H-0297 (H1101-005) January 23, 2011
Computer Science

IBM Research Report

A Standard Based Approach for Biomedical Knowledge Representation

Ariel Farkash[1], Hani Neuvirth-Telem[1], Yaara Goldschmidt[1], Costanza Conti[2], Federica Rizzi[3], Stefano Bianchi[4], Erika Salvi[3,5], Daniele Cusi[5], Amnon Shabo[1]

[1]IBM Research Division
Haifa Research Laboratory
Mt. Carmel 31905
Haifa, Israel

[2]IMS-Instituto di Management Sanitario SRL
Via Podgora
7-20122 Milano, Italy

[3]KOS Genetic SRL
Viale Ortles
22/A - 20139 Milano, Italy

[4]Softeco Sismat SRL
Via De Marini 1
WTC Tower
16149 Genoa, Italy

[5]Department of Medicine, Surgery and Dentistry
University of Milan
Milan, Italy

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

A Standard Based Approach for Biomedical Knowledge Representation

Ariel Farkash[a], Hani Neuvirth-Telem[a], Yaara Goldschmidt[a], Costanza Conti[b], Federica Rizzi[c], Stefano Bianchi[d], Erika Salvi[c,e], Daniele Cusi[e] and Amnon Shabo[a]
[a] *IBM Haifa Research Lab, Haifa Univ. Mount Carmel Haifa, 31905, Israel*
[b] *IMS-Istituto di Management Sanitario SRL - via Podgora, 7-20122 Milano, Italy*
[c] *KOS Genetic SRL- Viale Ortles, 22/A -20139 Milano, Italy*
[d] *Softeco Sismat SRL, Via De Marini 1, WTC Tower, 16149, Genoa, Italy*
[e] *Dept of Medicine, Surgery and Dentistry, University of Milan, Milan, Italy*

Abstract. The new generation of health information standards, where the syntax and semantics of the content is explicitly formalized, allows for interoperability in healthcare scenarios and analysis in clinical research settings. Genomic data as well may be saved in formats that enable computational analysis. Still, studies involve many artifacts that describe the accumulating knowledge as relationships between varied information items. These include genotype-phenotype associations and associations within the genomic and clinical worlds. Some involve analysis results targeted at a specific disease; others are of predictive nature specific to a patient and may be used by decision support applications. Representing knowledge is as important as representing data since data is more useful when coupled with relevant knowledge. Any further analysis and cross-research collaboration would benefit from persisting knowledge. This paper describes a methodology, used in Hypergenes (an EC FP7 project targeting Essential Hypertension), which captures data and knowledge using standards such as HL7 CDA and Clinical Genomics, aligned with the CEN EHR 13606 specification. We demonstrate the benefits of such an approach for clinical research as well as in healthcare oriented scenarios.

Keywords. HL7/ISO RIM, HL7/ISO CDA R2, HL7/ISO Clinical Genomics, CEN EHR 13606, Knowledge Representation.

Introduction

Standards are used for information exchange between disparate applications serving a variety of healthcare processes, often within the same enterprise. The goal is to have 'semantic interoperability' between the trading applications, so that an entity may deal with received data as its own. The new generation of Object Oriented standards facilitates this approach by serving also as the conceptual data models of persistency layers accessed by the applications. The HL7 v3 Reference Information Model[1] (RIM) is an ANSI and ISO-approved standard information model for healthcare data used to derive consistent health information standards such as laboratory, public health, clinical trials and clinical genomics. The Clinical Document Architecture[2] (CDA) model is derived from the RIM and specifies clinical documents structure. Similarly, the Clinical Genomics Genetic Variation[3] (GV) model captures genotype-phenotype relationships. These models are serialized to W3C XML Schemas. In order to allow for a wide variety of use cases, these models have a generic nature. Thus, in most applications, the

models are further constrained in implementation guides (also called 'templates') targeted at specific use cases (e.g., CDA Operative Note Implementation Guide). Traditionally, health standards were used for patient specific data exchange. With the emergence of decision support applications, it is evident that knowledge representation becomes crucial to enable evaluation of relevant knowledge and generate expected advice. Thus it would be more efficient to have data and knowledge represented over a common language. Indeed, several present efforts use the RIM to represent knowledge models, e.g. HL7 and CDISC clinical trials work in the Study Design model[4]. The HL7 Clinical Genomics workgroup develops a new Domain Information Model[5] with the Genome as the highest organizational entry point. Based on this effort, we have developed derivations of GV models to hold knowledge generated in the course of analyzing SNP variations of hyper- and normotensive subjects in a Genome-Wide Association Study (GWAS). Various approaches and analyses were applied yielding different results. The GV model can be instantiated so that each 'knowledge instance' holds results of a certain analysis. Thus, researchers can exchange and compare results. More importantly, decision support applications can use the results to combine patient data and disease knowledge for generating advice to the clinician.

In Hypergenes[6] (a European Commission FP7 project exploring the Essential Hypertension disease model) we built a set of templates for capturing the different artifacts created along the project. For clinical and environmental data we created an Essential Hypertension CDA-based template as a comprehensive data representation of data collected on Hypergenes subjects. For other artifacts such as genomic analysis results, subject genotyping and decision support information, we created templates based on the GV standard. These templates extend information sharing with the goal of being the underlying data model for representing interactions between environmental, clinical, and genomic factors relevant in studying the complex disease of essential hypertension. Moreover, these templates can be incorporated into the CEN EHR 13606 standard[7], when it is implemented over the RIM, where the CDA is a composition in the subject's EHR and GV instances are linked compositions in the same EHR folder where the CDA is placed. In this paper we depict the methodology used to capture data and knowledge artifacts in the Hypergenes project along with concrete examples.

1. Methods

We classify artifacts of our research to 3 categories: Data, Knowledge and Information. Data is raw clinical or genomic patient data. Knowledge is an understanding of the studied disease that is not specific to any patient. Knowledge may publically available or generated within the project scope, e.g. analysis on data. Information is a subject-specific analysis result that can be used as a prediction or for decision support purposes. The first step in any data driven research endeavor is data collection (top of Figure 1). Data integration is a multi-step process that involves harmonization, validation, normalization, and transformation into standard structures that can be accepted by the healthcare and medical research communities. Furthermore, the relationships among data items are often described implicitly, e.g., in some supplementary documentation or as tacit knowledge of human experts. These relationships must be expressed explicitly as to allow analysis algorithms, oblivious to implicit semantics, to use them effectively and avoid wrong conclusions based on missing implicit data. To that end, the HL7/ISO

RIM was used as the underlying information model coupled with the Web Ontology Language (OWL) and Resource Description Framework (RDF) as the semantic data integration technology. By constraining the CDA standard, we designed the Essential Hypertension Summary Document template (EH-CDA) tuned to the use case of Essential Hypertension. Details on data integration were previously described[8,9]. Genomic samples were genotyped using Illumina 1M-duo arrays. Raw intensity data was analyzed with Illumina Software Genome Studio for genotype calling. Raw data was converted to PLINK PED and MAP files for statistical analysis by PLINK[10], an open-source whole genome association analysis toolset.

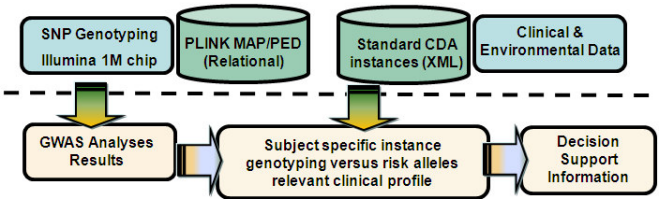

Figure 1. Data & Knowledge Representation Methodology Overview.

Next step involves generating disease-specific knowledge by employing analyses on the collected data. Whole-genome association studies (GWAS) are the state-of-the-art approach in genetic epidemiological studies of complex diseases. These diseases are caused by interaction between genetic, environment and lifestyle factors. GWAS aim to reveal the genetic basis for susceptibility to a disease, with the underlying assumption that diseases prevalent in the population are explained by common variations (Common Disease Common Variants hypothesis). A main source of such genomic variations is Single Nucleotide Polymorphisms (SNP). Current high-throughput technologies allow simultaneous genotyping of a million SNPs in a single chip. Thus, genomic data used in these studies consists of the genotypes for ~1 million SNPs for tens of thousands of individuals. In population-based case control studies, two groups of individuals are collected; a group that is *affected* by the disease and a *control* group. SNPs showing significantly different distribution between the two groups serve to predict a person's susceptibility to the disease and as candidates for further research of the disease mechanism, the final outcome of which is a custom chip for early diagnosis.

Normally, the artifacts of the research thus far, i.e. data and analysis results, would be used in one of two ways: data may be re-analyzed and analysis results may serve to support a clinical decision. While these are important, we want to extend this approach by capturing the amassed knowledge in a formal and standard representation allowing reusability of knowledge as well. As aforementioned, GV models serve as the basis of both data and knowledge in the genetic variation domain. This is made possible thru the Phenotype component used by the GV standard. The Phenotype model design is based on the distinction between observed and interpretive phenotype. The former represents phenotypes observed in the patient e.g., responsiveness to Gefitinib drug due to certain EGFR somatic mutations; the latter represents an interpretation of genomic observations e.g., patient might be resistant to Gefitinib drug due to EGFR somatic mutations. This way, genotype-phenotype associations in GV instances can be incorporated into a patient EHR and a knowledgebase serving clinical decision support applications. Moreover, in capturing analysis-results knowledge we enable further analysis to build on methodology and results, allow for reproducibility, and strengthen cross-research collaboration. Representation of analysis-results entails more than just

capturing the results themselves, i.e. SNP-disease risk assessment. One must capture analysis metadata such as performer, instance details (e.g. date), and methodology. Finally, to allow reproducibility, one must explicitly define the input dataset. We used a GV template to capture analysis results. The GV model is powerful enough to represent all of the above but one: methodology. Methodology is the scientific workflow that documents the analysis and may even allow execution given the appropriate input. We therefore used the encapsulation/reference mechanism of GV to reference a workflow markup that represents the algorithm, similar to the mechanism used for raw sequences. Having patient data and disease knowledge represented with the same HL7 v3 constructs enabled us to create instances that capture subject-specific information (i.e., analysis results specific to a subject). Thus, we designed another GV template to capture subject genotyping, analysis results as applied to subject's SNPs and references to clinical profile. In current efforts, we aim to combine the above data and knowledge to a patient report that may facilitate clinical decision at point of care.

2. Results & Discussion

The Hypergenes project provided us an opportunity to apply our approach to widely varying environmental and clinical datasets as well as genomic data of corresponding subjects. Clinical data included historical data spanning over 15 years and environmental measures based on questionnaires. There were 28 data sources with ~30,000 records for ~12,000 subjects divided into a discovery phase (3,603) and a validation phase (~8,000). Data was stored in a warehouse (using DB2 pureXML) containing CDA-compliant XML instances following the EH-CDA template model. Genomic data comprised of SNP genotyping performed in two centers, Milan and Lausanne, using Illumina 1M-duo arrays in the discovery phase and Infinium iSelectHD beadchip 15K in the validation phase. The raw genotyping data converted to PED and MAP was stored in both file format and relational database for random access.

Classic GWAS analyses test every SNP independently for association. Typically, a chi-square test is performed for every SNP comparing the genotype distribution in the case and the control groups, and a p-value is provided for every SNP. These p-values are used to rank the SNPs. The top scoring SNPs are selected for further research. Since the signals are weak, and many SNPs are being tested, this is a challenging task. In Hypergenes, we enhanced classic analysis by incorporating prior knowledge[11]. We used public SNP annotations and relied on former studies for associating SNP to various diseases. Then, a logistic regression model was trained, learning to utilize SNP annotations to identify a-priori potential of SNPs to be associated with a trait. The algorithm outputs the predicted prior probability of every SNP to be associated with a disease. This prior is used to re-rank the classic analysis results. The analysis can be described as a sequence of steps applied on the SNP data (feature selection, logistic regression, etc). This gave rise to a new tool for carrying out a sequence flow analysis - IBM Bio-clinical Data Mining[12] (BDM) tool. The BDM enables execution of machine learning and data mining algorithms on large datasets. A user may combine various algorithmic building blocks in a workflow to perform a desired task via an XML-based configuration file. Thus, users may utilize the BDM to build the required blocks to execute similar or different flows and analyze their own data.

Following the approach described in the Methods section of this paper, we applied the analysis-results GV template to generate an instance to capture all aspects of the

above analysis. SNP details, risk alleles and p-values were captured in appropriate GV geneticLocus XML constructs. The instance general constructs included metadata on the analysis, e.g. performer, date of execution. For methodology representation, we referenced the BDM workflow instance. Patient information was similarly captured. Subject's genotyping with encapsulated BSML Isoform XML constructs for SNPs and alleles. Top sections captured metadata, e.g. genotyping center, and the phenotype association mechanism of GV was used to reference clinical blood pressure observations from the subject's CDAs (useful mainly in validation phase). Finally, a geneticLoci component was used per analysis results of an individual, encapsulating BSML markup for subject's risk alleles as back-references to the genotyped alleles.

3. Conclusion

In this paper we depict a methodology to capture data, information and knowledge under a standard meta-model. We describe how we implemented this methodology in the scope of Hypergenes. Finally, we demonstrate the benefits of such an approach for clinical research as well as in more healthcare oriented scenarios. In future work we will investigate how to integrate information into an EHR following the CEN EHR 13606 standard.

4. Acknowledgement

The research leading to these results has received funding from European Community's Seventh Framework Program FP7/2007-2013 under grant agreement n° 201550.

References

[1] HL7 Reference Information Model, Health Level Seven, http://www.hl7.org/v3ballot/html/infrastucture/rim/rim.htm

[2] Shabo, A. "Integrating genomics into clinical practice: standards and regulatory challenges. Current Opinion in Molecular Therapeutics", June 2008, 10(3): 267-272.

[3] Dolin R. H. et al, "HL7 Clinical Document Architecture, Release 2", JAMIA 2006;13:pp30-39

[4] HL7 v3 Standards, Universal Realms, Regulated Studies, Study Design Topics. Available online at: http://www.hl7.org/v3ballot/html/welcome/environment/index.html.

[5] HL7 v3 Standards, Universal Realms, Clinical Genomics, Domain Information Model. Available online at: http://www.hl7.org/v3ballot/html/welcome/environment/index.html.

[6] EC FP7 Hypergenes, Available online http://www.hypergenes.eu/

[7] Health Informatics Part 1: Extended architecture. ENV13606-1, Committee European Normalisation, CEN/TC 251 Health Informatics Technical Committee, 2000. Online at: http://www.centc251.org/.

[8] Farkash A. et al. 2006. "Biomedical data integration - capturing similarities while preserving disparities." *Conf Proc IEEE Eng Med Biol. Soc.* 2006 1, 4654-4657.

[9] Carlson, D. et al. "A Model-Driven Approach for Biomedical Data Integration". *MEDINFO 2010*.

[10] Purcell, S., et al., PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet, 2007. 81(3): p. 559-75.

[11] H. Neuvirth-Telem et al. "Inferring Distributions of Trait-Associated SNPs with Application to Genetic Association Studies". ECCB 2010 poster presentation

[12] IBM Bio-clinical Data Mining (BDM) tool on AlphaWorks, http://www.alphaworks.ibm.com/tech/bdm