Towards an AI-based Genomic Medicine of Precision that Integrates Predictive and Explainable Knowledge Dimensions

Óscar Pastor[1], Salvador Navarro[1], Alberto García[1], Mireia Costa[1], Ana León[1]

[1] VRAIN Research Institute, Universitat Politècnica de Valencia, Spain

opastor@dsic.upv.es, sfnava@upv.es, algarsi3@pros.upv.es, micossan@vrain.upv.es, aleon@vrain.upv.es

Abstract. *Understanding the human genome and deciphering the Language of Life is a grand challenge that modern sequencing technologies are making feasible by generating huge amounts of data whose correct interpretation has yet to be accomplished. To do it, two knowledge dimensions must be integrated: the predictive one, Machine Learning-oriented, that obtain accurate information from data, and the explainable one, Conceptual Modeling-based, that uses a symbolic representation to provide meaning to the data in order to understand and explain the semantics behind predictions. This position report discusses the problem, contextualizes it under a Life Engineering perspective, and it proposes how to face the design of AI-based data management platforms that follows the introduced ideas.*

1 Introduction

Deciphering the language of life is a grand challenge that the modern genome sequencing technologies are making more and more feasible. This position paper discusses how combining two dimensions of Knowledge Engineering -a predictive one, Machine Learning (ML)-oriented, and an explainable one, Conceptual Modeling (CM)-based- can guide the process of understanding the genome, using the genomic medicine of precision as working context.

Software Engineering (SE) has made possible to create a software whose final code representation is a binary language represented by millions of 0s and 1s. This binary language is generated by compiling programming languages whose instructions we, Humans, know well because we have invented them. Programs written in these languages are obtained from Conceptual Models that represent the relevant aspects of the domain under analysis. Model-driven Development research and practice is still being developed to show to transform Conceptual Models into executable programs [Pastor and Molina 2007].

Making an analogy with SE, Life Engineering (LE) is exploring how to discover the instructions of the programming language of life that instead of being a binary language of 0s and 1s, is a quaternary, four-letter language (A, C, G, T) with a biological background, where those four letters represent four nucleotides. Millions of them together constitutes a program of life, which is our genomic sequence. A massive data generation process is continuously running thanks to the unstoppable progress of the genome sequencing technology, what is generating more and more data about the genome.

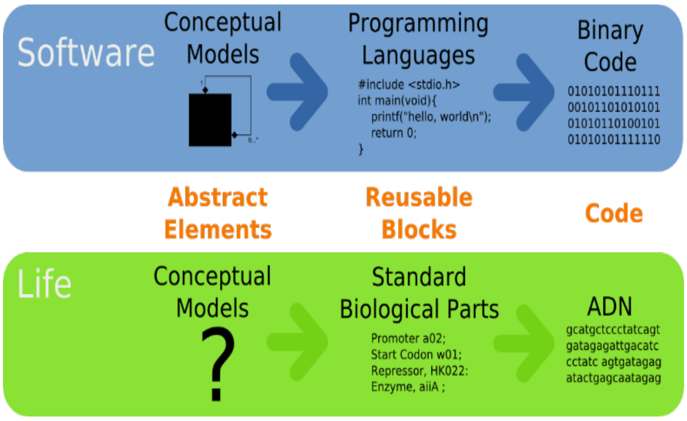
But data without semantics are useless, especially when our final purpose is to understand the genome and discover what are the instructions of the programming language that should guide that understanding process. ML techniques process data and train algorithms that predict valuable information. In our working domain, which is the genome-based medicine of precision this means that if a given patient (person) has a health problem (e.g. colorectal cancer (CRC)). Training a function with a sufficient number of samples from people with and without the considered disease (phenotype), a ML process can solve the predictive knowledge perspective of the game.

But to diagnose a sample with a CRC without explaining why this is the case, is an incomplete result. The explainable knowledge dimension is required. It is incorporated by applying an explainable AI process which is based on the use of a Conceptual Model of the Human Genome (CMHG) that delimits which is the relevant information to be considered to "explain" why the diagnosis is given. Navigating through the CMHG from genotype to phenotype information, the results provided by the ML trained function can be explained. For instance, selecting features related to genome variants that are relevant for the disease, the presence of these variants together with their associated effects in the protein synthesis process and their pathway participation's role can make the wished explanation process feasible.

Through the paper, these ideas are developed in more details. In section 2 we emphasize the analogy between the SE and LE perspectives, characterizing what do we mean by LE as a main process to face the grand challenge of deciphering the language of life. Taking the medicine of precision working domain of reference, and with the goal of articulating what is the valuable knowledge to be considered, Section 3 discusses the predictive knowledge dimension that is associated to a ML-based process, while Section 4 focuses on the explainable dimension, which is guided by the use of a CMHG as the main artefact, and that complements the ML perspective. Section 5 discusses the benefits of combining the predictive and the explainable knowledge dimensions, and how effective software platforms to manage health genomic data can be designed and implemented. Finally, conclusions and references close the work.

2 From Software Engineering to Life Engineering

Figure 1 depicts our comparison between SE and LE. Following a top-down perspective, SE has created programming languages that built programs with which a problem domain is represented. The software production process includes a set of models and models transformations from the conceptual model (that represents the relevant information of the domain under analysis) to the program (using instructions of a selected programming language) and to the code (binary code).



```
#include <stdio.h>
int main(void){
    printf("hello, world\n");
    return 0;
}
```

Figure 1: The analogy between Software Engineering and Life Engineering: from models, to programs, to code

LE follows a data-driven, bottom-up perspective. In this case, sequencing technologies allow us to get data and to know the lowest-level code (quaternary code-based), but we don't know yet which are the instructions of the programming language of life that is hidden behind this huge amount of genomic code that we accumulate. In a reverse way, we have to go from code to (biological) programs, and to models of life. Modeling life is still a further goal, but understanding the code of life is the problem that we are already facing. Our big challenge, a scientific challenge for humanity, is how to discover the instructions of that programming language of life. This is what we mean by deciphering the language of life. To do it, the SE analogy provides a useful approach, that has two main steps:

1. Massive data gathering, that is what modern sequencing techniques makes continuously feasible.
2. Discovery of programming patterns that are behind those data.

As it happens in SE, abstraction is an essential mechanism to make the information obtained from data explainable. The genomic-based medicine of precision provides a right working context to progress in that direction. Once a given disease (phenotype) is selected, comparing the genome (code) of healthy and unhealthy people (step 1), it becomes possible to identify programming patterns that are hidden in the data (step 2).

This is what we are doing in our PROS Research Center (within the VRAIN Research Institute at the Universitat Politècnica of València), and what we present here. Our strategy to advance in that direction is based on the need to distinguish clearly the two dimensions of knowledge that we discuss next:

- A predictive knowledge, ML-oriented because once a phenotype is selected, it trains a function with the corresponding clinical data in order to predict if a person is positively diagnosed.
- An explainable knowledge, that is CM-based: using a holistic conceptual model of the human genome, the relevant pieces of information are identified, and explanations associated to a provided prediction are given.

The correct integration of these predictive and explainable knowledge facets can be seen as the two sides of the coin that any AI-based software platform for managing genome data must accomplish. Such a software platform must achieve the combination of i) the adequate use of ML techniques to get a right prediction (predictive knowledge dimension) with ii) a trustworthy explainable AI, based on a conceptual model that provides a common, shared under-standing of the domain in order to establish cause-effect relationships for understanding the biological behavior and explaining a given prediction (explainable knowledge dimension).

3 The Predictive Knowledge Dimension (ML-oriented)

Using ML techniques, what we refer to as "the magic of mathematics" makes possible to train an algorithm with data in order to generate a model that predicts valuable information with a maximum accuracy [Swanson et al. 2023] In clinical terms, this means to follow the process of:

1. Selecting a phenotype of interest.
2. Understanding the task to be performed and selecting the right scope.
3. Creating the datasets by collecting the relevant data and improve its quality.
4. Selecting the appropriate ML technique intended to deliver the most accurate results in the domain under analysis.
5. Informing the final result.

The final result is the prediction provided by the selected ML technique. This predictive knowledge dimension states that a given diagnosis is associated to the patient whose data analytics process has been accomplished. Large initiatives as The Cancer Genome Atlas (TCGA) provides an effective data background to train different ML approaches [Liñares-Blanco et al. 2021]. This knowledge dimension covers the fast-thinking perspective discussed in [Guizzardi et al. 2023], where a clear distinction in made between "fast thinking" (data-driven) and "slow thinking" (theory-driven) characterization. The "fast thinking" side of the coin focuses on learning from data even if a precise conceptual counterpart is lacking. Paraphrasing Immanuel Kant, "Concepts without data are empty; data without concepts are blind": an explainable knowledge dimension is then needed to have the full picture (the “slow thinking” side of the coin).

4 The Explainable Knowledge Dimension (CM-based)

Explaining a prediction requires to use abstraction to understand what concepts are behind data representations, and what causal relationships explain natural effects. To do it well, any sound Explainable AI process must start with getting a shared understanding of the studied domain [Spreeuwenberg [S.d.]]. This is why the explainable knowledge dimension is CM-based: a holistic Conceptual Model of the Human Genome (CMHG) becomes a key artifact to identify relevant concepts that cover how genome, transcriptome and proteome are connected with the functional counterpart that pathways characterize. In several previous works we have presented and evolved such a CMHG, that provides the semantic basis to explain why a given prediction makes sense [Pastor et al. 2021] [Reyes Román et al. 2016] [Pastor 2008].

This explainable dimension is about concepts’ identification, about how they relate to each other, and about how previous, reported associations between concepts and effects in the context of a particular disease explain the identification of a precise diagnosis. Additionally, this explainable dimension discovers connections between functional biological effects, and their “code” counterpart. Navigating through the CMHG, it is for instance possible to associate genome variations with functional problems that allow us to infer those programming patterns that lead to the emergence of the programming instructions of the Programming Language of Life, the grand challenge behind the final goal of our long-term research work: deciphering the language of life.

Looking at health problems as “software” bugs that can be first identified, then corrected (by using the recently discovered genome edition techniques, the CRISPR-based approaches) opens a window of opportunity that has never before being possible to reach: to understand and subsequently manipulate the human genome. This is where SE and LE converge: LE-based solutions must now provide methods and tools to advance in how to make explainable the predictive, databased knowledge that is provided through genome sequencing. This is where integrating both types of knowledge with a sound, semantic ground becomes a need for any software platform to be designed and developed.

5 Integration & Discussion

We discuss in this section: i) how to integrate effectively the two knowledge dimensions that we have introduced before (predictive and explainable) in a software platform intended to manage valuable genomic data, and ii) what is our starting experience in developing such a software platform, called the “DELFOS Genome Oracle” [León Palacio et al. 2024](DELFOS in short from now on).

Making a reliable prediction and being able to explain it is the result of the proposed combination. AI-based software platforms that deal with sensitive data as the clinical ones must achieve the needed integration. Reusing the “coin” image for such a software platform, one side of it will provide the ML-based trained models that make accurate predictions, while a CM-based XAI process will articulate the other side of the coin.

Looking at a genome data management software platform from a holistic point of view, three steps conform its design process:

1. Selection of the phenotype (disease)
2. Development of an associated ML-based training model (predictive knowledge dimension)
3. Analysis of the relevant genomic data whose effects have been reported in a genomic data source with a concrete clinical significance (explainable knowledge dimension).

This what we are doing with our DELFOS platform in our Research Center (PROS@Vrain). For an initial set of phenotypes that include cardiopathies and pediatric oncology, with the support of clinical experts of three hospitals of the Valencian Community in Spain -Hospital La Fe and Hospital Clinic in Valencia, Hospital General in Alicante- and one local biotech company, we have developed the DELFOS software platform prototype, under the OGMIOS ResearchProject.

Reusing the DELFOS Oracle Greek mythology metaphor, we envision DELFOS as a platform that includes all the known genomic data for the phenotypes under study. Once a particular disease is selected from an initial set of available diseases, an individual user can enter DELFOS with her relevant information that can include two different types of genomic information:

i) the set of features required by a ML prediction process, that should for instance include gene expression or DNA methylation, and
ii) the information associated to the user’ sequenced genome, normally using standard file formats as VCF.

This set of information has a twofold use:

- The first one is used as input for the ML-based model that will give the predictive result, answering the question of “is the user affected by the disease?”.
- The second one is used to explain why that prediction is provided. The explanation is based on the information included in the CMHG, whose

concepts and relationships among concepts determine what effects explain why the reported health problem appears.

An initial prototype of DELFOS is being designed to diagnose (using the predictive perspective) and explain (using the explainable perspective) if an individual is affected by a given disease and why. The explanation generates a clinical report that, for the selected disease, states which relevant genomic variants have been identified with which clinical significance.

Several challenges must be addressed as immediate next steps. Firstly, the platform must evolve according to how new genomic knowledge is generated. As the genome understanding is a scientific field which is in continuous development, new information can lead to new knowledge that should automatically update previous clinical reports. Under the explainable dimension, this happens for instance when one new combination of variants is reported to be associated with a particular disease: having the source patient sequence with these variants whose connection was initially unknown, the former clinical report can be timely corrected/updated. Under the predictive dimension, the ML-based model training strategy can lead to datasets that update the set of selected features depending on that new genomic information.

Secondly, under the predictive knowledge dimension the idea is to design a generic, phenotype-independent method for the ML-based training model process. It should provide a template including the steps to be followed when a new predictive genome service (for a new disease) is incorporated in DELFOS, together with the methodological guidance needed to assist in the selection of the right ML technique by characterizing different possible working contexts.

Thirdly, under the explainable knowledge dimension, a XAI-based process is also to be defined. Strictly based on the CMHG, this process is intended to characterize where to search for the right genomic information for a selected phenotype, how to identify the relevant information from the selected data sources, how to make all that relevant information persistent in a DB engine, and what interaction mechanisms have to be provided for access and querying. We have initial results focused on a subset of the CMHG concerning the genomic variation information. A method called SILE (for Search, Identification, Load and Exploitation) [Palacio and López 2018] has been designed in our Research Center, and it has been initially used under a set of practical projects with our clinical partners obtaining positive results [Costa et al. 2022] [Costa et al. 2023].

Finally, coming back to our Grand Challenge of deciphering the language of life, we keep a long-term objective with our proposal: the more data we accumulate in our DELFOS Oracle, the more information can be discovered about those programming patterns that are hidden in the data. Connections among data which is associated to different phenotypes can provide clues to understanding what are those programming language of life' instructions.

6 Conclusions

Modern AI-based data management platforms must combine two essential knowledge dimensions:

- a predictive one, that is databased, and uses ML techniques to infer valid knowledge from data without being able to semantically explain with a precise conceptual ground why a given prediction is provided.
- an explainable one, that getting a shared understanding of a domain with a Conceptual Model, provides a symbolic knowledge with a precise semantics from where concrete explanations can be obtained.

We have shown in this position paper how Life Engineering can be a discipline conceptually equivalent to Software Engineering, where from a reverse engineering perspective we have huge amounts data that need to be understood. To accomplish this challenge, these two knowledges' dimensions (predictive and explainable) must be combined in a holistic software platform, as we are doing with our DELFOS approach. Our further work is oriented to extend our initial results to accumulating more data and generating more and more knowledge from that twofold knowledge integration strategy, having in mind the long-term Grand Challenge of deciphering the language of life with from our AI-based approach.

7 References

Spreeuwenberg, S. ([S.d.]). AIX: Artificial intelligence needs explanation. – Transparency increases the success of AI based decision support systems.

Costa, M., García, A. S. and Pastor, O. (1 dec 2023). The consequences of data dispersion in genomics: a comparative analysis of data sources for precision medicine. BMC Medical Informatics and Decision Making, v. 23, n. 3, p. 1–17.

Costa, M., García S, A. and Pastor, O. (2022). Conceptual Modeling-Based Cardiopathies Data Management. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), v. 13650 LNCS, p. 15–24.

Guizzardi, G., Pastor, O., Storey, V. C. and Kruchten, P. (1 nov 2023). Thinking Fast and Slow in Software Engineering. IEEE Software, v. 40, n. 6, p. 139–142.

León Palacio, A., García Simón, A., Reyes Román, J. F. and Costa, M. (mar 2024). The Delfos Platform: A Conceptual Model-Based Solution for the Enhancement of Precision Medicine [acepted for publication].

Liñares-Blanco, J., Pazos, A. and Fernandez-Lozano, C. (12 jul 2021). Machine learning analysis of TCGA cancer data. PeerJ Computer Science, v. 7, p. 1–47.

Palacio, A. L. and López, Ó. P. (27 dec 2018). Smart Data for Genomic Information Systems: the SILE Method. Complex Systems Informatics and Modeling Quarterly, v. 0, n. 17, p. 1–23.

Pastor, O. (2008). Conceptual modeling meets the human genome. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), v. 5231 LNCS, p. 1–11.

Pastor, Ó., León, A. P., Reyes, J. F. R., Garciá, A. S. and Casamayor, J. C. R. (18 jan 2021). Using conceptual modeling to improve genome data management. Briefings in Bioinformatics, v. 22, n. 1, p. 45–54.

Pastor, O. and Molina, J. C. (2007). Model-driven architecture in practice: A software production environment based on conceptual modeling. Model-Driven Architecture in Practice: A Software Production Environment Based on Conceptual Modeling, p. 1–302.

Reyes Román, J. F., Pastor, Ó., Casamayor, J. C. and Valverde, F. (2016). Applying conceptual modeling to better understand the human genome. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), v. 9974 LNCS, p. 404–412.

Swanson, K., Wu, E., Zhang, A., Alizadeh, A. A. and Zou, J. (13 apr 2023). From patterns to patients: Advances in clinical machine learning for cancer diagnosis, prognosis, and treatment. Cell, v. 186, n. 8, p. 1772–1791.