Facing the Challenges of Genome Information Systems: a Variation Analysis Prototype.

Ana M. Martinez, Ainoha Martín, Maria José Villanueva, Francisco Valverde,
Ana M. Levín, and Oscar Pastor

Centro de Investigación en Métodos de Producción de Software
Universidad Politécnica de Valencia
Camino de Vera S/N 46022, Valencia, Spain
{amartinez, amartin, mvillanueva, fvalverde, alevin, opastor}@pros.upv.es

Abstract. In Bioinformatics there is a lack of software tools that fit with the requirements demanded by biologists. For instance when a DNA sample is sequenced, a lot of work must be performed manually and several tools are used. The application of Information Systems (IS) principles into the development of bioinformatic tools, opens a new interesting research path. One of the most promising approaches is the use of conceptual models in order to precisely define how genomic data is represented into an IS. This work introduces how to build a Genome Information System (GIS) using these principles. As a first step to achieve this goal, a conceptual model to formally describe genomic mutations is presented. In addition, as a proof of concept of this approach, a variation analysis prototype has been implemented using this conceptual model as a development core.

1 Introduction

Thanks to the breakthrough of the Human Genome Project and the advances in DNA sequencing, an enormous amount of genetic data is being produced by researchers every day. Most of these experiments are focused on the understanding of the relationship between genotype (gene configuration and combination of a particular individual) and phenotype (expression of the genes in a specific human feature). As a consequence, the creation of biological databases and tools to exploit the produced data have grown drastically. However, these tools and databases have usually been defined to support an specific research area or experiment. Therefore, when biologists want to use them for a particular assay, it is very unlikely that they support their specific requirements. This issue leads to a situation where the researcher has to spend a lot of time and effort to perform a simple analysis. Since these bioinformatics tools are not developed using IS principles, they are not aligned with the user requirements. The main consequences of this issue are:

- Some biological databases are only human readable, thus cannot be processed properly in an automatic way.

- The extraction of relevant data is difficult because it is spread around different databases.
- Since several tools are required to analyze the data, the specification of the tooling workflow and integration is far from trivial.
- Inclusion of new studies and bibliography into the available tools turns into a hard task.

With the goal of facing these issues, some researchers have proposed [1] the development of Genomic information Systems (GIS), an IS specifically designed to handle a big amount of genomic data. In this work, a new approach to develop GIS is proposed: the use of conceptual models to organize genomic data and guide the development. Thanks to the close collaboration with biologists in the context of this research project, the gap between the disciplines of Software Engineering and Genetics is solved. The result of this interdisciplinary collaboration is a conceptual model that guides the alignment of concepts among both fields. Therefore the design and implementation of the software artifacts that made up a GIS becomes an easier process.

Following that idea, this paper presents a GIS prototype that analyzes a DNA sequence in order to find documented variations for a specific gene. Once all variations are located in the sequence, the prototype splits them in two groups: one group contains harmless variations and the other one contains variations that produce a change in gene or protein function. For those in the last group, their specific phenotype is reported as it has been described in the literature.

This information is bibliographically referenced and gathered in a report that helps the researcher to understand the genetic meaning of the variation and why it produces a certain phenotype. This is very useful because it can speed up the diagnosis of a specific disease. Furthermore, it is widely accepted that an early disease detection might be determinant. The main contribution of this work is that the GIS development is supported by a set of conceptual model entities that formalize the domain concepts related with genomic variations. As a consequence, the conceptual model plays an integration role to provide the genomic knowledge in an unambiguous way and independent from specific datasource details.

With that goal in mind, the rest of the paper is organized as follows. In section 2 a review of DNA variation analysis tools is presented. Section 3 details a conceptual model for describing genomic variations. Section 4 describes how the variation analysis prototype has been developed. Finally, in section 5 conclusions and future work are stated.

2 Related Work

In recent years, several commercial tools have been developed to provide genomic analysis. These tools can perform tests in order to estimate the customer probability to suffer certain diseases. Navigenics [2], 23andMe [3] and deCODEme [4] are the most relevant tools in this field. The differences between them are briefly summarized in Table 1. However, the accuracy of these tools is far from ideal.

Results are not reported in an unambiguous way because biological concepts are not precisely defined. Without a conceptual model that guides the precise definition of the domain, further integration with external tools is complex to achieve.

Another drawback of these tools is that the only variations reported are SNPs (Single Nucleotide Polymorphism). The conceptual model improves the reports quality because other complex variations such as repetitive insertions or deletions are classified. Furthermore the diseases detected by these commercial tools are constrained to the number of supported genes. The use of a conceptual model overcomes this constraint because provides guidelines to support several gene sequence references and their new discovered variations.

Table 1. Comparison of DNA analysis tools.

	Navigenics	23andMe	deCODEme
Analysis Type	Genotyping	Genotyping	Genotyping
Platform	Affymetrix [5]	Illumina [6]	Illumina [6]
Variations (million)	1 (only SNP)	0.5 (only SNP)	1.2 (only SNP)
Detected diseases	28	51	49

3 A Conceptual Model for Describing Variations

The main objective of the conceptual model presented in this paper is to establish a connection point between the genomic field and the GIS development domain. One of the main characteristics of the genomic field is heterogeneity. The unification of the relevant concepts is a difficult task, since genomic concepts are not precisely defined. Moreover, the field knowledge is still developing and these concepts are constantly evolving, making the organization of all the genetic data available more difficult.

Genetic databases are thus affected by this heterogeneity problem. Each database reflects the concepts according to the interpretation and terminology of a biologist. However, there are different definitions for the same concept; for example, a variation in the DNA sequence is referred under the terms: variation, mutation, polymorphism or SNP [7]. Even though all of them represent more or less the same concept, there are slight differences among them. The problem of heterogeneous data can be solved with the use of conceptual models, as some works have proposed [8]. The development of a conceptual model to represent the human genome is a useful approach to understand this complex domain since precise concepts are defined and related among them. If new concepts, relations or changes are discovered, they can be easily incorporated into the model.

The conceptual model presented here claims to be precise with genetic concepts and IS principles because it has been developed by software engineers and biologists specialized in the genomic field. The model presented in this section is focus on the description of genomic variations. However, it is an excerpt of a widest one [9], whose main goal is the specification of the required human genome concepts for developing GIS.

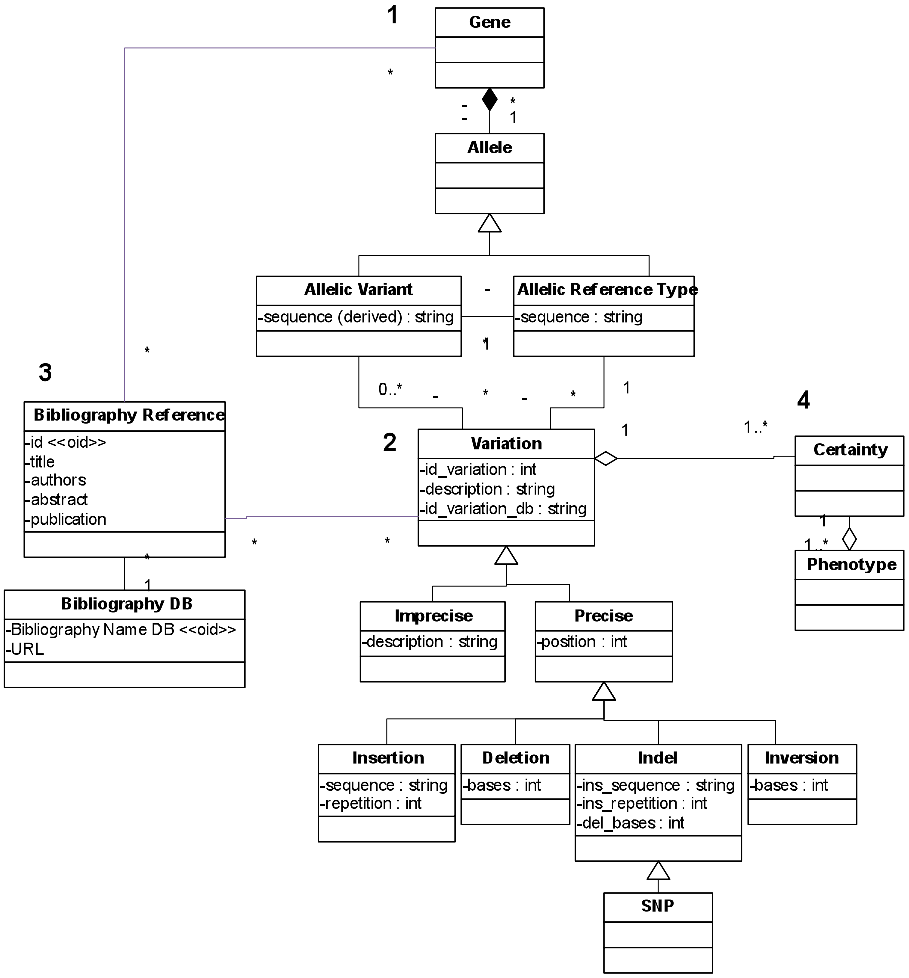

Fig. 1. Conceptual Model for describing variations

Figure 1 shows the proposed conceptual model. At the top of the picture (1) *Gene* and *Allele* modeling entities are defined. *Gene* entity models the generic concept of gene whereas *Allele* entitiy represents the individual instances of a gene. The *Allele* entity has two specializations: *Allelic Reference Type* and *Allelic*

Variant. *Allelic Reference Type* models the reference sequence that definesa "universal" gene to be used for comparison purposes. These reference sequences are extracted for trusted data sources as RefSeqGene database [10]. *Allelic Variant* represents a DNA sequence of an individual which has several variations from the allelic reference.

Each variation discovered by means of the comparison process performed over a sequence, is modeled by the *Variation* entity (2). The *Variation* entitiy stores all the variations documented in the genetic literature that are associated to some disease or to normal changes because of the intrinsic nature of an individual. This entity has two specializations: *Precise* variations, which define a variation that is completely located and *Imprecise* variations, whose location details are not specified. *Precise* variations are also categorized in four entities according to the change performed in the sequence : *Insertion*, *Deletion*, *Indel* (insertion/deletion) and *Inversion*. An indel can be categorized as *SNP* as well when it occurs at least in 1% of the population.

A variation that is specified in the model is always related to its phenotype, which is modeled by the *Phenotype* entity (3). The *Certainty* entity specifies the probability that a phenotype could show up because of a concrete variation on the genotype. In case is identified a genotype-phenotype association, it is essential to know information about the bibliographic reference and the original database where the discovery was stated. This data is defined by the *Bibliography Reference* and *BibliographyDB* entity (4) respectively. As a first result of this conceptual model, a genetic database (GDB) has been created to store the variation information that is used by the presented GIS prototype.

4 A GIS Proof of Concept: a Variation Analysis Prototype

The main goal of the prototype is to show how conceptual models can be useful to define a GIS. One of the most common tasks in the genomic area is the analysis of the genomic sequences [11]. Researchers perform the analysis by doing a comparison between a certain DNA sample from a concrete gene and its reference sequence. The comparison is done using an alignment tool that shows a list of differences among them. After that, an experienced researcher has to decide which variations are relevant and which not. Then, they have to dive into the vast and non-structured amount of information that is scattered across the Web and search the bibliography that justifies each relevant variation. Performing this work manually is a tedious and time consuming task.

The proposed prototype reduces this time by automating the major part of the manual work. This automation can be done thanks to the conceptualization of the domain by the presented conceptual model. Data such as genes, variations, phenotypes and bibliographic references is now represented as perfectly defined conceptual entities. Thanks to this conceptualization, heterogeneity and data dispersion problems are solved, avoiding the manual preprocess of some non-computer legible data and ensuring the quality of the data stored.

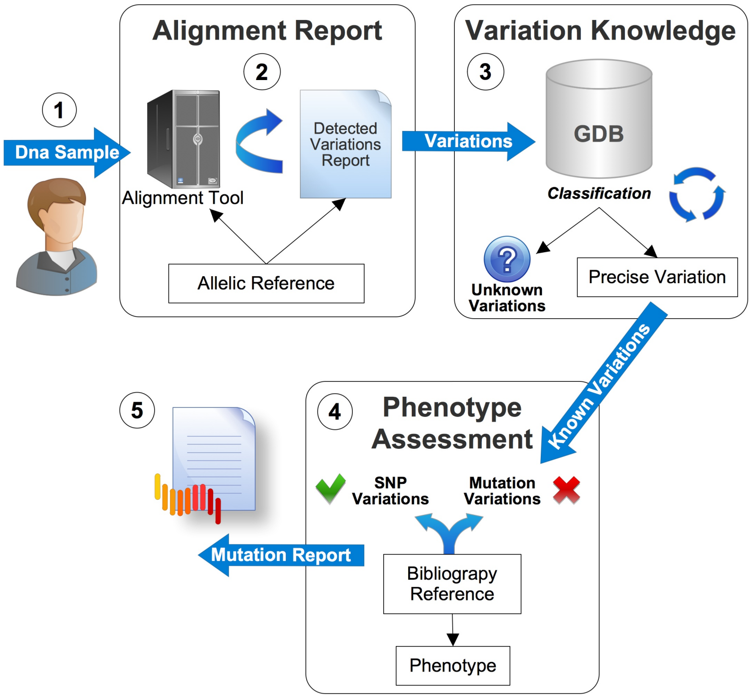

Fig. 2. Prototype Phases

The purpose of the presented GIS prototype is to receive a DNA sample from a patient and provide a report that helps the doctor to diagnose a certain disease. The experts only have to introduce the sample in the suitable format and review the provided results, forgetting everything about manual treatment and endless searches across the bibliography.

The analysis process performed by the prototype is summarized in figure 2. Some conceptual model entities that are used in the different steps are depicted in white rectangular boxes. The process is divided into five main steps:

1. Input data: The biologist selects a gene from the set supported by the prototype, for instance the BRCA1 gene, and introduces the DNA sample to be analyzed. The input of the sample can be performed manually or by uploading a file in FASTA format.
2. Alignment report: According to the selected gene, the prototype locates the suitable reference using the allelic reference entity. After that, an alignment process between the sample and the reference is carried out for finding variations. This alignment is performed using the BLAST algorithm [12], however importing results from DNA sequencing tools as Sequencher [13] will be supported in next versions. Using the defined conceptual model, each discovered difference is formalized as an instance of the variation entity. This formal-

ization, which it is not present at the moment in other tools or databases, is independent of the output from any alignment tool and provides a suitable way for exchanging variations. A report that summarizes all the changes is generated using these variation entities.

3. Variation knowledge: Thanks to the report generated in the previous phase the classification problem is simplified. Variations are located according to a well-know reference sequence and their positions match with the genomic data stored in GBD. Then, each variation is queried into the GDB to determine if it has been defined as a precise variation. If a variation cannot be found in our GDB is classified as unknown. At this point, known variations are classified into an specific type of sequence change. Unknown variations are classified as non-silent if the variation produces a change, in other words, an effect in the expected gene product (protein).
4. Phenotype Assessment: Variations classified as known may have some phenotype associated. In order to asses if the phenotype is related to an specific disease, a research publication is required to provide a trustful evidence. For those cases, the conceptual model describes the bibliographical reference that supports the phenotype for an specific variation. In the context of this work, variations with a pathogenic phenotype are classified as mutations whereas they are classified as SNPs if no negative phenotype is described.
5. Report creation: All the obtained information is gathered in a report. This report contains information about the variations found: mutations, variations whose phenotype is not a disease and unknown variations. Each variation is provided with the following information: the location where it was found in the sequence, its type (Insertion, Deletion, Indel or Inversion) and the number of nucleotides inserted or deleted. For the mutations found in the GDB their associated phenotype and its bibliography is added as well. Finally, the report file can be saved as a text document.

5 Conclusions and Future Work

This work proposes a GIS engineering solution in order to solve the problems of heterogeneity on the genomic domain. A conceptual model is presented which describes and defines formally the concepts related to genomic variations. As a proof of concept, a GIS prototype, which uses this conceptual model as background, has been implemented.

One of the advantages of using the presented GIS prototype is that the variation analysis can be performed using only one tool, avoiding the data workflow. In addition, using a conceptual model to guide the development simplifies the acquisition of the genetic data and can be precisely linked to the bibliography.

However, the study of the prototype performance working with real DNA samples must be analyzed. In order to fulfill this task, further studies related with sequencing algorithms and tools will be carried out.

Conceptual modeling of genes is not a completely novel research area. Some works [14] [15] [16] to organize the genomic data have also been proposed before. The main contribution of the presented work is that the conceptual model proposed here is specifically designed to guide the implementation of software artifacts using a model-driven development approach.

As further work it is planned to extend the GIS prototype with the aim of achieving a higher accuracy and to facilitate the input of sequences. As a final goal, the GIS prototype will be tested in a real environment by means of a collaboration with IMEGEN, a genomic medicine institute, and a couple of local hospitals.

Acknowledgments

This research work has been developed with the Generalitat Valenciana support under the project ORCA (PROMETEO/2009/015)

References

1. Gilbert, D.G.: Eugenes: a eukaryote genome information system. Nucleic Acids Research **30** (2002) 145–148
2. Navigenics. `http://www.navigenics.com` (2010)
3. 23andMe. `https://www.23andme.com` (2010)
4. deCODEme. `http://www.decodeme.com` (2010)
5. Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B., Speed, T.P.: Summaries of affymetrix genechip probe level data. Nucleic Acids Research **31** (2003) e15
6. Klein, R.: Power analysis for genome-wide association studies. BMC Genetics **8** (2007) 58
7. den Dunnen, J.T., Antonarakis, E.: Nomenclature for the description of human sequence variations. Human Genetics **109** (2001) 121–124
8. Richesson, R., Turley, J.P.: Conceptual models: Definitions, construction, and applications in public health surveillance. Journal of Urban Health **80** (2003) i128
9. Pastor, O., Levin, A.M., Casamayor, J.C., Celma, M., Villanueva, M.J., Eraso, L.E., Alonso, M.P. Enforcing conceptual modeling to improve the understanding of human genome. Research Challenges in Information Science (RCIS 2010)
10. NCBI: The RefSeqGene project. `http://www.ncbi.nlm.nih.gov/RefSeq/RSG` (2010)
11. Stevens, R., Goble, C., Baker, P., Brass, A.: A classification of tasks in bioinformatics. Bioinformatics **17** (2001) 180–188
12. Altschul, S., Gish, W., Miller, W., Myers, E.W., Lipman, D.: Basic local alignment search tool. Journal of Molecular Biology **215** (1990) 403–410
13. Gene Codes Corporation.: Sequencher. `http://www.genecodes.com` (2010)
14. Consortium, T.G.O.: Gene ontology: tool for the unification of biology. Nature genetics **25** (2000) 25–29
15. Paton, N.W., Khan, S.A., Hayes, A., Moussouni, F., Brass, A., Eilbeck, K., Goble, C.A., Hubbard, S.J., Oliver, S.G.: Conceptual modelling of genomic information. Bioinformatics **16** (2000) 548–557
16. Ram, S.: Toward Semantic Interoperability of Heterogeneous Biological Data Sources. In: Advanced Information Systems Engineering. Springer Berlin / Heidelberg (2005) 32