Based on the experience we have acquired during TREC competitions, we have developed and tuned an original search engine dedicated to patent search in the pharmaceutical domain. This paper describes the indexing and tuning of the engine to perform different types of search in a corporate patent collection.

Methods

A set of 1'004'868 patents has been randomly selected out of a collection of more than 13 millions of patents stored in an Oracle Database provided and maintained by IBM Almaden for Novartis. The content of the patents has been extracted using SQL queries and stored in files using an ad hoc XML format.

Evaluation of our methods is based on three sets of queries and relevance judgments. The first benchmark (B1) is used to evaluate the related patent search using the same methodology as proposed by TREC-Chem 2009 for the Prior Art Search task (Lupu et al., 2009). It is constituted of 96 topics or queries. Each topic corresponds to the title, abstract and claims of a given patent. For these experiments, the relevance judgments are generated out of the set of patents cited as prior-art by the given query. Only patents that are cited in the set of 1'004'868 patents are selected as many citations may concern patents not covered by the sample. The second benchmark (B2) is used to evaluate the engine in an *ad hoc* search task. In *ad hoc* search queries are usually limited to a few keywords. It is constituted of 24 topics, corresponding to the TREC-Chem 2010 and 2011 Technical Survey topics. Relevance judgments are provided by TREC and have been pre-processed to filter out patents not available in the 1'004'868 patent collection we are using. The last benchmark (B3) is used to evaluate a variant of the *ad hoc* search, where a single patent is targeted, using a *known-item search* methodology (Allen, 1989). It is constituted of 514 topics. Each topic contains ten words randomly selected from the title, abstract or claims of a given patent. In this set of experiments, the relevance judgments for each topic correspond to the patent from which the words are extracted. In that setting, a unique patent is considered as relevant for each query. The tuning of the system

Table1: P0 of the runs with the different strategies tested for each of the three benchmarks.

	B1	B2	B3
Baseline	2,20%	15,87%	23,63%
Remove description	2,87%	19,51%	33,59%
Remove description from metadata	3,63%	30,30%	35,02%
Use another weighting schema (BM25)	5,36%	20,05%	40,86%
Re-ranking based on citation network	6,76%	21,24%	40,87%
Injection of IPC codes	5,88%	23,28%	46,02%

is based on the maximization of the top precision, also called P0 or mean reciprocal rank. This measure evaluates the precision of the first returned result by the search engine. In our preliminary experiments, we focus on this metric since it provides a sound estimate of the retrieval effectiveness of the system for the three benchmarks. Indeed, other measures such as mean average precision cannot be applied to *known-item search* tasks.

We perform the indexing of the patent collection, using the Terrier search platform. Indexing is performed using baseline settings, with Porter stemming. First, we attempt to evaluate the impact of the description field – a time-expensive field to normalize and index – on the search effectiveness of the engine with the three use cases. Indeed, for sake of efficiency (in particular indexing time), we attempt to select only the most content-bearing sections of the patent. Second, we perform an ontology-driven normalization of the patent content. Three terminologies are used: Medical Subject Headings (MeSH), Gene Ontology (GO) and Caloha (Duek et al., 2012). Main terms and identifiers of mapped terms are stored as metadata. We evaluate the impact of the metadata field, to determine whether our onto-terminological normalization strategies bring useful additional information. Third, we evaluate the impact of the search models. Two search models are tested: the Okapi BM25 (Robertson et al., 2000) and PL2, a model based on Poisson estimation from randomness (Amati et al., 2002). Fourth, we evaluate the use of co-citation networks to improve our strategy. This approach consists in favouring the patents that are the most cited ones in the collection. We rank all patents by the number of time each patent

is cited by the others; thus building a large co-citation matrix. Then, we combine through linear combination this ranking with the results of the query as originally returned by the retrieval engine. Fifth, we attempt to evaluate the impact of the use of IPC classes. Some authors (Sternitzke, 2009) reported that using IPC codes with four values allowed retrieving the totality of the state of the art. Our method consists simply to add IPC codes to the topics and execute a new run.

Results and Discussion

The main, as well as most surprising result is that the *description* field did not improve our results for any of the three benchmarks (Table 1), but rather decreased significantly the precision at high ranks (P0). We thus decided to remove *description* fields from the engine's indexes, which resulted in faster indexing and reduced the size of indexes.

Second, we observed that the use of metadata, which was generated based only on the content of the title, abstract and claims, improved the precision of our results compared to the metadata including the description (Table 1). Thus, we can assume that descriptions should simply be discarded not only for indexing as mentioned in previous section, but also from the onto-terminology-driven normalization, which should result in a significant gain of time for the normalization process. It is to be noted that the next experiments are not based on this observation and use the onto-terminology-driven normalization of the full patent content (including description). As a further experiment, it would be interesting to evaluate the impact of the normalization by entity types. Indeed, (Ruch et al., 2005) reported that normalization and ex-

pansion of genes and gene products degraded the precision of search in MEDLINE during the TREC Genomic competition, while normalizing chemical, pathological, organism-related and anatomical concepts was moderately effective.

Third, concerning the weighting model, we observed that BM25 performed better than the deviation from randomness weighting schema we tested (Table 1). Our experiments focused on the feature selection and combination steps; therefore we assume the results reported here are rather weighting schema-independent; in particular because BM25 can be regarded as a strong baseline in the domain.

Fourth, we observed that the re-ranking based on citation networks improved results for the three benchmarks, but mainly for the related patent search (Table 1). We thus can assume that it is an appropriate functionality for prior art tasks (+26%, $p < 0.01$). The improvement for the *ad hoc* search task with the TREC benchmark was also significant (+5.9%).

Finally, we observed that IPC codes improved *ad hoc* search, but not related patent search (Table 1). Thus, we can assume that using an interactive IPC classifier (Teodoro et al., 2010) for *ad hoc* search could have a beneficial effect on the effectiveness of the search engine. In contrast, the length of the input for the prior art search makes obviously the use of IPC descriptors of less value.

We have thus presented the development of a search engine dedicated to patent search, based on the state of the research methods applied to patents. We have showed that a proper tuning of the system clearly increases the effectiveness of the system. We can also conclude that different search tasks, such as related patent search and *ad hoc* search, do demand to set up specific information retrieval models to be optimal.

Acknowledgements

Funding: Novartis Pharma AG.

References

1. Allen B (1989) Recall cues in known-item retrieval. *J Am Soc Inf Sci* 40(4), 246-252. doi:10.1002/(SICI)1097-4571(198907)40:4<246::AID-ASIA>3.0.CO;2-Z
2. Amati G and Van Rijsbergen CJ (2002) Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans Inf Syst* 20(4), 357-389. doi: 10.1145/582415.582416
3. Bonino D, Ciaramella A and Corno F (2010) Review of the state-of-the-art in patent information and forthcoming evolutions in intelligent patent informatics. *World Patent Information* 32(1), 30-38. doi:10.1016/j.wpi.2009.05.008
4. Duek PD, Gleizes A, Zwahlen C, Mottaz A, Bairoch A et al. (2011) CALOHA: A new human anatomical ontology as a support for complex queries and tissue expression display in neXtProt. *Bio-Ontologies* 2011. <http://bio-ontologies.knowledgeblog.org/196> (accessed 06 October 2012).
5. Grivell L (2002) Mining the bibliome: searching for a needle in a haystack? New computing tools are needed to effectively scan the growing amount of scientific literature for useful information. *EMBO Rep* 3(3), 200-203. doi: 10.1093/embo-reports/kvf059
6. Lupu M, Piroi F, Huang XJ, Zhu J and Tait J (2009) Overview of the TREC 2009 Chemical IR Track. In proceedings of TREC 2009. <http://trec.nist.gov/pubs/trec18/papers/CHEM09.OVERVIEW.pdf> (accessed 06 October 2012).
7. Robertson SE, Walker S and Beaulieu M (2000) Experimentation as a way of life: Okapi at TREC. *Information Processing & Management* 36(1), 95-108. doi: 10.1016/S0306-4573(99)00046-1
8. Ruch P, Müller H, Abdou S, Cohen G and Savoy J (2005) Report on the TREC 2005 Experiment: Genomics Track. In proceedings of TREC 2005. <http://trec.nist.gov/pubs/trec14/papers/uhsospital-geneva.geo.pdf> (accessed 06 October 2012)
9. Sternitzke C (2009) Reducing uncertainty in the patent application procedure – Insights from invalidating prior art in European patent applications 31(1), 48-53. doi: 10.1016/j.wpi.2008.04.007
10. Teodoro D, Pasche E, Vishnyakova D, Gobeill J, Ruch P et al. (2008) Automatic IPC Encoding and Novelty Tracking for Effective Patent Mining. In proceedings of NTCIR-8 Workshop Meeting. <http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings8/NTCIR/03-NTCIR8-PATMN-TeodoroD.pdf> (accessed 06 October 2012)

The ontogene system: an advanced information extraction application for biological literature

Fabio Rinaldi[✉]

Institute of Computational Linguistics, University of Zurich

Motivation and Objectives

The rapid expansion of the biomedical knowledge encoded in the scientific literature is proving to be a major bottleneck for the progress of biomedical sciences. It is increasing difficult even for the best experts to keep track of all relevant information pertinent to their domain of interest. It is becoming therefore imperative to explore solutions based on advanced text mining technologies in order to identify and extract the most relevant nuggets of information from the vastness of the literature.

Methods

The OntoGene system (<http://www.ontogene.org/>) is an advanced NLP-based pipeline capable of efficiently processing large quantity of textual documentation and extracting from it specific items of information, and in particular the biomedical entities of interest to the user, and their relationships.

Biomedical terminological resources can be leveraged for construction of large-scale knowledge bases. One example is KaBOB (Knowledge Base of Biology), a large RDF store based upon 17 prominent biomedical databases (Bada et al, 2011). Similar kinds of integrated data networks can be used for knowledge discovery purposes through usage of semantic web technologies (Chen et al, 2009). In our own work we have used such databases as knowledge sources for the process of semi-automated information extraction. In the rest of this section we describe the OntoGene Text Mining pipeline which is used to (a) provide all basic preprocessing (e.g. tokenization) of the target documents, (b) identify all mentions of domain entities and normalize them to database identifiers, and (c) extract candidate interactions.

We use in particular the following resources as terminology sources: UniProt Knowledge base (proteins), NCBI Taxonomy (species), Proteomics Standards Initiative Molecular Interactions Ontology (experimental methods), Cell Line Knowledge Base (cell lines), UMLS (diseases), etc.

Terms, i.e. preferred names and synonyms, are automatically extracted from the original database and stored in a common internal format, together with their unique identifiers (as obtained from the original resource). An efficient lookup procedure is used to annotate any mention of a term in the documents with the ID(s) to which it corresponds. A term normalization step is used to take into account a number of possible surface variations of the terms. The same normalization is applied to the list of known terms at the beginning of the annotation process, when it is read into memory, and to the candidate terms in the input text, so that a matching between variants of the same term becomes possible despite the differences in the surface strings (Rinaldi et al, 2008).

The system combines mentions of relevant domain entities (and their corresponding unique identifiers) from the same syntactic context in order to create candidate interactions. An initial ranking of the candidate relations can be generated on the basis of frequency of occurrence of the respective entities only. This ranking is further refined using a syntax-based approach, which is based upon an accurate parsing of all the sentences of the target document, and a machine learning approach which makes use of a maximum entropy classifier to boost candidate entities and interactions on the basis of the global distribution of information in the original database (Rinaldi, Schneider, et al, 2012).

Results and Discussion

The results of the text mining system are presented to the user through an intuitive and user-friendly interface, called ODIN (OntoGene Document Inspector). The ODIN interface allows the user to inspect entities and relationships identified by the text mining system, and see them in the context where they were originally found.

For example, the figure below shows an implementation of ODIN customized for curation of the Comparative Toxicogenomics Database (CTD, Mattingly et al, 2006). The left panel shows

The screenshot displays the ODIN interface. On the left, a document titled "Document PMID 10861484" is open, showing an abstract about cyclophosphamide's effect on p53-specific CTLs. The text is annotated with colored underlines: green for chemicals (e.g., cyclophosphamide), yellow for diseases (e.g., neoplasms), and blue for genes (e.g., p53, CTL). On the right, the "Annotation" panel shows a table of candidate interactions. The table has columns for "Conf", "Type 1", "Name 1", "Type 2", "Name 2", and "N". The interactions listed include relationships between Cyclophosphamide (chem) and Neoplasms (disease), CTL (gene), TRP53 (gene), and TP53 (gene), as well as relationships between Neoplasms (disease) and CTL (gene), TRP53 (gene), and TP53 (gene).

Conf	Type 1	Name 1	Type 2	Name 2				N
0.08	chem	Cyclophospha...	disease	Neoplasms	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
0.08	chem	Cyclophospha...	gene	CUTLET	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
0.08	chem	Cyclophospha...	gene	CTL	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
0.08	chem	Cyclophospha...	gene	TRP53	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
0.06	disease	Neoplasms	gene	CUTLET	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
0.06	disease	Neoplasms	gene	CTL	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
0.06	disease	Neoplasms	gene	TRP53	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
0.05	chem	Cyclophospha...	gene	TP53	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
0.04	chem	Cyclophospha...	gene	IFNB1	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
0.04	disease	Neoplasms	gene	TP53	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
0.04	disease	Neoplasms	gene	IFNB1	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
0.03	chem	Cyclophospha...	gene	P53	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Figure 1: ODIN: the OntoGene curation interface in the CTD application.

the original document, with entities underlined and color-coded (green: chemicals, yellow: diseases, blue: genes). The right panel shows the candidate relationships identified by the system. Selecting one of the interactions will highlight in the document the information that was used by the system to propose that interaction.

The results (interactions in this case) are presented according to a ranking which is based upon a score reflecting the confidence of the system in a given proposed interaction, thus allowing the user to stop inspecting them at an optional confidence threshold. The user can with a simple click then confirm or reject a candidate interaction. Additionally, all entities are easily editable, allowing correction of annotation errors.

The OntoGene pipeline has been applied to several Information Extraction tasks. In the context of the BioCreative challenges (Krallinger et al 2008), the system was capable of achieving the best results in extracting mentions of protein-protein interactions (2009) and mentions of experimental methods for protein interaction detection (2006).

Recently the system has been adapted for an experiment in assisted curation for the PharmGKB database (Klein et al 2001). This experiment, conducted in collaboration with PharmGKB curators, has lead to interesting re-

sults showing the reliability and usability of the system (Rinaldi, Clematide, et al, 2012).

In the "triage" task of BioCreative 2012 (ranking of documents according to relevance for the curation process of the CTD database), once again the OntoGene system obtained the best overall results among the participants (Rinaldi et al, 2013).

Acknowledgements

This research is partially funded by the Swiss National Science Foundation (grant 100014-118396/1) and Novartis AG, NIBR-IT, TextMining Services, CH-4002, Basel, Switzerland.

References

1. Michael Bada, Kevin Livingston, and Lawrence Hunter. An ontological representation of biomedical data sources and records. *Bio-Ontologies*, 2011.
2. Huajun Chen, Li Ding, Zhaohui Wu, Tong Yu, Lavanya Dhanapalan, and Jake Y. Chen. Semantic web for integrated network analysis in biomedicine. *Briefings in Bioinformatics*, 10(2):177–192, 2009.
3. T.E. Klein, J.T. Chang, M.K. Cho, K.L. Easton, R. Fergerson, M. Hewett, Z. Lin, Y. Liu, S. Liu, D.E. Oliver, D.L. Rubin, F. Shafa, J.M. Stuart, and R.B. Altman. Integrating genotype and phenotype information: An overview of the PharmGKB project. *The Pharmacogenomics Journal*, 1:167–170, 2001.
4. M. Krallinger, Alexander Morgan, Larry Smith, Florian Leitner, Lorraine Tanabe, John Wilbur, Lynette Hirschman, Alfonso Valencia. Evaluation of text-mining systems for biology: overview of the second BioCreative community challenge. *Genome Biology*, 2008, 9(Suppl 2):S1.

5. C.J. Mattingly, M.C. Rosenstein, G.T. Colby, J.N. Forrest Jr, and J.L. Boyer. The Comparative Toxicogenomics Database (CTD): a resource for comparative toxicological studies. *Journal of Experimental Zoology Part A: Comparative Experimental Biology*, 305A(9):689–692, 2006.
6. Fabio Rinaldi, Thomas Kappeler, Kaarel Kaljurand, Gerold Schneider, Manfred Klenner, Simon Clematide, Michael Hess, Jean-Marc von Allmen, Pierre Parisot, Martin Romacker, and Therese Vachon. OntoGene in BioCreative II. *Genome Biology*, 9(Suppl 2):S13, 2008.
7. Fabio Rinaldi, Simon Clematide, Yael Garten, Michelle Whirl-Carrillo, Li Gong, Joan M. Hebert, Katrin Sangkuhl, Caroline F. Thorn, Teri E. Klein, and Russ B. Altman. Using ODIN for a PharmGKB revalidation experiment. *The Journal of Biological Databases and Curation*, Oxford Journals, 2012
8. Fabio Rinaldi, Gerold Schneider, Simon Clematide. Relation Mining Experiments in the Pharmacogenomics Domain. *Journal of Biomedical Informatics*, 2012.
9. Fabio Rinaldi, Simon Clematide, Simon Hafner, Gerold Schneider, Gintare Grigonyte, Martin Romacker, Therese Vachon. Ranking of CTD articles and interactions using the OntoGene pipeline. *The Journal of Biological Databases and Curation*, Oxford Journals, 2013 (accepted for publication).
10. Thomas C. Wieggers, Allan Peter Davis, and Carolyn J. Mattingly. Collaborative Biocuration-Text Mining Development Task for Document Prioritization for Curation. *The Journal of Biological Databases and Curation*, Oxford Journals, 2013 (accepted for publication).

IntelliGenWiki: An Intelligent Semantic Wiki for Life Sciences

Bahar Sateli¹, Marie-Jean Meurs^{1,2}, Greg Butler^{1,2}, Justin Powlowski^{2,3}, Adrian Tsang^{2,4}, René Witte¹✉

¹Department of Computer Science and Software Engineering; Concordia University, Montréal, Canada

²Centre for Structural and Functional Genomics; Concordia University, Montréal, Canada

³Department of Chemistry and Biochemistry; Concordia University, Montréal, Canada

⁴Department of Biology; Concordia University, Montréal, Canada

Motivation and Objectives


The rapid growth of the scholarly literature makes the management and curation of the available information a labor-intensive and time-consuming task for researchers, during which significant knowledge can be easily missed. To address this problem, efforts have been made to use Natural Language Processing (NLP) techniques as a means to (semi-)automatically improve the exhaustive analysis of the available information. In order to make these NLP techniques more end-user friendly and integrate them with knowledge management workflows, we developed IntelliGenWiki, a novel combination of a wiki system with state-of-the-art techniques from the NLP and Semantic Computing domains. Wikis are well known as an easy-to-use, collaborative platform for creating and organizing knowledge. For example, the Gene Wiki project (Huss III et al, 2010) applies community intelligence to the annotation of gene and protein functions. However, existing approaches rely on a manual analysis of the literature. With IntelliGenWiki, we aim to leverage the collaborative nature of wikis by introducing new Human-AI collaboration patterns: Our goal is to provide text mining assistants that work together with humans on literature analysis tasks, like curation or the generation of semantic metadata, which can be used in an Linked Open Data context. IntelliGenWiki is based on an open service-oriented architecture: it can be applied to different projects by deploying custom NLP analysis pipelines suitable for the specific task and domain. Here, we demonstrate the benefits of this approach within a collaborative literature curation context.

Methods

We first describe the general workflow for working with NLP assistants, followed by a description of the underlying architecture.

Workflow. IntelliGenWiki provides a standard wiki user interface. From any wiki page (Fig. 1, top), users can ask for “Semantic Assistants” from the menu (Fig. 1, left), which will result in a dynamically injected user interface from which assistants can be selected and executed (Fig. 1, bottom). The user can now select an appropriate assistant from the list and invoke it on one or multiple pages of the wiki, gathered in a so-called “collection”. This will invoke the selected NLP pipeline on the set of wiki pages. The results (e.g., detected entities) are stored in the user’s place of choice and made persistent in the wiki repository (Fig. 1, middle). Thereby, all updated pages become immediately available to all wiki users for collaborative adjustment, modification and further refinement of the results.

Architecture. Technically, IntelliGenWiki combines NLP analysis pipelines developed in the General Architecture for Text Engineering (GATE) (Cunningham et al, 2011) with MediaWiki, <http://www.mediawiki.org> (Last accessed: 26.09.2012), a widely-used wiki engine. These pipelines are published as standard web services through the Semantic Assistants framework (Witte and Gitzinger, 2008). The Wiki-NLP integration is based on a service-oriented architecture that seamlessly introduces these NLP web services into wiki systems (Sateli and Witte, 2012). This allows wiki users to benefit from text mining techniques directly within their wiki environment, without the need for switching to an external application. Additionally, we support the generation of semantic metadata from NLP analysis results. This metadata is formally represented in the wiki through the Semantic MediaWiki (SMW) extension: <http://semantic-mediawiki.org/> (Last accessed on Sept 26, 2012). This formal representation of the available wiki knowledge can be exploited by exporting it in form of RDF triples. It can also be queried directly within the wiki using SMW inline queries. For example, users could write queries to retrieve



Sysop [my talk](#) [my preferences](#) [my watchlist](#) [my contributions](#) [log out](#)

page
discussion
edit
history
delete
move
protect
watch
refresh

PMID: 20709852

[Contents \[show\]](#)

Characterization of a Cellobiohydrolase (MoCel6A) Produced by Magnaporthe oryzae [\[edit\]](#)

PMID: 20709852

Authors: Machiko Takahashi,1‡ Hideyuki Takahashi,1‡ Yuki Nakano,1 Teruko Konishi,2 Ryohei Terauchi,1 and Takumi Takeda1*

Iwate Biotechnology Research Center, Kitakami, Iwate 024-0003, Japan,1 University of the Ryukyus, Department of Bioscience and Biotechnology, Faculty of Agriculture, 1 Senbaru Nishihara, Okinawa 903-0213, Japan2

*Corresponding author. Mailing address: Iwate Biotechnology Research Center, 22-174-4, Narita, Kitakami, Iwate 024-0003, Japan. Phone: 81 (197) 68-2911. Fax: 81 (197) 68-3811. E-mail: ttakeda@ibrc.or.jp

‡M.T. and H.T. contributed equally to this work.

Received March 10, 2010; Accepted July 30, 2010.

Full Text [\[edit\]](#)

Abstract [\[edit\]](#)

Three GH-6 family cellobiohydrolases are expected in the genome of Magnaporthe grisea based on the complete genome sequence. Here, we demonstrate the properties, kinetics, and substrate specificities of a Magnaporthe oryzae GH-6 family cellobiohydrolase (MoCel6A). In addition, the effect of cellobiose on MoCel6A activity was also investigated. MoCel6A contiguously fused to a histidine tag was overexpressed in M. oryzae and purified by affinity chromatography. MoCel6A showed higher hydrolytic activities on phosphoric acid-swollen cellulose (PSC), β-glucan, and cellooligosaccharide derivatives than on cellulose, of which the best

These results suggest that enhancement or inhibition of hydrolytic activities by cellobiose is dependent on the reaction mixture pH.

PMID: 20709852 [\[PubMed - indexed for MEDLINE\]](#) PMCID: PMC2950481 [Free PMC Article](#) [mycoMINE on PMID: 20709852_Abtract \(View\)](#) [\[View\]](#)

Content	Type	Start	End	Features
cellobiohydrolase	Enzyme	89	106	<ul style="list-style-type: none"> ■ enzyme_alias: cellobiohydrolase ■ BRENDA_SystematicName: 4-beta-D-glucan cellobiohydrolase ■ BRENDA_EcNumber: 3.2.1.91 ■ abbreviation_alias: - ■ google_search: http://www.google.com/search?q=cellobiohydrolase ■ BRENDA_RecommendedName: cellulose 1,4-beta-cellobiosidase ■ SwissProt_ID: O68438 ■ BRENDA's page: http://www.brenda-enzymes.org/php/result_flat.php4?ecno=3.2.1.91

navigation

- Main page
- Community portal
- Current events
- Recent changes
- Random page
- Help

search

toolbox

- What links here
- Related changes
- Special pages
- Printable version
- Permanent link
- Semantic Assistants
- Browse properties

Semantic Assistants Plug-in

Biomedical Literature

Annotations Extracted from Wiki Page

Wiki-NLP Integration Interface

This page was last modified on 28 March 2012, at 19:15.

This page has been accessed 24 times.

[Privacy policy](#) [About G-nWiki](#) [Disclaimers](#)

 Powered By MediaWiki

Available Assistants
Results Target
Global Settings
Console

Step 1. Select the service your wish to execute on your collection.
Once you add this page to your collection, you can continue browsing as your collection is saved.

Available Assistants

Runtime Parameters

- mycoMINE
- IR Information Extractor
- Information Extractor
- OrganismTagger

Collection

Fig. 1: The wiki interface with integrated text analysis services (bottom), showing automatically added, NLP-extracted entities (middle), together with original content (top)

literature that contains a certain type of entities, such as enzymes or organisms.

Results and Discussion

To test the effectiveness of NLP assistants in a wiki environment, we deployed an IntelliGenWiki installation within the Genozymes project: <http://www.fungalgenomics.ca> (Last accessed on Sept 20, 2012). The task we aimed to support in the project is biomedical literature curation for lignocellulose research. For this experiment, we deployed the mycoMINE NLP pipeline (Meurs et al, 2012), which automatically extracts knowledge from the literature on fungal enzymes by using semantic text mining approaches combined with ontological resources. We manually pre-filled the wiki with a corpus of 30 documents composed of PubMed abstracts and their corresponding full-text papers, selected by two expert biocurators. These biocurators provided us with their average time needed for curation without support on the same task. They performed the corpus curation through the wiki using mycoMINE to automatically extract relevant entities, and they kept track of the time spent on each document. The time for abstract selection (triage task) decreased from 1min. (without support) to 20sec. (using IntelliGenWiki), and from 37.5min (without support) to 30.6min (using IntelliGenWiki) for full paper selection (curation task), showing a productivity enhancement of 67% and 20%, re-

spectively. The results gathered from this experiment confirm the usability and the effectiveness of our approach.

The IntelliGenWiki system, including the NLP integration back-end, is available as open source software from <http://www.semanticsoftware.info/intelligenwiki>.

Acknowledgements

Funding for this work was provided by NSERC, Genome Canada and Génome Québec. Caitlin Murphy and Sherry Wu are acknowledged for their participation in the evaluation task.

References

1. Cunningham H, Maynard D, et al (2011) Text Processing with GATE (Version 6), University of Sheffield, Department of Computer Science
2. Huss III J. W., et al (2010) The Gene Wiki: Community Intelligence Applied to Human Gene Annotation, *Nucleic Acids Research* 38, p. 633–639. doi:10.1093/nar/gkp760
3. Meurs MJ, Murphy C, et al (2012) Semantic Text Mining Support for Lignocellulose Research, *BMC Medical Informatics and Decision Making* 12(Suppl 1):S5. doi:10.1186/1472-6947-12-S1-S5
4. Sateli B and Witte R (2012) Natural Language Processing for MediaWiki – The Semantic Assistants Approach, In 8th International Symposium on Wikis and Open Collaboration (WikiSym 2012). Linz, Austria.
5. Witte R and Gitzinger T (2008) Semantic Assistants – User-Centric Natural Language Processing Services for Desktop Clients, In Asian Semantic Web Conference (ASWC 2008), Springer LNCS 5367, pp.360–374. doi:10.1007/978-3-540-89704-0_25

ROCK: a resource for integrative breast cancer data analysis

Saif Ur-Rehman¹, Costas Mitsopoulos¹, Marketa Zvebil¹

¹Breakthrough Breast Cancer Research, Institute of Cancer Research, London, United Kingdom

Motivation and Objectives

There is currently an ongoing research effort to understand the molecular mechanisms underpinning breast cancer being undertaken worldwide. The generated data measure different

usage of the updated version of ROCK is to allow bench scientists to query the state of a given gene/genomic locus within a subset of samples defined by clinical annotation terms.

Methods

ROCK is implemented through the manual curation of publically available datasets containing data types such as gene expression, genomic copy number aberrations, microRNA expression, RNA interference, protein-protein interactions, and signalling protein networks. These data types are integrated using standard gene identifiers from Ensembl (Flicek et al 2011). It also utilises clinical annotation of samples in order to facilitate comparison of these data types between disease subtypes as illustrated in Figure 1. The new version of ROCK allows sub-setting of data in a multitude of clinical categories.

ROCK now contains an amalgamation of gene expression studies as measured by microarray across multiple platforms. This study is a meta-analysis of breast carcinoma samples from a number of public datasets, which have been normalised together (Boersma et al 2008; Desmedt et al 2007; Farmer et al 2005; Graham et al 2010; Loi et al 2008; Minn et al 2005; Pawitan et al 2005; Popovici et al 2010; Schmidt et al 2008; Sofiriou et al 2006; Wang et al 2005). It also implements a novel confidence score (CS) for binary protein-protein interactions, which is similar to the score, used by IntAct (Kerrien et al 2012), but also utilises structural parameters where available.

In addition to this ROCK allows survival analysis (using Kaplan-Meier plots) on both gene expression levels as well as the clinical features of a given tumour.

The system is implemented using a three-tier web architecture. Web browsers communicate with the interface using Javascript via Apache Tomcat. Server-side data requests are then primarily handled using a set of core Java classes that communicate with the backend relational database. This database is implemented using Oracle.t

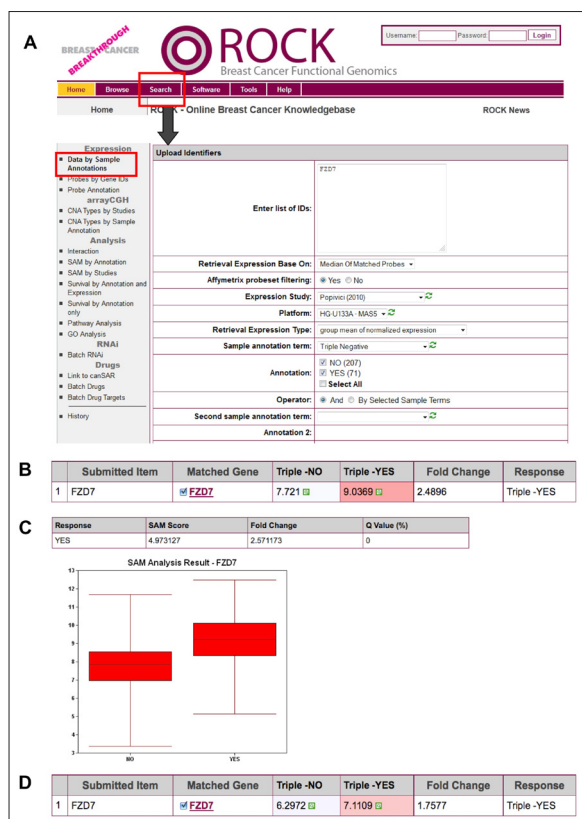


Figure 1: FZD7 differential expression in triple negative breast cancers. (A) ROCK interface for exploring gene expression fold changes. (B) Triple negative response for FZD7 in the Popovici dataset (Popovici et al 2010). (C) Significance of microarrays (SAM) (Tusher et al 2001) analysis of (B) performed in ROCK. The "SAM by annotation" option is under the Analysis header in the ROCK Search menu. (D) Triple negative response for FZD7 across amalgamated gene expression studies.

molecular characteristics that underlie the cancer phenotype. We present the updated ROCK (ROCK Online Cancer Knowledgebase) database, which now integrates these diverse data types allowing unique analyses of published breast cancer experimental data. The primary

Results and Discussion

ROCK provides a unique breast cancer analysis platform of integrated experimental datasets at the genomic, transcriptomic, and proteomic level. ROCK is available on rock.icr.ac.uk.

The data integration within ROCK allows for bespoke analyses to be carried out on a given data type, the results of which can be projected onto a different data type, e.g. a set of genes detected as differentially expressed between cancer subtypes can then be subjected to an analysis of the interactions of their resultant proteins. This allows a user to carry out an iterative analysis where the results of a preliminary investigation can be used as the input to a series of successive investigations. ROCK is used by an average of 2329 unique users annually.

We will present the recent and major functional updates and enhancements to the ROCK resource, including new analysis modules and microRNA and NGS data integration, and illustrate how ROCK can be used to confirm known experimental results as well as generate novel leads and new experimental hypotheses using the Wnt signalling cell surface receptor FZD7 and the Myc oncogene.

References

1. Boersma BJ, Reimers M, Yi M et al (2008) A Stromal Gene Signature Associated with Inflammatory Breast Cancer. *Int J Cancer* 122:1324
2. Desmedt C, Piette F, Loi S et al (2007) Strong Time Dependence of the 76-Gene Prognostic Signature for Node-Negative Breast Cancer Patients in the Transbig Multicenter Independent Validation Series. *Clin Cancer Res* 13:3207
3. Farmer P, Bonnefoi H, Becette V et al (2005) Identification of Molecular Apocrine Breast Tumours by Microarray Analysis. *Oncogene* 24:4660
4. Flicek P, Amode MR, Barrell D et al (2011) Ensembl 2011. *Nucleic Acids Research* 39:D800
5. Graham K, de las Morenas A, Tripathi A et al (2010) Gene Expression in Histologically Normal Epithelium from Breast Cancer Patients and from Cancer-Free Prophylactic Mastectomy Patients Shares a Similar Profile. *Br J Cancer* 102:1284
6. Kerrien S, Aranda B, Breuza L et al (2012) The Intact Molecular Interaction Database in 2012. *Nucleic Acids Research* 40:D841
7. Loi S, Haibe-Kains B, Desmedt C, Wirapati P et al (2008) Predicting Prognosis Using Molecular Profiling in Estrogen Receptor-Positive Breast Cancer Treated with Tamoxifen. *BMC Genomics* 9:239
8. Minn AJ, Gupta GP, Siegel PM et al (2005) Genes That Mediate Breast Cancer Metastasis to Lung. *Nature* 436:518
9. Pawitan Y, Bjohle J, Amler L et al (2005) Gene Expression Profiling Spares Early Breast Cancer Patients from Adjuvant Therapy: Derived and Validated in Two Population-Based Cohorts. *Breast Cancer Res* 7:R953
10. Popovici V, Chen W, Gallas BG et al (2010) Effect of Training-Sample Size and Classification Difficulty on the Accuracy of Genomic Predictors. *Breast Cancer Res* 12:R5
11. Schmidt M, Bohm D, von Tonne C, Steiner E et al (2008) The Humoral Immune System Has a Key Prognostic Impact in Node-Negative Breast Cancer. *Cancer Res* 68:5405
12. Sofiriou C, Wirapati P, Loi S et al (2006) Gene Expression Profiling in Breast Cancer: Understanding the Molecular Basis of Histologic Grade to Improve Prognosis. *J Natl Cancer Inst* 98:262
13. Tusher VG, Tibshirani R, Chu G (2001) Significance Analysis of Microarrays Applied to the Ionizing Radiation Response. *Proc Natl Acad Sci U S A* 98:5116
14. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM et al (2005) Gene-Expression Profiles to Predict Distant Metastasis of Lymph-Node-Negative Primary Breast Cancer. *Lancet* 365:671

TECHNOLOGICAL-INDUSTRIAL COMMUNICATIONS

Extracting knowledge from biomedical data through Logic Learning Machines and RuleX

Marco Muselli✉

Institute of Electronics, Computer and Telecommunication Engineering, National Research Council, Genoa, Italy

Motivation and Objectives

When dealing with biomedical data concerning a given problem, usually experts are required to infer specific conclusions about a pathology or a biological phenomenon of interest starting by a sample of previously collected observations. Besides conventional statistical techniques that allow to retrieve important indications about the characteristics of the system, machine learning methods have revealed to be very effective in predicting its behavior in cases different from those included in the observations at hand.

Among this last group of techniques rule generation methods build models described by a set of intelligible rules, thus permitting to extract important knowledge about the variables included in the analysis and on their relationships with the output attribute. Two different paradigms have been proposed in literature to perform rule generation: decision trees (Duda et al., 2001), which adopt a divide-and-conquer approach for generating the final model, and methods based on Boolean function reconstruction (Boros et al., 2000; Muselli and Ferrari, 2011), which follow an aggregative procedure for building the set of rules.

Available commercial software, such as SAS, SPSS or STATA, allows to employ a wide range of statistical techniques for the analysis of real world data, allowing also the application of some machine learning algorithms, among which neural networks and decision trees. However, the focus of these suites is more centered on conventional statistics rather than on machine learning and consequently it is difficult for a non expert to successfully extract knowledge from its own data by employing advanced techniques offered by commercial packages.

This is even more true when considering freely available software tools, such as Weka (www.cs.waikato.ac.nz/ml/weka), Orange (orange.biolab.si), or R (www.r-project.org); in these cases a wider range of machine learning approaches is generally made available, but the level of exper-

ience needed to achieve a satisfying result is often too high to allow their use by a non expert.

To overcome these difficulties a new suite for extracting knowledge from real world data has been developed by Impara srl (www.impara-ai.com); it is named RuleX (contraction of RULE EXtraction) since it is especially devoted to generate intelligible rules, although a wide range of statistical and machine learning approaches will be made available. An intuitive graphical interface allows to easily apply standard and advanced algorithms for analyzing any dataset of interest, providing solution to classification, regression and clustering problems.

Besides standard techniques, such as decision trees (DT), neural networks (NN), logistic (LOGIT), and k-nearest-neighbor (KNN), RuleX offers the possibility of applying an original proprietary approach, named Logic Learning Machine (LLM), which represents an efficient implementation of the switching neural network model (Muselli, 2006). LLM allows to solve classification problems producing sets of intelligible rules capable of achieving an accuracy comparable or superior to that of best machine learning methods.

The application of RuleX to the analysis of biomedical datasets included in the Statlog benchmark (Michie et al., 1994) permits to appreciate the good characteristics of this new analysis software. In particular, it is shown how different algorithms can be easily employed to extract knowledge from data at different levels of intelligibility, comparing results produced by corresponding models.

Methods

Although conventional statistical techniques or standard machine learning approaches allow to retrieve important indications about the characteristics of a system or of a phenomenon of interest, starting from a sample of observations regarding historical data, a deeper insight into the relationships among the considered variables can only be obtained by adopting rule generation methods. These techniques are capable of

constructing models described by a set of intelligible rules having the following form:

if **premise** then **consequence**

where **premise** is the conjunction of conditions on the input variables, whereas **consequence** contains information about the output of the model.

For instance, in a diagnosis problem rule generation techniques produce not only the subset of variables actually correlated with the pathology of interest, but also explicit intelligible conditions that determine a specific diagnosis. As a consequence, relevant thresholds for each input variable are identified, which represent valuable information for understanding the phenomenon at hand.

Most used rule generation techniques belong to the following two broad paradigms: decision trees and methods based on Boolean function synthesis. The approach adopted by the first kind of algorithms divides iteratively the training set into smaller subsets according to a divide and conquer strategy: this gives rise to a tree structure from which an intelligible set of rules can be easily retrieved. It is important to observe that the divide and conquer strategy leads to conditions and rules that point out differences among examples of the training set belonging to different output classes. In this sense we can say that the DT approach implements a discriminant policy: differences between output classes are the driver for the construction of the model.

In contrast, methods based on Boolean function synthesis adopt an aggregative policy: at any iteration some patterns belonging to the same output class are clustered to produce an intelligible rule. Suitable heuristic algorithms (Boros et al., 2000; Muselli and Ferrari, 2011) are employed to generate rules exhibiting the highest covering and the lowest error; a trade-off between these two different objectives has been obtained by applying the Shadow Clustering

(SC) technique (Muselli and Ferrari, 2011), which generally leads to final models exhibiting a good accuracy.

The aggregative policy allows to retrieve intelligible rules that better characterize each output class with respect to approaches following the divide-and-conquer strategy. As a matter of fact, clustering examples of the same kind permits to extract knowledge regarding similarities about the members of a given class rather than information about their differences. This is very useful in many applications and often leads to models showing a higher generalization ability.

LLM and DT represent two of the techniques available in Rulex for the analysis of real world data. In fact, Rulex can efficiently approach and solve supervised (classification, regression) and unsupervised (clustering) machine learning problems by allowing the creation of complex analysis processes through the composition of elementary tasks. A simple but powerful GUI permits to manage datasets providing advanced interactive visualization as well as complete control on the various computational phases. The software suite is in rapid evolution; therefore, the number and the functionalities of available tasks increase every day.

Results and Discussion

The functionalities of Rulex have been verified by considering three biomedical datasets included in the Statlog benchmark (Michie et al., 1994) and concerning as many classification problems:

Diabetes: it concerns the problem of diagnosing diabetes starting from the values of 8 variables; all the 768 considered patients are females at least 21 years old of Pima Indian heritage: 268 of them are cases whereas remaining 500 are controls.

Heart: it deals with the detection of heart disease from a set of 13 input variables concerning patient status; the total sample of 270 elements is formed by 120 cases and 150 controls.

Table 1: Results obtained by the application of five classification algorithms on biomedical datasets included in the Statlog benchmark.

	LLM			DT			NN	LOGIT	KNN
	Accuracy	# Rules	# Cond.	Accuracy	# Rules	# Cond.	Accuracy	Accuracy	Accuracy
Diabetes	76.52%	16	3.75	76.09%	42	4.77	75.65%	76.52%	68.70%
Heart	75.31%	19	4.26	64.20%	17	4.18	72.84%	74.07%	51.85%
Dna	91.98%	19	6.84	90.04%	67	6.26	87.09%	92.57%	40.68%

Dna: it has the aim of recognizing acceptors and donors sites in a primate gene sequences with length 60 (basis); the dataset consists of 3186 sequences, subdivided into three classes: acceptor, donor, none.

Five different classification algorithms have been considered for each dataset (LLM, DT, NN, LOGIT and KNN) and their results are compared both in terms of accuracy of the retrieved solution and of quantity of knowledge extracted from the dataset of examples at hand. For evaluating this last aspect the intelligibility of the rule set, measured by the number of rules and by the average number of conditions for each of them, has been taken into account.

Table 1 shows these two values for the rule sets produced by DT and LLM in the three considered datasets. To evaluate the quality of the resulting models the accuracy obtained by each method on an independent test set including 30% of data has also been reported.

Acknowledgements

This work has been partially supported by the Italian MIUR Flagship Project "InterOmics".

References

1. Boros E et al. (2000) An implementation of logical analysis of data. *IEEE Transactions on Knowledge and Data Engineering*, 12(2):292-306.
2. Duda RO, Hart PE, Stork DG (2001) *Pattern Classification*. New York: John Wiley & Sons.
3. Michie D et al. (1994) *Machine Learning, Neural, and Statistical Classification*. London: Ellis-Horwood.
4. Muselli M (2006) Switching neural networks: A new connectionist model for classification. In *WIRN/NAIS 2005*, vol. 3931 of *Lecture Notes in Computer Science*. Eds. Apolloni B et al., Berlin: Springer-Verlag, 23-30.
5. Muselli M, Ferrari E (2011) Coupling Logical Analysis of Data and Shadow Clustering for partially defined positive Boolean function reconstruction. *IEEE Transactions on Knowledge and Data Engineering* 23(1):37-50.

Data modeling: the key to biological data integration

François Rechenmann✉

Genostar Bioinformatics Solutions, Montbonnot, France

Motivation and Objectives

The advent of NGS technologies is focusing much of the attention towards the data management issue. However, more than their volume, it is the diversity of biological data which constitutes the real bioinformatics bottleneck; a bottleneck which cannot be solved through technological considerations only, such as cloud infrastructures for instance.

A bioinformatics platform must indeed store, organize and give access to a wide span of data and results. First of all, the experimental data and their transformations: not only the sequence data, such as the reads, the assembly files and the resulting contigs - to name the most important ones, but also spectra or metabolic flux measurements. Through the interpretation of these data, biological entities are predicted and characterized: coding regions, regulatory signals, polypeptides, enzymes classes, peptide tags, and so on. All these entities must also be properly described, connected each other and stored in adequately structured data repositories.

Conversely, the programs that implement the analysis algorithms must be able to access these data and these entity descriptions to produce new secondary data and predict new entities.

In this context, conceptual data modeling appears to be the very first task any project aiming at the design and the development of a bioinformatics integrated platform should perform.

Methods

Fortunately, computer scientists have designed a wide range of data modeling tools. Regarding databases, the relational data model provides both a formal well-founded framework, but also leads to efficient implementations as relational database management systems (DBMS). Regarding programs, object-oriented languages, such as Java or C++, allow for the definition of classes and subclasses that can capture the description of the entities these programs deal with.

Moreover, conceptual modeling tools allow designing a data schema independently of its implementation as relations or object hierar-

chies. Inspired by the entity-relationship model, UML is one of these modeling tools. Formally defined, UML offers a graphical view of a schema that is quite intuitive and may therefore be used during the design phase. Biologists, bioinformaticians and computer engineers can indeed efficiently interact on a shared UML diagram which progressively converges towards a consensual schema.

The Genostar bioinformatics platform for microbial genome annotation and comparison has been designed along these principles. The entities the various software modules had to handle have been identified, together with their relationships. Conceptual differences have been carefully taken into account. As an example, chromosomes, plasmids or segments are subclasses of the class "replicon", while reads and contigs are subclasses of "sequence". When sequences are annotated, features are added onto them. The class "feature" is the superclass of a quite deep hierarchy of classes and subclasses: genes, signals, etc. For example, an object CDS (i.e. coding sequence) is described through a list of variables (or fields), and is connected to the supporting sequence by a relation which is itself described by variables. One of them is the variable "location", which specified where the feature is located on the sequence. This information could not be stored in the feature, nor in the sequence, since a feature may be located on different subsequences, obtained through cut and paste operations from a common sequence.

All the methods provided by the software are also described as classes; the variables of a class are the input and output of the methods. Thus, the type of a variable in such a method description makes reference to the class of entity which can be accepted as values of this variable. This typing mechanism allows the software to check that the input data to a method are consistent with the method description. It also associates useful information to the results the method produces. Such consistency verifications are very useful in a dedicated integrated software platform, which is used by users who

wish to concentrate on their analysis process and not on software manipulation.

Results and Discussion

Once the complete data schema has been obtained, it has been implemented in a home-made entity-relationship modeling framework, AROM. This additional modeling level presents several advantages over the direct implementation of the schema as Java classes. It indeed supports query facilities. The software thus offers a large set of built-in queries over the entities and relations on a current workspace. A first example of such a query consists in selecting genes which, according to the computation of their sequence similarity, turned out to be specific to a strain within a set of strains. A second example is retrieving the genes which code for enzymes which catalyze a set of reactions which have been selected on a metabolic map. Moreover, specific queries can be expressed by the user. The software also supports browsing facilities, to explore the connections between the objects. The user can for instance follow the links between a gene, its product, its catalytic functions if any, the reactions it catalyzes, and the metabolic pathways in which these reactions occur.

A nearly identical data schema has been implemented in a relational DBMS. The resulting database MicroB thus integrates genomic, proteic and metabolic data on more than 1500 microorganisms, mainly bacteria at the present time. Since the two schemas, of the software and the database, nearly overlap, data exchanges between these two components are fluid and efficient. Reference data can be extracted from MicroB for comparative analyses in the software; conversely, fully annotated genomes can be stored in MicroB to be later retrieved. SQL queries can be expressed on the contents of the database, but built-in queries together with a dedicated user interface are provided for handling the most standard cases.

The association of the software module dedicated to genome annotation and comparison with the microbial database results in a powerful easy-to-use integrated bioinformatics platform. The explicit representation of objects and their relations offers friendly browsing and querying facilities, helpful type checking, and more generally efficient data management. The user of the platform can concentrate on the data analysis process and forgive all the time consuming and error prone issues resulting from data format conversion and method integration.

But computational biology is a fast evolving scientific domain. New types of data appear, new bioinformatics methods are designed, new methodologies emerge. Again, more than the volume of data, these increasing diversity and complexity appear to be the actual critical issues. Computational biology is a multidisciplinary domain. Multiple interpretations of a concept are frequent and must be resolved when designing software and databases. In this context, explicit and formal data modeling provides very appropriate tools for integrating heterogeneous data, for properly connecting methods and data, and for allowing computer scientists, bioinformaticians and biologists to interact fruitfully over an explicit data schema.

References

1. Peter P. Chen, The Entity-Relationship Model: Toward a Unified View of Data, *ACM Transactions on Database Systems*, Vol. 1, pp. 9-36, 1976
2. Michel Page et al., Knowledge representation with classes and associations: the AROM system (in French: "Représentation de connaissances au moyen de classes et d'associations : le système AROM"), LMO 2000, Actes des journées Langages et Modèles à Objets, Mont Saint-Hilaire, Québec, Canada; Christophe Dony, Houari A. Sahraoui (Eds.), January 25-28, 2000
3. Technical information on AROM, in English, at <http://www.inrialpes.fr/romans/arom/>
4. UML: Unified Modeling Language, <http://www.uml.org/>
5. GenoStar: A Bioinformatics Platform for Exploratory Genomics, François Rechenmann, ERCIM News, No. 51, October 2002, http://www.ercim.eu/publication/Ercim_News/enw51/

SHORT ORAL COMMUNICATIONS

BioQuery-ASP: querying biomedical databases and ontologies using answer set programming

Esra Erdem[✉], Umut Oztok

Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul, Turkey

Motivation and Objectives

Storing biomedical data in various structured forms, like biomedical databases or ontologies, and at different locations have brought about many challenges for answering complex queries about the knowledge represented in these resources. For instance, here are two queries about some genes, drugs and diseases: "What are the drugs that treat the disease Depression and that do not target the gene ACYP1?", "What are the 3 most similar drugs that target the gene DLG4?" One of the challenges of answering such complex queries is to represent the queries in a natural language and present the answers in an understandable form. Another challenge is to efficiently find answers to complex queries that require appropriate integration of relevant knowledge stored in different places and in various forms, and/or that require auxiliary definitions, such as, chains of drug-drug interactions, cliques of genes based on gene-gene relations, similarity/diversity of genes/drugs. Furthermore, once an answer is found for a complex query, the experts may need further explanations about the answer. We have developed novel computational methods and built a software system, called BioQuery-ASP, to handle all these challenges

Methods

We have addressed the challenges described above using a declarative programming paradigm, called Answer Set Programming (ASP) (Lifschitz, 2008; Brewka et al., 2011). ASP provides an expressive high-level knowledge representation formalism that allows recursive definitions, aggregates, default negation, etc. and efficient automated reasoners, such as Clasp (Gebser et al., 2007), which has recently won first places at ASP and SAT (Boolean Satisfiability) competitions in automated reasoning. Due to these attractive features, ASP has been used in various applications, such as phylogeny reconstruction (Brooks et al., 2006), systems biology (Gebser et al., 2011), service robotics (Aker et al., 2012), deci-

sion support systems (Nogueira et al., 2001), automatic music construction (Boenn et al., 2009), workforce management (Ricca et al., 2012).

To address the first challenge (i.e., representing queries in natural language), we have developed a controlled natural language (called BioQuery-CNL) for biomedical queries about drug discovery (Erdem and Yeniterzi, 2009; Oztok 2012). For instance, the queries above are in BioQuery-CNL. Then we have built an intelligent user interface that allows users to enter biomedical queries in BioQuery-CNL and that presents the answers with links to related webpages (Erdem et al., 2011b). Queries in BioQuery-CNL are translated into a set of ASP rules by a novel algorithm. For instance, the first query above is translated into the following ASP rules:

```
what _ drug(DRG) <-
  drug _ name(DRG),
  drug _ treats _ disease(DRG,"Depression"),
  not drug _ targets _ gene(DRG,"ACYP1")
```

which describe the drugs DRG that treat the disease Depression and that do not target the gene ACYP1.

To address the second challenge (i.e., efficiently answering complex queries), first we have developed a rule layer over biomedical ontologies and databases that not only integrates the concepts in these knowledge resources but also provides definitions of auxiliary concepts (Bodenreider et al., 2008). For instance, the predicate `drug _ treats _ disease` is defined in the rule layer as follows:

```
drug _ treats _ disease(DRG,DIS) <-
  drug _ treats _ disease _ pkb(DRG,DIS)
drug _ treats _ disease(DRG,DIS) <-
  drug _ treats _ disease _ ctd(DRG,DIS)
```

integrating the knowledge extracted from the knowledge bases PharmGKB (McDonagh et al., 2011) and CTD (Davis et al., 2011), about "which drug treats which disease." The auxiliary concept

of "chains of gene-gene relations" is defined recursively in the rule layer as well:

```
gene_reachable_from(X,1) <-
  gene_gene(X,Y),
  start_gene(Y)
gene_reachable_from(X,N+1) <-
  gene_gene(X,Z),
  gene_reachable_from(Z,N),
  N < L, max_chain_length(L)
```

to be able to answer queries like "What are the genes related to the gene ADRB1 via a gene-gene relation chain of length at most 3?" Then, for an efficient query answering, we have introduced an algorithm to identify the relevant parts of the rule layer and the knowledge resources with respect to the given query, and used automated reasoners of ASP to answer queries considering these relevant parts (Erdem et al., 2011a). Essentially, our algorithm identifies the relevant predicates that the query-predicates depend on (using a "dependency graph"), and considers the rules that contain these relevant predicates. For some queries, the relevant knowledge consists of about 500 thousand rules whereas the total size of all the knowledge resources (with the rule layer) is over 21 million rules; considering the relevant rules only decreases the computation time of answering a query by almost a factor of 100.

To address the third challenge (i.e., generating explanations), we have developed an intelligent algorithm to generate an explanation (i.e., a tree of "applicable" ASP rules) for a given answer, with respect to the query and the relevant parts of the rule layer and the knowledge resources. We have also developed algorithms to generate shortest/different explanations for a biomedical query taking into account the provenance information as well (Oztok 2012). For instance, an answer to the query "What are the genes that are targeted by the drug Epinephrine and that interact with the gene DLG4?" is ADRB1; and a shortest explanation that justifies this answer is as follows: "The drug Epinephrine targets the gene ADRB1 according to CTD and the gene DLG4 interacts with the gene ADRB1 according to BioGrid."

Based on these methods, we have developed a software system, BioQuery-ASP, that guides the user to represent a complex query in a natural language, finds answers to the query (if an answer exists), returns links to related web pages for

further information, and generates explanations (if the user asks for one). A demo of BioQuery-ASP is available at BioQuery-ASP Website: <http://krr.sabanciuniv.edu/projects/BioQuery-ASP/> (Last accessed on September 25, 2012)).

Results and Discussion

We have shown the applicability of BioQuery-ASP to answer complex queries that are specified by experts, over large biomedical knowledge resources about genes, drugs and diseases, such as PharmGKB, DrugBank (Knox et al., 2011), BioGrid (Stark et al., 2006), CTD, Sider (Kuhn et al., 2010), etc., using efficient solvers of ASP. BioQuery-ASP could find answers to most of the complex queries in 3-10 CPU seconds, over 10 million facts extracted from these knowledge resources and over 10 million rules integrating them (using a computer with two 1.60GHz Intel Xeon E5310 Quad-core Processors and 16GB RAM).

No existing biomedical query answering systems (e.g., web services built over the available knowledge resources, which answer queries by means of keyword search) can directly answer such queries, or can generate explanations for answers. In that sense, BioQuery-ASP is a novel biomedical query answering system that can be useful for experts in automating deep reasoning about knowledge about genes, drugs and diseases available via various biomedical databases and ontologies.

Acknowledgements

This work has been partly supported by TUBITAK Grant 108E229.

References

1. Aker E, Patoglu V, et al. (2012) Answer Set Programming for Reasoning with Semantic Knowledge in Collaborative Housekeeping Robotics. In Proc. of the 10th IFAC Symposium on Robot Control.
2. Bodenreider O, Coban Z, et al. (2008) A preliminary report on answering complex queries related to drug discovery using answer set programming. In Proc. of the 3rd International Workshop on Applications of Logic Programming to the Semantic Web and Web Services.
3. Boenn G, Brain M, et al. (2009) Anton: Composing logic and logic composing. In Proc. of the 10th International Conference on Logic Programming and Nonmonotonic Reasoning, pages 542-547.
4. Brewka G, Eiter T, et al. (2011) Answer set programming at a glance. *Communications of ACM* 54(12):92-103.
5. Brooks DR, Erdem E, et al. (2006) Inferring Phylogenetic Trees Using Answer Set Programming. *Journal of Automated Reasoning* 39(4): 471-511.

6. Davis AP, King BL, et al. (2011) The Comparative Toxicogenomics Database: update 2011. *Nucleic Acids Research* 39(Database issue):D1067-72.
7. Erdem E, Erdem Y, et al. (2011a) Finding answers and generating explanations for complex biomedical queries. In Proc. of the 25th Conf. on Artificial Intelligence (AAAI), pages 785-790.
8. Erdem E, Erdogan H, et al. (2011b) BioQuery-ASP: Querying biomedical ontologies using answer set programming. In Proc. of RuleML2011@BRF Challenge.
9. Erdem E and Yeniterzi R (2009) Transforming controlled natural language biomedical queries into answer set programs. In Proc. of BioNLP Workshop, pages 117-124.
10. Gebser M, Kaufmann B, et al. (2007) clasp: A Conflict-Driven Answer Set Solver. In Proc. of the 9th Int'l Conference on Logic Programming and Nonmonotonic Reasoning, pages 260-265.
11. Gebser M, Schaub T, et al. (2011) Detecting inconsistencies in large biological networks with answer set programming. *Theory and Practice of Logic Programming*, 11(2):1-38.
12. Knox C, Law V, et al. (2011) DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Research* 39(Database issue):D1035-41.
13. Kuhn M, Campillos M, et al. (2010) A side effect resource to capture phenotypic effects of drugs. *Molecular Systems Biology* 6:343.
14. Lifschitz V (2008) What Is Answer Set Programming? In Proc. of the 23rd Conference on Artificial Intelligence (AAAI), pages 1594-1597.
15. McDonagh EM, Whirl-Carrillo M, et al. (2011) From pharmacogenomic knowledge acquisition to clinical applications: the PharmGKB as a clinical pharmacogenomic biomarker resource. *Biomarkers in Medicine* 5(6):795-806.
16. Nogueira M, Balduccini M, et al. (2001) An A-Prolog decision support system for the space shuttle. In Proc. of the 3rd Int'l Symposium on Practical Aspects of Declarative Languages, pages 169-183.
17. Oztok U (2012) Generating Explanations for Complex Biomedical Queries. M.S. Thesis, Sabanci University.
18. Stark C, Breitzkreutz BJ, et al. (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Research* 34(Database issue):D535-9.
19. Ricca F, Grasso G, et al. (2012) Team-building with answer set programming in the Gioia-Tauro seaport. *Theory and Practice of Logic Programming* 12(3):361-381.

A strategy to reduce technical variability and bias in RNA sequencing data

Francesca Finotello¹✉, Enrico Lavezzo², Luisa Barzon², Paolo Mazzon¹, Paolo Fontana³, Stefano Toppo², Barbara Di Camillo¹

¹Department of Information Engineering, University of Padova, Padova, Italy

²Department of Molecular Medicine, University of Padova, Padova, Italy

³Edmund Mach Foundation, San Michele all'Adige, Trento, Italy

Motivation and Objectives

In the last decade, Next-Generation Sequencing (NGS) technologies have been extensively applied to quantitative transcriptomics, making RNA sequencing (RNA-seq) a valuable alternative to microarrays for measuring and comparing gene transcription levels (Wang et al., 2009). In this framework, the millions of sequences obtained through NGS are aligned to a reference genome or transcriptome, and *counts*, i.e. the number of reads aligned to each gene, give a digital measure of gene expression. Given that longer genes are more likely to be sequenced than shorter ones, gene counts depend not only on the true gene expression, but also on its sequence length. Several approaches have been explored to reduce length bias a posteriori, namely after that read counts have been computed (Mortazavi et al., 2008; Bullard et al., 2010; Hansen et al. 2012; Risso et al., 2011), or to provide a direct and unbiased estimate of transcript abundances (Trapnell, 2010). In addition, counts are biased toward highly transcribed genes, so most of the reads sequenced in a sample arise from a restricted subset of highly expressed genes (Robinson and Oshlack, 2009).

The present work is aimed at assessing technical variability and biases of RNA-seq counts, and exploring an alternative measure of exon expression, which is less biased toward long or highly expressed genes, thus requiring no length normalization, and characterized by a lower technical variability.

Methods

We consider two different experiments (Bullard et al., 2010; Griffith et al., 2010) with multiple technical replicates. Raw reads were aligned to the reference genomes using TopHat v1.2.0 (Langmead et al., 2009) and summarized on Ensembl exons using bedtools 2.15.0 (Quinlan and Hall, 2010) to compute read counts. We consider exon counts rather than transcript counts to

avoid introducing biases when dealing with alternatively spliced exons. We computed counts as the total number of reads that align to an exon (referred as *totcounts* in the following). As an alternative approach, we exploited the per-base read coverage to obtain counts for every position along each exon sequence. The measure of gene expression assigned to an exon, called *maxcounts* from here on, was then calculated as the maximum of its per-base counts. Both *totcounts* and *maxcounts* were normalized with the Trimmed Mean of M-values approach (TMM, Robinson and Oshlack, 2009) to correct differences in sequencing depth across libraries. In addition, we computed *Reads Per Kilobase of exon model per Million mapped reads* (RPKM, Mortazavi et al, 2008), calculated by dividing *totcounts*, not normalized via TMM, by the total number of reads mapped in each library, in millions, and by exon length, in kilobases.

Results and Discussion

To investigate the bias due to highly expressed exons, we computed cumulative counts for all replicates in MAQC-2 and Griffith's data sets. In MAQC-2 data (results not shown), when considering *totcounts*, about 3-5% of exons account for 50% of total exon counts and 27-32% of exons account for 90% of total exon counts, showing that a great fraction of counts belong to a restricted subset of exons. Differently, *maxcounts* are more evenly distributed across exons: 7-8% of exons account for 50% of total counts and 44-45% of exons account for 90% of total counts. RPKM distribution lies in between that of *maxcounts* and *totcounts*, with 5-7% of exons accounting for 50% of total RPKMs and 36-38% of exons accounting for 90% of total RPKMs. Also with Griffith's data (Figure 1A), *maxcounts* have the less steep cumulative distribution curves.

We also investigated length bias at single-exon level using smoothed scatter plots of counts/RPKMs versus exon-length, in log-log scale (see

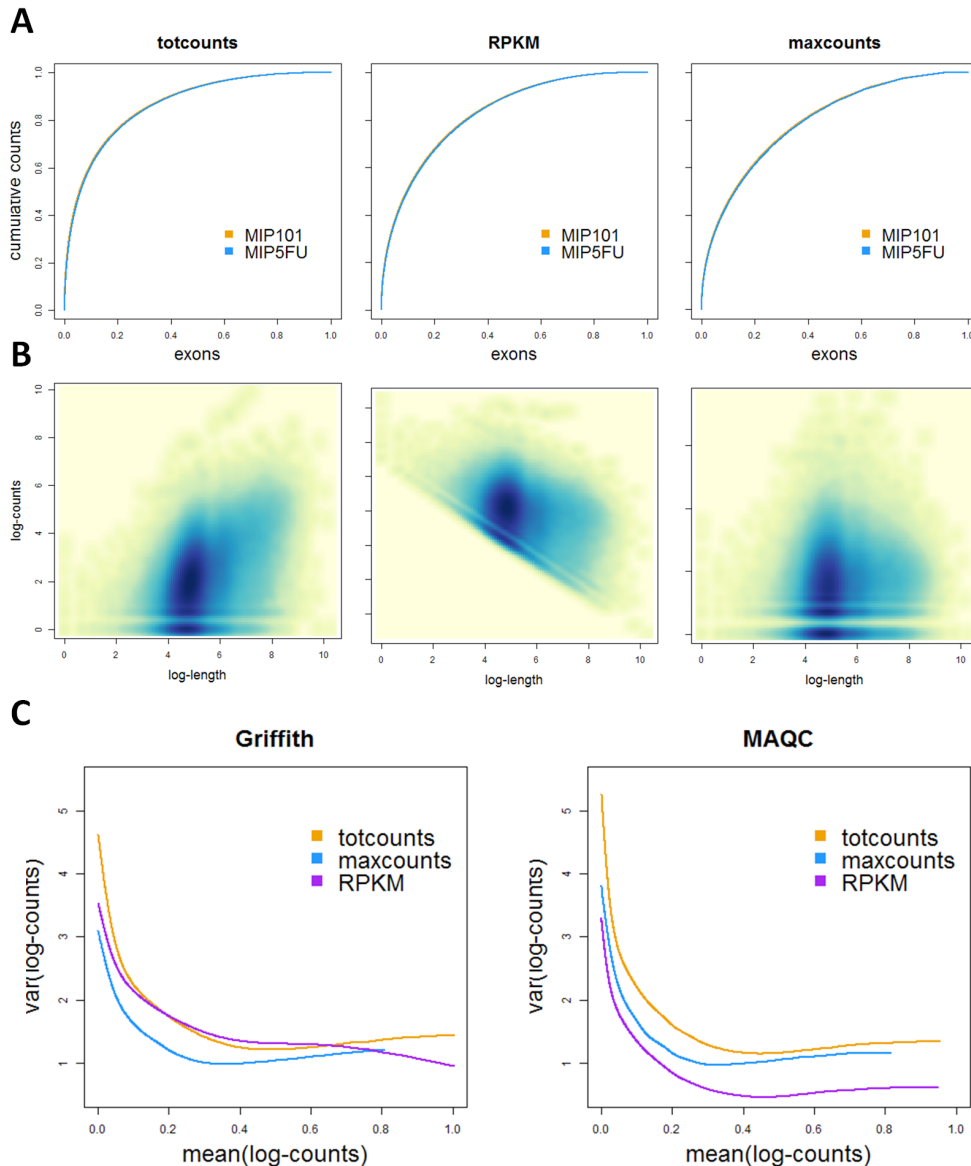


Figure 1: Diagnostic plots of *totcounts*, RPKMs and *maxcounts*: (A) distribution of exon counts/RPKMs in Griffith's data; (B) smoothed scatter plots showing dependence of counts/RPKMs over exon length for one Griffith's library; (C) variance of counts/RPKMs across technical replicates.

Figure 1B for results on Griffith's data). These plots show an increasing pattern of *totcounts* in dependence of exon-length, meaning that longer exons tend to have higher counts than shorter ones (Pearson's correlation $r=0.38$ for MAQC-2 and $r=0.43$ for Griffith). On the contrary, *maxcounts* are not correlated with exon-length (Pearson's correlation $r=0.10$ for MAQC-2 and $r=0.01$ for Griffith). RPKMs do not show the increasing pattern of *totcounts*, and are in fact characterized by negative correlation with exon length (Pearson's correlation $r=-0.28$ for MAQC-

2 and $r=-0.29$ for Griffith), meaning that dividing by exon length over-corrects length bias in shorter exons. Plots are reported for one library of Griffith's data set, but the same patterns are confirmed across all libraries of the two data sets (results not shown).

Finally, we assessed variance of *totcounts*, RPKMs and *maxcounts* across technical replicates, using a cubic-spline fit of the variance versus the mean of log-counts/log-RPKMs (Figure 1C); in both data sets *maxcounts* have a lower variance with respect to *totcounts*. Anyway, on

MAQC-2, RPKMs provide the lowest technical variance.

In summary, we confirm that *totcounts* strongly depends on the length of the feature they are summarized on, even when considering exons in place of genes. Using RPKMs, that normalize *totcounts* by exon length and sequencing depth, reduces technical variability but does not completely remove exon length bias. We propose an alternative measure of exon expression, *maxcounts*, which is less biased toward long or highly expressed genes than *totcounts* and RPKMs, and whose technical variability is lower than or comparable to that of *totcounts* and RPKMs, respectively.

We are now working on a refinement of this measure, to make it more robust to sequencing and mapping biases. In addition, we are assessing the accuracy and precision of *totcounts* and *maxcounts* in assessing the real RNA abundances using publicly available data sets for which spike-in RNAs measures are available. Future studies will focus on the definition of transcriptional models that could be used to aggregate *maxcounts* at gene or transcript level.

Acknowledgements

This research is supported by PRAT 2010 CPDA101217, "Models of RNA sequencing data variability for quantitative transcriptomics", and AACSE Project, "Algorithms and Architectures for Computational Science and Engineering".

References

1. Bullard JH, Purdom E, Hansen KD, et al. (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments. BMC Bioinformatics 11, 94. doi: [10.1186/1471-2105-11-94](https://doi.org/10.1186/1471-2105-11-94)
2. Griffith M, Griffith OL, Mwenifumbo J, et al. Alternative expression analysis by RNA sequencing. Nat Methods 7, 843. doi: [10.1038/nmeth1503](https://doi.org/10.1038/nmeth1503)
3. Hansen KD, Irizarry RA, Wu Z (2012) Removing technical variability in RNA-seq data using conditional quantile normalization. Biostatistics 13, 204. doi: [10.1093/biostatistics/kxr054](https://doi.org/10.1093/biostatistics/kxr054)
4. Langmead B, Trapnell C, Pop M, et al. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10, R25. doi: [10.1186/gb-2009-10-3-r25](https://doi.org/10.1186/gb-2009-10-3-r25)
5. Mortazavi A, Williams BA, McCue K, et al. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat methods 5, 621. doi: [10.1038/nmeth.1226](https://doi.org/10.1038/nmeth.1226)
6. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841. doi: [10.1093/bioinformatics/btq033](https://doi.org/10.1093/bioinformatics/btq033)
7. Risso D, Schwartz K, Sherlock G, et al. (2011) GC-Content Normalization for RNA-Seq Data. BMC Bioinformatics 12, 480. doi: [10.1186/1471-2105-12-480](https://doi.org/10.1186/1471-2105-12-480)
8. Robinson MD and Oshlack A (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol, 11, R25.
9. Trapnell C, Williams BA, Pertea G, et al. (2010) Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28, 511. doi: [10.1038/nbt.1621](https://doi.org/10.1038/nbt.1621)
10. Wang Z, Gerstein M, Snyder M. (2009) RNA-seq: A revolutionary tool for transcriptomics. Nat Rev Genet. 10, 57. doi:[10.1038/nrg2484](https://doi.org/10.1038/nrg2484)

Applications of a generic model of genomic variations functional analysis

Sarah Mapelli, Uberto Pozzoli[✉]

Scientific Institute I.R.C.C.S. "E. Medea", Bosisio Parini (LC)

Motivation and Objectives

Deep sequencing techniques, as well as the inherent equipment, are dramatically increasing their popularity in many scientific communities such as computational biology, "omics" and clinical research groups. The reasons are both scientific and budgetary: new experiments can be performed at a steadily decreasing cost. Nevertheless there are technical issues still to be addressed to make the results really useful in all the communities. We limit our interest to the "final" results of these experiments, usually a set of DNA/RNA variants or annotations relative to some reference assemblies. Except for those who are studying and developing algorithms and tools to produce them, these data are what people in different fields have to deal with. They are necessarily big and organized in a way that doesn't simplify their interpretation in terms of functional effect on phenotype. We speculated that a conceptual model of the connections between

these data and the "genomic objects" usually studied (i.e. transcripts, miRNA, chromosomes, and so on) can be useful to make analyses and could greatly simplify the development of tools and programs. After defining such a model we implemented it, as well as a number of utilities. The result of this work is a C++ library (namely GeCo++: Genomic Computation C++ library) still actively developed but already used in our institute.

Methods

In the GeCo++ library, a genomic reference is defined as a portion of DNA identified by a name. A genomic element is then defined as an interval with a given strand along a genomic reference. Positions along a genomic reference are defined as zero based unsigned values. Element positions are defined relatively to the beginning of a given element and therefore are represented by signed zero based values. Intervals for both references and elements are considered as

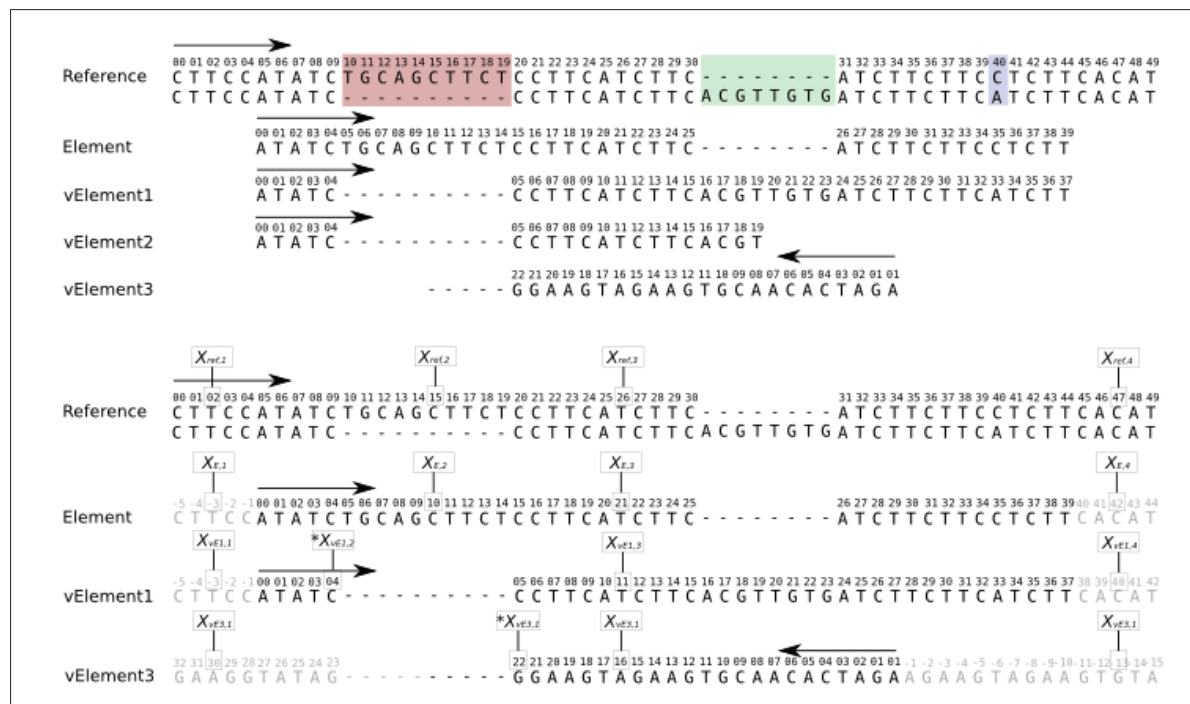


Figure 1: Reference, elements and variated elements definitions (upper panel). Positions and Interval mapping conventions are illustrated in the lower panel.

right open ones. A genomic element instance is defined by its reference boundaries (interval), its strand and possibly by a number of variations (insertions, deletions and substitutions) relative to the reference.

One of our goals was to allow forth and back conversion between element and reference positions. A set of simple and consistent rules has been defined and is applied throughout the library in order to allow interval/position conversion and mapping between reference and varied elements. Some examples of conversion from reference to element positions and intervals are reported in Figure 1.

"Sites" can be added to Element objects. Sites are positions along the element that have a particular biological meaning depending on the application one is developing. As well as sites, "Connections" can be added, representing meaningful directed links between two sites. Finally we define "Features" as numerical properties whose value varies along the element. Let's also define a "Feature Calculator" as an object that implements some algorithms to calculate/retrieve a specific feature along an element.

The objects defined above have been implemented in the GeCo++ library (Cereda et al., 2011). C++ was chosen because it's object oriented, faster than other languages (especially interpreted ones) and because a great number of high quality computational biology C++ libraries do exist and can be readily included in C++ programs.

Further implementations, with respect to the published core, include the definition of the class "Genotypes" as an object intended to represent genotype information deriving from both resequencing or genotyping experiments: it does so in terms of differences from a given reference. This object closely resembles to the kind of information one can find in a VCF (Variant Call Format) file. At this point we can formally define a pseudo function to assess the effect of one or more variations on a given feature along a genomic element as:

"mutated" element = method ("reference" element, genotypes , feature calculator)"

The resulting element allows to easily compare the feature values with the original ones even when insertions and deletions are present. Furthermore, this comparison is independent from the algorithm used to calculate the feature

and therefore the subsequent analyses can be performed in the same way for a given element type independently from the algorithm used to calculate the feature. Also, the genotypes object provided can derive from any kind of experiment: in this way, for example, it is straightforward to apply our function to data coming from Sanger sequencing to confirm NGS results.

A series of more specialized classes and functions have also been added to the library to retrieve elements from different sources (i.e. UCSC, Ensembl, gff files), to calculate a variety of features (PWM scores, RNA secondary structure) and to perform statistical tests on genotype information. To this purpose a database structure has been defined to hold genotype information that can be accessed through the genotypes class. In this way we can read genotype information from the VCF file resulting from a whole genome multi-sample experiment, store it in the database and later retrieve the information relative to the region/element and samples of interest.

By using the library it is particularly fast and easy to produce applications that implement complex tasks by using the method abstraction. Since C++ is not the most popular language (especially for those who have a biological background) we also developed a simple and lightweight framework which can produce command line applications that can be called from the R statistical package (R Core Team, 2012) and as web services. In this latter case a simple javascript library implements an Ajax interface.

Results and Discussion

The library is extensively used in our lab, we therefore have a way to store NGS as well as Sanger experiment results in a database. We also can analyze the results in different ways: from genome wide population genetics studies to single gene analyses performed by biologists in the molecular biology lab.

A set of applications has been developed to insert variations in the database and to analyze them. In particular through an application called deLorean it is possible to apply a variety of population genetics statistics to resequencing data. This application has been used in several population genetics studies recently published by our group to analyze the 1000 genomes project data. From the functional analysis point of view, a still provisionally named "testPWM" ap-

plication can evaluate variation effects on PWM scores of any JASPAR (Bryne et al, 2008) PFAM TFBS (Transcription Factor Binding Sites) matrix for any resequenced region in the database. Another application for which a web interface does already exist, allows researchers in our Institute to annotate their NGS sequencing variants with respect to a list of transcripts or to the transcripts overlapping a given genomic region.

The applications developed so far are extensively tested (especially some of the utilities) by all groups in our institute, some part still need to be fully developed and an effort should be made in the near future to exploit the parallel computing opportunities offered by the modern hardware. The library, as well as the applications are available upon request.

Acknowledgements

This research was funded by the Italian ministry of Health. We wish to thank Matteo Cereda for his great ideas and contribution to the library development.

References

1. Cereda M, Sironi M, et al. (2011) GeCo++: a C++ library for genomic features computation and annotation in the presence of variants. *Bioinformatics*. 1;27(9):1313-5. doi:10.1093/bioinformatics/btr123
2. R Core Team, (2012) R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing Vienna, Austria, ISBN 3-900051-07-0, <http://www.R-project.org>
3. Bryne JC, Valen E, et al. (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update., *Nucleic Acids Res.* 36 (Database issue):D102-6. doi:10.1093/nar/gkm955

Ranking-aware integration and explorative search of distributed bio-data

Marco Masseroli✉, Matteo Picozzi, Giorgio Ghisalberti

Dipartimento di Elettronica e Informazione, Politecnico di Milano, Milan, Italy

Motivation and Objectives

High-throughput production of both biomolecular data and their annotations is providing a rapidly increasing amount of very valuable information that can potentially help finding also long-searched answers to fundamental biomedical questions. Yet, such data deluge makes difficult to extract the information most reliable and most related to the increasingly complex biomedical questions to be answered, which can simultaneously regard many heterogeneous aspects of single or multiple organisms, biological tissues, cells or biomolecular entities. To address such complex questions, many bio-data about several heterogeneous topics, which are available but dispersed in different data sources, must be searched, extracted, integrated and comprehensively queried.

Different approaches have been proposed to combine individual search services available on the Web in order to support such heterogeneous searches (Hull et al., 2006; Nekrutenko, 2010). Yet, they rarely rely on a general model of the services to be integrated and none considers, in the integration process, the often available partial rankings of the data to be integrated. Lately, Search Computing (Ceri et al., 2010) has been proposed as a new software framework to build answers to complex search queries by interacting with a collection of cooperating search services and using ranking and joining of results as the dominant factors for service composition. By leveraging the peculiar features of search services, it offers query approaches, execution plans, plan optimization techniques, query configuration tools, and exploratory user interfaces.

Here, we report and discuss our work aimed at supporting the explorative search of heterogeneous distributed bio-data and the automatic integration and global ranking of their individual search results, also taking into account the partial rankings of individual searches. In so doing, we make a step towards the computational support required for complex biomedical question answering and biomedical knowledge discovery.

Methods

According to the Service Mart modeling approach of Search Computing (Ceri et al., 2010), we selected an initial set of typical biomolecular topics (i.e. Protein, Gene, Gene Expression and Biological Function) and modeled the Service Marts (i.e. the generalized and normalized conceptual description) of the bioinformatics services that provide data regarding such topics. We did so by identifying their main and common attributes and normalizing their names. We also defined the semantic Connection Patterns, i.e. the pair-wise coupling, between Service Marts of services that provide data about different topics. This was done by identifying pairs of normalized attributes of the connected Service Marts and defining their comparison predicates, as conjunctive Boolean expressions, that allow joining their values semantically. In so doing, we defined the Semantic Resource Framework (SRF) depicted in Figure 1, which constitutes the reference used by Search Computing to enable the exploration of the services registered in the framework and integrate the data that they provide (Ceri et al., 2010).

Then, using available Search Computing tools, we registered in the Search Computing framework five bioinformatics search services that provide data about the topics and semantic associations described in the biomolecular

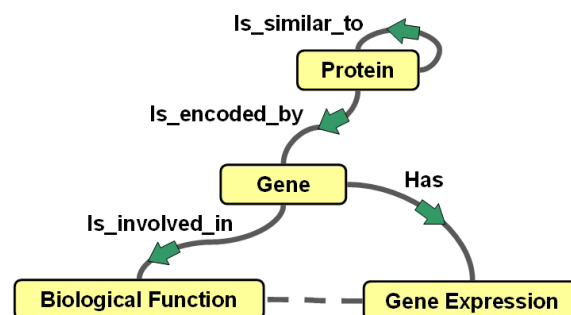


Figure 1: Biomolecular Semantic Resource Framework defined through modeling of data provided by bioinformatics search services and created through service registration. Boxes represent topics of the data provided by the search services registered in the Search Computing framework; lines represent the defined semantic connections created, at registration time, between the registered services.

SRF in Figure 1, i.e. the NCBI Blast (Johnson et al., 2008) and WU-BLAST (Lopez et al., 2003) protein sequence alignment search services, the Array Express gene expression search service (Parkinson et al., 2005), and two access services to the protein coding genes and their biological function annotations (e.g. Gene Ontology annotations) in our Genomic and Proteomic Data Warehouse (GPDW, <http://www.bioinformatics.dei.polimi.it/GPKB/>). Thus, through service registration, the biomolecular SRF in Figure 1, previously described at conceptual level, is created. To do so, for each service, first we created a wrapper, i.e. an adapter that matches the service attributes to their normalized version defined in a modeled Service Mart, and associated the wrapper with such a Service Mart. Then we defined one or more Access Patterns and Service Interfaces for the service. The latter ones map an access pattern to the end point of a concrete service data source, whereas the former ones are specific signatures of a Service Mart, with the characterization of each attribute as input (I) or output (O), depending on the role that the attribute plays in the service call; furthermore an output attribute can be characterized as ranked (R), if the service produces its results in an order that depends on the value of that attribute. An example Access Pattern for the GPDW Gene to Biological Function Feature (BFF) service is:

```
(GPDW_Gene2BFF-Name_byGeneID(GeneID',  
GeneIDName', BFFName', BFFIDo, BFFIDNameo,  
BFFNameo, BFFDefinitiono)
```

Specific Connection Patterns between individual registered services are then automatically derived from the Connection Patterns defined at conceptual level between the modeled Service Marts that have been associated with the registered services.

Results and Discussion

Leveraging the Search Computing framework and biomolecular SRF, which we constructed as previously reported in (Masseroli et al., 2011) and briefly described in the Methods section, we created the Bio Search Computing (Bio-SeCo) application. In particular, in the work here reported, we modeled and registered in Bio-SeCo two additional services and created a Web interface that offers public access to Bio-SeCo at <http://www.bioinformatics.dei.polimi.it/bio-seco/seco/>. It enables explorative search, automatic integra-

tion and global ranking of bio-data individually provided by the services registered in the framework. In this way and thanks to the additional services integrated, Bio-SeCo supports explorative answering of even more complex biomedical questions and biomedical knowledge discovery.

As an example, let us consider the following complex question: Which are the genes (if they exist) that encode proteins, in different organisms, with high sequence similarity to a protein X and have some biomedical features in common (e.g. up/down significantly co-expressed in the biological tissue or condition Y and involved in the biological function Z)? Using Bio-SeCo, a user can first input the UniProt ID of a protein X and run a sequence alignment search, by using the NCBI Blast or WU-BLAST service, to look for amino acid sequences similar to the protein X in a user selected protein database (e.g. UniProtKB Swiss-Prot). Then, he/she can select the most similar proteins found (or some of them, e.g. only those of some selected organisms) and automatically retrieve the coding gene of each of them by using the GPDW protein coding gene query service. Next, the user can search for biomedical features shared among the retrieved genes. For instance, by using the Array Express and GPDW gene biological function annotation services, he/she can explore if some of such genes are significantly co-expressed in the same biological tissue or condition Y and are known to be involved in the biological function Z. For example, the user can set the human *Paired box protein Pax-6 isoform a* protein (UniProt ID P26367) as input protein X, *tumor* as pathological biological condition Y, and *regulation of apoptotic process* as biological function Z. By doing so, unpredictably, on July 20th 2012, Bio-SeCo discovered the human PAX7 and PAX2, mouse Pax8 and human PAX8 genes, ranked by their global score of 0.90661, 0.90407, 0.90354 and 0.90289, respectively (with 1.0 as best score). This global score is computed by Bio-SeCo according to a score function defined as a combination of partial scores of intermediate ranked results, i.e. of the ranked sequence alignment expectation and gene expression p-value. To compute the global score, we adopted the Fagin method (Fagin et al., 2004), which resulted to be very fast and less computationally demanding than a recently proposed and very promising approach for ranking composition (Cohen-Boulakia et al., 2011). The 4 genes found

encode, respectively, the human Paired box protein Pax-7, human Paired box protein Pax-2, mouse Paired box protein Pax-8 and human Paired box protein Pax-8 (which have 1.35413×10^{-76} , 1.72295×10^{-70} , 3.22281×10^{-69} and 1.16475×10^{-67} expectation of sequence similarity to the input human Paired box protein Pax-6 isoform a protein) and all 4 genes are significantly co-expressed in tumor with a 1.0×10^{-11} p-value.

As the described methods and results demonstrate, Bio-SeCo provides a public extremely useful automated support for exploratory searches at the base of Life Science data driven knowledge discovery. It enables the user to explore the very large and very heterogeneous bio-data available, allowing he/she to easily make different attempts, inspect obtained partial results and move forward and backward in the construction of the global query that would eventually find the most relevant results, in case after several unsuccessful attempts.

Acknowledgements

This research is part of the Search Computing (SeCo) project (2008-2013) funded by the European Research Council (ERC), IDEAS Advanced Grant.

References

1. Ceri S, Abid A, et al. (2010) Search Computing: an approach for managing complex search queries. *IEEE Internet Comput* 14(6):14-22.
2. Cohen-Boulakia S, Denise A, Hamel S (2011) Using medians to generate consensus rankings for biological data. In: Cushing JB, French J, Bowers S (Eds.) *Scientific and Statistical Database Management*. LNCS, Vol. 6809. Springer, Heidelberg, D, pp. 73-90.
3. Fagin R, Kumar R, et al. (2004) Comparing and aggregating rankings with ties. *Proceedings ACM Symposium on Principles of Database Systems (PODS '04)*. pp. 47-58.
4. Hull D, Wolstencroft K, et al. (2006) Taverna: A tool for building and running workflows of services. *Nucleic Acids Res* 34(Web Server issue):729-732.
5. Johnson M, Zaretskaya I, et al. (2008) NCBI BLAST: a better web interface. *Nucleic Acids Res* 36(Web Server issue):W5-W9.
6. Lopez R, Silventoinen V, et al. (2003) WU-Blast2 server at the European Bioinformatics Institute. *Nucleic Acids Res* 31(13):3795-3798.
7. Masseroli M, Ghisalberti G, Ceri S (2011) Bio-Search Computing: Integration and global ranking of bioinformatics search results. *J Integr Bioinform* 8(2):166, p. 1-9.
8. Nekrutenko A (2010) Galaxy: A comprehensive approach for supporting accessible, reproducible, and transparent computational research in the Life Sciences. *Genome Biol* 11(8):R86.
9. Parkinson H, Sarkans U, et al. (2005) ArrayExpress - A public repository for microarray gene expression data at the EBI. *Nucleic Acids Res* 33(Database issue):D553-D555

DiGSNP: a web tool for Disease-Gene-SNP hierarchical prioritization

Carmen Navarro¹✉, Carlos Cano¹, Armando Blanco¹, Fernando García-Alcalde²

¹Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain

²Max Planck Institute for Infection Biology, Berlin, Germany

Motivation and Objectives

Understanding the genetic causes of human diseases is the major goal towards an effective personalized medicine. High-throughput technologies such as linkage analysis, association studies and array experiments allow to obtain experimental evidence of chromosomal regions associated with phenotypes. However, these technologies typically report a large number of results (i.e. genes, variants, etc.) associated with the conditions under study. In this context, tools supporting researchers in the process of prioritizing diseases, genes, and variations are highly desired to assist the scientific research and provide guidance on the most promising hypotheses. To this end, many gene-disease prioritization methods have been proposed in the literature (Moreau and Tranchevent, 2012). These methods describe computational approaches that use information retrieved from diverse sources in order to obtain prioritized lists of candidate genes to be related with a certain target disease. However, most of these tools do not consider gene variations, although they are known to be the main cause for many diseases. Proposals like AnnTools (Makarov et al., 2012) or SNPRank (Jadamba et al., 2012), based in genome-wide association studies (GWAS), relate variations directly to diseases, but leave gene-disease information out or implicit. Moreover, most currently available tools for associating genome variations to diseases focus on coding regions, disregarding relevant information present in the promoter regions of genes, such as variations that alter the binding affinity of transcription factor binding sites (TFBS), which have been shown to play an important role in the regulatory machinery of the cell.

In this work we present DiGSNP (Disease-Gene-SNP Prioritizer), a tool which allows to relate diseases, genes and variations in regulatory regions of the genome (particularly, those affecting TFBSs), simultaneously, helping researchers to understand how these relations work and focus on the most relevant regulatory elements in the early research stage of any disease.

Methods

DiGSNP prioritizer searches for relations between diseases, genes and variations in a two-level hierarchy. The first level builds an ordered list of genes related to a query disease. The second level adds a list of single-nucleotide polymorphisms (SNP) present in TFBSs in the regulatory regions of each gene (i.e. building a Disease-gene-SNP hierarchy), prioritized by the expected level of influence in the binding affinity of the TFBS.

Disease-gene prioritization method is based on ProphNet (Martinez et al., 2012, <http://genome2.ugr.es/prophnet/>). This method allows to prioritize biological entities from different domains (e.g. genes, diseases, protein domains) by integrating an arbitrary amount of heterogeneous sources of data represented as networks. The resultant super-graph is then mined using a Random Walk with Restart (RWR) algorithm for obtaining prioritized lists of elements associated to the user query. We have applied the ProphNet algorithm to obtain prioritized lists of genes for a query disease. The algorithm was applied on a network composed of three different types of nodes: genes/proteins, phenotypes and protein domains. The phenotype network and the phenotype-gene connections were extracted from OMIM using text-mining techniques; the gene network was obtained from the Human Protein Reference Database (HPRD, <http://www.hprd.org/>) and the protein domain network was derived from DOMINE and InterDom (<http://interdom.i2r.a-star.edu.sg/>), with the domain-gene and domain-phenotype relationships extracted from Pfam (<http://pfam.sanger.ac.uk/>).

After obtaining a prioritized list of genes related to the query disease, each gene is associated to a list of SNPs present in its promoter regions. SNPs are obtained from dbSNP (<http://www.ncbi.nlm.nih.gov/projects/SNP/>). Gene-SNP prioritization is applied to each SNP list based on two criteria. First, SNPs located in a TFBS are candidate regulatory SNPs. Second, a SNP that causes drastic changes in the binding affinity of a TFBS has a higher probability to affect the gene regulation

and therefore to be related to the query disorder. In order to assess whether a SNP is located in a known TFBS, SC_{intuit} (García-Alcalde et al., 2010), a sequence-motif similarity measure, is used. SC_{intuit} applies intuitionistic theory, an extension of fuzzy theory, to generate a similarity score. Motif information is retrieved from Jaspar (<http://jaspar.cgb.ki.se/>) and TRANSFAC. According to the mentioned two criteria, DiGSNP reduces the set of SNPs in regulatory regions of the gene to a set of SNPs located in a site matching a TFBS, i.e. sites showing a SC_{intuit} similarity score to a known TFBS above a provided threshold. Selected SNPs are ordered depending on the difference of similarity between the mutated and wild-type alleles. Therefore, SNPs that dramatically alter the binding affinity of a TFBS are ranked at the top by DiGSNP.

Table 1: Fragment of the information obtained with DiGSNP for Alzheimer disease. A top rank of 5 genes is shown, and for each of them the top ranked 3 SNPs. Each SNPs is also associated with the region of the gene where it was found and the motif that generated its score.

Disease	Gene	SNP	SNP score	Gene region	Motif ID
Alzheimer	APP	rs199610454	0,32	5' UTR	Kid3
		rs201528959	0,27	5' UTR	Churchill
		rs200990709	0,25	5' UTR	HNF4
PSEN2		rs200123803	0,34	5' near gene	ZNF354C
		rs200034334	0,32	5' UTR	Kid3
		rs150618255	0,29	5' near gene	Kid3
PSEN1		rs202004275	0,32	5' UTR	Kid3
		rs201506908	0,29	5' UTR	Kid3
		rs200531676	0,29	5' UTR	MAFB
TREM2		rs113167129	0,34	5' near gene	ZNF333
		rs187797067	0,34	5' near gene	C-MAF
		rs138222305	0,32	5' near gene	Kid3
HD		rs192838728	0,32	5' near gene	Kid3
		rs28616835	0,32	5' near gene	Kid3
		rs398691	0,32	5' near gene	Kid3

Integrating these two prioritization methods, we offer an approach to disease-gene-SNP prioritization. In table 1, a summary result relating Alzheimer disease to a list of prioritized genes and each gene relating to a list of prioritized SNPs is shown. This way of structuring the information

allows the user to visually infer possible relations among genes, diseases, SNPs and TFBSs. For instance, the frequent appearance of motif Kid3 as best result in many SNPs reveals a direct relation of Kid3 and Alzheimer (Acquaah-Mensah et al., 2012), which would have been otherwise difficult to discover.

Results and Discussion

The proposed methodology can be applied to any prioritization method that can score genes relating to diseases and SNPs related to genes. Due to the lack of information sources relating regulatory variations and diseases, validation becomes a difficult process, along with determining the biological impact of the results obtained. Relations between genes in Table 1 such as APP, TREM1 and TREM2 and Alzheimer disease can be found in the literature (Cruchaga et al. 2012). However, finding information in the literature about the SNPs that appear in table 1 was not possible. The main reason is probably that the focus of research relating to SNPs has relied on coding regions, searching for missense variations in exomic areas of the genome. Our main focus is transcriptional regulation, area from which the amount of available information is much less significant. Moreover, other tools, like SNPRank (Jadamba et al., 2012), focus on coding areas of the genome. This makes it unfeasible to compare the results, since the sets of SNPs obtained should always be different. Other approaches like regSNPs (Teng et al. 2012) focus on regulatory elements, but require experimental evidence from a GWAS and make an inverse process, starting from variations proven to be related to a disease by GWAS results. Furthermore, these tools relate directly diseases and variations, leaving gene information out or implicit.

We believe that DiGSNP can be helpful to researchers, who can see in a glance the relationship between a certain disease and a set of SNPs related to genes, probably involved in the regulation processes that affect the target disease. In addition, the second step of DiGSNP focuses on genomic information and our knowledge about TFBSs and their binding affinity, making it possible for researchers to obtain a set of probable candidates for any disease. Evidence of variation-disease association in the literature is not needed for placing a query in DiGSNP. This feature makes DiGSNP a helpful tool when trying to discover a

set of highly related SNPs and genes to a new query disease to boost further research.

Acknowledgements

This work has been carried out as part of projects P08-TIC-4299 of J. A., Sevilla and TIN2009-13489 of DGICT, Madrid

References

1. Acquah-Mensah GK, Taylor RC, et al. (2012) PACAP interactions in the mouse brain: implications for behavioral and other disorders. *Gene*, 491(2):224-231. doi:[10.1016/j.gene.2011.09.017](https://doi.org/10.1016/j.gene.2011.09.017).
2. Cruchaga C, Chakraverty S, et al. (2012). Rare Variants in APP, PSEN1 and PSEN2 Increase Risk for AD in Late-Onset Alzheimer's Disease Families. *PLoS One* 7(2). doi:[10.1371/journal.pone.0031039](https://doi.org/10.1371/journal.pone.0031039).
3. García-Alcalde F, Blanco A, et al. (2010). An intuitionistic approach to scoring DNA sequences against transcription factor binding site motifs. *BMC bioinformatics*, 11(1): 551. doi:[10.1186/1471-2105-11-551](https://doi.org/10.1186/1471-2105-11-551).
4. Jadamba E, Shin M. (2012). A SNP Prioritization Method Using Linkage Disequilibrium Network for Disease Association Study. *INTELLI 2012 (c)*, 86-88.
5. Makarov V, O'Grady T, et al. (2012). AnnTools: a comprehensive and versatile annotation toolkit for genomic variants. *Bioinformatics*, 28(5):724-5. doi:[10.1093/bioinformatics/bts032](https://doi.org/10.1093/bioinformatics/bts032).
6. Martínez V, Cano C, et al. (2012). Network-based gene-disease prioritization using ProphNet. *EMBnet.journal S18.B* (in press)
7. Moreau Y and Tranchevent LC (2012). Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nature Reviews Genetics* 13:523-536. doi:[10.1038/nrg3253](https://doi.org/10.1038/nrg3253).

The BioVel Project: Robust phylogenetic workflows running on the GRID

Saverio Vicario¹✉, Bachir Balech², Giacinto Donvito³, Pasquale Notarangelo³, Graziano Pesole^{2,4}

¹ Istituto di Tecnologie Biomediche, Consiglio Nazionale delle Ricerche, Bari, Italy

² Istituto di Biomebrane e Bioenergetica, Consiglio Nazionale delle Ricerche, Bari, Italy

³ Istituto Nazionale di fisica Nucleare, Bari, Italy

⁴ Dipartimento di Bioscienze, Biotecnologie e Scienze Farmacologiche, Università degli studi di Bari "Aldo Moro", Bari, Italy

Motivation and Objectives

Altered species distributions, the changing nature of ecosystems and increased risks of extinction all have an impact on important areas of social concern. Biologists and environmental scientists are asked to provide decision support for managing biodiversity components of our environment at multiple scales (genomic, organismal, habitat, ecosystem, landscape, etc...) to prevent and mitigate such losses. The BioVel project (www.biovel.eu) aspires to address these needs by offering a series of robust and reliable web services that could be managed with the tools suite of the myGRID project. The project proposes the building of workflows exploiting these services to ensure best practice and efficiency of use. These workflows provide the end users the capabilities to execute application easily accessible through several kind of resources such as EGI grid infrastructure, local batch farm or dedicated servers. Within the first round of services produced by the project, here we describe the phylogenetic inference workflows.

Phylogenetic inference is a summary of the evolutionary history of a group of organisms. The topology summarizes the relationships among the organisms, while branch lengths summarize the expected changes along a given section of them (Felsenstein, 2004). Therefore, phylogeny can be used as a basic tool to summarize biodiversity, categorize groups of organisms and study the impact of environmental change on biodiversity. Unfortunately, almost all phylogenetic methods are computationally intensive and sensitive to misuse (i.e. bad model choice could cause high support for wrong answer). For that, this workflow offers an easy way to use phylogenetic services that will allow a broad adoption of best phylogenetic inference practices in the current work of biodiversity scientists including not only ecologists and environmental scientists (Honeycutt et al., 2010) but also medical doctors interested in studying patients' biome (Cho and Blaser, 2012; Delzenne et al., 2011). In particular, in the field of environmental sequencing processing biosequences within a

phylogenetic context is a preliminary step for both taxonomic annotation and inferring evolutionary process from sequences within or across samples.

The usage of well designed workflow into Taverna workflow management system (Hull et al., 2006), is the key advantage of this work, as it will allow the end users to manage the execution of complex algorithms with simple interaction such as configuring simple parameters (input files and execution options), while the workflow will ensure the use of quality control steps and flag problematic inference at both the alignment level and then at the phylogenetic step itself.

It is important to note that, the implementation of the workflow within a workflow engine and editor publicly available, as in the myGRID suite of tools, allow two important practices to be implemented: 1) detailed peer review of the protocol implemented in a given work and 2) flexible update and/or modification of the workflow by the users without a specific coding capacities.

Methods

The workflow starts from a user defined list of biosequences (DNA/Amino Acids), access an alignment Web Service that implement HMMER3 align algorithm (Eddy, 2011) and uses, conditioned on the biosequences as queries, the correct PFAM as guiding profile chosen with 'HMMER3 scan' function. Using a supplied user threshold, DNA or Amino acids sites with lower posterior probability are filtered out. The alignment loaded in the workflow engine is then formatted to Nexus format. The MrBayes (Huelsenbeck et al., 2001; Altekar et al., 2004) model block is built following user supplied request, while the MCMCMC (Metropolis-coupled Markov Chains Monte Carlo) numerical integration options are in part specified by the user and in other part are fixed to maximize MPI efficiency on the farm system. MCMCMC numerical integration convergence is assessed by GeokS (Battagliero et al., 2011) software that estimates burn-in value and the reached convergence based on the tree parameter.

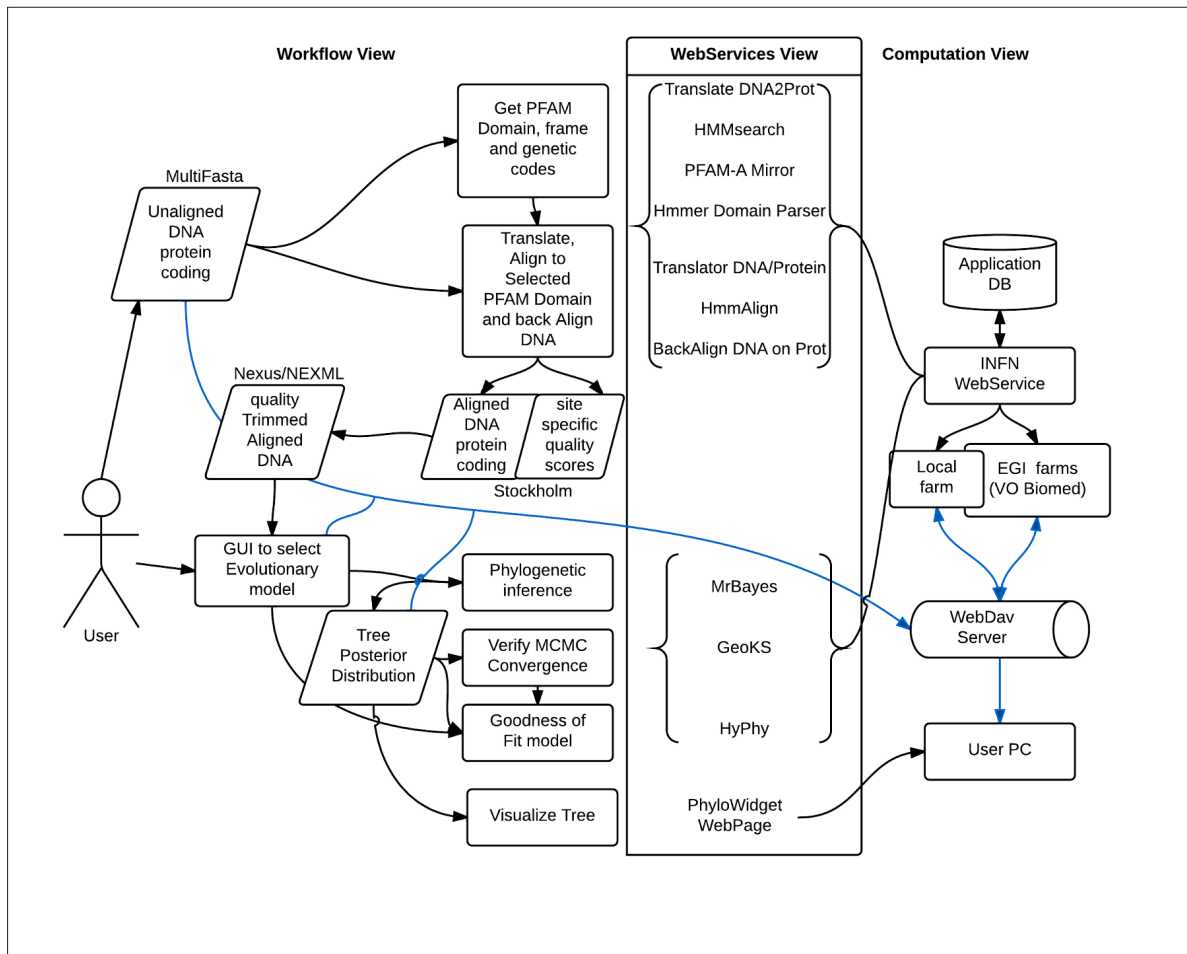


Figure 1. Schema Main Workflow showing underlying web services and computational resources. Gray arrows indicates real data flow, black arrows logic and symbolic link. Table 1: Fragment of the information obtained with DiGSNP for Alzheimer disease. A top rank of 5 genes is shown, and for each of them the top ranked 3 SNPs. Each SNPs is also associated with the region of the gene where it was found and the motif that generated its score.

The convergence information is back supplied to MrBayes to produce summary statistics submitted successively to the workflow. To control the molecular evolution model fit to the data, a web service implements a posterior predictive test within the software HyPhy (Pond et al., 2005) which uses as input the samples from the posterior distribution to simulate 200 data sets and compare the original data entropy with the distribution of simulated ones. The workflow is built within Taverna Workflow Management System, each of the described steps are executed in a distributed computational environment like EGI grid infrastructure. This is possible because we have built a REST-FUL web service that exploits the usage of JST (Job Submission Tool) (DeSario et al. 2009; Tulipano et al. 2011) in order to submit and monitor the jobs over the grid. In this work we will show how the same web service

built in Java and deployed over Tomcat server could be used to submit different applications and all procedures used to ensure the correct execution of the requested runs. We will also describe workflows provided to the final users and how they could help to use the grid infrastructure.

Results and Discussion

The use of JST helps in the management of jobs submitted to all computing infrastructure, and enables the Web Services to use all resources that are needed from the users. In these workflows, indeed, the user could need different computing resources: grid EGI infrastructure, local batch facilities and dedicated servers. By means of those workflows and the use of JST, the end user could exploit all the resources in a transparent and easy way. To solve the problem of

staging input and output, we choose a WebDav server in order to keep the interaction between the users and the service as simple as possible. In fact using the WebDav protocol the user could mount directly the remote server as a local file-system on his own personal computer, allowing a very easy transfer of single files or entire directory with a simple drag&drop.

The solution described in this work will allow also the very end user to exploit the power of a computing grid infrastructure like EGI, without the complexity of learning a new interface. Indeed, the community of BioVel, as many others communities are used to have Taverna as the only interface for their research. Expressing the high level formalization of the algorithm in a workflow language, allows scientists interested in setting up algorithm's parameters but not expert in grid technology to improve and update the system, and in same time non-expert scientists to use those services. In fact, using workflows, researchers could only focus the effort on scientific activities instead of learning complex procedures to execute their applications, and once the workflow is developed all others researchers can re-use a part of it or the entire workflow to build ad-hoc application according to their needs.

Acknowledgements

We would like to thanks FP7 funding that with the grant 283359 made possible this work

References

1. Altekar G, Dwarkadas S, Huelsenbeck JP, Ronquist F (2004) Parallel metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics* 20, 407-415.
2. Battagliero S, Puglia G, Vicario S, et al. (2011) An Efficient Algorithm for Approximating Geodesic Distances in Tree Space. *Ieee-Acm Transactions on Computational Biology and Bioinformatics* 8, 1196-1207.
3. Cho I, Blaser MJ (2012) APPLICATIONS OF NEXT-GENERATION SEQUENCING The human microbiome: at the interface of health and disease. *Nature Reviews Genetics* 13, 260-270.
4. De Sario G, Tulipano A, Donvito G, Maggi G, Gisel A (2009) High-throughput Grid computing for Life Sciences. In: M. Cannataro editor. *Handbook of Research on Computational Grid Technologies for Life Sciences, Biomedicine, and Healthcare*.
5. Delzenne NM, Neyrinck AM, Backhed F, Cani PD (2011) Targeting gut microbiota in obesity: effects of prebiotics and probiotics. *Nature Reviews Endocrinology* 7, 639-646.
6. Eddy SR (2011) Accelerated Profile HMM Searches. *Plos Computational Biology* 7.
7. Felsenstein J. 2004. *Inferring Phylogenies*. Sinauer Associates, Sunderland, Mass. 580pp.
8. Honeycutt LR, Hillis DM, Bickham JW (2010) *Molecular approaches in natural resource conservation and management* eds. DeWoody JA, Bickham JW, Michler CH, et al. Cambridge University Press.
9. Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP (2001) Evolution - Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* 294, 2310-2314.
10. Hull D, Wolstencroft K, Stevens R, et al. (2006) Taverna: a tool for building and running workflows of services. *Nucleic Acids Research* 34, W729-W732.
11. Pond SLK, Frost SDW, Muse SV (2005) HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 21, 676-679.
12. Tulipano A., Marangi C., Angelini L., Donvito G., Cuscela G., Maggi G., & Gisel A. (2011). GRID distribution supports clustering validation of large mixed microarray data sets. *EMBnet.Journal*, 17 (1), pp. 18-25.

Posters

Empowering web portal users with personalized text mining services

Fedor Bakalov¹, Marie-Jean Meurs^{2,3}, Birgitta König-Ries¹, Bahar Sateli², René Witte², Greg Butler^{2,3}, Adrian Tsang^{3,4}

¹Institute for Computer Science, Friedrich Schiller University of Jena, Jena, Germany

²Department of Computer Science and Software Engineering, Concordia University, Montreal, Canada

³Centre for Structural and Functional Genomics, Concordia University, Montreal, Canada

⁴Department of Biology, Concordia University, Montreal, Canada

Motivation and Objectives

Nowadays, many organizations use portals extensively as a single-point access to information, applications, and people. However, dealing with the constantly growing amounts of information available through web portals is difficult and time-consuming for users. Most of the current portal systems enable users to retrieve content statically defined as relevant - but reading and interpreting it remain a serious bottleneck. We propose to break this bottleneck with a personalized information system that integrates Natural Language Processing (NLP) to support users in analysing, transforming, and creating knowledge from large amounts of textual content. Our approach is a novel combination of web portal technology with the Semantic Assistants (SA) framework (Witte and Gitzinger, 2008), an extensible software architecture that allows invoking literally any NLP or text mining tool using either Web Services or Application Programming Interfaces (API). The whole system is designed to give users full control over personalization, and leverage visualizations to adjust the adaptive behaviour to the users' preferences in an easy-to-use way.

Methods

The proposed system relies on three major components: a web portal compliant with the Java Portlet Specification JSR286; the Semantic Assistants framework providing NLP services; and a module dedicated to user modelling and personalization. Web Portals are web applications providing users with unified access to various information resources and services. The most widely used industry standard for portal technology is the Java Portlet Specification JSR286. This standard defines an API for developing portlet applications in the Java programming language. A portlet is a pluggable user interface component that provides a specific piece of content or an application. Portlets can be aggregated into a portal page. Semantic Assistants are an existing open

source service-oriented framework that brokers NLP pipelines as W3C standard web services. This framework brings NLP techniques directly to end users by integrating them within desktop applications. User Modelling and Personalization components allow storing information on user interests, which are represented as an overlay of domain concepts defined in the domain ontology. For each concept, the user model stores the exact degree to which the user is interested in it. The user model is updated following our hybrid approach (Bakalov et al., 2009). The portal content is delivered to users through personalizable portlets that can be viewed in standard or personalized states. In a personalized state, users can choose between several personalization effects (e.g., sort content by interest or chronologically). *Genozymes Portal*: This biochemical literature portal has been developed for the Genozymes project at Concordia's Centre for Structural and Functional Genomics (CSFG) and is currently in use by a group of biologists, biochemists and geneticists working on lignocellulose research. The goal of this research is to find novel ways of creating bioproducts and biofuels from green waste. Part of this work is the curation of content regarding specific enzymes of fungal origin from the domain literature. Towards this end, literature from the PubMed portal needs to be evaluated for relevance, which is a time-consuming task. To support these researchers, we automatically import new articles appearing on PubMed into a portal (Figure 1), processing them with the mycoMINE NLP pipeline (Meurs et al, 2012), which extracts entities and facts related to fungal enzymes. The Query portlet displays user's search queries. These queries can be hierarchically organized and modified by adding, renaming or deleting keywords. The Listing portlet presents the most relevant papers found among new articles appearing on PubMed with regards to all or a selected subset of the user queries. In our example (Figure 1), the mention of cellulose percentage

The screenshot displays the Genozymes portal interface. On the left, the 'Query' portlet shows a search tree with 'Cellulose percentage' selected. The 'Listing' portlet displays a search result for 'Cellulose percentage' with a detailed abstract and a heatmap visualization. The heatmap shows the distribution of interest across various entities, with a hot zone in the center. A 'Personalization Options & Interest Profile' window is overlaid on the heatmap, showing a circular chart with slices representing different interest levels. The 'Index' portlet on the right shows a list of entities and their counts, including 'Enzyme' (47), 'Glycoside_Hydrolyase' (15), 'Substrate' (21), 'Organism' (35), 'Gene' (1), 'EnzymeStats' (1), 'Family' (3), 'Assay' (1), and 'OrganismStats' (1).

Figure 1: Genozymes portal with IntrospectiveViews

has been selected in the Query portlet and the user has requested the mycoMINE assistant on the papers appearing in the Listing portlet. The Index portlet displays the mycoMINE results in terms of entities and facts mentioned in the papers. The number of different occurrences is indicated for each type of entity and fact (e.g. 47 different enzymes). Each reported entity or fact is linked in the texts to their corresponding mentions, which are underlined, then highlighted when the user selects them in the Index portlet. Portal content and SA results can be personalized by users. In the personalized view, users can view and edit their interest profile as well as define how the portlet content should be personalized. This is done in a personalization options window (Figure 1 - bottom) displayed as an overlay over the portlet. The personalization options vary from portlet to portlet. For example, the Listing portlet supports three personalization effects which can be selected by checking the corresponding checkboxes: (1) sorting publications according to the user interest profile; (2) highlighting the most interesting of the user publications by a

colour marker; (3) highlighting mentions of items from the user interest profile in the publications list. User changes on the personalization options are immediately projected onto the portlet content. The personalization interface we proposed (Bakalov et al., 2010) visualizes user interests using a metaphor of circular zones partitioned into slices, where each zone represents items of certain interest degree and each slice represents items of a specific type. The hot zone in the centre displays items that users are strongly interested in, while the cold zone at the circle edge displays less interesting items. The visualization follows Shneiderman's information seeking mantra (Shneiderman, 1996) by providing functions for getting an overview, zooming in and out, filtering, searching and giving detailed information about items upon request (Figure 1 - bottom, details). The visualization also allows editing information in the model (adding and deleting items, changing interest degree, etc.). Similar to the changes of personalization options, all changes in the interest profile made through the visualization are immediately projected on the personalized con-

tent. For example, upon a change in the interest profile, the publications in the Listing portlet will be re-sorted and the colour markers of the most relevant publications will be updated.

Results and Discussion

To evaluate the impact of introducing personalized text mining services in our Genozymes portal, we conducted a user study with seven CSFG researchers. The results of this evaluation showed that providing users control over personalization and text mining services makes substantial impacts on the usefulness, usability, and user satisfaction of the personalized system. The Genozymes portal is available for demonstration at <http://www.minerva-portals.de:10040/wps/portal,user=demo,pswd=portaluser>.

We demonstrate how to enhance web portals with personalized text mining services that enable users to focus on the interesting sections of the presented contents. In such a way, portal users can apply NLP tools not only on publications, but also on a variety of other resources and applications that can be aggregated using portal technology, such as patents, databases, samples, or observation and sensor data.

Acknowledgements

We thank all the participants of the user study for their contribution. We thank Justin Powlowski for his expert advice on the biology content. Funding for part of this work was provided by Genome Canada and Genome Quebec. Part of this research was sponsored by the IBM Ph.D. Fellowship Awards Program and carried out in the framework of the Minerva Portals project in cooperation with IBM Deutschland Research & Development GmbH.

References

1. Bakalov F, König-Ries B, et al. (2009) Hybrid Approach to Identifying User Interests in Web Portals. Int. Conf. on Innovative Internet Community Systems, 2009.
2. Bakalov F, König-Ries B, et al. (2010) IntrospectiveViews: An interface for scrutinizing semantic user models, Int. Conf. on User Modeling, Adaptation, and Personalization.
3. Meurs MJ, Murphy C, et al., (2012) Semantic text mining support for lignocellulose research, BMC Medical Informatics and Decision Making 12(Suppl 1):S5. doi:10.1186/1472-6947-12-S1-S5
4. Shneiderman B (1996) The eyes have it: A task by data type taxonomy for information visualizations. IEEE Symposium on Visual Languages, 1996.
5. Witte R and Gitzinger T (2008) Semantic Assistants - User-Centric Natural Language Processing Services for Desktop Clients. Asian Semantic Web Conference, LNCS5367-360.

Ordering copy number alteration data to analyze colorectal cancer progression

Iuliana M. Bocicor¹✉, Giulio Caravagna², Alex Graudenzi², Claudia Cava², Giancarlo Mauri², Marco Antoniotti²

¹Department of Computer Science, Babes-Bolyai University, Cluj-Napoca, Romania

²Department of Informatics, Systems and Communication, University of Milan Bicocca, Milan, Italy

Motivation and Objectives

Cancer is a very complex disease and understanding its dynamics and evolution is one of the challenges of modern biosciences. As most available data on cancer is static, extracting dynamic information about its progression from "static" biological data would have a major significance.

We are approaching the Temporal Ordering Reconstruction (TOR) problem, that is the sorting of a collection of multi-dimensional biological data to reflect an accurate temporal progression of the target disease.

The most general form of the TOR problem has been studied from many points of view. Firstly, the TOR problem, as defined above has been tackled mostly in two works, which use **gene expression** data as the "raw" data in the samples (Gupta and Bar-Joseph, 2008; Magwene et al., 2003). Secondly, another series of works start by analyzing comparative genomic hybridization data to build a plausible tree of possible gene mutation events and continue towards a use of Bayesian models to assess pathways variations in a disease (Desper et al., 1999; Pathare et al., 2009; Gerstung et al., 2011; Beerenwinkel et al., 2005).

Our work is more focused on a specific approach to the TOR problem, previously proposed by Gupta and Bar-Joseph (Gupta and Bar-Joseph, 2008), which has been shown to work for gene expression data and we develop a methodology which enables us to apply this technique on a **Copy Number Alterations (CNAs)** data set. We also aim to provide a building block in an analysis pipeline that can be used to look at temporal reconstruction problems that assume an already (partially) ordered dataset (Ramakrishnan, 2010; Antoniotti, 2010).

Methods

The technique presented by Gupta and Bar-Joseph (Gupta and Bar-Joseph, 2008) is based

on the reduction of the sorting problem to the **Travelling Salesman Problem (TSP)**, under two biologically realistic assumptions over the gene expression data set. As we can assume that the CNAs data also fulfils these two assumptions, we develop a methodology which enables us to apply the technique on a CNAs data set.

In order to capture distinct aspects of the complex CNAs phenomenon, we define several chromosome-related measures and certain filters targeting significant portions of chromosomes. We also aim to identify which of these measures performs best regarding tumour progression or whether chromosomal gains (amplifications) or losses (deletions), considered separately, could influence the outcome.

As chromosome measures, we introduce the following notions: value, intensity, number and the averaged analogous: average of the values and average of the intensities, all these referring to alterations, deletions and amplifications. Furthermore, we propose two filtering methods to be applied on the initial data set, which could lead us towards obtaining more accurate orderings:

- recurrent CNAs - we consider those CNAs that belong to regions of the chromosomes that have suffered alterations in a higher number of different samples;
- recurrent CNAs, as well as CNAs belonging to regions that include at least one of the genes known to be involved in tumor progression (**cancer driver genes**).

In order to build the TSP instance, we consider the cities to be represented by the 22-dimensional samples (each dimension corresponding to one chromosome, not considering the gender-linked chromosome) and a distance matrix is used to define distances between any two samples. Two types of metrics are used: the **L_1 distance** and the **Euclidean distance**.

Figure 1 briefly illustrates our methodology, highlighting the most important steps that were

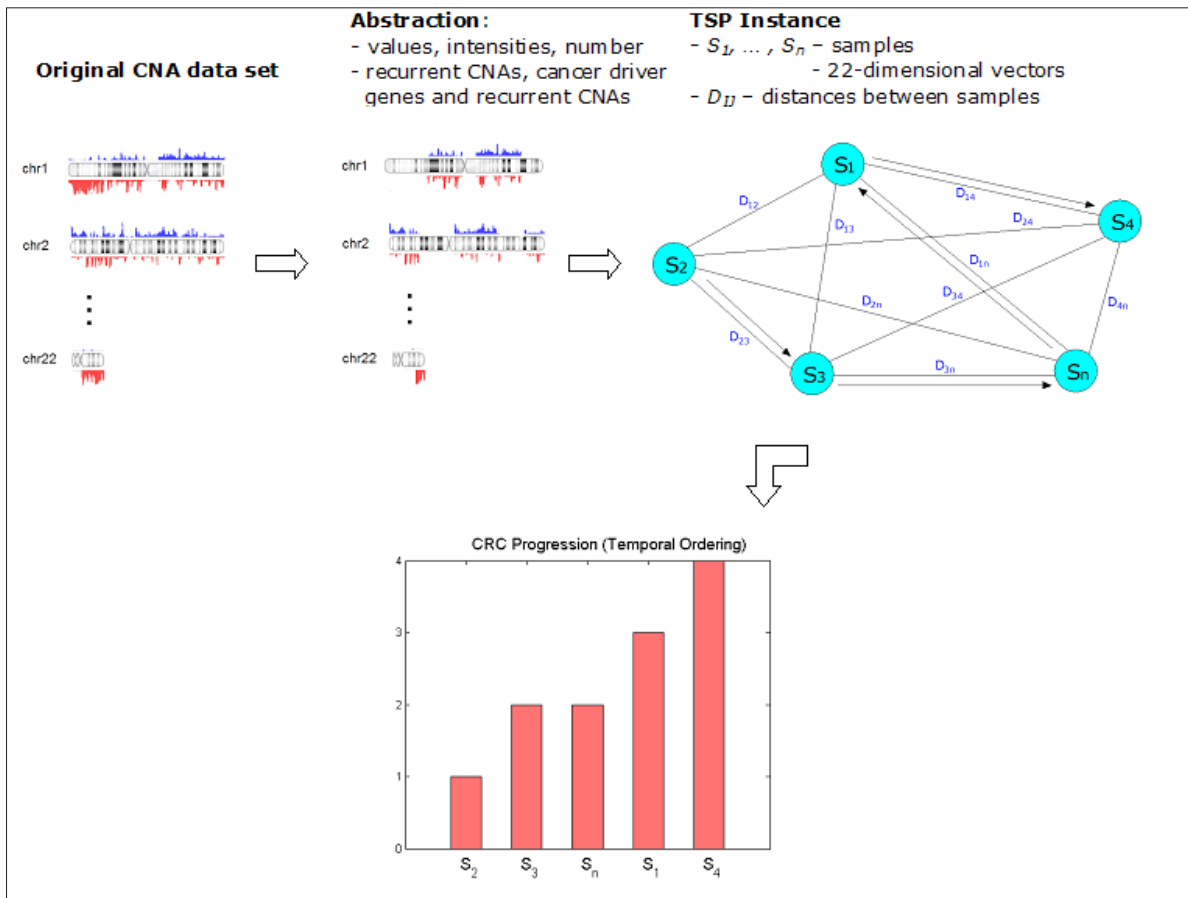


Figure 1: Representation of the proposed methodology. Starting from the input data set, different subsets are defined, using abstraction mechanisms. A TSP instance is built for each new data set and finally, the solution to the TSP represents the temporal ordering of the given samples.

used to determine a temporal ordering for a set of biological samples.

We tested the algorithm on a CNAs data set (Reid et al., 2009), consisting of 44 samples, in different stages of colorectal cancer (CRC). Three types of tests were made, one for the initial input data set and two for the subsets obtained by applying the above mentioned filters, therefore obtaining several different orderings. As a validation criterion, we used the survival time of each patient, after being diagnosed. We defined the “ideal ordering” as the one in which the first sample has the maximum, while the last one has the minimum overall survival time. Using the **Squared Deviation Distance (SDD)**, the distance from each obtained solution to the ideal one was computed. Therefore, the orderings having smaller SDDs (with regard to the ideal ordering) were considered to be more accurate.

Results and Discussion

Results show that the test in which recurrent CNAs are used in conjunction with CNAs belonging to cancer driver genes produces the highest similarities with respect to the ideal ordering, i.e., the lowest values of the SDD, in terms of minimum value. Therefore, the more filters we apply on the input data set, the closer the minimum obtained orderings are, with respect to the ideal one. This clearly outlines the importance of combining biological knowledge with mathematical techniques to achieve significant results.

The best result was obtained for the test that takes into account the CRC driver genes and recurrent CNAs, with the chromosome measure average of values of alterations and for the L_1 distance. The samples in the first half of this ordering belong to patients who have (on average) significantly higher survival times than those in the last half. Although in our data set the CRC histologi-

cal stages are not always directly correlated to the survival time, we observe that the best ordering is also compatible with the CRC stages, to a certain degree.

Concerning the other chromosome measures, we have noticed that, on average, in the case of values and intensities, amplifications and deletions, considered separately, induce a better ordering than all alterations; in the case of the number, deletions seemed to be more relevant, when considering the L_1 distance, while for the Euclidean distance, all alterations inferred orders with lower SDDs. For the averaged values and intensities, all alterations have proven to be more important than either gains or losses, when considering the L_1 distance. On average, the orderings obtained using the L_1 distance are more accurate, compared to those using the Euclidean distance.

We have presented a particular solution for the temporal ordering reconstruction problem. We have built our approach on a previously proposed solution (Gupta and Bar-Joseph, 2008), by adapting it to chromosomal CNA data and we tested it on a CRC data set. To the best of our knowledge, our work is the first to adapt the TSP approach to the TOR problem, in conjunction with CNA data.

Acknowledgements

The authors would like to thank Manuela Gariboldi for providing the colorectal cancer data set. We also acknowledge Regione Lombardia (project RetroNet, grant 12-4-5148000-40; U.A 053) and NEDD for financial support of this work. The work was also possible with the financial support of the Sectoral Operational Programme for Human

Resources Development 2007-2013, co-financed by the European Social Fund, under the project number POSDRU/107/1.5/S/76841 with the title "Modern Doctoral Studies: Internationalization and Interdisciplinarity".

References

1. Antoniotti M, Carreras M, Farinaccio A, Mauri G, Merico D et al. (2010) An Application of Kernel Methods to Gene Cluster Temporal Meta-Analysis. *Comput Oper Res* 37(8): 1361-1368. doi: [10.1016/j.cor.2009.03.011](https://doi.org/10.1016/j.cor.2009.03.011).
2. Beerenwinkel N, Rahnenfuhrer J, Daumer M, Hoffman D, Kaiser R et al. (2005) Learning multiple evolutionary pathways from cross-sectional data. *J Comput Biol* 12: 584-598. doi: [10.1089/cmb.2005.12.584](https://doi.org/10.1089/cmb.2005.12.584).
3. Desper L, Jiang F, Kallioniemi OP, Moch H, Papadimitriou CH et al. (1999) Inferring tree models for oncogenesis from comparative genome hybridization data. *J Comput Biol* 6(1): 37-51.
4. Gerstung M, Eriksson N, Lin J, Volgestein B, Beerenwinkel N. (2011) The temporal order of genetic and pathway alterations in tumorigenesis. *PLoS ONE* 6(11): 1-9. doi: [10.1371/journal.pone.0027136](https://doi.org/10.1371/journal.pone.0027136).
5. Gupta A and Bar-Joseph Z. (2008) Extracting dynamics from static cancer expression data. *IEEE/ACM Trans Comput Biol Bioinform* 5:172-182. doi: [10.1109/TCBB.2007.70233](https://doi.org/10.1109/TCBB.2007.70233).
6. Magwene PM, Lizardi P, Kim J. (2003) Reconstructing the temporal ordering of biological samples using microarray data. *Bioinformatics* 19(7):842-850. doi: [10.1093/bioinformatics/btg081](https://doi.org/10.1093/bioinformatics/btg081).
7. Pathare S, Schaffer AA, Beerenwinkel N, Mahimkar M. (2009) Construction of oncogenetic tree models reveals multiple pathways of oral cancer progression. *Int J Cancer* 124(12): 2864-2871. doi: [10.1002/ijc.24267](https://doi.org/10.1002/ijc.24267).
8. Ramakrishnan N, Tadepalli S, Watson LT, Helm RF, Antoniotti M et al. (2010) Reverse engineering dynamic temporal models of biological processes and their relationships. *PNAS* 107: 12511-12516. doi: [10.1073/pnas.1006283107](https://doi.org/10.1073/pnas.1006283107).
9. Reid JF, Gariboldi M, et al. (2009) Integrative approach for prioritizing cancer genes in sporadic colon cancer. *Genes, Chromosomes and Cancer* 48: 953-962. doi: [10.1002/gcc.20697](https://doi.org/10.1002/gcc.20697).

Glycans, the forgotten biomolecular actors of the big picture

Matthew P. Campbell¹✉, Julien A. Mariethoz², Catherine M. Hayes³, Pauline G. Rudd⁴, Niclas G. Karlsson³, Nicki H. Packer¹, Frédérique Lisacek²

¹Biomolecular Frontiers Research Centre, Macquarie University, Sydney, Australia

²Proteome Informatics Group, Swiss Institute of Bioinformatics, Geneva, Switzerland

³Department of Biomedicine, Gothenburg University, Gothenburg, Sweden

⁴NIBRT, Dublin, Ireland

Motivation and Objectives

Glycans or carbohydrates, both in the form of polysaccharides or glycoconjugates are increasingly recognised as being implicated in human health. Glycosylation is probably the most important post-translational modification in terms of the number of proteins modified and the diversity generated. Since glycoproteins, glycolipids and glycan-binding proteins are frequently located on the cell's primary interface with the external environment many biologically significant events can be attributed to glycan recognition. In other words, glycans mediate many protein-protein interactions. In spite of such a central role in biological processes, the study of glycans remains isolated, protein-carbohydrate interactions are rarely reported in bioinformatics databases and glycomics is lagging behind other -omics.

Recent progress in method development for characterising the branching structures of complex carbohydrates has now enabled high throughput technology. Automation then calls for software development. Adding meaning to large data collections requires bioinformatics means. Current glycobioinformatics resources do cover information on the structure and function of glycans, their association with proteins or their enzymatic generation. However, this information is partial, scattered and often inaccessible to non-glycobiologists.

In partnership with expert international research groups we are involved with the development of the UniCarb KnowledgeBase (UniCarbKB), an effort to develop and provide an informatic framework for the storage and the analysis of high-quality data collections on glycoconjugates, including informative metadata and annotated experimental datasets (Campbell et al., 2011). UniCarbKB is an initiative designed to support research in systems biology by complementing proteomics with glycomics

Methods

To achieve our goals, UniCarbKB is partnering with BCSDDB (Bacterial Carbohydrate Structure Database), GlycomeDB, GLYCOSCIENCES.de JCGGDB (Japan Consortium for Glycobiology and Glycotechnology Database), MonosaccharideDB to develop a standard Resource Description Framework RDF representation for carbohydrate structure, biological and bibliographic annotations and experimental evidence. Access to data stored in this format will allow users to perform queries that were not previously possible, and provide the ideal platform for connecting these disparate resources.

While we are still in the early development phases, we have designed a scalable web-friendly framework that integrates information from GlycoSuiteDB and EUROCarbDB. UniCarbKB is a representation of the tremendous growth in information available in glycomics and the adoption of leading-edge technologies to disseminate and query this knowledgebase.

UniCarbKB is based on the reengineering of GlycoSuiteDB and EUROCarbDB and built on the foundations of lightweight Java Rails architecture implementing new search features to explore the wealth of new data now available. The new version will be on-line late 2012. The framework adopts agreed standards to store structural and metadata content including the translation of GlycoSuiteDB structure entries into the GlycoCT format offering a comprehensive structure database (Herget et al., 2008). Significant improvements to the data schema have enabled the merger of these two databases in particular the rational adoption of taxonomic, tissue and disease ontologies. The schema is module in design to segregate the three components (i) structure (ii) informative metadata and (iii) supporting analytical data.

Results and Discussion

New information relevant to glycoproteins, notably the inclusion of glycosylated structures local-

ised in different tissues sourced from a literature exploration study was incorporated. This led to build an accessible database of qualitative and quantitative protein glycoprofilng data. In parallel, special effort is invested into linking this information with sugar recognition curated data (e.g., SugarBind and CFG Glycan Array) to allow deeper mining of the functional role of glycans. At this stage, our first focus is on infectious diseases.

The overall aim of the project is to access, query and mine the most comprehensive bio-curved overview of existing glycoinformation associated with proteins in a site-specific manner both from the attachment and the recognition perspective.

Acknowledgements

The initiative is supported by NECTAR (Australian National eResearch Collaboration Tools and

Resources), STINT (Swedish Foundation for International Cooperation in Research and Higher Education) and SNSF (Swiss National Science Foundation).

References

1. Campbell MP et al. UniCarbKB: putting the pieces together for glycomics research. *Proteomics* (2011) 11(21): 4117-21.
2. UniCarbKB: <http://unicarbkb.org/>
3. GlycosuiteDB: <http://glycosuitedb.expasy.org/glycosuite/glycodb>
4. EUROCarbDB: <http://www.eurocarbdb.org/>
5. SugarBind: <http://sugarbind.expasy.org/sugarbind/>
6. CFG Glycan Array: <http://www.functionalglycomics.org/glycomics/publicdata/primaryscreen.jsp>
7. Herget S, Ranzinger R, Maass K, Lieth CW. GlycoCT-a unifying sequence format for carbohydrates. *Carbohydr Res.* (2008) 343(12):2162-71.

Genomic and proteomic data integration for comprehensive biodata search

Arif Canakoglu[✉], Marco Masseroli

Dipartimento di Elettronica ed informazione, Politecnico di Milano, Milan, Italy

Motivation and Objectives

With high-throughput technologies in the life sciences, particularly in molecular biology, the amount of data available has grown exponentially. Yet, such data are stored in several different formats and spread into numerous databanks (Galperin and Fernández-Suárez, 2012). This scenario makes even more difficult to find and retrieve the data required to answer the scientists' questions, which usually are complex and regard multiple biological entities and several of their aspects. Consequently, in the last few decades biological data integration has become a major focus in bioinformatics. Data integration is essential to comprehensively evaluate and search information from different databanks. For example, no single data source exists that supplies association data between protein interactions and genetic disorders.

There are several approaches, with related implementations, to integrate heterogeneous data from different sources, such as information linkage (e.g. SRS (Etzold et al., 1996), NCBI Entrez (Tatusova et al., 1999)), federated databases (e.g. BioKleisli (Davidson et al., 1997), DiscoveryLink (Haas et al. 2001)), multi-databases (e.g. TAMBIS (Stevens et al., 2000), BACIIS (Miled et al., 2002)), mediator based solutions (e.g. BioDataServer (Freier et al., 2002), Biomediator (Cadag et al., 2007)) and data warehousing (e.g. EnsMart (Kasprzyk et al., 2004), BioWarehouse (Lee et al., 2006)). Data warehousing is the most convenient one when the data are very numerous and offline processing is a necessity to mine integrated data efficiently and comprehensively. Using such an approach, we created an integrative data warehouse, where integration is performed based on a predefined modular data model that provides a unified reconciled global view of the integrated data. Data warehouse creation and updating is performed by supervised automatic procedures, which also control variation of the integrated data in the original data sources (Davidson et al., 1995). The used modular data model supports both easy data warehouse extensibility, with the integration of new data sources,

and effective automatic querying on the integrated data for their search and extraction.

Methods

We built a Genomic and Proteomic Knowledge Base (GPKB), which is a relational, integrative and multi-organism data warehouse containing heterogeneous genomic and proteomic annotation data. We import them from several well known public databases, including Entrez Gene, UniProt, IntAct, MINT, BioCyc, KEGG, Reactome, GO, GOA and OMIM. The very numerous data integrated, which regard biomolecular entities (mainly genes and proteins) and their biomedical features and associations, are all checked for data correctness and consistency (Ghisalberti et al., 2010). By leveraging imported similarity and historical evaluation data available, we identify different IDs from different data sources as representing the same entity. This enables us to classify and extract different attributes available also from different data sources as referring to the same entity, feature or association, rather than as distinct attributes of different entities or of their features or associations.

For the GPKB, we designed a modular global data schema with abstraction and generalization of the main data features. It is characterized by a multi-level data architecture, which includes source-import level, instance-aggregation level and concept-integration level.

Leveraging on such data schema, we defined query templates to extract the integrated data. These query templates allow extracting the user required data from any version of the GPKB automatically. This supports different Web applications and services connected to the GPKB in automatically searching and extracting data from the data warehouse for different goals, including gene and protein annotation inference, annotation enrichment analysis and user query support for biomedical knowledge discovery.

The performed inference of gene and protein annotations is based on the "transitive closure" concept. It is inspired by Swanson work (Swanson, 1986) that is based on the transitive closure of het-

erogeneous extensive annotation data. The inference procedure is controlled by Standard Query Language (SQL) templates, which are applied to any relational biomedical molecular database.

Results and Discussion

With the data downloaded on May 28th, 2012, among others, the GPKB contained 9,537,645 genes of 9,631 organisms, 38,960,202 proteins of 338,004 species, 19,522 protein domains and 824,797 protein domains annotations, 28,889 biochemical pathways and 171,372 pathway annotations (77,812 gene and 93,560 protein annotations), 35,252 Gene Ontology terms and 64,185,070 Gene Ontology annotations (1,272,168 gene and 62,912,902 protein annotations), 10,212 human genetic disorders and their 27,705 gene annotations. Furthermore our GPKB integrates also other types of data regarding DNA sequences, transcripts, enzymes, small molecules of biological interest, and clinical synopses. In total it contains more than 103,006,922 gene annotations and 183,209,462 proteins annotations.

The great amount of biomolecular features and their association data that the GPKB contains makes it a unique valuable resource which can be used for different applications, in silico experiments and knowledge discoveries.

The created automatic query templates make possible to easily search and extract each of the integrated data, offering an efficient base for various data mining algorithms and applications. As an example, by leveraging the multi-source integrated data, we inferred new gene annotations through transitive closure on various association data regarding the features of the gene encoded proteins. The same approach enabled us also to infer possible associations between protein-protein interactions and genetic disorders. Towards this aim, protein-protein interaction data files downloaded from MINT (Licata et al., 2012) and IntAct (Kerrien et al., 2012) databases were automatically parsed. Data of 46,154 human protein-protein interactions (out of the contained 254,048 protein-protein interactions of 397 different organisms' proteins), regarding 12,178 distinct human proteins, were imported in the data warehouse. These human proteins, which represent 3.7% of all the 326,766 human proteins in the data warehouse, are encoded by 11,232 different human genes. By applying the transitive closure concept on the interacting

protein encoding gene and genetic disorder related gene association data, we identified 1,130 gene-gene interactions and found 1,136 human protein-protein interactions possibly associated with 628 genetic disorders (such as Alzheimer, Cystic fibrosis, Diabetes mellitus, Parkinson, etc.). Such genetic disorders resulted related to 86 clinical synopses and 3,481 phenotypes.

The created Genomic and Proteomic Knowledge Base, that is updated quarterly, can be freely accessible through an easy-to-use Web interface available at <http://www.bioinformatics.dei.polimi.it/GPKB/> where all integrated data in the GPKB can be comprehensively searched.

Acknowledgements

This research is part of the "Search Computing" project (2008-2013), funded by the European Research Council (ERC), under the 2008 call for "IDEAS Advanced Grants".

References

1. Cadag E, Louie B, et al. (2007), Biomediator data integration and inference for functional annotation of anonymous sequences. *Pac Symp Biocomput.* 343-354.
2. Davidson SB, Overton C, et al. (1995), Challenges in integrating biological data sources. *J. Comput. Biol.* 2(4):557-572.
3. Davidson SB, Overton C, et al. (1997), BioKleisli: a digital library for biomedical researchers. *Int. J. Digit. Libr.* 1997, 1(1):36-53. doi:10.1007/s007990050003
4. Etzold T, Ulyanov A, et al. (1996): SRS: Information Retrieval System for molecular biology data banks. *Meth. Enzymol.* 266:114-128.
5. Freier A, Hofestäd R, et al. (2002), BioDataServer: a SQL-based service for the online integration of life science data. *In Silico Biol.* 2(2):37-57.
6. Galperin MY, Fernández-Suárez XM (2012), The 2012 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. *Nucleic Acids Res.* 40(Database issue):D1-D8. doi:10.1093/nar/gkr1196
7. Ghisalberti G, Masseroli M, Tettamanti L (2010), Quality controls in integrative approaches to detect errors and inconsistencies in biological databases. *J Integr Bioinform* 7(3):199,1-13. doi: 10.2390/biecoll-jib-2010-119
8. Haas LM, Rice JE, et al. (2001), DiscoveryLink: a system for integrated access to Life Sciences data sources. *IBM Systems Journal* 40(2):489-511.
9. Kasprzyk A, Keefe D, et al. (2004), EnsMart: a generic system for fast and flexible access to biological data. *Genome Res.* 14(1):160-169. doi:10.1101/gr.1645104
10. Kerrien S, Aranda B, et al. (2012), The IntAct molecular interaction database in 2012. *Nucleic Acids Res.* 40(Database issue):D841-846. doi:10.1093/nar/gkr1088
11. Lee TJ, Pouliot Y, et al. (2006), BioWarehouse: a bioinformatics database warehouse toolkit. *BMC Bioinformatics*, 7:170,1-14. doi:10.1186/1471-2105-7-170

12. Licata L, Briganti L, et al., (2012), MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res.* 40(Database issue):D857-D861. doi:10.1093/nar/gkr930
13. Miled ZB, Li N, et al. (2002), Complex life science multidatabase queries. *Proc. IEEE* 90(11):1754-1763. doi:10.1109/JPROC.2002.804683.
14. Stevens R, Baker P, et al. (2000), TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources. *Bioinformatics*, 16(2):184-185.
15. Swanson DR (1986), Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect. Biol. Med.* 30(1):7-18
16. Tatusova TA, Karsch-Mizrachi I, et al. (1999), Complete genomes in WWW Entrez: data representation and analysis. *Bioinformatics* 15(78):536-543.

Generation of explicit rules predicting neuroblastoma patients' outcome

Davide Cangelosi¹, Fabiola Blengio, Rogier Versteeg², Angelika Eggert³, Alberto Garaventa⁴, Claudio Gambini⁵, Massimo Conte⁴, Alessandra Eva¹, Luigi Varesio¹

¹Laboratory of Molecular Biology, Gaslini Institute, Genoa, Italy

²Department of Human Genetics, Academic Medical Center, University of Amsterdam, Amsterdam, Netherlands

³Department of Pediatric Oncology and Hematology, University Children's Hospital Essen, Essen

⁴Department of Hematology-Oncology, Gaslini Institute, Genoa, Italy

⁵Departments of Pediatric Pathology, Gaslini Institute, Genoa, Italy

Motivation and Objectives

Neuroblastoma (NB) is the most common pediatric solid tumor characterized by clinical and molecular risk factors. The mortality is about fifty percent and this makes exploration of new and more effective risk factors for improving stratification mandatory. Hypoxia is a condition of low oxygen tension occurring in poorly vascularized areas of the tumor associated with poor prognosis. We had previously defined a robust gene expression signature measuring the hypoxic component of NB tumors (NB-hypo) that is a novel, independent risk factor (Fardin et al., 2010). Integrating classical risk factors with NB-hypo could improve the stratification of NB patients. We wanted to develop a prognostic classifier of NB patients' outcome blending existing knowledge on clinical and molecular risk factors with the prognostic NB-hypo signature. Furthermore, we were interested in the decision tree classifier that outputs explicit rules easily translated into the clinical setting.

Methods

A total of 182 NB patients were enrolled on the bases of availability of gene expression profile by Affymetrix GeneChip HG-U133plus2.0, clinical and molecular information. NB tumor stage was defined according to the International NB Staging System (INSS). Age at diagnosis was dichotomized as greater or equal than 1 year and less than one year. MYCN status was amplified or normal. Good and poor outcome were defined as the patient's status alive or dead 5 years after diagnosis respectively. The risk group was assigned according to the International Neuroblastoma Risk Group (INRG) Consensus Pretreatment Classification Schema. The 182 NB patients cohort was clustered in High and low hypoxia by k-means analysis of the 62 probsets constituting the NB-hypo signature previously described to measure tumor hypoxia (Fardin et

al., 2009). We utilized the k-means algorithm implemented in the WEKA software (Hall et al., 2009) setting up number of clusters to 2, 500 iterations, preserving instances order and using Manhattan distance.

The classification was performed by induction of decision trees. We utilized the Weka J48 implementation of the popular C4.5 algorithm (Kotsiantis, 2007; Murthy, 1998) and we set up the following options: pruning parameter was 0.25, pruning method was sub-tree raising and minimum number of instances per leaf was 2. Each leaf of the decision tree classifier identifies a non overlapping group of patients and each decision node identifies a branch which splits the dataset. We utilized Fisher's exact test to measure the statistical significance of groups and branches. Fisher's test was utilized in the context of decision trees to design a top-down approach to prune out non statistically significant branches (Liu et al., 2010). For each leaf we counted the number of correctly classified (named n) and the number of incorrectly classified instances (named m). We considered the marginal totals y and $\neg y$ which represent poor and good outcome patients respectively. We designed a 2x2 contingency table of the two possible outcomes (Good or Poor) against the number of instances included in a give leaf and the remaining instances. Application of the Fisher exact test to this table generates a p value giving the probability of observing 2x2 table, or more extreme tables, knowing the marginal totals (y , $\neg y$) and assuming independency among the patients in a specific leaf and those in other leaves.

Results and Discussion

Patients were stratified in good and poor outcome on the bases of the following risk factors: Age at diagnosis, INSS stage, MYCN status and NB-hypo. The algorithm generated a decision tree classifier composed by 3 decision nodes

and 7 leaves covering 87% of non overlapping good outcome patients and 100% of non overlapping poor outcome patients. Each path from the root to a leaf utilizes some, but not all, considered risk factors. Interestingly NB-hypo was included in the decision node that stratified stage 3 tumors demonstrating its usefulness in NB patients' stratification.

The leaves classifying good outcome patients had the very low error of 2% indicating a good performance of the algorithm in predicting this class. In contrast, the classification of poor outcome patients produced the leaf with the highest error of 13%. This leaf includes stage 4 tumors that are traditionally difficult to stratify by any known risk factor.

To test statistical significance of the splits performed by the algorithm and the groups of patients, we utilized the Fisher's exact test with a confidence set at 95%. The results showed that the groups obtained at each split were statistically significant. Furthermore, analysis of single groups of patients identified by leaves demonstrated that some, but not all reached the significance threshold. The significant leaves included the NB-hypo among the utilized risk factors. These results further strengthen the value of NB-hypo in predicting patients' outcome. Lack of significance was often associated with a rather low number of patients in the leaf. This is a limitation of the divide-and-conquer approach of the algorithm, applied to relatively small patients' cohorts, that recurrently splits the dataset.

We then assessed the concordance of the predictions with INRG risk assessment. High Risk patients were correctly included in the leaves classifying poor outcome patients and Low Risk patients mapped correctly in the good outcome leaves. Interestingly, NB-hypo generated a leaf identifying a new group of poor outcome patients, sharing the high NB-hypo, whose characteristic fell into both High and Intermediate Risk.

We collected and analyzed the results of 1000 10-fold cross validations and we observed that most classifiers had 7 leaves and only 5 out of 10^4 deviated from this pattern. The recurrence of 7 leaves demonstrated the high stability

of the decision tree classifier that we generated. Analysis of the pruning parameters revealed optimal performance in the range of 0.1-0.3 in line with what used in this study.

The path to reach each leaf can be easily transposed into a "if...than..." rule that, in turn provides an easy readout of the classifier, precious for translating the classification into the clinical setting. In conclusion, we demonstrated that the decision tree algorithm C4.5 can derive explicit rules for NB patients stratification if classical risk factors are blended with the NB hypo signature. These rules are statistically significant and quite stable and suitable to be conveyed to the clinic to design new therapies perhaps taking hypoxia into consideration as a potential target.

Acknowledgements

The work was supported by the Fondazione Italiana per la Lotta al Neuroblastoma, the Associazione Italiana per la Ricerca sul Cancro, the Società Italiana Glicogenosi, the Fondazione Umberto Veronesi and the Ministero della Salute Italiano. Davide Cangelosi and Fabiola Blengio are recipients of a fellowship from the Fondazione Italiana per la Lotta al Neuroblastoma.

References

1. Fardin P, Barla A, Mosci S, Rosasco L, Verri A, Varesio L. (2009) The l1-l2 regularization framework unmasks the hypoxia signature hidden in the transcriptome of a set of heterogeneous neuroblastoma cell lines. *BMC Genomics*, 10:474. doi:10.1186/1471-2164-10-474
2. Fardin P, Barla A, Mosci S, Rosasco L, Verri A, et al. (2010) A biology-driven approach identifies the hypoxia gene signature as a predictor of the outcome of neuroblastoma patients, *Journal of Molecular Cancer* 9:1. 185. doi:10.1186/1476-4598-9-185
3. Hall M, Elibe F, Holmes G, Pfahringer B, Reutemann P, Witten IH. (2009) The WEKA Data Mining Software: An Update. *SIGKDD Explorations*. doi: 10.1145/1656274.1656278
4. Kotsiantis S.B. (2007) Supervised Machine Learning: A Review of Classification Techniques.
5. *Informatica* 31:249-268.
6. Liu W, Chawla S, Cieslak D, Chawla N. (2010) A Robust Decision Tree Algorithm for Imbalanced Data Sets. In: *SDM*. 766-777.
7. Murthy S.K. (1998) Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey, *Data Mining Knowledge Discovery* 2:4. 345-389. doi: 10.1023/A:1009744630224

Simulation of caspases apoptotic signalling pathway in a tuple space-based bioinformatics infrastructure

Maura Cárdenas-García¹✉, Pedro P González-Pérez², Sara Montagna³

¹Facultad de Medicina, Benemérita Universidad Autónoma de Puebla, Puebla, Mexico

²Departamento de Matemáticas Aplicadas y Sistemas, Universidad Autónoma Metropolitana, México, D.F, Mexico

³Dipartimento di Informatica: Scienza e Ingegneria, Univesita' degli Studi di Bologna, Cesena, Italy

Motivation and Objectives

Understanding intracellular communication processes is essential, since they allow the cell to perform the totality of its functions. Among them each cell has a self-destruction system that starts and operates in a regulated manner. It is called apoptosis, and includes the decision to start self-destruction as well as the proper execution of the apoptotic program. Caspases, a family of cysteine proteases, are the central regulators of apoptosis. As such, it requires the coordinated activation and execution of multiple sub programmes. Historically, different modelling approaches have been developed to deal with intracellular signalling pathways, from mathematical models – mainly Ordinary Differential Equations (ODEs) – to computational models – process algebra such as stochastic π -calculus (Priami, 1995) and κ -calculus (Danos et al., 2007). Accordingly, different simulation tools have been developed, from mathematical ones – see a survey in (Alves et al., 2006) – to computational ones such as SPiM (Phillips, 2007). While they typically address scenarios with a single compartment, in recent years a trend has emerged which moves from the single global approach to mechanisms and constructs tackling the multi-compartment scenario. In this paper, we adopt a simulation approach based on the notion of Biochemical Tuple Spaces for Self-Organising Coordination (BTS-SOC), introduced in (Viroli and Casadei, 2009), and then show how it can be applied to the simulation of the caspases signalling pathway (MacFarlane and Williams, 2004), which plays a crucial role in the transduction and execution of the apoptotic signal induced by various stimuli.

Methods

A Biochemical Tuple Space for Self-Organising Coordination (BTS-SOC) is a tuple space working as a compartment where biochemical reactions take place, chemical reactants are represented as tuples, and biochemical laws are represent-

ed as coordination laws by the coordination abstraction. Technically, biochemical tuple spaces are built as ReSpecT (Reaction Specification Tuples) tuple centres (Omicini and Denti, 2001), running upon a TuCSoN (Tuple Centres over the Network) coordination infrastructure (Omicini and Zambonelli, 1999). Tuples are logic-based tuples, while biochemical laws are implemented as ReSpecT specification tuples. In particular, each biochemical tuple space is built around a ReSpecT chemical engine, whose core is an action selection mechanism based on Gillespie algorithm (Gillespie, 1977) – an algorithm typically used to simulate systems of chemical/biochemical reactions efficiently and accurately – to execute chemical reactions with the proper rate. The main components of our BTS-SOC model for simulating intracellular signalling pathways are the following: 1) tuple centres – representing extracellular milieu and intracellular compartments, i.e., extracellular space, membrane, cytosol, nucleus and mitochondria; 2) chemical reaction sets – modelling signalling components, i.e., proteins (membrane receptors, enzymes, regulators, adapters, etc.) and genes; and 3) elements recorded as tuples in a tuple centre – representing signalling molecules, e.g., ATP, inorganic phosphate, second messengers, etc. The work reported here represents the initial approach to the simulation of the caspases apoptotic signalling pathway. The work was performed as follows:

1. Review of the literature involving the caspases pathway and experimental kinetic data of them in humans (Roschitzki-Voser et al., 2012; Chowdhury et al., 2008).

2. Modelling the signalling components-e.g., chemical reactions-belonging to the caspases apoptotic signalling pathway. We start with a minimalist model where each signalling component is described by the following attributes: 1) identity; 2) concentration in each cellular compartment; 3) free concentration; 4) "bound"

concentration; 5) cellular compartment to which it belongs; 6) chemical reactions involving the component and the order in which they occur according to the affinity of the components; and 7) reaction temporality situation.

3. Simulation of the caspases apoptotic signalling pathway in the BTS-SOC-based bioinformatics infrastructure.

3.1. Creating cellular compartments. A tuple centre (BTS) is required for each cellular compartment involved in the signalling pathway to be simulated. In our study, four tuple centres (membrane, cytosol, mitochondria and nucleus) are required to model four intracellular compartments.

3.2. Introducing reactants. In order to set up the simulation system, reactants should be introduced in the BTS. First of all, each reactant belongs to a specific cellular compartment—so, it has to be put in the appropriate BTS. Initially, only the pre-existing reactants – i.e., those reactants al-

ready in the compartments before the signalling pathway is activated – have to be put in the BTS.

3.3. Setting chemical reactions. The last step in setting up the simulation is the introduction of the reactions modelling the behaviour of signalling pathway. In our model, based on the Gillespie algorithm, every chemical reaction has a rate that expresses (along with the concentration of the input elements) the probability of the transformation.

4. Getting simulation outcomes. After entering all required information and setting the initial parameters, the system is now ready to run the caspases apoptotic pathway simulation.

5. Analysis and parameter adjustment.

Results and Discussion

Our modelling and simulation methodology initially considers the caspases signalling pathway as shown in Figure 1A. The pathway begins with the death signals (hormones, growth factors, cytokines, stress, etc); these signals trigger two

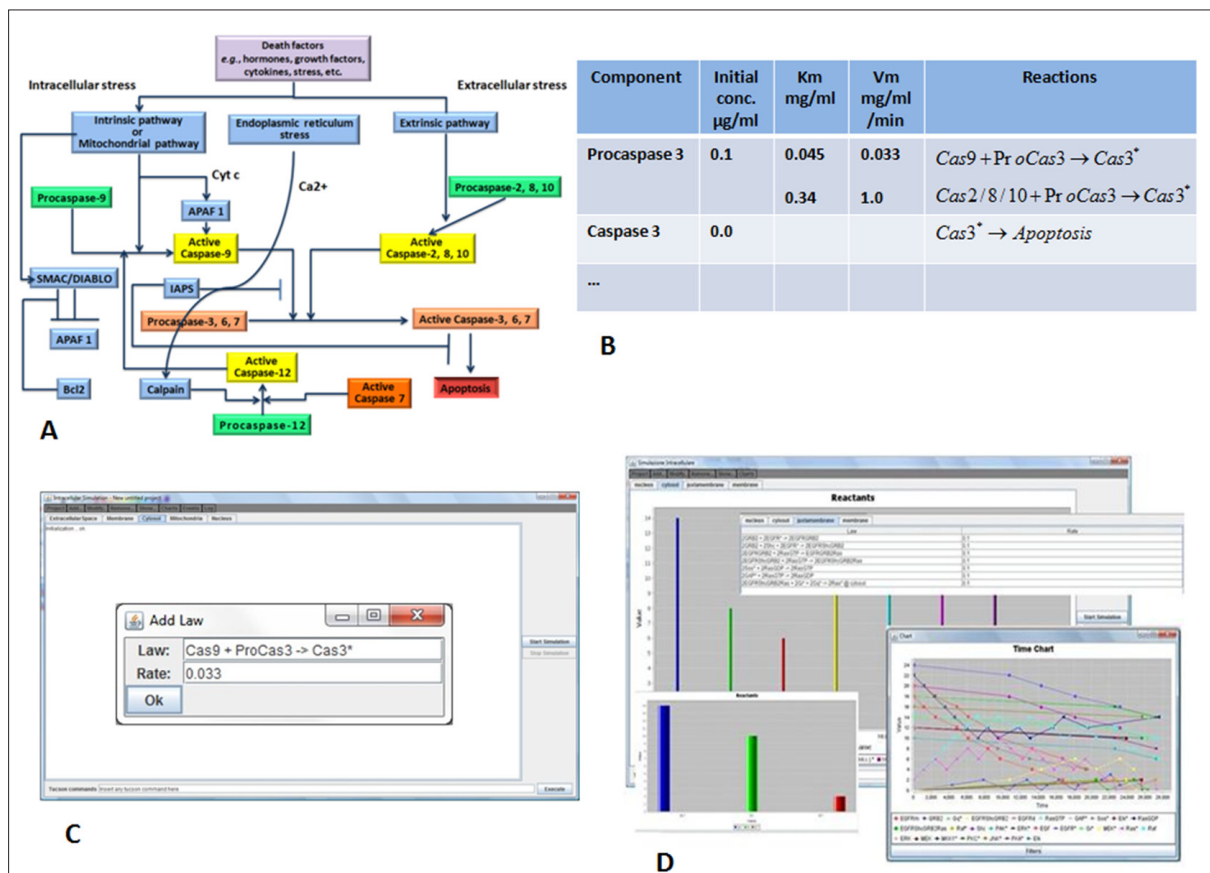


Figure 1. Basic steps of our modelling and simulation methodology. A. the effectors caspases 3, 6, 7 are activated as a consequence of the activation of the extrinsic or intrinsic pathway. B. Modelling of a signalling segment, from procaspase 3 to caspase 3*. C. Setting chemical reactions. D. Run the simulation and visualization of results.

types of response through extrinsic and intrinsic pathways. The modelling and simulation of these events are represented in Figures 1B and 1C, respectively. Take just one example for the simulation, as shown in Figure 1A. The effectors caspases 3, 6, 7 are activated as a consequence of the activation of extrinsic or intrinsic pathway. Caspase-3 is critical for apoptosis and it is activated in the cytoplasm, however, two hours after being activated it can be located at the plasma membrane in the cytoplasm and nucleus. Figure 1D shows the simulation results of these events in the BTS-SOC-based bioinformatics infrastructure. When running the simulation, we observe on the molecular level how cancer cells evade caspases signalling pathways, which allow us on the one hand to design an experiment and on the other to determine an invisible protein-protein interaction, even though it is evident in the model. According to the results obtained so far, our incremental model allows us to simulate the caspases path in a manner consistent with that reported in the literature. The next step is the refinement and adjustment of the kinetic data considering the experimental results obtained by our working group.

Acknowledgements

The authors would like to thank Andrea Boccacci for making a valuable contribution to this project.

References

1. Alves R, Antunes F, Salvador A (2006) Tools for kinetic modeling of biochemical networks. *Nat Biotechnol* 24(6), 667–672. doi:10.1038/nbt0606-667.
2. Chowdhury I, Tharakan B, Bhat GK (2008) Caspases — An update. *Comp Biochem Physiol B Biochem Mol Biol* 151(1), 10–27. doi:10.1016/j.cbpb.2008.05.010.
3. Danos V, Feret J, Fontana W, Harmer R, Krivine J (2007) Rule-based modelling of cellular signaling. *CONCUR* 2007, 17–41.
4. Gillespie, DT (1977) Exact stochastic simulation of coupled chemical reactions. *J Phys Chem* 81(25), 2340–2361. doi:10.1021/j100540a008
5. MacFarlane M, Williams AC (2004) Apoptosis and disease: a life or death decision. *Conference and Workshop on Apoptosis and Disease*. *EMBO reports* 5, 674–678.
6. Omicini A, Denti E (2001) From tuple spaces to tuple centres. *Science of Computer Programming* 41(3), 277–294. doi: 10.1016/S0167-6423(01)00011-9.
7. Omicini A, Zambonelli F (1999) Coordination for Internet application development. *Auton Agent Multi Agent Syst* 2(3), 251–269. Special Issue: Coordination Mechanisms for Web Agents.
8. Phillips A (2007) The Stochastic Pi Machine (SPiM). <http://research.microsoft.com/~aphillip/spim/> (accessed 5 November 2012).
9. Priami C (1995) Stochastic pi-calculus. *The Computer Journal* 38(7), 578–589.
10. Roschitzki-Voser H, Schroeder T, Lenherr ED, Frölich F, Schweizer A et al. (2012) Human caspases in vitro: Expression, purification and kinetic characterization. *Protein Expr Purif* 84, 236–246. doi:10.1016/j.pep.2012.05.009.
11. Virolli M, Casadei M (2009) Biochemical tuple spaces for self-organising coordination. In: Field J, Vasconcelos VT (Eds.) *Coordination Languages and Models*, ser. LNCS, Lisbon, Portugal: Springer, Jun. 2009, vol. 5521, 143–162.

Chromosome instability for tumor progression inference

Claudia Cava¹✉*, Italo Zoppis¹*, Manuela Gariboldi^{2,3}, James F. Reid^{2,3}, Marco Antoniotti¹, Giancarlo Mauri¹

¹Department of Informatics, Systems and Communication, University of Milan Bicocca, Milan, Italy

²Department of Experimental Oncology, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy

³Molecular Genetics of Cancer, FIRC Institute of Molecular Oncology Foundation, Milan, Italy

* These authors have contributed equally to this work.

Motivation and Objectives

The development and progression of Colorectal Cancer (CRC) - as for most other solid cancers, is a multi-step process leading to the accumulation of chromosomal instability (CIN) that occurs over the lifetime of a tumor (Shen et al, 2007; Vogelstein et al, 1988; Fearon et al, 1990). CINs include DNA copy number alterations (CNAs), i.e., regions of aberrantly increased or decreased DNA. For this reason, it is a challenge to identify both the regions that have CNAs and the genes whose expression could be deregulated (i.e., increased or decreased) because of gain or loss of their copies.

In this paper we focus on the role of copy number alteration in assessing prognosis of patients with CRC. Specifically, we show that the inference of the CRC progression benefits from exploiting CNA information when a particular relational representation of patients is given. The employed framework outperforms standard approaches where patients are represented through a set of available attributes. Documentation and software are available at <http://bimib.disco.unimib.it/people/claudia.cava/soft>. The data set for this analysis was provided by IRCCS Istituto Nazionale dei Tumori Milano (INT) and deposited at NCBI Gene Expression Omnibus (GSE16125).

Methods

Tissue specimens from 53 consecutive sporadic CRCs were obtained from previously untreated patients lacking family history and high-frequency microsatellite instability (MSI) who underwent surgical resection at the "Fondazione IRCCS Istituto Nazionale dei Tumori" (INT) Milano, between 1998 and 2000. After surgery all patients continued to be treated in INT, where their clinical course was constantly recorded. Tumor specimens containing more than 70% neoplastic cells and their surrounding normal mucosa were selected by

an experienced pathologist from cryopreserved tissue and used in a previous study of genetic features associated to colorectal carcinogenesis (Frattini et al, 2004). Microarray production was done following standard protocols by AROS Applied Biotechnology AS (Aarhus, Denmark). 51 DNA samples were hybridized to Affymetrix GeneChipVR Human Mapping 250K Nspl (SNP arrays). Raw intensity CEL files of the SNP arrays were processed with CNAG program v.2.0 (Copy-Number Analysis for Affymetrix GeneChips; Santa Clara, CA (Nannya et al, 2005) to detect chromosomal CNAs. Some samples were excluded due to poor quality hybridizations and unknown stage tumor progression (Reid et al, 2009). Also, stage-I patients were excluded because of the lack of instances in the considered data set. The selected cohort can be finally summarized as follows: 10 type-II patients, 10 type-III patients and 23 type-IV patients.

In order to quantify relationships between patients expressing the CCR progression, we first define a dissimilarity function over both an "advanced-stage" patient group and a specific "representative" base group e.g., patients with the lowest stage ("prototype"), then we classify patients according to the induced representations. In other words, the considered dissimilarity values quantify, by construction, subject differences due to different CNA information belonging to each subject. While in a "standard" case-control classification subjects are discriminated on its own set of attribute values, the dissimilarity function $D(f_x, f_y)$ is given through an estimation of the difference between the obtained CNA mean value frequency distributions f_x and f_y . In order to give a definition for D which can express dissimilarity between any pair of patients x and y (based on the CNA mean value frequency distribution f_x and f_y), we employ the symmetrised Kullback-Leibler (KL) divergence (Cover et al, 1991) between any pair of distribution f_x and f_y .

Table 1: a) Performances for the Standard Representation. b) Performances for the Dissimilarity Representation.

Test	Sensitivity	Specificity	PPV	NPV	Accuracy
a)					
stage II vs stage III	90,00%	80,00%	81.81%	88.88%	85,00%
stage II vs stage IV	100,00%	30,00%	76.66%	100,00%	78.79%
stage III vs stage IV	95.65%	20,00%	73.33%	66.66%	72.72%
b)					
stage II vs stage III	100,00%	90,00%	90.91%	100,00%	95,00%
stage II vs stage IV	91.30%	80,00%	91.30%	80,00%	87.88%
stage III vs stage IV	91.30%	80,00%	91.30%	80,00%	87.88%

Results and Discussion

The first issue of our investigation was to check the capability, for a given standard approach, of distinguishing patient groups. For this, we considered the following case - control study: (i) stage II (as control group) vs stage III (as case group); (ii) stage III (as control group) vs stage IV (as case group); (iii) stage II (as control group) vs stage IV (as case group).

All our evaluations employ a class of algorithms widely used in the machine learning community (i.e., the Support Vector Machine (Cristianini et al, 2000) within a k-fold cross-validation process. For the "standard" case, SVMs are given (input) matrices where patients are represented through the sequence of chromosomes as attributes, and each *i*-th component of the sequence is given by the CNA mean value associated to the chromosome *i*. Moreover, all experiments are evaluated by standard indices which are broadly applied in this context to measure the precision and recall capability of an inference system; i.e., sensitivity, specificity, positive (PPV) and negative predictive values (NPV), see for example (Davis et al, 2006). Table 1a) shows the performances when the classifiers are applied to the standard representations as discussed above. The standard approach is not able to discriminate both stage III from stage-IV patients (20% specificity) and stage II from stage-IV patients (30% specificity). On this basis, we used CNAs information to represent patients through dissimilarities as reported above. Table 1b) reports the inference performance when the dissimilarity representation is applied. We obtained substantially better accuracies reporting higher values of performances ($\geq 80\%$) for the whole set of the applied indices.

We showed that even a prediction analysis, concerning the progression of CRC, as characterized by the given staging classification system (Duke), benefits from exploiting CNA information when a specific representation of patients is considered. We point out that, in this work, the choice of a dissimilarity representation (i.e., the KL-divergence) has been addressed to obtain a function providing an estimation of the difference between the obtained CNA mean value frequency distributions for each pair of patients. More specific measures may be tested in future analysis.

Interesting questions on these arguments are reported in (Pekalska et al, 2005). Also the choice of a correct prototype set can be critical in this approach, and may change the results being investigated. This is another question which we are immediately interested for future analysis. Here we did not study the best possible prototype set, instead, the rationale for our choice was simply to employ a group of patients with a presumably low number of accumulated alterations. Numerical experiments indicate that the application of the applied representation for the considered data provide high precision and recall performances outperforming typical standard approaches where patients are represented through their set of available attributes. These results clearly suggest broader investigations either on different data sets or different CRC staging classification systems (Horton et al, 2005).

References

1. Cover T M, et al. (1991) Elements of information theory New York, NY, USA: Wiley-Interscience.
2. Cristianini N et al (2000). An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press.

3. Davis J et al (2006) The relationship between precision-recall and roc curves ICML '06: Proceedings of the 23rd international conference on Machine learning. New York, NY, USA: ACM, 233-240.
4. Fearon, E. and Vogelstein, B. (1990). Genetic model for colorectal tumorigenesis. *Cell*, 61:759–767.
5. Frattini M, Balestra D, et al. (2004) Different Genetic Features Associated with Colon and Rectal Carcinogenesis *Clin Cancer Res* 10(12):4015-4021.
6. Horton J K et al (2005) Staging of colorectal cancer: past, present, and future. *Clin Colorectal Cancer*, 4(5):302-12
7. Nannya Y, Sanada K (2005) A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. *Cancer Research* 65(14): 6071-6079.
8. Pekalska E et al (2005) The Dissimilarity Representation for Pattern Recognition: Foundations And Applications Machine Perception and Artificial Intelligence. World Scientific Publishing Company.
9. Reid J F, Gariboldi M, et al. (2009) Integrative approach for prioritizing cancer genes in sporadic colon cancer *Genes Chromosom. Cancer* 48(11):953-962
10. Shen, L., Toyota, M., et al(2007). Integrated genetic and epigenetic analysis identifies three different subclasses of colon cancer. *Proceedings of the National Academy of Sciences*, 104(47):18654-18659.
11. Vogelstein, B., Fearon, E., et al. (1988). Genetic alterations during colorectal-tumor development. *N. Engl. J. Med.*, 319:3526-3535.

Distilling structure in scientific workflows

Jiuqiang Chen^{1✉}, Christine Froidevaux², Carole Goble³, Alan R. Williams³, Sarah Cohen-Boulakia²

¹School of Information Science and Engineering, Lanzhou University, Lanzhou, Gansu, China

²AMIB group, INRIA, Saclay

³School of Computer Science, The University of Manchester, Oxford Road., Manchester

Motivation and Objectives

Scientific workflows management systems, (e.g., (Missier et al., 2010; Ludaesher et al., 2006; Goeck et al. 2011)) are increasingly used to specify and manage bioinformatics experiments. An experiment is then represented by a workflow in which a large number of bioinformatics tasks are linked to each other. A workflow specification is a framework for the execution of workflows. It specifies the order to be observed between the different tasks and their relationships with the workflow inputs and workflow outputs. According to the input data given to the workflow specification and assignments of values to the task parameters, different workflow runs are then obtained and may lead to different intermediate and final output data. Both workflow specifications and runs are represented by graphs.

Faced with the increasing complexity of runs and the need for reproducibility of results, provenance has become an important research topic. A significant number of tools for managing vast amounts of data provenance have been designed to assist the storage of provenance data (e.g., indexing), query the data (e.g., difference between executions, search for patterns), visualize the workflow provenance or (re)schedule executions... (See (Cohen-Boulakia and Leser, 2011) for a review on that topic). These tools all make intrinsically complex operations on graph structures (search for subgraphs in a graph, comparing graphs, ...), which, if carried out on Directed Acyclic Graphs (DAGs), with no other restriction of structure, lead to NP-hard problems. Instead, these problems can be solved in polynomial time when specific restrictions are imposed on graphs, such as considering series-parallel (SP) structures (Bein et al., 1992). Some provenance management approaches such as (Bao et al., 2009; Callahan et al., 2006) have therefore chosen to restrict workflow graphs to SP structures. As in general, workflows obtained using workflow systems are DAGs with any structure, graphs

transformation approaches such as (Escribano et al., 2009) can be exploited to transform workflow graphs into SP graphs. (Cohen-Boulakia et al, 2012) has recently designed SPFlow, the first algorithm able to rewrite any scientific workflow graph structure into an SP workflow structure while preserving provenance information. As expected, such an approach has a cost in that nodes and/or edges have to be duplicated in the rewritten workflow.

Determining the reasons why some workflows have non SP structures may help users to directly design workflows having a structure closer to SP structures. The rewriting process may then be used on less complex, distilled, workflows. The aim of this paper is to present the results obtained on the study that we have conducted on the set of Taverna workflows (Missier et al., 2010) available on myExperiment (De Roure et al, 2009) to analyze the reasons why workflows have non SP structures.

Methods

Our study has been conducted on a set of 1,014 distinct workflows extracted from the Taverna workflows available in myExperiment in May 2012. We have implemented the algorithm of (Valdes et al., 1979) to detect whether workflow graphs are SP. Intuitively, SP structures are graph structures having one main input (I in figure 1(a)) and one main output (O in Figure 1(a)), without loops and which can be synchronized. In particular the pattern highlighted in Figure 1 (b) is forbidden (in this pattern, arcs can be replaced by paths involving intermediate nodes). In this pattern, node w is responsible for the graph non to be SP. Such a node is called a reduction node (Bein et al., 1992) and is duplicated in SPFlow. In the workflow depicted in Figure 1(a) the getGeneInfo processor is a reduction node so that the workflow is not SP. Among the 390 workflows with non SP structures (38,5%), we have focused on identifying reduction nodes and analyzed the forbidden pattern in which they were involved. We have then driven two series of experiments:

- The first series of experiments has consisted in analyzing the structure of a subset of workflows having complex non SP structures.
- In the second series of experiments, we have considered all the non SP workflows of Taverna and we have conducted a study of the processors involved in non SP structures. We have identified the kinds of processors mostly involved in non SP structures and we have then made a more precise analysis by examining the processors themselves.

Results and Discussion

Trace links: The first series of experiments highlight the fact that some intermediate processors are directly linked to the workflow outputs merely for the sake of keeping track of intermediate results. We call such intermediate processors trace nodes and their outgoing edges linked to the workflow outputs are called trace links. On a total of 13,754 nodes in the set of non SP workflows, we found 1,524 reduction nodes including 631 reduction nodes that are also trace nodes (representing 41% of the reduction nodes) and

involved in 361 workflows (representing 92.6% of non SP workflows).

Interestingly, trace links could be removed by exploiting the powerfulness of the provenance module of Taverna that is in charge of collecting all intermediate and final results obtained and consumed during each execution.

Ongoing work includes focusing on the workflows for the BioVeL project and work in close collaboration with the workflow writers for potential improvement in the structure of some workflows when trace links may appear.

Non-SP-only processors: The second series of experiments revealed that most reduction nodes correspond to local processors (processors provided by Taverna to workflow designers) and web services processors. In particular, among a set of 92 web services, 40 services only appear in non SP workflows and occur at least once as reduction nodes. More interestingly, nine services appear only as reduction nodes in Non SP workflows. We call them Non-SP-only processors. As for local services, we found one Non-SP-only local processor.

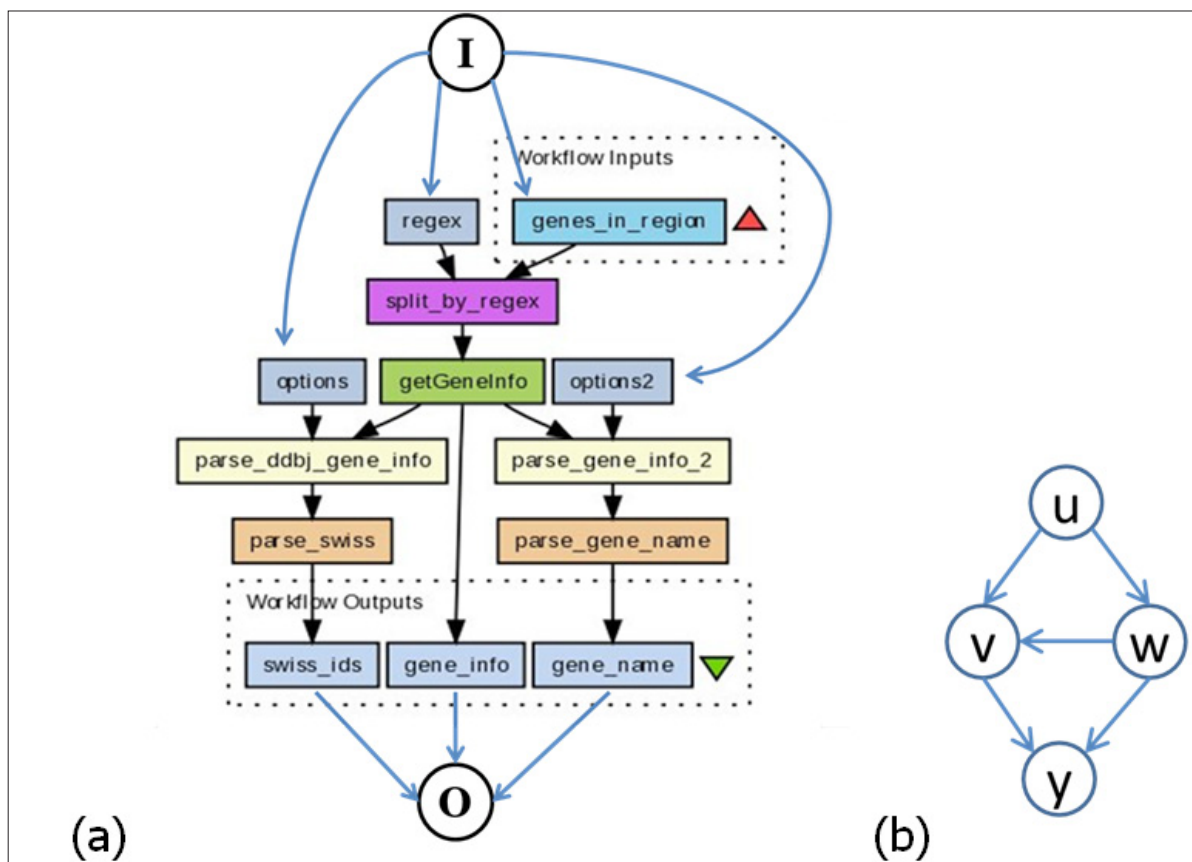


Figure 1: (a) Example of non SP structure from myExperiment; (b) forbidden pattern

Ongoing work includes investigating ways to modify the use of Non-SP-only processors (e.g., changing the processors ports, grouping several consecutive calls of the same processor, designing SP patterns of joint use) so that they are not anymore systematically associated to (and possibly responsible for) non SP structures.

In conclusion, we have identified several reasons why workflows may not have an SP structure. Following the solutions underlined, we will get distilled workflows in which the number of reduction nodes should importantly be reduced and we hope that a large part of workflows may become SP. In our approach, users do not have to consider structural constraints when they design workflows; our aim is instead to provide them with designing guidelines ensuring that distilled workflows are naturally produced.

Acknowledgements

This work has been partly supported by the INRA-INRIA ASAM project. J. Chen has been supported by the China Scholarship Council (CSC).

References

1. Bao Z, Cohen-Boulakia S, Davidson SB, Eyal A, Khanna S. (2009) Differencing provenance in scientific workflows, Proc. of ICDE 2009, 808-819. doi: [10.1109/ICDE.2009.103](https://doi.org/10.1109/ICDE.2009.103)
2. Bein W, Kamburowski J, Tallmann MF. (1992) Optimal reductions of two-terminal directed acyclic graphs, SIAM J. Comput. 21(6):1112--1129. doi: [10.1137/0221065](https://doi.org/10.1137/0221065)
3. Callahan SP, Freire J, Santos E, Scheidegger CE, Silva CT et al. (2006) Vistrails: visualization meets data management, Proc. of SIGMOD 2006, 745-747. doi: [10.1145/1142473.1142574](https://doi.org/10.1145/1142473.1142574)
4. Cohen-Boulakia S, Froidevaux C, Chen J. (2012) Scientific Workflow Rewriting while Preserving Provenance. Proc of the 8th IEEE Int. Conference on eScience. (In press)
5. Cohen-Boulakia S, Leser U (2011) Search, adapt, and reuse: the future of scientific workflows, SIGMOD Record 40(2):6-16. doi: [10.1145/2034863.2034865](https://doi.org/10.1145/2034863.2034865)
6. De Roure D, Goble C, Bhagat J, Cruickshank D, Goderis A et al. (2009) The Design and Realisation of the myExperiment Virtual Research Environment for Social Sharing of Workflows. Future Generation Computer Systems 25:561-567. doi: [10.1109/eScience.2008.86](https://doi.org/10.1109/eScience.2008.86)
7. Escribano A, van Gemund A J.C, Cardenoso-Payo V. (2009) Performance implications of synchronization structure in parallel programming, Parallel Computing 35(8-9) 455-474. doi: [10.1016/j.parco.2009.07.002](https://doi.org/10.1016/j.parco.2009.07.002)
8. Goecks J, Nekrutenko A, Taylor J, The Galaxy Team. (2011) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biology 11(8):R86. doi:[10.1186/gb-2010-11-8-r86](https://doi.org/10.1186/gb-2010-11-8-r86)
9. Ludaescher B, Altintas I, Berkley C, Higgins D, Jaeger E et al. (2006) Scientific workflow management and the Kepler system. Concurrency and Computation: Practice and Experience 18(10):1039-1065. doi: [10.1002/cpe.994](https://doi.org/10.1002/cpe.994)
10. Missier P, Soiland-Reyes S, Owen S, Tan W, Nenadic A et al. (2010) Taverna, Reloaded. SSDBM 2010, 471-481. DOI: [10.1007/978-3-642-13818-8_33](https://doi.org/10.1007/978-3-642-13818-8_33)
11. Valdes J, Tarjan RE, L. Lawler E. (1979) The recognition of series parallel digraphs, STOC, 1-12. doi: [10.1145/800135.804393](https://doi.org/10.1145/800135.804393)

CorrelaGenes: a new tool for the interpretation of the human transcriptome

Paolo Cremaschi¹, Sergio Rovida², Lucia Sacchi³, Antonella Lisa¹, Alessandra Montecucco¹, Giuseppe Biamonti¹, Silvia Bione¹, Gianni Sacchi²

¹Institute of Molecular Genetics, National Research Council, Pavia, Italy

²Institute of Applied Mathematics and Information Technology "Enrico Magenes", National Research Council, Pavia, Italy

³Dipartimento di Ingegneria Industriale e dell'Informazione, University of Pavia, Pavia, Italy

Motivation and Objectives

The comprehension of the molecular mechanisms involved in the physiology of human cells and in the pathogenesis of complex disorders, requires the development of new bioinformatic and biostatistic approaches able to integrate and interpret the huge amount of data derived from different kind of "omics" technologies. Nowadays, the interpretation of the transcriptional state of the cell and its alterations in particular experimental or pathological conditions is of particular interest. To this aim several technologies have been developed to identify and quantify the entire set of cellular transcripts, thus resulting in the availability of expression profiles of many different cell types in many different conditions. With the aim of contributing to the elucidation of transcriptional dynamics in the cell, we developed CorrelaGenes, a new bioinformatic tool that exploits the expression data available in the Gene Expression Omnibus (GEO <http://www.ncbi.nlm.nih.gov/geo/>) database. The main goal of this tool is to help identifying sets of genes whose expression appeared simultaneously altered in different experiments, thus suggesting co-regulation or coordinated action in the same biological process.

Methods

CorrelaGenes uses a PostgreSQL (<http://www.postgresql.org/>) 9.1.3: database initialized using the Curated DataSets in Homo sapiens cell lines publicly available in the GEO archive. The Extract Transform and Load process, described in Figure 1A, was created using the R language 2.14.1 available at The R Project for Statistical Computing (<http://www.r-project.org/>).

A total of 978 GEO DataSets were read using the GEOquery R package 2.21.9 (Davis and Meltzer, 2007) and transformed in objects suitable for the subsequent stages of the analysis. The DataSet design was manually analyzed to select 2120 biologically meaningful experimen-

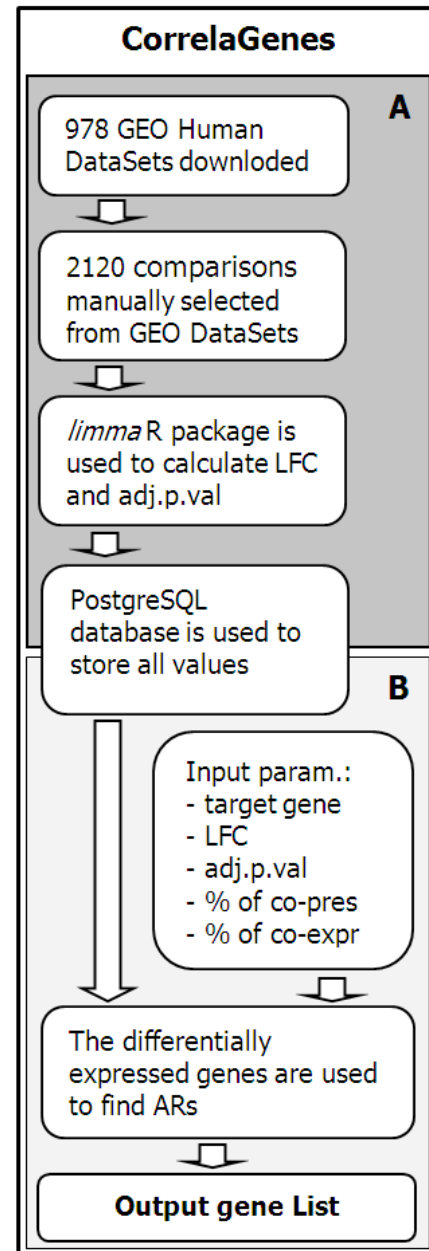


Figure 1: CorrelaGenes workflow. Panel A: PostgreSQL database initialization (R language). Panel B: Data Mining process (Fortran language).

tal comparisons. All the 2120 selected pairwise comparisons were analyzed with the limma R package 3.10.3 (Smyth, 2005) to calculate the log fold change (LFC) and the adjusted p values (adj.p.val) to identify the list of differentially expressed genes. All the values obtained from limma were stored in the PostgreSQL database.

The Association Rule Mining (ARM) is the unsupervised data mining technique that we used to discover genes that are frequently differentially co-expressed (Creighton and Hanash, 2003) in GEO DataSets. We used the standard ARM algorithms to look for Association Rules (ARs) limited to two genes (namely: IF Gene1 is differentially expressed THEN also Gene2 is differentially expressed) one of which is defined as an input parameter fixed for each search (i.e. target gene). The constraints used add a guided approach to the standard ARM technique with the aim of creating a list of genes sharing a coordinated expression with the target.

We defined two different criteria to select the most relevant ARs:

- percentage of co-presence (% of co-pres): as not all the comparisons include the same set of probes or some probes could be discarded for a not significant adjusted p value, we created an index to evaluate the percentage of comparisons where a gene is measured in relation of the whole number of comparisons where the target gene is measured;
- percentage of co-expression (% of co-expr): to evaluate the significance of the relationship between a gene and the target, we calculated the percentage of comparisons in which both genes are differentially expressed in relation of the number of comparisons where they are both measured.
- The procedure to perform the co-expression analysis, described in Figure 1B and implemented by a serial Fortran90 prototype code, can be summarized as follows:
- choice of the target gene and setup of the user defined indices for the analysis;
- initialization of the data structures (LFC and adj.p.val);
- identification of differentially expressed probes (a matrix of integer flags is defined, in order to select up-, down- and not-regulated or not-significant probes);
- selection of probes and comparisons associated to the target gene;
- evaluation of the percentage of significant values of both co-pres and co-expr for each single gene;
- creation of the list of all genes matching the selected criteria.

Results and Discussion

A total of 15 target genes (ACTG1, AFF3, APOE, APP, CDC5L, DIAPH2, EMD, FOXO1, HIF1A, IL8, MAPT, PRFP19, PSEN1, PSEN2, PTPN22) were used for the preliminary validation of the procedure with the following criteria: (i) adj.p.val \leq 0.05, (ii) absolute value of LFC \geq 0.65 (iii), % of co-pres \geq 40% and (iv) % of co-expr \geq 30%.

The simulations were carried out using a single blade of the CentOS IBM Cluster at IGM-CNR in Pavia. The cluster consists in six computational nodes, interconnected by Gigabit Ethernet and 10G Fiber Channel. Each node is a two processors Intel Xeon E5640 2.66 GHz, sharing 48 GB of RAM. The performance of the algorithm was evaluated using the execution time.

Averaging on the considered 15 genes, the whole procedure requires a mean execution time of 1221 sec for the co-expression analysis of a single gene. We evaluated the average cost of each phase as percentage of the total execution time. The profiling of the code showed that 64.3% of the total time is spent initializing the data, 35.5% is spent creating the different gene lists and only the 0.2% is actually spent gene-rating the ARs. The analysis algorithm exhibits an intrinsic data-parallelism at the level of the processing of the gene, a feature that will be further investigated in order to improve the performance of the whole procedure. A naïve approach to the parallelization consists in the multithreaded implementation for the creation of the gene lists by means OpenMP directives. Anyway, as the limiting step is the data initialization, a brand new approach to overcome this problem could be considered.

The gene lists created starting from the selected 15 target genes, were analyzed for their biological content in order to assess the relevance of the results obtained.

A first observation regards the highly variable number of associated genes extracted for each target gene (i.e. ranging from 99 to 2951) that could be due both to the different number of comparisons in which the target gene was mod-

ulated or to the different transcriptional behavior of the genes in the cell. Moreover, we found a quite large number of genes shared by all the 15 lists. This could either reflect the presence of constitutively modulated genes eventually involved in basic cell processes or be the consequence of a too tolerant choice of the parameters used in the simulation.

Some more detailed biological characterization was performed for the 2014 genes of the list extracted with PRPF19 as target the Database for Annotation, Visualization and Integrated Discovery (DAVID, <http://david.abcc.ncifcrf.gov/>). We used the Database for Annotation, Visualization and Integrated Discovery (<http://david.abcc.ncifcrf.gov/>) to query the Gene Ontology (GO, <http://www.geneontology.org/>) for the Biological Process subset of terms. Consistently with the literature data, the GO terms found significantly enriched (Benjamini corrected p value < 0,05) were related to the main known functions of PRPF19 in the cell (i.e. cell cycle, apoptosis, pre-mRNA splicing, DNA damage repair). We also investigated the gene list extracted for CDC5L (n=2794), a gene known to interact with PRPF19 in the pre-mRNA splicing complex (Grote et al., 2010). Despite the fact that the two genes were not selected as associated, a large overlap was found between the two lists (1531 genes in common). This list contains mainly genes related to cell cycle and splicing process. Moreover, an analysis with data obtained with the GeneMANIA (<http://www.genemania.org/>) and the STRING (<http://string-db.org/>) web tools for the two genes gave an independent confirmation for a number of genes extracted by our Correlagenes tool. Finally, a set of five genes involved in the pathogenesis of Alzheimer disease (Carter, 2007), a common human neurodegenerative disorder,

were included in our simulation (APOE, APP, PSEN1, PSEN2 and MAPT). A group of 952 genes were found in common among the five extracted lists thus suggesting the presence of shared pathways that could be exploited for further investigation of pathogenetic mechanisms.

The preliminary results of the simulation showed how Correlagenes could contribute to the characterization of transcriptional profiles in the cell and in the definition of molecular pathways and biological process. Moreover, it integrates expression results obtained from other available tools. The good performances shown during the simulation phase encourage us to plan wider validation steps to enhance the accuracy and the reliability of our instrument.

Acknowledgements

This work was supported by Cariplo Foundation grant n. 2010/0253 and by Progetto di interesse CNR-MIUR "Invecchiamento".

References

1. Carter CJ (2007) Convergence of genes implicated in Alzheimer's disease on the cerebral cholesterol shuttle: APP, cholesterol, lipoproteins, and atherosclerosis. *Neurochem Int.* 50(1) 12-38
2. Creighton C, and Hanash S (2003) Mining gene expression databases for association rules. *Bioinformatics* 19(1) 79-86. doi:10.1093/bioinformatics/19.1.79.
3. Davis S, and Meltzer PS (2007) GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* 23(14):1846-1847. doi: 10.1093/bioinformatics/btm254.
4. Grote M, Wolf E, Lemm I, Agafonov DE, Schomburg A, et al. (2010) Molecular Architecture of the Human Prp19/CDC5L Complex. *Mol Cell Biol.* 30(9) 2105-2119. doi: 10.1128/MCB.01505-09.
5. Smyth GK (2005) Limma: linear models for microarray data. In: Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W (Eds.) *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. Springer, pp. 397-420

Business intelligence for biopharmaceutical company

Anass El Haddadi[✉], Bernard Dousser

IRIT – UMR 5505, University of Toulouse III, Toulouse, France

Motivation and Objectives

Competition is a fundamental concept of economic market, which requires companies to practice Competitive Intelligence (CI) in order to be advantageously positioned on the market. Therefore, companies are required to monitor constantly the information's sources, for detect any change in order to make appropriate solutions in real time. However, for a successful monitoring, we should not be satisfied merely to monitor the opportunities, but before all, to anticipate risks. The external risk factors have never been so many: extremely dynamic and unpredictable markets, new entrants, mergers and acquisitions, sharp price reduction, rapid changes in consumption patterns and values, fragility of brands and their reputation.

To face all these challenges, our research consists in proposing a Competitive Intelligence System (CIS) designed to provide online services. Through in a descriptive and statistics exploratory methods of data, Xplor EveryWhere (XEW) display, in a very short time, new strategic knowledge such as: the profile of the actors, their repu-

tation, their relationships, their sites of action, their mobility, emerging issues and concepts, terminology, promising fields etc.

Methods

In our research team, we coordinate the process of CI around three concepts: strategic analysis, environmental scanning and information system. The CIS XEW (el. haddadi et al., 2011a. b.) aims to improve decision-making in all aspects in business life, particularly offensive and innovative decisions. XEW based on a multidimensional analysis model, whose objective analyzed the information environment in all dimensions of a decision problem, with the exploitation of information by analyzing the evolution of their interactions. Our approach combines two methods: knowledge discovering in text (KDT) and environmental scanning.

The dynamic aspect is vital to any analysis in the context of CI. These dynamics include continuous monitoring of the business environment in order to detect its changes and developments. The proposed information system, based on an exploratory multivariate analysis model: 'the re-

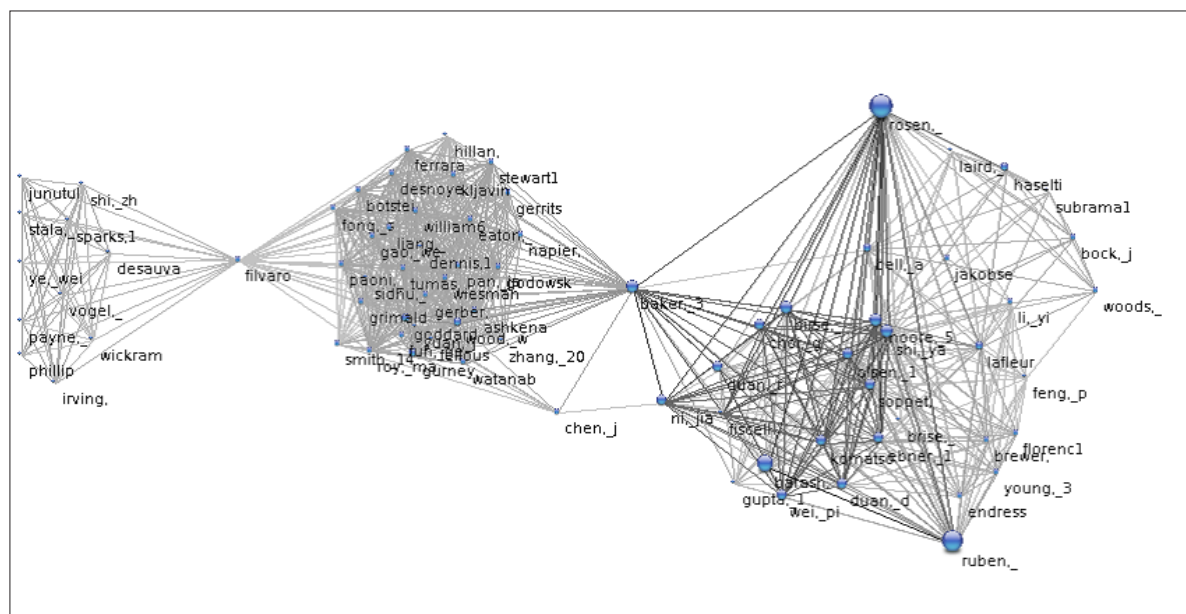


Figure 1. The inventor's networks for scaffolds research

lational aspect and the time dimension', which we call Xplor. It is based on extracting knowledge from textual data by analyzing relational data and their evolution. This model allows time specification to situate events, strategy and actions as well as in: the past by reconstructing the chronology; the present- oriented time to detect weak signal and the future to detect relationships in network, such as partnerships, alliances, mergers, acquisitions, co-citations, co-signatures, co-occurrences of all kinds

Results and Discussion

With the evolution of technology, such a CIS will enables us to increase efficiency and responsiveness, because at any moment, it is possible to gain access all strategic information by markers itself can be information back very quickly "field" which may possibly trigger other strategic analysis, for example it's possible to detect the social network, the semantic network, the company network and the others types of network's (figure 1).

The experimental system XEW shows a user satisfaction, regardless of their field: computer scientists, statisticians, analysts, decision-maker or watchmen, etc.. The majority describes the prototype in interactive, with ergonomic, collaborative and a information system for decision support.

For majority users the mobility is very important. They can now enjoy the advantage of data analysis everywhere. This experiment allowed us to validate the proposed prototype and consider measures to improve Xplor EveryWhere. We analyze these measures in our research perspectives.

References

1. El. Haddadi A, Dousset B, et al. (2011 a) Xplor EveryWhere - The competitive intelligence system for mobile. In Proceedings of the International Conference on Multimedia Computing and Systems, IEEE.
2. El. Haddadi A, Dousset B, et al. (2011 b) Discovering patterns in order to detect weak signals and define new strategies. In IGI Global.

Biomedical Text Mining for Disease Gene Discovery

Sarah ElShal^{1,2✉}, Jesse Davis³, Yves Moreau^{1,2}

¹Dept. of Electrical Engineering (ESAT-SCD) - K.U.Leuven, Leuven, Belgium

²IBBT-KULeuven Future Health Department, Leuven, Belgium

³Dept. of Computer Science- K.U.Leuven, Leuven, Belgium

Motivation and Objectives

Because of the amount of electronic literature now available, it is challenging for biologists to search biomedical corpora for any kind of desired information beyond simple text retrieval. Several tools have been developed to make text mining easier for them. Some of these tools focus on extracting biomedical terms; such as protein names and biological processes, given any input text. The tools COREMINE Medical (<http://www.coremine.com>, last accessed on 25 September 2012) and GoPubMed: (<http://www.gopubmed.com>, last accessed on 25 September 2012) are just two examples. Other tools apply rule-based strategies to relate biomedical concepts to each other. E.g., BITOLA (<http://ibmi.mf.uni-lj.si/bitola/>, last accessed on 25 September 2012) (Hristovski et al., 2005).

We have been developing a methodology and tool to discover genes implicated in any given disease or disorder. In fact, our tool takes from the user any free text query as an input and attempts to identify those genes most strongly linked to the query. As an output, the tool returns an ordered list of the best genes matching the query. The core work of our tool is based on text mining. Basically, each gene is linked to a profile that contains the biological terms that are most significant for it. Similarly, we link the input query to a corresponding keyword profile. The genes appearing at the top of the output list are the ones whose profiles are highly similar to that of the input query.

Methods

The text mining strategies we use in our work are applied to the biomedical abstracts published in PubMed (<http://www.ncbi.nlm.nih.gov/pubmed/>, last accessed on 25 September 2012). We divide our work into two phases: a background phase, and a live phase. In the background phase, we collect all the abstracts annotated to every gene described in *Entrez Gene* (<http://www.ncbi.nlm.nih.gov/gene/>, last accessed on 25 September

2012). For this, we use GeneRIF (<http://www.ncbi.nlm.nih.gov/gene/about-generif> last accessed on 25 September 2012), which provides functional annotation between genes and PubMed references. Afterwards, we index all the referenced abstracts via MetaMap (<http://metamap.nlm.nih.gov/>, last accessed on 25 September 2012) (Aronson, 2001), which maps the given biological text to the Unified Medical Language System (UMLS) Metathesaurus (<http://www.nlm.nih.gov/research/umls/>, last accessed on 25 September 2012) (Bodenreider, 2004). Thus for each gene, we could maintain a list of UMLS biomedical terms that functionally-describe it. We call this list a gene keyword profile. Then for each gene, we build another weighted profile in the form of a vector of Term Frequency-Inverse Document Frequency (TF-IDF). For a given gene, each entry in the vector measures how relevant a specific UMLS term is to the gene. We refer to the whole set of gene vectors as the "reference matrix". An example of this reference matrix is shown below in Table 1.

In the live phase of our work, we take a free text query as an input from the user (e.g., sleep disorders). Then, we use the E-utilities from PubMed to retrieve the corresponding abstracts that are relevant to the user query. And as we did with the genes in the background phase, we generate a keyword profile for the query and consequently a corresponding TF-IDF vector. Finally we match this query vector against all the gene vector entries in the reference matrix. Each match corresponds to a score that is calculated via a dot-product. The higher the matching score of a given gene vector entry, the more probably the gene relates to the user query. We also take into account the frequency of citation of a given gene. So genes appearing early in the ordered output list, do not only share the most similar profiles with the user query, but they are also cited by the highest number of references. Besides, we consider the fraction of common references between the query and the candidate genes as an additional scoring factor. As the number of com-

Table1: An example of the Reference Matrix. The heading row corresponds to a set of different UMLS terms (DNA-binding, cancers, tumours, ..., diabetes, and peptides). The heading column corresponds to two gene examples (Breast Cancer Type 1 (BRCA1), and Insulin (INS)). The numbers in each row (vector) correspond to the TF-IDF values of each UMLS term given the heading gene. For example, we observe that for BRCA1, the terms "Cancers" and "Tumours" have high TF-IDF values. This is related to the fact that they have high frequency of occurrence in the abstract texts annotated to BRCA1. Besides, we also observe that "DNA-Binding", "Diabetes", and "Peptides" have low TF-IDF values since they are not that frequent in the annotated text.

	DNA-Binding	Cancers	Tumors	...	Diabetes	Peptides
BRCA1	0.0	10.3	9.8	...	2.3	0.0
INS	0.0	3.7	0.0	...	10.5	9.3

mon references increases, the matching score of the candidate gene also increases.

Results and Discussion

The evaluation of the results is still ongoing. To validate the quality of our results, we use as a benchmark the phenotype-gene annotation provided by the Human-Phenotype-Ontology (HPO <http://human-phenotype-ontology.org/>, last accessed on 25 September 2012) (Robinson and Mundlos, 2010). For every phenotype in this annotation, a set of linked genes are recorded. The links are provided based on the information about the phenotypes of a given syndrome, and the genes known to cause that syndrome. Hence in our tool, we use each phenotype in the annotation file as a separate free text input. Then for each gene output list, we measure the percentage of recall against the HPO annotation.

To assess the power of our tool, we use some general biomedical search systems as a base-

line (e.g., Gene Ontology <http://www.geneontology.org/>, last accessed on 25 September 2012). We are expecting our tool to perform better. That is because such general systems rely on clear evidence to associate a gene product with a given query (e.g., inference from experiments or by curators), while our tool digs deeper in all the published literature as discussed in the Methods section.

References

1. Aronson AR (2001) Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp: 17.
2. Bodenreider O (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 32: D267. doi:10.1093/nar/gkh061
3. Hristovski D, Peterlin B, et al. (2005) Using literature-based discovery to identify disease candidate genes. Int J Med Inform 74: 289.
4. Robinson PN, Mundlos S (2010) The Human Phenotype Ontology. Clin Genet 77: 525

An ontological-based knowledge organization for bioinformatics workflow management system

Antonino Fiannaca, Massimo La Rosa, Salvatore Gaglio, Riccardo Rizzo, Alfonso Urso

ICAR-CNR, National Research Council of Italy, Palermo

Motivation and Objectives

In the field of Computer Science, ontologies represent formal structures to define and organize knowledge of a specific application domain (Chandrasekaran et al., 1999). An ontology is composed of entities, called classes, and relationships among them. Classes are characterized by features, called attributes, and they can be arranged into a hierarchical organization. Ontologies are a fundamental instrument in Artificial Intelligence for the development of Knowledge-Based Systems (KBS). With its formal and well defined structure, in fact, an ontology provides a machine-understandable language that allows automatic reasoning for problems resolution. Typical KBS are Expert Systems (ES) and Decision Support Systems (DSS). ESs gather and formalize the knowledge of a human expert of a domain in order to produce inferences and recommendations given an initial query. DSSs are more interactive KBS, in the sense they offer support, rather than replacement, for the decision making process during the execution of a task, suggesting one possible strategy or tool given a set of initial conditions. DSSs are mainly adopted in the clinical field, where they are called Clinical DSS (CDSS). Ontology specification, structure and organization are then of fundamental importance for the development of a KBS.

In this paper we present an improvement of our ontological approach for knowledge organization in DSS design. In our previous publication (Fiannaca et al., 2012) we defined a paradigm for ontology specification named Data Problem Solver (DPS) and we showed how our approach can be applied to bioinformatics domain, modeling the Protein-Protein Interaction Network extraction scenario. In the proposed approach, we aim at integrating into our ontology the concept of Workflow as a set of processes. Our main objective is to provide a general schema in order to add the functionalities and capability of a DSS to the more recent Workflow Management Systems, that especially in bioinformatics, with

the Taverna workbench (Hull et al., 2006), represent a powerful instrument for researchers. We called our extended ontological approach Data Problem Solver Workflow (DPSW).

Methods

DPSW ontology is shown, using UML notation, in Figure 1. The four main entities are, as the name suggests, Data, Problem, Solver and Workflow. Problem represents the set of Tasks to do in an application scenario, and it models the task decomposition from more complex goals to simpler ones. Data summarizes the type of information needed to perform a task belonging to a Problem. Data concept is specialized by Data _ Type class, representing the type of input and output data of a task, and each Data _ Type has one or more Data _ Format that encodes it. Solver concept fills the gap between a Problem to solve and the Tools that actual solve it. Each Solver is characterized by a computational Approach (probabilistic, topological, numerical approach for instance) and it models the expert knowledge (in terms of heuristics or strategies) on which Tool, or combination of Tools, are needed in order to accomplish a Task. Solver class is also characterized by a set of attributes, not shown in Figure 1, that specifies what are the pros and cons for using a solving strategy. Tool class identifies the generic entity that can be actually run, and it generalizes the concept of Algorithm, Web Service, Application, Device. Each Tool has a computational Paradigm (for example neural networks, graph analysis, etc...) and eventually a set of configuration Parameters; it requires a Data _ Format object (the input file), and of course other type of Tools can be further added. By considering separately Problem, Solver and Data, we want to clearly separate among the models of the problem itself, the way to resolve it, and the input data requested. This way we aim at enhancing the generalization, modularity and expandability features of the proposed ontology. The last main component of the proposed ontological approach is the Workflow entity. It rep-

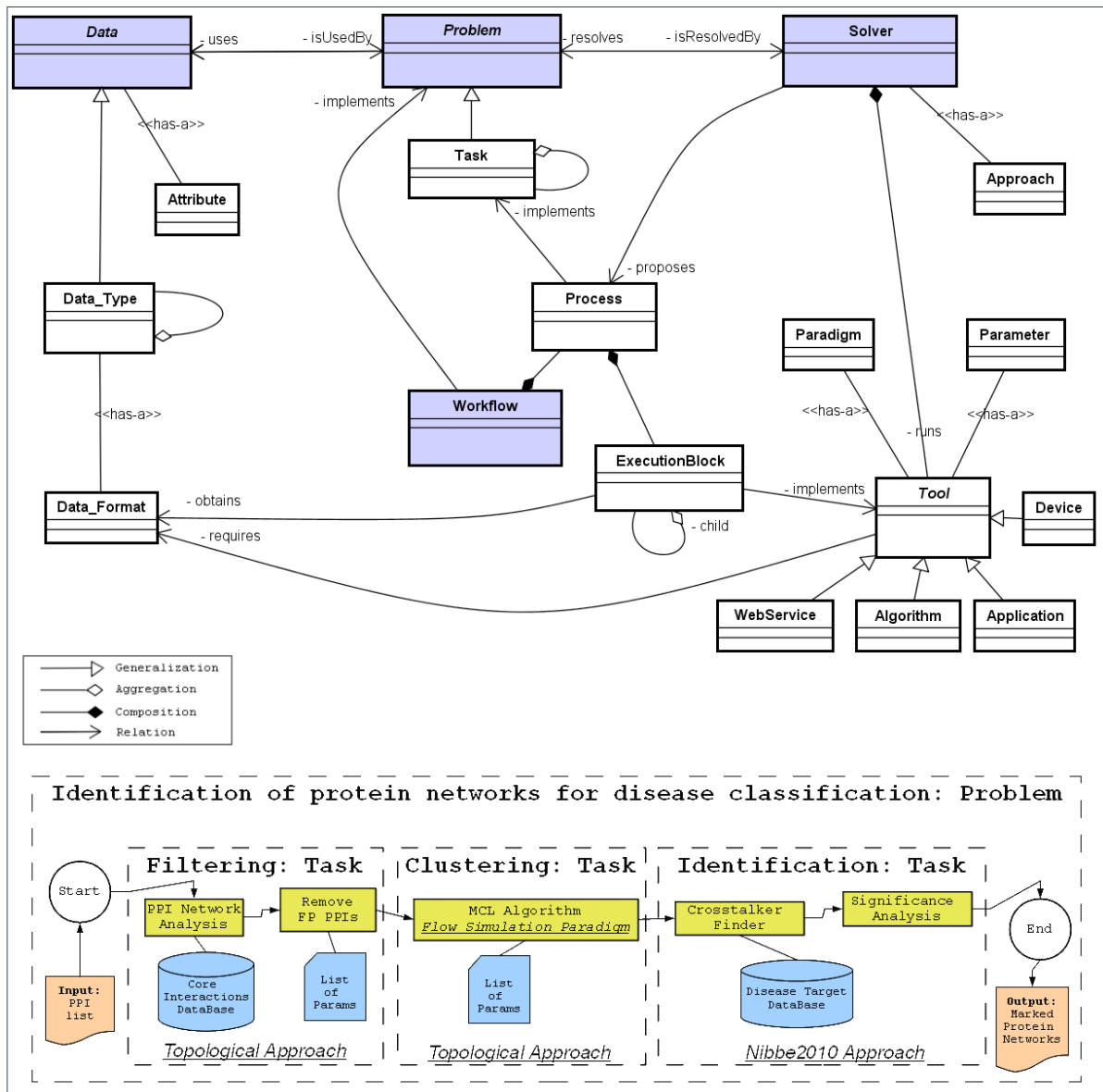


Figure 1 - The proposed DPSW ontology and its case of study.

resents the graphical view of a problem and its solving components. A Workflow is composed of one or more Processes, that can be seen as part of the global workflow implementing a specific Task, and each Process, in turn, is composed of ExecutionBlocks, that are the visual representations of an executed Tool. By embedding the concept of Workflow into our ontological structure, we want to provide a full Knowledge Base specification that can be used as building block of a DSS whose suggestions during a bioinformatics experiment can immediately be translated into an executive workflow.

Results and Discussion

In order to show how the proposed ontology can match with a real bioinformatics issue, we have taken into account a key challenge of cancer research, i.e., the detection of protein sub-networks that identifies markers correlated with metastasis. In facts, each protein complex is suggestive of a distinct functional pathway, that can provide novel hypotheses in organisms analysis (Sharan et al., 2007). A workflow related to this case of study is reported in the bottom of the Figure 1. In this example, we consider the "identification of protein networks for disease

classification", that, according to DPSW ontology, represents the problem concept; the implementation of this problem (the experiment) matches with the workflow concept. Here, we take as data input a list of protein-protein interactions (PPIs) and produces as data output a list of marked protein network, that could be responsible for some specific diseases. According to the related literature, this problem could be arranged in three main tasks: filtering, clustering and identification. For instance, the first task has been handled by some authors (Ucar et al., 2005) with a topological approach; in facts, they developed some graph-based algorithms in order to eliminate redundant false positive interactions from the original PPI dataset. This preprocessing strategy points to increase the reliability of PPI-Network. As regarding the second task, i.e. finding meaningful groups of biological units, a number of approaches have been proposed and a lot of them are based on clustering. A well-know algorithm is Markov Clustering Algorithm (MCL) (Enright et al., 2002), that divides the graph by means of "flow simulation paradigm". In facts, it separates the graph into different segments, with an iteration of simulated random walks within a graph. Once sub-networks are obtained, it is possible to identify those complexes that demonstrate a differential expression with respect to carcinogenesis phenotype, by means of an integrative -omics approach proposed in (Nibbe et al., 2010). Using these elements, we could obtain some putative disease protein sub-networks. Ultimately, in order to face with this case of study, we propose to use three tasks, two different approaches and six tools (both algorithms and applications). Each executed tool, with its proper input/output file and parameters, is stored into an instance

of the execution block concept, whereas a set of execution blocks that complete a single task are stored as an instance of process concept. Notice that workflow in Figure 1 has been defined using some different approaches that, we suppose, are contained into the knowledge base arranged according the DPSW ontology. Using the proposed ontology, the experimentalist can generate some novel workflows composed of both piece of well know techniques and some processes previously stored as instances of DPSW ontology. As future work, we will use this ontology for building an expert system for making reasoning in the analyzed case of study.

References

1. Chandrasekaran B, Josephson J, et al. (1999) What are ontologies, and why do we need them?. *IEEE Intelligent Systems and Their Applications* 14(1):20-26
2. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research* 30(7):1575-1584. doi: [10.1093/nar/30.7.1575](https://doi.org/10.1093/nar/30.7.1575)
3. Fiannaca A, La Rosa M, et al. (2012) An ontology design methodology for Knowledge-Based systems with application to bioinformatics. *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)* :85. doi: [10.1109/CIBCB.2012.6217215](https://doi.org/10.1109/CIBCB.2012.6217215)
4. Hull D, Wolstencroft K, et al. (2006) Taverna: a tool for building and running workflows of services. *Nucleic Acids Res* 34:W729-32
5. Hibbe RK, Koyuturk M, Chance M R (2010) An integrative -omics approach to identify functional sub-networks in human colorectal cancer. *PLoS Comput. Biol.* 6(1)
6. Sharan R, Ulitsky I, Shamir R (2007) Network-based prediction of protein function. *Mol Syst Biol.* 3:88. doi: [10.1038/msb4100129](https://doi.org/10.1038/msb4100129)
7. Ucar D, Parthasarathy S, Asur S, Wang C (2005) Effective Pre-processing Strategies for Functional Clustering of a Protein-Protein Interactions Network. *5th IEEE Symposium on Bioinformatics and Bioengineering* 129: 371-382. doi: [10.1109/BIBE.2005.25](https://doi.org/10.1109/BIBE.2005.25)

A tool for the extraction of new disease biomarkers from ex vivo and in vivo data

Francesca Gallivanone^{1✉}, Carla Canevari², Isabella Sassi², Alberto Marassi², Maria Picchio², Maria C Gilardi¹, Isabella Castiglioni¹

¹Institute of Molecular Bioimaging and Physiology (IBFM), CNR, Milan, Italy

²Ospedale San Raffaele, Milan, Italy

Motivation and Objectives

Recent studies have demonstrated the correlation among features of diseases obtained from ex vivo and in vivo Molecular Imaging (MI) studies (e.g. Genomics and Positron Emission Tomography, PET, Strauss et al., 2008; Genomics and Computerized Tomography, CT, Segal et al., 2007), opening a new role to non invasive clinical MI technologies in the current approach of personalized medicine.

At present, proper databases and statistical methodologies are on hand to deal with the different modalities of ex vivo and in vivo MI data but tools for the extraction of new disease biomarkers are still not available for the purpose of clinicians. The main interest of clinical specialists is in finding biomarkers of disease with diagnostic and prognostic values, and this can be performed with an interdisciplinary approach offered by ex vivo and in vivo MI and then translated into the clinical environment.

Aim of this work was the development of a software tool ("cOuch" Correlative and Collaborative Touch System) (Castiglioni et al., 2011) designed to be used by clinicians to find new diagnostic/prognostic biomarkers of disease from the comparison of ex vivo and in vivo data of patients. In this work, as representative example, "cOuch" has been applied to assess the prognostic and diagnostic value of the Standardized Uptake Value (SUV, Graham et al., 2000), a parameter of regional metabolic uptake measured by PET and 18F-labeled fluorodeoxyglucose for breast cancer lesions.

Methods

A. Couch functions and requirements

cOuch has been developed in Matlab R2008b. The Matlab standard toolbox was used, including Processing Toolbox, Statistic Toolbox and Curve Fitting Toolbox. Ad-hoc utilities were implemented with Java languages. A user friendly graphi-

cal interface (GUI) has been designed for clinical users. The GUI consists of different sections: the population of a Database of patient data (different modality data from different in vivo and ex vivo diagnostic tests), a section for Data Pre-processing (to select the category of the Variable Type for a specific correlation test) and a section for Data Processing (to select the specific correlation test). cOuch can be implemented with a standard or touchscreen hardware, on condition that a minimum Ram of 1GB (2GB is recommended) is installed for a 32-Bit system, due to the large software memory footprint required. A 64-Bit operating systems allows to achieve the best performances. Touchscreen system compliance offers the possibility to use cOuch by touchscreen tablets and mobile phones.

B. The Database

The cOuch database is a Relational MySQL Database that is populated through the GUI by registered and authorized clinical specialist users. In vivo and ex vivo MI data, including PET metabolic parameters of the oncological lesions, CT anatomical lesion volume, histopathological indexes and up/down regulated proteins extracted from surgical report samples, can be stored in the database. Every data is standardized and archived in standard formats. Patient data are anonymized and identified by an ID. Couch software is connected to the database and allows to perform statistical analysis and to save analysis reports.

C. Data Pre-processing

The variable type can be considered using its intrinsic type or can be transformed on the basis of proper options, in order to perform different statistical tests. Histological type, as assessed using the pTNM pathological classification, is treated as nominal data in all statistical analysis. Histological grade is treated as nominal data in some statistical analysis but reduced to nominal dichotomous data (only two categories) in

other statistical test. pT stage, as assessed using the pTNM pathological classification, is treated as categorical (four categories) in the statistical analysis. Lymph node status, as assessed using the pTNM pathological classification, is considered as nominal dichotomous variable, distinguishing the cases with no involvement of lymph node from the cases with involvement of lymph nodes. The expression of receptors (e.g. ER, PgR in the case of breast cancer), whose values range from 0% to 100, are considered as ordinal numerical continuous variables, except when not found expressed they are considered as nominal dichotomous data. Tumours are considered to over-express oncoproteins (e.g. c-erbB-2 positive in the case of breast cancer) if more than 30% of invasive tumour cells show definite membrane staining (Score 3 +) or if a definite membrane staining is found smaller than 30% (Score 2+), resulting in the so-called fishnet appearance. SUV is considered as an ordinal continuous variable.

D. Data Processing

Mann-Whitney tests and Kruskal-Wallis tests were implemented to assess the relationship between in vivo MI features with histopathological, immuno-histochemical parameters and up/down regulated proteins. Multivariate linear regression and clustering algorithms were implemented to assess multiple dependences of groups of parameters, also following some methods reported in literature considering matrix of protein expression as input for correlation analysis (Kim et al., 2012). In order to assess the diagnostic and prognostic role of the potential considered biomarkers, a ROC analysis was implemented with the purpose to find cut-off values for the biomarker. Specificity and Sensitivity were also calculated. This analysis allows to differentiate groups of patients on the basis of the values of the biomarker, higher or lower to the cut-off, being the two groups correlated to different biological characteristics of tumor (e.g. as assessed by histopathology).

E. Application of cOuch to real clinical studies

A protocol for the collection of in vivo and ex vivo MI data with the purpose of integration has been designed for one representative population of breast cancer patients. The study protocol, approved by the Institutional Review Board of the Scientific Institute H San Raffaele, involved 40 patients with biopsy-proven breast cancer designed for surgical intervention without per-

forming any treatment before surgery. Eligible patients underwent a total-body 18F-FDG PET/CT exam before surgery, and, during surgical intervention, biological samples were collected and analyzed at the Pathological Anatomy Unit and sent to the proteomics laboratory. For each patient, the senology, the anatomo-pathologist, and the biologist submitted their own reports to the nuclear medicine physician who calculated the SUV of the primary breast lesion from the PET/CT images and imported it in the database together with the other ex vivo data. The nuclear medicine physician performed the correlation analysis with cOuch between the considered parameters.

Results and Discussion

SUV was found correlated with histological type. Mann-Whitney test on SUV and the histological type (Invasive Lobular Carcinoma, ILC vs Invasive Ductal Carcinoma, IDC) showed that SUV was significantly ($p < 0.02$) lower in ILC (2.92 ± 0.94 g/cc) with respect to IDC (7.45 ± 6.16 g/cc). ROC curve analysis showed that a threshold of 3.87 g/cc for SUV allowed to distinguish ILC from IDC histological types with a Specificity of 67% and a Sensitivity of 100%. SUV was found correlated with histological grade. Kruskal-Wallis test on SUV and the histological grades (G1 vs G2 vs G3) showed that SUV was significantly ($p < 0.01$) different in G1 (3.90 ± 3.72 g/cc) with respect to G2 (5.61 ± 5.52 g/cc) and to G3 or G3 (11.30 ± 5.85 g/cc). Mann-Whitney test on SUV and the histological grade (G1 vs G3) showed that SUV was significantly ($p < 0.03$) lower in G1 (3.90 ± 3.72 g/cc) with respect to G3 (11.30 ± 5.85 g/cc). ROC curve analysis showed that a threshold of 3.99 g/cc for SUV allowed to distinguish G1 from G3 histological grades with a Specificity of 83.9% and a Sensitivity of 88.9%. No significant correlation were found between SUV of primary BC lesion and lymph node status either as detected by histopathology either as detected by PET. SUV was found correlated with ER and PgR hormone receptors. ER positive tumors showed a lower 18F-FDG (5.60 ± 5.14 g/cc) with respect to ER negative tumors (13.90 ± 5.65 g/cc) ($p < 0.005$) and PgR positive tumors showed a lower 18F-FDG (5.55 ± 5.13 g/cc) with respect to PgR negative tumors (14.04 ± 5.66 g/cc) ($p < 0.005$). As expected, the same relationship was found considering the total expression of hormone receptors

but with a lower statistical significance ($p < 0.02$). An analysis of ROC curves showed that a threshold of 7.75 g/cc for SUV allows to distinguish both ER positive from ER negative and PgR positive from PgR negative with a Specificity of 100.0% and a Sensitivity of 78.1%. No correlations were found between SUV and c-erbB2 but this could be due to the poor sample. In fact, excluding patient with incomplete information on c-erbB2 expression, only four patient presented an over-expression of c-erbB2 index. Using a threshold of 18% for MiB-1 proliferation index, we found that SUV was significantly ($p < 0.05$), correlated with MiB-1 proliferation index in particular $SUV = 4.52 \pm 2.92$ g/cc for tumor with an expression of MiB-1=18% and $SUV = 9.30 \pm 7.40$ g/cc for tumor with an expression of MiB-1 > 18%. An analysis of ROC curves was performed on the two clusters of data obtained using a 18% threshold. A value of 4.06 g/cc for SUV allows to distinguish positive or negative values of MiB-1 proliferation index, with a Specificity of 70.6% and a Sensitivity of 65.0%. Univariate linear regression analysis was performed on MiB-1 and SUV. Even if the significance of the estimated parameter for MiB-1 proliferation index was strong ($p < 0.001$), indicating the correlation already evaluated using Mann-Whitney test, the significance of the regression was low (R-square < 0.3) showing that linear relationship is not effective. Hierarchical clustering was applied involving SUV and the variables which were found one-to-one correlated with SUV by univariate tests: histological type and grade, hormone receptor status and MiB-1 proliferation index. K-means pre-processing on SUV was performed following the results obtained by univariate analysis. K-means algorithm allowed to define three intervals of SUV

values: a) SUV from 0.78 g/cc to 3.07g/cc; b) SUV from 3.40 g/cc to 4.38 g/cc; c) SUV from 7.03 g/cc to 27.53g/cc. A hierarchical cluster analysis allowed to define two different clusters of multiple-correlated indexes: a) a first cluster including ILC and IDC G1 tumors with a negative expression of MiB-1 and SUV in the first interval. Index of positive expression of hormonal receptors and SUV in the second interval are linked to this cluster but with a lower significance; b) a second cluster including IDC tumors G2 and G3 tumors with a positive expression of MiB-1, a negative expression of hormonal receptors and SUV in the third interval. In conclusion, cOuch allowed to prove that SUV is correlated with many features obtained from ex vivo histopathological tests, suggesting SUV is a good diagnostic/prognostic biomarker to be obtained in vivo by 18F-FDG PET. Further studies will be devoted to apply cOuch methodology to the analysis of proteins differentially expressed in breast cancer tissues.

References

1. Castiglioni I, Gallivanone F, Grosso E, and Stefano A. (2011) cOuch, SIAE registration number 008239 - D007436.
2. Graham MM, Peterson LM, et al. (2000) Comparison of simplified quantitative analyses of FDG uptake Nucl. Med. Biol. 27(7): 647-655. doi: [10.1016/S0969-8051\(00\)00143-8](https://doi.org/10.1016/S0969-8051(00)00143-8).
3. Kim BS, Sung SH. (2012) Usefulness of 18F-FDG uptake with clinicopathologic and immunohistochemical prognostic factors in breast cancer. Ann Nucl Med. 26(2): 175-183. Doi: [10.1007/s12149-011-0556-1](https://doi.org/10.1007/s12149-011-0556-1).
4. Segal E., Sirlin CB, et al. (2007) Decoding global gene expression programs in liver cancer by non invasive imaging Nat Biotechnol. 25(6): 675-680.
5. Strauss LG, Koczan D, et al. (2008) Impact of angiogenesis related gene expression on the tracer kinetics of 18F-FDG in colorectal tumors J Nucl Med, 49(8): 1238-1244. Doi: [10.2967/jnumed.108.051599](https://doi.org/10.2967/jnumed.108.051599).

Prediction and 3D visualization of biological networks using cytological disease mapping

Björn Sommer¹, Vladimir Ivanisenko², Patrizio Arrigo³, Ralf Hofestädt¹✉

¹AG Bioinformatics and Medical Informatics, University Bielefeld, Bielefeld, Germany

²Institute of Genetics, Siberian Branch of the Russian Academy of Sciences Novosibirsk, Novosibirsk, Russian Federation

³CNR ISMAC, Genoa, Italy

Motivation and Objectives

The subcellular localization of biochemical networks is one of the new challenges of the post-genomic era. It may be used in future as an additional criteria to judge the probability of potential proteomic interaction partners, to choose the appropriate experimental constellations or to create virtual cell environments. Today we can use database integration tools and text mining for the prediction of biological networks. However, we can use the same tools for the prediction of the localization values of each component of biological networks. Having this information calculated we can realize a 3D (2D) visualization of biochemical networks based on a 3D (2D) virtual cell. Therefore we extended our CELLmicrocosmos (<http://Cm4.CELLmicrocosmos.org>) project.

Methods

The CELLmicrocosmos 4.2 PathwayIntegration (short: CmPI) is a module developed based on the CELLmicrocosmos 1.1 CellExplorer (CmCX). CmCX provides a virtual cell environment. This cell environment, containing different cell components, can be associated with protein-related networks using localization information parsed from DAWIS-M.D. as well as ANDCell. Figure 1 shows the workflow between CmPI and the two aforementioned systems. CmPI connects now directly to ANDCell, which is a data mining tool for pathway reconstruction, and DAWIS-M.D., which is a bio data warehouse. The input is a set of proteins or a protein-related network. Based on the data mining and bio data warehouse approach, CmPI will search for localization information for each network component. Based on this data and an already created virtual cell, CmPI will semi-automatically map this information into the three-dimensional cell environment. The user-friendly mapping process is supported by different visualization techniques.

DAWIS-M.D. which is based on the Bioinformatics Data Warehouse (BioDWH) toolkit, is a platform-independent data warehouse approach developed by the Bio-/Medical Informatics Department of Bielefeld University. It contains a number of metabolic-disease-related databases. For this work, only a subset of those databases was used for the localization of proteins: Brenda, Reactome, The Gene Ontology (GO), and UniProt.

The ANDCell (Associative Network Discovery) system contains a large variety of different databases. It is a commercial product of PBSOFT Ltd., a start-up company from the Institute of Cytology and Genetics (Novosibirsk). It was developed for the automatic extraction of facts and knowledge of interactions between proteins, genes, metabolites, microRNA, cellular components, molecular processes, and their association with diseases from the texts of scientific publications and databases. For this work, only those information were taken into account which were ex-

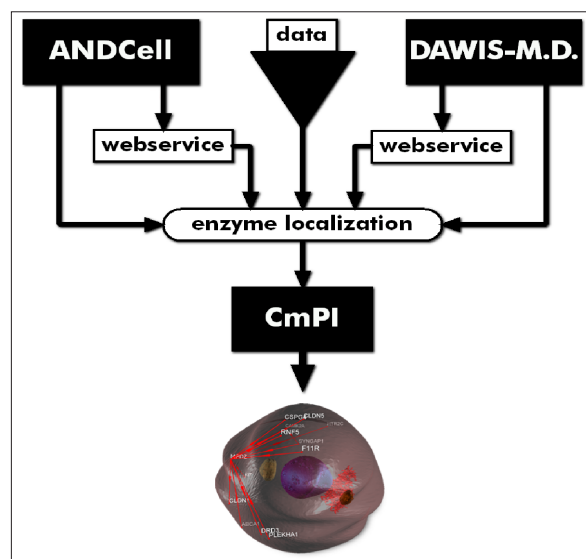


Figure 1: The Workflow between the databases and the CELLmicrocosmos 4.2 PathwayIntegration (CmPI)

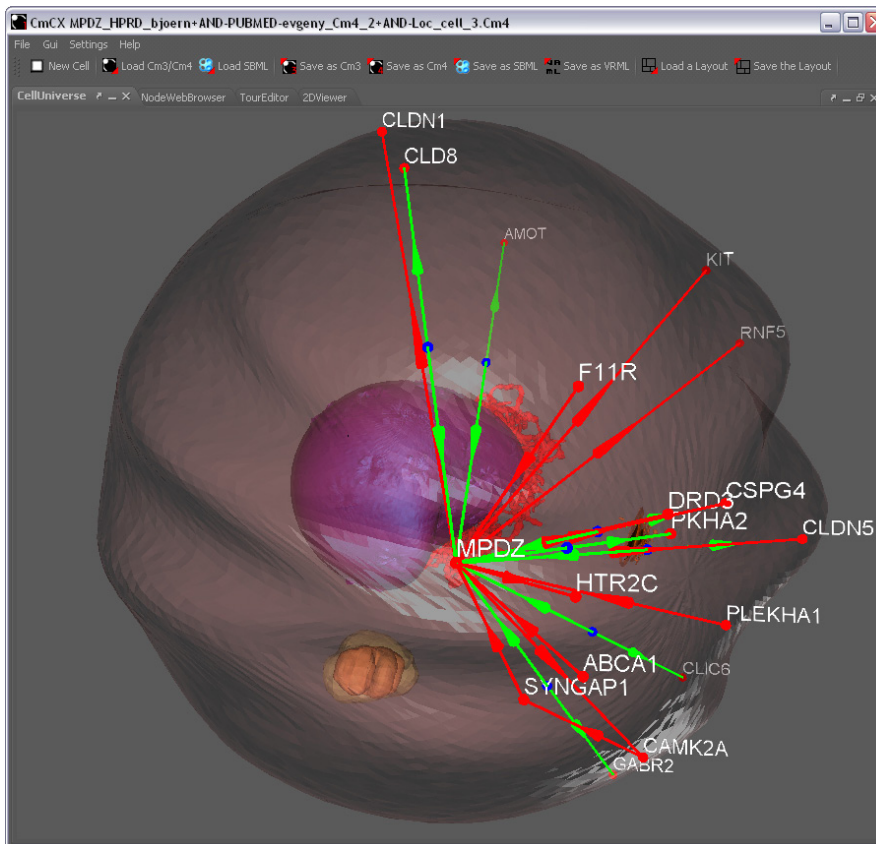


Figure 2: The MPDZ-protein-protein interaction network mapped onto a cell model

tracted from PubMed-abstracts which are freely available for CmPI.

Results and Discussion

These data integration techniques will underscore our understanding of the molecular mechanisms of diseases. In particular, this approach is important to investigate which cellular compartment is mainly involved in a specific pathological process. Based on this integrated knowledge, biological network prediction and reconstruction can be done for many investigated diseases even if the molecular knowledge of such disease is only rudimentary.

Figure 2 shows a MPDZ-related protein-protein interaction network associated with a three-dimensional cell model. This application case was discussed in Sommer et al. 2010. Experimental results suggested that MPDZ is involved in dilated cardiomyopathy. The interacting proteins were found to be associated with the tight junction complex of the cell membrane. Therefore it can

be suggested that this cell localization is affected by dilated cardiomyopathy and should be taken into account in future computational and experimental analyses.

Acknowledgements

This work was supported in part by: BMBF Internationale Zusammenarbeit in Bildung und Forschung mit Russland, RUS 08/005, EU-FP6, CARDIOWORKBENCH project, RAS Program "Living Nature: Current State & the Problems of Development", 14.740.11.0001 (PD, VI, NK); EU-FP7-260429-SYSPATHO.

References

1. B. Sommer, E. S. Tiys, B. Kormeier, K. Hippe, S. J. Janowski, T. V. Ivanisenko, A. O. Bragin, P. Arrigo, P. S. Demenkov, A. V. Kochetov, V. A. Ivanisenko, N. A. Kolchanov, R. Hofestädt. Visualization and Analysis of a Cardiovascular Disease-and MUPPI-related Biological Network combining Text Mining and Data Warehouse Approaches. *Journal of Integrative Bioinformatics*, 7(1):148, 2010.

A Grid-enabled web platform for integrated digital biobanking in paediatrics

Massimiliano Izzo¹✉, Andrea Schenone¹, Sara Barzaghi², Fabiola Blengio², Marco M Fato¹, Luigi Varesio²

¹Biolab, Department of Informatics Bioengineering Robotics and System Engineering, University of Genoa, Genoa, Italy

²Molecular Biology laboratory, Giannina Gaslini Institute, Genoa, Italy

Motivation and Objectives

A solid and integrated biobanking framework is an absolute requirement for high quality investigation in paediatric tumours. The overall goal of our activity is to design and develop a centralized Digital Biobank prototype able to integrate and interconnect an increasing number of local biobanks situated in various centres across Europe. As a first step, we are designing a web-based repository to store all tissue and genomic data from paediatric tumours collected by the G. Gaslini Children's Hospital, in Genoa. The repository satisfies flexibility and extensibility criteria, and is being deployed on a data Grid architecture (Bote-Lorenzo et al., 2004).

Methods

The repository is designed to contain data from all the tissue and blood samples obtained from infants and children affected by paediatric tumours, such as primary bone tumour and neuroblastoma. Many samples may be extracted from the same patient in a single visit or surgical operation; moreover from a single sample, nucleic acids (i.e. DNA and RNA) may be extracted for further analysis. These extractions could happen more than once, even at a distance of months or even years, if required.

In order to satisfy the strict requirements above and ensure the extensibility of the repository, we have adopted a process/event model, already used for designing data and image repositories in Neuroscience (Corradi et al., 2012). The process/event model is a multipurpose taxonomic schema composed by two main generic objects: processes and events. An event can be any 'atomic' operation that is performed on patients or samples, or any processing of data or everything else related to the repository administration and management. A process is defined as a group of sequential events or sub-processes related to an activity, allowing the creation of a sort of hierarchical structure. As an example, the storage of a DNA sample in a specified location

within a -80°C freezer and a post-processing step (such as differential expression, survival or correlation/anti-correlation analysis on microarray data) are single events, pertaining respectively to the more general 'Nucleic Acid Extraction' and 'Data Mining' processes.

Platform Architecture

The repository has a client-server architecture and it is composed by three main components, as shown in Figure 1:

- Repository portal
- Database
- Grid storage

The repository portal is designed to make the storage and the navigation of data and information easy, through a simple and transparent web interface. It is a Java Enterprise Edition web application based on several existing open source tools for the development of web applications. The basis of the portal consists in a framework that relies on an Apache Tomcat web application container. It incorporates a database interface layer built through MyBatis, a persistence framework that automates the mapping between SQL databases and objects in Java. To provide users with highly interactive interfaces, some components are designed using the Asynchronous Javascript and XML (AJAX) programming technique. Wherever possible information is exchanged in XML or JavaScript Object Notation (JSON) format. The web portal represents the main access point to all the functionalities available through the overall integration platform, and exposes both user and administrator interfaces.

The repository itself is based on a MySQL database. The database design is fundamental in order to make the repository highly flexible and easily extensible. The core of the database is formed by the two previously described entities: processes and events and their relationships to data and metadata. Existing processes and events are contained in two homonymous tables. Each element in the event table refers to an element in the data table. The information

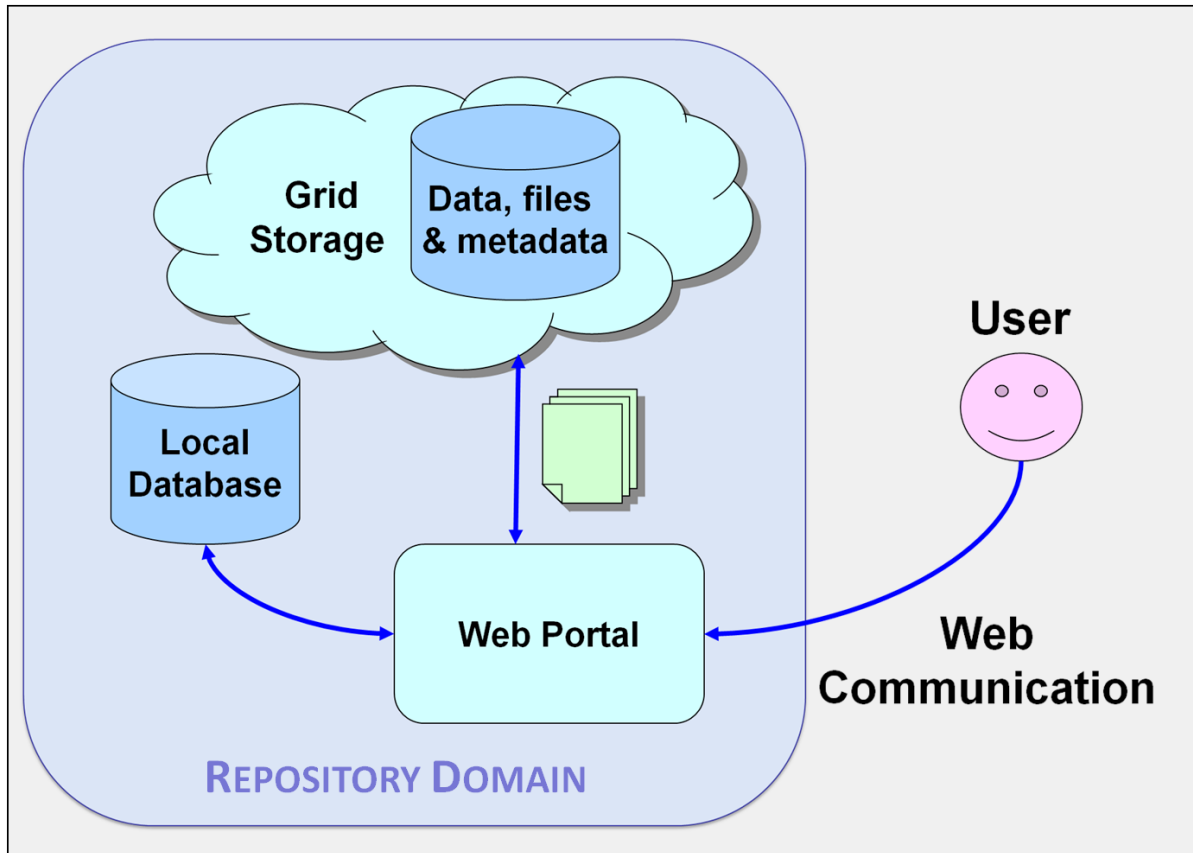


Figure 1. Repository overall architecture

inside the latter represents all the data inserted in the repository. These data can be associated with one or more files accordingly to their data type. The file table contains the logical path of all the stored files. The repository can be configured to store the metadata totally or partially within the database. In this latter case, the metadata are stored as XML descriptions inside the data table, to display the data in a rapid and dynamic way using XSL Transformations, and as records of specific metadata tables, to perform complex queries in an easier way. All data files are contained in the Grid storage, so the database doesn't really have to deal with hundreds of GB of data. Moreover, the number of operators should be quite small, thus making MySQL a reasonable choice as a database. The storage subsystem has been built around the iRODS data grid software (Rajasketar et al., 2010), chosen among others because it allows building a federated and distributed data storage system without the need of central components. Being able to deal with a huge amount of metadata, iRODS is widely used by

the research community, also for Next Generation Sequencing Projects (Chiang et al., 2011).

Careful attention has been given to security and privacy issues. All data are anonymised and cannot be linked in any way to patients' names, since the connection between clinical and personal data is done using unique identifiers managed exclusively by clinicians. Administrators are able to control users' access by creating groups and their association with pages and functions, define processes, events and all their relationships, define new data types and related metadata, associate them with the related events and manage available ontologies. Normal users, according to their assigned permissions, can insert new data, retrieve patients' information and view all the related data, download stored information, explore processes together with all the related events, data and metadata to have a global picture.

The integrated system we envision at a European level will take advantage of the data Grid features provided by iRODS. Each hospital or

biobank involved in the virtual community may have a local database and a dedicated separated iRODS system (called iRODS zone) where its own metadata and files can be saved. All the iRODS zones in the community will be federated. Federated iRODS zones are administered separately, but the users in the multiple zones, if given permission, will be able to access data stored in the other zones. If more hospital or research groups are working on the same project or using the same data structure, they may share a single iRODS zone and database. To provide access to the various local databases, federated database systems will be taken into account.

Results and Discussion

A first prototype of the repository is currently being tested at the Giannina Gaslini Institute, in Genoa. Information on over 1300 tissue samples, with their related DNA and RNA purified samples, have been stored together with administrative and clinical data from more than 700 patients. Three kinds of genomic analyses (i.e. event types) are currently provided, two for DNA samples - Comparative Genomic Hybridization (CGH) array and Multiplex Ligation-dependent Probe Amplification (MLPA) - and one for RNA - microarray analysis. For each analysis it is possible to store one or more files and user customized metadata. New data types can be configured via administrator interface, without additional programming, when new types of analyses or processing are required. The extensibility of our data

model with user-defined data types and metadata is a crucial aspect of our implementation.

As mentioned before, future developments will comprise the integration of our local biobank at the Gaslini Institute, with similar digital structures located across Europe. We are currently testing a distributed storage configuration, implementing data management policies expressed as rules that are interpreted by the iRODS Rule Engine.

Acknowledgements

Our research activity is performed in the framework of the 'European Network for Cancer Research in Children and Adolescents' (ENCCA) European project.

References

1. Bote-Lorenzo ML, Dimitriadis YA and Gomez-Sanchez E (2004) Grid characteristics and uses: a grid definition, Proceedings of the First European Across Grids Conference, ACG'03, Springer-Verlag, LNCS 2970, 291-298. doi:10.1007/978-3-540-24689-3_36
2. Chiang GT, Clapham P, Qi G, Sale K and Coates G (2011) Implementing a genomic data management system using iRODS in the Wellcome Trust Sanger Institute BMC Bioinformatics 2011, 12:361. doi:10.1186/1471-2105-12-361
3. Corradi L, Porro I, Schenone A, Momeni P, Ferrari , Nobili F, Ferrara M, Arnulfo G and Fato MM (2012) A repository based on a dynamically extensible data model supporting multidisciplinary research in neuroscience, BMC Medical Informatics and Decision Making (in press).
4. JSON (JavaScript Object Notation), [online], <http://www.json.org/>.
5. MyBatis, [online], <http://www.mybatis.org>.
6. Rajasketar A, Moore R, Hou C et al. (2010) iRODS Primer: Integrated Rule-Oriented Data Systems. Morgan & Claypool. doi:10.2200/S00233ED1V01Y200912ICR012
7. XSL Transformations [online], <http://www.w3.org/TR/xslt>.

MBLabDB: a social database for molecular biodiversity data

Flavio Licciulli¹, Domenico Catalano², Domenica D'Elia¹, Giorgio De Caro¹, Giorgio Grillo¹, Pietro Leo³, Giuseppina Mulè⁴, Paolo Pannarale³, Graziano Pappadà³, Francesco Rubino³, Antonella Susca⁴, Saverio Vicario¹, Gaetano Scioscia³

¹Institute for Biomedical Technologies (ITB), National Research Council (CNR), Bari, Italy

²Institute of Plant Genetics (IGV), National Research Council (CNR), Bari, Italy

³IBM GBS BAO Advanced Analytics Services, Bari, Italy

⁴Institute of Sciences of Food Production (ISPA), National Research Council (CNR), Bari, Italy

Motivation and Objectives

The biodiversity is nowadays one of the main scientific area of interest because of its importance for a sustainable development in many technological domains such as biotechnologies as well as for agriculture and human health. For instance, plant genetic resources are the basis of food security and consist of diversity of seeds and planting material of traditional varieties or modern cultivars and crop wild relatives. These resources are used as food, feed for domesticated animals and in recent years for the identification of new chemical compounds to be used in clinical therapeutic protocols.

Biodiversity research communities have to deal with data coming from many different domains (e.g., biology, geography, evolutionary studies, genomics, taxonomy, environmental sciences, etc.). Collecting and integrating data from so many disparate resources is not a trivial task, data are extremely scattered, heterogeneous in format and purpose, often protected in repositories of diverse research institutes.

With the advent of next generation technologies, molecular biodiversity research is producing large amounts of data that researchers use for complex comparative analyses exploiting information present both in public databases (like GenBank) and in their personal repositories. Improving the management of molecular data and their integration with related information present in the genetic resources databases such as morphologic, geographic and ecologic data will lead to new valuable biodiversity knowledge.

Driven by the widely diffused trend of the web of sharing information through aggregation of people with the same interests (social networks), and by the new type of database architecture defined as dynamic distributed federated database, here we present MBLabDB, a tool repre-

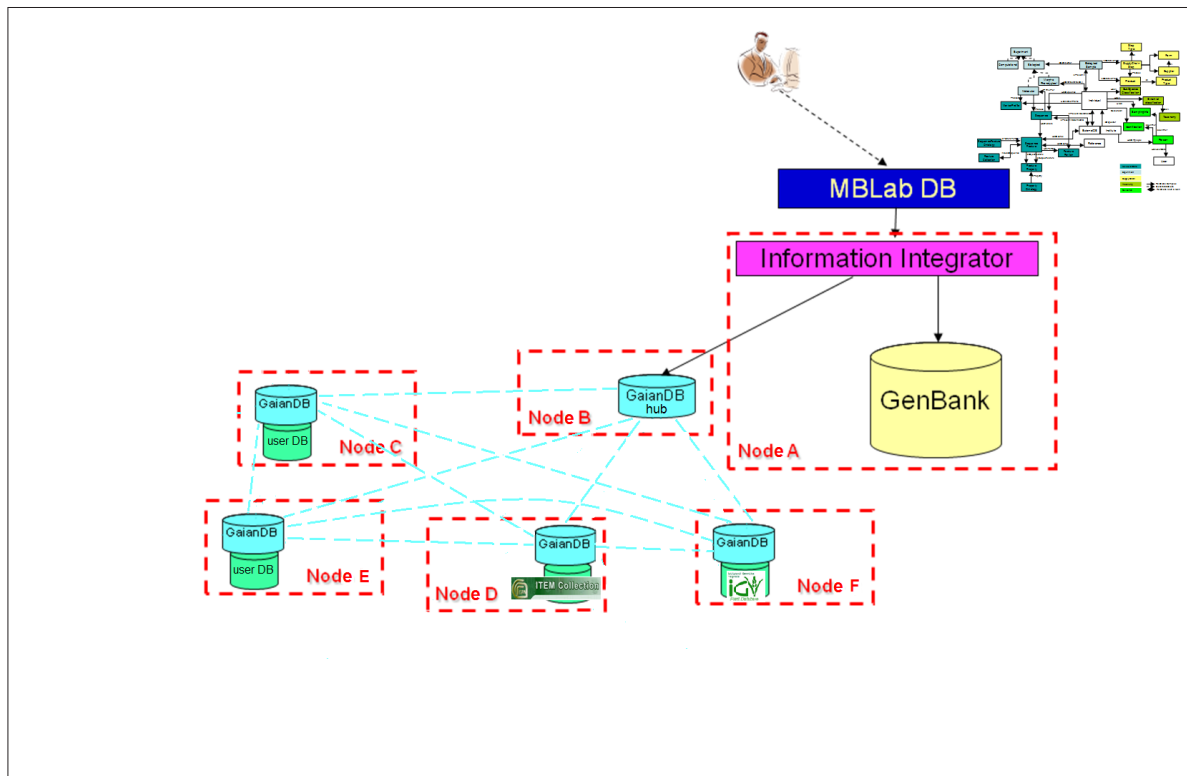
senting a new paradigm of data integration in the biodiversity domain.

Methods

MBLabDB uses a hybrid approach of data federation and data warehousing. The system architecture (Figure 1) is based on the integrated cooperation of several components: a robust Database Management System, managing the large volume of molecular data and information available in public resources such as GenBank; a set of federated databases implemented with GaiandB (Bent G. et al., 2008) tool, managing remote specialized biodiversity databases; the IBM Information Integrator, implementing the database conceptual schema and integrating all federated databases with public molecular data using a data warehouse approach.

The conceptual schema of MBLabDB named MolecularBiodiversity Database Schema (Pannarale et al., 2012), is tailored to biodiversity data collection, integration and analysis. It is modeled on six main sections: Individual, MolecularData, Experiment, Collection, Supply chain and Taxonomy. The MolecularData section is structured following a Chado-like model (Mungall CJ et al., 2007), using Sequence Ontology (Eilbeck K et al., 2005) entities and relations. Similarly the Taxonomy section has been designed in order to incorporate and integrate more than one taxonomy, because of different reference taxonomies that could be related to a taxonomic kingdom.

The federated databases have been implemented by GaiandB (Bent G. et al., 2008), a Dynamic Distributed Federated Database of sources whose growth is regulated by biologically inspired principles and graph theoretic methods. The idea is to create a network of database nodes, each containing specialised collections of biodiversity data, and to expose their content



by means of a GaianDB data server. Information coming from the network nodes are collected by a GaianDB hub and are integrated with public data by means of the Information Integrator server. Two steps are needed to add a new GaianDB node: the installation of a GaianDB server instance and the writing of a wrapper for the mapping of the local schema with the general MBLabDB schema.

An efficient and reliable ETL (Extraction, Transformation and Load) module, implemented with CLIPS Rule Based Programming Language (Pannarale et al., 2012), has been used to integrate GenBank data in MBLabDB. The ETL procedure extracts information from the GenBank entries and fits them into the MBLabDB schema.

The MBLabDB graphical user interface (GUI) has been developed as a Java platform web application. In the GUI the public-private data integration is highlighted through the implementation of taxonomic and ontology based queries.

Results and Discussion

Currently, MBLabDB integrates 4,360,218 entries from the GenBank database and two biodiversity data collections: the ITEM Collection (<http://www.ispa.cnr.it/Collection>), located at the ISPA-CNR

server (containing 9,181 specimen and 3,584 sequences), and the IGV Germoplasm Database (<http://www.igv.cnr.it>), located at the IGV-CNR server (containing 11,113 accessions). Furthermore the NCBI Taxonomy (www.ncbi.nlm.nih.gov/Taxonomy) and the Catalogue of Life (<http://www.catalogue-of-life.org/>) taxonomic classifications have been included in the Taxonomy section.

Two search and retrieval modalities are available in MBLabDB, an advanced query mode, where search criteria and results can be combined using an incremental composition of “querying & filtering”, and an ontology based retrieval that queries data using the biological concepts expressed by the Sequence Ontology.

Therefore, MBLabDB combines public molecular data with biodiversity data contained in genetic resource collections, that are typical of the biodiversity domain. By way of example, using MBLabDB a researcher can extract datasets of sequences related to specimen of his own interest using biodiversity criteria such as species/varieties, geolocation, morphology and passport data.

Using the MBLabDB paradigm of data integration, database hosting, management and information sharing strategy of specialised resources

are left to the research group owner of the data collection. So the biodiversity research groups can contribute to the information network by sharing their data sources with a reasonable effort.

In this network, named Social Database for Molecular Biodiversity Data, information remains scattered, but knowledge are shared.

Acknowledgements

This work was supported by DM19410 - Bioinformatics Molecular Biodiversity Laboratory - MBLab (www.mblabproject.it).

References

1. Bent G. et al. (2008) A dynamic distributed federated database. Second Annual Conference of ITA, Imperial College, London
2. Eilbeck K et al. (2005) The Sequence Ontology: A tool for the unification of genome annotations. *Genome Biology* 6:R44
3. Mungall CJ et al. (2007) A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics* 23: i337-i346
4. Pannarale P et al. (2012) GIDL: a rule based expert system for GenBank Intelligent Data Loading into the Molecular Biodiversity database. *BMC Bioinformatics* 13 Suppl 4:S4

A multivariate analysis of protein microarrays for signature selection profiles

Saveria Mazzara^{#✉}, Antonella Sinisi, Angela Cardaci, Sergio Abrignani, Mauro Bombaci^{#✉}

Istituto Nazionale Genetica Molecolare (INGM), Milan, Italy

These authors have contributed equally to this work

Motivations and Objectives

In recent years, protein microarrays have become one of the most invaluable research tools in the field of large-scale and high-throughput biology, and their use in basic research, diagnostics and drug discovery has emerged as a great promise of medicine. An interesting application of this technology is the identification of a serodiagnostic antigens ensemble whose expression profiles can effectively unveil discriminant patterns providing the classification of healthy and disease samples.

Nowadays, the analysis of protein microarray data for extracting biologically interpretable results is still an extremely complex process, and there is an increasing need for fully automated data mining approaches.

In the present study a Partial Least Squares Discriminant Analysis (PLS-DA) has been applied to protein microarrays aimed to reveal discriminative patterns between different clinical conditions. The method was evaluated to data generated from protein microarrays, including 1626 human recombinant proteins, probed with sera of patients with autoimmune liver diseases. Each array was depicted as a set of quantitative descriptors and analyzed by PLS-DA method in an attempt to classify samples according to their intrinsic protein expression profile. Moreover, the assessed model was able to extract antigens of interest representative of a different protein profile in Autoimmune Hepatitis (AIH) patients compared to Healthy donors (HD) (Zingaretti et al., 2012).

Here, the application of multivariate statistical techniques to protein microarray data represents an effective tool to identify informative protein profiles as of fully automatic strategy.

This kind of approach could lead to a more rapid and accurate development of diagnostic tests, providing useful factors able to discriminate different autoimmune diseases.

Methods

We proposed an innovative bioinformatic workflow, based on multivariate data analysis, for

identifying discriminative patterns between different clinical conditions. A schematic view of flow chart is shown in Figure 1. In the first step (Figure 1, panel A), protein arrays were developed to screen serum samples of patients affected by Autoimmune Hepatitis (AIH) and healthy controls (HD); characteristic of patients and protein platform generation were described in recent publication (Zingaretti et al., 2012). Briefly, fluorescence signals were detected by using a ScanArray Gx PLUS (PerkinElmer, Bridgeport Avenue Shelton, USA) and scanned images were imported in a house developed software for the successive image analysis. Normalization to the spotted human IgG curve was performed (Bombaci et al., 2009). Subsequently, the data were analyzed by partial least squares discriminant analysis (PLS-DA) (Wold et al., 2001; Eriksson et al., 2006) (Figure 1, panel B) with the aim of identifying the best candidates for the development of new application in clinical research (Figure 1, panel C). In recent years, projection methods are being successfully applied to biological data such as DNA microarrays and proteomic data but the combination of PLS-DA with the protein arrays represents a new and interesting approach for investigation of this type of proteomics data. The method is particularly suitable for analysis of data with numerous variables and is able to integrate information about the response matrix, Y, into the descriptor matrix, X (the antigens). PLS method is based on finding the latent variables that maximize the covariance between X and Y. The importance of each variable in the loadings of PLS-DA is given by the variable influence on projection (VIP) parameter. The VIP score reflects the influence of antigens on the classification, and predictors with score larger than one are considered relevant for explaining the differences in the two groups (Eriksson et al., 2006). The validation of the PLS-DA model was checked using cross-validation and response permutation testing. Cross-validation assesses the predictive power of the model by Q²Y while the response permutation test assess the statistical significance of the estimated predictive

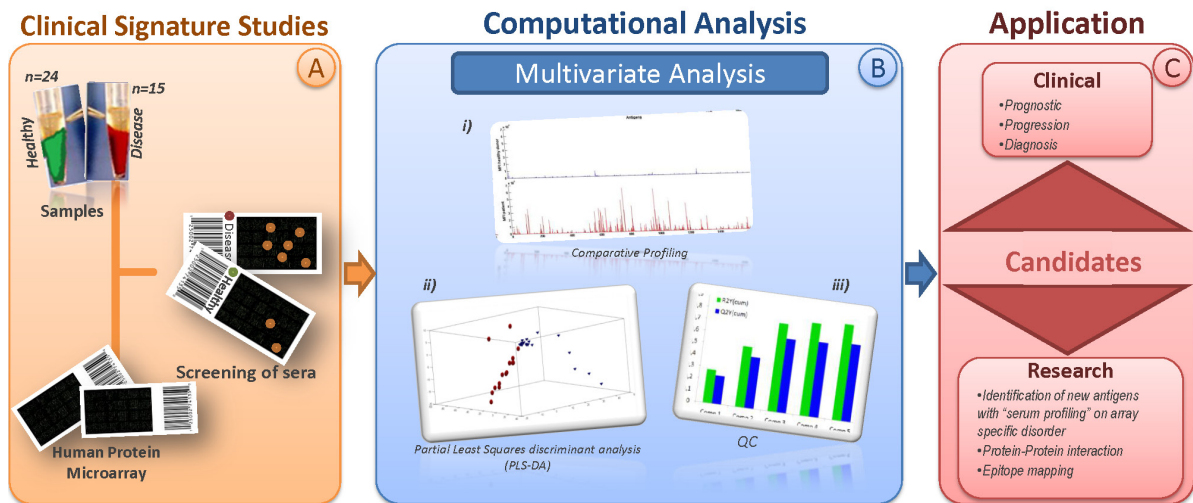


Figure 1: Schematic representation of strategy illustrating large scale serum autoantigen analysis (A) Screening of Healthy vs Patient sera by in house developed Protein microarray. (B) Multivariate Analysis: result interpretation and statistical analysis: (i) representative MFI distribution of AIH patients (bottom panel) compared with HD (top panel) subjects, (ii) PLS-DA projection of AIH and HD samples according t1,t2 and t3 coordinates. This projection was done to identify protein profiling that distinguishes between AIH (red spheres) and HD (blue cones), (iii) plot of R2Y (explained variation) and Q2Y (predicted variation); it shows how the considered parameters change as a function of increasing model complexity. According to the cross-validation, nine components resulted significant in order to explain the relationship between the descriptor matrix and the class response; nevertheless, three components were considered to allow score plotting. (C) Investigation and further applications on the identified markers.

power and test the model for overfitting due to the chance correlation. In this test, only the class labels is randomly reordered (50 times). A model is fitted to the new Y-data and new estimates of R2Y and Q2Y values are calculated. The distribution of the R2Y and Q2Y, based on random data, are useful for appraising the validity of the model (Eriksson et al., 2006).

Results and Discussion

In order to identify a set of protein signatures linked to autoimmune liver disease, we have processed protein microarray data generated from sera of 15 AIH patients and 24 HD subjects, as shown in Figure 1, panel A. AIH sera displayed a higher reactivity toward autoantigens than HD sera as documented by the intensity of recognition signals (MFI). To detect differences between the two clinical conditions a multivariate statistical analysis was performed. As a first step, an unsupervised approach by means of PCA was applied to the full data set. This preliminary exploration by PCA was done in order to screen for outliers and to survey possible groupings, useful for efficiently directing further modeling efforts with more innovative approaches such as

the PLS-DA. On the basis of the PCA score plot, a rough separation was observed owing to misclassification of one sample; thus, this sample was removed from further analyses due to its ambiguous behavior. The PLS-DA modeling has been based on the reduced data set of 38 samples described through 1296 features (X matrix). We created a dummy matrix of two Y-variables expressing diagnosis of the sera samples. Data were standardized to have mean 0 and standard deviation 1. The number of significant components was determined using cross-validation; this yielded nine components with an R2Y of 0.91 and a cross-validated R2 (Q2) of 0.75. However, a three component model was generated to enable the construction of the three dimensional score plot (Eriksson et al., 2006). There is clear discrimination between the two groups according to their clinical conditions. The model also gives the possibility to obtain a quantitative measure of the discriminating power of each autoantigens by means of VIP. X-variables characterized by VIP values larger than 1 have major importance for modeling the responses. After closer examination, autoantigens were, then, selected according to (i) VIP scores >1.0 and (ii) the recognition

frequency; self proteins were regarded as potential autoantigens if they were recognized by a delta difference recognition of 25% between AIH and HD population. In this way, a final list of 27 autoantigens was generated, that allowed good discrimination of the two populations of sera. At present, the study may be deepened at the biological level by further validation of these dominant features using proteomic analysis technique. Furthermore, we applied the response permutation testing to provide an estimate of the significance of a Q2Y value, we have permuted the response randomly 50 times and computed the new model with the original X-data matrix and reordered Y-data. For each derived model, both R2Y and Q2Y values were calculated and then compared with the estimates of the R2Y and Q2Y of the real model. On the basis of the validation plot, the Q2Y distribution is sign of high predictive validity of the original model indeed it is impossible to obtain a model with the same predictive value by chance.

In conclusion, we presented a multivariate approach as an effective alternative to classical univariate tools for the analysis of proteomics data for signature selection in autoimmune liver diseases. Combining multivariate modeling with protein microarray proves to be a successful tool for the discrimination of the different classes of samples and for the identification of the autoantigens responsible for class separation by means of VIP. This method could be applied for a fast screening of human protein microar-

rays to discriminate different clinical conditions representing a useful complementary analysis in the routine of a proteomic laboratory. However, further studies are necessary in order to extend the approach here described to different data set for verifying the chance to extrapolate and generalize classification rules.

Acknowledgements

This work was primarily supported by a Grant from Fondazione Cariplo and a grant FIRB from the Italian Ministry of University and Research (MIUR). We'd like to thank Fondazione IRCCS Ospedale Maggiore Policlinico, Milan; Policlinico Sant'Orsola, Bologna; Azienda Ospedaliera Universitaria Pisana, Pisa, for kindly providing the human sera used for the screening

References

1. Bombaci M, Grifantini R, Mora M, Reguzzi V, Petracca R et al. (2009) Protein array profiling of tic patient sera reveals a broad range and enhanced immune response against Group A Streptococcus antigens. *PLoS One* 4, e6332. doi:10.1371/journal.pone.0006332
2. Eriksson L, Johansson E, Kettaneh-Wold N, Trygg J, Wikström C et al. (2006) Multi- and megavariate data analysis. Basic Principles and Applications. Umetrics AB
3. Jain AK, Duin RPW, Mao J. (2000) Statistical pattern recognition: a review. *IEEE Trans Pattern Analysis Machine Intelligence* 22, 4-37. doi: 10.1109/34.824819
4. Wold S, Sjöström M, Eriksson L (2001) PLS-regression: a basic tool of chemometrics. *Chem Intell Lab System* 58, 109-130. doi:10.1016/S0169-7439(01)00155-1
5. Zingaretti C, Arigò M, Cardaci A, Moro M, Marabita F et al. (2012), *Mol Cell Proteomics*, Sep 20. [Epub ahead of print].

Extracting correspondences between terminologies for an easier access to biomedical information

Adila Merabti[✉], Lina F Soualmia, Stéfan J Darmoni

TIBS LITIS laboratory EA 4108, Rouen University Hospital, France

Motivation and Objectives

Biomedical terminologies play important roles in clinical data capture, annotation, reporting, information integration, indexing and retrieval. More particularly, genomic terminologies and ontologies are very useful for indexing genomic information. Several sources of information and terminologies have already been developed. For instance, the Gene Ontology (GO, <http://www.geneontology.org/>, last accessed on July 17, 2012), which is a controlled vocabulary widely used for the annotation of gene products; the Human Phenotype Ontology (HPO, <http://www.human-phenotype-ontology.org/>, last accessed on July 17, 2012) in which terms describe phenotypic abnormalities encountered in human disease, such as "atrial septal defect"; and ORPHANET, <http://www.orpha.net/consor/www/cgi-bin/index.php?lng=FR>, last accessed on July 17, 2012) the portal for rare diseases and orphan drugs. These knowledge sources have mostly different formats and purposes. For example, ORPHANET is a rare disease database whereas HPO is an ontology which supports the description of phenotypic information. Faced with this reality and the need to allow cooperation between various health actors and their related health information systems, it appeared necessary to link these terminologies by developing a semantic repository to integrate them. The most known repository is the Unified Medical Language System (UMLS) (Lindberg et al., 1993). Several works were based on the UMLS to align terminologies in French (Merabti et al., 2012) and in English (Bodenreider et al., 1998; Milicic Brandt et al., 2011; Mougjin et al., 2011). However, HPO and ORPHANET are not yet included in the UMLS. Thus, another solution is to find correspondences between these terminologies in French and in English using automatic methods. In (Merabti et al., 2012) we have proposed a lexical method to map biomedical terminologies either included or not into the UMLS. Nevertheless, these methods remain very dependent on the terminolo-

gies languages since they used NLP tools such as stemming or normalization. We propose in this study a string-based method to find correspondences between a subset of terminologies for an easier access to biomedical information. It is based on the combination of several string metrics and it is neither based on the UMLS, nor language dependent. Mixed with lexical or conceptual approaches developed in previous studies (Merabti et al., 2012), it could improve the number of correspondences between terminologies with a high precision. Semantic methods are also an envisaged issue to complete this study.

Methods

To map biomedical terminologies, we used string matching methods where concept names, terms and their labels are considered as sequences of characters. A string distance is determined to compute a similarity degree. Some of these methods can skip the order of characters. In this paper, the union of three metrics was used (i) Dice (Dice, 1945), (ii) Levenshtein (Levenshtein, 1965) and (iii) Stoilos (Stoilos et al., 2005).

The Dice's coefficient calculates the ratio between the number of bigrams of characters in common to both the strings x and y and the total number of bigrams for two strings defined by the following equation where $nb_big(x)$ is the number of bigrams of x :

$$Dice(x, y) = \frac{2 \times \text{number of common bigrams}}{nb_big(x) + nb_big(y)}$$

The Levenshtein distance between two strings x and y is defined as the minimum number of elementary operations that is required to pass from a string x to a string y . There are three possible transactions: replacing a character with another, deleting a character and adding a character. This measure takes its values in the interval $[0, \infty[$. The Normalized Levenshtein (Yujian and Bo, 2007) (LevNorm) in the range $[0, 1]$ is obtained by dividing the distance of Levenshtein $Lev(x, y)$ by the size of the longest string and it is defined by:

$$\text{LevNorm}(x, y) = 1 - \frac{\text{Lev}(x, y)}{\text{Max}(|x|, |y|)}$$

LevNorm(x,y) is element of [0,1] as Lev(x,y) < Max(|x|,|y|). |x| is the length of the string x.

The Stoilos distance has been specifically developed for strings that are labels of concepts in ontologies. It is based on the idea that the similarity between two entities is related to their commonalities as well as their differences. Thus, the similarity should be a function of both these features. It is defined by:

$$\text{Sim}(x, y) = \text{Comm}(x, y) - \text{Diff}(x, y) + \text{winkler}(x, y)$$

Where Comm(x,y) stands for the commonality between the strings x and y, Diff(x,y) for the difference between x and y, and Winkler(x,y) for the improvement of the result using the method introduced by Winkler in (Winkler, 1999). The function of commonality is determined by the substring function. The biggest common substring between two strings (MaxComSubString) is computed. This process is further extended by removing the common substring and by searching again for the next biggest substring until none can be identified. The function of commonality is given by the equation:

$$\text{Comm}(x, y) = \frac{2 \times \sum_i |\text{Max Com Sub String}_i|}{|x| + |y|}$$

The function of Difference is defined in the following equation where p is element of [0, ∞ [(usually p= 0.6), |ux| and |uy| represent the length of

the unmatched substring from the strings x and y scaled respectively by their length:

$$\text{Diff}(x, y) = \frac{|u_x| \times |u_y|}{p + (1-p) \times (|u_x| + |u_y| - |u_x| \times |u_y|)}$$

The Winkler parameter Winkler(x,y) is defined by the equation:

$$\text{Winkler}(x, y) = L \times P \times (1 - \text{Comm}(x, y))$$

where L is the length of common prefix between the strings x and y at the start of the string up to a maximum of 4 characters and P is a constant scaling factor for how much the score is adjusted upwards for having common prefixes. The standard value for this constant in Winkler's work is P=0.1. To evaluate the correspondences between the terminologies found using the proposed method we have calculated the precision on a sample set evaluated manually and defined as:

$$\text{Precision} = \frac{\{\{\text{Correct correspondences}\}\}}{\{\{\text{total correspondences}\}\}}$$

Results and Discussion

In this study we presented a combination of tree string matching methods to align several biomedical terminologies. The results showed that combining these methods on general terminologies such as MeSH and SNOMED provided more correspondences than only one method and with good results (with a precision > 99%). Aligning genomic terminologies provided also good results with high precision. However, we evaluated

	Dice	Levenshtein	Stoilos	Combination
MeSH with SNOMED INT	NB_align=75,176 P=99.82 % CI95%=[99.79-99.85]	NB_align=64,657 P=99.80% CI95%=[99.77-99.83]	NB_align=133,419 P=99.75% CI95%=[99.72-99.78]	NB_align=156,877 P=99.78% CI95%=[99.76-99.80]
HPO with GO (EN)	NB_align=161	NB_align=49	NB_align=207	NB_align=291
HPO with GO (FR)	NB_align=10 P=75.00%	NB_align=7 P=83.00%	NB_align=9 P=80.00%	NB_align=11 P=72.22%
HPO with ORPHANET (EN)	NB_align=2,593	NB_align=1,506	NB_align=3,718	NB_align=4,237
HPO with ORPHANET (FR)	NB_align=3,506 P=97.18% CI95%=[96.63-97.73]	NB_align=2,246 P=94.14% CI95%=[93.17-95.11]	NB_align=5,405 P=94.87% CI95%=[94.28-95.46]	NB_align=6,040 P=96.49% CI95%=[96.03-96.95]

Table 1: Total number of correspondences (NB_align) with a threshold of 0.8 and their associated precision (P%) according to each method. Only the correspondences in French were evaluated. We evaluated a sample of 100 correspondences.

here only “exact” correspondences and rated them as “correct” or “not correct”. Indeed, correspondences such as “broader–narrower” or “sibling” relations between terms were not considered. For example, when a correspondence is founded between two terms which one string is included in another one in most cases it is more general than the second, and a “broader–narrower” correspondence could exist (for example, correspondence between “insuffisance surrenale” term (Adrenal insufficiency) and all the terms such as “insuffisance surrenale aigue” (Acute Adrenal insufficiency), “insuffisance surrenale primaire” (Primary adrenal insufficiency)). These preliminary good results encouraged us to apply the combination of these string matching methods on other health terminologies. The correspondences found between two terminologies in their French version may be projected on their versions in other languages. As perspectives of this study, these methods will be completed with normalization techniques and the validation of the correspondences, manual here, will be done according to the UMLS semantic types for the terminologies included in it such as in (Mougin et al, 2011).

References

1. Bodenreider O, Nelson SJ, et al. (1998) Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. In Proc. AMIA Symp. 1998, pp.815–819.
2. Dice LR (1945). Measures of the amount of ecologic association between species. *Ecology* 26, pp.297–302.
3. Levenshtein VI (1965) Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Dokl.*10, pp.707–10.
4. Lindberg DA, Humphreys BL, et al. (1993) The Unified Medical Language System, *Methods Inf Med* 32(4): 281–291.
5. Merabti T, Soualmia LF, et al. (2012) Aligning Biomedical Terminologies in French: Towards Semantic Interoperability in Medical Applications. In Book *Medical informatics*, InTech, pp.41–68.
6. Millicic Brandt M, Rath A, et al. (2011) Mapping Orphanet terminology to UMLS. In Proc. AIME, LNAI 6747, pp.194–203.
7. Mougin F, Dupuch M, et al. (2011) Improving the mapping between MedDRA and SNOMED CT. In Proc. AIME. LNAI 6747, pp. 220-224.
8. Stoilos G, Stamou G, et al. (2005) A string Metric for Ontology Alignment. In Proc. ISWC, pp.624–37.
9. Winkler W (1999) The state record linkage and current research problems. Technical report: Statistics of Income Division, Internal Revenue Service Publication.
10. Yujian L, Bo L (2007) A normalized Levenshtein distance metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1091–1095.

Detection of allele-specific gene expression on Next Generation Sequencing data

Vladan Mijatovic¹✉, Luciano Xumerle¹, Alberto Ferrarini², Ilaria Iacobucci³, Chiara Pighi⁴, Antonio Mori¹, Chiara Zusi¹, Paola Prandini¹, Elisabetta Trabetti¹, Massimo Delledonne¹, Giovanni Martinelli³, Albert Zamò⁴, Pier F Pignatti¹, Giovanni Malerba¹

¹Department of Life and Reproductions Sciences, University of Verona, Verona, Italy

²Department of Biotechnology, University of Verona, Verona, Italy

³Department of Hematology and Oncological Sciences "L. and A. Seràgnoli", University of Bologna, Bologna, Italy

⁴Department of Pathology and Diagnostics, University of Verona, Verona

Motivation and Objectives

Many genetic variants mediate changes in gene expression. Some studies observed that variation of gene expression between alleles is common, and this variation may contribute to human variability of several traits (Lo et al, 2003, Main et al.,2009).

Next-generation sequencing (NGS) provides robust, comparable and highly informative expression profiling data (Shendure et al., 2008), and is rapidly replacing microarray methods in gene-expression (GE) studies (Wold et al., 2008, Wang et al., 2009). In contrast to microarrays, NGS expression profiling is based on sequencing and counting fragments of mRNA (Feng et al., 2010, van Iterson et al, 2009). The goal of our study is to develop a statistical framework aiming to measure and detect allele-specific GE differences from global GE experiments conducted using NGS technology.

Methods

We developed a statistical method for identification of allele-specific differential expression (ASDE). The method is based on the likelihood estimation of the observed data depending on the parameter θ . Assuming that each polymorphism biallelic locus presents the alleles A1 and A2 we define θ as $A1/(A1+A2)$. Therefore θ ranges from 0 to 1, and the expected value in the case of a fair expression of the two alleles is $\theta = 0.5$ whilst values departing from 0.5 indicate that one allele is more expressed than the other. The likelihood function (L) is based on the binomial model and depends on the θ value as follows: $L(\theta) = k * [(\theta)A1 * (1-\theta)A2]$ where k is constant of proportionality ($k > 0$). The hypothesis of $\theta \neq 0.5$ (i.e. ASDE) can be easily compared with the null hypothesis of $\theta = 0.5$ (i.e. the two alleles, A1 and A2, present the same expression value) through

a Likelihood Ratio Test (LRT): $LRT = -2 * \ln(L(\theta=0.5) / L(\text{tested-}\theta))$. The LRT from different samples can be summed up to a combined-LRT value that expresses the overall support of the model tested at θ value that maximizes the likelihood function. Therefore the LRT can be used to test different ASDE models on the available data. The arbitrary threshold of $LRT > 600$ was used to detect ASDE loci.

NGS expression data have been obtained using a pipeline (quality control, alignment, SNP detection and read count) of computer programs that has been developed in our laboratory.

The following software have been used: bowtie (Langmead B et al.,2012) and samtools (Li et al., 2009). The reference sequence of the human genome GRCh37 was used.

Only heterozygous single nucleotide polymorphisms (SNPs) were selected for the following analysis, defined as SNPs with a coverage of at least 10 reads for each allele.

LRT was then applied to the data results of each sample.

Currently we performed the analysis on a total of 7 mantle cell lymphoma libraries (MCL)(Pighi et al., 2011). One-hundred (100) base-pair (bp) sequence paired-end reads were generated for each sample using an Illumina sequencer Hi-seq-1000. The average coverage was 10.4.

Results and Discussion

On average, the MCL cohort (7 samples) contained nearly 70,000 heterozygous sites. Preliminary results suggested 501 ASDE loci of which 470 showed a $0 < \theta < 0.05$. We did not observed ASDE loci for $0.20 < \theta < 0.80$.

We plan to estimate a reliable threshold of LRT across the entire transcriptome of several samples by simulation studies. We shall also study in more detail if the suggested ASDE loci showing a $\theta < 0.05$ (or a $\theta > 0.95$) are true ASDE loci or NGS artifacts.

The method will be extended to compare the Θ values (ASDE status) among groups of individuals (i.e. cases versus controls). The molecular analysis will be extended to additional samples including leukemia (Iacobucci et al., 2012), heart and skeletal muscle cells as well as lymphoblastoid cell libraries of individuals suffering from Autism Spectrum Disorders (ASD)(Prandini et al., 2012). We developed a method able to detect ASDE loci from global gene expression NSG data. One of the features of this method is that it can easily measure the degree of ASDE through the parameter Θ . The method may also be applied to cancer research because an apparent ASDE locus might underlie the expression of a reduced amount of mutated cancer cells.

In conclusion we are developing a method for integration of information on allele variation with gene expression. This could increase our knowledge of hereditary factors involved in regulatory systems of gene expression.

References

1. Feng L, Liu H, Liu Y, et al. (2010) Power of deep sequencing and agilent microarray for gene expression profiling study. *Mol Biotechnol* 45:101. doi: [10.1007/s12033-010-9249-6](https://doi.org/10.1007/s12033-010-9249-6).
2. Iacobucci I, Ferrarini A, Sazzini M, et al. (2012) Application of the whole-transcriptome shotgun sequencing approach to the study of Philadelphia-positive acute lymphoblastic leukemia. *Blood Cancer J.* 2(3): e61. doi: [10.1038/bcj.2012.6](https://doi.org/10.1038/bcj.2012.6).
3. Lo HS, Wang Z, Hu Y, et al. (2003) Allelic variation in gene expression is common in the human genome. *Genome Res.*13(8):1855. doi: [10.1101/gr.1006603](https://doi.org/10.1101/gr.1006603).
4. Main BJ, Bickel RD, McIntyre LM, et al. (2009) Allele-specific expression assays using Solexa. *BMC Genomics.* 10:422. doi: [10.1186/1471-2164-10-422](https://doi.org/10.1186/1471-2164-10-422).
5. Pighi C, Gu TL, Dalai I, et al. (2011) Phospho-proteomic analysis of mantle cell lymphoma cells suggests a pro-survival role of B-cell receptor signaling. *Cell Oncol (Dordr).* 34(2):141. doi: [10.1007/s13402-011-0019-7](https://doi.org/10.1007/s13402-011-0019-7).
6. Prandini P, Pasquali A, Malerba G, et al. (2012) The association of rs4307059 and rs35678 markers with autism spectrum disorders is replicated in Italian families. *Psychiatr Genet.* 22(4):177
7. Shendure J, Ji H (2008) Next-generation DNA sequencing. *Nat Biotechnol* 26:1135. doi: [10.1038/nbt1486](https://doi.org/10.1038/nbt1486).
8. van Iterson M, 't Hoen PA, Pedotti P, et al. (2009) Relative power and sample size analysis on gene expression profiling data. *BMC Genomics* 10:439. doi: [10.1186/1471-2164-10-439](https://doi.org/10.1186/1471-2164-10-439).
9. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature Rev. Genet.* 10:57. doi: [10.1038/nrg2484](https://doi.org/10.1038/nrg2484).
10. Wold B, Myers RM (2008) Sequence census methods for functional genomics. *Nature Methods* 5:19. doi: [10.1038/nmeth1157](https://doi.org/10.1038/nmeth1157).

Network-based analysis of stem cells differentiation

Francesca Mulas¹✉, Lan Zagar², Blaz Zupan¹, Riccardo Bellazzi^{1,3}

¹Centre for Tissue Engineering, University of Pavia, Pavia, Italy

²Faculty of Computer Science, University of Ljubljana, Ljubljana, Slovenia

³Dipartimento di Ingegneria Industriale e dell'Informazione, Università di Pavia, Pavia, Italy

Motivation and Objectives

Understanding the real developmental stage of reprogrammed stem cells is still a demanding task for researchers in regenerative medicine. In fact, developmental biology is in need of methods that would accurately predict the developmental stage reached by cells cultured in non standard conditions, such as the induced Pluripotent Stem Cells (iPSCs). Bioinformatics approaches would be extremely useful for assessing the pluripotency status of the cells, and thus, their potential use in the clinics for repairing malfunctioning tissues and organs.

To this aim, several works have recently demonstrated the utility of applying dimensionality reduction techniques to genome wide expression data (Aiba et al, 2009). As we have shown (Zagar et al, 2011), these methods can be successfully used to predict the developmental stage of cells by mapping their transcriptional profile to a one-dimensional ruler, that we named differentiation scale. The proposed approach was also useful for identifying reduced subsets of genes that drive each developmental stage (Mulas et al, 2012).

A crucial issue in this field is the integration of the findings extracted from the data with the available knowledge coming from biological databases. For instance, genes that are surrounded by a high number of selected genes in the protein-protein interaction networks should be included in the analysis (Nitsch et al, 2010). In this work, we developed a network-based pipeline for analyzing temporal gene expression data coming from embryonic stem cells differentiation. The results highlighted the transcriptional changes occurring during development and allowed identifying known markers as well as novel gene candidates potentially involved in the regulation of stem cell differentiation.

Methods

Stem cell differentiation is characterized by an intense transcription activity where a number of transcription factors regulates the gene expression of specific targets (Zagar et al., 2011). These

transcriptional changes can be observed by analyzing the genome-wide expression profiles of m genes at n different samples along differentiation. Principal Component Analysis (PCA) may be used to assign a real number $p(s)$ to a sample s based on its expression profile. The result of this inference is a set of real numbers that can be placed in a 1D ruler, the differentiation scale. To construct a more robust predictive model, we combined the expression values from six data sets on embryonic stem cells differentiation provided by Gene Expression Omnibus with a meta-analysis method named Merging. Thanks to this approach, we obtained an integrated scale where a new uncharacterized sample can be projected to uncover its real stage of development with respect to the normal dynamics.

Reduced subsets of genes specifically activated in different stages of differentiation were identified by means of a novel gene selection procedure that assigns to each gene a score proportional to its PCA-inferred weight in the stage s and its expression value.

In order to obtain a complete picture of the gene transcription in each stage, we exploited the biological knowledge available through the STRING database (Szklarczyk et al., 2011). STRING imports and combines data gathered from heterogeneous sources to provide information about known and predicted protein-protein associations. A confidence score is also assigned to each predicted association.

In this work, we developed and analyzed a set of STRING-based networks, one for each stage of development, by applying the following procedure:

1. Network building. For each stage-specific list of genes, we mapped the gene symbols to their corresponding protein ids provided by the UNIPROT database. This step allowed retrieving the protein associations predicted by STRING, that we used to build a set of gene networks. A pair of genes in each network was connected if the confidence score of their proteins association in STRING exceeded a selected global threshold.

2. Topological analysis. The developmental stages measured in the considered experiments were grouped into three phases according to the clusters of projections that we observed in the differentiation scale. We analyzed the similarity of the networks obtained for each phase in terms of their topological properties. First, in order to take into account the number of common gene links in the networks, we computed the median jaccard index between networks of different phases. Moreover, to quantitatively characterize the importance of the nodes in our networks, we considered a topological index known as betweenness centrality. This measure is defined as the number of shortest paths that go through a considered node, and represents the influence of that node in the flow of information within the network. Nodes with a high betweenness typically make possible the communications in the network among clusters of nodes characterized by high internal connectivity. The betweenness values of the genes present in each phase were used to compare the different developmental phases and to identify the most relevant genes in each network.

3. Biological analysis. The expanded lists of genes obtained with STRING were further analyzed in light of the knowledge on developmental processes reported in the literature. Different works have recently pointed out the existence of a set of specific genes that are responsible for a particular pluripotency status of the cell (Zuccotti et al., 2011). First, in order to evaluate the effect of the network-based procedure on the gene selection, we compared the list of genes retrieved with our approach with the known markers.

The analysis then focused on identifying the most biologically significant genes for each differentiation phase. The importance of a gene in a developmental stage is represented by its role in the transcriptional regulatory circuitry of the process and is best quantified by its betweenness centrality value. We therefore looked for sets of significant genes among the bottlenecks of the developed networks. Phase-specific important genes were identified applying a selection procedure based on a global threshold value. For each phase, we determined a list of characterizing genes by extracting those that were contained in more than the 60% of the total number of networks present in the phase. All genes included in at least one phase-specific list were

assigned the median of all their betweenness centrality values. We then considered the distribution of such median values and computed the 95th percentile, which was assumed as the threshold. Finally, from each phase-specific gene list all genes whose betweenness centrality value exceeded the threshold were extracted.

Results and Discussion

The results of the presented procedure confirmed that a transcriptional wave is active during differentiation and influences the topological properties of the networks. While the analysis of the Jaccard index showed no significant phase characterization in terms of common edges, a phase comparison based on betweenness similarity highlighted instead a distance between the first phase and the last stages (Table1).

The biological analysis, i.e. the study of the expanded lists of genes in light of knowledge on developmental processes reported in the literature, identified 53 known markers included in at least one network. In particular, a number of known key genes retrieved with this method, such as Pou5f1, Nanog, Klf4, Sox2, were not previously selected by the data-driven procedure based on their expression profiles. This result confirmed how network-based approaches can integrate experimental findings contributing to the identification of significant genes, whose importance is due to the crucial role they play in the regulation of the global network.

Gene characterization based on betweenness centrality selected a higher number of genes (24) in the first phase if compared to the others (11 and 9 genes for the second and the last phase, respectively), confirming a distinction of the early stages, where the majority of transcriptional changes are known to occur. Known markers as well as novel yet uncharacterized genes were identified. Starting from these results, future experiments will be focused on the application of network-based prioritization procedures, that would help to automatically retrieve the most significant genes in the networks.

Table1: Topological similarity of the networks for the three phases of differentiation.

Phases	1:2	1:3	2:3
Jaccard	0,78	0,8	0,75
Betweenness	0,58	0,58	0,91

Acknowledgements

This work was supported by the Fondazione Cariplo grant (2008–2006) “Bioinformatics for Tissue Engineering: Creation of an International Research Group” and by EU FP7 project “CARE-MI” and grants from Slovenian Research Agency (P2-0209, J2-9699, L2-1112). Camilla Colombo is gratefully acknowledged for her help in software development.

References

1. Aiba K et al (2009) Defining developmental potency and cell lineage trajectories by expression profiling of differentiating mouse embryonic stem cells, *DNA Res* 16:73-80. doi:10.1093/dnares/dsn035
2. Mulas F et al (2012) Supporting Regenerative Medicine by Integrative Dimensionality Reduction, *Methods Inf Med*. 51(4). doi: 10.3414/ME11-02-0045
3. Nitsch D et al (2010) Candidate gene prioritization by network analysis of differential expression using machine learning approaches. *BMC Bioinformatics* 11:460. doi: 10.1186/1471-2105-11-460
4. Szklarczyk D et al (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res* 39: D561-8. doi: 10.1093/nar/gkq973
5. Zagar L et al (2011) Stage prediction of embryonic stem cell differentiation from genome-wide expression data. *Bioinformatics* 27(18):2546-53. doi: 10.1093/bioinformatics/btr422
6. Zuccotti M et al (2011) Gatekeeper of pluripotency: a common Oct4 transcriptional network operates in mouse eggs and embryonic stem cells. *BMC Genomics* 12:1-13. doi: 10.1186/1471-2164-12-345

AnnotateGenomicRegions: a web application

Heiko Muller[✉], Luca Zammataro, Gabriele Bucci

Computational Research, Center for Genomic Science of IIT@SEMM, Istituto Italiano di Tecnologia (IIT), Genova, Italy

Motivation and Objectives

A common denominator for all applications of New Generation Sequencing technology is the need to annotate genomic regions of interest. Tools such as Galaxy (Giardine et al., 2005), CisGenome (Ji et al., 2008), or the Bioconductor ChIPpeakAnno package (Zhu et al., 2010) have been published to perform this task. However, using these tools often requires a significant amount of bioinformatics skills and/or downloading and installing dedicated software. A widely accepted, web-based annotation tool available to bioinformaticians and biologists with widely varying skill levels is not available. Here we present AnnotateGenomicRegions, a web application that accepts genomic regions as input and outputs overlapping and/or neighboring genome annotations chosen on a simple web-form.

Genomic data sets are diverse. However, a common denominator of all studies is the possibility to represent the data as a set of genomic regions identified by "chromosome name : start base - end base", followed by some quantitative or qualitative measure characteristic of the data set. This data format is also used by genome browsers to display known genome features and is called browser embedded format (.bed). Therefore, the most straight-forward way of annotating a genomic data set is based on using genomic regions of interest as genome browser queries.

Tools performing this task have been developed in the past. For example, a bioinformatician with programming skills may use the EnsEMBL core API or the Bioconductor ChIPpeakAnno package. Slightly less demanding is the use CisGenome or Galaxy. All of these options require considerable programming skills, the download of dedicated software, or both. A simple web tool that accepts genomic regions as input and outputs annotations in a format ready to be pasted into an Excel sheet is, to the best of our knowledge, not available. Here we address this need by presenting AnnotateGenomicRegions.

AnnotateGenomicRegions is an open-source web application that can be installed on any computer running the Glassfish web server. This might

be a personal laptop or an institute's Linux cluster. AnnotateGenomicRegions is available at: <http://bioserver.iit.ieo.eu/AnnotateGenomicRegions>

Methods

AnnotateGenomicRegions uses a set of simple Java servlets to process the annotation queries and returns the annotations as zipped, tab-delimited tables. It has been developed using Java Enterprise technology on the NetBeans 6.9 Integrated Development Environment and the Glassfish version 3 web server. This choice is motivated by the better scalability and portability of Java Enterprise as opposed to common gateway interface based web applications. AnnotateGenomicRegions is a Sourceforge project and can be downloaded from <http://sourceforge.net/projects/annotatelocus/> along with detailed descriptions of input and output formats.

Results and Discussion

The design of AnnotateGenomicRegions is based a few simple requirements:

1. Genomic regions shall be used as input query.
2. The output shall be pastable into an Excel table.
3. The application shall be web-based.
4. No programming skills required to use the application.
5. It must be fast enough to annotate hundreds of thousands of genomic regions within seconds

The steps to be followed by the user to annotate his/her data are: on the "Annotate" pane (Figure 1 A) choose the genome, choose the desired features for annotation and whether the feature shall be overlapping and/or neighboring the query regions, paste or upload the query regions, and finally submit the query. The results of an annotation query are displayed in tabular form (Figure 1 B). The results can be downloaded in zip format and pasted into an Excel spreadsheet.

For non-standard annotations, a "CUSTOM" menu option has been provided. Here, the user

Web annotation of genomic regions.

HOME HOW **ANNOTATE** CUSTOM DISTANCE NEWS CONTACT

Annotation of genomic regions

genome: Oct2012/hg19

Annotations for Oct2012/hg19

annotation	overlap	neighbor
hg19/simpleRepeat	<input type="checkbox"/>	<input type="checkbox"/>
hg19/refgene_ID	<input type="checkbox"/>	<input type="checkbox"/>
hg19/phastConsElements	<input type="checkbox"/>	<input type="checkbox"/>
hg19/all_mRNA_ACC	<input type="checkbox"/>	<input type="checkbox"/>
hg19/cpgIslandExt	<input type="checkbox"/>	<input type="checkbox"/>
hg19/refgene_TSSpm1kb_ID	<input type="checkbox"/>	<input type="checkbox"/>
hg19/ensGene_TSSpm1kb	<input type="checkbox"/>	<input type="checkbox"/>
hg19/refgene_TSSpm1kb_Symbol	<input type="checkbox"/>	<input type="checkbox"/>
hg19/ensGene	<input type="checkbox"/>	<input type="checkbox"/>
hg19/refgene_Symbol	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
hg19/all_mRNA_TSSpm1kb_ACC	<input type="checkbox"/>	<input type="checkbox"/>

Input regions. Formats:
 chr1:1000000-1100000 or
 chr1tab1000000tab1100000 or
 chr1space1000000space1100000

chr1:879422-879422
 chr1:881892-881892
 chr1:883516-883516
 chr1:892306-892306
 chr1:892511-892511
 chr1:892634-892634
 chr1:897050-897050
 chr1:949444-949444
 chr1:977447-977447
 chr1:979353-979353
 chr1:982818-982818
 chr1:985349-985349
 chr1:985360-985360
 chr1:987181-987181
 chr1:989314-989314
 chr1:990201-990201
 chr1:1018348-1018355
 chr1:1115602-1115602
 chr1:1115604-1115604
 chr1:1115604-1115604

Clear

Or upload data from a file:

Or paste URL (http://...) to a file in the correct format:

Copyright 2011-2012 by IIT@EMBL. All rights reserved.

[Home](#) [How](#) [Annotate](#) [Custom](#) [Distance](#) [News](#) [Contact](#)

can upload an annotation file in bed format along with the queries. The user chooses the number of desired annotation files, browses to the local files with the annotations, specifies the column numbers for chromosome, start, end, and annotation name, and chooses whether overlap or neighbors queries are desired. When submitting the queries, the annotations will be uploaded to the server, processed for fast annotation, and annotations will be provided as a zipped output file. Distances can be calculated using the "DISTANCE" pane. The annotations used for distance calculations must be provided by the user including strand information.

Design criterion 5 regards the speed and the scaling of the application. Without going into too much detail, the core of the application is located in a Java class called Query. This class ensures that both the query regions and the annotations of interest are sorted first by chromosome and then by start position. For each chromosome, a separate Hashtable object is created that holds the query regions sorted by start position in an ArrayList. Similar Hashtables are created for each annotation. Then, auxiliary Hashtables are generated that make sure that querying a chromo-

download

region	hg19/refgene_Symbol_of	hg19/refgene_Symbol_In	hg19/refgene_Symbol_rn
chr1:69538-69538	OR4F5	FAM138A	LOC729737
chr1:874447-874447	SAMD11	LOC100130417	NOC2L
chr1:874456-874456	SAMD11	LOC100130417	NOC2L
chr1:874465-874466	SAMD11	LOC100130417	NOC2L
chr1:879422-879422	SAMD11	LOC100130417	NOC2L
chr1:881892-881892	NOC2L	SAMD11	KLHL17
chr1:883516-883516	NOC2L	SAMD11	KLHL17
chr1:892306-892306	NOC2L	SAMD11	KLHL17
chr1:892511-892511	NOC2L	SAMD11	KLHL17
chr1:892634-892634	NOC2L	SAMD11	KLHL17
chr1:897050-897050	KLHL17	NOC2L	PLEKHN1
chr1:949444-949444	ISG15	HES4	AGR
chr1:977447-977447	AGR	ISG15	RNF223
chr1:979353-979353	AGR	ISG15	RNF223
chr1:982818-982818	AGR	ISG15	RNF223
chr1:985349-985349	AGR	ISG15	RNF223
chr1:985360-985360	AGR	ISG15	RNF223
chr1:987181-987181	AGR	ISG15	RNF223
chr1:989314-989314	AGR	ISG15	RNF223
chr1:990201-990201	AGR	ISG15	RNF223
chr1:1018348-1018355	C1orf159	RNF223	LOC254099
chr1:1115602-1115602	TTL10	MIR429	TNFRSF18
chr1:1115604-1115604	TTL10	MIR429	TNFRSF18
chr1:1115604-1115604	TTL10	MIR429	TNFRSF18
chr1:1115604-1115604	TTL10	MIR429	TNFRSF18
chr1:1115604-1115604	TTL10	MIR429	TNFRSF18
chr1:1159233-1159233	SDF4	TNFRSF4	B3GAL76
chr1:1164118-1164118	SDF4	TNFRSF4	B3GAL76
chr1:1192497-1192497	UBE2J2	FAM132A	SCNN1D

Figure 1: Screenshot of AnnotateGenomicRegions. A) Annotation pane. B) output example

somal region in the vicinity of a previous query does not result in searching a region that has already been searched by the previous query, which is guaranteed to have a start position smaller than or equal to the start position of the current query. The Query class performs searches for hundreds of thousands of query regions and tens of annotations in a matter of seconds and the scaling with the number of query regions or the size of annotation files is linear.

ChIP-Seq analysis tools have been developed that comprise functional annotation, for example CisGenome, W-ChIPeaks, Sole-Search, or CASSys (Ji et al., 2008; Blahnik et al., 2010; Lan et al., 2011; Alawi et al., 2011). These tools are focusing on the identification of enriched regions in ChIP-Seq experiments and annotation of genomic regions is provided as a side-aspect. Therefore, using these tools for annotation purposes only is cumbersome. Command-line tools such as BEDtools (Quinlan and Hall, 2010) are extremely powerful at identifying overlapping regions in two bed formatted files. But being command-line tools, they are off-limits for most biologists. The same is true for the BioConductor ChIPpeakAnno package (Zhu, 2010). Tools such as the EnsEMBL Ruby API (Strozzi and Aerts, 2011) require considerable programming skills, which precludes widespread use by biologists.

Galaxy (Giardine et al., 2005) is a sophisticated web-based suite of genome analysis tools

that can also perform annotation of genomic regions as part of the "Operate on Genomic Intervals" menu option. It is an expert tool that requires some familiarity. The option "Fetch closest non-overlapping feature" will find annotations that have been defined as "neighbors" in this work. The file defining the neighbors must be uploaded along with the query regions. No default annotations for neighbor fetching are provided. Only one annotation can be fetched at the time. Identification of overlapping features requires the use of a different menu option ("Intersect"). In contrast to AnnotateGenomicRegions, none of the above mentioned tools can be used easily by non-experts..

Acknowledgements

We thank Dr. Davide Cittaro for helpful discussions on application design and implementation.

References

6. Alawi M, Kurtz S, Beckstette, M (2011) CASSys: an integrated software-system for the interactive analysis of ChIP-seq data. *J Integr Bioinform.* 8, 155. doi:[10.2390/biecoll-jib-2011-155](https://doi.org/10.2390/biecoll-jib-2011-155)
7. Blahnik KR, Dou L, et al. (2010) Sole-Search: an integrated analysis program for peak detection and functional annotation using ChIP-seq data. *Nucleic Acids Res.* 38, e13. doi:[10.1093/nar/gkp1012](https://doi.org/10.1093/nar/gkp1012)
8. Giardine B, Riemer C, et al. (2005) Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.* 15, 1451-5. doi:[10.1101/gr.4086505](https://doi.org/10.1101/gr.4086505)
9. Ji H, Jiang H, Ma W, Johnson DS, Myers RM et al. (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat Biotechnol.* 26, 1293-300. doi:[10.1038/nbt.1505](https://doi.org/10.1038/nbt.1505)
10. Lan X, Bonneville R, et al. (2011) W-ChIPeaks: a comprehensive web application tool for processing ChIP-chip and ChIP-seq data. *Bioinformatics* 27, 428-30. doi:[10.1093/bioinformatics/btq669](https://doi.org/10.1093/bioinformatics/btq669)
11. Quinlan AR, Hall IM. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841-2. doi: [10.1093/bioinformatics/btq033](https://doi.org/10.1093/bioinformatics/btq033)
12. Strozzi F, Aerts J (2011) A Ruby API to query the Ensembl database for genomic features. *Bioinformatics* 27, 1013-4. doi:[10.1093/bioinformatics/btr050](https://doi.org/10.1093/bioinformatics/btr050)
13. Zhu LJ, Gazin C, et al. (2010) ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics* 11, 237. doi:[10.1186/1471-2105-11-237](https://doi.org/10.1186/1471-2105-11-237)

G-SNPM - A GPU-based SNP mapping tool

Alessandro Orro¹✉, Andrea Manconi¹, Emanuele Manca², Giuliano Armano², Luciano Milanesi¹

¹Institute for Biomedical Technologies, National Research Council, Milano, Italy

²Department of Electrical and Electronic Engineering, University of Cagliari, Cagliari, Italy

Motivation and Objectives

In genotyping analysis often researchers need to merge together genetic datasets coming from different genotyping platforms that use different sets of Single Nucleotide Polymorphisms (SNPs) to represent genetic polymorphisms. In order to do this, it is necessary to know the exact position of a SNP in a chromosome and update this information when new builds of the reference genome are available.

In this work, we present G-SNPM (GPU SNP Mapping) a GPU-based tool to map SNPs on a genome.

Methods

G-SNPM is a tool that maps a short sequence (read) representative of a SNP against a reference DNA sequence in order to find the absolute position of the SNP in that sequence.

Several tools have been devised to perform short-read mapping. Without aiming to be exhaustive, we can cite some solutions: MAQ (Li and Durbin, 2008), RMAP (Smith et al., 2008; Smith et al., 2009), Bowtie (Langmead et al., 2009), BWA (Li and Durbin, 2009), CloudBurst (Schatz, 2009), and SHRiMP (Rumble et al., 2009). A comparative study aimed at assessing the accuracy and the runtime performance of six state-of-the-art next-generation sequencing read alignment tools (Ruffalo et al., 2011) highlighted that among all SOAPv2 (Li et al., 2009) is the one that shows the higher accuracy.

Recently, it has been proposed SOAPv3 (Liu et al., 2012) the GPU-based evolution of the SOAPv2 aligner. Experimental results shown that SOAPv3

outperforms notably both BWA and Bowtie. When tested to align millions of 100-bp read pairs to the human genome it resulted at least 7.5 times faster than BWA, and 20 times faster than Bowtie. Moreover, SOAPv3 that not exploits heuristics is able to align correctly slightly more reads than BWA and Bowtie. The current release of SOAPv3 supports alignments with up to four mismatches while it does not support indels.

In G-SNPM each SNP is mapped on its related chromosome by means an automatic three stage pipeline. In the first stage, G-SNPM uses SOAPv3 to parallel align on a reference chromosome its related reads representative of a SNP. Due to the fact that SOAPv3 does not support indels, it might not be able to align some reads. Then, in the second stage G-SNPM uses another short-read mapping tool to align the unmapped reads. In particular, in this stage it is used SHRiMP which exploits specialized vector computing hardware to speed-up the dynamic programming algorithm of Smith-Waterman. Finally, in the third stage, G-SNPM analyses the alignments of the reads mapped by SOAPv3 and SHRiMP to calculate the absolute position of each SNP. An output file is generated which for each SNP reports its name, the related chromosome, the original SNP position, and the mapped SNP position. Moreover, information about the alignment as the strand, number of mismatches, and indels are also provided (see Figure 1).

In G-SNPM reference DNA sequences are accepted in standard FASTA format, whereas SNPs must be represented through two files: a FASTA file with the representative reads of the SNPs, and

Name	CHR	SNP	Map	S	M	I	D
rs13305024	Y	17038316	17038318	-	1	0	0
rs9786448	Y	17115298	17115299	+	1	0	0
rs9785704	Y	17175506	17175507	+	0	0	0
MitoA9073G	MT	9073	9073	+	1	0	0
MitoA9094G	MT	9094	9094	+	1	0	0

Figure 1: Screenshot of the generated output file.

another flat file with information about the SNP, in particular the original absolute SNP position and its alleles. Currently, automatic generation of these files is provided for SNP probes of the Illumina Chip. G-SNPM analyses Illumina files to automatically generate the previous described files for each chromosome.

Results and Discussion

The tool has been tested in the problem of re-mapping all the SNP probes of the Illumina Chip HumanOmni 1S (version 1), in order to find the map positions of each SNP in the build 37.3 of the refseq.

To assess the performance of G-SNPM we compared its performance with those obtained by mapping the same SNPs with the state-of-the-art short-read mapping tool BWA. Experimental results shown that in the task of mapping around 1.2 million of SNPs BWA has been unable to map 55 SNPs (maximum edit distance 4% and up to two gap opening), whereas G-SNPM mapped correctly all SNPs. In particular, 178 SNPs has been mapped with SHRiMP in the second stage of the pipeline.

Results shown that BWA has been able to map more reads than SOAPv3. Since SOAPv3 does not support indels, it might be unable to align some reads. However, it should be pointed out that differently that SOAPv3, BWA is designed not to miss any potential alignment resulting in many incorrect mapped reads (Ruffalo et al., 2011).

Currently, G-SNPM runs on linux and it is freely available as a standalone application at the address <http://www.itb.cnr.it/web/bioinformatics/g-snpm>.

To use G-SNPM is required a computer equipped with a CUDA enabled GPU card based on the Fermi architecture. We assessed G-SNPM with a NVIDIA GeForce GTX 480 card.

Acknowledgements

This work has been supported by the Italian Ministry Education and Research (MIUR) through the Flagship "InterOmics", ITALBIONET (RBPR05ZK2Z), HIRMA (RBAP11YS7K) and the European "MIMOMICS" projects.

References

1. Langmead B, Trapnell C, et al. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10:R25.
2. Li H, Ruan J, Durbin R (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, 18(11):1851–8.
3. Li H and Durbin R (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25(14):1754-1760
4. Li R, Yu C, Li Y, et al. (2009). SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, 25(15).
5. Liu CM, Wong T, et al. (2012). SOAP3: ultra-fast GPU-based parallel alignment tool for short reads. *Bioinformatics*, 28(6):878-9.
6. Ruffalo M, LaFramboise T, Koyutürk M (2011). Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics* 15;27(20):2790-6.
7. Rumble SM, Lacroute P, et al. (2009). SHRiMP: Accurate Mapping of Short Color-space Reads. *PLoS Comput Biol* 5(5):e1000386. doi:10.1371/journal.pcbi.1000386.
8. Schatz MC (2009). CloudBurst: highly sensitive read mapping with MapReduce. *Bioinformatics*, 25(11):1363–9.
9. Smith AD, Chung WY, et al. (2009). Updates to the RMAP short-read mapping software. *Bioinformatics*, 25(21):2841–2842.
10. Smith AD, Xuan Z, Zhang MQ (2008). Using quality scores and longer reads improves accuracy of solexa read mapping. *BMC Bioinformatics*, 9:128.

HARP: an automated platform for targeted resequencing data analysis

Fernando Palluzzi✉

Dipartimento di Elettronica e Informazione, Politecnico di Milano, Milan, Italy

Motivation and Objectives

The fast-evolving scenario of Next Generation Sequencing (NGS) technologies caused an increasing demand of ready-to-use, costless and computationally powerful analysis systems, that could both represent a straightforward way to analyze huge amounts of data and offer a set of well assessed protocols to guide the user into an extensive landscape of different standards. In this session, I will present a simple tool called Hierarchical Assisted Resequencing Platform (HARP). HARP is an integrated NGS analysis platform, oriented especially towards resequencing experiments. HARP features allow the user to create personalized resequencing pipelines, using different tools and simplifying their usage and tuning; lead multiple projects at the same time; produce, manipulate, analyze and store data; a user-friendly interface, and finally create graphs, reports and benchmark protocols to assess the final results. Many general purpose platforms, such as Crossbow (Langmead et al. 2009), CloudBurst (Shatz 2009) or Galaxy (Goecks et al. 2010), have been successfully created, providing instruments for computationally intensive analyses directly on internet, without the need of huge hardware facilities. HARP has been prepared with the same purposes, but with the final goal of providing a risk evaluation parameter connected with the clinical and personal genetic profile of breast cancer affected patients.

Methods

HARP is almost completely implemented in Python (<http://www.python.org/>, last accessed on 22/09/2012) and Biopython (http://biopython.org/wiki/Main_Page (last accessed on 22/09/2012), with a minor part of code written in _ Bash (<http://www.gnu.org/software/bash/>, last accessed on 22/09/2012) and R (<http://www.r-project.org/>). A set of internal Python scripts has been used to create the HARP core. The core is composed by functions for environment management, format conversion, data pre-processing and wrappers for a set of third-party dependencies. Bash scripts has

been used for sanity check, while R for statistics and graphics creation.

HARP interface has a modular structure, in which each module is presented to the user as a different menu, i.e. an independent task manager. In addition, there is a panel called experiment design. The experiment design manager allow the user to create personalized pipelines employing and interact with the whole HARP functionalities, by adjusting a relatively low number of basic parameters.

Third-party software include: Fastx toolkit (http://hannonlab.cshl.edu/fastx_toolkit/, last accessed on 22/09/2012), for some data cleaning and manipulation procedure; SMALT (<http://www.sanger.ac.uk/resources/software/smalt/>, last accessed on 22/09/2012), for reference-based alignments; Samtools (<http://samtools.sourceforge.net/>, last accessed on 22/09/2012), to call variants; and finally the simulation tools set called ART (<http://www.niehs.nih.gov/research/resources/software/biostatistics/art/>, last accessed on 22/09/2012), to perform results benchmarking. The different dependencies has been chosen regarding at different characteristics: the capability of working correctly with the main NGS standards, i.e.: fastq, SFF, SAM, BAM, BCF, VCF; the possibility of analyzing data from different experiments, e.g.: single-end or paired-end libraries; flexibility for different purposes, such as whole genome resequencing, amplicon resequencing or exome sequencing; and finally a straightforward usage.

Results and Discussion

In the table below are reported the results of a test analysis performed on a multiplexed sample, containing BRCA1/2 sequences from seven patients affected by breast cancer (BC). All these patients presented at least one BC variant. Specificity is expressed as the number of verified variants (i.e. the variants present in dbSNP) over the number of detected variants.

The risk assessment tool has been developed in R, but currently is not tested due to delays in obtaining real clinical data. However, the HARP risk assessment tool is an implementation of the Gail

Table 1: example of test analysis. This table shows the results of a test performed on an SFF file produced by a multiplexed sequencing experiment with a 454 GS Junior instrument (MID stands for Multiplex ID). The table reports all the detected variants, all those found in dbSNP, those variants that failed the quality check (QC) and the BC-related variants. The specificity is expressed as the ratio between the verified variants (i.e. present in dbSNP) over all the detected ones. Among these patients, only MID4 presents a novel variant (an indel), with unknown relation with BC.

MID	All variants	dbSNP	QC-failed	BC SNPs	Specificity
2	9	7	2	1	77.78%
3	8	8	0	2	100%
4	6	5	0	1	83.33%
5	15	15	0	5	100%
6	16	16	0	4	100%
7	9	9	0	2	100%
8	9	9	0	3	100%

model for absolute risk evaluation, that is a parametric model based on a series of clinical parameters, that can be enhanced using genome information (Gail et al. 1989; Gail 2009). The limitation in using such parametric approaches is the low discriminatory accuracy achieved, that is around 63% when genome information is included (Gail 2010).

Currently, HARP interface is still not available on the web, but a command-line version of HARP, called Breast Cancer risk Pipeline (BCP), is available for testing on Sourceforge, at <https://sourceforge.net/projects/bcpipeline/>.

Acknowledgements

The author thanks Professor Jordi Villa-Freixa, Professor Elena Maestrini and the Biocoputing Group of University of Bologna, for their constant support.

References

1. Ellsworth R E, Decewicz D J, et al. (2010). Breast cancer in the personal genomics era. *Curr. Genomics*, **11**: 146. doi:10.2174/138920210791110951.
2. Gail M H, Brinton L A, et al. (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J. Natl. Cancer Inst.*, **81**: 1879.
3. Gail M H (2009). Value of adding Single-Nucleotide Polymorphism genotypes to a breast cancer risk model. *JNCI*, **101**: 959. doi:10.1093/nci/djp130.
4. Gail M H (2010). Personalized estimates of breast cancer risk in clinical practice and public health. *Stat. Med.*, **30**: 1090. doi:10.1002/sim.4187.
5. Goecks J, Nekrutenko A, et al. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, **11**: R86. doi:10.1186/gb-2010-11-8-r86.
6. Langmead B, Schatz M C, et al. (2009). Searching for SNPs with cloud computing. *Genome Biol.*, **10**: R134. doi:10.1186/gb-2009-10-11-r134.
7. Shatz M C (2009). CloudBurst: highly sensitive read mapping with MapReduce. *Bioinformatics*, **25**: 1363. doi:10.1093/bioinformatics/btp236.

The integration of microRNA target data by biclustering techniques opens new roads for signaling networks analysis

Gianvito Pio¹✉, Michelangelo Ceci¹, Corrado Loglisci¹, Donato Malerba¹, Domenica D'Elia²

¹Department of Computer Science, University of Bari "Aldo Moro", Bari, Italy

²CNR, Institute for Biomedical Technologies, Bari

Motivation and Objectives

MicroRNAs (miRNAs) are key modulators of gene expression. In addition to their recognised role in embryonic and adult cell proliferation and differentiation (Ren et al., 2009), many recent studies on diverse types of human cancer have demonstrated that miRNAs are functionally integrated into those oncogenic pathways that are central to tumorigenesis (Olive et al., 2010). Although microarray profiling and next generation sequencing technologies have allowed researchers to discover much of their structural and functional features as well as many new miRNAs, the current challenge is to understand their specific biological functions and mechanisms through which they are able to ensure cell homeostasis and to control developmental timing and cancer progression. This is not a trivial task because the post-transcriptional regulation of gene expression mediated by miRNAs is rarely resolved by a simple one-to-one interaction between a miRNA and a target gene. It is much more complex, often involving multiple binding of the same miRNA and/or of different miRNAs in a co-operative manner. The combinatorial effects of different miRNAs on the same gene, or on different genes of the same pathway, is an essential part of the mechanism through which they are able to fine-tune signaling pathways (Inui et al., 2010). Indeed, the effect of a miRNA may change depending on which other miRNAs are co-expressed or silenced, which in turn depends on the specific context in which the cell, the tissue or the organism is considered. This makes the interpretation of miRNAs expression profile really difficult and a mere analysis of the list of differentially expressed genes cannot provide enough information to elucidate the multiplicity of potential miRNA:mRNA interactions. In this context, the exploitation of data mining techniques, and in particular of biclustering algorithms, is considered as a useful approach to search the correlations among miRNAs and mRNAs. However, as each miRNA may target hundreds of genes, the

selection of the most significant results for further experimental validations still remains a challenging task for many biologists.

The proposed method, which is implemented in the system HOCCLUS2, has been designed to analyse data of miRNA:mRNA interactions (derived from expression arrays or from large sets of predictions) in order to detect significant co-regulatory partnerships. In particular, the aim is to provide the biologists with a tool which can support them in two challenging tasks, that is, the detection of actual miRNAs target genes and the identification of the context-specific co-associations of different miRNAs. A further contribution to the considered research consists in the ranking of the extracted biclusters on the basis of the semantic similarity between the target genes, which allows the biologists to easily select the most significant results, from a biological view point.

Availability: <http://www.di.uniba.it/~ceci/micFiles/systems/HOCCLUS2/index.html>

Methods

HOCCLUS2 exploits and integrates multiple resources. In particular: i) a novel biclustering algorithm specifically designed for the task in hand; ii) existing SVM-based classification algorithms; iii) large sets of validated or predicted miRNA:mRNA interactions; iv) gene classification ontologies (i.e. Gene Ontology) (Ashburner et al., 2000).

The analysis of miRNA:mRNA interactions consists of three steps:

1. the extraction of a set of non-hierarchically organised biclusters in form of bicliques;
2. an iterative process in which, at each iteration, two operations are performed: i) overlap identification, in which miRNAs or mRNAs belonging to a bicluster can be added to another bicluster, by exploiting an SVM-based classification algorithm; ii) merging, in which biclusters are merged when some (distance- and density-based) heuristic criteria are satisfied. Merging implicitly defines a hierarchy of clusters;

3. a ranking of the extracted biclusters. Ranking is performed on the basis of the p-values obtained by the Student's T-Test through which we compare the intra- and inter- functional similarity of miRNA targets. The similarities between miRNA targets (belonging to the same and to different biclusters, respectively) are pairwise computed according to a semantic similarity measure which takes into account the gene classification provided in GO.

Results and Discussion

In order to identify miRNA:mRNA meaningful interactions, HOCCLUS2 has been specifically designed to identify biclusters which are:

- possibly overlapping, since mRNAs and miRNAs can be involved in multiple regulatory networks. Ignoring this aspect would lead to the identification of incomplete interaction networks;
- hierarchically organised. A hierarchical arrangement facilitates the biological interpretation of results, even when a high number of biclusters is extracted from large datasets of miRNA:mRNA interactions. More importantly, this allows us to exploit the intrinsic hierarchical organisation of miRNAs, where it is possible to distinguish among miRNAs involved in many signaling pathways (universe miRNAs) and pathway-specific miRNAs (intra-pathway miRNAs) (Shirdel et al., 2011);
- highly cohesive. This means that miRNAs and mRNAs in the same bicluster should be highly related and show (only) reliable interactions.

The results reported in this paper are referred to the application of HOCCLUS2 on miRTarBase (Hsu et al., 2011) and miRDIP (Shirdel et al., 2011) selected datasets.

By comparing the results of HOCCLUS2 with those of other biclustering algorithms we have verified that HOCCLUS2 performs significantly better in terms of biclusters cohesiveness, interpretability of the results (thanks to the hierarchical organisation) and biological significance of the extracted biclusters (according to the statistical test on GO).

We have found confirmation of multiple miRNAs co-associations in experimental results reported in the current literature for many of the most significant biclusters produced by HOCCLUS2. Moreover, mRNAs in these biclusters are significantly enriched in the same or

related pathways (Reactome mapping and over-representation statistical analysis) (Haw et al., 2011). Much importantly, we have also identified potential miRNAs combinatorial associations (likely context-specific) and specific miRNA targets (potential new target genes) not yet reported in the literature and that well correlate with existing functional hypothesis. These results suggest that the proposed method is appropriate to easily identify meaningful biological correlations otherwise impossible to discover because of the huge amount of data to deal with. Indeed, the amount of data produced by experimental approaches, if from one hand provides an invaluable resource, on the other hand requires complex and exhausting procedures for their analysis. Searching for the target genes of a miRNA in miRTarBase or in miRDip, or in any other similar database, returns thousands of potential targets and to correlate these results to those of other co-expressed miRNAs is a very complex task. Such a type of analysis may greatly benefit by the application of HOCCLUS2 because of its ability to extract and rank biologically significant interaction networks. Furthermore, the possibility to dissect functional components (miRNAs and target genes) of biclusters at higher level of the hierarchy in smaller co-regulative units (biclusters at lower levels of the hierarchy), provides the key for the interpretation of multiple and diverse co-associations of specific miRNAs which could be responsible for their context-dependent activity. These data are almost impossible to obtain by other biclustering algorithms (Caldas and Kaski, 2010; Yoon and De Micheli 2005; Prelic et al., 2006; Cheng and Church, 2000; Deodhar et al., 2000) and, at our knowledge, no similar approaches have been developed and applied in the miRNAs research domain.

The HOCCLUS2 software, the user manual, all the datasets and detailed results are available from the HOCCLUS2 web site. HOCCLUS2 is currently available as a stand-alone software. The results are available in textual format and can be used for searching significant miRNA co-associations in biclusters as well as specific miRNAs gene targeting. A web-based tool for the analysis of a given set of miRNAs (or mRNAs) which renders biclusters obtained by HOCCLUS2 is under development. A further improvement envisages the integration of gene related pathways information from Reactome.

Acknowledgements

This work is partial fulfillment of the research objective of "DM19410 - Laboratorio di Bioinformatica per la Biodiversità Molecolare" and "PON01_02589 - MicroMap project "Caratterizzazione su larga scala del profilo metatrascrittomico e metagenomico di campioni animali in diverse condizioni fisiopatologiche".

References

- Ashburner M et al. (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25:25-29. doi: [10.1038/75556](https://doi.org/10.1038/75556)
- Caldas J and Kaski S (2010) Hierarchical Generative Biclustering for MicroRNA Expression Analysis. In *Research in Computational Molecular Biology*, vol. 6044 of LNCS 2010:65-79.
- Cheng Y and Church GM (2000) Biclustering of Expression Data. In *Proc. of ISMB'00* 2000:93-103.
- Deodhar M et al. (2009) A scalable framework for discovering coherent co-clusters in noisy data. In *Proc. of ICML'09*:31.
- Haw R et al. (2011) Reactome pathway analysis to enrich biological discovery in proteomics data sets. *Proteomics*, 11 (18):3598-3613. doi: [10.1002/pmic.201100066](https://doi.org/10.1002/pmic.201100066)
- Hsu SD et al. (2011) miRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic Acids Research*, 39 (Database issue):D163-9. doi: [10.1093/nar/gkq1107](https://doi.org/10.1093/nar/gkq1107)
- Inui M et al. (2010) MicroRNA control of signal transduction. *Nat. Rev. Mol. Cell Biol.*, 11:252-263. doi: [10.1038/nrm2868](https://doi.org/10.1038/nrm2868)
- Olive V et al. (2010) mir-17-92, a cluster of miRNAs in the midst of the cancer network. *Int J Biochem Cell Biol*, 42 (8):1348-1354. doi: [10.1016/j.biocel.2010.03.004](https://doi.org/10.1016/j.biocel.2010.03.004)
- Prelic A et al. (2006) A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22 (9):1122-1129. doi: [10.1093/bioinformatics/btl060](https://doi.org/10.1093/bioinformatics/btl060)
- Ren J et al. (2009) MicroRNA and gene expression patterns in the differentiation of human embryonic stem cells. *J Transl Med.*, 7 (20). doi: [10.1186/1479-5876-7-20](https://doi.org/10.1186/1479-5876-7-20)
- Shirdel EA et al. (2011) NAViGaTing the Micronome - Using Multiple MicroRNA Prediction Databases to Identify Signalling Pathway-Associated MicroRNAs. *PLoS ONE*, 6 (2):e17429. doi: [10.1371/journal.pone.0017429](https://doi.org/10.1371/journal.pone.0017429)
- Yoon S and De Micheli G (2005) Prediction of regulatory modules comprising microRNAs and target genes. *Bioinformatics*, 21 (2):93-100. doi: [10.1093/bioinformatics/bti1116](https://doi.org/10.1093/bioinformatics/bti1116)

ONCO-i2b2: improve patients selection through CBR techniques with heterogeneous distance functions

Daniele Segagni¹✉, Matteo Gabetta², Valentina Tibollo¹, Arianna Dagliati³, Alberto Zambelli², Cristiana Larizza², Silvia G Priori¹, Riccardo G Bellazzi²

¹Laboratorio di Informatica e Sistemistica per la Ricerca Clinica, IRCCS Fondazione Salvatore Maugeri, Pavia, Italy

²Dipartimento di Ingegneria Industriale e dell'Informazione, University of Pavia, Pavia, Italy

³IUSS, Istituto Universitario di Studi Superiori, Pavia, Italy

Motivation and Objectives

The University of Pavia (UNIPV) and the IRCCS Fondazione Salvatore Maugeri hospital (FSM) in Pavia have recently started an information technology initiative to support clinical research in oncology called ONCO-i2b2. This project aims at supporting translational research in oncology and exploits the software solutions implemented by the Informatics for Integrating Biology and the Bedside (i2b2) research center. The ONCO-i2b2 software is designed to integrate the i2b2 infrastructure with the hospital information system, with the pathology unit and with a cancer biobank that manages both plasma and cancer tissue samples. Exploiting the medical concepts related to each patient, we have developed a novel data mining procedure that allows researchers to easily identify patients similar to those found with the i2b2 query tool, so as to increase the number of patients, compared to the patient set directly retrieved by the query. This allows physicians to obtain additional information that can support new insights in the study of tumors.

Methods

ONCO-i2b2 is based on the software developed by the Informatics for Integrating Biology and the Bedside (i2b2) research center. i2b2 has delivered an open source suite centered on a data warehouse, which is efficiently queried to find sets of interesting patients through a query tool interface.

The ONCO-i2b2 system gathers data from the FSM pathology unit (PU) database and from the hospital biobank, and integrates them with clinical information from the hospital information system (HIS).

One of the main functionalities of the ONCO-i2b2 project is related to the ability of gathering data about patients and samples, collected during the day-to-day activities of the Oncology I department of the FSM. ONCO-i2b2 also makes these data available for research purposes in an

easy, secure and de-identified way. When a patient is hospitalized, he/she is invited to sign an informed consent to make available for research the samples, specimens and data collected for clinical purposes. Specimens obtained surgically are first analyzed by the pathologists of the PU, who may decide to send the specimens exceeding their expertise (together with the signed informed consent) to the laboratory of experimental oncology. The next step consists in the biobank storage of bio-specimens. ONCO-i2b2 is activated when a biopsy is performed to obtain a detailed diagnosis, and a report is generated. The report contains the cancer diagnosis, including the cancer 'stages' and the size of the tumor. These pieces of information are extracted using a dedicated natural language processing (NLP) module and will be used as concepts for running queries within the i2b2 web client. During this phase the selected samples are de-identified through the use of a new barcode, which does not contain any direct information about the donor.

At the same time the system integrates clinical data automatically from the FSM HIS and matches this information to the biobank samples. This information is then stored in the i2b2 Clinical Research Chart (CRC), the star schema data warehouse on which i2b2 is based.

Within the ONCO-i2b2 project, a case-based reasoning procedure has been developed, in order to allow researchers to enhance the patient selection process with an information retrieval procedure that uses the whole medical concept space related to a patient set to identify a group of similar patients. This functionality supports the extension of the original patient set obtained with the i2b2 query tool, and allows the extraction of the most similar patients to a specific patient on the basis of a set of variables.

At the current stage of the project we are mainly focused on the comparison between patients'

clinical data and we are going to expand the CBR system to allow analysis based on heterogeneous (binary, nominal and continuous) variables. Binary variables refer to the presence/absence of diseases or signs/symptoms. Nominal and continuous variables, instead, represent discrete/continuous values of clinical observations.

Concerning binary variables, to calculate the distance between two patients we exploit the Unified Medical Language System (UMLS) Metathesaurus to model their relationship, in order to create a uniform structure that can be used to compare patients, based on a normalized layer. After a patient set has been retrieved using the i2b2 query tool, the procedure finds all concepts related to patients' binary observations (disease, signs, symptoms) by means of an array containing UMLS concepts. Each concept is represented by its Concept Unique Identifier (CUI), a code that identifies concepts in the UMLS Metathesaurus, and by a boolean modifier, which indicates if the variable referring to the CUI is asserted or negated. The distance computed between cases exploits the semantic similarity between concepts in the UMLS ontology. For this reason we consider such a distance, a Semantic Distance (SD).

The CBR system we are developing computes the distance between patients considering both SD and the Interpolated Value Difference Metric (IVDM) distance proposed by Wilson and Martinez (1997) designed to handle applications with nominal attributes, continuous attributes, or both. It combines the two distances to derive the distance between two patients on the basis of any combination of binary, nominal and continuous variables. The distance function for the Interpolated Value Difference Metric for an attribute a on two patients x and y is defined as:

$$IVDM(x, y) = \sum_{a=1}^m ivdm_a(x_a, y_a)^2$$

where $ivdm_a$ is defined as:

$$ivdm_a(x, y) = \begin{cases} \sum_{c=1}^C |P_{a,x,c} - P_{a,y,c}|^2 & \text{if } a \text{ is discrete} \\ \sum_{c=1}^C |P_{a,c}(x) - P_{a,c}(y)|^2 & \text{otherwise} \end{cases}$$

Results and Discussion

In this phase of the project we have tested the accuracy of the IVDM metric using a specific dataset derived from the amount of cancer data the ONCO-i2b2 CRC contains. Data used in the test phase are related to breast cancer patients, classified by histopathological attributes and concerning cells receptor status: estrogen receptor (ER), progesterone receptor (PR) and HER2. Cells with or without these receptors are called ER positive (ER+), ER negative (ER-), PR positive (PR+), PR negative (PR-), HER2 positive (HER2+), and HER2 negative (HER2-). Cells with none of these receptors are called basal-like or triple negative (TN).

We used as Case Base a cohort of 300 patients, classified in Luminal A (ER+ and low grade), Luminal B (ER+ but often high grade) and TN and a set of 60 patents has been used to validate the IVDM method. Table 1 shows the results of the validation phase.

Table1: this table describes the accuracy of the IVDM method and the number of similar patients rightly predicted using a test set of 60 patients (20 cases for each class).

Class	Similar patients found	Accuracy
Luminal A	12	60%
Luminal B	16	80%
Triple Negative	16	80%
Total	44	73%

The future step of this work consists in combining the SD with the IVDM in order to be able to handle in the distance computation any kind of variables. The next effort will be to combine the two distances in a function that weights them through a coefficient λ to be defined in dependence on the relevance of the two set of variables as:

$$dist = \lambda \times SD + (1-\lambda) \times IVDM$$

At this time the ONCO-i2b2 CRC contains the data of about 7,000 patients related to breast cancer diagnosis (about 600 of them have at least one biological sample in the cancer biobank), totaling about 50,000 visits and 120,000 observations recorded using 960 concepts. This very huge data set will represent a very relevant mean for validating our CBR system. The patient retrieval time is in the order of a

few seconds for patient sets up to 1000 patients. The performance decreases for larger patient sets. The implementation of such heterogeneous distance function we expect will enhance the i2b2 framework allowing the exploitation of the overall patient set for a most flexible patients retrieval from the ONCO-i2b2 CRC. Further evolutions of the system are related to import clinical data coming from the ordinary medical activity like haematochemical or instrumental.

Acknowledgements

The Onco-i2b2 project is funded by the "Regione Lombardia" in Italy. We gratefully acknowledge Prof. Carlo Bernasconi and the Collegio Ghislieri in Pavia for their active support. This paper revises and extends the paper "ONCO-i2b2: improve patients selection through case-based information retrieval techniques", by D. Segagni et al, presented at the DILS 2012 conference in Washington DC.

References

1. Betsy L Humphreys, Donald A B Lindberg, Harold M Schoolman, G Octo Barnett (1998) The Unified Medical Language System: An Informatics Research Collaboration. *J Am Med Inform Assoc.* 5, 1-11
2. Caviedes J, Cimino J (2004) Towards the development of a conceptual distance metric for the UMLS. *J. Biomed. Inform.* 37, 77-85
3. Mate S, Bürkle T, et al. (2011) Populating the i2b2 database with heterogeneous EMR data: a semantic network approach. *Stud Health Technol Inform.* 169,502-506
4. Melton GB, et al. (2006) Inter-patient distance metrics using SNOMED CT defining relationships. *J.Biomed. Inform.* 39(6), 697-705
5. Murphy SN, Weber G, et al. (2010) Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J Am Med Inform Assoc.* 17(2), 124-130
6. Segagni D, et al. (2012) ONCO-i2b2: improve patients selection through case-based information retrieval techniques. *Lecture Notes in Computer Science 7348*, 93-99
7. Strauss JA, Chao CR, Kwan ML, Ahmed SA, Schottinger JE, Quinn VP (2012) Identifying primary and recurrent cancers using a SAS-based natural language processing algorithm. *J Am Med Inform Assoc.* [Epub ahead of print] PubMed PMID: 22822041
8. Wilson DR, Martinez TR (1997) Improved Heterogeneous Distance Functions. *Journal of Artificial Intelligence Research* 6, 1-34

Bioinformatics approach for data management about bone cells grown on substitute materials

Federica Viti¹✉, Ivan Merelli¹, Silvia Scaglione², Luciano Milanesi¹

¹Institute for Biomedical Technologies, National Council of Research, Segrate, Italy

²Institute of Electronics Computer and Telecommunication Engineering, National Council of Research, Genoa, Italy

Motivation and Objectives

Tissue engineering, the research field aimed at finding high technological biomaterials able to restore, maintain, or improve tissue function, concentrates many efforts in the contest of bone and cartilage, due to their possible wide-spread clinical applications. The main target is the design of well performing scaffolds suitable to promote the development of natural tissue in implant conditions, without generating rejection and, hopefully, degrading *in vivo* at the same rate of tissue formation.

Concerning both bone and cartilage, one of the most important research aspects is determining the biochemical and topological factors that induce cell differentiation and tissue ingrowth. The scaffold material composition is crucial and many of them have been already tested, including alginate (Duggal et al., 2009), collagen/chitosan (Ravindran et al., 2012), polycaprolactone (PCL) and hydroxyapatite (HA) (Scaglione et al., 2010), Poly-L-Lactide Acid (PLLA) (Ciapetti et al., 2012), polymethylmethacrylate (Bombonato-Prado et al., 2007), bioactive glasses (Leven et al., 2004), carbon nanotubes (Van der Zande et al., 2004), etc.

To better evaluate material performance, the biomolecular characterization of the cellular response is becoming a common practice among researchers. Nonetheless, experimental data usually remain sparse in literature: the collection of high-throughput gene expression profiles from samples on different materials could allow data comparisons and formulation of new hypotheses about the effectiveness of bone/cartilage substitute.

In this context authors extended the existing OsteoChondroDB database (Viti et al., 2012), which collects data and metadata from microarray gene expression of cells cultured in different conditions onto diverse materials, and allows analyzing differentially expressed genes (DEG) from the available knowledge base.

Methods

The OsteoChondroDB relies on MySQL database, while the web interface has been developed using php and javascript technologies, and this improved version of the systems is based on the same infrastructure.

Manual research has been performed on papers containing biomolecular data about osteochondral tissue developed on different scaffolds. Data have been retrieved from known public repositories such as Gene Expression Omnibus (<http://www.ncbi.nlm.nih.gov/geo/>) and ArrayExpress (<http://www.ebi.ac.uk/arrayexpress/>), whenever experiments were available, or directly contacting papers authors. The amount of data produced in this field is new, scarce, and variegated, although the importance of biomolecular aspects related to the tissue growth on materials can be of great importance to design improved scaffolds.

To exploit the available data in an integrated fashion, the strategy described by Kodama et al. (Kodama et al., 2012) appears useful, because it proposes a multi-species and multi-platform approach for gene expression microarray data meta-analysis. In this way, it is possible to increase the number of evidences, by considering mouse and rat data together with human experiments and by mixing different versions, brands and designs of microarray chips. The information mapping between species can be performed through AILUN system (Chen R et al., 2007), which converts ids of different platforms.

Results and discussion

A new section of the OsteoChondroDB (freely accessible to users at url: <http://www.itb.cnr.it/osteochondrogene/>) has been created to maintain the collected information. The database is organized into tables containing the considered biomaterial types, the organism and the exploited cells, the array platforms, the PubMed identifier of the paper, some notes about the experi-

ment and the obtained results. Contextually, the database web interface has been extended to suitably visualize this new information. Collected experiments have been grouped according to the reference biomaterial, and metadata and data about experimental conditions can be browsed from the related section. Data are stored in the file-system and accessed through a link from the database web interface, while metadata are hosted in the database itself. Data retrieval is possible starting from array type, organism and biomaterial.

In order to infer new knowledge about mechanisms involved in bone and cartilage generation, a suitable pipeline (Figure 1) has been designed to mine knowledge from collected information. Microarray data, retrieved from different sources, are maintained into a single database table, whose main fields are the name of the gene, and the expression value in each experiment. Samples are grouped into treatments (diverse biomaterials) and controls, and inter-microarray normalization is performed on the basis of recognized housekeeping genes (de Jonge et al., 2007) that are preserved in each microarray platform. The core of the strategy relies on the approach implemented by Kodama et al.: considering collected datasets as belonging to

a single experiment, it produces an integrated model of genes expression values from which a selection of DEG can be statistically inferred.

Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) ontology have been exploited to annotate genes, and define the functional enrichment of the results set (on the basis of the hypergeometric distribution), in order to elucidate the mostly activated functional mechanisms involved in bone tissue generation according to each biomaterial.

Moreover, the database interface has been extended to perform data annotation. The OsteoChondroDB maintains many references related to bone and cartilage that can be easily exploited to annotate DEG with the collected knowledge base. A suitable interface has been developed to upload microarray results, to automatically perform annotation and to retrieve the list enriched by one or more literature references for each DEG.

The OsteoChondroDB improvements discussed in this work offer users the possibility to access data that previously were sparse in literature, providing the possibility of statistically analyzing them in an integrated fashion. Two main features contribute to add value to the application: the automatic annotation of osteochondro

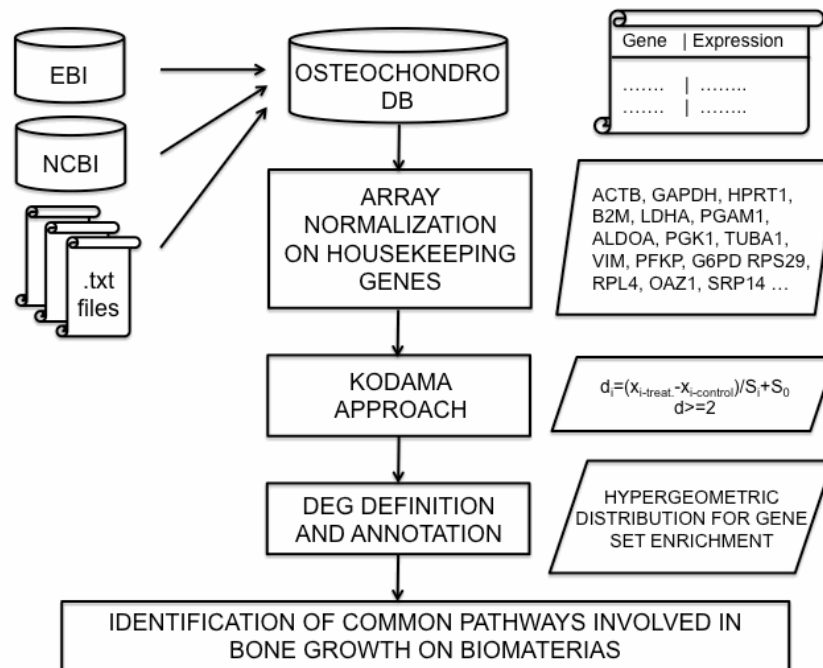


Figure 1: Schema of the designed analysis pipeline.

genes according to literature references, which improves the information about physiological pathways involved in the development of these tissues, and the exploitation of many integrated microarray data to statistically analyze bone and cartilage ingrowth, that represents a useful support for tissue engineering applications.

Acknowledgements

This work has been supported by the Italian Ministry Education and Research (MIUR) through the Flagship "InterOmics", ITALBIONET (RBPR05ZK2Z), HIRMA (RBAP11YS7K) and the European "MIMOMICS" projects.

References

1. Bombonato-Prado KF, Bellesini LS, et al. (2007) Microarray-based gene expression analysis of human osteoblasts in response to different biomaterials, *J Biomed Mater Res A* 88(2), 401-8.
2. Chen R., Li L., and Butte AJ (2007) AILUN: Reannotating Gene Expression Data Automatically, *Nature Methods* 4(11), 879
3. Ciapetti G, Granchi D, et al. (2012) Enhancing Osteoconduction of PLLA-Based Nanocomposite Scaffolds for Bone Regeneration Using Different Biomimetic Signals to MSCs, *Int. J. Mol. Sci.* 13, 2439-2458; doi:10.3390/ijms13022439.
4. de Jonge HJM, Fehrmann RSN et al. (2007) Evidence Based Selection of Housekeeping Genes, *PLoS ONE* 2(9), e898, doi:10.1371/journal.pone.0000898
5. Duggal S, Fronsdal KB, et al. (2009) Phenotype and Gene Expression of Human Mesenchymal Stem Cells in Alginate Scaffolds, *Tissue Engineering: Part A* 15(7). doi: 10.1089/ten.tea.2008.0306.
6. Kodama K, Horikoshi M et al. (2012) Expression-based genome-wide association study links the receptor CD44 in adipose tissue with type 2 diabetes, *Proc Natl Acad Sci U S A* 109(18), 7049-54
7. Leven RM, Virdi AS et al. (2004) Patterns of gene expression in rat bone marrow stromal cells cultured on titanium alloy discs of different roughness, *J Biomed Mater Res A* 70(3), 391-401
8. Ravindran S, Gao Q, et al. (2012) Biomimetic Extracellular Matrix-Incorporated Scaffold Induces Osteogenic Gene Expression in Human Marrow Stromal Cells, *Tissue Engineering: Part A* 18(3-4), doi: 10.1089/ten.tea.2011.0136.
9. Scaglione S, Lazzarini E, Ilengo C, Quarto R. (2010) A composite material model for improved bone formation. *J Tissue Eng Regen Med.* 4(7):505-13.
10. Van der Zande M, Walboomers F, et al. (2010) Genetic profiling of osteoblast-like cells cultured on a novel bone reconstructive material, consisting of poly-L-lactide, carbon nanotubes and microhydroxyapatite, in the presence of bone morphogenetic protein-2, *Acta Biomater.* 6(11),4352-60.
11. Viti F, Merelli I and Milanesi L (2012) OsteoChondroDB: a database about biomolecular chondral-bone development in physiological and diseased conditions, *EMBnet.journal* 18

A first RDF implementation of the COSMIC database on mutations in cancer

Achille Zappa¹✉, Paolo Romano²

¹Department of Informatics Bioengineering Robotics and Systems Engineering (DIBRIS), University of Genoa, Genoa, Italy

²Bioinformatics Laboratory, IRCCS AOU San Martino - IST, Genoa, Italy

Motivation and Objectives

Within a living organism, genome and proteome variations may influence many molecular interactions and biochemical pathways, leading to deleterious effects in the proper activity of cells, tissues, and organs; ultimately, this may be the cause of many syndromes and diseases. It is now well known that tumors may arise as a result of a series of DNA sequence abnormalities and mutations. It is then not surprising that there is a vast amount of information available in the scientific literature and that a lot of information systems devoted to the management of related data exist. Among these, of particular interest are the many Locus Specific Data Bases (LSDB) and the COSMIC (Catalogue of Somatic Mutations in Cancer) database (Forbes et al., 2011). Such data, however, are not yet sufficiently integrated with other molecular, biomedical, and clinical databases. New efforts are therefore needed in this direction.

Data retrieval, search and integration solutions in bioinformatics are increasingly making use of a set of standards and technologies which are the basis of the Semantic Web (Berners-Lee et al., 2001) framework. This framework is intended to evolve the web into a distributed knowledge-base and a first step in this evolution is the generation of a Web of Data (Bizer et al., 2009). In this view, we can see Linked Data as an approach to data integration that employs ontologies, terminologies, Uniform Resource Identifiers (URIs), and the Resource Description Framework (RDF) to connect pieces of data, information and knowledge on the Semantic Web (Belleau et al., 2008). In particular, RDF describes semantic rich information on the web through a composition of simple triples (predicates), such as ('Subject', 'Property', 'Object'), that link entities through relations which are expressed by using ontologies, and are defined by using URIs. See the RDF reference site: <http://www.w3.org/RDF/>, last accessed on October 3, 2012). A relevant contribution to this vision comes from the conversion of data stored in relational databases (RDB) into RDF. There is a vast amount of information on human

variation in the literature and several mutation and variation databases, but, to our knowledge, this kind of information is still scarce in the Web of Data. Various motivations can be depicted for using Semantic Web technologies and publishing Linked Data life sciences datasets; this allows to improve data and information integration, share ability of openly accessible data through standard and programmatic interfaces, semantic normalization, data discoverability and query federation from distributed sources.

A first work carried out by our group led to the implementation of an RDF version (Zappa et al., 2012) of the IARC TP53 Somatic Mutation database (IARCDB) (Petitjean et al., 2007). Here, we present the initial development of an RDF version of the COSMIC (Catalogue of Somatic Mutations in Cancer) database by means of Semantic Web technologies.

Methods

COSMIC was developed, and is currently maintained, at the Wellcome Trust Sanger Institute. It is designed to gather, curate, and organize information on somatic mutations in cancer and to make it freely available on-line. It combines cancer mutation data, manually curated from the scientific literature, with the output from the Cancer Genome Project (CGP). Genes are selected for full literature curation using the Cancer Gene Census. COSMIC datasets are freely available as common CSV flat files. However these files don't contain all the available information and they don't reflect the original schema and table contents of the database. For this reason, and also due to the huge amount of data, we started from a basic automatic RDB to RDF mapping of a relational version of COSMIC. Many research works have been focused on mapping data from RDB to RDF. They have led to the implementation of both mapping tools and domain specific applications. The structure of an RDB database may provide a partial characterization of semantics of the domain it refers to. Some tools rely on this property to generate an

approximate mapping to RDF, which can then be manually tuned and thus brought to be in line with a shared conceptualization.

Mapping is the process of making explicit correspondences or relationships between entities in the relational database and the RDF graph. In our case, the mapping was first created by using D2RQ, a platform for treating relational databases as virtual RDF graphs. See the D2RQ web site: <http://www.d2rq.org/>, last accessed on October 3, 2012). This tool also allows on-the-fly generation of RDF triples from the database. The relational database was then published using a D2R server. D2R enabled us to publish a first SPARQL endpoint on top of the relational database, build an RDF data dump, and make it possible browsing the generated RDF triples through a standard web interface.

One of the most important aspects of the RDB to RDF conversion is, however, the capability of representing the semantics that is not explicitly defined in the relational schema. After a careful analysis of the database schema, we were able to map our resources into separated well defined classes and sub-graphs and to define the relationships and properties of our statements. For instance, D2RQ generates predicate names which are based on the RDB column names: it has no way to know when a predicate refers to a property for which a shared representation (ontological concept) exists. By customizing predicates we have been able to improve the representation of data semantics, according to shared ontologies. Where shared relations were not available to express the content of our database, we have used ad-hoc defined properties.

The final RDF dataset is being deployed according to Linked Open Data (LOD) principles with external links set to datasets such as DBpedia, a system including all structured information which is present in Wikipedia pages (see DBpedia web site: <http://www.dbpedia.org/>, last accessed on October 3, 2012), PubMed, the Human Genome Nomenclature Committee (HGNC) database (see HGNC web site: <http://www.genenames.org/>, last accessed on October 3, 2012), the On-line Mendelian Inheritance in Man (OMIM) system, UniProt (Belleau et al., 2008) and Linked Life Data.

In order to improve performances, the RDF export must be imported into a native RDF triple store system. The RDF dump of COSMIC was then uploaded in a Jena TDB triple store. See the Jena and TDB web sites at: <http://openjena.org/>

[index.html](#) and at <http://jena.sourceforge.net/TDB/>, last accessed on October 3, 2012). A Fuseki server was implemented to make available our data through a SPARQL endpoint. See the Fuseki web site at: <http://fuseki.sourceforge.net/>, last accessed on October 3, 2012).

Since one of the main use cases and aim of COSMIC is to provide somatic mutation frequencies and distributions via plots and histograms, it is then a good practice to deploy a web interface able to graphically visualize such kind of information also in a Semantic Web context. A web interface based on javascript and some graphical libraries can then display results of SPARQL queries to improve visualization of this kind of information by means of charts.

Results and Discussion

Prototype servers are available on-line. The D2R server web site is available at http://bioinformatics.istge.it/D2R_CosmicRDF_proto/. The SPARQL endpoint, that is only meant for SPARQL queries and cannot therefore be used as-is by researchers, is available at the following URL: http://bioinformatics.istge.it/CosmicRDF_protosparql/cosmic/sparql. Currently, servers present only a subset of the database, corresponding to the "full export" that may be downloaded from the COSMIC web site. This dataset, however, does not reflect the database schema, whose analysis is an ongoing effort.

A Linked Data view, an HTML view and a SPARQL endpoint are available. The latter can be explored by any Semantic Web browser or application. These are building blocks for data integration solutions incorporating mutation data. The standard web interface includes graphical visualization features of results of some specific SPARQL queries. These prototypes demonstrate how an RDF representation of relational database contents can be easily provided.

Although a great value of our system would lie on the identification of a shared, semantically meaningful, ontology-based representation of variation information, that could only be defined through a collaboration with the community of curators of variation databases, our approach already allows to carry out queries on the database, as well as some graph-analysis for validation of data and elucidation of implicit relations among data, relations that could not be exploited with the current system.

Relying on dereferenceable URIs, that is URIs that may be redirected to a unique existing Internet address usually accessible via HTTP, existing predicates and ontologies allows our system to be a part of the growing Web of Data with the aim to be in integrated and interlinked part of the Linked Open Data Cloud.

An improved and extended version of our prototypes and interfaces are under development. A new web interface with demo queries and specific use-cases is also under development, with the aim of building a prototype user-friendly interface that can be more proficiently used by such users as biologists and clinicians.

Acknowledgements

This work has been partially supported by the Liguria region (project Liguria eScience).

References

1. Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *Journal of Biomedical Informatics* (2008) 41(5):706-716.
2. Berners-Lee T, Hendler J, Lassila O. The semantic web. *Scientific American*, May 2001.
3. Bizer C, Heath T, Berners-Lee T. Linked Data – The Story So Far. *International Journal on Semantic Web and Information Systems* (2009) 5(3):1-22.
4. Petitjean A, Mathe E, Kato S, Ishioka C, Tavtigian SV, Hainaut P, Olivier M. Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: lessons from recent developments in the IARC TP53 database. *Human Mutation* (2007) 28(6):622-629.
5. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, Jia M, Shepherd R, Leung K, Menzies A, Teague JW, Campbell PJ, Stratton MR, Futreal PA. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucl. Acids Res.* (2011) 39 (Suppl 1): D945-D950. doi: 10.1093/nar/gkq929.
6. Zappa A, Splendiani A, et al. (2012) Towards linked open gene mutations data. *BMC Bioinformatics* 13(Suppl 4):S7. doi:10.1186/1471-2105-13-S4-S7.

National Nodes

Argentina

IBBM, Facultad de Cs. Exactas, Universidad Nacional de La Plata, Buenos Aires

Brazil

Lab. Nacional de Computação Científica, Lab. de Bioinformática, Petrópolis, Rio de Janeiro

Chile

Centre for Biochemical Engineering and Biotechnology (CIByB), University of Chile, Santiago

China

Centre of Bioinformatics, Peking University, Beijing

Colombia

Instituto de Biotecnología, Universidad Nacional de Colombia, Edificio Manuel Ancizar, Bogota

Costa Rica

University of Costa Rica (UCR), School of Medicine, Department of Pharmacology and ClinicToxicology, San Jose

Finland

CSC, Espoo

France

ReNaBi, French bioinformatics platforms network, Villeurbanne

Greece

Biomedical Research Foundation of the Academy of Athens, Athens

Hungary

Agricultural Biotechnology Center, Godollo

Italy

CNR - Institute for Biomedical Technologies, Bioinformatics and Genomic Group, Bari

Luxembourg

Luxembourg Centre for Systems BioMedicine (LCSB), Luxembourg

Mexico

Nodo Nacional de Bioinformática, EMBnet México, Centro de Ciencias Genómicas, UNAM, Cuernavaca, Morelos

Norway

The Norwegian EMBnet Node, The Biotechnology Centre of Oslo, Oslo

Pakistan

COMSATS Institute of Information Technology, Chak Shahzaad, Islamabad

Poland

Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Warszawa

Portugal

Instituto Gulbenkian de Ciencia, Centro Portugues de Bioinformatica, Oeiras

Russia

Biocomputing Group, Belozersky Institute, Moscow

Slovakia

Institute of Molecular Biology, Slovak Academy of Science, Bratislava

South Africa

SANBI, University of the Western Cape, Bellville

Spain

EMBnet/CNB, Centro Nacional de Biotecnología, Madrid

Sri Lanka

Institute of Biochemistry, Molecular Biology and Biotechnology, University of Colombo, Colombo

Sweden

Uppsala Biomedical Centre, Computing Department, Uppsala

Switzerland

Swiss Institute of Bioinformatics, Lausanne

The Netherlands

Centre for Molecular and Biomolecular Informatics, Radboud University Nijmegen Medical Centre, Nijmegen

Specialist- and Assoc. Nodes

CASPUR

Rome, Italy

EBI

EBI Embl Outstation, Hinxton, Cambridge, UK

Nile University

Giza, Egypt

ETI

Amsterdam, The Netherlands

IHCP

Institute of Health and Consumer Protection, Ispra, Italy

ILRI/BECA

International Livestock Research Institute, Nairobi, Kenya

MIPS

Muenchen, Germany

UMBER

Faculty of Life Sciences, The University of Manchester, UK

CPGR

Centre for Proteomic and Genomic Research, Cape Town, South Africa

The New South Wales Systems

Biology Initiative Sydney, Australia

for more information visit our Web site

www.embnet.org

EMBnet.journal

ISSN 2226-6089

Dear reader,

If you have any comments or suggestions regarding this journal we would be very glad to hear from you. If you have a tip you feel we can publish then please let us know. Before submitting your contribution read the "Instructions for authors" at <http://journal.EMBnet.org/index.php/EMBnetnews/about> and send your manuscript and supplementary files using our on-line submission system at <http://journal.EMBnet.org/index.php/EMBnetnews/about/submissions#onlineSubmissions>.

Past issues are available as PDF files from the Web site:

<http://journal.EMBnet.org/index.php/EMBnetnews/issue/archive>

Publisher:

EMBnet Stichting p/a
CMBI Radboud University
Nijmegen Medical Centre
6581 GB Nijmegen
The Netherlands

Email: erik.bongcam@slu.se

Tel: +46-18-4716696