Software Engineering and Genomics: The Two Sides of the Same Coin?

José Fabián Reyes Román[1,2], Ana León Palacio[1] and Óscar Pastor López[1]
[1]*Research Center on Software Production Methods (PROS), Universitat Politècnica de València, Valencia, Spain*
[2]*Department of Engineering Sciences, Universidad Central del Este (UCE), San Pedro de Macorís, Dominican Republic*

Keywords: Precision Medicine, Conceptual Modeling, Software Engineering, Genomics, CMHG.

Abstract: Programs are historically the basic notion in *Software Engineering* (SE) that represent the final artefact to be executed in a machine. These programs have been created by humans, using a silicon-based code, whose final components use a binary code represented by 0s and 1s. If we look at life as a program with a DNA-based genetic code and a final representation that uses four essential units (A, C, G and T), one challenging question emerges. Can we establish a correspondence between life *-from a genomic perspective-* and programs *-from a Software Engineering perspective-*? This paper assumes a positive answer to this question and goes further into this mapping by proposing how *conceptual models* (CM) are not only required to understand life but to manage the huge amount of data generated in the genomic domain day after day. The main contributions focus on i) showing how to design such a *Conceptual Model of the Human Genome* (CMHG), analysing how it evolves as knowledge accumulates on the domain, and ii) how these ideas can be applied in an advanced, genome-based, precision medicine, under the assumption that this medicine will only reach our health systems if these sound SE practices are properly applied in the genomic domain.

1 INTRODUCTION

Understanding life as we know it on our planet can probably be considered the biggest challenge of our century. However, *can Software Engineering (SE) help us to achieve this?* Answering this question becomes a relevant issue that affects how modern *Precision Medicine* (PM) can reach our society, changing and improving medicine, as we historically know it. As in our previous work, we try to answer this question in this paper by looking at life from an SE perspective. Our position is easy to explain: humans build programs executed by a silicon-based binary code. These programs are the written representation of conceptual models (CM) (Olivé, 2007) that abstractly represent a relevant part of the real work we are interested in. The upper part of Figure 1 depicts this process, following a pure SE perspective.

It is interesting that a similar metaphor can be applied to achieve the desired clear understanding of life. In this case, programs are living beings whose genetic code includes the instructions that explain life, as we perceive it. Instead of having the SE materialization of a binary executable code, in this case we have what we could call a quaternary executable code, based on four letters (A, C, G, T) that represent the four nucleotides that form the basic components of this *"carbon-based"* executable. (See the lower part of Figure 1).

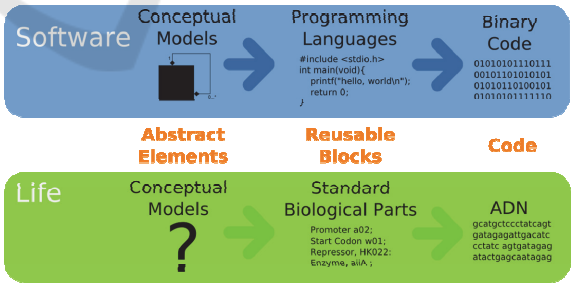

Figure 1: From conceptual modes to code: a SE-perspective and a life understanding perspective.

If we want to develop this idea, one immediate question that arises is: *What is then the language of life that would allow us to understand and manage life as we understand and manage SE-based programs?* We are perfectly aware of the magnitude of the challenge that arises from this question. But at the same time, we are aware that the race to face this

Román, J., Palacio, A. and López, Ó.
Software Engineering and Genomics: The Two Sides of the Same Coin?.
DOI: 10.5220/0006368203010307
In *Proceedings of the 12th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE 2017)*, pages 301-307
ISBN: 978-989-758-250-9

challenge has not only started but is proceeding at full speed.

Although DNA is the basis of all life as we know it on Earth, we focus here on the human genome, where rapid progress is being made especially in the context of PM. It is in this context that we want to focus our work, and where we want to report the experience accumulated in the last years in how essential it is to have a *Conceptual Model of the Human Genome* (CMHG) for structuring the huge amount of data and knowledge that day after day is generated in the genomic domain.

What we want to indicate with the selected title of this paper is that we can draw a parallel between SE and genomics, by considering live beings a particular kind of programs whose (*genomic*) code is started to be known, but whose CM are still to be discovered. In our work, we are not simply applying one SE technique (*conceptual modeling*) to a complex domain (*human genomics*). We go much further: what we want to show is how genomics and SE can share a same picture (as Figure 1 represents), and particularly, how genomics can benefit from SE by applying conceptual modeling to determine those relevant data that life represents in order to manage those data accordingly, with especial emphasis in the health domain.

Following this argument, the general characteristics of PM are introduced in Section 2. Next, how a CMHG is introduced as the basis of any IS intended to manage genome data. The final section contains our conclusions and future work.

2 BACKGROUND: PRECISION MEDICINE (PM)

The social context of our work is PM, an emerging approach for disease treatment and prevention intended to change what we could refer to as *"historical medicine practices"*. It takes into account the variability in *genes*, *environment*, and *lifestyle* of each person to provide individualized treatments and disease prevention (Aguilar, 2015).

PM is a way of treating the patient which allows doctors to identify an illness and select treatments that are more likely to help patients according to a genetic concept of the disease suffered by a given patient (also called "*Personalized Medicine*") (Instituto Nacional del Cáncer, 2015).

As shown in Figure 2, this approach is based on detailed knowledge of the genomic domain, and on the information derived from the large amount of data generated in recent years, which is in constant growth as research is providing more and more information every day.

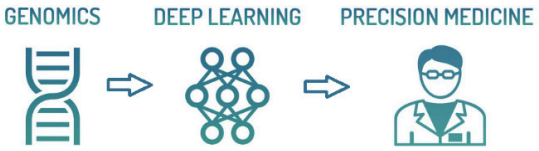

Figure 2: From Genomics to Precision Medicine.

In practical terms, *Fowler et. al.* describes the advances in genomics that will provide information about diseases, explaining which people are most likely to suffer this diseases, and how to apply a more successful treatment for each individual (Jiménez, 2014). For example, although there are many causes of lung cancer, only people who have a mutation in the gene "*EGFR*" respond to treatment with *tyrosine kinase inhibitors* (Paez et al., 2004). Even when the cause of a disease is known, different genetic variants can affect treatment efficacy by altering the way in which drugs are metabolized or by increasing the likelihood of adverse events (Aronson & Rehm, 2015). Another advantage of PM is that it can combine *epidemiology*, *clinical*, *genomics* and *personal preferences* in a bold revolution to prevent and treat disease.

Research on PM can help to improve survival rates in diseases like *cancer*, or could find new treatment options for rare diseases for which a specific treatment is not available. Prevention also has an important impact not only on avoiding the development of disease, but also on reducing medical costs. To achieve these goals (*prevention*, *treatment*, *knowledge*) the different technologies from different areas must be combined. After the introduction of *NGS-technologies*, DNA sequenciation has become cheaper and accessible to many research centers. Consequently, the amount of information has grown considerably and *Big Data technologies, Information Systems (IS)* and *Conceptual Modeling* play a key role in managing it.

Having a well-structured knowledge representation based on high-quality genomic information is crucial to ensuring the success of PM. When we face a problem where a lot of data are accumulated, and the information is to be inferred by discovering those valid patterns that have a precise meaning, conceptual modeling becomes the essential approach to structure and manage fruitfully all this data. Nevertheless, proper inferences cannot be drawn from low-quality data. Errors in the knowledge base will lead to invalid conclusions, so *Conceptual Modeling* and *Data Quality Assessment* (DQA) must evolve together (León et al., 2016). As we said before, if we want to understand the complete and diverse set of information related to

the human genome and interpret it correctly, the first requirement is a precise CM able to identify and represent precisely all the relevant information in this particular domain. It is surprising to realize how far conceptual modeling is in what we could call conventional Bioinformatics practice. When we look at genome data platforms, we usually find low-level, solution space-oriented answers where the conceptual data perspective is mostly ignored. This creates strong interoperability problems that are especially serious in a domain with a huge amount of diverse data sources. This is why our main contribution in this context is to propose a simple, holistic conceptual view of genome data through the CMHG that we introduce next.

3 CONCEPTUAL MODEL OF THE HUMAN GENOME

It is widely accepted that applying CM facilitates the understanding of complex domains (e.g., *genomics*). In our case, we used this approach to define a model representing the characteristics and behavior of the human genome (Ram & Wei, 2004).

One of the essential benefits of CM is that it accurately represents the relevant concepts of the analyzed domain. After performing an initial analysis of the problem domain, the next step is to design a domain representation in the form of a CMHG. Our CMHG evolved with the new discoveries made in the field of genomics (Reyes, 2016) in order to improve the processing of data to ensure effective PM.

We can thus see how CM gives positive support to the knowledge on which PM is based. In a case study, we used this approach to define the structure of the human genome. It is important to highlight that such a model has to evolve over the years due to the changes and developments that continually occur in this domain. Indeed, the advantage of CM for representing this domain is that it eases the integration of new knowledge into the model.

Representing conceptually genome data is a complex task that requires a strong exercise of data abstraction to select the relevant data that must be included in the target CM. A set of meeting with geneticists allowed us to develop a first version of such a CM, that shown soon to be very dynamic in terms of the selected representation for the basic genomic concepts. This is why we find significant to discuss the basics of that CM evolution that was the consequence of acquiring a more in-depth knowledge of the domain.

Below we describe the evolution of our initial version of the CMHG v1 up to the current version (CMHG v2), with the purpose of showing how significant such a CM-based discussion is in order to reach a precise understanding of the domain.

3.1 An Initial Conceptual Model for the Human Genome: CMHG v1

In this first representation, important decisions are taken to adequately represent the concepts that are basic to understanding the domain. The first important decision is how to structure the representation of the analyzed domain. Considering the complexity of the information contained in the human genome, we decided to divide CM representation into three main parts, each one related to a specific domain view:

- The *Gene-Mutation* view: is focused on the *gene* structure, together with its possible, relevant variations and the determination of the data sources.
- The *Genome* view: is focused on how we go from the whole genome to its relevant component (*chromosomes*) and the type of DNA segments they are made of.
- The *Transcription* view: is centered on the actors that participate in the essential processes of transcription and translation, in order to identify the components that guide the process of going from the DNA-based genotype to the protein synthesis, which is related to the phenotype (*external gene manifestation*).

Our first version of the CMHG is a combination of these three views. For more information, *see the full view and description in* (Reyes et al., 2016).

3.2 From v1 to v2: CMHG v2

Once our CMHG v1 was considered finished, we started to think about its ability to deal with the real data that are managed in the domain. While doing this, we identified some questions to address:

- We were not sure about the suitability of mixing a Genome view related to the storage of individual genomes –the so-called Genome view in v1-, with a more theoretical, structural Genomic view related to the Genome configuration and characterization as a whole –the so-called Gene-Mutation and Transcription view-.
- Concerning the core concept of gene, it is not always feasible to describe DNA structure in terms of genes as molecular basic notions. We concluded that the safest structural description

should be based on chromosome elements as the basic building DNA element.
- Incorporating more detailed relevant information in the CMHG is a need, especially when basic concepts are involved in the discussion. For instance, the concept of SNPs.
- We detected the need for extending the v1 with more significant genome-related information. To go from genotype-to-phenotype in a complete -sound way-, we needed the specification of the *pathway* description perspective.

The development of these four ideas led us to the introduction of a new CM that we explain in detail below (called *CMHG v2*) (Pastor et. al., 2016).

3.2.1 Removing Individual Genomes Data Bank

Reviewing the knowledge represented in our v1, our first idea was that the generic genome template –*which is the precise human genome structure and how to characterize it*- and the genome data bank perspective –*how to store individual genomes that are going to be analysed*- was mixed. The gene-mutation and transcription views on the one hand, and the genome view on the other appear together in the CM, and these two concerns should be separated.

We concluded that this is not the best way of representing the domain knowledge, as the generic properties of the genome and the individual samples should be clearly distinguished. In this way, to develop a software platform to generate a genome clinical report would be much easier, by separating the individual sample of a patient from the genome template taken as the reference to find (i.e., significant variations in terms of disease implications). The v2 thus omits the so-called Genome view, focusing on a more precise description of a generic genome template intended to collect all the relevant genome information. We decided to organize it into five main parts (*views*)*:*

- *Structural:* basic elements of the DNA sequence.
- *Transcription*: components involved in going from DNA to the diversity of RNAs.
- *Variation:* describes the changes in the sequence of reference.
- *Pathways:* intended to enrich the CM with information about metabolic pathways to join genome components that participate in pathways with phenotype expressions.
- *Bibliography references:* to assess the source of any information in order to pinpoint the data source.

3.2.1 The Chromosome Elements As Conceptual Modeling Units

The use of chromosome elements as basic DNA building elements had a direct influence on the way in which variations and their DNA origin were represented in the CM. In v1, the notion of allele was represented as an explicit derived notion – through the class *Allelic Variant*-. Additionally, all the variations were related to genic segments, as it was not possible to register variations whose source were in other –*non genic*- genome parts. To overcome this problem, our conceptual proposal for the next version (v2) was to directly relate a variation with a specific DNA chromosome position, as this solution better represents the real genome structure. The benefit of not having the variation directly related to an *Allelic Variant* is twofold:

- Firstly, it allows the variation to be defined with more precision and less dependence, as it is associated with a unique genome sequence just where the variation occurs. The variation is not dependent on the Allelic Variant or on the corresponding many-to-many association, as occurred in v1. This relationship was indeed something of a problem. For instance, *how do we determine that different variations of a common allele are not incompatible?*
- Secondly, the Allelic Variant concept is no longer needed explicitly. As we have no individual genomes in the model, the absence of individual genomes eliminates the need for managing *Allelic Variants*. As our knowledge of the genomic domain increased, we wondered if reference *Allelic Variants* do really exist. This would mean that there is a catalogue of well-determined variants whose structure and behaviour should be perfectly known. The introduction of this knowledge into the model could be accomplished at any time. But while a precise answer to this question does not exist, we conclude that omitting the *Allelic Variant* class provides a clearer description, conceptually speaking.

In any case, it is possible to generate allele instances using the appropriate combinations of variations, because it can be seen as derived information obtained by applying a set of selected variations to the source sequence of reference. To have instances of an *Allelic Variant* class involves

characterizing the specific set of variations that "*create*" the considered allele. We argue that this v2 representation is more precise because the separation of these conceptual concerns is made explicit, the CM is in a –semantically-speaking- clearer state, and it enables incorporating new knowledge, as satisfactory answers to the open questions are provided by the progress in the genome understanding process. Considering that a set of variations are identified as a semantic whole, an allele would be the result of applying this specific set of variations to the sequence of reference that makes up a chromosome element as a DNA chain of nucleotides. The representation of this allelic knowledge is left out until the next version (CM).

3.2.3 Modeling SNPs

In the initial version, a highly relevant genome concept such as the SNP (*Single Nucleotide Polymorphism*) was not explicitly represented in our conceptual definition. The specialization of different variations accomplished in v2 is more precise and distinguishes between two categories: the *frequency* of the variation, and its known, *precise* or *imprecise description*. Beyond this conceptual simplification, it is important to take into account how SNPs are stored in current, widely-used data sources (as *dbSNP* (Bhagwat, 2001)). Looking at these current representations, we performed a reverse conceptual engineering exercise to include a set of classes in the CMHG that represents this knowledge. We discovered that an SNP is seen in this domain as a potential set of variations in which one nucleotide may appear changed by another. This change is open, meaning that the notion of variation in this case is that one position in the sequence of reference may have different values according to the population studied, and with a given frequency. This is what we have included in v2, through the specialization hierarchy introduced for SNPs. This change leaded to a new discussion. Any precise variation is modelled as an individual variation, where the sequence of reference "*suffers*" a change. However, the way in which SNPs are treated is somewhat different: an SNP defines which nucleotide is altered. It appears in the source reference sequence (through the attribute *"allele"* for the homozygous case, *"allele1"* and *"allele2"* for the heterozygous case).

This representation preserves the way in which SNP data appears in real genomic settings. But the view of SNPs as a set of individual variations suggests that a better representation would be to model SNP as an aggregation of precise (*indel*) variations. This change will better represent conceptually what an SNP is, but the change has to be carefully analysed because the data management of current SNP data repositories should be properly adapted to the new data representation.

3.2.4 Introduction of Pathways-related Knowledge

One of the important innovations in v2 is the extension of the CM with the integration of the *Pathways*. Within the most important biological pathways, we find three main types: (1) *Metabolic*: these make possible the chemical reactions that occur in the organism (i.e., the process of converting food into energy). (2) *Genetic regulation:* these are responsible for the regulation of genes (are responsible for the generation of proteins, which are required for each of the tasks of our body). (3) *Transmission of signals:* these pathways enable the signal to pass from outside to inside the cell and vice-versa.

The pathways play a key role in advanced genomic studies, and that is why their inclusion in this version of the CM is necessary. In this version 2 includes the first of the three types of biological pathways, "*metabolic pathways*". These are a series of chemical reactions leading an initial substrate to one or more final products. The final product of a metabolic pathway can be used in three different forms: (1) to be used immediately, (2) to initiate a new metabolic pathway, and (3) to be stored in the cell. The metabolic pathway is represented in the CM as a combination of events, represented by the relationship between the concepts of "*Pathway*" and "*Event*", which can be of two types. The first type is a single atomic process, or in other words, a process of the simplest type and not reliable enough to decompose into smaller ones (represented in the CM by the "*Process*" class). The second type is a complex process consisting of a sequence of other processes of complex or single type, represented in the CM as "*Pathways*". The association between Pathways and events represents the composition of a pathway (provides information about other previous events that form part of this metabolic pathway). To know the order of the composition of the events in a metabolic pathway, a reflexive relation is defined on the "*Event*" class. The chemical substances that take part in a process are represented in the CM by the "*Entity*" class, which is related to the "*Process*" class by the "*Takes_part*" class. This can happen in different forms: (a) being the main chemical, (b) as a

result of the process and (c) being a regulator of the process two types (*activator* or *inhibitor*), represented in the CM by the "*Input*", "*Output*" and "*Regulator*" classes respectively. Through our CMHG we incorporate genomic data currently used (e.g., *dbSNP*, *Ensembl*, *etc*.), achieving a conceptual representation that meets the needs of the bioinformatics domain. As we said earlier, this evolution aims to improve the conceptual definition of the human genome, and thus leave a conceptual framework for further improvements.

4 CONCLUSIONS AND FUTURE WORK

PM is going to change the way we have historically understood medicine. The new practical context associated with it requires a sound working environment, and the correct application of the adequate SE practices. We face this problem in this work focusing on the need to design a holistic CM intended to capture structurally all the relevant domain information, together with the conceptual complexity associated with the continuously changing context of Precision Medicine.

We assume that conceptual modeling techniques are the basic strategy to design and develop the required sound and efficient *Genomic Information Systems* (GeIS). Most of this work is devoted to reporting how complicated keeping "*alive*" such a CM is, especially due to the rapidly evolving knowledge. The conceptual representation of basic notions has been discussed, emphasizing that the CM applied to this type of environment facilitates the generation of systems that support decision-making processes in the Bioinformatics domain. The domain knowledge must always be prepared to incorporate any required extension in order to meet new needs. This is why the CM is not only useful but also necessary. The initial version (v1) focused on modeling "*Genotyping*" then sought to create a semantic and content description. However, we had to discuss multiple decisions before moving on to our next CMHG v2. Version 2 is characterized by the change in its central axis based on "*genes*" and takes as its axis the concept of "*Chromosome (and chromosome elements)*". This change was made to simplify the schema and provide a more flexible approach to extend it according to the domain evolution. This new version gives us greater precision, and allows us to manipulate data in a more direct way. All these decisions have a direct implication on how data are managed and consequently on how data quality is to be assessed.

Future research work will focus to three main goals: (1) the evolution of the CM by adding new genomic concepts into the CM (i.e., *haplotypes*). (2) The implementation of a complete ETL process, using our CM. The ETL should be able to identify relevant data for a particular phenotype, and to load them conveniently in the DB that represents the conceptual model. (3) develop a proper, unified framework specifically for GeIS. The aim of this framework is to complement the conceptual model with a DQA procedure in order to ensure the quality of the information represented and loaded by the ETL. The achievement of these three goals will provide the required support to the knowledge on which PM is based.

ACKNOWLEDGEMENTS

This work has been supported by the *MESCyT* of the Dominican Rep. and also has the support of Generalitat Valenciana through project *IDEO* (PROMETEOII/2014/039), the MICINN through project *DataME* (ref: TIN2016-80811-P) and the *Research and Development Aid Program* (PAID-01-16) of the UPV under the FPI grant 2137.

REFERENCES

Aguilar, A. (2015). Medicina personalizada, medicina de precisión, ¿cuán lejos estamos de la perfección?. *Carcinos* 5 (2): 32-33.

Aronson, S., and Rehm, H. (2015). Building the foundation for genomics in precision medicine. *Nature* 526 (7573): 336–342.

Jiménez, N. (2014). Una medicina nueva, más inteligente y menos invasiva. *FARMAESPAÑA INDUSTRIAL* pages 72-73.

León, A., Reyes R., J. F., Burriel, V., and Valverde, F. (2016). Data Quality problems when integrating genomic information. *3rd. QMMQ Workshop 2016 (Gifu, Japan)*, pages 173-182.

Instituto Nacional del Cáncer (2015). *Medicina de precisión en el tratamiento del cáncer*. from https://www.cancer.gov/espanol/cancer/tratamiento/tipos/medicina-de-precision.

Olivé, A. (2007). *Conceptual modeling of information systems*, Springer-Verlag. Berlin Heidelberg.

Paez, J.G., Jänne, P.A., Lee, J.C., Tracy, S., Greulich, H., Gabriel, S., Herman, P., Kaye, F.J., Lindeman, N., Boggon, T.J., and Naoki, K. (2004). EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science* 304(5676): 1497–1500

Pastor, O., Reyes J. F., and Valverde, F. (2016). Conceptual Schema of the Human Genome (CSHG). *Tech. Rep.* from http://hdl.handle.net/10251/67297.

Ram, S., and Wei, W. (2004). Modeling the semantics of 3D protein structures. *ER2004*, pages 696-708.

Reyes R., J. F., Pastor, O., Casamayor, J. C., and Valverde, F. (2016). Applying Conceptual Modeling to Better Understand the Human Genome. *ER2016 (Gifu, Japan)*, pages 404-412.

Bhagwat, M. (2010). Searching NCBI's dbSNP database. *Current Protocols in Bioinformatics*, pages 1-19.