EMBnet.journal
Volume 19
Supplement B
October 2013

NETTAB 2013
Workshop on "Semantic, Social, and Mobile Applications for Bioinformatics and Biomedical Laboratories"
16-18 October 2013, Lido of Venice, Italy
http://www.nettab.org/2013/

Editorial

Dear reader, welcome to a new supplement of *EMBnet.journal*, an international, open access, peer-reviewed bioinformatics journal.

This supplement is dedicated to the NETTAB 2013 workshop focused on

"*Semantic, Social and Mobile Applications for Bioinformatics and Biomedical Laboratories*", held 16-18 October 2013 in Venice Lido, Italy.

NETTAB 2013 is the thirteenth in a series of international workshops on Network Tools and Applications in Biology. NETTAB workshops are aimed at presenting and discussing emerging ICT technologies whose adoption in support of biology could be of particular interest.

EMBnet.journal is very pleased to publish the contributions to the workshop presented at this important event. For this special issue (19.B), the selection of articles was overseen by the Conference Scientific Committee, while the lay-out 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. In this series, we have also published the First Scientific Meeting of the COST Action, SeqAhead (issue 17.B), proceedings of the Bioinformatics Italian Society (BITS) 9th Annual Meeting, 2-4 May 2012 (issue 18.A), proceedings of the NETTAB 2012 Workshop on "*Integrated Bio-Search*", 14-16 November 2012, Como, Italy (issue 18.B), and the SeqAhead "*The Next NGS Challenge Conference: Data Processing and Integration*", 14-16 May 2013, Valencia, Spain (issue 19.A).

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

EMBnet.journal Editorial Board

Cover picture: Sunset on the Venice's lagoon. Venice, Italy, 2008. [© Nicola Cannata]

Contents

Preface

NETTAB 2013

Workshop on "Semantic, Social, and Mobile Applications for Bioinformatics and Biomedical Laboratories"


Nicola Cannata[1], Paolo Romano[2]

[1]University of Camerino, Camerino (MC), Italy
[2]IRCCS AOU San Martino IST, Genoa, Italy

Due to the distributed nature of biomedical information resources, and to their heterogeneity and size, it is now clear that bioinformatics plays an essential role in every domain of biomedical research, from experimental to clinical studies. In this context, network tools, platforms and applications are one of the most relevant research and development domain for bioinformatics, because they may indeed allow an efficient and effective integration of information, that is the basis for any further data analysis.

Since 2001, a series of workshops has been devoted to this topic and held annually in Italy, under the name of "Network Tools and Applications in Biology" (NETTAB). These workshops have been aimed at presenting and discussing some of the most innovative Information and Communication Technologies (ICTs) and their application in the biomedical domain.

Each workshop has been focused on a different topic. Since 2001, many different focus themes have been explored, including Standardization for data integration (Genoa, 2001), Multi Agent Systems (Bologna, 2002), Scientific workflows (Naples, 2005), Grid and Web Services (Santa Margherita di Pula, 2006), Semantic Web (Pisa, 2007), Collaborative research and development (Catania, 2009), and Biological Wikis (Naples, 2010).

The NETTAB 2013 workshop, the thirteenth in the series, is held in Lido of Venice, Italy, on October 16-18, 2013. It is focused on "Semantic, Social, and Mobile Applications for Bioinformatics and Biomedical Laboratories".

Human beings are social animals and ICT have recently permeated human society of newer forms and ways of participation in social activities. In the Internet, the hype has shifted from Web2.0 to Social Media; in science too, this development is evident. Beside facilitating communication and making easier the sharing of information, social platforms and technologies are enhancing learning, problem solving and crowdsourcing. In biology, and especially in the "-omic" disciplines, we already rely on a wide diffusion of social tools and applications, e.g. for distributed annotations, Wiki knowledge bases, documentation and productivity.

On the other hand, access to the Internet is nowadays increasingly happening through mobile devices. Mobile Internet access is expected to soon overtake access from standard PCs and workstations. Moreover, mobile phones are expected to become the main personal computing devices. Smartphones and tablets represent the most practical computing devices in biomedical laboratories and they actually are the ideal companions for "always on the move" scientists. While we can observe a widespread diffusion of health and lifestyle mobile applications and a rapid adoption of mobile solutions in medicine and healthcare we cannot say the same for life sciences and bioinformatics.

Semantic methodologies and technologies are instead well established in "-omic" projects. It can even be proudly observed that the bioinformatics community was an early adopter of Semantic Web technologies.

In the NETTAB 2013 workshop, mobile and/or social and/or semantic solutions for bioinformatics and laboratory informatics problems will be explored. It is our opinion that a savvy combination of these three technologies could greatly enhance the research outcome of life scientists and markedly simplify the workflows in biomedical laboratories.

The workshop is open to all aspects of the focus theme, including issues, methods, algorithms, and technologies for the design and development of tools and platforms able to provide semantic, social, and mobile applications supporting bioinformatics and the activities carried out in a biomedical laboratory.
The Call for abstracts was able to attract 20 quality submissions that are after peer review and revision included in these proceedings, grouped by submission type. Oral communication abstracts are listed according to the presentation in the programme, while poster abstracts are ordered by first author name.

The Chairs of the NETTAB 2013 workshop hope that this is a great meeting for all participants. Ideas, and doubts, on the perspectives of the widespread application of these new technologies will be discussed with outstanding scientists such as Antony Williams, Ross King, Barend Mons, Andrea Splendiani, Christine Chichester, Dominique Hazaël-Massieux, and Alex Clark, who present invited keynotes and tutorials, and many others who enthusiastically join the workshop and actively participate in it.
We wish to thanks all participants, but also all Supporting Institutes and companies that made this workshop possible and successful.

Speakers

Keynote Speakers

Antony Williams
Royal Society of Chemistry

Ross D. King
School of Computer Science, University of Manchester, Manchester, United Kingdom

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

Tutorials

Andrea Splendiani
IntelLeaf, United Kingdom, and Digital Enterprise Research Institute, Ireland

Christine Chichester
Swiss Institute of Bioinformatics, CALIPHO group, Geneva, Switzerland

Dominique Hazaël-Massieux
W3C/ERCIM, Sophia Antipolis, Biot, France

Alex Clark
Molecular Materials Informatics, Inc

Oral presentations

Christine Chichester
René Witte
Ismael Navas-Delgado
Fabio Rinaldi
Rebecca Lawrence
Tim P. Eyres
Steffen Möller
Stian Soiland-Reyes
Alejandra Gonzalez-Beltran
Martijn Devisscher
Larisa N. Soldatova
Mario Cannataro

Editors

Chairs and Conference Committees

Chairs

Scientific committee

Organising Committee

Supporting Institutes, Scientific Societies and Projects

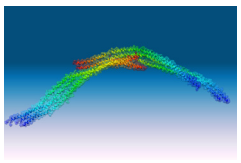
The workshop is held under the patronage of:

ISCB International Society for Computational Biology
http://www.iscb.org/

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

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

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



Rete Ligure di Bioinformatica
https://sites.google.com/site/reteliguredibioinformatica/

Polish Bioinformatics Society
http://www.ptbi.org.pl/

and with support from:

University of Camerino (MC), Italy
http://www.unicam.it/

IRCCS AOU San Martino IST, Genoa, Italy
http://www.hsanmartino.it/

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

Sponsors

This workshop has been supported by:

BMR Genomics
http://www.bmr-genomics.com/

Genostar
http://www.genostar.com/

F1000 Research
http://f1000research.com/

Life Technologies
http://www.lifetechnologies.com/

Aldebra
http://www.aldebra.com/

EMC²
http://www.ecm2.it/

Conference Programme

NETTAB 2013
Workshop on "Semantic, Social, and Mobile Applications for Bioinformatics and Biomedical Laboratories"

16-18 October 2013, Lido of Venice, Italy

http://www.nettab.org/2013/

Conference Programme

Wednesday	October 16, 2013 TUTORIAL DAY
10.30 - 17.30	*Registration and poster hang-up*
11.00 - 13.00	**Tutorial** 1
	Semantic Web for Life Sciences: vision, aims, tools, platforms *Andrea Splendiani* *IntelLeaf, United Kingdom, and Digital Enterprise Research Institute, Ireland*
	Demonstration on Tutorial 1
13.00 - 13.30	**Brunch Break**
13.30 - 15.30	**Tutorial 2**
	Open PHACTS and NanoPublications *Christine Chichester* *Swiss Institute of Bioinformatics, CALIPHO group, Geneva, Switzerland*
	Demonstration on Tutorial 2
15.30 - 17.30	**Tutorial 3**
	Standards for Web Applications on Mobile: current state and roadmap *Dominique Hazaël-Massieux, W3C/ERCIM,* *Sophia Antipolis, Biot, France*
	Demonstration on Tutorial 3
17.30 - 19.30	**Tutorial 4**
	Mobile applications for life sciences: perspectives, limitations, and real examples *Alex Clark* *Molecular Materials Informatics, Inc*
	Demonstration on Tutorial 4

Thursday	October 17, 2013 WORKSHOP DAY 1
8.30 - 18.30	*Registration and poster hang-up*
9.15 - 9.30	*Welcome and Introduction*
9.30 - 10.10	**Scientific Session 1**
	Mobile applications driven by Open PHACTS semantic web technology *Christine Chichester, Lee Harald, Tim Harder*
	TagCurate: Crowdsourcing the verification of biomedical annotations to mobile users *Bahar Sateli, Sebastien Luong, René Witte*
10.10 - 11.00	**Invited Lecture**
	Facilitating Scientific Discovery through Crowdsourcing and Distributed Participation *Antony Williams* *Royal Society of Chemistry*
11.00 - 11.30	**Coffee Break**
11.30 - 12.10	**Scientific Session 2**
	iNGS: a prototype tool for genome interpretation and annotation *Ismael Navas-Delgado, Maria Jesús García Godoy, Fátima Arjona-Pulido, Trinidad Castillo-Castillo, Ana Isabel Ramos-Ostio, José F. Aldana-Montes*
	The OntoGene literature mining web service *Fabio Rinaldi*
12.10 - 13.10	**Poster Session**
13.10 - 14.30	**Lunch Break**
14.30 - 15.10	**Industrial - Technological Session**
	New publishing opportunities for digital life science *Rebecca Lawrence*
	Extracting more value from data silos: Using the semantic web to link chemistry and biology for innovation *Tim P Eyres*
15.10 - 15.30	**Introduction to open group discussions**
	Semantic, social & mobile technologies at the forefront of bioinformatics research *Speakers and Chairs propose the topics to be discussed by groups*
15.30 - 16.30	**Open group discussions**
	Semantic, social & mobile technologies at the forefront of bioinformatics research: selected topics *All participants discussing in the topic group of their greatest interest*
16.30 - 17.00	**Coffee Break**
17.00 - 17.50	**Invited Lecture**
	Semantic technologies for the automation of research in biomedicine *Ross D. King* *School of Computer Science, University of Manchester, Manchester, United Kingdom*
17.50 - 18.30	**Plenary reports of group discussions**
20.00 - 23.00	**Social Dinner at Nicelli Airport Restaurant, Lido of Venice**

Friday	October 18, 2013 WORKSHOP DAY 2
8.30 - 9.30	*Registration*
9.00 - 9.40	**Scientific Session 3**
	Sprints, Hackathons and Codefests as community gluons in computational biology *Steffen Möller, Enis Afgan, Michael Banck, Peter J. A. Cock, Matus Kalas, Laszlo Kajan, Pjotr Prins, Jacqueline Quinn, Olivier Sallou, Francesco Strozzi, Torsten Seemann, Andreas Tille, Roman Valls Guimera, Toshiaki Katayama, Brad Chapman*
	Taverna Mobile: Taverna workflows on Android *Hyde Zhang, Stian Soiland-Reyes, Carole Goble*
9.40 - 10.30	**Invited Lecture**
	SCIMOBS; the million minds approach revisited in mobile context *Barend Mons, Leiden University Medical Center, Leiden, and Netherlands Bioinformatics Center, The Netherlands*
10.30 - 11.00	**Coffee Break**
11.00 - 12.20	**Scientific Session 4**
	Bio-GraphIIn: a graph-based, integrative and semantically-enabled repository for life science experimental data *Alejandra Gonzalez-Beltran, Eamonn Maguire, Pavlos Georgiou, Susanna-Assunta Sansone, Philippe Rocca-Serra*
	An ontology based query engine for querying biological sequences *Martijn Devisscher, Tim De Meyer, Wim Van Criekinge, Peter Dawyndt*
	The representation of biomedical protocols *Larisa N. Soldatova, Ross D. King, Piyali S. Basu, Emma Haddi, Nigel Saunders*
	The role of parallelism, web services, and ontologies in bioinformatics and omics data management and analysis *Mario Cannataro, Pietro Hiram Guzzi*
12.20 - 13.00	**Closing session**
	Assignment of Best Poster Awards
	Announcement of NETTAB 2014
	Farewell

Keynote Lectures

Facilitating scientific discovery through crowdsourcing and distributed participation

Antony Williams
Royal Society of Chemistry

Antony Williams is the Vice President, Strategic Development at the Royal Society of Chemistry and leads their eScience efforts. He was one of the founders for ChemSpider, one of the world's primary internet resources for chemists now hosted by RSC and providing access to over 28 million unique chemicals linked out to over 400 data sources on the internet. He was the Chief Science Officer for Advanced Chemistry Development (ACD/Labs) focusing his cheminformatics skills on structure representation, nomenclature and analytical data handling. He is an NMR spectroscopist by training with a PhD from the University of London and is an accomplished author with over 140 publications, and many book chapters or books. He is known as the ChemConnector in the chemistry social network and is passionate about providing tools for data sharing, collaboration and alternative metrics for recognizing the contributions of scientists.

Science has evolved from the isolated individual tinkering in the lab, through the era of the "gentleman scientist" with his or her assistant(s), to group-based then expansive collaboration and now to an opportunity to collaborate with the world. With the advent of the internet the opportunity for crowd-sourced contribution and large-scale collaboration has exploded and, as a result, scientific discovery has been further enabled. The contributions of enormous open data sets, liberal licensing policies and innovative technologies for mining and linking these data has given rise to platforms that are beginning to deliver on the promise of semantic technologies and nanopublications, facilitated by the unprecedented computational resources available today, especially the increasing capabilities of handheld devices. The speaker will provide an overview of his experiences in developing a crowdsourced platform for chemists allowing for data deposition, annotation and validation. The challenges of mapping chemical and pharmacological data, especially in regards to data quality, will be discussed. The promise of distributed participation in data analysis is already in place.

Semantic technologies for the automation of research in biomedicine

Ross D. King
School of Computer Science, University of Manchester, Manchester, United Kingdom

Dr Ross King moved to the School of Computer Science, University of Manchester, in February 2012. Before that, he was at the University of Wales, Aberystwyth, for fifteen years. His first degree was in Microbiology, but he also has a M.Sc. and Ph.D. in Computer Science. The research achievement he is most proud of is developing Robot Scientists, a physical implementation of the task of Scientific Discovery in a microbiology laboratory, representing the merging of increasingly automated and remotely controllable laboratory equipment and knowledge discovery techniques from Artificial Intelligence.

The use of computers is changing the way that science is described and reported. Scientific knowledge is best expressed in formal logical languages with associated probabilities. Only formal languages provide sufficient semantic clarity to ensure reproducibility and the free exchange of scientific knowledge. Despite the advantages of logic, most scientific knowledge is expressed only in natural languages. This is now changing through developments such as the Semantic Web and ontologies.

A Robot Scientist is a physically implemented robotic system that applies techniques from artificial intelligence to execute cycles of automated scientific experimentation. A Robot Scientist can automatically execute cycles of hypothesis formation, selection of efficient experiments to discriminate between hypotheses, execution of experiments using laboratory automation equipment, and analysis of results. We have developed the Robot Scientists Adam (functional genomics), and Eve (drug design).

Robot Scientists provide excellent test-beds for the development of methodologies for formalising science. Using them it is possible to completely capture and digitally curate all aspects of the scientific process. We attempted to record and formalise Adam's experiments. For the core organization of this we used the ontology of scientific experiments EXPO. This ontology formalizes generic knowledge about experiments. We then developed LABORS, a customized version of EXPO. Application of LABORS produces experimental descriptions in the logic-programming language Datalog. In the course of its investigations, Adam observed 6,657,024 optical density (OD595nm) measurements (forming 26,495 growth curves). These data are held in a MySQL relational database. Use of LABORS resulted in a formalization of the scientific argumentation involving over 10,000 different research units (segments of experimental research). This has a nested treelike structure, 10 levels deep, that logically connects the experimental observations to the experimental metadata. This structure resembles the trace of a computer program and takes up 366 Mbytes. Making such experimental structures explicit renders scientific research more comprehensible, reproducible, and reusable.

My vision of the future is a collaboration between human and Robot Scientists will produce better science than either can alone, and the scientific knowledge produced will be primarily expressed in logic with associated probabilities and published using the Semantic Web.

SCIMOBS: the million minds approach revisited in mobile context

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

Barend Mons is Professor of Biosemantics at the LUMC and is one of the scientific directors of NBIC, the Netherlands Bioinformatics Center. 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 head of Node of the developing Dutch node in 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 and is an advisor for several companies as well. From the year 2000 onward he increasingly focuses on the development of semantic technologies to manage big data and he founded the Biosemantics groups, first at Erasmus University in Rotterdam and later also in Leiden. Both groups collaborate very closely.

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.

In this talk, he will revisit the 'need to engage a million minds in expert crowd sourcing' based on his highly cited 2008 paper "Mons et al. Calling on a million minds for community annotation in WikiProteins, Genome Biology 2008, 9:R89 ".

Now that mobile technology is so much further in development and social acceptance, and the first real scientific applications are coming on the market, there is a need to revisit how best to engage people in expert crowdsourcing.

Tutorials

Semantic Web for Life Sciences: vision, aims, tools, platforms

Andrea Splendiani
IntelLeaf, United Kingdom, and Digital Enterprise Research Institute, Ireland

Andrea Splendiani is an independent professional (intelliLeaf ltd) and adjunct lecturer at DERI.

He earned a Laurea Degree in Information Technology from the Politecnico di Milano, and a PhD in Computer Science from the University of Milano-Bicocca. He has been working in functional genomics for immunology (Genopolis) and Systems Biology (Institut Pasteur) where he has been an active contributor to the BioPAX standard. He has then worked on biomedical ontologies (University of Rennes 1) and data integration (Rothamsted Research, BBSRC).

He is co-chair and organizer of SWAT4LS (Semantic Web Applications and Tools for Life Sciences) and Guest Editor of the Journal of Biomedical Semantics.

He is currently collaborating with different companies on data modeling for healthcare and life sciences, as well as semantic solutions for knowledge and data management.

This tutorial is an introduction to Semantic Web technologies for Bioinformaticians, Researchers and other professionals working on biomedical and life sciences data. It is composed of two parts: an introduction to technologies and tools, and an overview of how these technologies fit in the current bioinformatics landscape.

The first part will briefly introduce key concepts and technologies (RDF, OWL, Rules, SKOS, SPARQL, Linked-Data, Triplestores) and then present tools and techniques to address two prototypical use cases: publishing information on the Semantic Web, and consuming information that can be found in RDF.

The second part of the tutorial will present an overview of the state of adoption of Semantic Web technologies in bioinformatics (which databases publish information in RDF and different representation patterns). It will also present features and limitations that characterize these technologies, to provide an understanding of which data and applications they can practically benefit.

Perspective participants are invited to get in contact with the presenter so that the content of the tutorial can be tuned to their specific interests. See his LinkedIn page.

Open_PHACTS and NanoPublications

Christine Chichester
Swiss Institute of Bioinformatics, CALIPHO group, Geneva, Switzerland

Christine Chichester belongs to the CALIPHO group at the Swiss Institute of Bioinformatics and is the co-leader of the work package dealing with the data in the Open PHACTS platform. Open PHACTS project is a unique public-private partnership, established by the Innovative Medicines Initiative (IMI), between the European Community European Federation of Pharmaceutical Industries and Associations (EFPIA). The goal is to provide a platform that will support key drug discovery tasks by the integration of pharmacological and other biomedical data using open standards such as RDF.

Christine earned her PhD in Molecular Pharmacology (University of California, Davis) and has since expanded her skills by working many years in the Bioinformatics domain. She is now attempting to establish discussions between the Bioinformatics and Semantic Web communities principally through data modeling and data integration projects.

This tutorial is an introduction to the Open PHACTS platform and API tailored to the wide community of potential data providers, data consumers, and service providers. The tutorial will include an overview of the Open PHACTS architecture, data sources, and example applications followed by a hands-on session detailing the use of the API.

The session will include discussions concerned with signing up for a developer API key, making API calls, and the data returned. Secondly, the tutorial will also include some discussion on the modeling of scientific assertions as nanopublications and their use within the Open PHACTS platform.

Standards for Web Applications on Mobile: current state and roadmap

Dominique Hazaël-Massieux
W3C/ERCIM, Sophia Antipolis, Biot, France

Dominique Hazaël-Massieux supervises the development and standardization of Web technologies that are most relevant to mobile devices in W3C, and is in charge of the W3C groups that are developing APIs to access more device capabilities from the Web (camera, addressbook, etc) and to enable peer-to-peer audio-video communication in Web browsers (WebRTC).

He also regularly puts in practice these technologies and guidelines as a developer of a number of sites and applications.

Web technologies have become powerful enough that they are used to build full-featured applications; this has been true for many years in the desktop and laptop computer realm, but is increasingly so on mobile devices as well.

This tutorial will introduce the various technologies developed in W3C that increase the capabilities of Web applications, and how they apply more specifically to the mobile context.

Mobile applications for life sciences: perspectives, limitations, and real examples

Alex Clark
Molecular Materials Informatics, Inc

Alex Clark is the founder of Molecular Materials Informatics, which is dedicated to bringing cheminformatics software to future platforms such as mobile, cloud & web.

Since 2010 the company has produced a number of cutting edge apps and cloud-based webservices for manupulating chemical structures for iOS and Android, and is able to demonstrate powerful chemistry and life sciences workflows for the post-PC era.

The tutorial will demonstrate manipulation of chemical structures and associated data using the mobile platform. Apps can be used to create or import content and organise it on the device, and from there it can be browsed, visualised, modified, shared and published, as well as for searching online databases, building models and calculating properties.

The union of gesture-based mobile interfaces with cloud-hosted webservices can be used to accomplish a broad variety of tasks that were formerly only practical for the desktop platform.

The new computing platform provides a number of advantages in addition to mobility, such as inexpensive modular components, shorter learning curves, and ultimate ease of deployment based on the "app store" model.

Demonstrations will be performed using the iOS platform (iPhone/iPod/iPad), and focus on real-world problems encountered in chemical informatics and computer-aided drug design.

Oral Communications

Mobile applications driven by Open PHACTS semantic web technology

Christine Chichester[1]✉, **Lee Harald**[2], **Tim Harder**[3]

[1]Swiss Institute of Bioinformatics, CALIPHO group, University of Geneva, Geneva, Switzerland
[2]Connected Discovery Ltd., London, United Kingdom
[3]Center for Bioinformatics, University of Hamburg, Hamburg, Germany

Received 1 August 2013; **Accepted** 8 August 2013; **Published** 14 October 2013

Competing interests: the authors have declared that no competing interests exist.

Abstract

The Open PHACTS project is a large-scale public-private partnership funded under the European Innovative Medicines Initiative. The aim of the project is to create a stable infrastructure that combines diverse biomedical databases and enables scientists to answer complex questions of relevance to drug discovery and human health. The data integration is achieved through semantic web technologies, with data modeled in RDF and served through a high performance triple store. The main point of interaction with the system is an API, designed to be "developer friendly" by using familiar web technologies such as REST services and JSON. The vision is the Open PHACTS API becomes the foundation of an "application ecosystem", enabling anyone to create applications targeting different use-cases within this domain. Here, we present two examples of the use of the Open PHACTS API to create mobile scientific applications. The ChemBioNavigator is a tool that allows researchers to analyze and triage sets of bioactive small molecules. Physiochemical properties are used to separate the molecules into different areas of a graph, upon which biological information from the Open PHACTS database is overlaid. This visualization allows users to understand the relationship between these data and select the molecules with the best overall combinations of these properties. The second example is the iPharm application which is designed to address a common situation where a researcher needs a high-level overview on a drug or target. The user simply opens up the application, enters a drug name and is presented with key clinical and molecular data summarizing the action of that drug. An important aspect of both applications is that while they are based on semantic technologies, the user interfaces are specifically designed to look like "normal" applications. We demonstrate the combination of semantic data with natural user interfaces provides a powerful mechanism to address scientific data navigation.

Motivation and Objectives

The Open PHACTS project is a large-scale public-private partnership funded under the European Innovative Medicines Initiative (IMI). The aim of the project is to create a stable infrastructure that combines diverse biomedical databases and enables scientists to answer complex questions of relevance to drug discovery and human health. The data integration is achieved through semantic web technologies, with data modeled in RDF and served through a high performance triple store. The main point of interaction with the system is an API, designed to be "developer friendly" by using familiar web technologies such as REST services and JSON. The vision is the Open PHACTS API becomes the foundation of an "application ecosystem", enabling anyone to create applications targeting different use-cases within this domain. Here, we present two examples of the use of the Open PHACTS API to create mobile scientific applications. The ChemBioNavigator is a tool that allows researchers to analyze and triage sets of bioactive small molecules. Physiochemical properties are used to separate the molecules into different areas of a graph, upon which biological information from the Open PHACTS database is overlaid. This visualization allows users to understand the relationship between these data and select the molecules with the best overall combinations of these properties. The second example is the iPharm application which is designed to address a common situation where a researcher needs a high-level overview on a drug or target. The user simply opens up the application, enters a drug name and is presented with key clinical and molecular data summarizing the action of that drug. An important aspect of both applications is that while they are based on semantic technologies, the user interfaces are specifically designed to look like "normal" applications. We demonstrate the combination of semantic data with natural user interfaces provides a powerful mechanism to address scientific data navigation.

Methods

The mobile applications presented in this work are based on the data retrieved from the Open PHACTS platform, which is essentially a semantic data integration platform. This semantic approach to data integration has been pioneered by other research efforts in the biomedical domain (Belleau *et al,* 2008; Chen *et al*, 2010; Hardy *et al*, 2010; Hassanzadeh *et al*., 2009), as

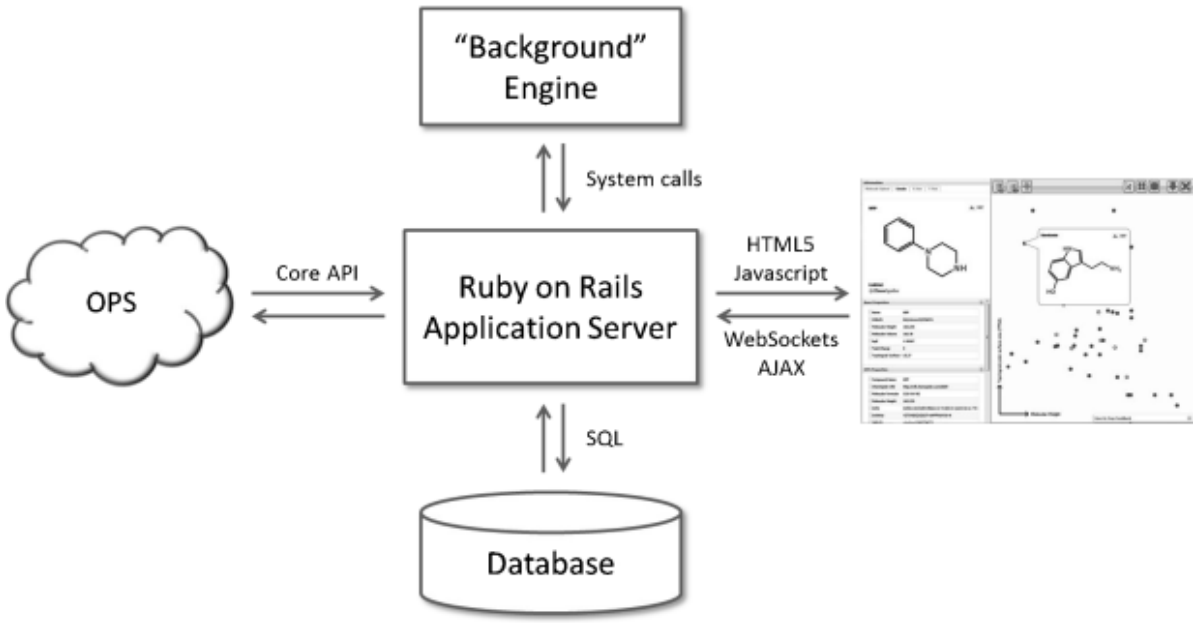


Figure 1. Schematic overview of the CBN technical architecture (Stierand *et al.*, 2012). The central application server connects the different parts of the system and handles the client communication.

well as by LinkedLifeData[1] and the World Wide Web Consortium Semantic Web Health Care and Life Sciences Interest Group[2]. The platform provides a semantically integrated view of the data by exploiting an identity mapping service (IMS) to construct appropriate responses based on the contextual aspects of each query. For instance, will users differentiate to the same level of granularity? *I.e.* do they want results for both genes and proteins. The platform also contributes to performance by caching the semantic representations of frequently used datasets. Additionally, an important aspect of the platform is the ability to integrate datasets and services; for example, the integration of the ConceptWiki for resolving textual queries and the possibility for community curation.

Although the behind the scenes semantic technology is required to provide the data, very important to creating the mobile applications is the rich Open PHACTS API. This interface provides REST-style based interfaces for common queries. These queries are defined in agile cycles working with application and user interface developers. By driving API development from the user perspective, Open PHACTS reduces the complexity of data integration by focusing on those core semantic types and properties that are necessary for the end-users. A key difficulty in many linked data solutions today has been the threat to performance caused by difficult or challenging queries and it is not yet clear in many current prototypic applications how to deal with complexities in the representation of facts in RDF. By adopting a well-defined API, the SPARQL queries can be optimised to ensure that such performance difficulties do not occur. Finally, the API provides rich information and access to the provenance of the results it returns. This includes the data sources from which those results are provided including the ability to track licensing information. Currently, the platforms includes datasets from: DrugBank[3], ChEMBL[4], SwissProt/UniProt[5], ChEBI[6], Gene Ontology[7], GOA[8], Wikipathways[9], ChemSpider[10] and ConceptWiki[11].

1 http://linkedlifedata.com/
2 http://www.w3.org/blog/hcls/
3 http://www.drugbank.ca/
4 https://www.ebi.ac.uk/chembl/
5 http://www.uniprot.org/
6 http://www.ebi.ac.uk/chebi/
7 http://www.geneontology.org/
8 http://www.ebi.ac.uk/GOA
9 http://wikipathways.org/
10 http://www.chemspider.com/
11 http://ops.conceptwiki.org/

Technically, the main challenges for the mobile applications using Open PHACTS is to design an architecture that is able to handle remote querying of the platform and at the same time provide an intuitive and responsive user interface. Further, the applications should be flexible enough to be able to adapt to possible future request such as additional data sources like proprietary algorithms or databases. Specifically for the ChemBioNavigator (CBN) example, the architecture is designed around a Ruby on Rails web-application server, connecting the different parts of the application (Figure 1) and modern web technologies such as HTML5, JavaScript and AJAX/JSON.

Results and Discussion

Delivering an API for building applications based on the Open PHACTS platform has resulted in much knowledge into the technical challenges concerned with the large amounts of data, the diverse and often non-standard formats along with the use of different biological identifiers in most data sources and what is needed for ease of adoption for developers. Clearly, these results: developing a deeper understanding of the application of Semantic Web technologies to the life sciences, the integration and mapping of disparate data types and sources, the influence of data quality, the crowd-sourcing opportunities and as well as the processes and approaches necessary to deliver a groundbreaking technology platform to application developers are results that will help refine and augment the Open PHACTS platform going forward.

Applying modern web technologies in the development of mobile applications allows the seamless integration across different platforms and even device types as shown with CBN and iPharm. These technologies accommodate the growing amount of tablet computers and other mobile devices for easy access to the necessary web services and enable the average bench scientist to easily explore large amounts of data, augmenting their research results with new information. The mobile applications can thus make the information accessible wherever needed, at the office computer or the tablet in a meeting or at the lab bench.

Acknowledgements

This work was supported by the Innovative Medicines Initiative Joint Undertaking under grant agreement n° 115191 for Open PHACTS, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in kind contribution. We greatly appreciate the efforts from all the partners in the entire Open PHACTS consortium.

References

Belleau F, *et al.* (2008) Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *J. Biomed. Inform.* **41**, 706–716. doi:10.1016/j.jbi.2008.03.004

Chen B, *et al.* (2010) Chem2Bio2RDF: A semantic framework for linking and datamining chemogenomic and systems chemical biology data. *BMC Bioinformatics* 11, 255. doi:10.1186/1471-2105-11-255

Hardy B, *et al.* (2010) Collaborative development of predictive toxicology applications. *J. Cheminform.* **2**, 7. doi:10.1186/1758-2946-2-7

Hassanzadeh O, *et al.* (2009) LinkedCT: A linked data space for clinical trials. http://arxiv.org/abs/0908.0567

Stierand K, Harder T, *et al.* (2012) The Internet as Scientific Knowledge Base: Navigating the Chem-Bio Space. *Molecular Informatics* Special Issue: Open Innovation in Drug Discovery **31** (8), 543-546. doi:10.1002/minf.201200037

TagCurate: crowdsourcing the verification of biomedical annotations to mobile users

Bahar Sateli, Sebastien Luong, René Witte✉
Concordia University,Montréal , Canada

Received 7 August 2013; **Accepted** 12 September 2013; **Published** 14 October 2013

Competing interests: the authors have declared that no competing interests exist.

Abstract

We present TagCurate, a distributed system that allows for disseminating biomedical annotations to users on Android-enabled devices for further verification. A web-based interface provides Task Managers with the ability to supervise the crowdsourcing process, as well as viewing the results gathered from the TagCurate Android app installed on the crowd's devices. We believe that the results of this research is beneficial to both curators and the NLP development communities. The efforts of expert curators will be efficiently allocated to resolving controversial annotations, while NLP pipeline developers can further train their algorithms from gold standard corpora solicited from a large group of contributors.

Motivation and Objectives

Robust, automatic approaches are being developed by industrial and academic communities specifically targeting the well-known problem of information overload, caused by the overabundance of available scholarly publications. State-of-the-art approaches, in particular from the computational linguistics and Natural Language Processing (NLP) domains, aim at aiding the laborious task of manually extracting structured knowledge from the unstructured free-style text found in scientific publications. However, the inherent complexity of natural languages used by researchers in communicating their findings makes the knowledge extraction an intricate task in need of human verification to produce effective results. Based on the division of labor principle, we are proposing a novel system to *crowdsource* the verification of automatically extracted annotations from biomedical literature to mobile users. The hypothesis behind our research is that providing a synchronised, distributed system for human verification of annotations generated during the literature curation process helps to decrease the time needed to accomplish the task, while improving the curators' productivity by providing an ubiquitous environment available both in the web and mobile context.

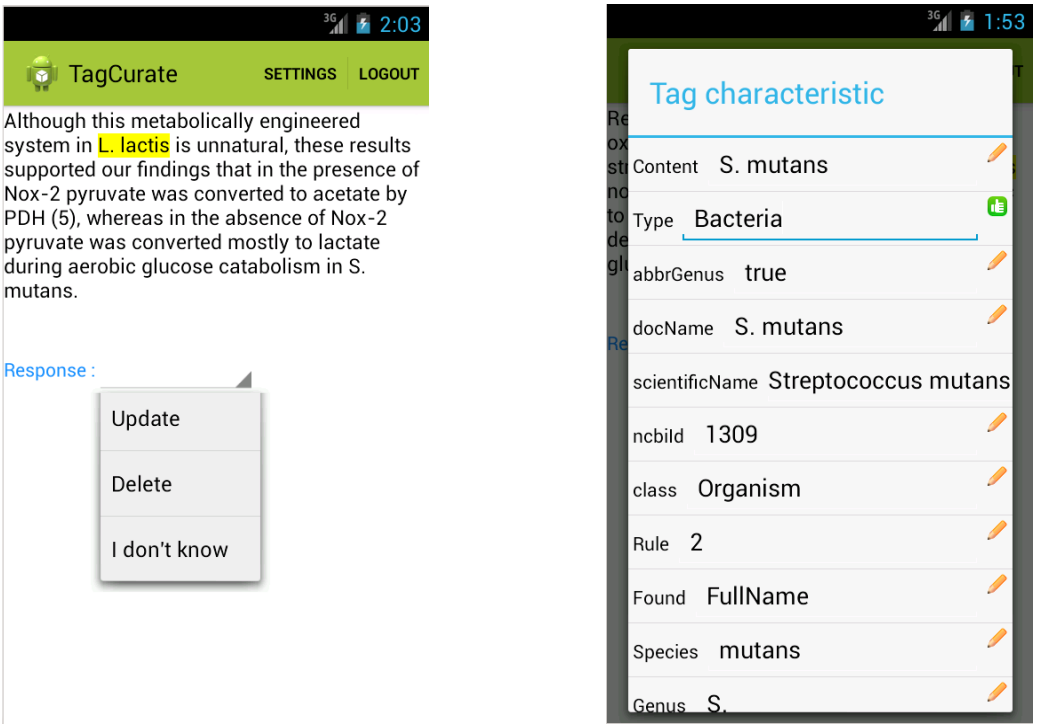
We present TagCurate, a distributed system that allows for disseminating biomedical annotations to users on Android-enabled devices for further verification. A web-based interface provides *Task Managers* with the ability to supervise the crowdsourcing process, as well as viewing the results gathered from the TagCurate Android app installed on the crowd's devices. We believe that the results of this research is beneficial to both curators and the NLP development communities. The efforts of expert curators will be efficiently allocated to resolving controversial annotations, while NLP pipeline developers can further train their algorithms from gold standard corpora solicited from a large group of contributors.

Methods

The TagCurate system is composed of a server-side component responsible for distributing and managing annotations and an Android app through which users verify the annotations assigned to them. Conforming to a client/server model, both components communicate over the HTTP protocol through a message passing mechanism.

As an extension to the Semantic Assistants framework (Witte and Gitzinger, 2008), the server-side component is implemented using the J2EE Servlet technology and provides a RESTful endpoint to interact with the TagCurate Android app. Featuring a web-based user interface, the TagCurate system allows so-called Task Managers to define a verification task by uploading annotated documents, provided that they have been annotated by either NLP pipelines or human annotators based on the General Architecture for Text Engineering (GATE) framework[1] (Cunningham *et al.*, 2011). The TagCurate system then generates an internal representation of the existing annotations, in form of an XML document, that are ultimately distributed to the TagCurate Android apps installed on the crowd's devices.

1 http://gate.ac.uk/



(a) *(b)*

Figure 1. TagCurate Android App annotation verification (a) and modification (b) activities.

The TagCurate Android app, designed based on the Android 4.3 Jelly Bean API, allows users to authenticate themselves on the server-side component and *pull* annotations that are assigned to them for verification. Through an interactive interface, users can view each annotation in the context of the sentence that they appear, in order to determine whether the annotation is correctly tagged. A long-click on each annotation allows users to view features, i.e., further information provided in the annotation representation, like the scientific name of an organism (Naderi *et al.*, 2011), where applicable. Based on the available information, users can then decide whether the annotation is entirely correct or should be removed from the document in case of a false positive (Figure 1-a). In addition, if an annotation is partially correct, e.g., if only some features are wrong or the character offset (span) of the annotation in the text needs to be changed, users can directly edit the annotation features (Figure 1-b) and submit an updated representation to the server-side, where it is made persistent.

The gathered feedback from the crowd is aggregated on the TagCurate server-side component that can provide overview reports of the crowdsourcing progress, as well as reporting annotations with high percentage of disagreement between users, once a specified threshold is passed.

Results and Discussion

TagCurate is the first open source project that targets the distributed verification of (biomedical) annotations to mobile users. Rather than relying on expert annotators only, it is now possible to distribute document annotations, whether they are manually created or computed by an NLP pipeline, to a large user base – for example, students in a university setting. TagCurate will soon be available both as open source software and on the Android "Google Play" market for direct installation.

In future research, we will apply our mobile crowdsourcing platform to a large-scale annotation task. In particular, we will investigate incentives for mobile users to participate in verification tasks and analyse the quality differences that can be obtained from mobile users with varying backgrounds vis-à-vis expert curators. In addi-

tion to the verification task, we are also planning to provide mobile users with the ability to automatically annotate the documents using our Semantic Assistants Android Open Intents (Sateli *et al.*, 2013) that allows for remote execution of NLP pipelines on provided content, thereby enabling users to perform the complete literature curation task entirely through a mobile interface.

References

Cunningham H, Maynard D, et al. (2011) *Text Processing with GATE (Version 6)*, University of Sheffield, Department of Computer Science. 15 April 2011. ISBN 0956599311.

Sateli B, Cook G, and Witte R (2013) Smarter Mobile Apps through Integrated Natural Language Processing Services. In *10th International Conference on Mobile Web Information Systems (MobiWIS 2013)*, Paphos, Cyprus, Springer Lecture Notes on Computer Science LNCS 8093, pp. 187--202. August 26--28, 2013, doi:10.1007/978-3-642-40276-0_15

Naderi, N, Kappler T, Baker CJO, and Witte, R (2011) OrganismTagger: Detection, normalization, and grounding of organism entities in biomedical documents. *Bioinformatics* **27**(19), 2721-2729. doi:10.1093/bioinformatics/btr452

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 Lecture Notes on Computer Sciences LNCS* **5367**, 360–374. doi:10.1007/978-3-540-89704-0_25

iNGS: a prototype tool for genome interpretation and annotation

Ismael Navas-Delgado✉, **Maria Jesús García Godoy, Fátima Arjona-Pulido, Trinidad Castillo-Castillo, Ana Isabel Ramos-Ostio, Sarai Infantes Díaz, Ana Medina García, José F. Aldana-Montes**

University of Málaga, Spain

Received 31 July 2013; **Accepted** 10 September 2013; **Published** 14 October 2013

Competing interests: the authors have declared that no competing interests exist.

Abstract

Currently, clinical interpretation of whole-genome NGS genetic findings are very low-throughput because of a lack of computational tools/software. The current bottleneck of whole-genome and whole-exome sequencing projects is in structured data management and sophisticated computational analysis of experimental data. In this work, we have started designing a platform for integrating, in a first step, existing analysis tools and adding annotations from public databases to the findings of these tools. This platform can be used to produce tools for different kind of users. As a first experiment with this platform, we have developed a Web tools for running multiple analysis tasks, completing the findings with public data and producing a simple report similar to blood test reports.

Motivation and Objectives

Personalised medicine can be considered as a medical approach which proposes the customisation of healthcare involving medical decisions, treatments, etc., that are applied to a patient individually and tailored to that patient. This medical methodology is possible due to the rapid advances in technology in areas such as genomics, transcriptomics, proteomics, metabolomics, etc. In this context, it is important to mention that the development of sequencing approaches of personal human genomes and the detection of DNA variations by means of a reference human genome that was unveiled in 2003-2004 (Human Genome Sequencing Consortium International, 2004) are both huge contributions that should be integrated into the personal "omics" (in this case, genomics) of each patient.

Nonetheless, clinical interpretation of whole-genome and NGS (Next Generation Sequencing) genetic findings are currently very low-throughput because of a lack of computational tools/software to integrate all this information. In this sense, the reason for the current bottleneck of whole-genome and whole-exome sequencing projects is the management of structured data and sophisticated computational analysis of the experimental data obtained.

Therefore, we have started designing a platform for integrating, firstly, existing genome analysis tools and then adding more annotations than those currently provided from the findings of these tools in public databases. As a first experiment with this platform, we have developed iNGS[1], a Web tool for running multiple analysis tasks. All findings of these analysis tools are completed with public data to generate a simple report to complete information provided by genome interpretation and annotation tool results.

Method

In the context of tools for analysing the whole-genome, there are many available tools that are able to provide users with NGS genetic findings. Some of these tools, described in more detail below are Annovar (Wang et al., 2010), GATK (McKenna et al., 2010), SeatleSeq (Ng et al., 2009), VAAST (Yandell et al., 2011) and Galaxy (Goecks et al., 2010).

Annovar is a command-line tool widely used to annotate functional effects of variants with respect to genes, genomic region-based annotations (which refer to those regions that are different to genes) and compare variants with those variations stored in databases. Furthermore, an interesting feature of this tool is the detection of diseases associated with the regional annotations of GWAS[2] (Genome-Wide Association Studies) catalogue.

GATK is a Genome Analysis Toolkit that presents five main functionalities: 1) initial read mapping; 2) local realignment around INDELs (insertions or deletions); 3) base quality score recalibration; 4) SNP (single-nucleotide polymorphism) discovery and genotyping to find all potential variants; and 5) machine learning to separate true segregat-

1 http://khaos.uma.es/iNGS
2 http://www.genome.gov/gwastudies/

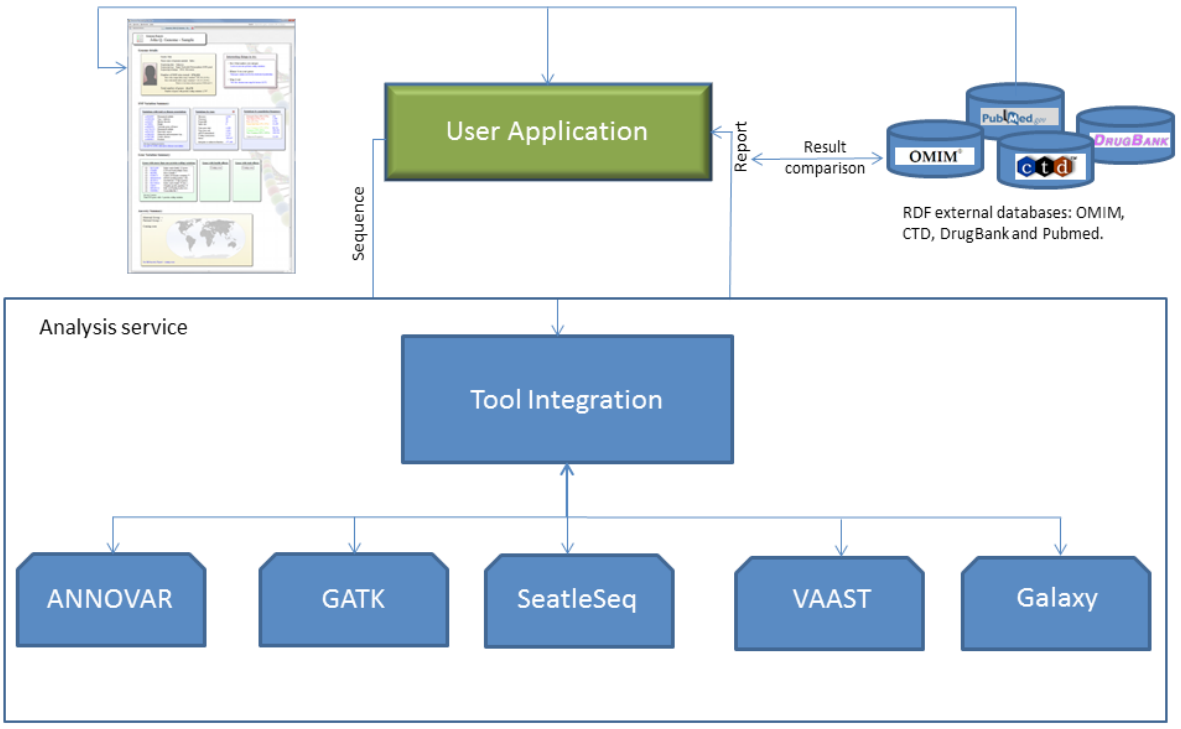


Figure 1. Platform and Pilot User Tool.

ing variation from machine artefacts common to NGS technologies.

SeatleSeq Annotation provides annotations of known and novel SNPs. The annotations include dbSNP (rs IDs), gene names and accession numbers, SNP functions, protein positions, amino-acid changes, conservation scores, HapMap frequencies, PolyPhen predictions and clinical associations. Furthermore, this tool has been tested to identify rare and common variants in over 300 megabases (Mb) of coding sequence using exomes from Freeman-Sheldon syndrome patients.

VAAST is a probabilistic tool that has been designed to identify damaged genes and their disease-causing variations in personal genome sequences. This tool has been benchmarked in multiples studies, which include 100 Mendelian conditions (Ng et al., 2009) and also the identification of genes responsible for common diseases (Lesage et al, 2002).

Galaxy is a framework that provides a set of web-based tools including different analysis variation tasks. This tool is used to look for disease SNPs in a full genome, to detect SNPs differing between populations, to look for disease-associated SNPs in a pedigree and for population structures and selective sweeps.

Besides the aforementioned tools, there are many databases that provide up-to-date findings that can help to provide interpretations for the genome analysis of these tools. Examples of such findings could be DNA variants related with diseases, genes, bibliography, drugs and drug targets. This information is publicly available in different data sources such as OMIM (Hamosh et al., 2005), CTD (Mattingly et al., 2003), DrugBank (Wishart et al., 2006) and PubMed (Roberts, 2001). Furthermore, as Figure 1 shows, the output files obtained contain information on DNA variants detected by the genotype analysis that can be completed with information from different repositories. To do this, we have selected a set

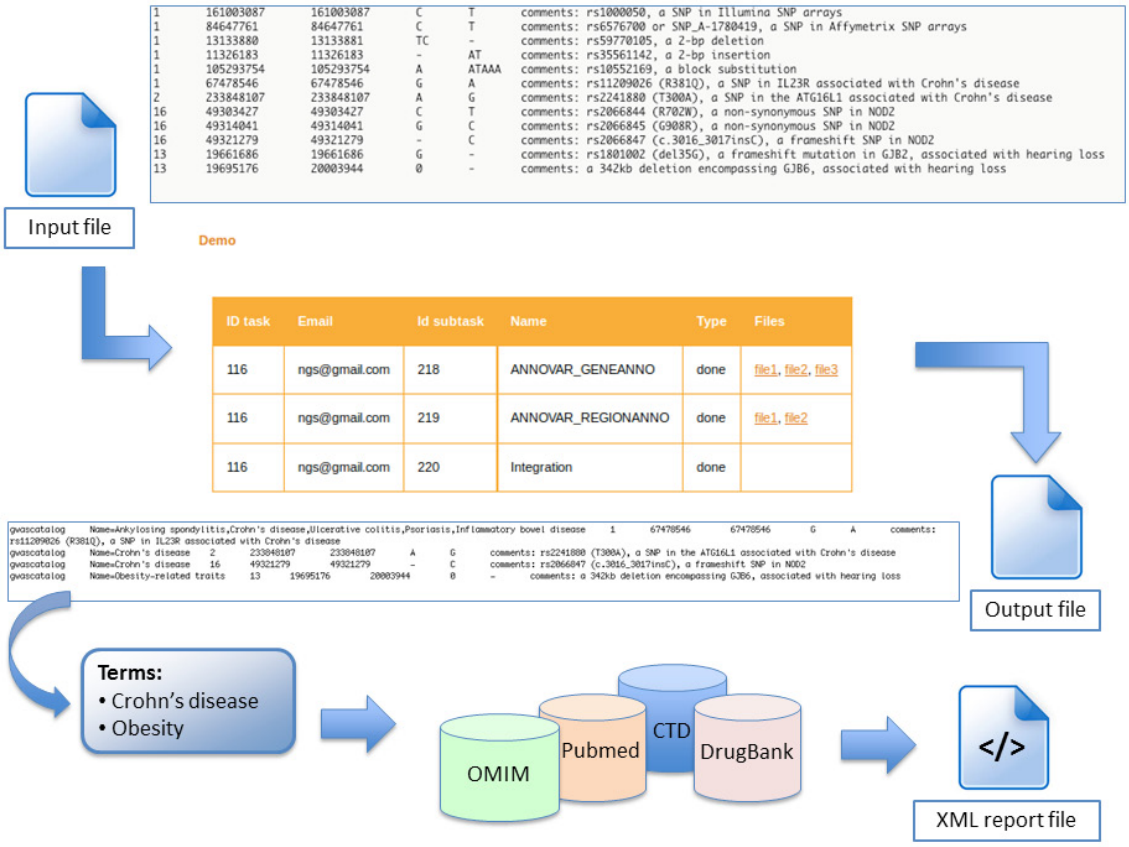


Figure 2. General scheme of how the genome-analysis tool platform works. The user introduces an input file (e.g. a whole-exome sequence of a patient). The analysis tools previously selected by the user process the file generating an output file. The interesting terms (e.g., disease and gene annotations) written in the output file are extracted automatically from the output file and included as parameters in a parameterized SPARQL query. The information retrieved is included in a report in XML format to complete the information obtained from the genomic analysis tools. The user with a biological or medical background can interpret this report.

of RDF (Resource Data Framework) data sources, information of which is stored following the principles of Open Linked Data.

OMIM endpoint provided by Bio2RDF server[3], contains data classified into different classes (such as allelic variant, clinical synopsis, gene, phenotypes gene-phenotypes etc.) that can be useful for phenotype-genotype analysis.

CTD endpoint[4] provides information on annotated associations between genes, diseases, proteins, bibliographic references and toxic agents.

DrugBank[5] is a Bio2RDF endpoint the data structure of which is divided into several interesting classes such as drug-drug, drug-target, drug-enzyme and drug-transporter interactions.

We have also included the PubMed repository[6] which is an important source of bibliographic data. Furthermore, it is worth mentioning that PubMed resources are mostly associated with other PubMed resources stored in OMIM, CTD and DrugBank endpoints.

The availability of these databases allows a platform to be designed that integrates analysis tools and public data, to be used by end-users without requiring them to install and configure complex software packages. Figure 1 depicts a general view of the proposed platform and how it can be used to develop a tool for combining analysis tools and public data. In this case, the tool presented aims to provide simple to understand reports. The analysis processes will run on

3 http://s4.semanticscience.org:16019/sparql
4 http://s4.semanticscience.org:16004/sparql
5 http://s4.semanticscience.org:16006/sparql
6 http://pubmed.bio2rdf.org/sparql

the served side, freeing the user from having to rely on computational resources.

The first prototype built using this platform will run several analysis tools in parallel, generating a set of DNA annotations results: element variants in genes (either exonic and intronic regions) and intergenic regions and their relationships with diseases according to GWAS studies, transcription binding site annotations, DNA variation annotations that fall within conserved genomic regions etc. Then, these results will be complemented with information retrieved from biological endpoints through a set of SPARQL queries. To do this, the outputs of the analysis tools (e.g., annotated diseases, gene names, segmental duplication annotations, etc.) will be used to retrieve information from the data sources to complete the analysis findings of the platform tools. Figure 2 shows how this platform works.

In this regard, this work attempts to ease the task of understanding DNA variants such as short and large INDELs, SNPs, especially disease-associated SNPs, large-scale rearrangements, CNVs (copy number variation), and their functional consequences in target genome sequence(s) by completing these findings with a report containing information on genotypes, phenotypes, diseases, bibliography, drugs and drug targets.

Results and Discussion

To test the functionality of the prototype tool, we have selected an input (CEU.low _ coverage.2010 _ 07.indel.sites) in VCF format from the 1000 genomes repository[7]. This input file was loaded into the analysis tool platform. The region-based annotation analysis using GWAS catalogue demonstrated that CEU population contains a 1bp insertion (C) in exonic region of NOD gene that is related with Crohn's disease (among other relevant DNA mutations). The NOD genes and Crohn's disease terms were extracted automatically from the output files and were included as parameters in a previous parameterized SPARQL query to retrieve information from GWAS, OMIM, DrugBank and Pubmed data in order to complete the information obtained by the analysis tools of iNGS. The data retrieved contain phenotype, genetic, bibliographic and pharmacogenomic information related with the extracted terms. Such information is included in an XML report to be visualised by the user such as that shown in the schema in Figure 2. This task represented a challenge due to the lack of standards in this context. Regarding automating the term to be extracted from each output file (e.g. disease or gene annotations), it was necessary to analyze how the annotations detected by each analysis tool are structured. This implies that the annotation structure of the outputs of each tool integrated in the iNGS platform had to be analysed according to the different output file formats (e.g., Annovar has its own output and input file formats).

The aforementioned use case indicates that this platform could be useful for the analysis of genome sequence inputs. The generated reports are completed with integrated information from different data sources. As mentioned in the Method section, such information is retrieved through these SPARQL queries. At this point it is important to mention that another important challenge was the design of this set of SPARQL queries. Initially, we have designed a set of queries that retrieves non-redundant and repetitive information. In this way, the information is represented in an XML report making it more comprehensible for users.

For future work, we are planning to integrate a query federation system in order to efficiently retrieve information from more than one data source. Moreover, we are considering integrating other DNA variation analysis tools that focus on the prediction of the impact of DNA variations at DNA level the use of which is much extended throughout the user community in the field of personalised medicine. Finally, once tested by users we hope to open up this project to involve more developers in the production of an open platform.

7 ftp://ftp.sanger.ac.uk/pub/1000genomes/PILOT/REL-1007/indels/

Acknowledgements

The Project Grant [TIN2011-25840) (Spanish Ministry of Education and Science) and P11-TIC-7529 (Innovation, Science and Enterprise Ministry of the regional government of the Junta de Andalucía).

References

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**(8), R86. doi: 10.1186/gb-2010-11-8-r86

Hamosh A, Scott AF, *et al.* (2005) Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**(Database issue), D514-517. doi:10.1093/nar/gki033

Lesage S, Zouali H, *et al.* (2002) CARD15/NOD2 mutational analysis and genotype-phenotype correlation in 612 patients with inflammatory bowel disease. *Am J Hum Genet* **70**(4), 845-857. doi:10.1086/339432

Mattingly CJ, Colby GT, *et al.* (2003) The Comparative Toxicogenomics Database (CTD) *Environ Health Perspect.* **111**(6), 793-795. doi:10.1289/txg.6028

McKenna A, Hanna M, *et al.* (2010) The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**(9), 1297-1303. doi: 10.1101/gr.107524.110

Ng SB, Buckingham KJ, *et al.* (2009) Exome sequencing identifies the cause of a mendelian disorder. *Nature Genetics* **42**(1), 30-35. doi: 10.1038/ng.499

Ng SB, Turner EH, *et al.* (2009) Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**(7261), 272-276. doi:10.1038/nature08250

Roberts RJ. (2001) PubMed central: The GenBank of the published literature. *Proc. Natl. Acad. Sci. U. S. A.* **98**(2), 381-382. doi: 10.1073/pnas.98.2.381

Wang K, Li M, *et al.* (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**(16):e164. doi: 10.1093/nar/gkq603

Wishart DS, Knox C, *et al.* (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **34**(Database issue), D668-D672. doi: 10.1093/nar/gkj067

Yandell M, Huff C, *et al.* (2011) A probabilistic disease-gene finder for personal genomes. *Genome Res* **21**, 1529–1542. doi: 10.1101/gr.123158.111

The OntoGene literature mining web service

Fabio Rinaldi

University of Zurich, Switzerland

Received 1 August 2013; **Accepted** 10 September 2013; **Published** 14 October 2013

Competing interests: the authors have declared that no competing interests exist.

Abstract

Text mining technologies are increasingly providing an effective response to the growing demand for faster access to the vast amounts of information hidden in the literature. Several tools are becoming available which offer the capability to mine the literature for specific information, such as for example protein-protein interactions or drug-disease relationships. The biomedical text mining community regularly verifies the progress of such systems through competitive evaluations, such as BioCreative, BioNLP, i2b2, CALBC, CLEF-ER, BioASQ, etc. The OntoGene system is a text mining system which specialises in the detection of entities and relationships from selected categories, such as proteins, genes, drugs, diseases, chemicals. The quality of the system has been tested several times through participation in some of the community-organised evaluation campaigns. In order to make the advanced text mining capabilities of the OntoGene system more widely accessible without the burden of installation of complex software, we are setting up a web service that will allow any remote user to submit arbitrary documents. The results of the mining service (entities and relationships) are then delivered back to the user as XML data, or optionally can be inspected via a flexible web interface.

Motivation and Objectives

Text mining technologies are increasingly providing an effective response to the growing demand for faster access to the vast amounts of information hidden in the literature. Several tools are becoming available which offer the capability to mine the literature for specific information, such as for example protein-protein interactions or drug-disease relationships. The biomedical text mining community regularly verifies the progress of such systems through competitive evaluations, such as BioCreative, BioNLP, i2b2, CALBC, CLEF-ER, BioASQ, etc.

The OntoGene system is a text mining system which specializes in the detection of entities and relationships from selected categories, such as proteins, genes, drugs, diseases, chemicals. The quality of the system has been tested several times through participation in some of the community-organized evaluation campaigns.

In order to make the advanced text mining capabilities of the OntoGene system more widely accessible without the burden of installation of complex software, we are setting up a web service which will allow any remote user to submit arbitrary documents. The results of the mining service (entities and relationships) are then delivered back to the user as XML data, or optionally can be inspected via a flexible web interface.

Methods

The text mining pipeline which constitutes the core of the OntoGene system has been described previously in a number of publications (Rinaldi, 2008; Rinaldi, 2010; Rinaldi, 2012). We will only briefly describe the core text mining technologies, and instead focus mainly on the novel web service which allows remote access to the OntoGene text mining capabilities.

The first step in order to process a collection of biomedical literature consists in the annotation of names of relevant domain entities in biomedical literature (currently the systems considers proteins, genes, species, experimental methods, cell lines, chemicals, drugs and diseases). These names are sourced from reference databases and are associated with their unique identifiers in those databases, thus allowing resolution of synonyms and cross-linking among different resources. A term normalization step is used to match the terms with their actual representation in the text, taking into account a number of possible surface variations. Finally, a disambiguation step resolves the ambiguity of the matched terms.

Candidate interactions are generated by simple co-occurence of terms within the same syntactic units. However, in order to increase precision, we parse the sentences with our state-of-the art dependency parser, which generates a syntactic representation of the sentence. This is in turn used to score and filter candidate interactions based on the syntactic fragment which connects the two participating entities.

The ranking of relation candidates is further optimized by a supervised machine learning

```
<collection>
<source>PUBMED</source>
<date>20130422</date>
<key>ctdBCIVLearningDataSet.key</key>
<document>
<id>10617681</id>
<passage>
<infon key="type">title</infon>
<offset>0</offset>
<text>
Possible role of valvular serotonin 5-HT(2B) receptors in the cardiopathy associated with
fenfluramine.
</text>
</passage>
<passage>
<infon key="type">abstract</infon>
<offset>104</offset>
<text>
Dexfenfluramine was approved in the United States for long-term use as an appetite suppressant until
it was reported to be associated with valvular heart disease. The valvular changes (myofibroblast
proliferation) are histopathologically indistinguishable from those observed in carcinoid disease
or after long-term exposure to 5-hydroxytryptamine (5-HT)(2)-preferring ergot drugs (ergotamine,
methysergide). 5-HT(2) receptor stimulation is known to cause fibroblast mitogenesis, which could
contribute to this lesion. To elucidate the mechanism of "fen-phen"-associated valvular lesions,
we examined the interaction of fenfluramine and its metabolite norfenfluramine with 5-HT(2)
receptor subtypes and examined the expression of these receptors in human and porcine heart valves.
Fenfluramine binds weakly to 5-HT(2A), 5-HT(2B), and 5-HT(2C) receptors. In contrast, norfenfluramine
exhibited high affinity for 5-HT(2B) and 5-HT(2C) receptors and more moderate affinity for 5-HT(2A)
receptors. In cells expressing recombinant 5-HT(2B) receptors, norfenfluramine potently stimulated the
hydrolysis of inositol phosphates, increased intracellular Ca(2+), and activated the mitogen-activated
protein kinase cascade, the latter of which has been linked to mitogenic actions of the 5-HT(2B)
receptor. The level of 5-HT(2B) and 5-HT(2A) receptor transcripts in heart valves was at least 300-
fold higher than the levels of 5-HT(2C) receptor transcript, which were barely detectable. We propose
that preferential stimulation of valvular 5-HT(2B) receptors by norfenfluramine, ergot drugs, or
5-HT released from carcinoid tumors (with or without accompanying 5-HT(2A) receptor activation) may
contribute to valvular fibroplasia in humans.
</text>
<annotation>
<infon key="type">disease</infon>
<text>HEART VALVE DISEASES</text>
<id>MESH:D006349</id>
</annotation>
</passage>
</document>
</collection>
```

Box 1. The output of the system is generated in the the BioC specification format. This oputput was generated by a query aiming at retrieving the diseases from pubmed abstract 10617681.

method. Since the term recognizer aims at high recall, it introduces several noisy concepts, which we want to automatically identify in order to penalize them. Additionally, we need to adapt to highly-ranked false positive relations which are generated by our frequency based approach. The goal is to identify some global preference or biases which can be found in the reference database. One technique is to weight individual concepts according to their likeliness to appear as an entity in a correct relation, as seen in the target database.

The OntoGene web service has been implemented as a RESTful service (Richardson and Ruby, 2007). It accepts simple XML files as input, based on the BioC specification[1]. The output of the system is generated in the same format. For example, a query aiming at retrieving the diseases from pubmed abstract 10617681 would generate the output presented in Box 1.

Options can be used in the input query to select whether the result should contain in-line annotations (showing where exactly in the text the term was mentioned), or stand-off annotations (as in the example above). Currently the system uses pre-defined terminology, and only allows the users to decide whether they want to use or not to use one of the pre-loaded vocabularies. However we foresee in future the possibility to upload own terminologies.

Since the OntoGene system not only delivers the specific terms found in the submitted articles, but also their unique identifiers in the source

1 http://www.ncbi.nlm.nih.gov/CBBresearch/Dogan/BioC/

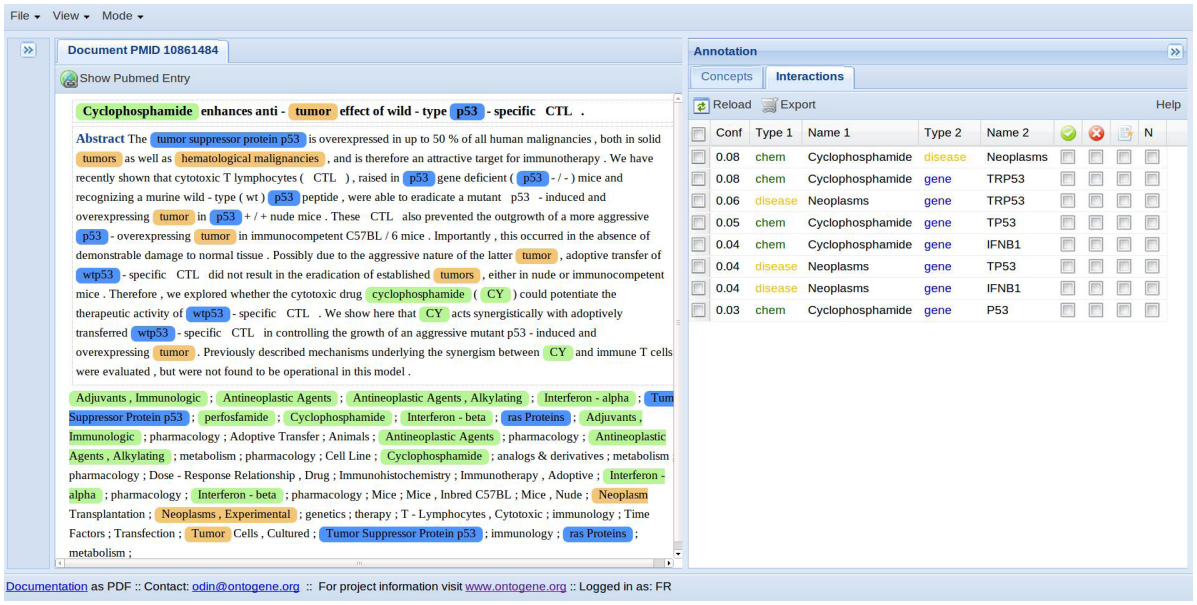


Figure 1. Example of visualization of text mining results using the ODIN interface.

database(s), it is relatively easy to turn its results in a semantic representation, as long as the original databases are based on a standardized ontology. Any term annotation can be turned into a monadic ground fact (possibly using a suitable URI), and interactions can be turned into RDF statements, which could then potentially be integrated across a large collection of documents.

Results and Discussion

Users can submit arbitrary documents to the OntoGene mining service by embedding the text to be mined within a simple XML wrapper. Both input and output of the system are defined according to the BioC standard [Comeau et al., 2013]. However typical usages will involve processing of PubMed abstracts or PubMed Central full papers. In this case the user can provide as input simply the PubMed identifier of the article. Optionally the users can specify which type of output they would like to obtain: if entities, which entity types, and if relationships, which combination of types.

The OntoGene pipeline identifies all relevant entities mentioned in the paper, and their interactions, and reports them back to the user as a ranked list, where the ranking criteria is the system own confidence in the specific result. The confidence value is computed taking into account several factors, including the relative frequency of the term in the article, its general frequency in PubMed, the context in which the term is mentioned, and the syntactic configuration among two interacting entities (for relationships). A detailed description of the factors that contribute to the computation of the confidence score can be found in (Rinaldi et al, 2010).

The user can chose to either inspect the results, using the ODIN web interface (see figure 1), or to have them delivered back via the RESTful web service in BioC XML format, for further processing locally. The usage of ODIN as a curation tool has been tested within the scope of collaborations with curation groups, including PharmGKB, CTD, RegulonDB (Rinaldi, 2012).

The effectiveness of the web service has been recently evaluated within the scope of one of the BioCreative 2013 shared tasks. The official results will be made available at the BioCreative workshop (to be held at the NIH, Bethesda, Maryland, 7-9 October 2013), where only two groups have been invited to present their results, thus showing that the OntoGene/ODIN system is among the top achievers, and will be discussed at the NETTAB workshop when this paper is presented. The system can currently be tested via the ODIN interface[2].

As a future development we envisage the possibility that ODIN could be turned into a tool for collaborative curation of the biomedical literature, with input from the text mining system

2 http://kitt.cl.uzh.ch/kitt/ontogene/bc2013-ctd/

aimed only at facilitating the curation process but not at fully replacing the knowledge of the human experts. It is already possible in ODIN for any user to easily add, remove or modify annotations provided by the system. Such social application could help address the widening gap between the amount of published literature and the capabilities of curation teams to keep abreast with it.

Acknowledgements

The OntoGene group is partially supported by the Swiss National Science Foundation (grants 100014-118396/1 and 105315-130558/1). A continuation of this work is planned within the scope of a collaboration with Roche Pharmaceuticals, Basel, Switzerland.

References

Comeau DC, Islamaj Doğan R, *et al.* (2013) BioC: A Minimalist Approach to Interoperability for Biomedical Text Processing, *Database (Oxford)* **2013**, bat064. doi:10.1093/database/bat064

Richardson L and Sam R (2007), *RESTful Web Services*, O'Reilly, ISBN 978-0-596-52926-0.

Rinaldi F, Kappeler T, *et al.* (2008). OntoGene in BioCreative II. *Genome Biol* **9**:S13. doi:10.1186/gb-2008-9-s2-s13

Rinaldi F, Schneider G, et al. (2010) OntoGene in BioCreative II.5 *IEEE/ACM Trans Comput Biol Bioinform* **7**(3), 472-480. doi:10.1109/TCBB.2010.50

Rinaldi F, Clematide S, et al. (2012) Using ODIN for a PharmGKB revalidation experiment. *Database (Oxford)*, bas021; doi:10.1093/database/bas021

Rinaldi F, Schneider G, and Clematide S. (2012) Relation Mining Experiments in the Pharmacogenomics Domain. *J Biomed Inform.* **45**(5), 851-861. doi:10.1016/j.jbi.2012.04.014

Extracting more value from data silos: using the semantic web to link chemistry and biology for innovation

Tim P. Eyres

Syngenta AG, Switzerland

Received 5 July 2013; **Accepted** 16 September 2013; **Published** 14 September 2013

Competing interests: the authors have declared that no competing interests exist.

Abstract

In order to maximize the chances of finding novel crop protection molecules, that are safe for humans and the environment, it is necessary to bring together biological and chemical information from both inside and outside of an organisation. The integrated use of biological data can help eliminate false positive molecular candidates and improve the chances of finding the correct candidates for development.Information about the biological activity of compounds is captured in disparate systems within Syngenta and in the public domain. This research showed how highly curated bioactivity data from ChEMBL was linked to the Syngenta corporate chemical catalogue, along with other Syngenta research data and commercial patents indexes, using the Resource Description Framework (RDF).

Motivation and Objectives

To maximise the chances of finding novel and safe crop protection molecules it is necessary to join biological and chemical information located inside and outside an organisation. The integrated use of biological data provides a more holistic description of the activity of molecules, and thus helps eliminate false positive molecular candidates and improve the chances of finding the correct candidates for development.

Information about the biological activity of compounds is stored in disparate systems both within Syngenta and in the public domain. This research shows how semantic technologies were used to join public domain bioactivity data with Syngenta corporate data and commercial patents indexes.

The resulting linked data was used to support mode of action, spectrum and selectivity competency questions used in herbicide discovery.

The key outcomes of the research included:

1. increased speed and decreased effort required to retrieve the desired information related to chemical substances of interest (from weeks to days);
2. improved quality of the results when compared to existing approaches leading to large savings in terms of efforts and business benefits, particularly:
   a. discovery of new insights which would otherwise be missed, potentially reducing the number of late stage candidate failures and increasing the likelihood of identifying successful projects early;
   b. avoiding duplicating research efforts.

Methods

Technology

Federated search is an information retrieval pattern (Shokouhi and Si, 2011) allowing the simultaneous search of multiple disparate content sources with one query. A single query request is distributed to the multiple databases in real time, and upon collection the results are arranged in a useful form prior to being presented back to the user.

Prior to this research Syngenta integrated R&D data through the use of data warehouse systems. Although the Data Warehouse pattern offers very fast reporting capabilities on the available data, the approach has *cost and time to market* challenges due to complex processes of extraction, transformation and loading. The rapid increase in data volumes seen in Life Sciences further compounds the integration challenge when keeping internal and external data up to date.

Semantic web technologies provide a novel approach to the federated search pattern. Specific technologies used in this work included TopBraid from Top Quadrant, D2RQ[1], RDF[2], XML[3], SPARQL[4] and Web browsers.

Data

The data shown in Table 1 was integrated.

1 http://d2rq.org/
2 http://www.w3.org/RDF/
3 http://www.w3.org/TR/2008/REC-xml-20081126/
4 http://www.w3.org/TR/rdf-sparql-query

Table 1. This table reports essential information for all integrated data sources.

Data type	Source	Provider	Integration technique
Compound identifiers and activity	Syngenta chemistry repository (internal)	Syngenta	D2RQ federation
Document metadata	Syngenta document repository (internal)	Syngenta	XML import
Protein crystal	Syngenta protein crystal repository (internal)	Syngenta	D2RQ federation
Small molecule activity	Syngenta small molecule repository (internal)	Syngenta	D2RQ federation
Patent metadata	Derwent world patent index (http://thomsonreuters.com/derwent-world-patents-index/)	Thomson Reuters	XML import
Compound identifiers, activity, target and toxicity	ChEMBL (https://www.ebi.ac.uk/chembl/)	EBI	SPARQL

Architecture

The solution faced two key technical challenges of distributed queries and data connectivity. Distributed queries were implemented through the exploitation of SPARQL inferencing following a map/reduce pattern. Data connectivity required a variety of patterns and technologies due to the connectivity offered by the source systems. This included native SPARQL access, D2RQ connectivity to SQL based systems and file export/import. The native SPARQL access proved to be the simplest with the file transfer being the most time and resource consuming. An overview of the Syngenta Federated Search Architecture is shown in Figure 1.

Measuring Benefits

The research benefit was measured by answering competency questions defined in terms of:

1. increased efficiency in lead finding, reducing the time and effort for the scientists;
2. improved access to the internal data sources by linking them and providing a uniform way of accessing them;
3. improving the quality of generated leads and lead evaluation by giving the scientists insights into the external databases and research data.

A selection of example questions is given below:

1. Which chemical structures that are inhibitors of a given enzyme have a potency (threshold) above the potency threshold?
2. What tox data is available for a given enzyme inhibitor?
3. What references have been published for this structure or similar structures?
4. Which species for the given enzyme have IC50 data on plant weed varieties?
5. Find the given enzyme binding site of structures similar to a particular enzyme inhibitor.
6. Which species and compounds have reported measured bioactivity values for a given enzyme?
7. What X-ray crystallographic data is available for a given enzyme?

A method to assess and measure benefits was defined. Several iterations were repeated over four identical phases. Iteration 1 was performed without the tool to establish a baseline. Subsequence iterations were performed with the tool to measure the benefits in comparison to the baseline.

Phase 1: working with the information and available tools to answer competency questions.

Phase 2: completion of an evaluation questionnaire capturing experiences.

Phase 3: collection and formatting of the answers.

Phase 4: open discussion on the results to agree a score per question.

The questionnaire measured the business benefits in the following categories:

1. General tool usability
2. Efficiency gains
3. Quality of results compared to the current ways of working.

Results and Discussion

The Federated Search implementation was tested by five scientists, biochemists and chemists. They had no prior experience of the tool and developed experience during the benefit measurement activity. Where applicable results were on the scale of 1-4 where 4 is the best. The over-

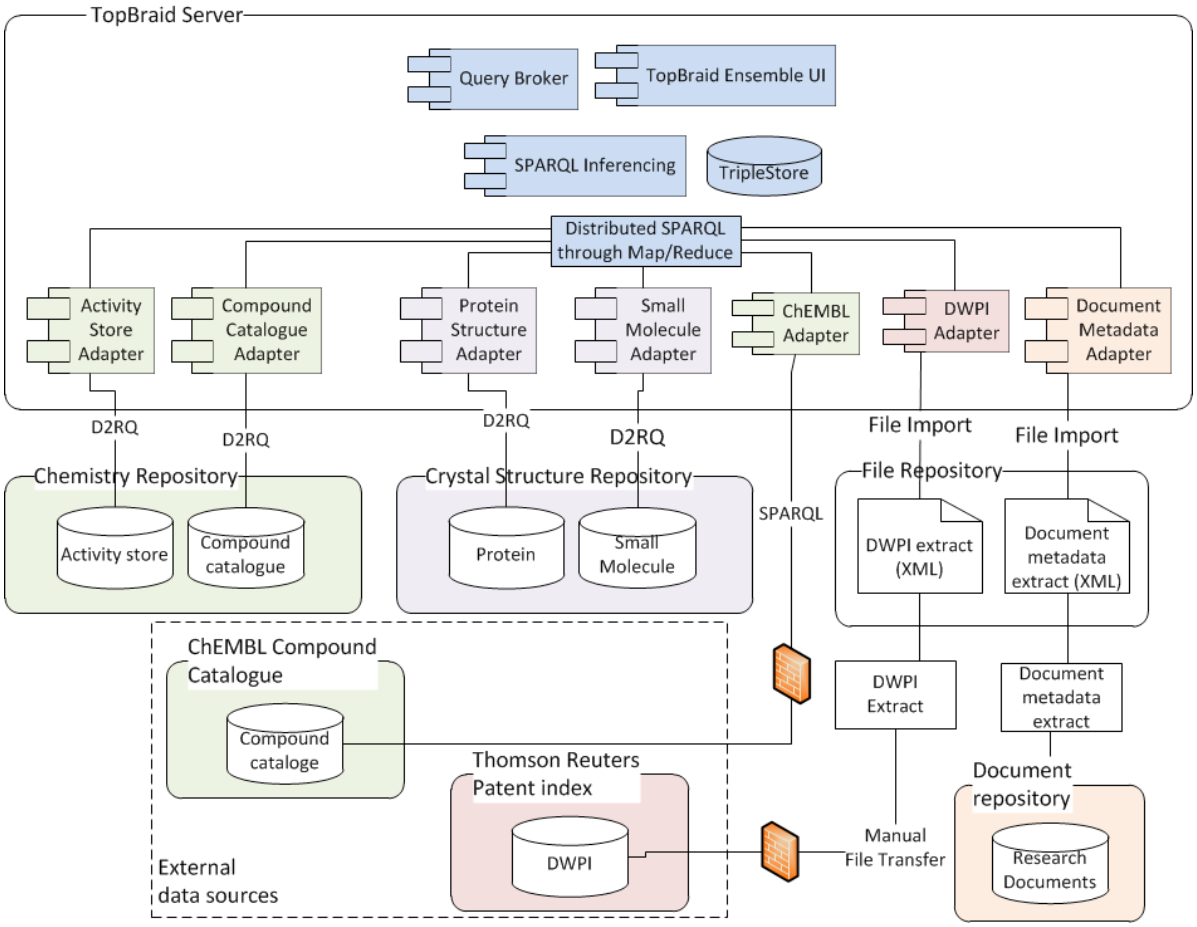


Figure 1. Syngenta Federated Search Architecture overview.

all scores were the final results after all iterations were complete.

How simple was the tool compared to current alternatives?
Overall score: 3. with the following caveats:

1. the tests have been done on a prototype of a Federated Search tool;
2. explanations were missing to form the query correctly;
3. scripted queries worked perfectly, but looking for other things was tricky.

Was there an increase in speed to complete the process from query to result compared to current alternatives?
The efficiency gains were in range of 1:6 for the tool. However, comparison against the baseline was difficult due to certain parts of the existing process being too time consuming without the tool.

The Federated Search solution was recognised as much simpler than the current alternatives.
The proposed Federated Search solution tool effectively replaces a longer manual process.

How does the quality of output compared to current alternatives?
Overall score: 4.

While the output was comprehensive, the presentation of the results made them difficult to analyse.

How was the level of detailed compared to current alternatives?
Overall score: 4.

Complicated search formulation for the x-ray related competency questions made the solution difficult to use with ad hoc queries.

Overall, the approach has been shown to be very valuable, both for integrating data sets and

answering key scientific questions. The ability to apply semantic web technologies to a federated search approach has opened up new avenues in exploiting the large volumes of R&D data within Syngenta. This novel approach is being extended to other parts of R&D in Syngenta.

Acknowledgements

The project team at Syngenta. Our collaborators at ChEMBL. Our collaborators at Thomson Reuters.

References

Shokouhi M, Si L. (2011) Federated Search. *Foundations and Trends® in Information Retrieval* **5**(1), 1-102. doi:10.1561/1500000010

Sprints, Hackathons and Codefests as community gluons in computational biology

Steffen Möller✉, **Enis Afgan, Michael Banck, Peter J. A. Cock, Matus Kalas, Laszlo Kajan, Pjotr Prins, Jacqueline Quinn, Olivier Sallou, Francesco Strozzi, Torsten Seemann, Andreas Tille, Roman Valls Guimera, Toshiaki Katayama, Brad Chapman**

Universtiy of Lübeck Department of Dermatology, Lübeck, Germany

Received 31 July 2013; **Accepted** 10 September 2013; **Published** 14 October 2013

Competing interests: the authors have declared that no competing interests exist.

Abstract

Sprints, Hackathons and Codefests are all names used for informal software developer meetings, especially popular in open source communities. These meetings, along side more traditional conferences, are a vital part of the international network of interactions between software developers working in bioinformatics and computational biology, and complement purely online interactions such as project mailing lists, online chat, web forums and more recently voice and video calls. This paper lays out how the events are organised and presents an overview on their achievements.

Motivation and Objectives

The challenge for everyone is to be aware of existing implementations of a particular desired functionality and the compatibility with the local infrastructure. Strategically, it is beneficial to know other contributors to the externally maintained library, and to ensure that contributions are integrated with the remaining code in the best future-compatible way and with the least possible redundancies.

To help achieve these goals, the Bioinformatics Open Source Conference (BOSC) was established in 2000 by the Open Bioinformatics Foundation Bio* project members as an international venue for showcasing new projects and progress, and for developers world-wide to meet in person. To support team building and help communication, BOSC adopted Birds-of-a-Feather (BoF) sessions, i.e. group meetings of one-two hours.

Methods

A series of longer BioHackathons have been held since 2002 (Stajich *et al.*, 2002). This is short for "biologically motivated code hacking marathons". This initiative evolved into the annual BioHackathons in Japan, organised every year since 2008 with Japanese and key foreign Open Source developers attending (Katayama *et al.*, 2010; Katayama *et al.*, 2011; Katayama *et al.*, 2013).

BOSC's Codefests run as a precursor to an international conference, i.e. BOSC and ISMB, and so is more international. The Sprints take particular effort to invite bioinformaticians local to the event. The BioHackathons have been organised as an invitational event with the loose intention of encouraging the participants to collaborate towards a given theme.

The Codefest is about new developments, but also about helping legacy code to remain compatible with new file formats and/or libraries. The role of Debian Med and Bio-Linux is largely that of an observer, re-distributor, extra pairs of eyeballs during packaging (involving the recompilation) and that of a bridge while answering or forwarding reports by users and/or downstream developers. The same individuals that package for the distribution may also contribute to the packaged project itself.

A main driving force for bringing all those biological tools to a Linux distribution is to save resources by avoiding the compilation, know the versions installed to be tested with the right set of dependencies, and thus allow for more complex combinations of those tools – to create and refine biological workflows. The binaries can already be integrated (Krabbenhöft *et al.*, 2008) with Taverna (Wolstencroft *et al.*, 2013) and remote resources be added (Möller *et al.*, 2010). Bio-Linux ships with Taverna as it comes from the developers' website. And it offers Galaxy (Goecks *et al.*, 2010), the latter packaged so nicely that it should also ship with Debian.

The original BioHackathons in 2002 and 2003 were mainly dedicated to interoperability in handling sequence data amongst the Bio* projects. BioPerl (Stajich *et al.*, 2002), BioJava (Prlić *et al.*, 2012), Biopython (Cock *et al.*, 2009), and BioRuby and BioGems (Goto *et al.*, 2010; Bonnal *et al.*, 2012) groups worked together to develop

common sequence object models, APIs for the BioSQL database and Web services. This ensured fundamental bioinformatic functionality would be compatible among those four programming toolkits.

The first years of the BioHackathon meetings in Japan focused on Web services and interoperability (Katayama *et al.*, 2010; Katayama *et al.* 2011) and later moved to improving life science data integration with Semantic Web technologies (Katayama *et al.*, 2013), reflecting the perceived needs of the biomedical community to move from work flows towards integration of data resources, ontology, semantics and reasoning.

Debian Med (Möller *et al.*, 2010) and Bio-Linux (Field *et al.*, 2006) provide the necessary glue for distribution of individual tool updates back to the wider community. This is achieved by packaging and distributing the tools in the context of these larger tool repositories. Debcamp is an unconference, a place where people meet and work on specific topics, either alone or in teams. The Debcamp concept was later generalised to the Debian Sprints, weekend gatherings of a small number of individuals to address a specific technical challenge. The Debian Med Sprints, because of the heterogeneity of applications while working with many similar types of data, take the idea further. Every winter, general invites are sent to the mailing list to convene at a European coastal town in a family-run hotel, historically resulting in 20 to 25 attendees with expertise across many scientific fields. The first Sprint in 2011 achieved the admirable goal of synchronising Bio-Linux with Debian Med.

Results and Discussion

Every event held keeps a description of its progress and achievements on a dedicated web page: Codefest[1], Debian Med[2], BioHackathon[3]. Having 20+ talented and motivated individuals with shared interests together for two or more days is always special. Such events can be organised at any level with a large enough user base, like in universities, and in all regions of the world. They combine individualised training, social networking, technical contributions and help prepare scientific discoveries.

The two day Codefests and Sprints are often too short to allow every issue to be resolved or to complete forming a consensus. One commonly observes subgroups to dive deeply at one particular topic and stick to it throughout the event. This is excellent for the participants, but difficult for other contributors to synchronise and approach with their concerns. At one week long, as their name suggests, the BioHackathon events in Japan are more of a marathon than a sprint, and allow more interaction between groups - but are more expensive to organise.

Since Open Source software developers spread across the globe already collaborate by communicating online via distributed source-code repositories, mailing lists, chat and other means, the time and expense of travelling to meet up in person may seem like a waste - even case of the BOSC Codefest there is no additional travel for those already attending the main conference. However, physical meetings bring an edge to productivity, including temporarily avoiding day to day workplace duties, and the opportunity to see software and infrastructure problems from outside your local needs. Also, meeting in person temporarily solves the problems of cross time zone collaborations. This is particularly acute for contributors in Australasia communicating with Europeans or Americans, where live interactions like conference calls must be often scheduled outside normal office hours, and any conversation by E-mail can takes days. This is often inefficient, and can be a barrier for promoting international collaboration on Open Source projects when the development speed matters and intensive communication is needed in early brain storming.

Meeting physically also helps build inter-personal relationships and can motivate attendees to follow-up on issues they might not tackle otherwise. One feels a joint strength and confirmation. However, there is also a joint network of remote experts, also outside pure Bioinformatics, that one can rely on.

We feel that to further increase acceptance of the Open Source infrastructures, even though these are already accepted as a commodity, we need to find ways to further ease an adoption of the technology and pave the way for user contributions. The Debian Med Sprints have tutorials and general overviews on software. Future Codefests will likely consider including these too,

1 http://www.open-bio.org/wiki/Codefest
2 http://wiki.debian.org/DebianMed/Meeting
3 http://www.biohackathon.org/

and possibly borrow an idea from the Google Code-in: small well-described yet unresolved tasks tackling real-life problems for the participants to complete within a few hours.

We can point to specific examples of software developments and bug fixes made during the developer meetings described, and in some cases meeting report publications. However, the true worth is more intangible in the form of the community itself, new and strengthened collaborations, and the spread of ideas and best practice - both scientific and for software development.

Acknowledgements

All scientific groups are thanked that support contribution to Open Source software. The events' hosting institutions and funding agencies are thanked for providing facilities and support. O'Reilly, Electric Genetics, Apple Asia, the Debian project and the employing institutions are thanked for contributing to the travel costs of the events. The authors are listed alphabetically by surname, with the exception of the first and last authors who initiated the meetings.

References

Bonnal RJ, Aerts J, *et al.* (2012) Biogem: an effective tool-based approach for scaling up open source software development in bioinformatics. *Bioinformatics* **28** (7):1035-1037. doi: 10.1093/bioinformatics/bts080

Cock PJ, Antao T, *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**(11):1422-1423. doi: 10.1093/bioinformatics/btp163

Field D, Tiwari B, *et al.* (2006) Milo: Open software for biologists: from famine to feast, *Nat Biotechnol.* **24**:801-803. doi:10.1038/nbt0706-801

Goecks J, Nekrutenko A, Taylor J and The Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* **11**(8):R86. doi:10.1186/gb-2010-11-8-r86

Goto N, Prins P, *et al.* (2010) BioRuby: bioinformatics software for the Ruby programming language. *Bioinformatics* **26** (20):2617-2619. doi: 10.1093/bioinformatics/btq475

Katayama T, Arakawa K, *et al.* (2010) The DBCLS BioHackathon: standardization and interoperability for bioinformatics web services and workflows. The DBCLS BioHackathon Consortium*. *J Biomed Semantics* **1**(1):8. doi: 10.1186/2041-1480-1-8.

Katayama T, Wilkinson MD, *et al.* (2011) The 2nd DBCLS BioHackathon: interoperable bioinformatics Web services for integrated applications. *J Biomed Semantics* **2**(2):4. doi: 10.1186/2041-1480-2-4.

Katayama T, Wilkinson MD, *et al.*, (2013) The 3rd DBCLS BioHackathon: improving life science data integration with Semantic Web technologies. *J Biomed Semantics* **4**(1):6. doi: 10.1186/2041-1480-4-6.

Krabbenhöft HN, Möller S, Bayer D. (2008) Integrating ARC grid middleware with Taverna workflows. *Bioinformatics* **24**(9): 1221-1222. doi:10.1093/bioinformatics/btn095

Möller S, Krabbenhöft HN, *et al.* (2010) Community-driven computational biology with Debian Linux. *BMC Bioinformatics* **11**(S-12): S5. doi:10.1186/1471-2105-11-S12-S5

Prlić A, Yates A, *et al.* (2012) BioJava: an open-source framework for bioinformatics in 2012. *Bioinformatics* **28**(20):2693-2695. doi: 10.1093/bioinformatics/bts494

Stajich JE, Block D, *et al.* (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.* **12**(10):1611-1618. doi: 10.1101/gr.361602

Wolstencroft K, Haines R, *et al.* (2013) The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. *Nucleic Acids Res* **41** (W1): W557-W561. doi:10.1093/nar/gkt328

Taverna Mobile: Taverna workflows on Android

Hyde Zhang, Stian Soiland-Reyes✉, Carole Goble
University of Manchester, United Kingdom

Received 10 September 2013; **Accepted** 14 September 2013; **Published** 14 October 2013

Competing interests: the authors have declared that no competing interests exist.

Abstract
Researchers are often on the move, say at conferences or projects meetings, and as workflows are becoming ubiquitous in the scientific process, having access to workflows from a mobile device would be a significant advantage. We therefore have developed Taverna Mobile, an application for Android phones which allows browsing of existing workflows, executing them, and reviewing the results.

Motivation and Objectives

We present Taverna Mobile, a mobile application for browsing and executing Taverna Workflows on Android phones. Taverna is a mature scientific workflow tool suite, designed to combine distributed web services and local tools into complex analysis pipelines (Wolstencroft *et al.*, 2013). Taverna can be executed on local desktop machines, using the Taverna Workbench, or larger infrastructures such as grid or cloud installations using the Taverna Server. Although Taverna has been growing in popularity within fields such as astronomy, chemistry, biodiversity and text mining, its largest user base remains within the bioinformatics community.

The social networking site myExperiment facilitates sharing and reuse of scientific workflows within the scientific community (Goble *et al.*, 2010) and currently includes more than 1,000 Taverna workflows[1], most of which are freely accessible under a Creative Commons license.

Researchers are often on the move, say at conferences or projects meetings, and as work-

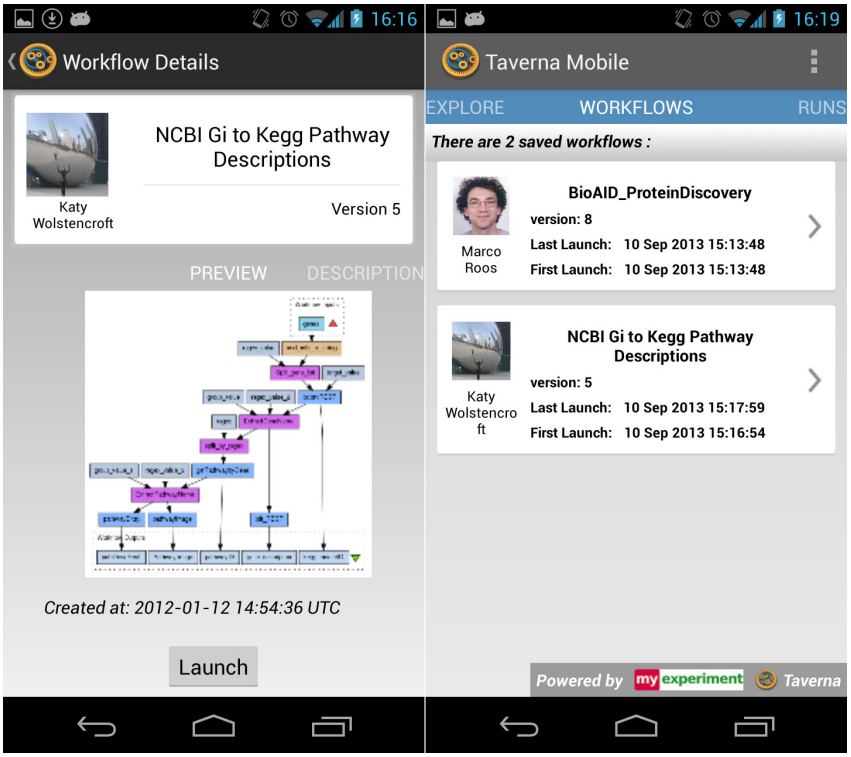


Figure 1. Left: Taverna Mobile browsing myExperiment workflow #2659 NCBI Gi to Kegg Pathway Descriptions. Right: Previously run workflows are saved on the device, with their earlier inputs.

1 http://www.myexperiment.org/workflows

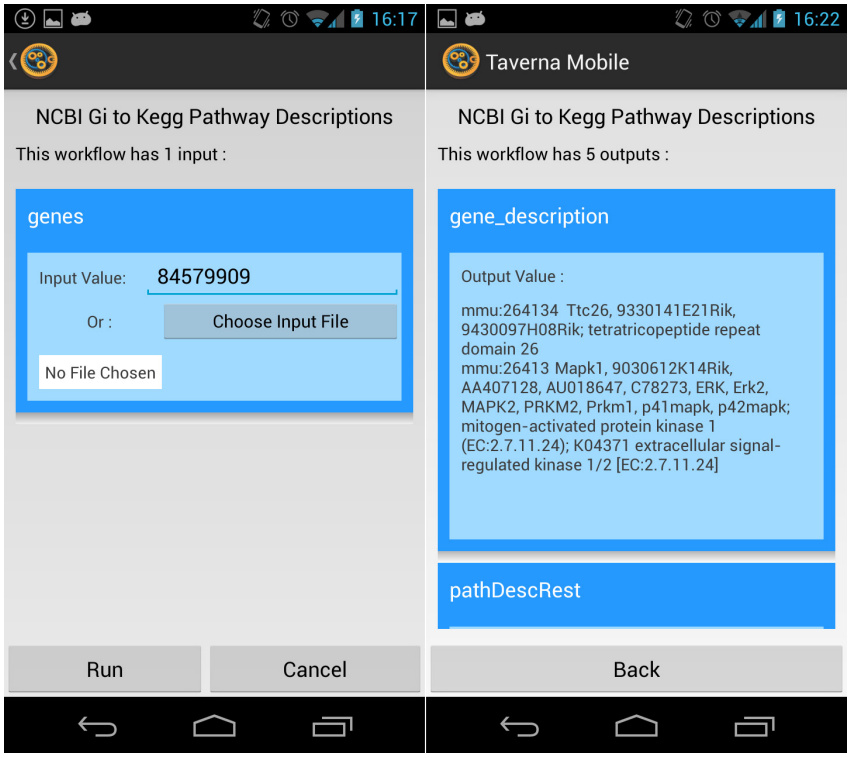


Figure 2. Left: Taverna Mobile prompting for inputs before launching workflow #2659, filled with Gene Identifier (GI) 84579909. Right: Results from workflow run, showing KEGG Pathways with descriptions.

flows are becoming ubiquitous in the scientific process, having access to workflows from a mobile device would be a significant advantage. We therefore have developed Taverna Mobile, an application for Android phones which allows browsing of existing workflows, executing them, and reviewing the results.

Methods

The mobile user can browse and search existing workflows from myExperiment. Metadata about workflows, including uploader, description and rating are displayed together with a diagram of the workflow. From here the user can either "favourite" a workflow to explore it in detail later, download the workflow definition, or click "Run" to execute the workflow.

For running workflows, Taverna Mobile connects to a remote installation of the Taverna Server. The user is prompted for workflow inputs; which can either be typed in or selected from a file on the device. The Android application initiates workflow execution on the server, which status can be monitored from the mobile. Finally, workflow outputs are retrieved and stored as files on the Android device, from where they can be viewed, saved to external locations such as Dropbox and Google Drive, or shared by email.

Frequently accessed workflows are remembered by Taverna Mobile, including their previous input data, which allow the user to quickly repeat workflow runs on the go. Long-running workflow runs that have been started on the Taverna Server through other means, such as through a portal, can also be monitored in the application.

See Figures 1 and 2 for detailed examples making reference to myExperiment workflow #2659 NCBI Gi to Kegg Pathway Descriptions[2] and #74 BioAID Protein Discovery[3].

Results and Discussion

Taverna Mobile does not aim to reproduce the full experience of building workflows in the Taverna Workbench, rather it focuses on tasks we have deemed relevant to a scientist that is not at her desk. For instance, when visiting a conference she might hear about someone's workflow, which she can quickly locate and mark for later exploration. When in the biology lab, faced with updated scientific data, the scientist can rerun

2 http://www.myexperiment.org/workflows/2659/versions/5
3 http://www.myexperiment.org/workflows/74/versions/4

her own workflow with new inputs. While commuting, she can monitor the status of a long-running job.

Taverna Mobile support researchers in these situations by providing immediate access to workflows and workflow runs. The Taverna Mobile application is planned to be released on the Google Play market by the end of 2013, and its source code is available on the Taverna-Mobile[4] project.

Acknowledgements

Hyde Zhang developed the Taverna Mobile application as part of his BSc third year project and a summer placement at School of Computer Science, University of Manchester, and was supervised by Professor Carole Goble. He would also like to thank Robert Haines, Alan Williams and Donal K. Fellows from the myGrid team for their great technical support.

Stian Soiland-Reyes wrote this abstract together with Hyde Zhang, and is funded for the Wf4Ever project by the European Commission's 7th FWP FP7-ICT-2007-6 270192.

References

Goble C, Bhagat J, *et al.* (2010). myExperiment: a repository and social network for the sharing of bioinformatics workflows. *Nucleic Acids Res.* **38**(Suppl 2), W677-W682. doi: doi:10.1093/nar/gkq429

Wolstencroft K, Haines R, *et al.* (2013). The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. *Nucleic Acids Res.* **41**(W1), W557-W561. doi:10.1093/nar/gkt328

4 https://github.com/myGrid/taverna-mobile

Bio-GraphIIn: a graph-based, integrative and semantically-enabled repository for life science experimental data

Alejandra Gonzalez-Beltran✉, **Eamonn Maguire, Pavlos Georgiou, Susanna-Assunta Sansone, Philippe Rocca-Serra**

University of Oxford, United Kingdom

Received 9 July 2013; **Accepted** 10 September 2013; **Published** 14 October 2013

Competing interests: the authors have declared that no competing interests exist.

Abstract

We present the design and architecture of Bio-GraphIIn or "Biological Graph Investigation Index", an integrative and semantically-enabled repository for heterogeneous biological and biomedical experimental metadata.

Motivation and Objectives

Biological and biomedical experiments often rely on a multiplicity of methods to monitor distinct biological signals from a given sample. This is the case, for example, in multi-omic experiments, where samples are studied using several post-genomic techniques (e.g. proteomics, transcriptomics). This variety of methods produce heterogeneous data, whose analysis results should be considered in an integrated manner to provide new insights at the systems biology level. Interpretation of results, as well as potential reuse of data, demands access to the provenance of data and sample information from the overall metadata payload.

Major primary databases, such as those maintained by the US National Center for Biotechnology Information (NCBI) and the European Bioinformatics Institute (EBI), support technology-centric formats and, often, single data types. This state of affairs hampers realising truly integrative work. NCBI and EBI maintain centralised BioSample databases (Barrett *et al.*, 2012; Gostev *et al.*, 2012) to link data back to the originating samples. Yet, owing to the current one-way cross reference, technology-specific databases remain insular.

A second observation is that most of the existing databases support retrospective submissions. In other words, metadata and associated data are deposited in one go, as a bundle when experiments have been completed, and submission systems seldom allow for incremental deposition. Thus, when errors (from spelling mistakes to more serious issues) in the metadata arise, edits to existing repositories are not particularly straightforward, often requiring deleting the submission and re-submitting an entire dataset.

In this work, we present the design and architecture of Bio-GraphIIn (pronounced "biographene"), or "Biological Graph Investigation Index", an integrative and semantically-enabled repository for heterogeneous biological and biomedical experimental metadata that:

1. relies on the ISA-TAB format for the description of experiments and their provenance, as it is multi-purpose and supports multiple data types (Rocca-Serra *et al.*, 2010) ; this makes Bio-GraphIIn an integrative repository;
2. exploits ISA2OWL project (Gonzalez-Beltran *et al.*, 2012) to provide a semantically explicit representation of ISA-TAB formatted experimental information, expressed in a graph model provided by semantic web technologies such as the Resource Description Framework (RDF);
3. supports semantically-rich and traversal queries within and across experiments; for example, involving elements from their design, such as retrieve all investigations whose treatment groups sizes are greater than four, or experiments corresponding to control animals, or retrieving all the data associated with certain samples;
4. enhances user experience by taking advantage of experimental design information to improve experimental metadata in more meaningful ways, such as the study groups defined as combination of factor values or as dynamic groupings;
5. supports Create, Read, Update, Delete (CRUD) operations in a web-interface, allowing for the creation of metadata from the experimental planning phase up to the data analysis. It also offers third party curation/correction possibilities that often lack.

This work improves over existing components from the Investigation/Study/Assay (ISA) Infrastructure (Rocca-Serra *et al.*, 2010), described briefly in the following paragraphs. The ISA infrastructure is a metadata tracking framework that was designed to deal with multi-omic experiments and it is based on three pillars:

1. the multi-purpose ISA-TAB format for the description of the experimental design, factors, what is being measured, the characteristics of the samples, the technology used, the assays, and so on;
2. a software suite allowing for the curation, creation, conversion to other formats such as those supported by public data repositories and RDF/OWL, links to analysis platforms and publication to data journals (Rocca-Serra *et al.*, 2010; Maguire *et al.*, 2013; Gonzalez-Beltran *et al.*, 2013);
3. an international and active user community grouped in the ISA commons (Sansone *et al.*, 2012).

The ISA infrastructure has implemented components for data persistence, namely: the Bio-Investigation Index[1] (BII) web-application, database and database manager tool. These components have been successfully used in systems such as the Stem Cell Discovery Engine (Ho Sui *et al.*, 2012). Contrary to other omics data repositories, as the BII is based on ISA, it supports multiple data types, without the need for a federated infrastructure where each data type is stored in a different endpoint. Similar to other data repositories, BII supports browsing of the stored experiments, comprehensive free text search, filtering according to organism, measurement, technology and platform, and programmatic access. However, BII is 'read only' and does not exploit any semantic features, nor does it allow 'slicing and dicing' across datasets. Bio-GraphIIn is the new generation of the BII and is designed to extend BII's functionality as per the five points above, addressing requests obtained from our user community.

Methods

Graphical user interface (GUI) and database backend - requirements elicitation.

In order to gather the requirements for the Bio-GraphIIn system, we analysed existing data repositories and their functionalities. We also contacted a number of biologists, some within the ISA commons and others not familiar with the ISA infrastructure, and performed semi-structured interviews. The main outcomes of this requirement analysis phase are the basis for the new functionality. This includes the support for CRUD operations, the GUI views at each level of the ISA hierarchy, the ability to retrieve raw and derived data files from samples that satisfy certain conditions on their characteristics across multiple experiments, which depicts not only explicit metadata from ISA-TAB but also elements from the experimental design, such as study groups. In terms of the interface design, biologists have once again stated their preference for spreadsheet-like interfaces, so Bio-GraphIIn relies on a tabular format for ISA-TAB creation.

Service-oriented architecture and implementation details

Figure 1 depicts the modular software architecture of the Bio-GraphIIn system. The web application is based on the Django Web Framework[2] and it relies on two RESTful web services when creating or uploading datasets: one for validation of the ISA-TAB files (wrapping the ISAvalidator code) and another one for conversion to RDF (wrapping the ISA2OWL conversion code). In addition, the system persists the graph representation of the ISA-TAB datasets into a graph database, and relies on SPARQL queries through a Storage And Inference Layer (SAIL), using the TinkerPop open-source stack[3] to interact with the Graphical User Interface.

Experimental metadata representation using semantic web and graph technologies.

The ISA-TAB format lends itself very well for a graph representation, as it describes the experimental workflow: material entities and data files can be represented as graph nodes, whose transformations are described by processes, specified in experimental protocols. The ISA2OWL project (Gonzalez-Beltran *et al.*, 2012) has developed a semantic representation of the ISA-TAB syntax, where the relationships between the highly interconnected ISA elements is made explicit and tagged with ontology terms. These include, for example, the relationships among material and data nodes and their related ISA processes. This

1 https://github.com/ISA-tools/BioInvIndex

2 http://djangoproject.com/

3 http://www.tinkerpop.com/

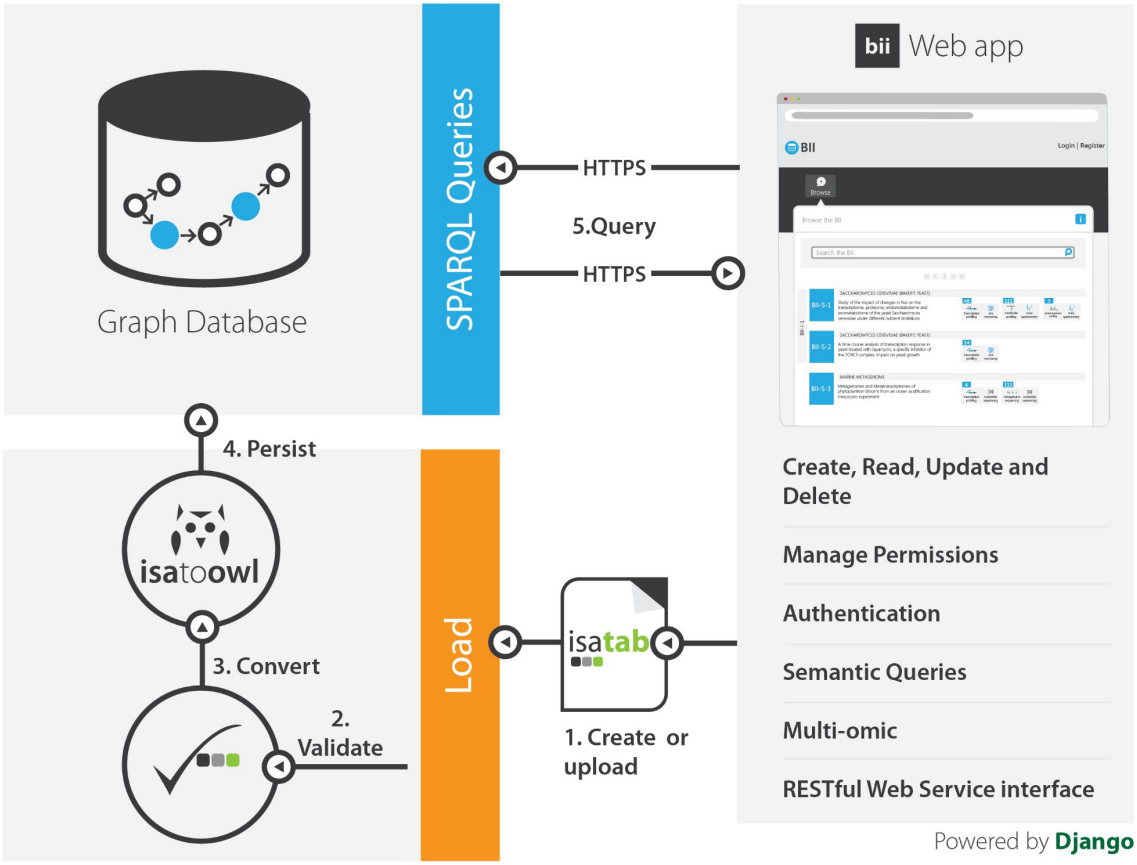


Figure 1. the software architecture of the Bio-GraphIIn system.

representation also allows to build links to external resources (e.g. publications, chemical compounds used in the description of experiments). The ISA2OWL component has been designed to support multiple semantic frameworks, which are specified through mapping files.

Graph databases, CRUD and query operations

For data persistence, Bio-GraphIIn relies on graph database technologies to exploit their ability to deal with highly interconnected data, their scalability and performance. In particular, their use was chosen due to the requirement to perform traversal queries such as those that relate samples to their associated data files. In order to be able to evaluate different existing technologies, the implementation relies on the TinkerPop Blueprints[4], a generic property graph model analogous to Java DataBase Connectivity (JDBC) but for graph databases. This implementation decision will allow us to evaluate different graph database implementations, including neo4j and RDF triple stores (e.g. Sesame).

As regards the CRUD operations, we expect update/delete operations to be more efficient with the underlying graph representation than storing the tabular representation (Brandizi *et al.*, 2012). For the query operations, we will also evaluate their performance for different underlying databases and present the preliminary results.

Results and Discussion

In this work, we presented the design and architecture of Bio-GraphIIn, a graph-based, integrative and semantically-enabled repository for heterogeneous biological and biomedical experimental data. Bio-GraphIIn is composed of a web application interface and a graph database back-end. It relies on a graph data model, as offered by semantic-web technologies such as RDF and OWL, to represent ISA-TAB datasets that describe biological and biomedical experiments relying on a variety of technologies.

4 http://www.tinkerpop.com/

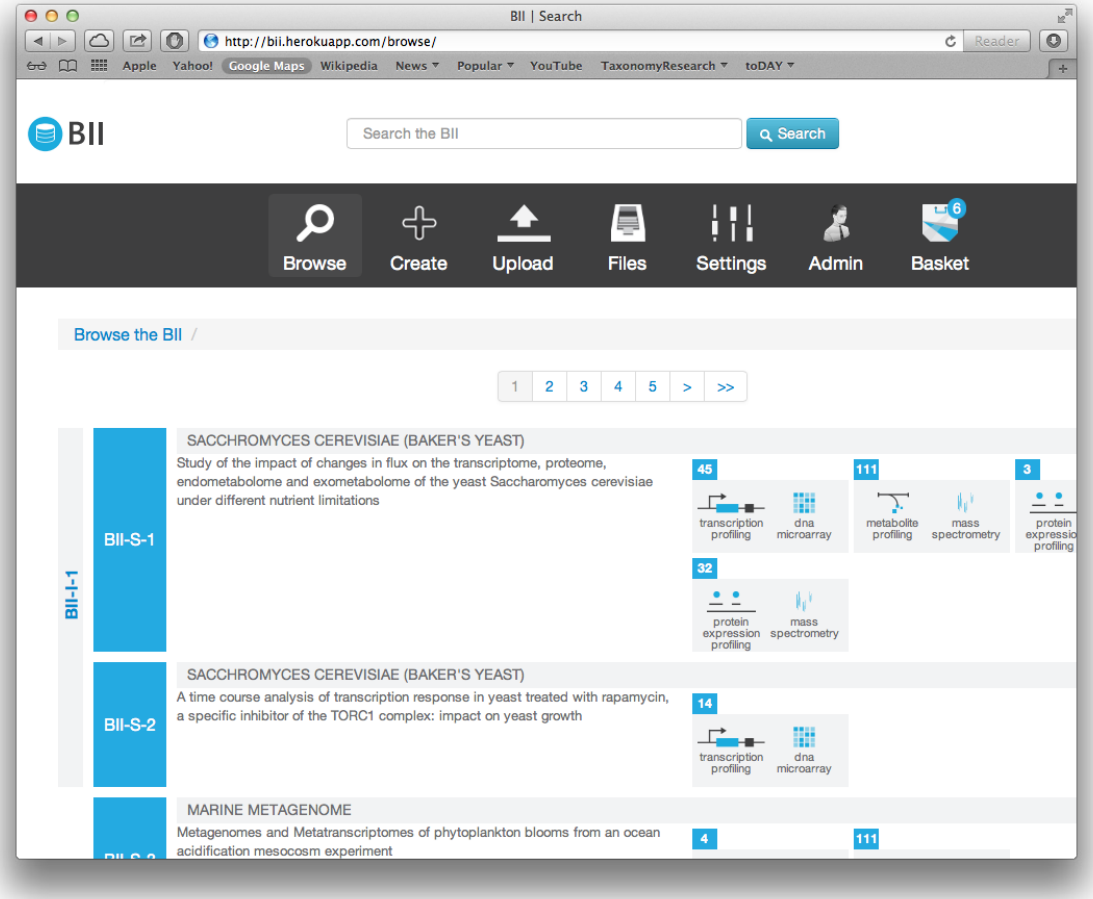


Figure 2. a screenshot of the Bio-GraphIIn system.

As ISA-TAB describes the experimental workflows, from the characteristics and preparation of the samples up to the data analysis performed and summary of results, Bio-GraphIIn supports the tracking of data provenance. Given the graph representation, traversal queries from samples to associate metadata are easily implemented. The latest instantiation of the Bio-GraphIIn database[5] is available on-line (see Figure 2 for a screenshot).

During the presentation, we will show the application using concrete multi-omic datasets and the operations that can be performed with them, and show our preliminary results on the performance analysis for uploading ISA-TAB datasets and for the CRUD operations implemented.

As future work, we will add a versioning feature. We will also investigate ways to associate the Bio-GraphIIn system with resources such as the Refinery Platform[6], a Django-based system for the integration of visualization and analysis of large-scale biological data based on the ISA-TAB format.

Acknowledgements

AGB, EM, SAS and PRS would like to thank their funding support to BBSRC BB/I000771/1, BB/I025840/1 and BB/J020265/1, EU COSMOS EC312941 and the University of Oxford e-Research Centre.

References

Barrett T, *et al.* (2012) BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acid Research* **40**(Database issue), D57–D63. doi: 10.1093/nar/gkr1163

5 http://bii.oerc.ox.ac.uk

6 http://refinery-platform.org/

Brandizi M, *et al.* (2012) graph2tab, a library to convert experimental workflow graphs into tabular formats. *Bioinformatics* **28**(12), 1665-1667.doi:10.1093/bioinformatics/bts258

Gonzalez-Beltran A, *et al.* (2012) The open source ISA software suite and its international user community: knowledge management of experimental data. *EMBnet.journal* **18**, Suppl B, 35-37.

Gonzalez-Beltran A, *et al.* (2013) The Risa R/Bioconductor package: integrative data analysis from experimental metadata and back again. *BMC Bionformatics* In Press.

Gostev M, *et al.* (2012) The BioSample Database (BioSD) at the European Bioinformatics Institute. *Nucleic Acid Research* **40**(Database issue), D64–D70. doi: 10.1093/nar/gkr937.

Ho Sui, *et al.* (2012) The Stem Cell Discovery Engine: an integrated repository and analysis system for cancer stem cell comparisons. *Nucleic Acid Research* **40**(Database issue), D984-991. doi: 10.1093/nar/gkr1051.

Maguire et al (2013) OntoMaton: a BioPortal powered ontology widget for Google Spreadsheets. *Bioinformatics* **29**(4), 525-527. doi: 10.1093/bioinformatics/bts718

Rocca-Serra P, *et al.* (2010) ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics* **26**(18), 2354-2356. doi: 10.1093/bioinformatics/btq415

Sansone S-A, *et al.* (2012) Towards interoperable bioscience data. *Nature Genetics* **44**, 121–126. doi:10.1038/ng.1054

An ontology based query engine for querying biological sequences

Martijn Devisscher[1]✉, **Tim De Meyer**[2], **Wim Van Criekinge**[2], **Peter Dawyndt**[2]

[1]Genohm, Gent
[2]Ghent University, Gent

Received 5 July 2013; **Accepted** 10 September 2013; **Published** 14 October 2013

Competing interests: the authors have declared that no competing interests exist.

Abstract

This work presents the design and proof of principles of Boinq, a flexible query engine for querying and analysing sequence data based on bio-ontology based annotations.The Boinq framework is a web application that allows querying sequencing data in a user friendly way. The application includes a genome browser, and a query builder component that builds SPARQL queries to interrogate endpoints providing sequence annotations.It contains a visualization component for inspection of the data using a genome browser, and an interface for defining the analysis that needs to be done. The analysis will be split up in two steps: (1) Definition of a region of interest by combining a number of simple match operators, and (2) Definition of the analysis [still under construction]The framework also offers a number of SPARQL endpoints that act as sources for delivering feature information as RDF data, and a SPARQL endpoint providing metadata about the feature datasources. These endpoints are queried by the framework, both to fetch the features, and to compose the queries for filtering these feature based on the match operators.

Motivation and Objectives

This work presents the design and proof of principles application of Boinq, a flexible query engine for querying and analysing sequence data based on bio-ontology based annotations.

The bandwidth of sequence data generation has increased spectacularly since the advent of so-called next generation sequencing techniques, now approx. eight years ago (Metzker, 2009). This rate is still increasing today due to single molecule techniques, which are expected to increase data rates even further (Blow, 2008). These developments have spawned development of data processing workflows. At some stage, these pipelines result in a set of reads from the instrument, mapped to a reference genome assembly.

In many applications (such as RNASeq or ChIPSeq) counting these reads in a certain region is the start of further analysis. In research environments, these private read data are combined with publicly available datasets to perform numerous integrative queries over dispersed and highly heterogeneous datasets. An example of such questions is to compare read counts between treatment A and treatment B in regions upstream of genes annotated with a certain gene ontology (GO) term. Such an analysis requires counting reads from two data sources, consulting the GO, and finding gene annotations from a public database. Such analyses still requires hacking together a combination of queries, data conversions and ad hoc scripts. This is time consuming and error prone, and furthermore requires specialised personnel.

An analysis of a set of example questions revealed that there is a need for a rapid analysis pipeline to:

- quickly specify and visualise regions of interest based on a number of criteria. In these criteria, interoperability with bio-ontologies is required;
- perform simple aggregating or ranking analyses in these regions.

The boundary condition imposed by working with a collection of distributed, heterogeneous data is the natural ecosystem for semantic web technologies. This fact, and the required interoperability with bio-ontologies led to the decision to leverage semantic web technologies for disclosing, integrating and querying sequence data. The use of semantic web technologies offers clear advantages. As sketched above, the technologies are ideally suited to deal with distributed, heterogeneous data sources, and a growing body of (molecular) biological knowledge is being disclosed using well defined bio-ontologies.

Drawbacks can be identified as well. First, the inherent freedom associated with exposing data as RDF creates challenges. Obviously, using a common technology is not sufficient for guaranteeing interoperability. Therefore, a way to describe data sources that describe annotation features needs to be agreed upon. While such an initiative needs discussion in a wider

group, some minimum requirements for such a standard are put forward in what follows. A second drawback originates from the complexity for the layman to create queries for a liberal data space. For this reason, we felt the need to include a query builder into the platform, as discussed further on.

Methods

The Boinq framework is composed of the following components:

- an RDF store with SPARQL endpoint documenting available data sources for features. An ontology is available for this meta dataset and is discussed further;
- a set of local SPARQL endpoints for exposing feature sets from various sources as RDF data, either directly or through mapping of the underlying SQL data. An endpoint is available for querying a subset of a locally running ensembl core data set for homo sapiens (Flicek *et al.*, 2012);
- an interface for exploring the feature data sources. It contains a visualisation component for inspection of the data using a built-in genome browser, and an interface for defining the analysis that needs to be done. The analysis is split up in two steps:
  - definition of a region of interest;
  - definition of the analysis (still under construction).

The Boinq application is offered as a web application, and is built on a Java software stack. The following technologies were used to develop the framework:

- the client interface is built using SmartGWT[1];
- ontologies were built using Protégé[2];
- the server software is composed of individual components orchestrated using Spring[3], and persistence is achieved using Hibernate[4];
- the RDF data and the ontologies used are exposed as a SPARQL endpoint using the Apache Jena framework[5], more specifically the fuseki component. This framework is also used as a SPARQL client;
- mapping relational data to RDF dynamically is done using d2rq[6] (Bizer & Seaborne, 2004). The mapping from the ensembl core to FALDO[7] is documented on-line;
- Apache tomcat is used as application server;
- asynchronous jobs are handled using Quartz[8].

The architecture is depicted graphically in Figure 1.

Results and Discussion

The boinq tool in its current version assumes the presence of data sources providing genome annotations (or features) through a SPARQL endpoint. We limit our application to features that are mapped to a publicly available reference genome and with exactly known positions on this reference.

Describing RDF feature data sources

Currently, exposing features as RDF data is not yet common practice, yet several initiatives are in place for doing just this in the near future. As common standards are not yet in place, we define in this section the assumptions about a feature data source. We have tried to make use of publicly available ontologies as much as possible:

- the position of a feature on a reference assembly is described using FALDO[9];
- the types of features are described using the sequence ontology (Eilbeck *et al.*, 2005).

All further information about feature types that are offered by a data source are described in the meta dataset, that is an RDFS/OWL repository, available at the boinq website. A direct SPARQL endpoint to this meta dataset is also available. In the current version, we are using a central repository, but we solicit cooperative efforts to ensure that each data source adheres to a common ontology describing its fields and feature types.

In addition to providing sufficient information about the resource for human interpretation, the data source RDF store needs to provide assistance to the region builder component (see further). It needs to detail the available (filterable) fields for each feature type, and their target data types. The query builder uses SPARQL to query the data source metadata repository.

1 http://code.google.com/p/smartgwt/
2 http://protege.stanford.edu
3 http://www.springsource.org/
4 http://www.hibernate.org/
5 http://jena.apache.org/
6 http://www.d2rq.org/
7 http://github.com/mr-tijn/d2rq__ensembl__faldo
8 http://quartz-scheduler.org/
9 https://github.com/JervenBolleman/FALDO-paper

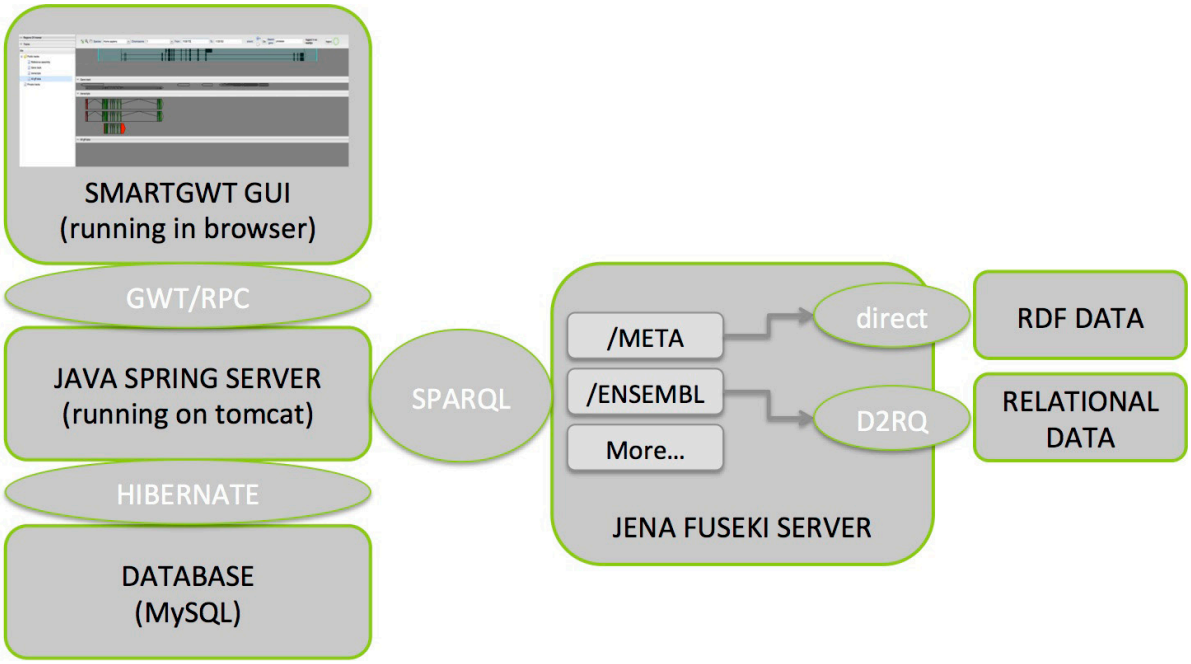


Figure 1. Architecture of the boinq application.

The region builder

While SPARQL is relatively straightforward to use for a technical audience, the tool is intended for biologists, not computer scientists. In addition, querying pure RDF without a reasoner in place will fail to return desired results. Therefore, a query builder was designed to aid the user in defining "Match trees", a combination of match operators that result in queries on SPARQL data sources for computing regions of interest.

An additional benefit of defining queries on a meta-level rather than directly in SPARQL is that generators can be built on top of the match tree that target other protocols for fetching remote features. For example, the design of the match tree is kept compatible with the DAS specification (Jenkinson *et al.*, 2008), so a generator for DAS query URLs may be plugged in at a later stage.

The region of interest that is used to perform the analysis, is defined by setting a number of criteria on known features. The region of interest for the analysis is then composed of a set of regions bound to the set of matching features.

The criteria are built by combining Match operators. The following Match operators are currently implemented:

- Match Location - these are criteria regarding the location of the feature;
- Match Type - to only return features of the specified type;
- Match All - this is an aggregate operator that requires all subcriteria to match;
- Match Any - this aggregate operator requires at least one subcriterion to match;
- Match Field - this operator restricts the features to those that fit within a restriction on a property of the feature. This operator is discussed in further detail.

The match field operator makes extensive use of the data source metadata. Indeed, as RDF data have inherently little restrictions on data types or the type of statements that can be made, there is great freedom in the definition of SPARQL queries. In order to guide the user in defining queries that are relevant to the data at hand, some restrictions are necessary. The field match operator first queries, for the type of feature at hand, the fields that are known. For this, a SPARQL query is used that takes advantage of rdfs:domain statements on properties of superclasses of the feature type. To avoid a too wide selection however, superclasses are restricted to avoid generic classes such as owl:Class. For the property field that is chosen from this selection by the user, the rdfs:range statements are inspected for the target type of the selected property.

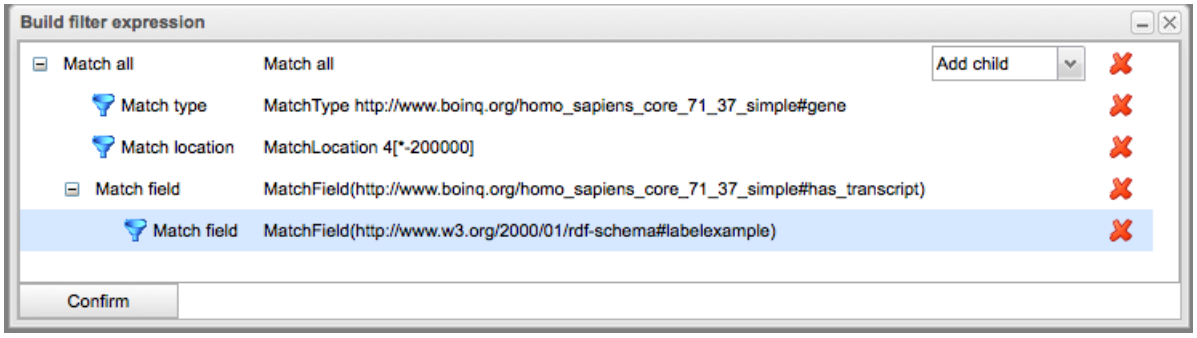


Figure 2. Example of a query under construction.

If no information is found, the user can enter either an URI or a literal of a supported type (string and some numeric types). If range statements do provide information about the target entity, there are two possibilities:

- the entity is further described in the data source metadata - in this case, the interface will recurs into the target entity, creating a sub-Match on a field of the target entity, resulting in a chain of Matches;
- the entity is unknown - in this case, again the user can choose between a URI or a literal;
- the entity is a literal - in this case, the user can enter a restriction on the literal value.
- For literal values, the following restrictions are supported:
- strings will be interpreted as a regex that must match, and optionally be case insensitive;
- for numeric values, comparison operators can be constructed (=,<,>,...).

An example of a query under construction is given in Figure 2. In Figure 2, features of type "gene" are to be fetched form the ensembl *Homo sapiens* datasource, that have a transcript whos label matches "example", and are located on chromosome 4, before position 200000. After construction of the Match tree, a SPARQL expression is generated that is used to fetch the set of matching features, and calculate the region of interest from that set. The user is presented the expression before use, so advanced users can make changes to the query, if necessary. For example, the query generated from Fig. 2 is

```
SELECT [...]
WHERE
  { ?feature
    rdfs:label            ?featureId ;
   faldo:begin            ?featureBegin .
 ?featureBegin
    faldo:position         ?featureBeginPos.
  ?feature
    faldo:end              ?featureEnd.
  ?featureEnd
    faldo:position         ?featureEndPos.
  ?featureBegin
    faldo:reference        ?featureRef.
  ?featureRef
    rdfs:label             ?featureRefName.
  ?featureBegin
    rdf:type
?featurePositionType.
  ?feature
    rdf:type               ensembl:gene;
    ensembl:has_transcript  ?entity1.
  ?entity1
    rdfs:label             ?entity2 .
    FILTER (str(?featureRefName) = "4")
    FILTER (?featureBeginPos <= 200000)
    FILTER regex(str(?entity2),
"example", "i")
  }
ORDER BY
ASC(?featureBeginPos)
```

External terms (terms from external ontologies) are handled in a specific way. The metadata repository specifies for each external term:

- a SPARQL endpoint;
- a SPARQL graph;
- optionally a term that is the superclass of any terms that can be used;
- optionally additional restrictions on the terms from the target ontology.

The system currently refers to Bioportal (Whetzel *et al.*, 2011) for external terms. Upon encountering an external term, the user interface will use the Bioportal SPARQL endpoint to find and retrieve the relevant terms, and present them as a tree to the user. The user can also use full text search to find terms. The user has an option to retrieve all subclasses to be included in the match.

The visualization component

A built-in genome browser is available to inspect the computed regions visually. The browser is integrated in the interface and supports track based visualisation of generic features, and also shows ENSEMBL data as a background.

The analysis component

The analysis part is under development, but currently supports computing counting reads from a track in a predefined region of interest. Features to be included here are:

- ranking individual regions in a region of interest based on read count comparison. While further analysis is needed for validating findings based on these comparisons, it is already useful in focusing the effort;
- comparing read counts between tracks in a region of interest. For this, further research has to be performed into methods for computing statistical significance based on unknown data sources.

Availability

The tool is still under active development, but a prototype version is available at http://www.boinq.org/. This version currently supports basic queries based on Homo sapiens assembly v. 37. At this URL, you will find links to:

- a brief user manual
- the actual webapp
- the SPARQL endpoint
- the RDFS/OWL files

Acknowledgements

The work was made possible by a Baekeland scholarship from the Agency for Innovation by Science and Technology in Flanders (IWT). The first author furthermore wish to thank all Genohm colleagues, and in particular David De Beule and Ruben Simoens for their expertise and kind assistance with programming in SmartGWT and Spring.

References

Bizer C and Seaborne A. (2004) D2RQ-treating non-RDF databases as virtual RDF graphs. Proceedings of the 3rd International Semantic Web Conference, Hiroshima, Japan, November 7-11, 2004.

Blow N (2008) DNA sequencing: generation next-next. *Nature Methods* **5**, 267-274. doi:10.1038/nmeth0308-267

Eilbeck K, Lewis SE, *et al.* (2005) The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biology* **6**(5). doi:10.1186/gb-2005-6-5-r44

Flicek P, Ahmed I, *et al.* (2012) Ensembl 2013. *Nucleic Acids Research* **41**(D1), D48–D55. doi:10.1093/nar/gks1236

Jenkinson AM, Albrecht M, *et al.* (2008) Integrating biological data - the Distributed Annotation System. *BMC Bioinformatics* **9**(Suppl 8), S3. doi:10.1186/1471-2105-9-S8-S3

Metzker ML (2009) Sequencing technologies - the next generation. *Nature Reviews Genetics* **11**(1), 31–46. doi:10.1038/nrg2626

Whetzel PL, Noy NF, *et al.* (2011) BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Res.* **39**(Web Server issue), W541-545. doi:10.1093/nar/gkr469.

The representation of biomedical protocols

Larisa N. Soldatova✉, **Ross D. King, Piyali S. Basu, Emma Haddi, Nigel Saunders**

Brunel University, London, United Kingdom

Received 15 July 2013; **Accepted** 10 August 2013; **Published** 14 October 2013

Competing interests: the authors have declared that no competing interests exist.

Abstract

An explicit and logically consistent model for the representation of biomedical protocols would enable researchers in the Life Sciences to better record, execute, share, and report experimental procedures and results. The model we propose is based on the ontology of EXperimental ACTions (EXACT). The EXACT model is designed to define typical actions performed by biologists in labs and their essential attributes to enable recording of biomedical protocols in a computer processable form. This oral communication aims to report on the recent progress with the EXACT-based representation of biomedical protocols.

Motivation and Objectives

An explicit and logically consistent model for the representation of biomedical protocols would enable researchers in the Life Sciences to better record, execute, share, and report experimental procedures and results. The model we propose is based on the ontology of EXperimental ACTions (EXACT) (Soldatova et al., 2008). The EXACT model is designed to define typical actions performed by biologists in labs and their essential attributes to enable recording of biomedical protocols in a computer processable form.

EXACT was originally developed to support the protocols executed by the Robot Scientist "Adam". This is a physically implemented robotic system that applies techniques from artificial intelligence to execute cycles of automated scientific experimentation (King *et al.*, 2009). A Robot Scientist can in a fully automatic manner: originate hypotheses to explain observations, devise experiments to test these hypotheses, physically run the experiments using laboratory robotics, interpret the results, and then repeat the cycle. Adam is capable of running in parallel thousands of experiments with yeast strains. The first ever fully automated scientific discovery made by Adam has captured public imagination and was listed by the Times Magazine[1] as one of the most important scientific discoveries of 2009. While the EXACT approach has been proved successful to represent and record experiments with yeast, it is not sufficient to support the representation and recording of a wider range of biomedical protocols. The proposed EXACT model is built on the success of EXACT ontology and extends its representations to support a wide range of biomedical protocols.

Related works

Typically, ontological representations are focussed on modelling declarative knowledge about principal physical objects, and their qualities and relations with other objects. Process entities are included to represent the processes in which physical objects participate, *e.g.* gene-gene interactions. Representations where procedural knowledge plays the central role are rare. The Ontology for Biomedical Investigations[2] (OBI) includes both existential and procedural knowledge, but the main focus is on the representation of entities participating in biomedical investigations. For example, OBI Core contains only 17 procedural entities (occurrents), and about 100 continuants. OBI is sufficient to formally capture information about typical assays. However, standard operating procedures remain largely non-formalised, and are usually in the form of natural language with links to ontological classes to specify participating entities.

Initially the EXACT ontology contained only 45 experiment actions limited to the representation of biomedical lab automation protocols in yeast biology, and not all of them had well defined properties. Moreover, some of the defined experiment actions are not suitable for most of biomedical laboratories. For example, it is important to instruct a robot to remove a lid from a plate. However, such an action would be implicitly understood by a human researcher. At a laboratory standards workshop in Stockholm in December 2011 it was decided to modify the EXACT approach to suit the needs of the

[1] http://www.time.com/time/specials/packages/article/0,28804,1945379_1944416_1944423,00.html

[2] http://obi-ontology.org/

Molecular Methods database[3] (MolMeth) for the recording of protocols (Klingström *et al*, 2013).

This oral communication aims to report on the recent progress with the EXACT-based representation of biomedical protocols.

Methods

Analysis of biomedical protocols

We are manually inspecting thousands of published and also commercial biomedical protocols from several areas of biomedicine, including neurology, epigenetics, metabolomics, stem cell biology. We are analyzing instructions, what properties an experiment action has, what conditions are required and what goals are specified. We are also populating the EXACT ontology by newly identified experiment actions. We are modifying existing EXACT classes by specifying their properties. For example, the class mix has been defined as "to put together or combine (two or more substances or things) so that the constituents or particles of each are interspersed or diffused more or less evenly among those of the rest" (the Oxford English Dictionary, 1989). However, only one property equipment has been specified. The following new information about the experiment action mix has been added to the EXACT model:

```
has-participant (mix, entity) AND min
cardinality = 2
has-participant (mix, container)
```

to specify that at least two entities have to participate in the experiment action mix, and this action has to be carried out in some container. If a user while entering a lab protocol to a system would miss any of these properties, then the system would request to specify the missed properties of the action *mix*.

While (semi-) automated text mining methods are available, we judged that expert analysis of how Life Science practitioners express their procedural knowledge would output a higher quality knowledge model. We will use text mining tools to check if our model covers at least 95% of domain procedural knowledge. We will continue to analyze protocols till the coverage is sufficient.

Assessment of biomedical protocols by experts

Unfortunately, as often happens with natural language, the instructions in biomedical protocols are not always consistent or complete, and therefore do not always guarantee full reusability of the protocols (Soldatova *et al.*, 2008). For example, based on the analysis of existing biomedical protocols we have identified that the following attributes of the action store are typically recorded:

- an entity (what will be stored),
- duration (for how long it will be stored),
- condition (*e.g.* humid air),
- a location and/ or a container (where it should be stored).

However, it is not obvious what attributes are essential and must be recorded for each action store, and what attributes are optional. Some statements in published protocols, *e.g.* "store working solution at -20°C until use", specify the entity and the condition, but not a duration or a location. There are also some statements about the action *store* in other protocols that specify locations and durations, but not conditions. We aim to capture all essential information about typical experiment actions, and also what information is optional and useful to record. We wish to strike the right balance between ensuring that all the essential information is recorded, and at the same time not requiring unnecessary or optional information from our users. Therefore we are consulting with experts in Life Sciences in order to define what properties of experiment actions are essential and what are optional.

Observation of the execution of experiment actions

Much procedural knowledge is implicit and difficult to verbalize, and therefore hard to capture and model. Therefore, a high quality representation of biomedical protocols can only be achieved if knowledge engineers directly observe how Life Sciences practitioners perform experiment actions in their labs. So far we have observed the execution of experiment actions in two different labs, one in the University of Aberystwyth (Wales), and the other in Brunel University (London). We are negotiating with three further labs to provide us with an access to their lab facilities and also to interview their biologists in order to capture implicit procedural knowledge.

Knowledge re-use

Previously defined relevant classes will be imported to the EXACT model. For example the OBI classes as *cell fixation* (definition: a protocol

[3] http://www.molmeth.org

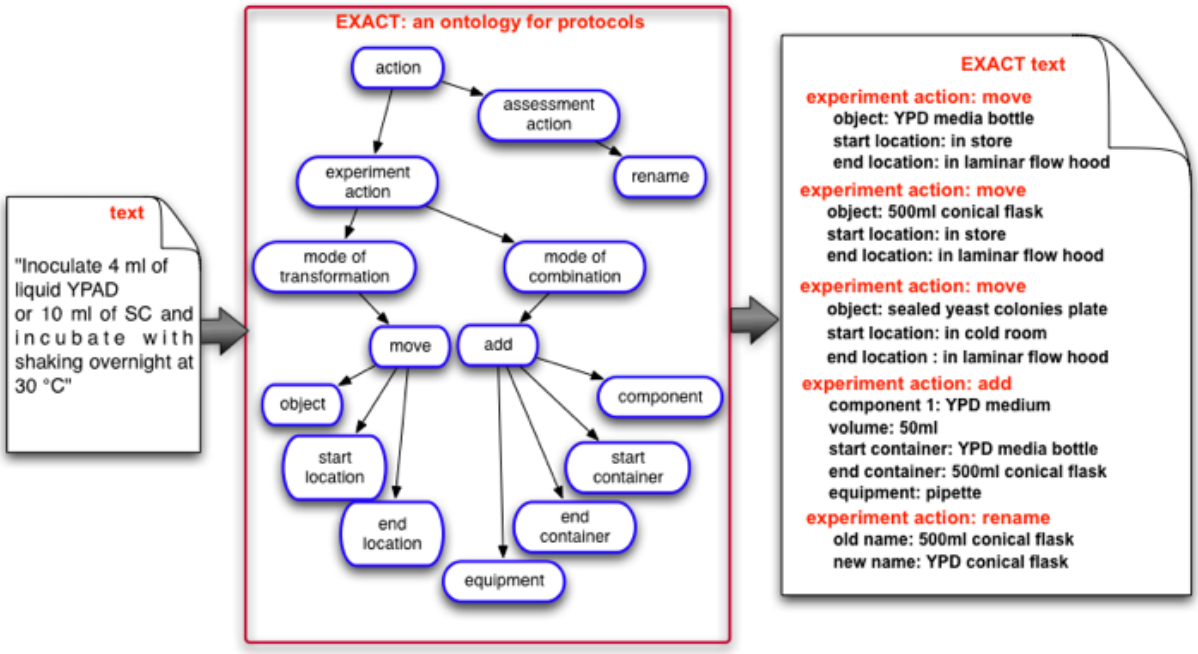


Figure 1. An example of the EXACT representation of biomedical protocol instructions.

application to preserve defined qualities of cells or tissues (sample) which may otherwise change over time), *decapitation* (definition: decapitation is a process by which the head of a living organism is physically removed from the body, usually resulting in rapid death), *labeling* (definition: the addition of a labeling reagent to an input biomaterial in order to detect the labeled material in the future) will be re-used in EXACT with the OBI URIs (Unique Resource Identifiers).

Results and Discussion

The number of experiment actions in the EXACT model has been increased significantly, and new properties had been defined. The EXACT model has been harmonised with the OBI representations. Currently the EXACT model is being verified by experts, and we are checking how well it covers the domain (see the methods section). By the end of this process we will deposit an updated EXACT model to BioPortal[4].

We aim to provide an intuitive and easy representation of biomedical protocols and ensure that experimental procedures are fully reproducible (see Fig. 1). Through the use of the EXACT model reporting tools we will be able to provide biologists with more intelligent support. It will be possible to check if a protocol contains all the required information about experiment actions, suggest how to fill in any identified gaps or remove inconsistences, provide templates for typical experiment actions, and help to re-use already recorded protocols.

Acknowledgements

This work has been partially funded by the Brunel University BRIEF award and a grant from Occams Resources.

References

King RD, Rowland J, *et al.* (2009) The Automation of Science. *Science* **324**, 85-89. doi:10.1126/science.1165620

Klingström T, Soldatova L, *et al.* (2013) Workshop on laboratory protocol standards for the molecular methods database. *New Biotechnology* **30**(2), 109-113.

Soldatova LN, Aubrey W, *et al.* (2008) The EXACT description of biomedical protocols. *Bioinformatics* (Special issue ISMB)**24**, i295-i303. doi:10.1093/bioinformatics/btn156

The Oxford English Dictionary (1989). Oxford University Press, 2nd ed.

[4] http://bioportal.bioontology.org/

The role of parallelism, web services and ontologies in bioinformatics and omics data management and analysis

Mario Cannataro✉, **Pietro Hiram Guzzi**

University of Catanzaro, Italy

Received 14 July 2013; **Accepted** 12 September 2013; *Published* 14 October 2013

Competing interests: the authors have declared that no competing interests exist.

Abstract

The increasing availability of omics data poses new challenges to bioinformatics applications regarding the efficient storage and integration of experimental data, their efficient and high-throughput preprocessing and analysis, the building of reproducible "in silico" experiments, the integration of analysis results with pre-existing knowledge repositories stored into ontologies, like Gene Ontology, or into specialised databases, as those available in pharmacogenomics. This paper presents an overview of how parallelism, service orientation, and knowledge management techniques can be used to face those challenges presenting some recent bioinformatics tools and projects that employ such technologies in different stages of the bioinformatics analysis's pipeline.

Motivation and Objectives

The increasing availability of *omics* data poses new challenges to bioinformatics applications that need to face an overwhelming availability of raw data. Main challenges regard: (i) the efficient storage, retrieval and integration of experimental data; (ii) their efficient and high-throughput preprocessing and analysis; the building of reproducible "in silico" experiments; (iii) the integration of analysis results with pre-existing knowledge repositories stored into ontologies or into specialized databases.

This paper presents an overview of how parallelism, service orientation, and knowledge management techniques can be used to face those challenges presenting some recent bioinformatics tools and projects that employ such technologies in different stages of the bioinformatics analysis's pipeline.

Methods

Main *omics* disciplines are gaining an increasing interest in the scientific community due to the availability of high throughput platforms and computational methods which are producing an overwhelming amount of *omics* data.

The increased availability of *omics* data poses new challenges both for the efficient storage and integration of the data and for their efficient preprocessing and analysis.

Hence, managing *omics* data requires both support and spaces for data storing as well as procedures and structures for data preprocessing, analysis, and sharing. The resulting scenario comprises a set of methodologies and bioinformatics tools, often implemented as web services, for the management and analysis of data stored in geographically distributed biological databases.

As the storage, preprocessing and analysis of raw experimental data is becoming the main bottleneck of the analysis pipeline, due to the increasing size of experimental data, **high-performance computing is playing an important role in all steps of the life sciences research pipeline**, from raw data management and processing, to data integration and analysis, up to data exploration and visualization.

Web services and workflows are used to face the complexity of the bioinformatics pipeline that comprises several steps. Finally, **ontologies** and knowledge management techniques **are used to connect pre-existing knowledge** in biology and medicine **to the omics experimental data and analysis results**.

We present an overview of how parallelism, service orientation, and knowledge management techniques can be used to face those challenges presenting some recent bioinformatics tools and projects that employ such technologies in different stages of the bioinformatics analysis's pipeline, with special focus on the analysis of *omics* data. Moreover, we briefly introduce some recent emerging architectures (multicore systems, GPUs) and programming models (MapReduce, Cloud Computing) that will have a key role to face the overwhelming volumes of data generated by *omics* platforms.

Results and Discussion

The role of parallelism, web services and ontologies in bioinformatics

In these last years, both well-known high performance computing techniques such as Parallel and Grid Computing, as well as emerging computational models such as Graphics Processing and Cloud Computing, are more and more used in bioinformatics and life sciences (Cannataro, 2009).

The huge dimension of experimental data is the first reason to implement large distributed data repositories, while high performance computing is necessary both to face the complexity of bioinformatics algorithms and to allow the efficient analysis of huge data. In such a scenario, novel parallel architectures (e.g. multicore systems, GPU, FPGA, hybrid CPU/FPGA, CELL processors) coupled with emerging programming models (e.g. Service Oriented Architecture, MapReduce) may overcome the limits posed by conventional computers to the mining and exploration of large amounts of data.

On the other hand, the modeling of complex bioinformatics applications as collections of web services composed through workflows, is an emerging approach to face the high complexity of bioinformatics applications and enabling the repeatability of "in silico" experiments, and thus the reproducibility of the same experiment by different research groups.

Workflows are used to combine such web services forming reusable bioinformatics applications that may be deployed on several distributed or parallel architectures, such as Grids or clusters. Moreover, using parameter sweep technology, a single workflow may be instantiated in various forms to test in parallel different algorithms on some of the steps of the bioinformatics pipeline.

Knowledge management techniques and especially ontologies are more and more used to model pre-existing knowledge in medicine and biology. For instance Gene Ontology[1] (GO) is used to annotate experimental data or results data with external information.

Ontologies are not only useful to annotate data, but also to support the composition of bioinformatics workflows. By modeling the application domain of a bioinformatics application and the analysis techniques used to analyse data, ontologies may be used to guide the development of bioinformatics workflows, suggesting tools needed to implement specific steps of preprocessing or analysis or alerting the user when some constraints are going to be violated, e.g. when the user tries to apply a wrong preprocessing tool or a wrong sequence of tools.

Parallel preprocessing of gene expression microarray data.

The dimension of microarray datasets is becoming very large since the dimension of files encoding a single chip and the number of the arrays involved in a single experiment, are increasing. The system developed in (Guzzi and Cannataro, 2010a) uses a master/slave approach, where the master node computes partitions of the input dataset (i.e., it sets a list of probesets intervals) and calls in parallel several slaves each one wrapping and executing the apt-probeset-summarize program, that is applied to the proper partition of data. Such system showed a nearly linear speedup up to 20 slaves.

Web Services-based preprocessing of gene expression microarray data

micro-CS (Microarray Cel file Summarizer) (Guzzi and Cannataro, 2010b) is a distributed tool for the automation of the microarray analysis pipeline that supports the automatic normalization, summarization and annotation of Affymetrix binary data, providing a web service that collects on behalf of the user the right and most updated libraries.

Workflow-based preprocessing and analysis of mass spectrometry-based proteomics data

The analysis of mass spectrometry proteomics data requires the combination of large storage systems, effective preprocessing techniques, and data mining and visualisation tools. The management and analysis of huge mass spectra produced in different laboratories can exploit the services of computational grids that offer efficient data transfer primitives, effective management of large data stores, and large computing power.

MS-Analyzer (Cannataro, 2007) is a software platform that uses ontologies and workflows to combine specialized spectra preprocessing algorithms and well known data mining tools, to analyze mass spectrometry proteomics data on the Grid. Data mining and mass spectrometry

1 http://www.geneontology.org/

ontologies are used to model: (i) biological databases; (ii) experimental data sets; (iii) and bioinformatics software tools.

MS-Analyzer uses the Service Oriented Architecture and provides both specialised spectra management services and publicly available data mining and visualisation tools. Composition and execution of such services is performed through an ontology-based workflow editor and scheduler, and services are classified with the help of the ontologies.

Ontology-based annotation and querying of protein interaction data

Protein-protein interaction (PPI) databases store interactions among proteins and offer to the user the possibility to retrieve data of interest through simple querying interfaces. Thus, even simple queries like "retrieve all the proteins related to glucose synthesis" are usually hard to express.

OntoPIN[2] (Cannataro *et al.*, 2010) is a software platform that uses ontologies for automatically annotating proteins interactions and for querying the resulting annotated interaction data. OntoPIN includes a framework able to extend existing PPI databases with annotations extracted from GO and an interface for querying the annotated PPI database using semantic similarity in addition to key-based search.

Semantic similarity-based visualisation of protein interaction networks

The use of such annotations for the analysis of protein data is a novel research area. Semantic similarity measures evaluate the similarity of two or more terms belonging to the same ontology, thus they may be used to evaluate the similarity of two genes or proteins measuring the similarity among the terms extracted from the same ontology and used to annotate them (Guzzi *et al.*, 2012).

Recently, we used semantic similarity measures among proteins to develop a novel visualization method for protein interaction networks implemented into CytoSevis[3], a plugin of Cytoscape[4]. CytoSevis visualizes protein interaction networks in a semantic similarity space (Guzzi and Cannataro, 2012). CytoSevis exploits semantic similarity analysis and provides a graphical user interface that enables the visualization of networks in such a semantic space.

Emerging architectures and programming models

High performance computing is more and more used in biology, medicine and bioinformatics, to face the increasing amount of available experimental data. In the following we briefly introduce some emerging architectures and programming models that will have a key role to face the overwhelming volumes of data generated by *omics* platforms.

Multi-core and many-core systems

A multi-core processor is a single computing element containing two or more independent CPUs, called "cores", which read and execute program instructions in parallel. A multi-core processor usually comprises two, four, six or eight independent processor cores on the same silicon chip and connected through an on-chip bus.

Multi-core processors execute threads concurrently and often use less power than coupling multiple single-core processors. On the other hand, when increasing the number of cores the on-chip bus becomes a bottleneck, since all the data travel through the same bus, limiting the scalability of multi-core processors. Many-core processors put more cores in a thermal container than the corresponding multi-core processors.

General Purpose Graphics Processing Units

The Graphics Processing Unit (GPU) is a specialised electronic device initially used to accelerate the building of images to be sent to a display. The term General Purpose GPUs (GPGPU) indicates GPUs that are used for general-purpose computation. GPGPUs are mainly used for embarrassingly parallel computations and they are well suited to applications that exhibit large data-parallelism. One of the main manufacturers of GPUs and GPGPUs is NVIDIA.

MapReduce and Apache Hadoop

MapReduce is a recent programming model well suited for programming embarrassingly parallel applications that need to process large volumes of data. MapReduce is also the name of the Google programming model (Dean and Ghemawat, 2008). The MapReduce model is inspired by the **map** and **reduce** functions used in functional programming. A very popular free

2 http://www.ontopin.org/
3 http://sites.google.com/site/cytosevis/
4 http://www.cytoscape.org/

implementation of MapReduce is Apache Hadoop[5].

Cloud Computing

Cloud Computing allows to access computers, services and eventually infrastructures as a utility, through the Internet. Developers can build novel Internet services without the need to buy large and costly hardware to deploy them as well as the human expenses to operate them (Ahmed *et al.*, 2012). Cloud Computing encompasses technology, economics and business model aspects so finding a complete definition is an issue.

According to Ahmed *et al.* (Ahmed *et al.*, 2012): "Cloud computing is a way of leveraging the Internet to consume software or other information technology services on demand". Using Cloud computing both resources and costs are shared. Also Ahmed *et al.* conclude that Cloud computing is more a business model than a computing paradigm.

Cloud-based Bioinformatics

Clouds are more and more used to host and deploy bioinformatics applications. Amazon EC2 has made available two main bioinformatics datasets in its publicly available repository[6]: the Annotated Human Genome Data provided by ENSEMBL, and UniGene provided by the National Center for Biotechnology Information. Dudley and Butte (Dudley and Butte, 2010) point out that clouds not only can offer elastic computational power to bioinformatics applications, but the availability of instances of bioinformatics applications that are stored and shared in the cloud can make computational analyses more reproducible. Schatz *et al.* (Schatz *et al.*, 2010) report a list of bioinformatics resources made available through the cloud[7]. Recently, several cloud-based platforms for bioinformatics and biomedical applications have been deployed.

Conclusion

Parallelism, web services and ontology technologies are key tools for modern emerging bioinformatics applications. The paper discussed the role of such technologies in many case studies regarding especially the management, preprocessing and analysis of *omics* data, with special focus on genomics, proteomics and interactomics data. The discussion is completed through the presentation of several bioinformatics tools exploiting those technologies.

Future work will regard a more comprehensive assessment of such technologies through the definition of quantitative and qualitative application requirements that can be fulfilled by adopting those technologies.

References

Ahmed M, Chowdhury ASMR, *et al.* (2012) An Advanced Survey on Cloud Computing and State-of-the-art Research Issues. *International Journal of Computer Science* **9**, 201-207.

Cannataro M (2009) *Computational Grid Technologies for Life Sciences, Biomedicine and Healthcare*, Medical Information Science Reference, IGI Global Press, Hershey, USA, May 2009.

Cannataro M, Guzzi PH, *et al.* (2007) Using Ontologies for Preprocessing and Mining Spectra Data on the Grid. *Future Generation Computer Systems* **23**(1), 55-60. doi:10.1016/j.future.2006.04.011

Cannataro M, Guzzi PH, Veltri P (2010) Using ontologies for querying and analysing protein-protein interaction data. *Procedia CS* **1**(1), 997-1004. doi:10.1016/j.procs.2010.04.110

Dudley JT, Butte AJ (2010) In silico research in the era of cloud computing. *Nature Biotechnology* **28**, 1181-1185. doi:10.1038/nbt1110-1181

Guzzi PH, Cannataro M (2010a) Parallel Pre-processing of Affymetrix Microarray Data. Euro-Par Workshops 2010, *Springer Lecture Notes in Computer Sciences* LNCS 6586, 2011, 225-232. doi:10.1007/978-3-642-21878-1_28

Guzzi PH, Cannataro M (2010b) μ-CS: An extension of the TM4 platform to manage Affymetrix binary data. *BMC Bioinformatics* **11**, 315. doi:10.1186/1471-2105-11-315.

Guzzi P, Cannataro M (2012) Cyto-Sevis: semantic similarity-based visualisation of protein interaction networks. *EMBnet. journal* **18**(A), 32-33.

Guzzi PH, Mina M, Guerra G, Cannataro M (2012) Semantic similarity analysis of protein data: assessment with biological features and issues. *Briefings in Bioinformatics* **13**(5), 569-585. doi:10.1093/bib/bbr066

Schatz MC, Langmead B, Salzberg SL (2010) Cloud computing and the DNA data race. Nature Biotechnology 28, 691–693. doi:10.1038/nbt0710-691

5 http://hadoop.apache.org/
6 http://aws.amazon.com/publicdatasets
7 http://www.nature.com/nbt/journal/v28/n7/fig_tab/nbt0710-691_T1.html

Posters

SstmpDB: a database of single-spanning transmembrane proteins

Olga Bejleri, Zoi Litou, Stavros Hamodrakas✉

Department of Cell Biology and Biophysics, Faculty of Biology, Panepistimiopolis, University of Athens, Athens, Greece

Received 30 July 2013; **Accepted** 6 September 2013; **Published** 14 October 2013

Competing interests: the authors have declared that no competing interests exist.

Abstract

Membrane proteins represent ca. 20–30% of both eukaryotic and prokaryotic proteomes. They play crucial roles in cell survival and cell communication, as they function as transporters, receptors, anchors and enzymes. More than 30% of all prescribed drugs are targeting membrane proteins. Transmembrane proteins are either single-spanning membrane proteins or multi-spanning proteins. Single-spanning proteins can be classified into four types I, II, III and IV, depending on their topology and membrane targeting. They are very important functionally, involved in the presentation of antigens to the immune system, they are calcium-dependent cell adhesion proteins, they play a role in septum formation and they have many more specific, crucial roles.The purpose of this work was the construction of a database containing all single-spanning membrane proteins and their functional classification. This database is available at http://aias.biol.uoa.gr/sstmpdb.

Motivation and Objectives

Membrane proteins represent ca. 20–30% of both eukaryotic and prokaryotic proteomes. They play crucial roles in cell survival and cell communication, as they function as transporters, receptors, anchors and enzymes. More than 30% of all prescribed drugs are targeting membrane proteins. Transmembrane proteins either span the membrane once (single-spanning membrane proteins) or several times (multi-spanning membrane proteins). Single-spanning proteins are classified into four types I, II, III and IV, depending on their topology and membrane targeting (Hedin *et al.*, 2011). They are very important functionally, involved in the presentation of antigens to the immune system, they are calcium-dependent cell adhesion proteins, they play a role in septum formation and they have many more specific, crucial roles.

The key objective of this project was the collection of all available to date single pass transmembrane proteins and the construction of a database and a web interface for storing and handling these proteins. Also, a functional clustering was performed, aiming at the creation/discovery of novel functional clusters/families, for all single-pass transmembrane protein types.

Methods

For data collection, the database used was UniProtKB/SwissProt, release 2012_11[1]. From all initially collected data, fragments were removed and the remaining data set was further filtered by subcellular location, keeping only single spanning proteins. Then all virus proteins were removed and the final data set contained only proteins with clear experimental evidence at protein and transcript level. Isoforms were not kept as separate entries in the database. Data was grouped by type, organism, and subcellular location. All data pre-possessing has been done using Perl scripts. The main database was built using MySQL on a Apache server and the web interface for SSTMPdb, created with PHP and javascript, is located at http://aias.biol.uoa.gr/sstmpdb/.

For functional clustering, modern NLP algorithms (e.g., Latent Semantic Analysis, LSA) (Landauer *et al.*, 1998) and common techniques for statistical data analysis/clustering, such as k-means clustering using MATlab (Zeimpekis *et al.*, 2006), were used. As input, pre-processed datasets of the field *Function* of the Uniprot/Swiss-Prot files, for all single-pass transmembrane proteins were utilised.

Results and Discussion

SstmpDB currently contains 10,250 proteins from 344 organisms and provides information such as their sequence, their type, the functional family they belong to, isoforms, etc. From the web interface of the database, the user has the ability to search entries by Uniprot AC, type and organism and a more advanced search is also available. All data are downloadable in FASTA, text and tab delimited format for each entry or several entries, at will. The web site also allows BLAST searches against the database and contains a detailed manual as supporting material. SstmpDB is the

1 http://www.uniprot.org/

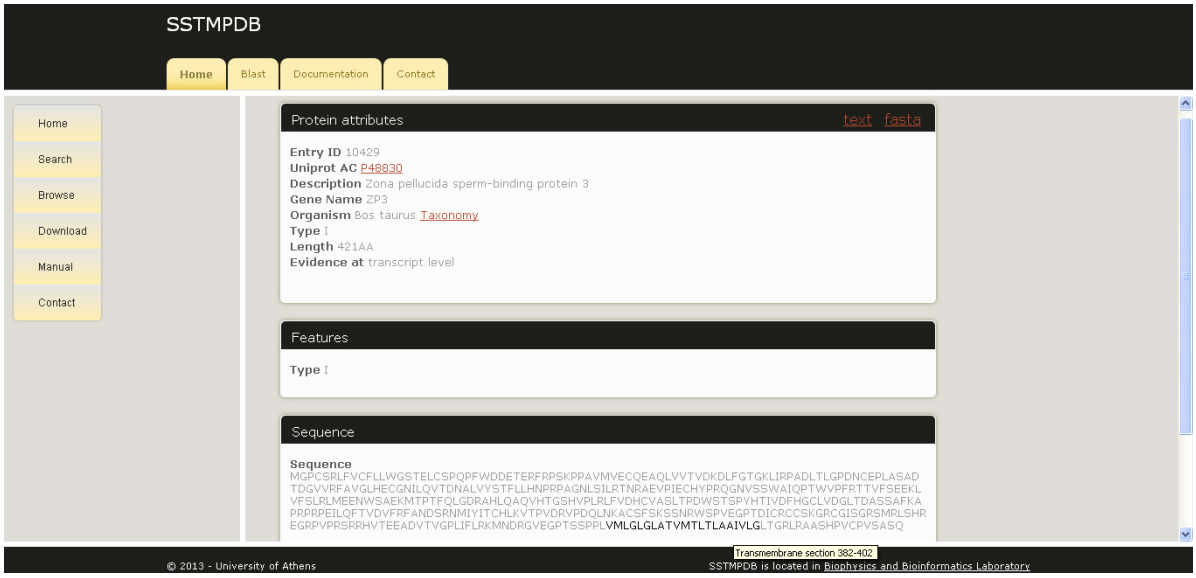


Figure 1. Display of data entry 10429. Protein attributes, features and sequence with transmembrane section (382-402) are shown.

first publicly available database that collects and provides information about single-spanning membrane proteins.

Acknowledgements

This work was funded by the SYNERGASIA 2009 PROGRAMME, co-funded by the European Regional Development Fund and National resources (Project Code 09SYN-13-999), General Secretariat for Research and Technology of the Greek Ministry of Education and Religious Affairs, Culture and Sports.

References

Hedin L, Illergård K, *et al.* (2011) An introduction to membrane proteins. *J Proteome Res* **10**(8), 3324-3331. doi: 10.1021/pr200145a.

Landauer T, Foltz P, *et al.* (1998) Introduction to Latent Semantic Analysis. *Discourse Processes* **25**, 259-284.

The UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). (2012) *Nucleic Acids Res.* **40**, D71-D75. doi: 10.1093/nar/gkr981.

Zeimpekis D, Gallopoulos E (2006) TMG: A MATLAB Toolbox for Generating Term-Document Matrices from Text Collections. In: Kogan J, Nicholas C, Teboulle M, (Eds.), *Grouping Multidimensional Data: Recent Advances in Clustering*, Springer, Berlin, 187-210.

Sinergy: how semantic can improve early prevention of skin cancers

Diletta Romana Cacciagrano[1] ✉, Flavio Corradini[1], Leonardo Vito[2], Laura Cavalieri[3]

[1]Science and Tecnology, Computer Science Division, University of Camerino, Italy
[2]Researcher funded by L.I.L.T., Gagliole, Italy
[3]University of Camerino - ADiTech s.r.l., Ancona, Italy

Received 17 September 2013; **Accepted** 19 September 2013; **Published** 14 October 2013

Competing interests: the authors have declared that no competing interests exist.

Abstract

Melanoma is the major source of deaths related to skin cancer. The project "Sinergy", in cooperation with Lega Italiana Lotta Tumori (L.I.L.T.) and ADiTech s.r.l., aims at realising a Cloud-based and semantic-driven framework for screening and analysis of skin lesions. The framework aims at being an ubiquitous laboratory in Cloud, equipped with semantic-driven tools for making analysis, defining/executing/sharing in-silico experiments, mining laboratory data and activities. The engine kernel, a multi-scale and multi-physics skin lesion ontology, not only makes intuitive any operation for non-expert users, but also improves diagnostic processes thanks to the possibility ofglobally reasoning on skin lesion variables relative to different biological scales (e.g., genetic, molecular, cellular, tissutal ones) and to not-biological ones (e.g., age, gender, location, geography, race).

Motivation and Objectives

Cancer is a class of diseases characterised by out-of-control cell growth and tissue invasion.

Melanoma is the major source of deaths (75%, with lifetime risk for Caucasians of one in 39 for men and one in 58 for women) related to skin cancer.

When melanoma progresses to metastatic stage, powerful mechanisms to resist chemotherapy, radiation and biological intervention are established in the neoplastic lesions, thus hampering the efficacy of current medical therapies and drugs (Helmbach *et al.*, 2001).

Because advanced skin cancers remain incurable, early prevention is mandatory. Research focused on genetic level, diagnosis based on conventional screening tests and treatment by surgical excision are currently the only approaches to reduce mortality.

Research

Since cancer initiation seems to depend on a series of genetic mutations affecting intrinsic cellular programs, the vast majority of cancer research to date has focused on the identification of these genetic and molecular properties of cancer cells. Two highly related genes, KN2A and CDK4, were discovered to harbour germline mutations in roughly 50% of melanoma pedigrees.

However, genetic predisposition can only be found in 3% of all cases. It means that the inheritance of CDKN2A and CDK4 mutations looks insufficient to lead to melanoma in all carriers.

Empirical studies confirm such a conjecture, suggesting that the focus of the research has to necessarily include also other parameters, related to (1) other biological scales (not only the genetic one) and to (2) other physical system (not only the biological one). A model is considered to be multi-scale if it spans two or more different spatial scales and/or includes processes that occur at two or more temporal scales. It is considered to be multi-physics if it involves multiple physical models or multiple simultaneous physical phenomena

For what concerns (1), consider the angiogenesis (Ziemys *et al.*, 2011), a significant transforming phase in tumour growth. Drugs delivered to tissues will not only change the behaviour of melanoma cells (secretion of cytokines, proliferation, differentiation, apoptosis, or migration) in the intracellular drug-triggered cell division process, but also inhibit the development of new capillary sprouts by preventing sprouts from receiving vascular endothelial growth factors. In turn, inadequate glucose and oxygen transported from the blood vessel will drive even more melanoma cells towards apoptosis.

The crossed synergy among organ, tissue and cell scales in angiogenesis hardly interferes on drug distribution and therapeutic effects (i.e., on molecular scale). For instance, several molecular drugs developed to treat melanoma cancer did not work as well *in vivo* as *in vitro* because of absorption, distribution, metabolism, or toxicity problems (Soengas and Lowe, 2003). Many synergistic drug delivery methods have been devel-

oped to increase the drug effect *in vivo*, but it is difficult to quantitatively evaluate their performance. All that highlights the need of defining and exploiting models and methodologies for studying skin melanoma cancer at different biological scales.

Many mathematical models have been proposed to address the challenge mentioned above. These models studied one or more phases of cancer progression, including tumour growth, angiogenesis, and drug treatment, with the purpose of better understanding the pathophysiology of cancer, mechanisms of drug resistance and the optimisation of treatment strategies. Although biologists have already obtained many experimental data sets at the molecular, cellular, micro-environmental and tissue levels, only a few scientists have integrated these data into a n-scale tumor model, where n is greater than two (e.g., atomic-molecular (Liu *et al.*, 2007), molecular-microscopic (Athale and Deisboeck, 2006), microscopic-macroscopic (Zheng *et al.*, 2005)).

For what concerns (2), several heterogeneous factors (e.g., age, gender, location, geography, race) seem to play an important role in melanoma predisposition, incidence and distribution. In some way, such factors should be also related to biological variables (and to genetic ones, in particular). However, it cannot be formally proved yet. In detail:

- *Age:* the relationship between the incidence of melanoma and age is unusual in comparison to other common cancers. There is not an exponential increase in risk with age but rather a more even distribution across age groups.
- *Gender:* men are more likely than women to develop melanoma (67% higher incidence) and their prognosis is worse (136% higher risk of death from melanoma).
- *Location:* basal cell cancers arise exclusively from cutaneous sites and are closely related to sites of skin that receive the most sun exposure, such as the scalp, face, neck, and arms. A small percentage of melanomas arise on acral surfaces of the hands and feet, which tend to be diagnosed at a later stage.
- *Geography:* the rates of melanoma and other skin cancers are highest where fair-skinned Caucasians migrated to lower latitudes.
- *Race:* Caucasians are by far the most susceptible race for melanomas. Hispanics have a lower incidence but represent the group at next highest risk. Asians and African-Americans have the lowest rates of skin cancer. All racial groups are equally likely to develop melanoma on the acral surfaces of the hands and feet or mucosal surfaces.

Diagnosis and treatment

Conventional screening tests are based on a naked-eye examination by an experienced clinician. ABCD (Soyer *et al.*, 2004) - looking at asymmetry of the skin lesion, irregular edges (borders), color variegation, and diameter - is one of the easiest and most widely used algorithm for evaluating suspicious pigmented skin lesions (PSLs). Since a naked-eye ABCD may fail to detect many borderline PSLs that are small or/and regular in shape or colour, the algorithm is often combined to dermoscopy - a diagnostic technique for an *in vivo* better visualisation of PSLs - and/or implemented in computer-aided image analysis systems (CaIASs).

Dermoscopy improves diagnostic sensitivity by 20–30% compared with a naked-eye diagnosis (Soyer *et al.*, 2004). However the results of dermoscopic examination have limitations, especially for the inexperienced, and they are effective only if the user is formally trained.

CaIASs allow PSLs to be classified by automatically quantifying ABCD attributes and other texture-based characteristics. However, CaIASs effectiveness depends largely on dataset, feature selection and classification methods used (Boldrick *et al.*, 2007). Consequently, conventional CaIASs, although pivoting on objective parameters, cannot allow more reliable decision processes than the ones of an experienced clinician (pivoting on experience, complex inferences and extensive knowledge).

"Putting semantic" into CaIASs could enable more reliable decision processes, extraction of implicit information from data, integration of quantitative and functional information to infer relevant new knowledge, as well as standardisation of terminology, verification of data consistency and integration of heterogeneous biomedical databases.

Methods

The project "Sinergy", in cooperation with Lega Italiana Lotta Tumori (L.I.L.T.) and ADiTech s.r.l., started in March 2013, aims at realising a Cloud-based and semantic-driven framework for

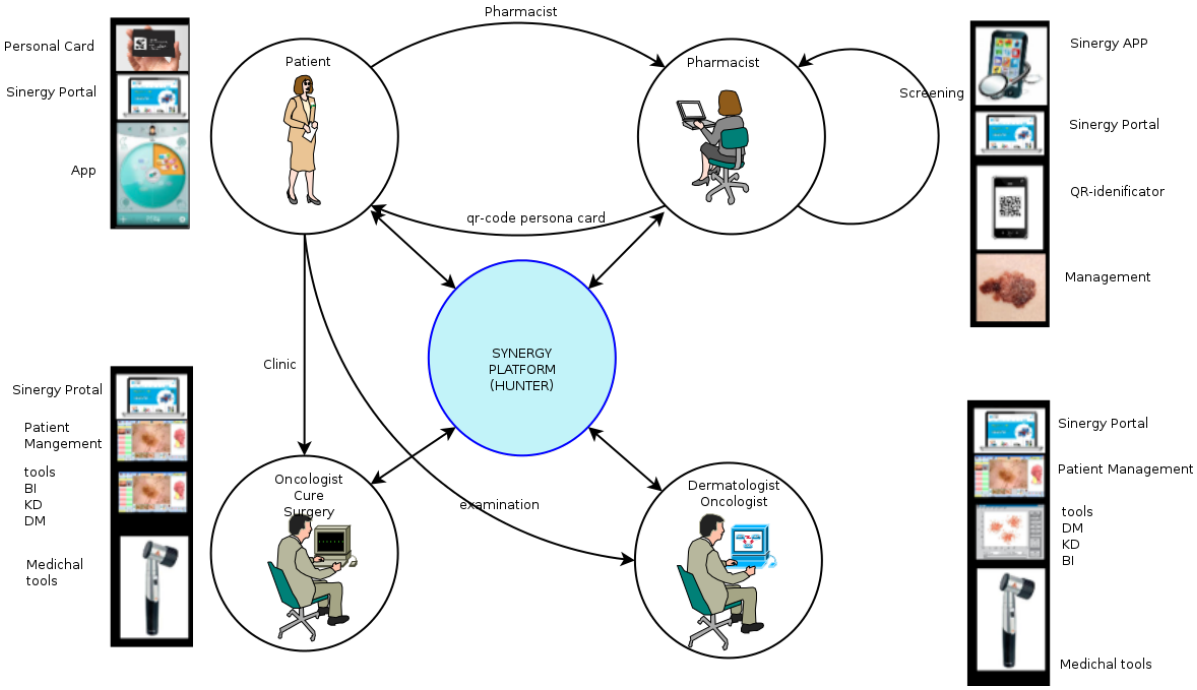


Figure 1. A Sinergy Platform scenario.

screening and analysis of skin lesions. The framework is conceived as the integration of two existing prototypes, UBioLab (Bartocci *et al.*, 2012) and OWL-meaning (Cacciagrano *et al.*, 2012), following the architectural schema described in (Bartocci *et al.*, 2007). In detail, the framework:

- is conceived as a semantic Enterprise Resource Planning (ERP) platform, equipped with (i) a semantic knowledge base storing semantically-annotated information, and (ii) semantic-driven services for managing, analysing, meaning, querying and clustering stored information;
- provides a semantic-driven interface for managing and semantically-annotate the knowledge base, as well as for programming, executing and storing (in the knowledge base as a declarative knowledge) workflow-based procedural knowledge (e.g., business processes, in-silico experiments, new services). This feature, in particular, makes possible mining operations on heterogeneous types of knowledge (e.g., it is possible to reason on in-silico experiments altogether with the involved resources);
- relies on a Cloud-based middleware providing the needed integration mechanism between services and knowledge base and among services themselves. This allows also further services (developed by the framework programming interface) to be plugged in the framework without changing its current architecture.

The core of the framework is a smart knowledge model (Cannata *et al.*, 2005) that allows at contextualising resources w.r.t. a given domain, and activities w.r.t. given resources, making the framework domain-independent. Selecting a specific domain ontology from a repository suffices to customise the framework for a specific domain (that one conceptualised by the loaded ontology).

Moving information from people to doctor (and not the vice versa) is the first step to spread screening facilities and to considerably improve the early detection of skin cancers.

This is the reason why a smartphone-based screening service has been already developed to be plugged into the framework. It enables pharmacists to acquire skin images and to store them in the (remote) knowledge base using conventional smartphones equipped with cheap dermoscopic cameras.

As shown in Figure 1, different actors can participate at the process of data entry and analysis. Pharmacists are driven in the rule of data

entry by the screening service. Data are annotated (on the basis of a suitable multi-scale and multi-physics domain ontology loaded from the framework repository) by the semantic annotation service.

Once information of patients and acquired images are stored on the knowledge base, dermatologists and oncologists can (1) visualise them by the knowledge management service; (2) use image filter services; (3) classify skin lesions by the clustering service (that, differently from a conventional ABCD, can mix images features with different type of information, like age, social aspect, job, location, family etc.); (4) plan altogether a response and, if necessary, call a specific patient for the frontal examination.

Results and Discussion

Currently, we are testing the screening service over one hundred patients, with the collaboration of L.I.L.T. voluntaries and ASP - Azienda Publlica di Servizi alla Persona "A. Chierichetti" of Gagliole (Macerata, Italy).

For what concerns the Dermatology domain (with focus on melanoma cancer and skin lesions), defining and engineering a semantic multi-scale and multi-physics skin lesion ontology is crucial.

At the moment, the domain ontology takes into account only demographic and social concepts linked to a FMA Ontology subgraph.

To improve the domain conceptualisation, we plan to extend the basis integrating other well-known standard biomedical vocabulary and ontologies. In detail, XML-based medical standards like HL7/CDA, standards for medical image acquisition like DICOM, nomenclature for describing dermatological disorders as PROVOKE, ONTODerm and DermLex, vocabulary for describing diseases as ICD (version 9 and 10) and SNOMED CT, ontologies like UMLS and Gene Ontology.

Acknowledgements

This work has been supported by Lega Italiana Lotta Tumori (L.I.L.T.) grant - Project Sinergy (2013).

References

Athale C, Deisboeck TS (2006) The effects of EGF-receptor density on multiscale tumor growth patterns. *J Theor Biol* **238**(4), 771–779. doi:10.1016/j.jtbi.2005.06.029

Bartocci E, Cacciagrano D, *et al.* (2007) An Agent-based Multilayer Architecture for Bioinformatics Grids. *IEEE Trans Nanobioscience* **6**(2), 142-148. doi:10.1371/journal.pcbi.0010076

Bartocci E, Cacciagrano D, *et al.* (2012) UBioLab: a web-LABoratory for Ubiquitous in-silico experiments. *J Integr Bioinform* **9**(1), 192. doi:10.2390/biecoll-jib-2012-192

Boldrick JC, Layton CJ, *et al.* (2007) Evaluation of digital dermoscopy in a pigmented lesion clinic: clinician versus computer assessment of malignancy risk. *J Am Acad Dermatol* **56**(3), 417–421. doi:10.1016/j.jaad.2006.08.033

Cacciagrano D, Merelli E, *et al.* (2012) Semantics on the Cloud: toward an Ubiquitous Business Intelligence 2.0 ERP Desktop. *Proceedings of the Sixth International Conference on Advances in Semantic Processing 2012 (SEMAPRO 2012)* 23-28 September 2012, Barcelona, Spain. Cacciagrano D, Dini P (eds). Curran Associates Inc., Red Hook, NY, USA. Pp. 42-47.

Cannata N, Merelli E, Altman RB (2005) Time to organize the bioinformatics resourceome. *PLoS Comput Biol* **1**(7), e76. doi:10.1371/journal.pcbi.0010076

Helmbach H, Rossmann E, *et al.* (2001) Drug-resistance in human melanoma. *Int J Cancer* **93**(5), 617–622. doi:10.1002/ijc.1378

Liu Y, Purvis J, *et al.* (2007) A multiscale computational approach to dissect early events in the Erb family receptor mediated activation, differential signaling, and relevance to oncogenic transformations. *Ann Biom Eng* **35**(6), 1012–1025. doi:10.1007/s10439-006-9251-0

Soengas MS, Lowe SW (2003) Apoptosis and melanoma chemoresistance. *Oncogene* **22**(20), 3138–3151. doi:10.1038/sj.onc.1206454

Soyer HP, Argenziano G, *et al.* (2004) Three-point checklist of dermoscopy. A new screening method for early detection of melanoma. *Dermatology* **208**(1), 27–31. doi:10.1159/000075042

Ziemys A, Kojic M, *et al.* (2011) Hierarchical modeling of diffusive transport through nanochannels by coupling molecular dynamics with finite element method. *J Comput Phys* **230**(14), 5722–5731. doi:10.1016/j.jcp.2011.03.054

Zheng X, Wise SM, Cristini V (2005) Nonlinear simulation of tumor necrosis, neo-vascularization and tissue invasion via an adaptive finite-element/level-set method. *Bull Math Biol* **67**(2), 211–259. doi:10.1016/j.bulm.2004.08.001

JavaEE for breakfast: start off on the right foot developing biological Web applications

Arnaud Ceol✉, **Heiko Mulle**

Center for Genomic Science of IIT@SEMM, Istituto Italiano di Tecnologia (IIT), Milan, Italy

Received 15 September 2013; **Accepted** 17 September 2013; **Published** 14 October 2013

Competing interests: the authors have declared that no competing interests exist.

Abstract

Bioinformaticians may be intimidated by the apparent complexity of programming languages such as Java and stick to efficient but less suited languages for building rich applications and web sites. Nevertheless, with Java Enterprise Edition (EE), the part of programming is drastically reduced and many tools and frameworks allow for building professional applications rapidly. Indeed in order to access local or remote data, to extend existing software or build user friendly, and good looking, web interfaces, most of the pieces are already provided. The developer can assemble and finalise the applications with his/her own business classes, focusing on the biological concepts. Here we present an overview of some useful and easy to use tools and frameworks based on Java programming, and provide as an example a simple web interface to browse annotations associated with cancer samples provided by the Cancer Genome Atlas (TCGA).

Motivation and Objectives

Bioinformaticians may be intimidated by the apparent complexity of programming languages such as Java and stick to efficient but less suited languages for building rich applications and web sites. Nevertheless, with Java Enterprise Edition (EE) the part of programming is drastically reduced, and many tools and frameworks allow for building professional applications rapidly. Indeed in order to access local or remote data, to extend existing software or build user friendly, and good looking, web interfaces, most of the pieces are already provided. The developer can assemble and finalise the applications with his/her own business classes, focusing on the biological concepts. Here we present an overview of some useful and easy to use tools and frameworks based on Java programming and provide as an example, a simple web interface[1] to browse annotations associated with cancer samples provided by the Cancer Genome Atlas (TCGA).

Methods

Many tools and frameworks help Java programmers to build professional applications. Thanks to these tools, bioinformaticians with less experience in software development may forget about complex implementations of many features that are already provided and focus on the goal and design of the applications.

Managing data easily

Biological data do not always require the complexity of a relational database. Solr[2], a Java open source search platform, provides a complete and easy solution for storing data from heterogeneous formats (including text, XML, doc and pdf). The data is indexed and searchable in a "Google" like fashion. In addition, Solr provides a web interface to query and manage the data, and a web service to access it locally or remotely from a browser or a remote application (Figure 1a).

Remote access to data

In the era of cloud computing, it is increasingly less necessary to store local copies of external voluminous databases. Indeed, many biological repositories provide computational access (web service) to their data that can be used to query and retrieve information from them directly from within a program or a web site. Java EE[3] provides a simple interface to implement a client to a web service (Figure 1b).

Building an application and managing dependencies

One of the most tedious tasks in software programming, and even more so for the end user during the installation, is the management of dependencies. Maven[4], a software management and comprehension tool allows to drastically simplify both processes. A configuration file stores the names of the packages which the ap-

1 https://cru.genomics.iit.it/TcgaAnnotationBrowser

2 https://lucene.apache.org/solr/
3 http://www.oracle.com/technetwork/java/javaee/
4 https://maven.apache.org/

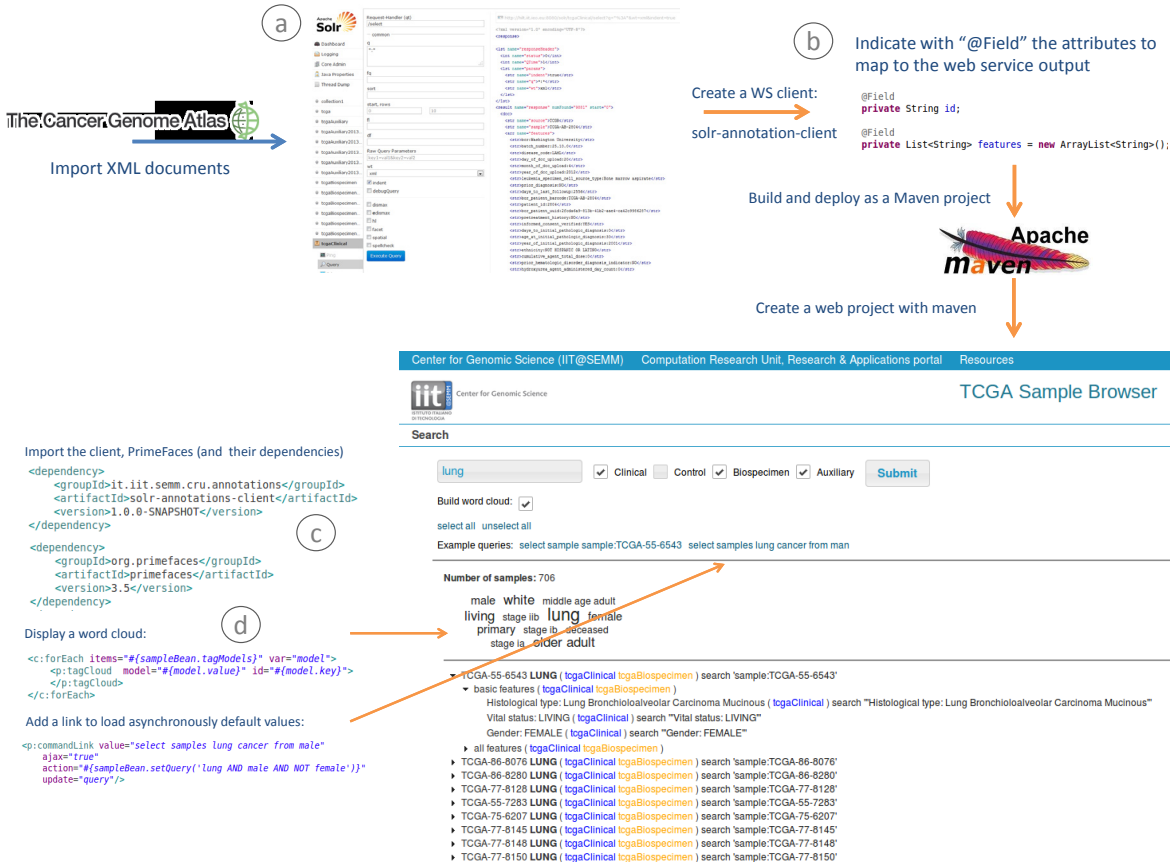


Figure 1. Building a web application. a) Annotations from TCGA are downloaded in XML format and imported into Solr. Solr includes a web interface that allows to query the data. b) Create a client to access the data on Solr. An annotation is enough for the client to understand which attributes should be mapped to the output of the web service (WS). c) Import the client and PrimeFace into the web application. With Maven, it is enough to add them into the configuration file. Maven will resolve and download all the dependencies for those packages. d) Create asynchronous links (ajax) or word cloud with PrimeFaces.

plication directly depends on. Maven then takes care of downloading all those packages and their own dependencies. Any piece of software developed with maven can be successively imported in other Maven projects (Figure 1c).

Building dynamic and good looking, web applications

When it comes to building dynamic and professional websites, the inexperienced bioinformatician may soon be lost in developing workflow of data between the server and the client, or digging and adapting complex JavaScript pieces of code. The last versions of Java EE include the JavaServer Faces (JSF) specification to help building web sites. On top of JSF, PrimeFaces[5] for example provides a user interface kit that can be easily integrated into the user web pages to allow for drawing tables that can be ordered or filtered by the user, to update asynchronously part of the pages, or displaying word clouds (Figure 1d). In addition, it includes a component tool kit oriented to mobile devices.

Results and Discussion

We have presented a series of tools based on Java technology that can help bioinformaticians to build professional applications. The tools and their usage is not exhaustive. We did not mention the possibility of extending existing Java applications with plugins (e.g. Cytoscape) (Saito *et al.*, 2012) and the Integrated Genome Browser (Nicol *et al.*, 2009), or the advantages of mapping relational databases to objects. These tasks have been made easy with the Java Enterprise Edition.

5 http://primefaces.org/

To illustrate this, we developed a web interface for browsing the annotations of cancer samples from TCGA[6]. The XML documents where imported into a local installation of Solr and implemented a client for the web service provided that was with Maven. This package is used by a web page that we developed using the PrimeFaces toolkit. Our client allows the web application to retrieve the annotations in a table, and the components from PrimeFaces display the results both as a table and as a word cloud that highlight the annotations most commonly associated to the cancer samples associated to the query.

Other tools and frameworks exist for other languages. Django[7] and Symphony[8], for instance, make it easy to build dynamic websites with Python or PHP, and JavaScript libraries like jQuery[9], on which PrimeFaces relies, can be integrated in any web page. But in our opinion none of those languages provides the same amount of tools and frameworks as Java does. They are easy to install and manage and are well supported by bioinformatics institutes like the European Bioinformatics Institutes (EBI). EBI's Maven repository[10] contains packages for accessing and managing data about proteins (Patient *et al.*, 2008), chemicals (Deng *et al.*, 2011), molecular interactions (Aranda *et al.*, 2011; Kerrien *et al.*, 2012) and many others. Biological libraries like BioJava (Prlic *et al.*, 2012) or the structure visualization tool Jmol (Herráez, 2006) are similarly available.

We believe that the adoption of the technologies presented here can benefits to the whole scientific community by improving the quality of web tools while diminishing the time spent on software development.

References

Aranda B, Blankenburg H, *et al.* (2011) PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nat Methods* **8**(7), 528-529. doi:10.1038/nmeth.1637

Deng N, Zhang J, *et al.* (2011) Phosphoproteome analysis reveals regulatory sites in major pathways of cardiac mitochondria. *Mol Cell Proteomics* **10**(2), M110.000117. doi:10.1074/mcp.M110.000117

Herráez A (2006) Biomolecules in the computer: Jmol to the rescue. *Biochem Mol Biol Educ.* **34**(4), 255-61. doi:10.1002/bmb.2006.494034042644

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

Nicol JW, Helt GA, *et al.* (2009) The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. *Bioinformatics (Oxford, England)* **25**(20), 2730-2731. doi:10.1093/bioinformatics/btp472

Patient S, Wieser D, *et al.* (2008) UniProtJAPI: a remote API for accessing UniProt data. *Bioinformatics* **24**(10), 1321-1322. doi:10.1093/bioinformatics/btn122

Prlic A, Yates A, *et al.* (2012) BioJava: an open-source framework for bioinformatics in 2012. *Bioinformatics* **28**(20), 2693-2695. doi:10.1093/bioinformatics/bts494

Saito R, Smoot ME, *et al.* (2012) A travel guide to Cytoscape plugins. *Nat Methods* **9**(11), 1069-1076. doi:10.1038/nmeth.2212

6 https://tcga-data.nci.nih.gov/tcga/
7 https://www.djangoproject.com/
8 http://symfony.com/
9 http://jquery.com/
10 http://www.ebi.ac.uk/~maven/m2repo/

Describing the genes social networks relying on chromosome conformation capture data

Ivan Merelli[1] ✉, **Pietro Liò**[2], **Luciano Milanesi**[1]

[1]Institute for Biomedical Technologies - National Research Council of Italy, Italy
[2]University of Cambridge, United Kingdom

Received 31 July 2013; **Accepted** 6 September 2013; **Published** 14 October 2013

Competing interests: the authors have declared that no competing interests exist.

Abstract

The analysis of the chromosomal conformation in the nucleus is critical for understanding mechanisms that regulate genes colocalisation and coexpression. As for social networks that combine information about positions and activities of users, we show how it is possible, by using an R package developed for this purpose, to correlate the three-dimensional organisation of chromosome in the nucleus and gene activity during stem cell differentiation.

Motivation and Objectives

In the social network society it is often difficult to organise information in a systemic view. In all fields, from information technology and sociology to biomedical science, and in particular genetics, the description of the interactions established by vertices in a network is incredibly important (Barabási *et al.*, 1999). Nowadays, through mobile devices localization it is possible to provide suitable information about facilities and opportunities in the neighbourhood of the users. Social networks such as FourSquare make of positioning the core of their business. Information about localizations combined with a constant encouragement of people to describe their activities, allows tracing mass movement and behavior (Noulas *et al.*, 2011). The collection of such information can be extremely useful, for example, in marketing, in order to target advertisements and promotions.

These concepts have been already applied to medicine, which in future will be participatory and personalised (Hood, 2013) by mean of social networks, such as PatientsLikeMe. Moreover, future medicine is expected to be predictive and preventive (Hood, 2013), by fully exploiting the integration with omics science, in particular for understanding how the 3D nuclear maps of genes can be exploited for precisely targeting new drugs. Noteworthy, both social networks and molecular networks could use semantic information, part of which could be shared across these domains. From a human cognitive behavior, it would be ideal to have similar ways for querying different networks and perhaps molecular and cellular information could be even linked to patients' social networks. Can we design 3D nuclear information the same way we design social networks?

Despite several hundreds human genomes have been sequenced, we know very little about three-dimensional chromosome conformation beyond the scale of the nucleosome. Considering the number of evidences about colocalisation and coregulation of genes, this is very important in describing the social behavior of genomic actors (Di Stefano *et al.*, 2013). In particular, recent advances in high throughput molecular biology techniques and bioinformatics have provided insights into chromatin interactions on a larger scale (Lieberman-Aiden *et al.*, 2009). A novel sequencing technique called Chromosome Conformation Capture (3C) allows analysing the organisation of chromosomes in a cell's natural state (Duan *et al.*, 2012). While performed genome wide, this technique is usually called Hi-C. Clearly, studying the structural properties and spatial organisation of chromosomes is important for the understanding and evaluation of the regulation of gene expression, DNA replication and repair, and recombination (Lin *et al.*, 2012).

Methods

Inspired by social networks like FourSquare, we developed NuChart (Merelli *et al.*, PLoS One, accepted), an R package that integrates Hi-C information, describing the chromosomal neighborhood, with predicted CTCF binding sites (Botta *et al.*, 2010), isochores (Marculescu, *et al.*, 2006), potential cryptic RSSs (Varriale and Bernardi, 2010), and other user desired genomics features, such as methylation and chromatin conformation, to infer how the three-dimensional organisa-

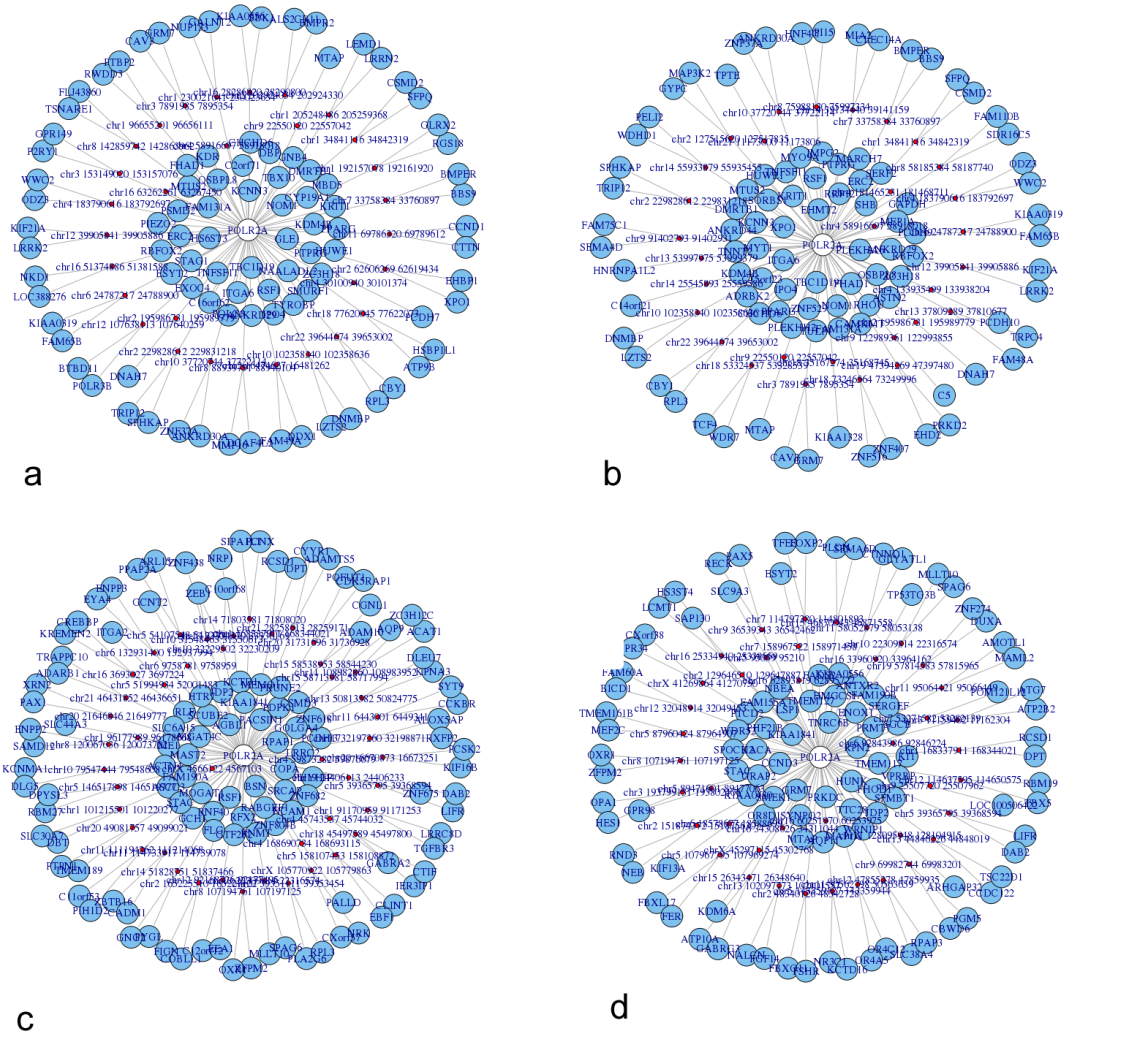


Figure 1. Neighborhood graph of the gene POLR2A in four different runs from the Hi-C experiments of Dixon to show inter and intra run modifications. In the panel a) and b) on the top part of the figure, the sequencing runs are from the same cell line hESC, while panel c) and d) are from IMR90.

tion of DNA works in controlling gene expression. This can be very useful while studying the differentiation of stem cells or to identify chromosomal reorganisations in cancer cells. For example, Philadelphia translocation is a specific chromosomal abnormality that is associated with chronic myelogenous leukemia (CML). It is the result of a reciprocal translocation between chromosome 9 and 22. The software has been designed to answer question such as: are chromosomal translocations occurring between nearby chromosomes?

Beside functions for loading and normalizing data, the core of NuChart is the creation of the neighborhood graph of the user provided list of genes or pathway. This package provides the possibility of analysing Hi-C data in a multi-omics context, by enabling the capability of mapping on the graph vertices expression data, according to a particular transcriptomics experiment, and on the edges genomic features that are known to be involved in chromosomal recombination, looping and stability. At the same way of FourSquare, the relative positions of the social actors and their functional activities are the core information for describing the neighborhood global behavior.

NuChart also provides functions to describe, compare, and analyse statistically the neighborhood graphs after their creation, which can be very important to highlight local and global characteristics of the Hi-C fragment distributions in the nucleus and of the multi-omics features in the context of the DNA three-dimensional topology. The possibility of analysing data to infer structural-activity relationships in a social network is of critical importance (Reagans and McEvily, 2003).

Results and Discussion

In the following example, we discuss an interesting analysis related to the data of Dixon *et al.* (Dixon *et al.*, 2012) experiments. The idea is to show the different chromosomal organisations that occur in the nucleolus, while gene expression is heavily characterising the differentiation of stems cells. Respectively, the graphs in Figure 1 in the top part are from two different sequencing runs performed on human embryonic stem cells (SRA:SRR400261 and SRA:SRR400262), while the graphs in the bottom part are from human lung embryonic fibroblast (SRA:SRR400264 and SRA:SRR400265).

In particular, these graphs show the neighborhood of the POLR2A gene, which catalyses the transcription of DNA to synthesize precursors of mRNA and most snRNA and microRNA, in the different cell lines. Noteworthy, the variability in the neighborhood of this gene, computed as correlation between the lists of adjacent genes, in the different cell lines is significant. While the similarity between two different runs of sequencing performed on the same cell type is quite high (respectively 60% and 67%), there are very importance differences between the two different cell lines (correlation below 30%), which witness the importance that chromosomal re-organisations at nucleolus level has for co-expression. This is very important to understand, in a particular moment, what the cell is going to express, by re-organising its internal chromosomal structure in the three dimensional space.

Recalling the parallelism with FourSquare, the power of these tools relies in capturing and describing the colocalisation and co-activation of entities in the social network. Moreover, the interaction of the social actors with the environment is of critical importance for understanding dynamics of the whole system. Future medicine will require the integration of different social and genetic networks in a multilevel approach: therefore, the possibility of having topological coherent graph descriptions and overlapping semantics for annotations across these two domains will be very important.

Acknowledgements

This work has been supported by the Italian Ministry of Education and Research (MIUR) through the Flagship (PB05) InterOmics, HIRMA (RBAP11YS7K) and the European MIMOMICS projects.

References

Barabási A-L, Albert R (1999) Emergence of Scaling in Random Networks. *Science* **286**(5439), 509-512. doi:10.1126/science.286.5439.509

Botta M, Haider S, *et al.* (2010) Intra- and inter-chromosomal interactions correlate with CTCF binding genome wide. *Mol Syst Biol.* **6**, 426. doi:10.1038/msb.2010.79

Di Stefano M, Rosa A, *et al.* (2013) Colocalization of Coregulated Genes: A Steered Molecular Dynamics Study of Human Chromosome 19. *PLoS Comput Biol* **9**(3), e1003019. doi:10.1371/journal.pcbi.1003019

Dixon JR, Selvaraj S, *et al.* (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**(7398), 376-380. doi:10.1038/nature11082

Duan Z, Andronescu M, *et al.* (2012) A genome-wide 3C-method for characterizing the three-dimensional architectures of genomes. *Methods* **58**(3), 277-288. doi:10.1016/j.ymeth.2012.06.018

Hood L (2013) Systems Biology and P4 Medicine: Past, Present, and Future. *Rambam Maimonides Med J.* **4**(2), e0012. doi:10.5041/RMMJ.10112

Lieberman-Aiden E, van Berkum NL, *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**(5950), 289-293. doi:10.1126/science.1181369

Lin YC, Benner C, *et al.* (2012) Global changes in the nuclear positioning of genes and intra- and inter- domain genomic interactions that orchestrate B cell fate. *Nat Immunol.* **13**(12), 1196-1204. doi:10.1038/ni.2432

Marculescu R, Vanura K, *et al.* (2006) Recombinase, chromosomal translocations and lymphoid neoplasia: targeting mistakes and repair failures. *DNA Repair (Amst)* **5**(9-10), 1246-1258.

Noulas A, Scellato S, *et al.* (2011) An Empirical Study of Geographic User Activity Patterns in Foursquare. *Proceedings of the Fifth International Conference on Weblogs and Social Media ICWSM 2011*, Barcelona, Catalonia, Spain, July 17-21, 2011. The AAAI Press. Palo Alto, CA, USA. pp. 70-73.

Reagans R, McEvily B (2003) Network structure and knowledge transfer: The effects of cohesion and range. *Administrative science quarterly* **48**(2), 240-267.

Varriale A, Bernardi G (2010) Distribution of DNA methylation, CpGs, and CpG islands in human isochores. *Genomics* **95**(1), 25-28. doi:10.1016/j.ygeno.2009.09.006

An ontology describing congenital heart defects data

Charalampos Moschopoulos[1,2]✉, **Jeroen Breckpot**[3], **Yves Moreau**[1,2]

[1]Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, Katholieke Universiteit Leuven, Leuven, Belgium
[2]iMinds Future Health Department, KU Leuven, Belgium
[3]Center for Human Genetics, KU Leuven, Leuven

Received 9 September 2013; **Accepted** 9 September 2013; **Published** 14 October 2013

Competing interests: the authors have declared that no competing interests exist.

Abstract

Congenital heart defects (CHDs) are a group of diseases characterized by a structural anomaly of the heart that is present at birth. They are considered as the commonest cause of childhood death in developed countries. The causes of congenital heart disease are still under investigation, but their strongly presumed to be genetic or a combination of genetic and environmental factors. In order to store the derived knowledge, a collaborative knowledge base called CHDWiki has been developed, which collects all the information about the genetic basis of CHDs. However, this dedicated web resource suffers the same problems with similar portals where heterogeneous information is hosted. Further steps should be taken in order to ensure the interoperability of the available data. Also, the hosted data should be offered in a machine-readable format in order to direct be used by other Bioinformatic tools. In order to solve these problems, the life science scientific community tends to use semantic web technologies, which have proved their efficiency through numerous examples including VariO ontology, Gene Ontology (GO), Orphanet Ontology of rare diseases (OntoOrpha) and many more. In this contribution, we present an ontology, which describes CHD data, developed around three main data categories: genotype, phenotype (CHDs) and clinical reports. Retrieving data from CHDWiki, this ontology describes the relationships between genes and human phenotypes, derived from published data or single clinical cases, providing a useful tool to geneticists, molecular biologists and clinicians. This ontology hosts information about syndromic genes, chromosomal aberrations that may cause CHDs and associations between genes and CHDs. Further information is included such as structural variations and single point mutations.

Motivation and Objectives

Congenital heart defects (CHDs) are a group of diseases characterised by a structural anomaly of the heart that is present at birth. They are considered as the commonest cause of childhood death in developed countries; see, e.g., the web site of the American Heart Association[1]. The causes of congenital heart disease are still under investigation, but they are strongly presumed to be genetic or a combination of genetic and environmental factors (Jenkins *et al.*, 2007). In order to store the derived knowledge, a collaborative knowledge base called CHDWiki (Barriot *et al.*, 2010) has been developed, which collects all the information about the genetic basis of CHDs.

However, this dedicated web resource suffers the same problems with similar portals where heterogeneous information is hosted. Further steps should be taken in order to ensure the interoperability of the available data. Also, the hosted data should be offered in a machine-readable format in order to directly be used by other bioinformatic tools. In order to solve these problems, the life science scientific community tends to use semantic web technologies, which have proved their efficiency through numerous examples including VariO ontology[2], Gene Ontology (GO) (Ashburner *et al.*, 2000), Orphanet Ontology of rare diseases (OntoOrpha) (Aime *et al.*, 2012), and many more.

In this contribution, we present an ontology, which describes CHD data, developed around three main data categories: genotype, phenotype (CHDs) and clinical reports. Retrieving data from CHDWiki, this ontology describes the relationships between genes and human phenotypes derived from published data or single clinical cases, providing a useful tool to geneticists, molecular biologists and clinicians. This ontology hosts information about syndromic genes, chromosomal aberrations that may cause CHDs and associations between genes and CHDs. Further information is included such as structural variations and single point mutations.

Methods

As mentioned before, the CHD ontology is composed by three main concepts (genes, CHDS and clinical reports) and the between them relationships. These three concepts are further classified using a `is_a` relation as it can be seen in Figure 1. The CHD family of diseases is hierarchi-

1 http://www.heart.org/HEARTORG

2 http://variationontology.org/

cally structured, starting from the CHD concept which is further refined in 11 subclasses (e.g., "Abnormalities of atriums and atrial septum", "Abnormalities of great veins", "Rhythm and conduction disturbances", etc). These concepts, in some cases, are further refined to more specific congenital heart defects. The other two main classes of the CHD ontology are also refined: the gene entities, which are subclasses of the concept Gene, are belonging either to the subclasses of the "Non Syndromic Genes" or "Syndromic Genes". The clinical case entities fall into one of the following categories: "ASHG" which stands for American Society of Human Genetics, "CME _ Leuven _ Bench" which refers to unreported patients with causal CNVs from Leuven Hospital, and "PMID" which stands for case reports in literature (PubMed ID).

Two further relationships were used in our ontology: `connected _ with`, which describes an association between a gene and a CHD, and `involved _ in`, which describes an association between a gene or a CHD and a case report. It has to be noted that CHDWiki users have manually curated each association between entities, providing high quality data.

We also created a vocabulary that describes the relationships between the ontology terms. As we consider the extensibility of ontologies a very important property, we used words into our vocabulary derived from two well-known ontologies: the Dublin Core Metadata Initiative[3] (DCMI) and the RDF Schema[4] (RDFS). Each class includes a set of properties such as structure variation (where deviation, chromosome, start and end region are referred) and single point mutation (where DNA mutation, peptide mutation and relevant reference are referred). Moreover, whenever was possible, we used external URIs for each class property. For instance, the property "label" of a gene refers to the corresponding gene page at Ensemble web portal. Also a variety of external URLs are provided under the "seeAlso" property of each Gene and CHD subclass.

To annotate the basic CHD ontology classes using unique URIs, we used well-known IDs such as the shortlist of the Association for European Pediatric Cardiology[5] (AEPC) for the CHDs, the Hugo IDs (Gray *et al.*, 2013) for the associated genes and the CHDWiki case report IDs for the recorded clinical cases. Specifically, the AEPC URI annotation for the CHDs is considered more accurate than using OMIM[6] annotation as this disease family is not so well characterised by it. Besides that, it has to be noted that the created URIs of the CHDs and the clinical cases refer to CHDWiki pages, while the associated gene URIs refer to the HUGO Gene Nomenclature Committee[7] (HGNC) web pages.

Table1. The class hierarchy of the CHD ontology (snapshot from Protégé 4.2).

- CaseReport
  - ASHG_2007
  - CME_Leuven_Bench
  - PMID
- CHD
  - Abnormalities_of_atriums_and_atrial_septum
  - Abnormalities_of_AV_valves_and_AV_septal_defect
  - Abnormalities_of_coronary_arteries_and_arterial_duct_and_pericardium
  - Abnormalities_of_great_veins
  - Abnormalities_of_position_and_connection_of_heart
  - Abnormalities_of_VA_valves_and_great_arteries
  - Abnormalities_of_ventricles_and_ventricular_septum
  - Diagnostic_congenital_and_generic_cardiac_codes
  - Heart_muscle_disease_and_cardiomyopathies
  - Rhytm_and_conduction_disturbances
  - Tetrallogy_of_Fallot_and_variants
- Gene
  - Non_Syndromic_Genes
  - Syndromic_Genes

Results and Discussion

In our project, we used Protégé 4.2[8] and the ontology was built in OWL format. We chose not to use OBO format as OWL is a more expressive language and we could also describe the relations between our ontology classes, which were not always an `is _ a` instance. As Protégé reasoner presents some limitations concerning the class properties querying, we also created our ontology in simple RDF turtle format. Our ontology is also hosted at the Bioportal (Whetzel *et al.*, 2011), which is the most popular open repository of biomedical ontologies.

As a future work, we will continue the enrichment of the created ontology and will test its biological usefulness by applying queries on very specific research questions. Also, we plan to connect the CHDWiki data ontology with other well-known bio-ontologies such as the Human Phenotype Ontology (HPO) (Robinson *et al.*, 2008). This way, queries with increased information value for doctors and geneticists could be applied. Finally, we plan to create a SPARQL[9] endpoint that, accompanied with a friendly in-

3 http://dublincore.org/
4 http://www.w3.org/TR/rdf-schema/
5 http://www.aepc.org/
6 http://omim.org/
7 http://www.genenames.org/
8 http://protege.stanford.edu/
9 http://www.w3.org/TR/rdf-sparql-query/

terface, will be hosted on CHDWiki portal. The goal is to encourage untrained users to query the available data without having any previous knowledge regarding semantic technologies such as RDF and SPARQL. While the scientific interest on rare diseases is continuously increasing, ontologies could play a vital role into the integration and further analysis of the generated data

Acknowledgements

We wish to thank Dr. Peter Robinson (Charité - Universitätsmedizin Berlin) for the helpful discussions.

Funding: Research supported by: Research Council KU Leuven: GOA/10/09 MaNet, KUL PFV/10/016 SymBioSys, IOF: HB/12/022 Endometriosis. Flemish Government: FWO: PhD/postdoc grants, projects: Hercules Stichting: Hercules III PacBio RS, iMinds: SBO 2013. Federal Government: FOD: Cancer Plan 2012-2015 KPC-29-023 (prostate). Action BM1006: NGS Data analysis network.

References

Aime X, Charlet J, *et al.* (2012) Rare diseases knowledge management: the contribution of proximity measurements in OntoOrpha and OMIM. *Studies in health technology and informatics* **180**, 88-92. doi:10.3233/978-1-61499-101-4-88

Ashburner M, Ball CA, *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29. doi:10.1038/75556

Barriot R, Breckpot J, *et al.* (2010) Collaboratively charting the gene-to-phenotype network of human congenital heart defects. *Genome Med* **2**, 16. doi:10.1186/gm137

Gray KA, Daugherty LC, *et al.* (2013) Genenames.org: the HGNC resources in 2013. *Nucleic Acids Res.* **41**, D545-552. doi:10.1093/nar/gks1066

Jenkins KJ, Correa A, *et al.* (2007) Noninherited risk factors and congenital cardiovascular defects: current knowledge: a scientific statement from the American Heart Association Council on Cardiovascular Disease in the Young: endorsed by the American Academy of Pediatrics. *Circulation* **115**, 2995-3014. doi:10.1161/CIRCULATIONAHA.106.183216

Robinson PN, Kohler S, *et al.* (2008) The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am J Hum Genet* **83**, 610-615. doi:10.1016/j.ajhg.2008.09.017

Whetzel PL, Noy NF, *et al.* (2011) BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Res.* **39**(Web Server Issue), W541-545. doi:10.1093/nar/gkr469

Towards a semantic wiki for human and animal cell lines

Paolo DM Romano[1]✉, **Dan M Bolser**[2]

[1]IRCCS AOU San Martino IST, Genova, Italy
[2]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, United Kingdom

Received 16 September 2013; **Accepted** 18 September 2013; **Published** 14 October 2013

Competing interests: the authors have declared that no competing interests exist.

Abstract

Wiki systems have emerged as a social network tool able to stimulate users to contribute to the collaborative building of a common knowledge base. Some of the specific aims of wikis for biology (bio-wikis) include collaborative efforts for the development and sharing of knowledge, the annotation of database contents, and the creation of database contents. Indeed, the increasing volumes of biological data cannot be adequately managed by the existing centralised databases and database curators. On the contrary, information and data must be exchanged by the whole community. Special features may be required to adapt to the specificity of biological data. Moreover, integration with the most relevant biological databases must be implemented. In this context, Semantic technologies promise to be of paramount relevance. In this abstract, we present a preliminary analysis for the implementation of a semantic wiki system for human and animal cell lines derived from the Cell Line Data Base (CLDB) and tightly connected to external relevant resources.

Motivation and Objectives

It is well known that human cell lines constitute one of the most useful biological resources for biomedical research since they represent an optimal model for many assays. For their best exploitation, it is essential that cell lines are properly stored, characterised, maintained, distributed and used. Biological Resources Centres offer an adequate infrastructure for these aims. More and more, their resources and services are promoted through the Internet, by means of web sites, databanks and catalogues. However, knowledge concerning cell lines is not restricted to Biological Resources Centers, and a great wealth of knowledge, information, and data, is owned by researchers throughout the world: for example, those developing cell lines, investigating the properties of cell lines, or using cell lines in their own experiments. Not to mention the knowledge and facts relevant to cell lines that are to be found in the scientific literature. This additional information may be highly relevant and interesting for researchers, stimulating the search and investigation of known characteristics of resources before their selection, the comparison of different resource features, the analysis of previous behaviours and responses from cell lines, and, finally, the selection of the most adequate and effective research tool. To this end, a knowledge base, able to store and organise this collective information could be a valid contribution.

Wiki systems have recently emerged as a network tool able to stimulate users to contribute to the collaborative building of a common knowledge base. Notorious examples exist, e.g., Wikipedia[1], that demonstrates this concrete opportunity. In the Life Sciences, it has already been demonstrated that wiki systems offer a variety of advantages for the management of biological data and information. These include, just to mention a notable few, Gene Wiki[2] (Huss *et al.*, 2008; Huss *et al.*, 2010; Good *et al.*, 2012), a specialised section of Wikipedia[3] aimed at re-organising, extending, and completing its articles related to human genes, WikiGenes[4] (Hoffmann, 2008), a wiki system whose main goal is to encourage the collaborative creation of scientific papers by associating every single text to its author, and WikiPathways[5] (Pico *et al.*, 2008; Kelder *et al.*, 2012), a wiki system aimed at complementing the existing databases of metabolic pathways (KEGG, Reactome, Pathway Commons). A wiki based database of biological databases was also implemented (Bolser *et al.*, 2011).

Some of the specific aims of wikis for biology (bio-wikis) include collaborative efforts for the development and sharing of knowledge, the annotation of database contents, and the creation of database contents. These aims stem from the realisation that the increasing volumes of biological data cannot be adequately managed by the existing centralised databases and database curators. Information and data must be exchanged by the whole community. In the

1 http://www.wikipedia.org/
2 http://en.wikipedia.org/wiki/Gene_Wiki
3 http://en.wikipedia.org/wiki/Portal:Gene_Wiki
4 http://www.wikigenes.org/
5 http://www.wikipathways.org/

future, many collaborative wiki systems can be envisaged.

However several important issues still need to be addressed. For example, i) how reliable are user contributions? ii) what format should annotations take? iii) how can user provided information be feed back into 'authoritative' databases? Special features may be required to cater for the specificity of biological data: textual information is only a small part of biological data, we must cater for the numerous and heterogeneous biological data types, for example, images, plots, and diagrams.

In this abstract, we present some considerations and a preliminary analysis of a possible implementation of a wiki system for human and animal cell lines tightly connected to the Cell Line Data Base[6] (CLDB).

Methods

CLDB (Romano, 2009) and its hypertextual version HyperCLDB[7] are a long established service, offering information on human and animal cell lines since 1990. Included in CLDB, are data on availability of cell lines in some of the most well known European collections and in many Italian research laboratories, together with their main biological characterisation.

We are developing a wiki system (CellLinesWiki) as a collaborative knowledge base for human and animal cell lines for the community of Biological Resources Centers, biobanks, collections and researchers active in this area. CellLinesWiki consists of three layers. Database information is automatically uploaded from CLDB and constitutes the first layer of authoritative information. The second layer is composed of a curated set of contributions about the cell lines described in the database. It is maintained by a limited number of nominated experts in the field, and the creators of cell lines. The third layer is built from end users, authenticated, but not necessarily trusted at the same level, authorised to provide less specific information on cell lines.

MediaWiki[8], an open source package for wiki development written in PHP, was chosen as the starting point for the development of CellLinesWiki due to its wide user and implementation base.

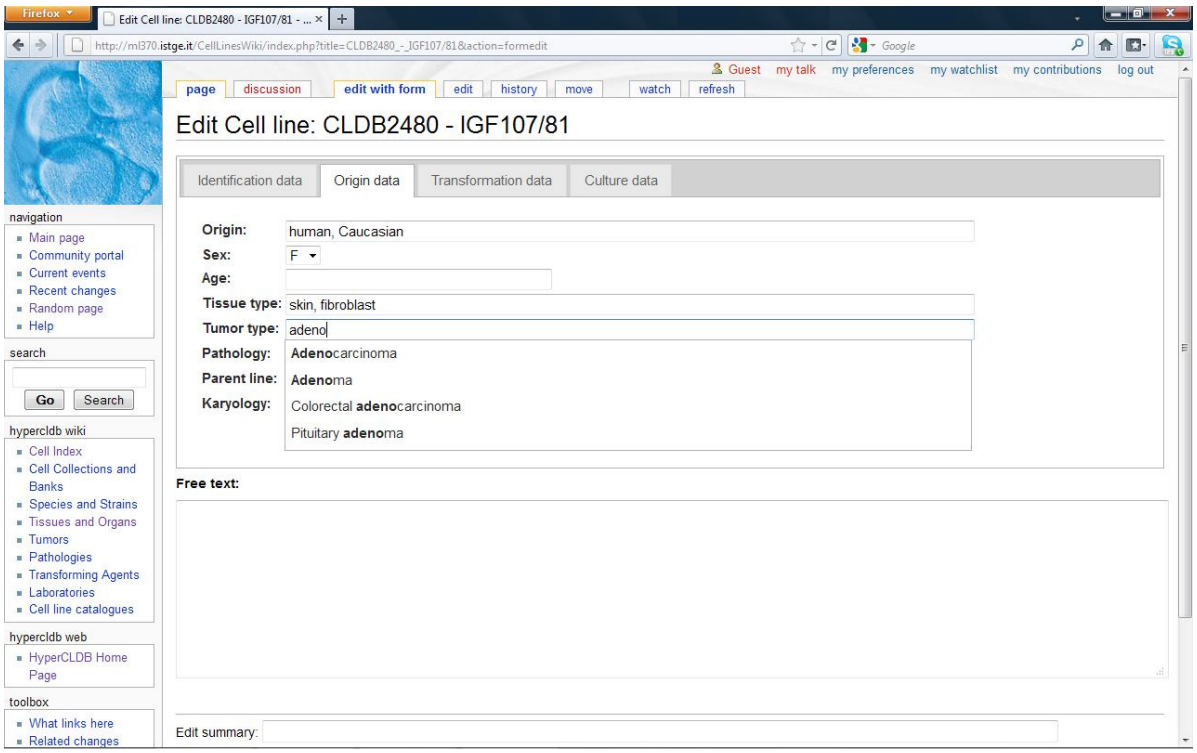


Figure 1. Display of data entry 10429. Protein attributes, features and sequence with transmembrane section (382-40v2) are shown.

6 http://bioinformatics.istge.it/cldb/cldb.php
7 http://bioinformatics.istge.it/hypercldb/
8 http://www.mediawiki.org/wiki/MediaWiki

As we required to store structured data, its extension Semantic MediaWiki[9] (SMW) was also used. Together these tools provide for the collaborative authorship of structure data within the wiki system. Using various wiki extensions we designed a system whereby the data in CLDB can be browsed, queried and updated by known users at the three different levels described.

Results and Discussion

The CellLinesWiki currently contains already information for the 6,632 cell lines, housed within 82 laboratories.

Each cell line may carry information about its associated literature, quality control information, purchasing information and treatments, as well as the associated biological meta-data about the cell lines origin, and transformation details.

Each section of the wiki can be edited by the administrators, while recognised researchers can add further details to their lines of interest. Finally, general users can comment on the information found. Information coming from these three sources is clearly delineated.

In Figure 1, the form that allows granted users to update information on a given cell line is shown. Four distinct tabs include information on data subsets. A contextual help allow users to select one item value from a list as they key some preliminary characters (in the figure, having typed "adeno" four possible values are presented).

As the wiki is in the early stages of development, so far it has only been used internally. The preliminary version of the CellLinesWiki will soon be made publicly available on-line. By promoting the wiki, we hope to engage the community in contribution.

References

Bolser DM, Chibon P-Y, *et al.* (2012) MetaBase - The wiki-database of biological databases. *Nucl. Acids Res.* **40**(Database issue), D1250-D1254. doi:10.1093/nar/gkr1099

Good BM, Clarke EL, *et al.* (2012) The Gene Wiki in 2011: community intelligence applied to human gene annotation. Nucleic Acids Res. 40(Database issue), D1255-1261. doi:10.1093/nar/gkr925

Hoffmann R. (2008) A wiki for the life sciences where authorship matters. *Nat Genet* **40**, 1047-1051. doi:10.1038/ng.f.217

Huss JW, Lindenbaum P, *et al.* (2010) The Gene Wiki: community intelligence applied to human gene annotation. *Nucleic Acids Res.* **38**(Database issue), D633-639. doi:10.1093/nar/gkp760

Huss JW III, Orozco C, *et al.* (2008) A Gene Wiki for Community Annotation of Gene Function. *PLoS Biology* **6**(7), e175. doi:10.1371/journal.pbio.0060175

Kelder T, van Iersel MP, *et al.* (2012) WikiPathways: building research communities on biological pathways. *Nucleic Acids Res.* **40**(Database issue), D1301-1307. doi:10.1093/nar/gkr1074

Pico AR, Kelder T, *et al.* (2008) WikiPathways: Pathway Editing for the People. *PLoS Biol* **6**(7), e184. doi:10.1371/journal.pbio.0060184

Romano P, Manniello A, *et al.* (2009) Cell Line Data Base: structure and recent improvements towards molecular authentication of human cell lines. *Nucleic Acids Res.* **37**(Database issue), D925-D932. doi:10.1093/nar/gkn730

9 http://semantic-mediawiki.org/wiki/Semantic_MediaWiki

SEBSem: simple and efficient biomedical semantic relatedness measure

Maciej Rybinski✉, **José Francisco Aldana-Montes**

Departmento de Lenguajes y Ciencias de la Computacion, Universidad de Málaga, Spain

Received 31 July 2013; **Accepted** 6 September 2013; **Published** 14 October 2013

Competing interests: the authors have declared that no competing interests exist.

Abstract

Calculating semantic relatedness between terms is crucial in numerous knowledge and information processing tasks highly relevant to the biomedical domain. Examples include semantic search and automated processing of scientific texts. Most available methods rely heavily on highly specialised resources, which substantially limits their reusability in various applications within the domain. In this work we present a simple semantic relatedness measure that relies only on very general resources and its design features allow minimising the costs of online computations. The relatedness is computed through comparing automatically extracted key-phrases relevant to respective input terms. This simple strategy provides a method that gives promising early test results, comparable to those of human annotators and state-of-the-art methods, on a well established benchmark.

Motivation and Objectives

Calculating semantic relatedness between terms is vital in numerous knowledge and information processing tasks of much relevance to the biomedical domain, such as named entity disambiguation (Hoffart *et al.*, 2012), ontology population (Shen *et al.*, 2012), word sense disambiguation (McInnes *et al.*, 2011). Being able to relate entities of interest is crucial in processes of semantic search, information extraction from texts and in building similarity databases. Many successful methods use specialised knowledge bases and lexicon-style resources (Lin, 1998), preparation of which is very tiresome and time consuming. Moreover, in many domains it is not possible to create resources that capture all the possible relationships between the entities of interest. Typically, it is much easier to assemble a fairly large repository of possibly relevant documents that span the domain with their implicit knowledge.

There are also corpus based approaches, whose downsides are often related with computational intensity (Pedersen *et al.*, 2007) and heavy dependence on a specific corpus and its specific features. This paper presents a simple measure designed to provide solutions to those problems, while still being able to produce results comparable with state-of-the art methods.

The design goal was to design a measure that could be used successfully in less-than -ideal availability of knowledge-rich resources. The method takes advantage of a fairly general document corpus of medium size (several orders of magnitude less than Web scale) with very limited use of background knowledge without depending on specific structural features of the knowledge base. Following those design goals should increase robustness and applicability of the new measure, which in terms of quality provides results comparable to those of a human annotator.

Methods

The measure relies on the idea of computing relatedness based on comparing key-phrases related to respective terms. The outline of the method is presented in Figure 1, along with key components used for the similarity computation.

Most relevant documents for the input terms are chosen from the public subset of PubMed articles[1]. Linked Data (Bizer *et al.*, 2009) flavored version of Wikipedia, DBPedia[2], is used as a complementary knowledge base (KB) in the process of query expansion. These are very general resources, selection of which is aimed at obtaining a robust and flexible tool for resolving the relatedness calculation within the whole biomedical domain.

Relatedness of two terms is defined as an overlap measure between key-phrase sets of those terms, as shown in Formula 1. Key-phrases are extracted from K most relevant documents, where document relevance for a term is defined as cosine distance between the document vector and the vector of an expanded query formed around the term. Vectors are defined for a TF-IDF weighted Vector Space Model. For each document N most frequent key-phrases are extracted with a one-pass sliding window T-GSP algorithm

1 http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/
2 http://dbpedia.org/sparql

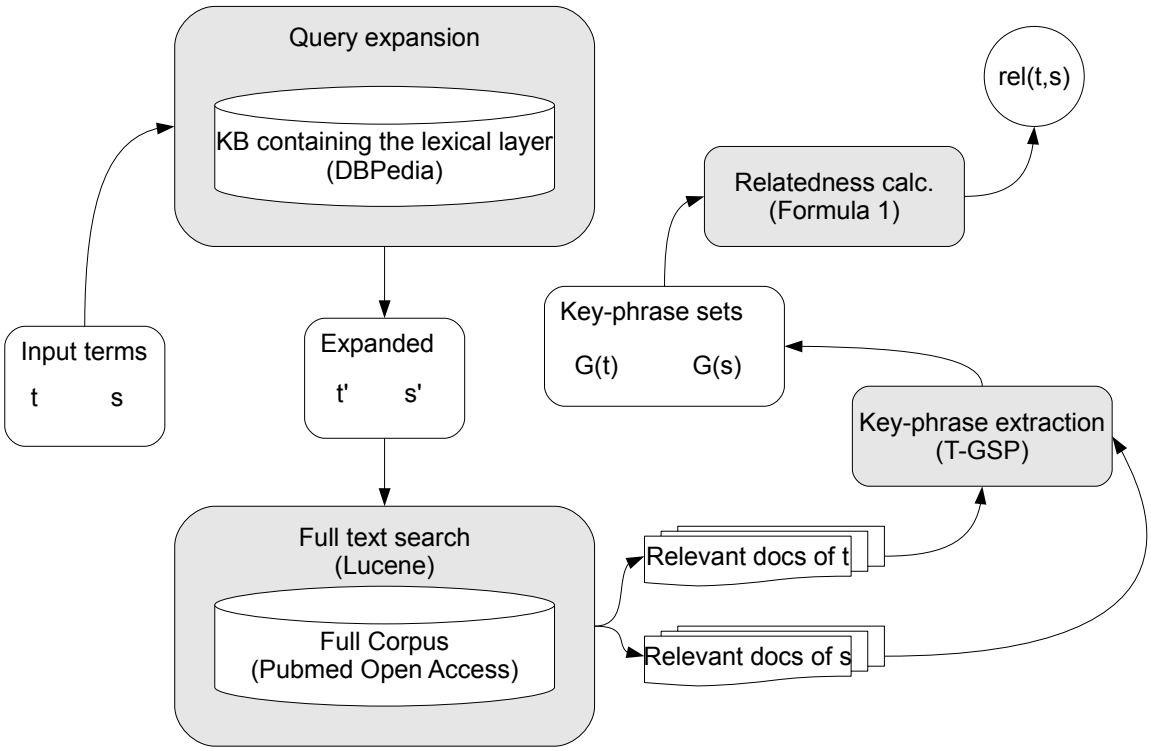


Figure 1. General vision of the system for computing semantic relatedness.

(Protaziuk *et al.*, 2007) that uses several common grammatical patterns to take advantage from shallow parsing information. Query extraction is an optional step that uses the general KB in order to provide wider query context.

Formula 1:

$$\rho(s,t,K,N) = \frac{|G^{sKN} \cap G^{tKN}|}{max(|G^{sKN}|, |G^{tKN}|)}$$

where ρ(s,t,K,N) is semantic relatedness of string s to string t under given parameters K and N and G^{xKN} denotes a set of key-phrases related to string x, with K being the number of documents and N being the max number of most common phrases extracted from a single document.

Query expansion for input terms is executed on the fly and aggregates the results of a DBPedia query that retrieves disambiguation/redirect labels and other synonyms.

For better clarity, the flow of the method is presented in its 'on-demand' implementation, where all computations are executed on-demand for a pair of arguments and additional parameters (*K* and *N*). Nonetheless, it is worth taking notice that the steps of identifying relevant documents for a term and T-GSP analysis of article contents are independent. This means that the key-phrase extraction can be done efficiently off-line for the entire corpus (given the *K* and *N* parameters) in order to speed up the on-line portion of the process. In this case the actual relatedness computations are very simple and are limited to a reduced vector space, which makes the method suitable for large-scale processing, very much present in the biomedical domain.

Results and Discussion

The presented method has been tested on a small benchmark presented in (Pedersen *et al.*, 2007). The benchmark consists of 30 term pairs annotated by a group of physicians and the same pairs annotated by a group of medical coders. Our method was tested with *K* and *N* pa-

rameters ranging from 1 to 10 (integers only), and in it achieved best results for the physicians case with K=6 and N=7, giving correlation rp=0.68 with respect to the average answers, and with K=6 and N=10 it achieved its best result for coders set, achieving the correlation rc=0.77 with respect to the average answers. Those scores were achieved for documents matched to terms through expanded queries, a method that provides 95% coverage for the terms included in the benchmark.

The method without query expansion achieves best case scores of rp=0.39 and rc=0.46 respectively, while also providing 95% coverage for single terms. Best scores without expansion are achieved for a yet another parameter pair (K=5, N=2), which shows the need for further testing in order to fine-tune the algorithm. The version with query expansion seemed more robust in terms of parameter dependence, as it would generally score reasonably high for higher values of K and N parameters.

Additionally, shifting the key-phrase extraction to offline processing will allow involving whole-corpus statistics in the relatedness computation process without excessive additional cost. Doing so should also improve the algorithm in terms of its parameter sensitivity, which shows in the preliminary tests, especially for the evaluations without query expansion.

In general, the presented method in its optimal settings (with query expansion) achieves promising results, which are comparable to those of a human annotator (correlation higher than inter-annotator agreement). Additionally, according to a comparative of various measures presented in (Zhang *et al.*, 2011), our method ranks well against other state-of-the-art related measures for biomedicine, while using very general knowledge sources. During the evaluation it has also been established that the presented measure provides much better (correlation-wise) results than a commonly used Wikipedia-based measure, which relies on comparing class labels of objects most related to the input terms. Furthermore, the automatically extracted key-phrases seem to be more meaningful in terms of semantic relatedness than the actual keywords associated with the documents. Using PubMed keywords as input for the relatedness computation would tend to cause a substantial decrease in result quality.

As a part of the future work the measure should be tested against a much larger (by at least two orders of magnitude) benchmark, possibly with various subdomains. Such an evaluation would certainly help in terms of validating the robustness of the approach and showing the real importance of tuning the K and N parameters. It would also certainly be beneficial for further development of the measure formula itself, as a large benchmark will provide more significant answers.

Acknowledgements

Part of this work was financed under project grants TIN2011-25840 (Spanish Ministry of Education and Science) and P11-TIC-7529 (Innovation, Science and Enterprise Ministry of the regional government of the Junta de Andalucía).

References

Bizer C, Heath T, Berners-Lee T (2009) Linked data-the story so far. Int. *J. Semant. Web Inf. Syst.* **5**(3), 1-22. doi:10.4018/jswis.2009081901

Hoffart J, Seufert S, *et al.* (2012) KORE: keyphrase overlap relatedness for entity disambiguation. In: *Proceedings of the 21st ACM international conference on Information and knowledge management (CIKM '12)*. ACM, New York, pp. 545-554. doi:10.1145/2396761.2396832

Lin D (1998) An information-theoretic definition of similarity. In: *ICML '98 Proceedings of the Fifteenth International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, pp. 296-304.

McInnes BT, Pedersen T, *et al.* (2011) Knowledge-based method for determining the meaning of ambiguous biomedical terms using information content measures of similarity. In: *AMIA Annual Symposium Proceedings*. American Medical Informatics Association, Vol. 2011, pp. 895–904.

Pedersen T, Pakhomov SVS, *et al.* (2007) Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics* **40**(3), 288-299. doi:10.1016/j.jbi.2006.06.004

Protaziuk G, Kryszkiewicz M, *et al.* (2007) Discovering compound and proper nouns. In: *Rough Sets and Intelligent Systems Paradigms*. Springer, Berlin, pp. 505-515.

Shen W, Wang J, *et al.* (2012) A graph-based approach for ontology population with named entities. In: *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*. ACM, pp. 345-354. doi:10.1145/2396761.2396807

Zhang Z, Gentile AL, Ciravegna F (2011) Harnessing different knowledge sources to measure semantic relatedness under a uniform model. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 991-1002.

SLIMS: a LIMS for handling next-generation sequencing workflows

Francesco Venco[1], **Arnaud Ceol**[2], **Heiko Muller**[2]✉

[1]Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milan, Italy
[2]Computational Research, Center for Genomic Science of IIT@SEMM, Istituto Italiano di Tecnologia (IIT), Milan, Italy

Received 16 September 2013; **Accepted** 18 September 2013; **Published** 14 October 2013

Competing interests: the authors have declared that no competing interests exist.

Abstract

Next-generation sequencing (NGS) is becoming a standard method in modern life-science laboratories for studying biomacromolecules and their interactions. Methods such as RNA-Seq and DNA resequencing are replacing array-based methods that dominated the last decade. A sequencing facility needs to keep track of requests, requester details, reagent barcodes, sample tracing and monitoring, quality controls, data delivery, creation of workflows for customised data analysis, privileges of access to the data, customised reports etc. An integrated software tool to handle these tasks helps to troubleshoot problems quickly, to maintain a high quality standard, and to reduce time and costs needed for data production. Commercial and non-commercial tools called LIMS (Laboratory Information Management Systems) are available for this purpose. However, they often come at prohibitive cost and/or lack the flexibility and scalability needed to adjust seamlessly to the frequently changing protocols employed. In order to manage the flow of sequencing data produced at the IIT Genomic Unit, we developed SLIMS (Sequencing LIMS).

Motivation and Objectives

Next-generation sequencing is becoming a standard method in modern life science laboratories for studying biomacromolecules and their interactions. Methods such as RNA-Seq and DNA resequencing are replacing array-based methods that dominated the last decade. A sequencing facility needs to keep track of requests, requester details, reagent barcodes, sample tracing and monitoring, quality controls, data delivery, creation of workflows for customised data analysis, privileges of access to the data, customised reports etc. An integrated software tool to handle these tasks helps to troubleshoot problems quickly, to maintain a high quality standard, and to reduce time and costs needed for data production. Commercial and non-commercial tools called LIMS (Laboratory Information Management Systems) are available for this purpose (Melo *et al.*, 2010; Stocker *et al.*, 2009; Triplet and Butler, 2012; Scholtalbers *et al.*, 2013). However, they often come at prohibitive cost and/or lack the flexibility and scalability needed to adjust seamlessly to the frequently changing protocols employed. In order to manage the flow of sequencing data produced at the IIT Genomic Unit, we developed SLIMS (Sequencing LIMS).

Methods

SLIMS is a web application with a MySQL backend that was developed in continuous interaction with the wet-lab scientists running the sequencing facility and with database experts from the University of Milan (Politecnico). SLIMS is written in the Java programming language, runs on a Glassfish web server, uses Hibernate for object-relational mapping, follows the Model-View-Controller model, and employs the Java-Server-Faces and the PrimeFaces framework for the frontend. Maven is employed to manage dependencies and to build the software. Message bundles are used for easy internationalisation. Workflows are provided to the system as XML files that list the protocol steps in detail. A script generator reads the workflow and generates all the commands needed to perform demultiplexing, quality control, alignments, and visualisation steps. SLIMS can run in a completely automated fashion, although human interaction is helpful at times, especially when non-standard tasks are to be performed. Access to the system is provided via an LDAP realm. LDAP authentication permits a single login policy for users of computational resources. Once authenticated, five roles are being assigned: admin, groupleader, user, analyst, and guest. Admins have full access and can submit or delete sequencing reagents, submit or delete sequencing requests, etc. Users can submit, view, and modify their requests. Supervisors can view the requests of their users. Analysts can create, modify, and launch workflows that ultimately lead to tracks viewable by the biologist in a web browser. Guests can view the general guidelines and the statistics regarding requests, waiting lists, and data delivery.

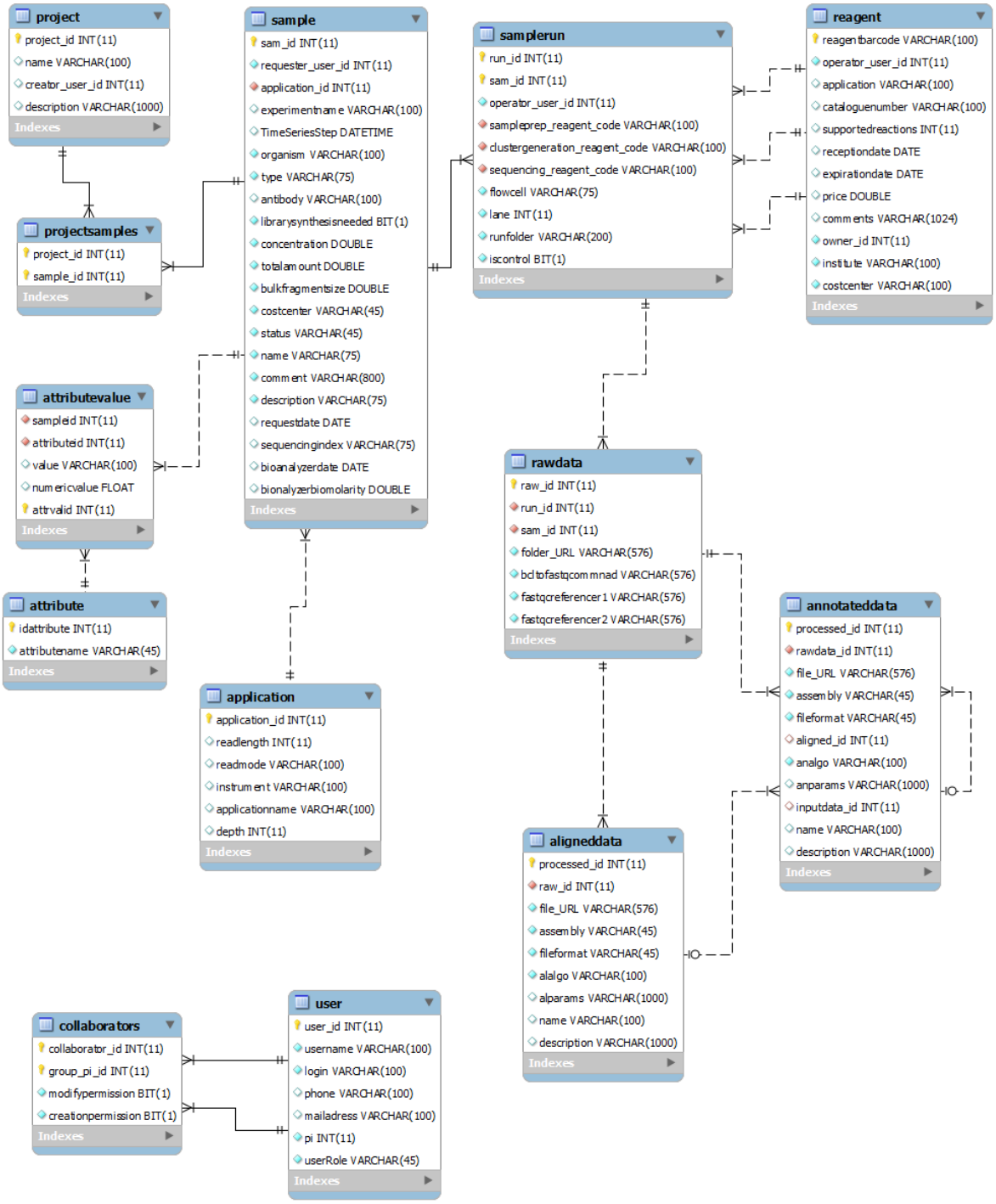


Figure 1. The data model of SLIMS.

Results and Discussion

SLIMS has been developed since August 2011. The first version has been deployed in September 2011. Since then, 2,700 samples have been processed. Samples processed before September 2011 were also inserted into the system. Since the use of SLIMS, data delivery procedures have been standardized making it easier for biologists and analysts to navigate the data. Furthermore, data delivery times have been drastically re-

duced. Currently, when the transfer of data from the sequencing machine to the data archive is complete, demultiplexing starts and provides data in Fastq format within hours, e.g. 4 hours for 50 base pair single read protocols. In summary, SLIMS is handling all the sequencing requests that are processed by the IIT Genomic Unit located at the IFOM-IEO-campus and has been adapted to necessities identified during the processing of numerous sequencing runs. SLIMS is available at http://cru.genomics.iit.it/SLIMS.

Acknowledgements

We are indebted to Prof. Stefano Ceri, Prof. Marco Masseroli, and Dr. Fernando Palluzzi for numerous discussions regarding the data model. This work was supported by a PRIN awarded to Prof. Stefano Ceri.

References

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

Scholtalbers J, Rößler J, *et al.* (2013) Galaxy LIMS for Next Generation Sequencing. *Bioinformatics* **29**(9), 1233-1234. doi:10.1093/bioinformatics/btt115

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

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

National Nodes

Argentina
IBBM, Facultad de Cs. Exactas, Universidad de Buenos Aires, 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

Egypt
Nile University, Giza

Finland
CSC, Espoo

France
ReNaBi, French bioinformatics platforms network, Villeurbanne Cedex

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

The Netherlands
Centre for Molecular and Biomolecular Informatics (CMBI), Nijmegen

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

United Kingdom
The Genome Analysis Centre (TGAC), Norwich

Specialist- and Assoc. Nodes

CASPUR
Rome, Italy

EBI
EBI Embl Outstation, Hinxton, Cambridge, UK

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 1023-4144

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