




# Towards an Effective and Efficient Management of Genome Data: An Information Systems Engineering Perspective

Alberto García S. , José Fabián Reyes Román , Juan Carlos Casamayor, and Oscar Pastor 

PROS Research Center, Universitat Politècnica de València, Valencia, Spain  
{algarsi3,jreyes,jcarlos,opastor}@pros.upv.es

**Abstract.** The Genome Data Management domain is particularly complex in terms of volume, heterogeneity and dispersion, therefore *Information System Engineering* (ISE) techniques are strictly required. We work with the *Valencian Institute of Agrarian Research* (IVIA) to improve its genomic analysis processes. To address this challenge we present in this paper our *Model-driven Development* (MDD), conceptual modeling-based, experience. The selection of the most appropriate technology is an additional relevant aspect. NoSQL-based solutions where the technology that better fit the needs of the studied domain – the IVIA Research Centre using its Information System in a real-world industrial environment – and therefore used. The contributions of the paper are twofold: to show how ISE in practice provides a better solution using conceptual models as the basic software artefacts, and to reinforce the idea that the adequate technology selection must be the result of a practical ISE-based exercise.

**Keywords:** Conceptual modeling · Genomics · Neo4J · Data management

## 1 Introduction

Understanding the internals of the genome is one of the greatest challenges of our time. The complexity of the domain is tremendous and we have just begun to glimpse the vast knowledge we can extract it. Technology has improved over the years, allowing to sequencing in mass not only the genome of humans but of any organism at an increasingly lower cost. Not only the fast growing of the domain information, but the complexity, variety, variability or velocity makes it a Big Data domain. Special attention has been paid to the understanding of the human genome projected on the potential benefits on our health in the field of *Precision Medicine* (PM) [6]. But not only the study of the human genome can report us a benefit. Keeping the interest in the relationship between genotype and

phenotype, the focus can be put on other species in order to achieve also relevant results. For instance, the improvement in food characteristics has a positive effect on our health. Increasing the concentration of vitamins or improving resistance to weather anomalies are just some examples of possible benefits [13]. Through the years, the costs of DNA sequencing have dropped severally. In 2017 the cost of sequencing a genome has reduced from 100 million dollars in 2007 to 1.000 dollars in 2018. This decreasing in cost has come with increasing speed of sequencing thanks to new techniques [11]. Because of these two factors, the amount of generated data is massive. As an example, the US healthcare system reached 150 exabytes in 2013 [18]. This high volume has followed *Conceptual Modelling* (CM) principles in a very limited way. The result is the existence of a very heterogeneous and disperse data from thousands of data sources. Each one of these data sources stores the information following its own defined structure. Great challenges arise and make necessary the execution of data science-oriented projects to integrate this heterogeneous data. Even though some standards are defined [7], most of the genomics data is mainly unstructured. Dealing with a variety of structured and non-structured data greatly increases complexity. These characteristics hinder the value of the data and make mandatory to accomplish complex data analysis tasks with the goal of extract the hidden value.

The IVIA Research Centre is a recognized centre of reference in the domain of Citrus genome data management [22]. Based on our previous experience on structuring the process of discovering relevant genomic information in the PM domain [14], in this paper we report our work on how their genomic studies methods and the corresponding data management strategies have been improved. This is what we mean by improving efficacy and efficiency from an *Information Systems Engineering* (ISE) perspective: firstly, to apply a rigorous CM process in order to identify and delimit the domain core concepts. This task is essential in such a complex and disperse domain. Secondly, to design and implement the corresponding *Information System* (IS) emphasizing the need to select the most appropriate data management technology.

To accomplish this goal, after this introduction the paper discusses in Sect. 2 the state of the art by studying how CM is applied to the genomic domain and how graph-oriented databases, our selected technology, are used in this domain. Section 3 describes the characteristics of the data that the IVIA manages, and how they are currently working with it, emphasizing the complexity of the working domain. Section 4 introduces the proposed solution, namely, a *Citrus Conceptual Schema* (CiCoS). This CiCoS is the key artefact used to design and develop the associated IS. Up to three types of representative complex queries are used to guide their genomic processes and the selection of the most appropriate data management technology. We present a brief real-world use case in which the schema is successfully used through the implementation of a prototype to obtain value from the data. Section 5 exposes the conclusions and future work.

## 2 State of the Art

We study the state of the art in two different fields. Firstly, we describe how CM is applied in the genomics domain in order to provide deep knowledge to generate value. Secondly, we study the use of Neo4J in the Genomic Data Management domain.

### 2.1 Conceptual Modeling in the Genomic Domain

The CM defines the activities of obtaining and describing the general knowledge needed for a particular IS [12]. The main goal of CM is to obtain that description, called “*conceptual schema*”. The conceptual schemas (CS) are written under so-called “*Conceptual Modelling Languages*” (CML). CM is an important part of requirement engineering, the first and most important phase in the development of an IS [1, 12]. The use of this approach in previous work showed how CM grant a clear definition of the domain, allowing a deeper understanding of the involved activities and their relations. For this reason, it is widely accepted that using CM eases the comprehension of complex domains, namely, genomic domain. One of the first presented papers regarding the application of CM to the genome was written by Paton [17]. His work focused on describing the genome from different perspectives. These perspectives were the description of the genome of a eukaryote cell, the interaction between proteins, the transcriptome and other genomic components, though this work was discontinued.

CM has been also used to model proteins [19], which included a great amount of data with a deeply complex structure. This study was based on search and comparison through the 3D structure of a protein and this goal was easier to achieve thanks to the use of CM. Other approaches arose with the objective of representing genome concepts, i.e., the representation offered by GeneOntology<sup>1</sup>. It aims the unification of terms used on the genomic domain and obtains a thesaurus of terms (see more in [2]). Despite the huge amount of genomic data sources publicly available, it is not usual to find underlying stable CS. This is mainly caused because the accessible data is focused on the solution space and do not tackle the process of conceptualizing the analyzed domain. CM is not only used as an approach to describe and represent a specific domain but also helps on software production. Particularly, MDD has already been used on bioinformatic domain [16, 20]. Gardwood et al. (2006) created user interfaces to examine biological data sources using MDD [9]. Note that CS are rare to find on the genomics domain. On the citrus domain, no other intent of defining a conceptual schema has been found.

The relational databases are a mature, stable and well-documented technology. It has been around several decades and can face and solve almost every use case. Nevertheless, relational databases struggle with highly interconnected data. Relational databases deal poorly with relationships. Relationships are generated at modelling time as a result of the execution of multiple joins over the tables.

---

<sup>1</sup> <http://www.geneontology.org/>.

As the model complexity and stored data size increases, so does the impact on performance. The more complex and interconnected is a domain, the more the efficiency will drop when interrogating that domain. This is caused, in part, by how relational databases physically store the relationships between entities, namely, by using *foreign keys*. The relational model makes some kind of queries very easy, and others more complex. Join tables add extra complexity since it mixes business data with foreign key metadata. Foreign key constraints add additional development and maintenance overhead. With exploratory queries, relations are translated into expensive join operations. This is even worse with recursive questions, where the complexity increases as the degree of recursion increases.

## 2.2 Graphs as Genomic Modelling Entities

Graph structure fits particularly well the genomic domain. Life can be modelled as a dense graph of several biological interactions that semantically represents the core concepts.

Pareja-Tobes et al. created Bio4J (2015) [15]. Bio4J is a bioinformatic graph platform that provides a framework for protein related information querying and management. It integrates most data available in Uniprot, Gene Ontology, UniRef, NCBI Taxonomy and Expasy Enzyme DB webs. Storing information on a graph-oriented solution allowed them to store and query in a way that semantically represents its own structure. McPhee et al. (2016) [10] used graph DB to record and analyse the entire genealogical history of a set of genetic programming runs and demonstrated the potential of Neo4J as a tool for exploring and analyzing a rich matrix of low-level events. Balaour et al. (2017) [3] used Neo4J to implement a graph database for colorectal cancer. This database is used to query for relationships between molecular –*genetic* and *epigenetic*– events observe at different stages of colorectal oncogenesis. They probed that graph DB facilitate the integration of heterogeneous and highly connected data over relational databases. Besides, it offers a natural representation of relationships among various concepts and helped to generate a new hypothesis based on graph-based algorithms. The same author created Recon2Neo4J, that offers a computational framework for exploring data from the Recon2 human metabolic reconstruction model (2017) [21]. Using graph-oriented DB facilitated the exploration of highly connected and comprehensive human metabolic data and eased the identification of metabolic subnetworks of interest.

As we can observe, bioinformatic and genomic domains rely heavily on graph database technology and in Neo4J in particular in a significant way. Neo4J is widely used in the genomic domain for the benefits it brings when modeling the information and eases working with complex and highly interconnected data. These examples show the usage of Neo4J as the technology to implement in the genomics context.

### 3 Data Characteristics and IVIA Context

The IVIA research centre has as main goals the development of plant improvement programs in order to obtain agricultural products with greater resilience and adaptation to increase their diversification and competitiveness. It has several lines of investigation and we collaborated with the one dedicated to obtaining, improvement and conservation of citrus varieties. They get new citrus varieties by irradiation and selection directed by genomic methods and establish phenotype-genotype *relations* through genomic analysis. They work with hundreds of citrus varieties and with a very heterogeneous and diverse set of external data sources. The studies they perform can be of two types. Firstly, to compare citrus groups to determine their differences at a DNA level and then try to establish a correlation with their phenotype: from genotype to phenotype. Secondly, starting from a protein, enzyme or pathway –*functional annotations*– of interest identify the variations that directly intervene with them: from phenotype to genotype. Based on these studies, three types of queries have been defined. The first one obtains differences on the genome of groups of citrus that have different characteristics –*phenotype*– from a global perspective to determine which variations cause these differences. The second one starts from a particular functional annotation and identifies only the variations that directly affect it. The third, unlike the first type of query, focuses the search not on the global genome but in very specific regions of interest. Two challenges arise. The first challenge is how to store and retrieve the data in a quick and, especially, cost-effective way; here lays the importance of selecting the right technology. The second challenge is to integrate all the heterogeneous data they work with in order to be able to extract and analyse it together. This heterogeneity leads to three different problems: technological heterogeneity, schema technology and instance heterogeneity. Based on these conflicts, three strategies can be used: (i) conflict ignoring, (ii) conflict-avoiding and (iii) conflict resolution [4].

The technological heterogeneity is resolved by integrating the data. The schema heterogeneity is resolved by defining a conceptual schema. This schema defines the final data structure and guides the data transformations. It acts as a global source of knowledge and helps on the process of resolving technological heterogeneity by easing the data science project. This conceptual schema is explained in more detail in Sect. 4.1. Regarding the instance heterogeneity, until now the strategy used to deal with was conflict ignoring. This changed to conflict resolution by the application of filters to allow scientists to dynamically remove not sufficiently truthful data from their analysis by defining parametrizable quality filters.

The objective is to extract knowledge by building genotype-phenotype relations in order to establish clear and direct relations between a desired or undesired effect and its genetic source. Let the following idea be a simple example to recognise the potential value of the data that is been working on: determine what variations cause acidness in a specific citrus variety and be able to revert that condition obtaining a sweeter version of that variety.

## 4 Proposed Solution

This section illustrates the generated conceptual schema (CiCoS), the result of an iterative process of discussions with the experts of PROS and IVIA Research centres, and the implemented IS, that is being used in a real industrial environment.

### 4.1 Citrus Conceptual Schema

CM has a vital role in the development of conforming applications. In a such a complex domain, the importance of accurately identifying and define concepts is essential. CM arises as the solution to properly generate the knowledge domain that is needed. The generated schema serves us as an ontological basis and provides all the necessary information. Several sessions were needed to implement the schema. On one hand, the IVIA experts provided their vast biological knowledge to correctly understand and interpret the available data. This knowledge allowed to successfully transform an immense amount of data into a well defined conceptual schema in order to extract value using data analysis. On the other hand, the PROS researchers provided their proved years of experience in CM to properly design and implement the conceptual model. In the course of these sessions, all data was carefully analyzed in order to identify the elements of higher importance. Through an iterative process multiple models where created and expanded until accomplishing a stable version. The resulting schema is a precise representation of the genomic domain tailored to the specific needs of the IVIA. This schema can be grouped into three main views: (i) *the functional annotation view*, (ii) *the structural view* and (iii) *the variations view*.

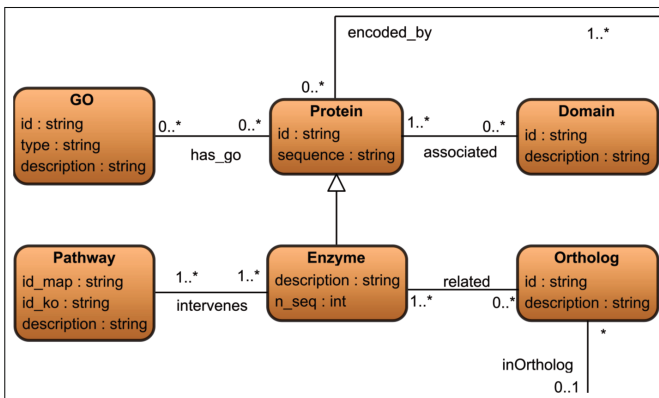
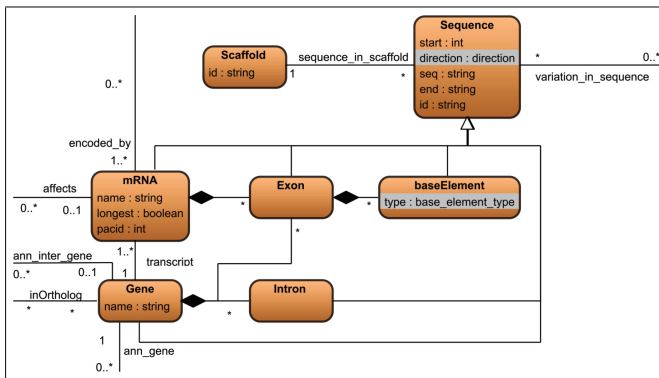


Fig. 1. Functional annotations view

On Fig. 1 the functional annotations view can be found. These view groups information related to functional annotations. Functional annotations are defined

as every element that is related to a gene and has a specific function inside the biological domain (*main elements*). On this view auxiliary elements that add additional information (*secondary elements*) are also included. The objective of this view is to get effective information about how a gene interacts with the organism. Three main elements can be found, namely, the *proteins* –and the *enzymes*, by extension, since an enzyme is a type of protein– and the *pathways*. A protein can be understood as a molecule made of a set of amino acids that are physically arranged in a particular way. A pathway can be understood as chains of biological reactions that happen inside a cell and modifies it in some way. Secondary elements are *domains*, to characterize proteins, *GO* to specify protein functionality and *orthologs* that relates genes with a common ancestor with the enzyme that encode them.



**Fig. 2.** Structural view

On Fig. 2 the structural view can be found. This view establishes a hierarchical structure of the identified genomic elements. This view is not intended to be a comprehensive organizational model but rather tries to ease the analysis process that will be carried on. The most important elements are the *gene* and the *messenger RNA (mRNA)* that is transcribed from the genes. These two elements act as hooks between the other two views. These elements are affected by the elements of the variations view and modify the behaviour of the elements of the functional annotations view. The elements move from general to specific, that is, from larger sequences to smaller ones. All these elements are a kind of sequence. A *sequence* is the parent of all the elements and allows us to abstract the structural hierarchy details from the variations view. Since a variation can be located in multiple sequences we can range between the level of specificity. These sequences are grouped into *scaffolds*. We can understand the concept of the scaffold as an equivalent of a chromosome although a scaffold does not contain the full chromosome sequence. Instead, it contains some gaps of known distance where the sequence content is unknown. On top of this organization, the gene

and the mRNA can be found, a gene can be transcribed to multiple mRNAs and each mRNA translated into a protein. The second stage is formed by the *exon* and *intron* elements, the exon is the part of the gene that is transcribed while the intron is the part that is not transcribed. Finally, the third stage contains the so-called base elements. These elements are the parts that are part of an exon, namely the *coding sequence* –*CDS*–, the *5 prime untranslated region* and the *3 prime untranslated regions*.

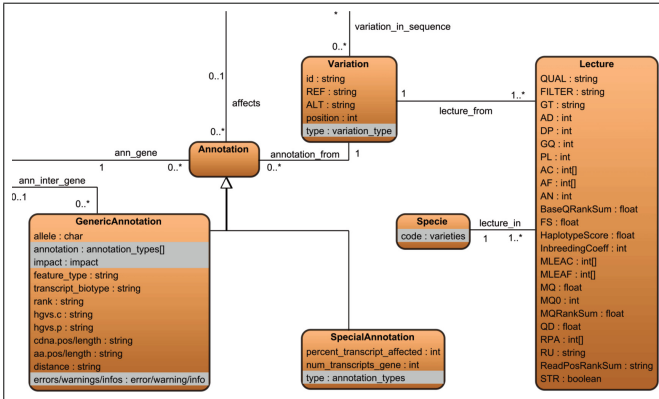


Fig. 3. Variations view

On Fig. 3 the variations view can be found. On this view, identified *variations* on the different citrus varieties are modelled. The different variations can be found in one or more citrus varieties and each occurrence has some values related to how the variation was identified and isolated. These values depend, among others, on the technology used to perform the sequencing, the curation filters and specific varieties characteristics such as the allele frequency or the allele depth. The variations are annotated using SnpEff software [5]. This software predicts the effect an *annotation* will produce on an organism. These annotations are incorporated into the model differentiating the different types of annotations the software can produce: the generic, most common, annotation, and *loss of functionality* –the variation produces a change that provokes a loss of functionality of the protein translated, also known as LOF– and *Nonsense mediate decay* –when the variation degrades the mRNA causing not to be translated into a protein, also known as NMD– annotations, represented as special annotations since they are rather rare. Generic annotations has an attribute called “*annotation\_impact*” that determines the degree of impact of the variation.

#### 4.2 Use Case: IVIA Research Centre Genomic Analysis

After developing the CiCoS, a prototype has been implemented. This prototype is being currently used by the IVIA Research Centre to perform genomic analysis



in a real industrial environment. This helps us get feedback to improve and fine-tune the prototype to finally release a stable version. The prototype helps this analysis process by defining a set of queries that generate value from the stored data. The prototype can be defined as a web application that interacts with an API to execute the defined queries. For this example, we will focus on one of the three types of queries defined.

The data analysis process carried out on our genomic data has proven to be useful to properly identify core concepts of the domain and help in the process of generating effective knowledge, namely, genomic data value. We proceed to evince a use case where the implemented conceptual schema has helped to obtain valuable knowledge through the execution of the first query defined previously.

To detail more in-depth the chosen query, we start by defining two arbitrary groups of citruses. Typically, these groups will share some common characteristic, namely *sweetness* versus *acidness*. The goal is to determine how different the groups are from a genomic perspective. This objective is accomplished by finding which variations are present on one of the groups –the first– and not on the other one –the second–. The aim is to find variations present in all the varieties of the first group and not present in any of the varieties of the second group.

In order to fine-tune the query, we need to be able to apply restrictions regarding multiple parameters. Throughout the sequencing process, each variety comes characterized by a set of quality attributes that defines the degree of truthfulness of it. These attributes can be used to filter additional variations in order to be included or not on the sets depending on the degree of strictness we want to use. Three attributes are used for this purpose, namely, Approximate read depth, Conditional genotype quality and Allele Depth. More information about these attributes can be found at [7].

Additionally, there are three more criteria to filter variations. We refer to the concept of positional depth as the first criterion. It is used to filter variations based on their physical position. As seen on the structural view, multiple types of sequences are defined based on a hierarchical organization. It is of interest to filter variations based on any of the elements of the defined hierarchical organization. This additional filter allows the researchers to focus on more limited regions of the genome. There is another interesting point regarding data quality and the second criterion addresses it. Due to sequencing failures, false positives or false negatives may arise. To deal with these undesirable events, it is imperative to provide flexibility to the group selection: suppose two groups of varieties. We may want to indicate that the variations of interest must be present not in 100% varieties, but in 90% of them. Another possibility is that the variations may be present in at most one of the varieties where they should not appear. The third and last criterion deals with the degree of impact of a variation predicted by an annotation, making it possible to discard variations based on the impact they have on the varieties and allowing us to focus on critical, undefined or neutral variations.

The result is displayed on a table showing information about the selected variations, including the gene or genes they affect and the functional annotations

related to that genes. A more detailed approach on how the data is displayed can be found in [8].

## 5 Conclusions and Future Work

The complexity of the problem that we have solved cannot be faced without using sound ISE techniques. We want to emphasize with our work that using an ISE perspective is essential to provide valid and efficient solutions to complex problems. We have done it working in two main directions. Firstly, the use of CM and MDD has facilitated the understanding of a given domain. Secondly, having a precise, well-defined and standardized conceptual base on which to discuss has eased the knowledge transference that the project required. These techniques also improve data and processes management and allow a more efficient exploitation.

After a study of the genomic domain, analyzing its characteristics and the available technologies, it has been determined that graph-oriented databases are the technological environment that better fits this domain. CM allowed us to define the key artefact to develop the associated IS. Using a technology that makes easier to manage the conceptual model of the relevant domain data allowed us to rapidly adapt and evolve the schema as the understanding of the domain increases. The Graph-oriented databases allow to *capture* and *model* the genomic domain in a more natural way since the domain is composed of highly interconnected interdependent data.

The generated IS is currently being used by the IVIA Research Centre to obtain valuable feedback to implement improvements that allow to speed up more their genomic analysis process. The objective is to extend the CiCoS with the IVIA researchers feedback. Likewise, we want to compare this schema with the PROS Research Centre *Conceptual Schema of the Human Genome* (CSHG) [20]. This will allow us to point out similarities and differences and start the process of unifying models in order to generate a conceptual schema of the genome (CSG) independent of the species.

**Acknowledgments.** This work was supported by the Spanish Ministry of Science and Innovation through Project DataME (ref: TIN2016-80811-P) and the Generalitat Valenciana through project GISPRO (PROMETEO/2018/176). The authors would like to thank members of the PROS Research Centre Genome group, the IVIA Research group, especially Manuel Talón and Javier Terol for the fruitful discussions and their valuable assistance regarding the application of CM in the genomics domain and José Marín Navarro for his advice and help.

## References

1. Aguilera, D., Gómez, C., Olivé, A.: Enforcement of conceptual schema quality issues in current integrated development environments. In: Salinesi, C., Norrie, M.C., Pastor, Ó. (eds.) CAiSE 2013. LNCS, vol. 7908, pp. 626–640. Springer, Heidelberg (2013). [https://doi.org/10.1007/978-3-642-38709-8\\_40](https://doi.org/10.1007/978-3-642-38709-8_40)

2. Ashburner, M., et al.: Gene ontology: tool for the unification of biology, May 2000. <https://doi.org/10.1038/75556>
3. Balaur, I., et al.: EpiGeNet: a graph database of interdependencies between genetic and epigenetic events in colorectal cancer. *J. Comput. Biol.* **24**(10), 969–980 (2017). <https://doi.org/10.1089/cmb.2016.0095>
4. Batini, C., Scannapieco, M.: *Data and Information Quality*. DSA. Springer, Cham (2016). <https://doi.org/10.1007/978-3-319-24106-7>
5. Cingolani, P., et al.: A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* **6**(2), 80–92 (2012)
6. Collins, F.S., Varmus, H.: A new initiative on precision medicine. *N. Engl. J. Med.* **372**(9), 793–795 (2015). <https://doi.org/10.1056/NEJMp1500523>
7. Danecek, P., et al.: The variant call format and VCFtools. *Bioinformatics* **27**(15), 2156–2158 (2011). <https://doi.org/10.1093/bioinformatics/btr330>
8. García Simón, A.: Gestión de datos genómicos basada en Modelos Conceptuales (2018). <https://hdl.handle.net/10251/111666>
9. Garwood, K., et al.: Model-driven user interfaces for bioinformatics data resources: regenerating the wheel as an alternative to reinventing it. *BMC Bioinf.* **7**(1), 532 (2006). <https://doi.org/10.1186/1471-2105-7-532>
10. McPhee, N.F., Donatucci, D., Helmuth, T.: Using Graph Databases to Explore the Dynamics of Genetic Programming Runs, pp. 185–201 (2016). [https://doi.org/10.1007/978-3-319-34223-8\\_11](https://doi.org/10.1007/978-3-319-34223-8_11)
11. NHGRI: DNA Sequencing Costs: Data - National Human Genome Research Institute (NHGRI) (2015). <https://www.genome.gov/sequencingcostsdata/>
12. Olive, A.: *Conceptual Modeling of Information Systems*, 1st edn. Springer, Heidelberg (2007)
13. Paine, J.A., et al.: Improving the nutritional value of Golden Rice through increased pro-vitamin a content. *Nat. Biotechnol.* **23**(4), 482–487 (2005). <https://doi.org/10.1038/nbt1082>
14. Palacio, A.L., López, Ó.P., Ródenas, J.C.C.: A method to identify relevant genome data: conceptual modeling for the medicine of precision. In: Trujillo, J.C., et al. (eds.) *ER 2018*. LNCS, vol. 11157, pp. 597–609. Springer, Cham (2018). [https://doi.org/10.1007/978-3-030-00847-5\\_44](https://doi.org/10.1007/978-3-030-00847-5_44)
15. Pareja-Tobes, P., Pareja-Tobes, E., Manrique, M., Pareja, E., Tobes, R.: Bio4j: an open source biological data integration platform. In: Rojas, I., Guzman, F.M.O. (eds.) *IWBBIO*. p. 281. Copicentro Editorial (2013). [http://iwbbio.ugr.es/papers/iwbbio\\_051.pdf](http://iwbbio.ugr.es/papers/iwbbio_051.pdf)
16. Pastor, O.: Conceptual modeling of life: beyond the homo sapiens. In: Comyn-Wattiau, I., Tanaka, K., Song, I.-Y., Yamamoto, S., Saeki, M. (eds.) *ER 2016*. LNCS, vol. 9974, pp. 18–31. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-46397-1\\_2](https://doi.org/10.1007/978-3-319-46397-1_2)
17. Paton, N.W., et al.: Conceptual modelling of genomic information. *Bioinformatics* **16**(6), 548–557 (2000). <https://doi.org/10.1093/bioinformatics/16.6.548>
18. Pérez, J.A., Poon, C.C.Y., Merrifield, R.D., Wong, S.T.C., Yang, G.: Big data for health. *IEEE J. Biomed. Health Inform.* **19**(4), 1193–1208 (2015). <https://doi.org/10.1109/JBHI.2015.2450362>
19. Ram, S., Wei, W.: Modeling the semantics of 3D protein structures. In: Atzeni, P., Chu, W., Lu, H., Zhou, S., Ling, T.-W. (eds.) *ER 2004*. LNCS, vol. 3288, pp. 696–708. Springer, Heidelberg (2004). [https://doi.org/10.1007/978-3-540-30464-7\\_52](https://doi.org/10.1007/978-3-540-30464-7_52)

20. Reyes Román, J.F.: Diseño y Desarrollo de un Sistema de Información Genómica Basado en un Modelo Conceptual Holístico del Genoma Humano. Ph.D. thesis, Universitat Politècnica de València (2018). <https://riUNET.upv.es/handle/10251/99565>
21. Swainston, N., et al.: Recon 2.2: from reconstruction to model of human metabolism. *Metabolomics* **12**(7), 109 (2016). <https://doi.org/10.1007/s11306-016-1051-4>
22. Wu, G.A., et al.: Genomics of the origin and evolution of Citrus. *Nature* **554**(7692), 311–316 (2018). <https://doi.org/10.1038/nature25447>