Omic Data Modelling for Information Retrieval

Chloé Cabot[1], Julien Grosjean[1], Romain Lelong[1], Arnaud Lefebvre[1], Thierry Lecroq[1], Lina F. Soualmia[1,2], and Stéfan J. Darmoni[1,2]

[1] CISMeF & TIBS, LITIS EA4108, Rouen University Hospital, France
[2] INSERM, LIMICS UMR 1142, Paris, France

Abstract. This study aims at designing a data model managing the vast majority of omic data types, dedicated to information retrieval on two dimensional levels, single patient and multi-patients.
The exploitation of scientific health research data, in the search for new biomedical applications, is a promising challenge. Data integration generated by scientific research, particularly *omics*, with clinical data stored in the Electronic Health Record (EHR) can lead to significant progress in the development of new diagnostic tests and therapies as well as improve our understanding of complex genetic diseases and cancers.
Currently, a few solutions already exist such as i2b2 or Transmart. However, they do not handle all main omic data types. Moreover, integrating omic analysis results in EHRs is today mandatory to help clinicians in decision making.
This study proposes a generic omic data model dedicated to managing main omic data types (expression data, DNA-methylation, variants, etc.), omic data representation, and information retrieval.
Integrating omics data within clinical data involves: (i) identify the different types of omic data, (ii) select relevant information in the context of integration with clinical data then (iii) design an effective omic data model.
Four levels of data have been defined according to their level of interpretation. The second level representing interpreted data has been selected to build the data model. Various omic data types have been integrated into a database coupled to clinical data.

1 Introduction

During the past decade next generation sequencing (NGS) has become more accessible and gradually replaced microarray techniques in laboratories. While human genome sequencing had taken more than ten years to be completed and cost billions dollars, today, scientists can perform genome or exome sequencing in about a week and for less than a thousand dollars [1]. NGS techniques are currently used to answer many biological issues at the genome scale: variations identification, expression analysis, or even chromatin modifications. Omic data generated by the increasing use of such techniques open up new perspectives in the research of biomedical applications.

Currently, the medical community is facing a new paradigm in the way they have to interact with clinical data. In fact, medical data have become more

and more dependent on decision support tools [2]. Electronic Health Records (EHRs) allow to manage and share all different clinical data types (e.g. dates, numerical, boolean, chronological, etc.). A few clinical data warehouse projects offer architecture, tools and services which permit the use of clinical data within EHRs, especially for biomedical investigation.

To date, i2b2 (Informatics for Integrating Biology and the Bedside), a National Institutes of Health (NIH) funded National Center for Biomedical Computing based at Partners HealthCare System, is considered the most important project. The i2b2 Center is developing a scalable informatics framework that will enable clinical researchers to use existing clinical data for discovery research and, when combined with Institutional Review Board (IRB)-approved genomic data, facilitate the design of targeted therapies for individual patients with diseases of a genetic origin. This platform currently enjoys wide international adoption by the Clinical and Translational Science Awards (CTSA) network, academic health centers (about 70 around the world), and industry [3]. i2b2 is a commonly adopted software in the scientific community, whereas i2b2 data model does not include a single patient point of view and does not handle sequence data.

Transmart [4] is a translational research platform based on the i2b2 data model funded by the pharmaceutical company Johnson and Johnson and supported by a growing developer community. It enables to explore phenotypic data, run meta-analysis, test and validate new hypotheses. However, currently, only expression data are covered.

Since 2011, an ongoing project called Retrieval and Visualization in Electronic Health Records (RAVEL) dedicated to the development of effective and efficient tools has permitted users to locate, in real time, relevant elements of the patients EHR and visualize them according to synthetic and intuitive presentation models. Three academic teams (including the CISMeF team) and two private companies are members of the RAVEL consortium.

The aim of this study was to build a specific omic data model, based on the RAVEL clinical data model, in order to complete several tasks: (i) manage and store the vast majority of omic data types, (ii) omic data representation allowing specific human interface, (iii) information retrieval on two dimensional levels: one dedicated to patient care and for several patients dedicated to epidemiology or quality indicators. The two last tasks should improve clinical research.

2 Material and Methods

2.1 Omic data source

Omic data are coupled with reference data concerning genes, proteins and phenotypes. Omic data have been obtained from several sources such as international repositeries (Gene Expression Omnibus (GEO) [5], ArrayExpress [6], The Cancer Genome Atlas (TCGA) [7]) and local collaboration.

2.2 Reference data source

Reference data from National Center for Biotechnology Information (NCBI) Genes for gene information and from Uniprot Swissprot/KB for protein information were used. NCBI Genes and Uniprot SwissprotKB are two well-known curated and comprehensive international databases. These two databases were filtered to retain only human and human-related genes and proteins.

Phenotypes description relies upon Online Mendelian Inheritance in Man (OMIM) compendium and Orphanet database. OMIM provides information about genetic phenotypes and disorders. Orphanet provides information about rare genetic orphan diseases.

Gene Ontology (GO) was also used to complete gene and protein description.

2.3 RAVEL EHR Model

The RAVEL EHR model is based on a generic EHR conceptual data model integrated to a generic physical model, optimized for information retrieval, developed during a PhD thesis [8]. This model is focused on stays. It is composed of eight entities among which: patients, stays, analyses, medical procedures, etc.

2.4 Data Model

The RAVEL EHR data model is based on a generic EAV (Entity-Attribute-Value) [10] model composed of two parts: the model defining a conceptual data model and the model instance storing the data. This model is dedicated to information retrieval. It enables to manage heterogeneous data types. The database management system used is Oracle 11g r2, including the partitioning option.

2.5 Omic Data Types and Levels

Designing a generic data model gathering clinical and omic data needs to establish a comprehensive and consistent review of all omic data types. Then relevant data for integration with clinical data has to be selected. Then, some challenges appear, including data volume and the lack of consensus on relevant information.

Four data levels to describe these data types, according to conventions adopted by international repositeries like ArrayExpress, GEO or TCGA were used (see Table 1). The first level corresponds to raw, not processed, data. It can be for example Affymetrix CEL scan files from microarray experiments, or BAM (Binary Alignment/Map) files containing sequence alignment data. This data can range from a few thousands of megabytes to several gigabytes per sample. The second level indicates processed data by normalizing raw data, for example in microarray experiments, this level represents normalized signals per probe per sample. This data can represent thousands of megabytes for a single sample. The third level fits interpreted data, resulting from the aggregation of processed data, for example validate variants per sample. This data can represent a few megabytes per sample. Finally level four data represents quantified associations accross several classes of samples and subsequently various regions of interest.

Table 1. Omic data levels

Level	Type	Description	Example
1	Raw data	Low-level data for single sample, not normalized	BAM or CEL files raw signals per probe
2	Processed data	Normalized single sample data	Normalized signals per probe or probe set
3	Interpreted data	Aggregate of processed data from single sample	Expression calls for genes, per sample
4	Region of interest	Quantified associations across classes of samples	A gene X is involved in 10% of lymphomas

3 Results

3.1 Reference data source

Approximately (i) 9 GB data from NCBI Gene (ii) 530 MB from Uniprot/KB and (iii) 165MB from OMIM were initially integrated. A batch program updates this data on a daily basis.

For this study, 80% of OMIM diseases have been translated in French and included in a health cross-lingual terminology portal [9][3].

3.2 Omic data model

The four different levels of data have been assessed to determine which data are relevant. The first two levels of data bring together raw and processed data which are too low-level or voluminous to be considered. This information does not match patient level as they are neither aggregated nor interpreted. However, the third level gathers aggregated and interpreted data like expression calls per gene per single sample or validated variants calls. This information fits to be integrated with clinical data as they are of a high level and consistent with patient data. Moreover, data volume and therefore storage costs are limited. Finally, the fourth level does not correspond with a patient level but fits research purposes, as it gathers observations across several patients and samples. Although this level is valuable, it is not relevant in this case.

The omic data model (Fig. 1) has been designed according to level 3 data, which is considered suitable for integration with clinical data. This model is composed of three parts, managing (i) laboratory and study data, (ii) variants data and (iii) expression data. Detailed managed data types are shown in Table 3.2.

[3] http://www.hetop.eu

Managing laboratory and study data. The first part of this model aims at managing laboratory, study and submitters data. Information about laboratory are name, code, address, e-mail and phone number. Information stored regarding studies aims at recovering and tracking protocol, equipment, sample and genome build used in the experiment as well as data source. This part also manages submitters related data (names, coordonates).

Managing variants data. This part handles variant information like Single Nucleotide variants (SNVs) and insertion/deletions (indels). For each variant, systematic names (nucleic and proteic), reference and mutated base and codon, variation category, localisation and involved region are stored. For each individual, detected variations and genotype status for the corresponding variation are stored.

Managing expression, copy number variations (CNV), DNA methylation data. The database contains data about genes and proteins from respectively NCBI Gene and Uniprot/KB [11]. This data are used to reference gene or protein expression data. Other analysis like DNA methylation, loss of heterozygoty (LOH) or CNV are managed within a single entity `OMI_SEGMENT`. This entity has several attributes such as the type of the genomic segment analyzed and genomic positions. Each gene, protein or generic segment is related to a result for each patient and a reading (if applicable). Each part of this model is linked to the clinical data model through patient entity.

Table 2. Managed omic data types

Data type	**Level 3 description**
Structural variants	Modifications by segmented region by sample
Copy number variations	Copy number variations by segmented region by sample
DNA methylation	Beta values by genomic region by sample
Expression: exon	Normalized expression call by exon by sample
Expression: gene	Normalized expression call by gene by sample
Expression: miRNA	Normalized expression call by miRNA by sample
Expression: junction	Normalized expression call by junction by sample
Expression: transcript	Normalized expression call by transcript by sample
Expression: protein	Normalized expression call by protein by sample
Variants (SNV, indels)	Confirmed variants by sample

3.3 Data integration

The TGCA portal is the only international repositery to offer wide studies with publicly available level 3 data. Therefore, omic data (expression analyses, CGH-array and DNA methylation data) from the TGCA portal and variants from University of Rouen, INSERM U1079 [12] have been integrated. This data are related to clinical data already integrated within the RAVEL clinical data model. Approximately 1GB of data have been integrated in total, using a parser developped to integrate relevant data to corresponding patients in the database (see Fig. 2).

3.4 Omic data visualization

A specific graphical user interface dedicated to omic data has been developped and integrated into the RAVEL prototype, allowing to visualize and retrieve both clinical and omic data (Fig. 3). A specific tab is dedicated to omic analysis results, among clinical analysis results or stays dedicated tabs. All omic data displayed are normalized, interpreted and curated data and each data table can be filtered and sorted.

Variant data. This interface allows to visualize variants data for one patient (Fig. 3) or several patients. Information about systematic names (nucleic and proteic), reference and mutated base and codon, variation category, related gene, localisation or involved region can be displayed and retrieved.

Expression, CNV, DNA methylation data. This interface allows to visualize expression data (gene, protein, miRNA...), CNV, CGH and DNA methylation data. For each omic analysis type all the measures performed for a patient are displayed with an interpretation and a comment.

Cross references Each analysis is linked to the related study, which information can be consulted (title, protocol, material, submitters...). For each gene, NCBI Gene database information are available as well as Uniprot/Swissprot information are available for each protein. Information about gene and protein are completed with the related GO annotations. Cross references with OMIM phenotypes, Orphanet diseases or HPO phenotypes are also available within the interface.

4 Discussion and conclusion

Although a few software solutions already exist in translational sciences to integrate clinical and biological data (such as Transmart), none of them handle all different omic data types, such as sequence data, expression data or variants.

While there is a certain challenge to integrate different omic data types from different kind of studies into a same data model, this type of model would interest both clinical research and care. Indeed, gathering clinical and omic data could lead to innovative applications i.e. new diagnoses tests or targeted therapies. Moreover, this could bring a decisive progress about our understanding of complex genetic diseases and mecanisms involved in cancers.

The omic data model proposed in this study handles the most common omic data types. It has been tested with several omic data types from different omic studies. Data from expression analysis (gene, proteins and miRNAs), cgh-array and DNA methylation analysis have been successfully integrated. Moreover, about 25,000 variants, including SNV and indels have been also successfully integrated in an Oracle database implementing the described data model. However, variant data has been extracted from only one study, due to the lack of available data.

While the reference solution i2b2 is widely adopted in both academic community and industries, this model brings some decisive advantages. In fact, this omic data model within RAVEL clinical data model is able to manage many data types (numeric, dates...) and is highly extensible and adaptable to future new omic data. Based on this model, the graphical user interface is dedicated to data visualization and retrieving. This interface allows to retrieve both clinical and related omic data. Moreover, the combined search engine currently developped in RAVEL project handles logic operators able to manage numeric data ($<$, $>$, $=$) and keywords handling chronological queries. This search engine can process both multi-patients queries and one-patient queries including querying at patient and stay levels, whereas i2b2 handles only multi-patients queries.

It could be also interesting to determine a standard for level 3 data, based on HL7 RIM V3. Such a standard will be essential to industrialize this solution. Finally, this solution warrants further assessment and confirmation with a dataset containing both clinical and omic data.

References

1. Fernald, G.H., Capriotti, E., Daneshjou, R., Karczewski, K.J., Altman, R.B.: Bioinformatics challenges for personalized medicine. Bioinformatics 27(13) (Jul 2011) 1741-8
2. Wyatt, J.: Medical informatics, artefacts or science? Methods Inf Med 35(3) (Sep 1996) 197-200
3. Murphy, S. N., Weber, G., Mendis, M. and Gainer, V. and Chueh, H. C. and Churchill, S. and Kohane, I.: Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc 17(2) (2010) 124-30
4. Sarkar, I.N., Butte, A.J., Lussier, Y.A., Tarczy-Hornoch, P., Ohno-Machado, L.: Translational bioinformatics: linking knowledge across biological and clinical realms. J Am Med Inform Assoc 18(4) (2011) 354-7
5. Edgar, R., Domrachev, M., Lash, A. E: Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30(1) (Jan 2002) 207-10

6. Rustici, G., Kolesnikov, N., Brandizi, M., Burdett, T., Dylag, M., Emam, I., Farne, A., Hastings, E., Ison, J., Keays, M., Kurbatova, N., Malone, J., Mani, R., Mupo, A., Pedro Pereira, R., Pilicheva, E., Rung, J., Sharma, A., Tang, Y.A., Ternent, T., Tikhonov, A., Welter, D., Williams, E., Brazma, A., Parkinson, H., Sarkans, U.: Arrayexpress update-trends in database growth and links to data analysis tools. Nucleic Acids Res 41(Database issue) (Jan 2013) D987-90
7. NIH: The genome cancer atlas. `cancergenome.nih.gov` (Jul 2013)
8. Dirieh Dibad, A.-D.: Recherche d'Information Multi Terminologique au sein d'un Dossier Patient Informatisé. PhD thesis University of Rouen, 2012.
9. Grosjean, J., Merabti, T., Soualmia, L. F., Letord, C., Charlet, J., Robinson, P. N., Darmoni, S. J.: Integrating the human phenotype ontology into HeTOP terminology-ontology server. Stud Health Technol Inform 2013, 192:961
10. Stead, W.W., Hammond, W.E., Straube, M.J.: A Chartless Record-Is It Adequate? Proceedings of the Annual Symposium on Computer Application in Medical Care 7 (2 November 1982) 89-94
11. The UniProt Consortium: Update on activities at the Universal Protein Resource (UniProt). Nucleic Acids Res. 41: D43-D47 (2013).
12. Coutant, S., Cabot, C., Lefebvre, A., Léonard, M., Prieur-Gaston, E., Campion, D., Lecroq, T., Dauchel, H.: EVA: Exome Variation Analyzer, an efficient and versatile tool for filtering strategies in medical genomics. BMC Bioinformatics 2012;13 Suppl 14:S9

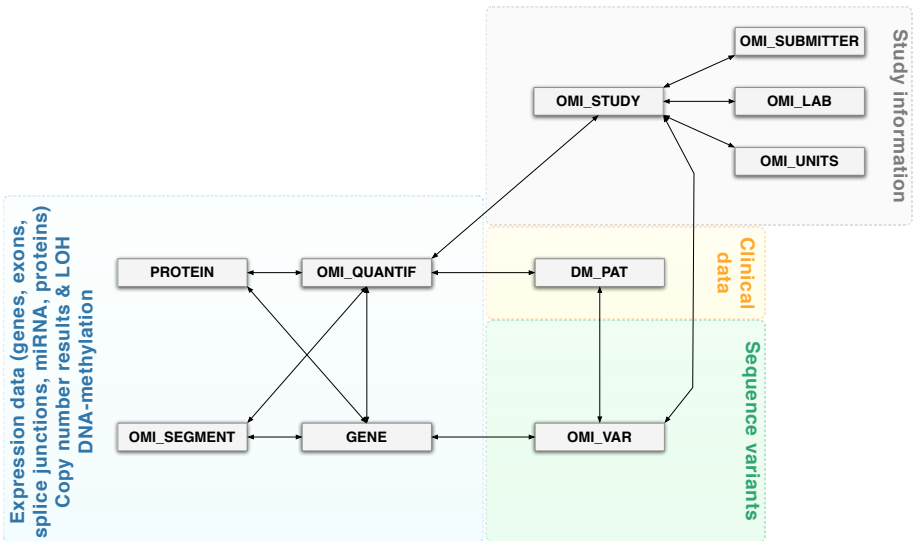

Fig. 1. Simplified omic data model

This omic data model is composed of four parts: (i) a first part, in grey, is dedicated to information about omic studies, laboratories and study submitters, (ii) the green part handles information about variants, SNVs and indels, (iii) the blue part handles information about expression analysis (arrays or RNAseq), copy number variations and DNA methylation studies and finally (iv) the yellow part represents patients data within the RAVEL clinical data model (not shown here). The detailed model can be consulted at http://www.chu-rouen.fr/cismef/papers/omic_mld.pdf.

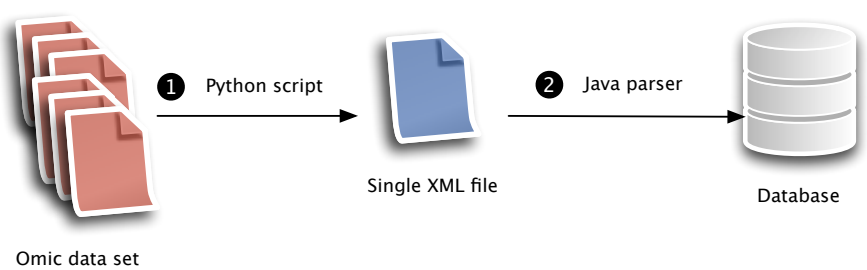

Fig. 2. Omic data integration workflow

(1) Omic data sets are usually composed of tabulated text files, one file corresponding to a patient's analysis results. A set is converted into a single XML file, following a XML schema (XSD) which matches the omic data model. (2) This single XML file is then parsed to integrate data in the database using a specific parser developed to process XML files conforming to the given XSD schema.

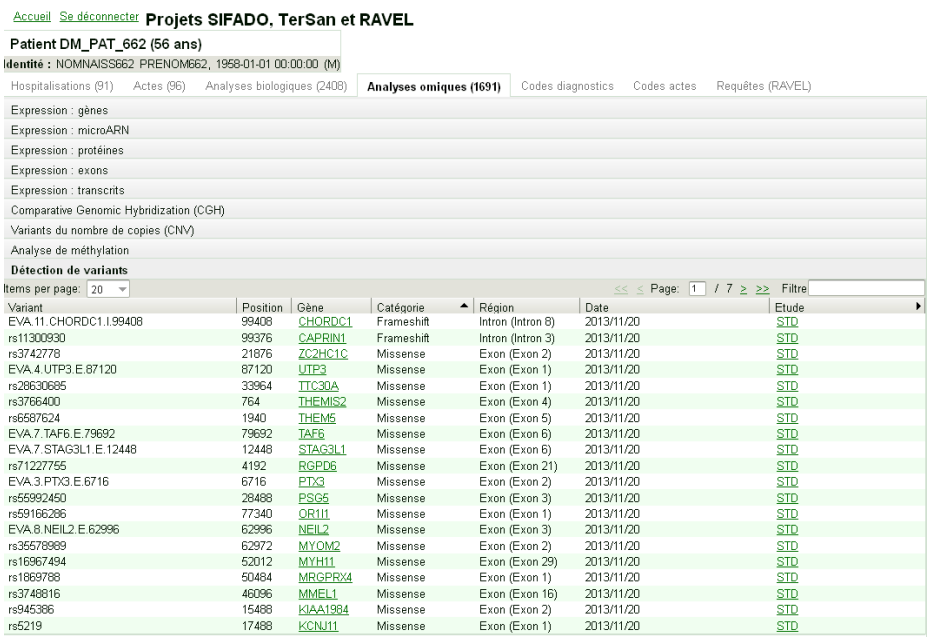

Fig. 3. Screenshot of the RAVEL prototype showing the omic data dedicated tab, including all omic results for a patient