Genomic Information Systems applied to Precision Medicine: Genomic Data Management for Alzheimer's Disease Treatment

Ana León Palacio *aleon@pros.upv.es*
Research Center of Software Production Methods (PROS)
Universitat Politècnica de València,
Valencia, Spain

Ignacio Pascual Fernández *i.pascual@pros.upv.es*
Research Center of Software Production Methods (PROS)
Universitat Politècnica de València,
Valencia, Spain

Óscar Pastor López *opastor@pros.upv.es*
Research Center of Software Production Methods (PROS)
Universitat Politècnica de València,
Valencia, Spain

Abstract

The Alzheimer's Disease is one of the most prevalent neurological disorders in our current society. The study of the genetic characteristics of every patient, makes possible the study of significant DNA variations in order to ease an early diagnosis, essential to stop the progression of the disorder. The problem is that the vast amount of available information makes necessary the use of a method designed to adequately store and manage this data in an optimal way for its exploitation. In this context, the Information Systems Engineering in general and the conceptual modelling techniques in particular, provide a suitable solution in order to determine which data is relevant and how to manage the corresponding information. With these fundamentals in mind, this paper introduces a particular example to bear the methodological treatment of the search, filter and load of genomic variations related to Alzheimer's Disease for its later exploitation with clinical purposes.
Keywords: Conceptual Modelling, Alzheimer's Disease, Genomics, Information Systems.

1. Introduction

Personalized medicine has been gaining strength in the last years with the aim of specializing the care of patients improving their quality and expectancy of life. The study of the individuality of each patient can be addressed by reading (sequencing) its DNA. Thanks to the continuous progress achieved by genomic sequencing technologies it is possible to know the predisposition of developing certain diseases (phenotypes), what is called genotype-phenotype association.

The huge amount of data that the continuous advances in sequencing technologies are generating, together with the heterogeneity and diversity of the existing data sources, makes the process of searching and identifying the relevant data, extremely hard. In this context, the study case cited in this article explains how relevant data have been obtained from a research about the phenotypical expression of Alzheimer's Disease (AD), and how to manage them effectively and efficiently. By using appropriate information systems techniques and conceptual modelling techniques is the only way to move in the right direction, managing data correctly by providing

i) a conceptual model to characterize relevant data, and
ii) a sound method to search and identify those data that are considered clinically significant for the selected phenotype (AD in our case).

Data about genotype-phenotype associations are stored in specific repositories for its exploitation in population studies. What makes this domain so challenging is the huge amount of potentially valid data sources. In its last update, The NAR Online Molecular Biology Database Collection summarizes information about 1,737 repositories [1]. But not all of them are suitable to understand the genetic characteristics of a specific disease. As this process involves a great quantity of information, an efficient method is strictly required to obtain the relevant clinical data and their optimal management for its later exploitation.

The method followed to select relevant data, explained in detail later, is based on an optimal database selection looking for the characteristics and goals of our work. A first selection was performed based on the scope, content and disease specificity of the databases. The data of 25 databases storing information about Alzheimer, human DNA mutations and gene-disease associations were examined in order to evaluate its quality and to make a filtering with the purpose of excluding those repositories that do not add enough information. In the end, 4 repositories were selected as reliable enough for the task at hand: Clinvar, Alzforum Mutations, DisGeNET, and ENSEMBL.

By using these criteria, truthful and contrasted information of quality was obtained to be exploited in the context of the clinical diagnosis focused on AD. Making possible this clinical early diagnosis constitutes an extremely relevant social value of this work, as AD is a degenerative neuronal disorder, and stopping it "in time" becomes a key to the expectations of controlling the disease. This is only possible through accurate studies of its genotype-phenotype association, by managing the right data with the right method, what conforms the contribution reported in this work.

Concretely, this paper introduces as essential contribution the process (together with the accumulated experience of its use) to determine how to search, identify and manage in a software platform for the genomic diagnosis, a set of relevant genomic variations for the AD. With this purpose in mind, after this introduction, in section 2 the problem is contextualized, introducing genome-based characteristics of the disease and analyzing in section 3 data repositories where information about its relevant genomic variations can be found. Section 4 presents the conceptual model that plays an essential role to fix which are the relevant data to be considered. Then, in section 5 all the previous pieces are put together in the methodological work that was carried on and whose practical application experience the paper discusses. Finally, conclusions, future works and references complete the paper.

2. Problem Statement

Alzheimer's Disease (AD) is a type of dementia that mainly affects to the capabilities and functionalities of the brain, decreasing them and avoiding the patient's normal life development.

This type of dementia is characterized by affecting memory, followed by the language, thinking, orientation and behavior among other factors. Nowadays, there is not complete certainty of Alzheimer's cause. Physiologically it is observable that when AD is advancing, the number of neurons is reduced, the accumulation of an abnormal protein (β-amyloid) increases and the brain starts to show injuries as senile plaques and neurofibrillary tangles [2]. These characteristics make clinical imaging and genome diagnosis essential to treat the disease: clinical imaging to detect visually the physiological, structural brain problems, and genomic diagnosis to identify accurately the genome variation sources of the abnormal protein production.

The social value of adopting correct information systems design techniques to better diagnose and treating AD becomes obvious: it is the most common dementia, contributing in a 60% to all of them. Considering the rise of the expectancy of life addressed by population during decades, it is possible to realize how serious this disease is in our current society. According to World Health Organization (WHO), 47 million people suffer from dementia and every year over 10 million more are registered [3].

The prevalence of AD in low and medium incomes countries is high, but in high income countries is also a disturbing topic and it has become a big social and economic problem for

health systems. That is because it is the main cause of dependency of old people and because the irreversibility of the disease, as nowadays it is only possible to reduce the speed of the disease's advance. In economic terms, some research carried out in 2015 placed the social in US$ 818 000 million. WHO also estimates that from 60 years old, from 5 to 8% of the population will suffer dementia once in their life and the number of affected patients will grow up yearly in the next century. At this point, it is possible to understand the importance of the reduction of these significant statistical numbers. For that task it is important to know the problem to face.

To close this brief section dedicated to understand the clinical working context, it is important to remark that AD can be classified, attending to the symptomatology showed, in early-onset and late-onset.

- Early-onset AD: The disease starts to show symptoms before 65 years old (commonly around 50). This variant of the disease is less frequent than late-onset, however, genes which intervene in the development of the disease are well known nowadays [4]. This is why many studies talk about early-onset AD as Familial Alzheimer Disease.
- Late-onset AD: Symptomatology is visible after 65 years old approximately. This variant of AD and the early-onset one, have been studied to determine that both could be hereditary. Nevertheless, the association between late-onset AD and the genes involved are not as studied as the early-onset ones.

The genes affected in early-onset AD are well known. However, late-onset AD has not a wide and contrasted list of confirmed genes, so due to our strict reliability criteria for effectiveness data exploitation the work will be focused on different forms of early-onset AD disease and familial AD. The genes affected by this disease are three:

- APP, responsible gene of the amyloid precursor protein synthesis [5].
- PSEN1 and PSEN2, which intervene in regulation and moderation of β-amyloid proteins [4, 14].

If we look at the diversity of the data that the genomic characterization of the studied disease has, it is clear that we face an extremely complex data management problem, where the diversity in the data sources, the data heterogeneity, redundancies and inconsistencies, make strictly required the use of sound Information System-based approaches. It has been surprising for us to detect that this too often is not the case in applied bioinformatics, where ad-hoc, solution-space oriented results are much more common than a careful conceptual-based data analysis. This damages the effectiveness of this particular complex working domain, and it is indeed the problem that we want to better solve with this work.

The first step for the methodological support that we want to characterize must be to select adequately the relevant data sources to obtain information from. Let's analyze how the problem has been solved for the Alzheimer case.

3. An overview of Alzheimer-related Data Sources

Nowadays, many repositories with genomic information are publicly available, but finding repositories where contrasted direct associations between variations and diseases (the genotype-phenotype connection) may be quite more difficult. The work to be done is strongly dependent on this information, so it is important to know the sources where relevant information for the research process is going to be found. Through the method that we present in next section, up to 25 possible repositories were examined, and only 4 of them were finally selected: Alzforum Mutations, DisGeNET, ClinVar and ENSEMBL.

Why have we selected those 4 public data sources commented above and not others? DisGeNET is one of the biggest public collections of genes and variations linked to human diseases. Thus, it is possible to select the disease and through a list of genes it is possible to find those which are relevant for this study and analyze, variation by variation, the gen associated to the chosen disease. The information offered can be contrasted by accessing to its associated

bibliography. Moreover, it is possible to know the quality of the genotype-phenotype relation through a score which introduces how strong or weak is this link [6].

Maybe ClinVar is the most well-known repository of this work. ClinVar is a freely accessible, public archive of reports of the relationships among human variations and phenotypes, with supporting evidence. It belongs to National Center for Biotechnology Information, what makes its control and quality more regulated [7].

Finally, ENSEMBL presents itself as a "genome browser for vertebrate genomes that supports research in comparative genomics, evolution, and sequence variation and transcriptional regulation". It owns a specific browser which filters information from the repository and makes the information appear more visual and useful [8].

These public data repositories can be complemented with others specialized in a particular disease that can provide additional valuable data. Indeed, Alzforum is a webpage where a lot of contrasted information about Alzheimer is exposed. It contains various repositories, where the Mutations database can be found [9]. Mutations is the only specific repository about AD that is also used in this article. Its goal is to introduce a list of variations about specific genes and relate them to clinical, neuropathological characteristics and functional effects. This repository was selected as it works on the specific genes APP, PSEN1, PSEN2 and MAPT, being three of them the main genes studied in this article.

In next section we present the conceptual model that plays an essential role to identify the relevant data to be considered from each selected data source.

4. Conceptual Model

Conceptual models let conceptualize a specific domain in order to ease its understanding. Corresponding relational schemas (if the represented database of the conceptual model is relational) provide an operational base which allows a more efficient and correct data management. Due to the huge amount of data that must be stored and managed when we talk about disease genomic diagnosis, a precise conceptual model plays an essential role to represent the complexity of the genomic variations, their association with the phenotypes and the way to structure the information.

Fig 1. introduces the conceptual model that determines what relevant data attributes must be captured from the selected data sources. The selection of data is then conceptual model-guided, and this is a basic characteristic of the method presented in the next section. The attributes of the conceptual model fix which data must be captured from the corresponding data sources. Once the mappings between conceptual model attributes and their associated counterparts in the data sources are specified, the method is ready to accomplish a process of data load in the database used by the final software platform that will perform the genomic diagnosis report.

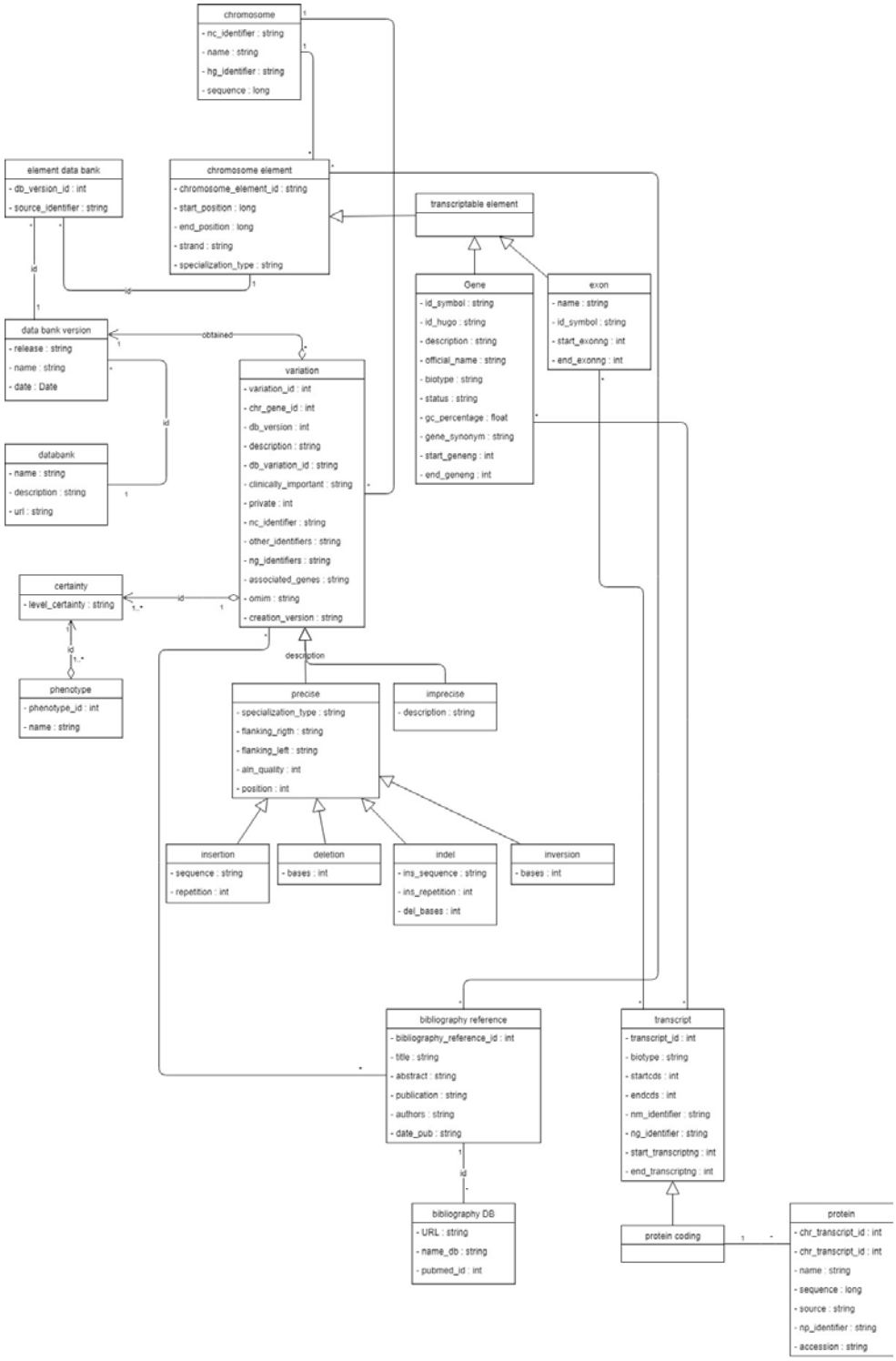

Fig. 1. View of the Human Genome Conceptual Model used in this work [10].

This Conceptual Model itself is not a contribution of this work; It is part of the Human Genome Conceptual Model designed by the PROS Research Center in UPV. It is currently in its 3rd version (HGCM v3) and (according to [10]) it maintains the essential genome information through its 5 main views (structure, transcription, variations, chromosome and metabolic routes). The contribution of this paper is how the Conceptual Model becomes the key component of the method presented next, by determining what information from the selected data sources must be captured and loaded in the database.

5. SILE: The Method for the Alzheimer's Disease Research

As mentioned before in this paper, genomic diagnosis requires a truthful and reliable high quality data [11]. Thus, a set of filters which provide desired characteristics for our data must be defined. In next section, the sequence of actions carried out to make effective and efficient the filtering task, is exposed, with the aim of obtaining a curated dataset that includes the right variations for clinical purposes. This series of steps is based on the SILE method, whose application for the AD case is reported here as a proof of concept.

The SILE method is proposed as a solution to the bottleneck that currently affects the identification of relevant DNA variations with clinical purposes. The process is performed by experts in the field who extracts the information from the available literature by reading hundreds of studies. This constitutes a high effort and the use of public databases is not widely extended due to the huge number of repositories and the increasing difficulty of integrating the heterogeneous data they store. SILE provides a systematic method to extract knowledge from these repositories based on well-defined quality criteria that assure the reliability of the results.

There are some proposals which integrate information from multiple repositories or extract data from literature, such as DisGeNet[1], SNPedia[2] or LitVar[3], but they don't take into account the veracity or relevancy of the information managed, a key aspect when the information is going to be used in a clinical environment.

The SILE method goal is to accomplish the optimal information identification and its efficient load in the Human Genome Database. The method follows 4 essential steps named Search, Identification, Load and Exploitation. Each SILE's step is necessary to ensure the right application of the next step.

5.1. Search

The Search step is a process intended to select certain databases and sources that provide reliable quality information, i.e., it is addressed directly to data sources. As we have advanced in section 3, AD has specific disease repositories (as Alzforum Mutations), but most of the information is located in general databases about genotype-phenotype variations which add extra knowledge.

In the AD case, the Search step started with a list of 25 possible repositories, 21 of which were discarded by the application of filters, resulting in the final selection of 4 data sources. The list of filters is summarized in Table 1.

Table 1. Set of filter used to select reliable databases.

	Filter description
Trustworthiness	The information must be reviewed by a group of experts in the field.
Reputation	The data source must be supported by any research center, association or institute with national or international relevance.
Timeliness	There must be date stamps about when data were entered as well as when last modified.
Currency	The database must be active (less than 1 year from the last update).
Availability	The information must be public and freely accessible.
Usefulness	The information must be relevant and useful for the task at hand.

One of the first selected databases was Mutations from Alzforum because of its AD specificity. The amount of data offered by the database about AD was bigger due to the main goal of the repository. Moreover, data curation is provided by a team with backgrounds in science journalism, information technology, design, and data science. Together with a

[1] http://www.disgenet.org/web/DisGeNET/menu;jsessionid=1a9s465nw2fyxbalv0dyqb90m

[2] https://www.snpedia.com/index.php/SNPedia

[3] https://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/LitVar/#

distinguished Scientific Advisory Board and the active participation of a global network of scientists, they produce high quality and rigorous data. This argument is supported by the proportion of articles which references each variation.

Another useful repository is DisGeNET. As also mentioned in Section 3, this source is specialized in the genotype-phenotype relation in human diseases. It means that with a simple search in the database about AD a great amount of data about the disease can be retrieved. The information is organized attending to curation levels where we can distinguish curated data from other sources (ClinVar, UniProt, PsyGeNET, Orphanet…), animal models and a wide variety of information about variations extracted through software directly from the literature. Data mining can be done in many ways but the easiest one is through its interface which provides a disease browser where finding associated genes, and variations is possible. The added value of DisGeNET is the classification of variations according to a score. In this repository, information about updates of the database and its information can be easily found.

ClinVar also was selected for the task as it offers reports about relations between human genomic variations and phenotypes with demonstrated evidence. Moreover, ClinVar eases access to and communication about the relationships asserted between human variation and observed health status, and the history of that interpretation. This repository is reviewed weekly and big updates are released monthly, so its updating tax is higher than other repositories. Interpretation of variations is not carried out by ClinVar, but the curation is carried out by experts who propose guides and practices to control the submission of results to the database. The quality classification of curation is through the number of submitters attending coincidences or discrepancies. For that task, it is used a star rank based on a hierarchy where 1 star is less curated than the maximum of 4 stars.

Finally, Ensembl database was chosen because it is a repository specialized in the human genome. Among all available databases, Ensembl has the best interface to filter information and select the required characteristics to ensure quality data according to the conceptual model. The database provides information and detailed dates about future patches which can modify the database and its data. As mentioned before, it owns an advanced browser to find variations associated with AD and to obtain the needed information for the Load process.

5.2. Identification

Once the data sources have been selected, the Identification step starts. At this point, the information must be extracted in order to be analyzed later to obtain a curated dataset for the exploitation. The result is a candidate set of relevant variations, that must follow a process of filtering in order to determine the final, reduced subset of totally reliable data. Data reliability is essential considering its clinical, diagnosis-oriented purpose. This part of the global method has the higher cost as it depends on the application of different filters that we detail below.

It is important to remark that the “Identification” step is strictly linked to the conceptual model. Inside the model, there are many attributes for a variation. In this work, an analysis of the essential attributes that a variation must have was carried out in order to determine what a complete variation is and how to solve the problem in a real case. In Table 2 the set of characteristics selected from the Conceptual Model is shown. Each attribute in the conceptual model has a counterpart attribute coming from the corresponding selected data source. In the repositories that support this study, different values for AD variation attributes were found. As it happens, it was firstly decided to select all variations which show up the following characteristics:

1. The criteria used to stablish the association between the variation and the disease must be provided (e.g. at least one star in ClinVar).
2. The condition or phenotype must be showed.
3. The variations must be identified by using the *rs identifier* (i.e. rs2918276).
4. They must have clinical significance.

In that way, variations which does not comply these requirements would be discarded and the final dataset would be more valid. Starting by this criteria, 205 valid variations were

obtained from Alzforum Mutations, 97 from DisGeNET, 110 from Clinvar and 59 from Ensembl. The result was 471 variations in our initial dataset where other filters must be applied.

Table 2. Correspondence between the attributes of the HGCM and the attributes provided by each database.

	ClinVar	**Ensembl**	**DisGeNet**	**AlzForum**
CHROMOSOME_NAME	Genomic location	Location	Chr	Position
CLINICALLY_IMPORTANCE	Clinical Significance	Clinical Significance		Pathogenicity
DBSNP_IDENTIFIER	dbSNP	Archive dbSNP	dbSNP	dbSNP ID
STRAND		Location		
POSITION_START	HGVS	VCF	Position	Position
REF_SEQUENCE	HGVS	VCF	Alleles	Position
ALT_SEQUENCE	HGVS	VCF	Alleles	Position
SYMBOL	Affected gene	Gene	Gene	Gene
LEVEL_CERTAINITY	Review Status	Evidence Status	Score	
PHENOTYPE_NAME	Condition(s)	Phenotype	Disease	Clinical Phenotype
PUBMED_ID	Citations	PMID	PMID	Citation

At this point, a new filtering step was applied to ensure even more the veracity of the dataset, discarding every variation which has no citations or references to studies where the information could be contrasted. After applying this filter, the final set was stablished in 176 variations. With this final high quality dataset generated, and having the connection between attributes of the conceptual model and attributes of the data sources precisely specified, the method was ready to move to the next stage intended to load the generated data in a specific database.

5.3. Load

In this step of the process, the extracted dataset of AD variations was loaded in the human genome database (HGDB), created for clinical purposes oriented to the precision medicine. The concrete database management system is completely independent of the conceptual model. In our case study the HGDB is a relational database implemented in MySQL. Nevertheless, the model could have been implemented in any other type of database by making the corresponding transformation from the conceptual schema to the logical and physical ones. This transformation is out of the scope of this work, but it is important to emphasize that the SILE method is totally “agnostic” with respect to any selection of database technology. There is indeed a transformation using a Graph Database (Neo4J) being currently under developed in our research team.

The reported work is in any case an illustrative example of the method to find and organize clinical genomic data from repositories with the purpose of exploiting them efficiently. The goal of this “Load” step is just to fix the structural organization of the information inside the database, in a conceptual model-based process.

As said before, Conceptual Modelling is important in the Identification process to know which attributes of the repository will be useful; nevertheless, it is in the Load stage where they are organized to obtain a working solution, and to make the data accessibility intuitive for the clinical context.

VARIATION_ID	CHR_GENE_ID	DB_VERSION_ID	DESCRIPTION	DB_VARIATION_ID	CLINICALLY_IMPORTANT
2	1	2	(NULL)	98101	Pathogenic
3	2	1	(NULL)	rs63750197	Likely Benign
4	2	2	(NULL)	8852	Likely Benign
5	1	1	(NULL)	rs362344	Likely benign
6	1	2	(NULL)	314000	Likely benign
7	1	1	(NULL)	rs362384	Likely benign
8	1	2	(NULL)	313954	Likely benign
9	2	1	(NULL)	rs6759	Likely benign

Fig. 2. Example of loading 8 variations with its identifiers in the database for the gene, the identifiers for the version of the database, the identifier of the variation in its corresponding database and its clinical importance.

Fig. 2 shows how variations can be accessed and managed when they are available through a functional database support. The precise link between conceptual model, genome data sources and selected databases assures the understanding of the final software product where data are to be interpreted and managed in real clinical settings.

5.4. Exploitation

This is the final step of the developed example of this project. With the data load completed, the subsequent database is the core of the software platform intended to make possible the genomic diagnosis by professionals. The selected tool is called GenesLove.me [12] and its purpose is to analyze the information obtained from a patient sample and determine if there are variations associated to a certain disease, according to the data stored in the database.

6. Results and Discussion

The final result of this process consists on a dataset, whose quality, veracity and validity are assessed by the methodological background provided by the SILE method. Data are loaded in a database and are ready to be used in an effective precision medicine scope. The reported work shows up the main objective: to demonstrate the current difficulty of the treatment of genomic data related to a characteristic disease such as the Alzheimer's Disease is. The reach of the desired goal of the study –the set of AD-related relevant variations-, is a complex task. To fix a concrete set of requirements has been essential to exclude those variations that could not fit the required quality criteria. This set of quality requirements is dynamic: potential changes are to be considered to adapt the filtering process to the information that is continuously appearing in the analyzed domain.

During the initial process of searching, most of the databases were excluded as they did not add precise information, i.e. they were giving references to other repositories, generating data duplication when different sources really referred to the same result. That is a good example of the need to specify strict filters to avoid future quality compromises.

During the identification process, most of the difficulties were found in relation to the acceptation or rejection of variations. Genomic data repositories are very different. The lack of common criteria to represent the same information generates the existence of different attributes with a wide variety of nomenclatures. This makes the Identification process more difficult, and makes automation become complicated, requiring an intense handwork by the researcher. Additionally, non-existing information from the conceptual model has to be managed, finding out the right data sources where the lacking information could be recovered. The use of the conceptual model provides the needed semantic reference to identify the required data, and the methodological background provided by SILE structures the process to be followed to conform a correct, genomic IS management.

More examples can be discussed here. Different repositories may contain different amounts of information, sometimes even distinct. In the same way that information about updates have been found for ClinVar or ENSEMBL, Mutations did not show any information about it. Talking about DisGeNET, it was necessary the interpretation of all the bibliographic references

in order to extrapolate the clinical significance of each variation (pathogenic, benign or protective).

Another important aspect related to the identification process was the presence of contradictory information. Variations which show up a determined clinical significance in a specific database could introduce the opposite result in a different repository. An example of that is variation rs140501902, pathogenic for AD in DisGeNET but in Alzforum Mutations it appears as not pathogenic. Data discrepancies implies in our method an automatic discard due to the doubts in the veracity of the result, although it is important to remark that the dynamic property of the selected criteria makes a research option to select variation fulfilling or not fulfilling any particular criteria: each individual SILE execution generates a result that is fully dependent on the selected criteria.

Finally, it is worth to emphasize the absence of fully contrasted information about AD. A great quantity of data existed in the diverse repositories which, a priori, could suggest that the study will be very complete. Nevertheless, when quality filters were applied, results decreased rapidly and in a significant way, what shows that a very relevant part of the data was associated with not valid information for the exploitation process.

7. Conclusions and Future Work

In the scope of the modern, genomic-based Medicine of Precision, an effective and efficient information systems development strategy is essential. This paper shows how to face this problem for a concrete clinical domain –the Alzheimer's Disease-, by using a method conceptual model-centered to characterize and structure how relevant data sources must be selected, how to obtain from them a set of candidate variations that are significant in clinical terms and how to filter them in order to refine the set to obtain the required set of valid variations.

The role of the conceptual model is to precise the information that must be captured, defining the mappings between the attributes of the model and their corresponding genomic data source counterpart. The automation of this essential part of the process is now under study, with the goal of designing a software platform based on a database whose data could be updated automatically by developing the corresponding access mechanisms with the data sources.

How the method was applied for the AD case is reported in the work. One of the most relevant results is the selected set of variants, that have been loaded in a data-base that is used by a genomic diagnosis software platform (GenesLove.Me [12]) that can now include in its catalog of clinical services the diagnosis of this disease.

As the application of the SILE method in different diseases can have different properties, it is our next step to apply the method to other disease in order to compare the obtained results, analyzing commonalities and differences of the process, and adapting the quality criteria used by the method to the particular information that is available for the case under study. The final attempt is to propose a holistic, universal method that could be applied to manage correctly Genome Information Systems for the challenging Medicine of Precision.

Acknowledgements

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

References

1. Rigden, D. J., Fernández, X.: The 2018 Nucleic Acids Research Database Issue and the Online Molecular Biology Database Collection. In Nucleic Acids Research 46. Database issue, D1–D7. PMC (2018)
2. Medline, https://medlineplus.gov/spanish/alzheimersdisease.html. Accessed September 14, 2017
3. Dementia stats by WHO, http://www.who.int/mediacentre/factsheets/fs362/es/. Accessed September 15, 2017
4. Waring, SC., Rosenberg, RN.: Genome-wide association studies in Alzheimer disease. In Arch Neurol, vol 65(3), pp. 329–334 (2008) doi:10.1001/archneur.65.3.329
5. APP data, https://www.ncbi.nlm.nih.gov/gene/351. Accessed September, 20 (2017)
6. DisGeNET - a database of gene-disease associations, http://www.disgenet.org/web/DisGeNET/menu;jsessionid=1iqtm0exl63vbmkkluw07jeea. Accessed December, 3 (2017)
7. Introduction - ClinVar – NCBI, https://www.ncbi.nlm.nih.gov/clinvar/intro/. Accessed December,3 (2017)
8. Ensembl genome browser 91, http://www.ensembl.org/index.html. Accessed December, 3 (2017)
9. Mutations | ALZFORUM, http://www.alzforum.org/mutations. Accessed December, 3 (2017)
10. Reyes Román, J. F., Pastor, Ó., Casamayor, J. C., Valverde, F.: Applying Conceptual Modeling to Better Understand the Human Genome. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Berlin Heidelberg, pp. 404–412, 2016
11. León, A., Reyes, J.F.R, Burriel, V., Valverde, F.: Data Quality Problems When Integrating Genomic Information. In Advances in Conceptual Modeling. ER 2016. Lecture Notes in Computer Science, Springer, Champ, pp. 173-182, 2016
12. Román, J.F.R., Iñiguez-Jarrín, C., López, O.P.: GenesLove.Me: A model-basedweb-Application for direct-To-consumer genetic tests. In ENASE 2017 - Proceedings of the 12th International Conference on Evaluation of Novel Approaches to Software Engineering, pp. 133–143 (2017).