Towards an Effective Medicine of Precision by using Conceptual Modelling of the Genome*

Short Paper

Ana León Palacio
Óscar Pastor López
aleon@pros.upv.es
opastor@pros.upv.es
Research Center on Software Production Methods (PROS), Universitat Politècnica de València
Valencia, Spain

ABSTRACT

The continuous improvement in our understanding of the human genome is leading to an increasing viable and effective Precision Medicine. Its intention is to provide a personalized solution to any individual health problem. Nevertheless, three main issues must be considered to make Precision Medicine a reality: i) the understanding of the huge amount of genomic data, spread out in hundreds of genome data sources, with different formats and contents, whose semantic interoperability is a must; ii) the development of information systems intended to guide the search of relevant genomic repositories related with a disease, the identification of significant information for its prevention, diagnosis and/or treatment and its management in an efficient software platform; iii) the high variability in the quality of the publicly available information. This paper presents a conceptual framework for solving these problems by i) using a precise conceptual schema of the human genome, and ii) introducing a method to search, identify, load and adequately interpret the required data, assuring its quality during the entire process.

CCS CONCEPTS

• **Information systems → Extraction, transformation and loading**; *Data analytics*;

KEYWORDS

Precision Medicine, Conceptual Modelling, Data Quality

ACM Reference Format:
Ana León Palacio and Óscar Pastor López. 2018. Towards an Effective Medicine of Precision by using Conceptual Modelling of the Genome: Short Paper. In *SEHS'18: SEHS'18:IEEE/ACM International Workshop on Software Engineering in Healthcare Systems , May 27, 2018, Gothenburg, Sweden.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3194696.3194700

*Produces the permission block, and copyright information

SEHS'18, May 27, 2018, Gothenburg, Sweden

ACM ISBN 978-1-4503-5734-0/18/05…$15.00
https://doi.org/10.1145/3194696.3194700

1 INTRODUCTION

Precision Medicine is a new paradigm of prevention, diagnosis and treatment of diseases. The Precision Medicine approach is based on the individuality of each human being. It considers the genetic predisposition, lifestyle and the influence of the environment over the health to take the right decisions for each patient [1]. In order to succeed when applying Precision Medicine in the clinical practice, it is necessary to integrate information coming from diverse areas of knowledge. But they have been traditionally studied independently. They are the so-called "omic sciences", such as Genomics, Proteomics, Epigenomics and Pharmacogenomics. All these sciences have experimented a great progress during the last two decades, especially Genomics. Advances in research technologies such as Next Generation Sequencing have allowed us to read (sequence) DNA in a faster and cheaper way. The sequencing of DNA has become a routine research tool and it has revolutionized our understanding of human biology. We are starting to understand how changes (variations) in the DNA are involved in the risk of suffering a certain disease. Nevertheless, to take advantage of it, we must be able to provide mechanisms to enrich this knowledge with information coming from other research areas. In order to achieve this aim, two main issues must be faced:

- The available information is heterogeneous and dispersed: hundreds of different genomic data sources are publicly available, allowing biologists and clinicians to tackle complex diseases in a multidisciplinary way [10]. However, they have been commonly developed ad-hoc, focused on addressing specific knowledge requirements and not designed to share information among them.
- The complexity of biological processes, the noisy nature of experimental data and the diversity of sequencing technologies results in a great variability in the quality of the available information. That is why a huge amount of information is ready to be used, but only part of it is relevant to be applied with clinical purposes.

In order to face the aforementioned issues, the use of different software engineering practices must be combined to provide a proper technological background to drive the improvement of Precision Medicine. The use of conceptual modelling techniques allows to connect the information coming from heterogeneous data sources under a holistic perspective for a better understanding of the human biology. The use of data quality techniques helps to identify

the information which is really relevant and of enough quality to understand the causes of the disease and make reliable clinical diagnosis. In this paper, we propose a framework based on conceptual models and data quality assessment techniques to guide the identification of reliable information, and provide the required holistic view and specific software platforms for an efficient management in daily work. In section 2, we present the Conceptual Schema of the Human Genome (CSHG), specially developed to comprehend and integrate information related to the human genome. In section 3, we present the SILE method which guides the selection and validation of genomic data, taking into account information quality requirements. Finally, concluding remarks are exposed.

2 USING CONCEPTUAL MODELS TO INTEGRATE GENOMIC INFORMATION: THE CSHG

The integration of information must be based on a well-defined ontological background to avoid mistakes such inconsistencies or redundancy when representing the information. In particular, genetic diagnosis critically depends on accurate and standardized description and sharing of the new findings. One of the most important problems when trying to understand the core concepts of the genomic domain is the lack of an ontological commitment to define them. As a result, situations such as discrepancies in the use of critical concepts arise. In some disciplines the term "mutation" is used to indicate "a change" while in other disciplines it is used to indicate "a disease-causing change". Similarly, the term "polymorphism" is used both to indicate "a non-disease-causing change" or "a change found at a frequency of 1% or higher in the population" [5] [6]. These issues hinder the process of retrieval, annotation and integration of heterogeneous datasets. When the research community realized that this issue was becoming a remarkable problem the first approach to solve it was to construct ontologies, with the aim of unifying knowledge and make it interoperable through consistent vocabularies. But these ontologies became essentially large terminological resources, used as a glossary of genomic terms that are often heterogeneous and even inconsistent when compared among them [4]. Examples of such ontologies are Gene Ontology (GO), Sequence Ontology (SO) and Variation Ontology (VariO). Each of them describes a specific part of the genomic domain, but when we look for a common conceptual schema in order to have a holistic view of all these knowledge, there is not a clear solution.

The use of conceptual modelling techniques applied to the biological domain has been introduced by Paton in 2000 [7]. In his work, a set of models to describe transcriptional and translational processes is proposed. Later on, Ram and Weis modelled the 3D structure of proteins [9] and Bernasconi the meta data related to experimental information [2]. Nevertheless, these proposals focus on specific parts of the domain. To provide a global conceptualization we propose the use of the CSHG designed by the PROS Research Center in UPV. It is currently in its 3rd version and has been developed in collaboration with biologists and experts in the genomic domain. The CSHG is based on the existing biological knowledge and provides a global view of the different parts which compound the human genome through its five main parts or views [8]:

- The *Structural* view describes the genome structure.
- The *Transcription* view represents the mechanisms involved in the synthesis of proteins.
- The *Variation* view represents the changes that may occur in the DNA sequence.
- The *Pathway* view describes the information related with metabolic pathways.
- The *Bibliography* and data bank view describes where the data comes from.

As an example, Figure 1 shows how a variation is represented in the CSHG as part of the Variation view.

According to the schema, a variation is a change which occurs in a certain position in the DNA sequence. Depending on its frequency and description it can be classified into different types. For a deeper explanation about the content of the schema see [8]. The CSHG allows to store each piece of knowledge in the right place, regardless of its origin.

3 METHODOLOGICAL BACKGROUND: THE SILE METHOD

The CSHG is the basis to manage genomic information. The SILE (Search-Identification-Load-Exploitation) method supports the search of relevant repositories as well as the identification, load and exploitation of high quality data through its four levels:

- *Search*: Selection of the most suitable data sources by using the CSHG as a guide to determine the information required.
- *Identification*: Selection of the most relevant data from each data source.
- *Load*: The identified data is loaded into a database, based on the structure provided by the CSHG.
- *Exploitation*: The data stored in the database is efficiently analysed and managed by a set of tools developed to be applied in the clinical practice.

All the process is supported by certain tasks specifically design to assure the quality on each level of the SILE method. The need to assure quality of genomic data is key due to two main aspects: i) to achieve competitive advantages through its analysis by an IS and ii) because decision making based on low genomic data quality may involve serious mistakes with important consequences when applied with clinical purposes. Even when the study of data quality standards began in the 1990s, research about quality in genomics has just started and there are not sound results yet. Data Quality has been defined by Wang and Strong [11] as data that are "fitness for use". Data Quality is a multidimensional concept where a data quality dimension is defined as a set of attributes. These attributes can be assessed using specific metrics in order to get a quantitative measure that represents the quality of the data being managed. There are hundreds of publicly available databases storing genomic information. Thus, to establish the data quality dimensions and metrics required to assure the suitability of the information it is important to identify the problems affecting these repositories. To accomplish this goal, we have performed a study which involves a set of well-known databases [3]. The study has allowed us to classify the most common types of errors into six data quality dimensions: Accuracy, Completeness, Consistency,

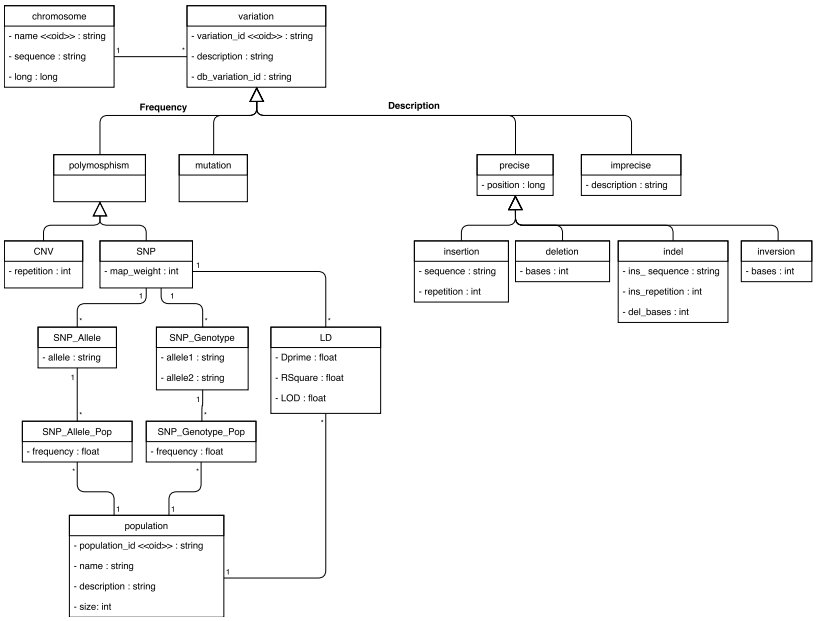

Figure 1: Representation of a variation according to the CSHG.

Uniqueness, Currency and Reliability. Some examples of the errors found related to each dimension are:

- *Accuracy* errors: Conflicts in the DNA sequence, incorrect annotations and syntactic errors.
- *Completeness* errors: Missing values, due to lack of knowledge or automatic curation.
- *Consistency* errors: Different ways of representing the same concept or different types to classify the same property.
- *Uniqueness*: Duplicated information due to consistency errors.
- *Currency*: Inactive databases and information out of date.
- *Reliability* errors: Annotation of DNA sequences using automatic algorithms which are error-prone.

In order to minimize the impact of the detected errors, the measurement of the data quality dimensions identified has been incorporated to SILE. A summary of the different levels of SILE and the quality dimensions checked on each one are shown in Figure 2 Believability, Relevancy and Reputation are dimensions related to Reliability, providing a higher level of granularity which helps to achieve more accurate results.

As mentioned before, in order to measure quality, specific metrics related to the dimensions must be clearly defined. For example, one of the metrics used to measure "Believability" of a database is: "The information stored in the database is examined and curated by a group of experts in the field before it becomes publicly available". Only if the minimum data quality requirements established for the current level are fulfilled, the activities of the next one begin. Following the previous example, once the relevant data sources that satisfy the quality criteria are found the identification of relevant information starts. This process assures the selection of reliable data sources as well as the detection and correction of errors before data are loaded into the database.

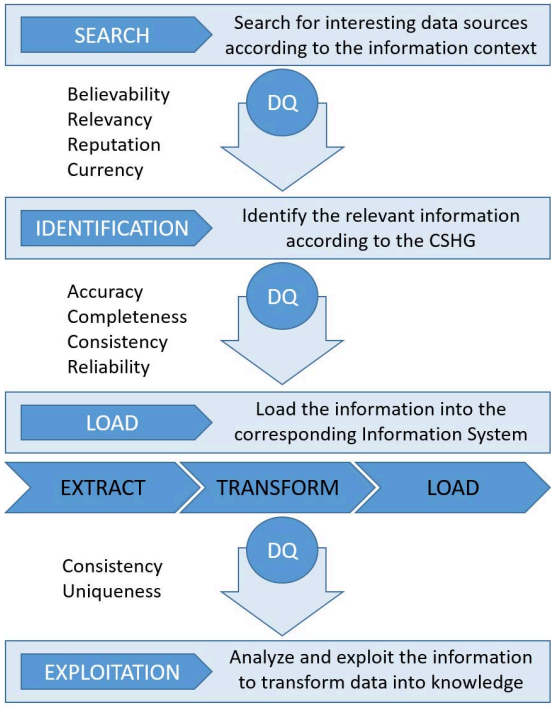

Figure 2: SILE method and Data Quality Dimensions.

4 CONCLUDING REMARKS

Some research has been done on conceptual modelling applied to the biological domain. But this research is not enough to satisfy the needs of Precision Medicine. The lack of consensus on the definition of core biological concepts and the absence of interoperability among the available data sources require the use of new approaches. By using the CSHG, we are able to clearly establish the ontological background needed to understand complex domains such as genomics. Moreover, the structure provided by the CSHG is key to relate the information coming from unconnected areas of knowledge under the global view required by Precision Medicine. Additionally, the use of techniques to assure data quality is crucial to manage relevant data in clinical practice. The presented SILE method serves as a guide to select and manage genomic information, taking into account the variability of the information available in public repositories. Besides, it provides data quality support to identify and prevent the inherent errors in these data sources. As a consequence, it is guaranteed that the results derived from the data analysis are reliable and accurate enough to be applied in clinical practice.

ACKNOWLEDGMENTS

The authors would like to thank the members of the PROS Research Centre Genome Group for the fruitful discussions regarding the application of Conceptual Modelling in the medicine field. This work has been supported by the Spanish Ministry of Science and Innovation through project DataME (ref: TIN2016-80811-P) and the Research and Development Aid Program (PAID-01-16) of the Universitat Politècnica de València under the FPI grant 2137.

REFERENCES

[1] M. Turcotte A. Alyass and D. Meyre. [n. d.]. From big data analysis to personalized medicine for all: challenges and opportunitiesâĂİ, journal =. ([n. d.]).
[2] A. Campi A. Bernasconi, S. Ceri and M. Masseroli. [n. d.]. Conceptual Modeling for Genomics: Building an Integrated Repository of Open DataâĂİ, booktitle =.
[3] V. Burriel A. León, J. Reyes and F. Valverde. [n. d.]. Data Quality Problems When Integrating Genomic InformationâĂİ, booktitle =.
[4] M. Donato A. Splendiani and S. Draghici. [n. d.]. *Ontologies for Bioinformatics.*
[5] I. Lauer C. M. Condit, P. J. Achter and E. Sefcovic. [n. d.]. The changing meanings of àĂIJmutation:âĂİ A contextualized study of public discourseâĂİ, journal =. ([n. d.]).
[6] J. T. den Dunnen et al. [n. d.]. HGVS Recommendations for the Description of Sequence Variants: 2016 updateâĂİ, journal =. ([n. d.]).
[7] N. W. Paton et al. [n. d.]. Conceptual modelling of genomic informationâĂİ, journal =. ([n. d.]).
[8] J. C. Casamayor J. F. Reyes Román, O. Pastor and F. Valverde. [n. d.]. Applying Conceptual Modeling to Better Understand the Human GenomeâĂİ, booktitle =.
[9] S. Ram and W. Wei. [n. d.]. Modeling the Semantics of 3D Protein StructuresâĂİ, journal =. ([n. d.]).
[10] D. J. Ridgen and X. M. Fernández. [n. d.]. The 2018 Nucleic Acids Research database issue and the online molecular biology database collectionâĂİ, journal =. ([n. d.]).
[11] R. Y. Wang and D.M. Strong. 1996. Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems* 12, 4 (1996), 5–33.