BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 11 2013, pages 1477–1480
doi:10.1093/bioinformatics/btt162

Databases and ontologies

Advance Access publication May 2, 2013

RegaDB: community-driven data management and analysis for infectious diseases

Pieter Libin[1,2,*], Gertjan Beheydt[1], Koen Deforche[2], Stijn Imbrechts[1], Fossie Ferreira[1,3], Kristel Van Laethem[1,3], Kristof Theys[1], Ana Patricia Carvalho[4], Joana Cavaco-Silva[5], Giuseppe Lapadula[6], Carlo Torti[6], Matthias Assel[7], Stefan Wesner[7], Joke Snoeck[1], Jean Ruelle[8], Annelies De Bel[9], Patrick Lacor[10], Paul De Munter[3], Eric Van Wijngaerden[1,3], Maurizio Zazzi[11], Rolf Kaiser[12], Ahidjo Ayouba[13], Martine Peeters[13], Tulio de Oliveira[14], Luiz C. J. Alcantara[15], Zehava Grossman[16], Peter Sloot[17], Dan Otelea[18], Simona Paraschiv[18], Charles Boucher[19], Ricardo J. Camacho[4,5] and Anne-Mieke Vandamme[1,5,*]

[1]Department of Microbiology and Immunology, Rega Institute for Medical Research, KU Leuven, 3000 Leuven, Belgium, [2]Mybiodata, Biomedical IT solutions, 3110 Rotselaar, Belgium, [3]Internal Medicine, University Hospitals Leuven, 3000 Leuven, Belgium, [4]Centro Hospitalar de Lisboa Ocidental, 1349-019 Lisbon, Portugal, [5]Centro de Malária e Outras Doenças Tropicais, Instituto de Higiene e Medicina Tropical, Universidade Nova de Lisboa, 1349-008 Lisbon, Portugal, [6]Department of Clinical and Experimental Sciences, University of Brescia, 25121 Brescia, Italy, [7]High Performance Computing Center, University Stuttgart, 70049 Stuttgart, Germany, [8]Université Catholique de Louvain, 1348 Louvain-la-Neuve, Belgium, [9]Clinical Biology, Universitair Ziekenhuis Brussel, 1090 Brussels, Belgium, [10]Internal Medicine, Universitair Ziekenhuis Brussel, 1090 Brussels, Belgium, [11]Department of Biotechnology, University of Siena, 53100 Siena, Italy, [12]Institute of Virology, University of Cologne, 50923 Cologne, Germany, [13]Institut de recherche pour le développement, 34394 Montpellier, France, [14]Africa Centre for Health and Population Studies, Nelson R Mandela School of Medicine, University of KwaZulu-Natal, 4041 Durban, South Africa, [15]Gonçalo Moniz Research Center, FIOCRUZ, 40296-710 Salvador, Brazil, [16]Sackler School of Medicine, Tel-Aviv University, 6997801 Tel-Aviv, Israel, [17]Department of Computational Science, University of Amsterdam, 1012 Amsterdam, the Netherlands, [18]Department of Molecular Diagnostics, Institutul National de Boli Infectioase 'Prof. Dr. Matei Bals', Sector 2 Bucharest, Romania and [19]Department of Virology, Erasmus Medical Center, Erasmus University, 3015 Rotterdam, the Netherlands

Associate Editor: Jonathan Wren

ABSTRACT

Summary: RegaDB is a free and open source data management and analysis environment for infectious diseases. RegaDB allows clinicians to store, manage and analyse patient data, including viral genetic sequences. Moreover, RegaDB provides researchers with a mechanism to collect data in a uniform format and offers them a canvas to make newly developed bioinformatics tools available to clinicians and virologists through a user friendly interface.

Availability and implementation: Source code, binaries and documentation are available on http://rega.kuleuven.be/cev/regadb. RegaDB is written in the Java programming language, using a web-service-oriented architecture.

Contact: pieter.libin@rega.kuleuven.be

Received on November 21, 2012; revised on March 13, 2013; accepted on April 1, 2013

1 INTRODUCTION

Advances in infectious diseases research require efficient collaboration and exchange of clinical and virological data. Researchers need access to large amounts of data to test hypotheses or extract valuable information through data mining (Sloot *et al.*, 2008, 2009). For this purpose, RegaDB was developed as a free and open source data management and analysis environment for infectious diseases (Libin *et al.*, 2007).

RegaDB runs on Windows, Linux or Mac OS X. The system can be installed within a hospital or institute so that the data stays within the clinical environment. RegaDB follows the idea of an integrated environment for bioinformatics analysis, such as the Genetic Data Environment (de Oliveira *et al.*, 2003), ViroLab (Assel *et al.*, 2009) and Geneious (Drummond *et al.*, 2011). The difference is that RegaDB uses a relational database, and can be locally or remotely accessed. This allows RegaDB to be used for clinical management and/or research in one locality or for long-term data-sharing collaborations between different institutes.

2 DATABASE STRUCTURE AND TOOLS

RegaDB's database enforces the data abstraction paradigm (Fig. 1). This approach ensures flexibility, as the database can be conveniently extended as needed without upgrading its schema in most of the cases (Imbrechts *et al.*, 2009). All abstract data entities are connected to a central patient entity, including

*To whom correspondence should be addressed.

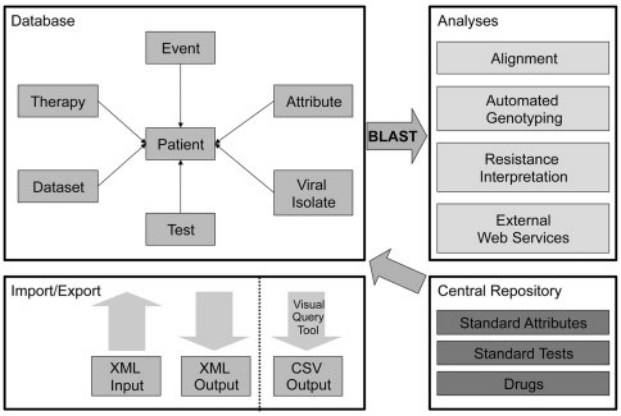

Fig. 1. An overview of RegaDB's database entities and functionalities

attributes, tests, events, therapies and viral isolates. Attributes annotate a patient with information, which is typically of a clinical or epidemiological nature, e.g. the gender or transmission risk group. RegaDB implements tests as values that are obtained at a given moment in time, i.e. there is only one date associated with it. The results can be *in vivo* or *in vitro* measurements, appointments, as well as computational results obtained from a web service. General tests are used to store data extracted from patient samples, e.g. cell counts and viral loads. Tests can also be linked to viral isolates, e.g. typing and subtyping results, to drugs, e.g. therapeutic drug monitoring, or to a combination of an isolate and a drug, e.g. phenotypic and genotypic resistance interpretations. Events cover a specific time interval in the patient's history, i.e. have a start and end date, e.g. AIDS-defining illnesses or pregnancy. The default list of attributes, tests and events available in the system can be extended via the user interface. In this way, RegaDB can be tailored to the user's needs or research interests. Attributes, tests and events are annotated with a data type (numbers, strings, nominal values, etc.), which allows the user interface and data access layer to maintain data integrity. The therapy entity allows users to store the medication history of a patient. A single therapy consists of a start date, a stop date and a combination of drugs, i.e. a regimen, which the users can select from a list of both generic and commercial drug names. When the therapy has a stop date, the clinician can indicate a reason for ending or switching the treatment, e.g. resistance, side effects or adherence issues.

A viral isolate contains one or more nucleotide sequences, allowing multiple sequences extracted from one viral genome to be grouped together. Once an isolate is added to RegaDB, the corresponding pathogen is determined by invoking a web service that implements a BLAST search procedure (Altschul *et al.*, 1990). When RegaDB supports the pathogen, the appropriate reference sequence is loaded and used to perform a codon-correct alignment with frame-shift detection and correction. The alignment procedure finds the protein reading frames encoded by the sequences that make up the isolate. This information, together with all detected point mutations, insertions and deletions, is stored in the database. The alignment web service implements the Needleman–Wunsch algorithm in C++ (Needleman and Wunsch, 1970) to analyse large sequences efficiently. Depending on the pathogen determination returned by the web service, the viral isolate is directed to a typing web service (Alcantara *et al.*, 2009; de Oliveira *et al.*, 2005) and/or resistance interpretation web service (Liu and Shafer, 2006). Table 1 shows detailed information on reference sequences and bioinformatics tools available for the supported pathogens. RegaDB supports the use of bioinformatics tools published on the web as web services.

All data can be viewed and edited through a web-based interface. Key parameters of a patient's clinical history are visualized in a patient chart as a time-line annotated with viral loads, CD4 counts, regimens and viral isolate time points. RegaDB can export patient details into a report document by replacing variables in a user-designed RTF template.

Several tools are already available or are being developed, some of which by the users. Drug resistance interpretation can be performed according to several algorithms. For HIV, various versions of the Stanford algorithms (HIVdb, Liu and Shafer, 2006), the Rega algorithms (Van Laethem *et al.*, 2002) and the ANRS algorithms (Meynard *et al.*, 2002) are implemented. For each algorithm, a cumulative overview is available, whereby resistance detected in a patient is taken forward to the last sample. Evolution of a virus isolate is tabulated as amino acid changes compared with the previous isolate from the same patient. Another tool allows plotting a phylogenetic tree constructed from a set of sequences with a pre-defined similarity to a query sequence. To ensure the quality of the sequence database, a tool was developed that can be used to flag potential contaminations, errors in sampling or data entry, super infection or transmission chains, by detecting unusual intra- or inter-patient evolutionary distances.

Attributes are synchronized with a central repository to ensure compatibility between different RegaDB instances. The central repository contains a collection of standardized data fields and corresponding values such as demographic information (country of origin, transmission risk group, etc.), test results (viral load, cell count, etc.) and drug names (both generic and commercial). In addition, this repository also provides access to the latest versions of drug resistance algorithms. Compatibility functionalities allow the system to be updated, with minimal effort, as new content becomes available.

3 OPPORTUNITIES FOR RESEARCHERS

When the development of RegaDB started, several custom-made databases were available that allowed users to enter ambiguous representations of data, for example, different representations for the same medical compound. However, to facilitate efficient data exchange and to make the execution of aggregate queries possible, it is important that data are available in a structured format. By providing support for explicit data types and enforcing these data types through the user interface, RegaDB circumvents many difficulties that might complicate the exchange of data.

RegaDB allows data to be exported in XML format from local data sources (hospitals, institutes), and these exports can be combined in a research database.

Data from other databases can be imported via a generic import tool. RegaDB also provides a programming interface,

Table 1. Pathogens currently supported by RegaDB, annotated with the reference sequence used for alignments and with the subtyping and resistance interpretation bioinformatics tools applied to new isolates of the respective pathogen

Pathogen	Reference sequence (Genbank accession)	Genotyping	ASI resistance interpretation
HIV-1	HXB2 (K03455)	Rega HIV Subtyping Tool	REGA, HIVDB, ANRS
HIV-2a	ROD (M15390)	Rega HIV Subtyping Tool	REGA, ANRS
HIV-2b	EHO (U27200)	Rega HIV Subtyping Tool	–
HCV	H77 (AF009606)	Oxford HCV Subtyping Tool	–
HTLV	HTLV-1 (J02029)	LASP HTLV-1 Subtyping Tool	–

which can be used to develop custom import programs to support more complicated data sources. A procedure to import data encoded in the HICDEP (hicdep.org) format directly into RegaDB is currently under development.

A research database will generally be accessed via the Internet; therefore, authentication is an important security aspect. RegaDB supports password-based authentication by default. The authentication module abstraction allows for a straightforward implementation of alternative authentication back-ends (OpenId, Kerberos, etc.), which makes it possible for RegaDB to connect to existing user management systems. The application will only allow registered users to access the system. Once granted access to the system, a user is only able to access patient information that belongs to a dataset connected to the user's profile. The owner of the dataset can configure the access of users to this dataset, and revoke the access after a certain analysis or assignment is finished.

Researchers can query RegaDB using the visual query tool, which allows users to define complex queries guided by a user interface. Query definitions can be saved and re-run every time an update of the data becomes available. Work is in progress to support the use of predefined SQL-based queries via the user interface. Query results can be exported to a CSV and/or FASTA file. It is possible to set-up an analysis workflow by configuring a query to execute a python post processing script. If the script generates statistical data in a graphical format, this is visualized in the query user interface after the query has been executed.

When researchers make their tools available as web services, they can be easily integrated in RegaDB, lowering the threshold for clinicians and virologists to use such tools.

RegaDB has been used in several collaborations including the Virolab EC project (virolab.org). Data from several European hospitals were stored in one RegaDB instance, resulting in a combined dataset of >8000 sequences. During the last phase of the project, we were able to combine our efforts with another EC project, EUResist (euresist.org), resulting in a combined database of >55 000 sequences.

Another example of the utility of RegaDB is the collaborative database used within the Southern African Treatment and Resistance Network (SATuRN). This network has 24 member institutions working in Southern Africa, the region at the epicentre of the HIV epidemic. Currently there are >10 institutions using the SATuRN RegaDB for patient data management, data curation and research. Under SATuRN, >7000 genotypes with treatment and monitoring data have been collected. Using the built-in customized report and query functionality, data of specific attributes are selected, analysed and used to answer specific clinical and research questions (de Oliveira *et al*., 2010; Manasa *et al*., 2012). In addition, members of the SATuRN project recently published a book (Rossouw *et al*., 2013) containing a series of case studies used for training. More than 1450 physicians and nurses have been trained through conferences, workshops and online web-tutorials.

4 AVAILABILITY AND USAGE

RegaDB is a software application that can be downloaded from the Internet and installed in a health care or research institute. Documentation, source files and binaries are available on http://rega.kuleuven.be/cev/regadb. Because of its modular and flexible design, RegaDB can be used in many different contexts and settings, from managing patient data in a clinical environment to setting up large-scale research collaborations. Currently, all RegaDB instances are private instances that can only be accessed by a restricted user base. Some of these instances are accessible on the Internet; others are only accessible from within the institute's intranet.

The current version of the software is already used for storing genetic data of HIV-1, HIV-2, HTLV (Araujo *et al*., 2012) and HCV isolates and related patient and clinical information.

ACKNOWLEDGEMENTS

The authors would like to thank the AIDS Reference Laboratory of Leuven that receives support from the Belgian Ministry of Social Affairs through a fund within the Health Insurance System, the 'SPIRALES' program from IRD and the mybiodata company.

Funding: This work was supported by the Fonds voor Wetenschappelijk Onderzoek (FWO) Flanders [grants to K.T. and J.S., G.0611.09, 1.5.236.11N, G.A029.11]; the Research Fund of the KU Leuven [OT/08/047, PDMK/10/204 to K.T.]; the Institute for the Promotion of Innovation through Sciences and Technology in Flanders (IWT) [PhD grant to G.B.]; the Interuniversity Attraction Poles Programme, Belgian State, Belgian Science Policy [IAP-VI P6/41]; the Virolab project [EU IST STREP Project 027446]. The research leading to these results

has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under the project 'Collaborative HIV and Anti-HIV Drug Resistance Network (CHAIN)'—grant agreement n°223131.

Conflict of interest: none declared.

REFERENCES

Alcantara,L.C.J. *et al.* (2009) A standardized framework for accurate, high-throughput genotyping of recombinant and non-recombinant viral sequences. *Nucleic Acids Res.*, **37**, W634–W642.

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Araujo,T.H. *et al.* (2012) A public HTLV-1 molecular epidemiology database for sequence management and data mining. *PLoS One*, **7**, e42123.

Assel,M. *et al.* (2009) A collaborative environment allowing clinical investigations on integrated biomedical databases. *Stud. Health Technol. Inform.*, **147**, 51–61.

de Oliveira,T. *et al.* (2003) An integrated genetic data environment (GDE)-based Linux interface for analysis of HIV-1 and other microbial sequences. *Bioinformatics*, **19**, 153–154.

de Oliveira,T. *et al.* (2005) An automated genotyping system for analysis of HIV-1 and other microbial sequences. *Bioinformatics*, **21**, 3797–3800.

de Oliveira,T. *et al.* (2010) Public database for HIV drug resistance in southern Africa. *Nature*, **464**, 673.

Drummond,A.J. *et al.* (2011) Geneious v5.4. http://www.geneious.com (27 April 2013, date last accessed).

Imbrechts,S. *et al.* (2009) Extending the RegaDB data and analysis management software environment towards HIV-1, HIV-2 and HCV. *Rev. Antiviral Ther.*, **1**, 104–105.

Libin,P. *et al.* (2007) RegaDB: an open source, community-driven HIV data and analysis management environment. *Rev. Antiviral Ther.*, **2**, 82–83.

Liu,T.F. and Shafer,R.W. (2006) Web resources for HIV type 1 genotypic-resistance test interpretation. *Clin. Infect Dis.*, **42**, 1608–1618.

Manasa,J. *et al.* (2012) Primary drug resistance in South Africa—data from 10 years of surveys. *AIDS Res. Hum. Retroviruses*, **28**, 558–565.

Meynard,J.L. *et al.* (2002) Phenotypic or genotypic resistance testing for choosing antiretroviral therapy after treatment failure: a randomized trial. *AIDS*, **16**, 727–736.

Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.

Rossouw,T. *et al.* (2013) *HIV & TB: Drug Resistance & Clinical Management Case Book*. MRC Press, Capetown, South Africa.

Sloot,P.M.A. *et al.* (2008) Virolab: a collaborative decision support system in viral disease treatment. *Rev. Antiviral Ther.*, **3**, 4–7.

Sloot,P.M.A. *et al.* (2009) HIV decision support: from molecule to man. *Phil. Trans. R. Soc. A*, **367**, 2691–2703.

Van Laethem,K. *et al.* (2002) A genotypic drug resistance interpretation algorithm that significantly predicts therapy response in HIV-1-infected patients. *Antiviral Ther.*, **7**, 123–129.