STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LXII**, Number 1, 2017
DOI: 10.24193/subbi.2017.1.06

A BIG DATA APPROACH IN MUTATION ANALYSIS AND PREDICTION

SILVANA ALBERT

Abstract. Although the technology advancement in the last few years has been exponentially growing, there are still a lot of medical problems that don't have an accessible solution. One of these problems is the one that genetics is facing: the absence of a solution for inspecting the previously reported genetic mutations. In order to confirm a mutation, the specialists need to narrow it down based on their experience and, if present, the few documented precedent cases. This paper focuses on presenting a solution for analyzing big amounts of historical genetic data in an efficient, fast and user-friendly way. As a proof of concept, it demonstrates the huge role that Big Data has in genetic mutations aggregation and it can be considered a starting point for similar solutions that aim to continuously innovate genetics. The effectiveness of our proposal is highlighted by comparing it with similar existing solutions.

1. Introduction

The volume of aggregated medical information has increased exponentially in the last few years and it will keep increasing. DNA Sequencing is just a few years away from becoming affordable. When that happens, the technology has to be prepared to analyze it, extract patterns, prevent anomalies and provide accurate predictions, which will rely heavily on using big data [21]. In our opinion, the genetic advancement should go hand in hand with the technological advancements in order to exploit the full potential of both branches.

The need for interdisciplinary collaboration between the genetic specialists and software engineers has been recently identified and it is proved to be effective [22]. This paper is the result of a brief collaboration that started by putting the needs of the medical community on the first place and the concern of how to develop the solution on second.

Received by the editors: April 24, 2017.

2010 *Mathematics Subject Classification.* 68N01, 68T05.

1998 *CR Categories and Descriptors.* D.2.11 [**Software**]: Software engineering – *Software Architectures*; I.2.6[**Computing Methodologies**]: Artificial Intelligence – *Learning*.

Key words and phrases. Big data, genetics, software, machine learning.

Genetic mutation aggregation refers to the process of gathering any relevant information about genetic mutations and storing it for future use in analysis and visualization.

The contribution of our paper is twofold and is summarized in the following. First, we are proposing a solution for analyzing big amounts of historical genetic data in a very efficient and fast way, using a big data approach. The proposed solution demonstrates the huge role that Big Data [6, 14] has in genetic mutations aggregation and it can be considered a starting point for similar solutions that aim to continuously innovate genetics [5]. Our second aim is to highlight the potential of using supervised *machine learning* [16] models in predicting future genetic mutations. The overall purpose of the paper is to provide demographics and metrics regarding prophylaxis, and diagnosis of different genetic disorders and to offer a solution that allows the medical personnel to access a comprehensive history of patients screened/diagnosed with certain chromosome anomalies or gene mutations.

The rest of the paper is structured as follows. Section 2 presents the fundamental background concepts related to genetics, as well as the applicability of Big Data in mutation analysis. Our approach in mutation analysis and prediction using a Big Data approach is introduced in Section 3. Section 3.2 details the prediction component of the proposed solution and provides several experimental results. An analysis of our proposal and comparison with existing similar work is given in Section 4. Section 5 presents the conclusions of our paper and outlines directions for further improvement and extension.

2. Background

We are presenting in the following section the main concepts involved in our approach.

2.1. **Genetic background.**

DNA, short from deoxyribonucleic acid is a molecule present in all living things that contains instructions needed by organisms in order to develop and reproduce. RNA, abbreviated from Ribonucleic acid is a molecule which plays an important role in creating proteins from DNA.

A *mutation* is a natural process that occurs when a cell copies the DNA before dividing and that changes the DNA sequence.

Mutations are unavoidable; most of them arise along with the natural process of DNA transcription and therefore corrected by efficient DNA repair mechanisms. Usually, mutations are perceived as something bad that happened or that something got broken but there are a lot of cases when it has no impact (the changes are in the areas of the genome that is between the genes).

Nucleotides are structural components of the DNA and RNA. There are approximately 3.000.000.000 nucleotides in a human genome. When a cell

divides, it is supposed to make a copy of its own DNA but sometimes, something bad happens and the result is a similar sequence that has one different nucleotide. That small difference is called a mutation and it is depicted in Figure 1.

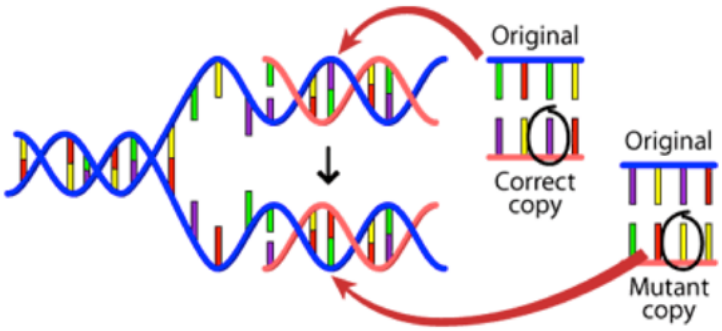

FIGURE 1. Mutation genesis. Figure source: Understanding Evolution [3].

Mutations can arise also from the interaction with external factors. If the person was exposed to certain chemicals or radiation, the chances of a mutation occurring are higher. The agents are breaking the DNA and when the cells try to repair it, some differences can happen.

Gene data is stored in lots of different ways depending on the database. Three of the existing disease-related variation databases are described in section 4. Usually, mutations are stored by gene symbol with very detailed information about the place in the DNA sequence where it happened, all kinds of locus coordinates and all kinds of classifications based on the mutation type (deletion, insertion, splicing, etc.).

At the time of this research, we are not aware of a system that stores both gene data and information about the demographics of its origin because this data is usually confidential and anonymous due to its sensitive nature. As a result of sequencing, data can be stored in standard recognized formats. Some of the most common ones are: Plain sequence format (containing one or more sequences with no extra information), FASTQ (which stores both biological sequence and quality scores and is produced by advanced sequencing instruments) and EMBL (contains an id per sequence and other relevant annotation lines before and after the sequence) [18].

2.2. **Big data and NoSQL.** *Big data* [20, 6] is a notion for storing large collections of data sets and further analyzing, visualizing and transferring them. Collecting data from all kinds of devices leads to storing a lot of data but what is truly impressive about it is not the quantity of information but what we can do with it. From a few terabytes of collected data that we had stored in 2012, we got to entire petabytes now and it is still growing.

Big data can be described as large pools of data that is captured and aggregated with advantages that lead to increasing modern economics, health care, transportation and many other industries.

NoSQL [10] means Not Only SQL, implying that when designing a software solution or product, there is more than one storage mechanism that could be used based on the needs [19]. There are a lot of NoSQL database management systems (at this moment, there are approximately 150) and they can be classified based on their data model [19]: *Key-value* (Dynamo, Riak), *Graph* (Allegro, Infinite Graph), *Multi-model* (OrientDB, FoundationDB), *Document* (MongoDB, Couchbase) and *Column* (Accumulo, Cassandra).

One main characteristic of NoSQL databases is that it is schema agnostic; this means that there is no need for an upfront schema design for allowing data storage. Another important aspect is strong consistency that can be translated in: all the clients should see the same version of data [19]. The last characteristic that NoSQL databases need to have is partition tolerance: the complete system should keep its properties even when deployed on separate servers [3].

In the context of genetic mutations, the information related to a gene and all it's possible anomalies that need to be stored in order to perform a relevant analysis is what constitutes Big Data. Each mutation entry has a series of characteristics that will be described in Section 3.3 and the number of characteristics changes constantly as the science advances. It is important to be able to save different characteristics and not be constrained by an existing rigid schema and that's why NoSQL and Genomics go hand in hand.

3. Our approach

The general idea behind our proposal is to offer a solution that provides demographics and metrics about diagnostics and mutations. It started with the idea of creating a solution that allows the medical personnel to browse through a comprehensive history of patients screened/diagnosed with certain chromosome anomalies or gene mutations.

Visualizing the number of mutations by countries needs a complete database parsing and for 1 million results, it takes a couple of minutes. One adjustment that could make this interrogation faster is splitting the work into a number of threads equal to the number of countries and each thread would count the

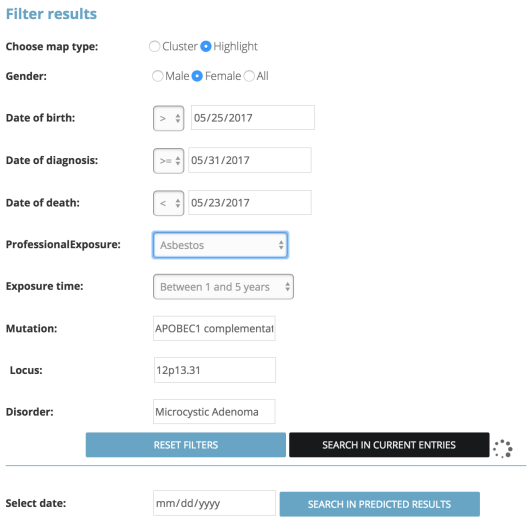

FIGURE 2. Screen shot from the application with possible filtering options.

number of mutations occurred in the given country. In the proposed solution, inserting one million entries without optimizing performance with threads, takes less than 12 minutes. Further optimization can be done and the time can be decreased by using multiple servers and batch inserting using threads. There is some bench-marking performed by Netflix that claims to have inserted 1.1 million client writes per second using Cassandra [4].

The novelty of our solution is that it stores and aggregates genetic, demographic and geographic data. The solution presented in this section is a proof of concept and the stored data is mock data based on real genes and mutations obtained by scraping existing databases.

From the visualization perspective, the originality factor is that based on the multiple filters from Figure 2, the number of people that match are displayed on the world map and the intensity of the color reflects the number of entries per country as seen in Figure 6. The solution paints a very graphic picture of the current situation on the globe, while other solutions are just listing individual mutations in tabular views. The other approach requires a lot of time, attention and work to scroll and comprehend the displayed data; However, if that approach is still needed for reports, our solution also allows exporting the data to Excel files.

3.1. **The proposed solution.** Our solution makes the following scenario possible:

"*As a doctor, I want to see the count of women that were diagnosed at the age of 25 with Breast Cancer, have the mutation KRAS, are currently under treatment and were professionally exposed to chemical agent benzyl.*"

The possible capabilities from the doctors perspective are represented in the use case diagram from Figure 3.

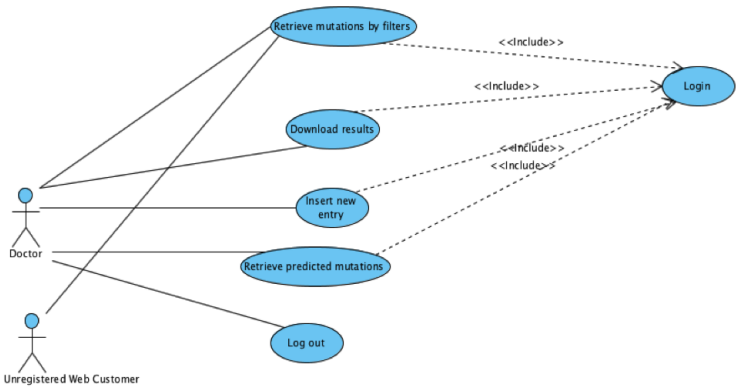

FIGURE 3. Use Case Diagram of the proposed solution.

From the technical perspective, the proposed architecture contains four main components:

(1) Creating a database that stores information about patients and their found mutations and diseases.
(2) Creating a solution for interrogating that database with regard of performance and scalability.
(3) The prediction engine that can help the doctors gain a better overview of the expansion of a mutation in a given period of time.
(4) Once the results are retrieved, they are displayed in a user friendly web interface in a way that means something for the user (position mutations geographically with the possibility of zooming in and further filtration based on gender, age, environmental conditions and so on) instead of simply listing the results in a table.

The component diagram of the proposed solution is depicted in Figure 4.

3.2. **The component for predicting future mutations.** One of the main components of the proposed solution is the one for future mutations prediction.

From a *machine learning* perspective, *predictive modelling* refers to analyzing historical information to make predictions about future [7]. *Machine*

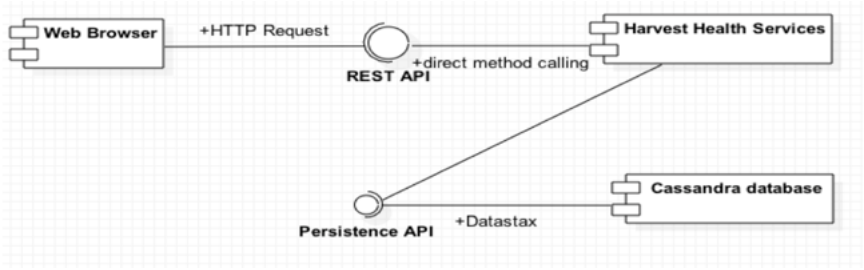

FIGURE 4. Component diagram of the proposed solution.

Learning (ML) [16] is a challenging field of *Artificial Intelligence* in which the focus is to develop adaptive computer systems, able to improve their *performance* from experience and through learning some specific domain knowledge.

Within the machine learning domain, a major emphasis is on *supervised learning*. The systems which learn from an external supervisor are connected to *predictive modelling*. The *predictive models* are able to make predictions based on some training data (i.e. historical data). In supervised learning, the learner is provided with a set of labeled examples (inputs with their known outputs) and then it will be able to generalize from the received examples and to predict the output when faced with an input instance unseen during training. The chosen software for predicting mutations is Weka [9]. The prediction engine workflow starts with the query executed on Cassandra [12] based on the users filters. The results from the database are parsed and stored into a csv file.

The *Training Model* contains 6 fields: the *country code*, the *count of mutations* (retrieved based on the country and exposure factor), the *date of diagnosis* (when the mutation was discovered and entered into the system), the *gender* of the patient, the *exposure time*(in years because it is split in intervals: less than 1 year, between 1 and 5 years, between 5 and 10, between 10 and 20, and over 20) and the most important field: the *professional exposure factor* (the chemical element to which the person was exposed).

Examples of exposure factors: Arsenic, Asbestos, Asphalt fumes, Benzene, Beryllium, 1-Bromopropane and many more.

The *Prediction Service* receives as training data the current aggregated entries from the database. The prediction flow is depicted in Figure 5.

Counts are computed for each country based on the exposure factor and other search criteria and saved as training data in an .arff file. That file is read and classified and based on it, the predicted counts are computed, then converted to JSON and finally passed on as a response to the Prediction

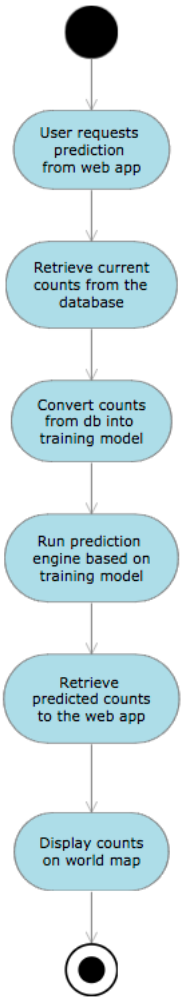

FIGURE 5. Activity diagram of the prediction flow.

Endpoint. This returned JSON will be handled by the UI similar to any other result set from Cassandra and will display the results on the World Map. As an example, Figure 6 illustrates the World map of mutation frequency by professional exposure Arsenic.

The counts from the database for professional exposure to Arsenic for less than one year are: Canada: 18, China: 1509, France: 1354, Norway: 18, Niger: 975, New Zeeland: 413 and so on.

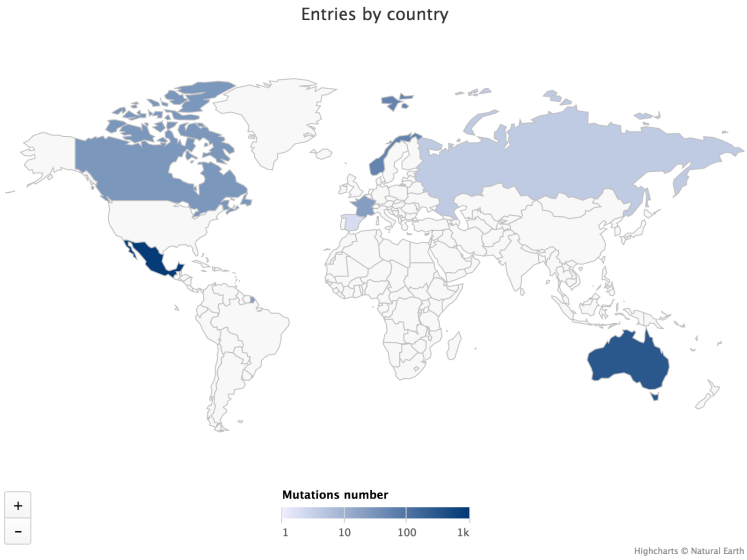

FIGURE 6. World map of mutation frequency by professional exposure Arsenic.

The results vary based on existing counts and number of years we perform the prediction. When changed to 100 years from now, the world map looks significantly different (because the current date and prediction dates are factored into the generated training data)

The predicted counts in 10 years for professional exposure to Arsenic for less than one year are depicted in Figure 7: Canada: 28, Mexico: 875, France: 22, Norway: 56, Spain: 2, Russia: 4 and Australia: 300.

3.3. **Implementation details.** *Apache Cassandra* was used for this proof of concept because of its scalability and reliability. The reading speed is more important than the insert because that is the main use case of the proposed solution: analyzing existing data.

3.3.1. *Persistence model.* The selected model for storage contains the following data: *Name*, *Identification number*, *Gender*, *Country*, *Date of birth*, *Professional exposure* (if present, with possible categories and time of exposure), *Age at diagnosis*, *Date of death*, *Submitted by* (name of the doctor), *Details*, *Disorder 1* (Mutation 1_locus22, Mutation 2_locus30 ... Mutation n_locusxy), *dots Disorder m* (Mutation 1_locus15, Mutation 2_locus13 ... Mutation p_locuskz).

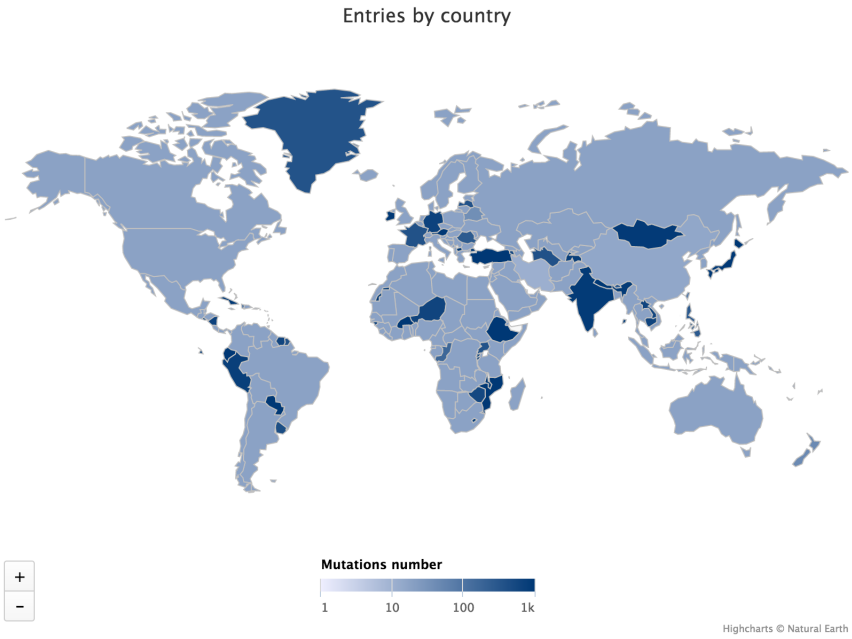

FIGURE 7. World map of mutation frequency by professional exposure Arsenic in 10 years.

It will allow extracting information about frequency of mutation in different regions with different characteristics. The number of stored rows can be dynamic for each entry and this is achieved by using a non-relational database.

Having a different number of columns for each entry is important for building a system for quickly previewing mutations. That is because a patient can have 1 disease with only one mutation while other can have 9 diseases with 1000 mutations.

The primary key contains the identification number of the person, the country identifier and the mutation entries.

A mutation entry has the following form:

$$A2M\_12p13.31\_AlzheimersDisease\_AACS_{1}2q24.31\_Traheal\ Cancer$$

This means that on the gene $A2M$ (which is stored with the complete name in a separate table), on locus $12p13.31$, there is a mutation that causes the disorder: *Alzheimers disease*. But this patient has another mutation entry that is separated by comma. So the mutations column contains a string that respects the previously defined format. This way, complex information about a

patient can be stored on a single row. The same logic applies to the professional exposure field also; this is how a professional exposure entry looks like:

$$Asphaltfumes\_34000$$

Unlike mutations, this field is optional for most patients. The first part of the entry is the name of the substance to which the patient was exposed while the former represents the duration of exposure in milliseconds (the patient worked in a contaminated facility for 3 years).

The database was created with replication $=\{'class' :' SimpleStrategy', 'replication\_factor' : 3\}$. This means that the used strategy is enough for evaluation purposes. The alternative is NetworkedTypologyStrategy which needs to be used when multiple data centers are connected. The replication factor describes the number of replicas of data on multiple nodes. It has to be specified only when using the simple strategy. The personal identification number, the mutation string and the country code compose the primary key. Those are the most important properties that define an entry and are most frequently used in searches. For searches on other criteria, indexes are created. Although indexes are not as performant as searching only for what is contained in the primary key, they were created in order to give the searches flexibility because not all the parameters are filled in for every search query.

4. Discussion and comparison to related work

The solution proposed in this paper facilitates the analysis of genetic data using a big data approach.

As a proof of concept, our proposal demonstrates the huge role that Big Data has in genetic mutations aggregation and it can be considered a starting point for similar solutions that aim to continuously innovate genetics. The presented solution allows the doctors to filter the mutations, visually inspect their frequency on the world map, predict future mutations, insert new entries and export data in various formats. It helps by aggregating all the precedent mutations correlated with a series of external factors. The doctor is able to narrow it down to a reasonable number of possibilities based on the cases that were already solved. This leads to making an informed decision of which mutations to test for. After successfully determining the current case, the specialist will introduce it to the global database, this way, helping future doctors.

We chose the NoSQL approach for implementing our prototype based on existing literature and also the following reasoning.

From a relational database perspective, this same issue could have been resolved by having a table with all genes, another with all possible mutations

by gene and a third one with all the disorders linked to multiple mutations in various genes. When trying to filter by any of the stored information, multiple joins would need to happen.

Our intuition is that performing multiple joins would take more time than the proposed solution but no concrete experiments were performed. Back to our example, the patient with 1000 mutations would be inserted into the database 1000 times having each mutation id as a foreign key or inserted once and store in column or mutation ids separated by commas. Either way, we assume performance would suffer because of the necessary joins that would need to happen.

By using a non relational database instead, the number of columns can be dynamic and it does not matter which of the stored information will be used as a filter. Using the alternative relational approach, if we want the counts of people that have certain disorders would mean joining with mutations to the the different disorder ids and then with disorders to get the names. Filtering by disorders would take more time than filtering by mutation code because the latter means a single join operation while the first means two. By using dynamic columns, it doesn't matter if the entry is searched by mutation or disorder because they are both columns on each row and it takes the same amount of time to retrieve.

We are describing in the following existing solutions for mutation analysis and prediction, comparing them with our proposal.

4.1. **Cosmic-Catalogue of somatic mutations in cancer** [15]**.** *Cosmic* [2] is a tool that allows inspecting various mutations and their frequency. It contains data from The Cancer Genome Atlas and the International Cancer Genome Consortium portals but also, data can submitted directly through their website. It is a comprehensive database that contains so far 20,981 mutations and details about each. Cosmic provides statistics about mutations but it does not contain demographics. There is no way of accessing any information about the people that have these mutations. The main capability that our solution provides and Cosmic doesn't is the visualization of incidence of mutations based on geographic location.

4.2. **The Human Gene Mutation Database** [11]**.** The *Human Gene Mutation Database* (HGMD) [17] is a database at the Institute of Medical Genetics in Cardiff, from BIOBASE that contains over 152.000 mutations. It is a comprehensive data on human inherited disease mutations to genetics and genomic research.

HGMD [11] provides all kinds of features for analyzing mutations but it does not have the capability of linking the characteristics of the person with the discovered genetic characteristics. However, it allows inspecting various

aspects of a mutation from the strictly technical point of view. It also analyzes candidate genes for finding disease linkage and predisposition.

However, there is no apparent correlation between exposure to exposure factors and the subsequent mutations, like in the solution proposed in this paper.

4.3. **Orphanet** [8]**.** *Orphanet* [8] is the solution that doctors in Romania currently use. It is a European website that has the office location in Paris and its main focus is providing content regarding rare diseases and orphan drugs [1]. Orphanet was funded by Inserm (the French National Institute of Health and Medical Research), the French Directorate General for Health and the European Commission. It contains a database that is an encyclopedia of rare diseases and it encourages the collaboration between research teams. It is basically a search engine that provides raw information that is updated annually. It also provides details about ongoing trials [13].

Although it is vastly used by Romanian specialists, Orphanet doesn't provide statistics about the place where the mutations and rare conditions occurred and in which circumstances (age, professional exposure, gender and so on). There is no way to determine the likeliness of the same mutation happening to the current patient.

Compared to the previously mentioned solutions, our approach has the already mentioned advantages: keeps the link between the occurred mutation and it's details to the person that developed it. That way, advanced correlation there can be made based on the factors that led to a mutation, a person's gender and age, diagnosis date, recovery rate and so one.

From the performance perspective, a direct comparison between the proposed solution and other gene databases is not possible because of the way other databases display the results: they are paginated and not aggregated; Users can retrieve a limited amount of data at the time (depending on the database).

Our solution's biggest contribution that none of the alternative solutions provide is the geographic clustering of mutations.

5. Conclusions and future work

We proposed in this paper a *big data* approach in mutation analysis and prediction. As a proof of concept, the presented solution demonstrates the huge role that big data has in genetic mutations aggregation and it can be considered a starting point for similar solutions that aim to continuously innovate genetics.

Future work may be done in order to enhance performance and scalability of the proposed system in order to increase the reading speed. Concrete experiments using a relational approach should be performed. The capabilities of the prediction engine could also be increased and this would lead to more accurate predictions. The basic linear prediction that is used now, can be enhanced to handle complex scenarios that take into consideration environmental factors that may lead to a mutation spreading.

Acknowledgments

The author thanks Dr. Cătană Andreea who contributed to the paper by describing the current methodologies used for genetic diagnosis and helped identifying the need for this analytical software. She also came up with the list of particularities that each entry in the database should have and provided feedback on the solutions main capabilities.

References

[1] S. Ayme and J. Schmidtke. Networking for rare diseases: a necessity for europe. *Bundesgesundheitsblatt*, 2007.

[2] S. Bamford, E. Dawson, S. Forbes, J. Clements., R. Pettett, A. Dogan, A. Flanagan, J. Teague, P.A. Futreal, M.R. Stratton, and R. Wooster. The cosmic (catalogue of somatic mutations in cancer) database and website. *Br. J. Cancer*, 91(2):355–8, July 2004.

[3] Berkeley University. Understanding Evolution - The causes of mutations. http://evolution.berkeley.edu/evolibrary/article/evo_20. Online; 2017.

[4] A. Cockcroft and D. Sheahan. The Netflix Technology Blog. https://medium.com/netflix-techblog/benchmarking-cassandra-scalability-on-aws-over-a-million-writes-per-second-39f45f066c9e. Online; 2011.

[5] B. Feldman, E.M. Martin, and T. Skotnes. Big data in healthcare hype and hope. Technical report, Dr. Bonnie 360, October 2012.

[6] L. Fernandes, M. O′Connor M., and V. Weaver. Big data, bigger outcomes. *AHIMA*, 83(10):38–43, 2012.

[7] S. Finlay. *Predictive Analytics, Data Mining and Big Data: Myths, Misconceptions and Methods*. Business in the Digital Economy. Palgrave Macmillan UK, 2014.

[8] French National Institute for Health and Medical Research. The portal for rare diseases and orphan drugs. http://www.orpha.net/consor/cgi-bin/index.php. Online; 2017.

[9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten. The WEKA data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November 2009.

[10] R. Hecht and S. Jablonski. NoSQL evaluation: A use case oriented survey. In *2011 International Conference on Cloud and Service Computing*, pages 336–341, Dec 2011.

[11] Institute of Medical Genetics in Cardiff. The Human Gene Mutation Database. http://www.hgmd.cf.ac.uk/ac/index.php. Online; 2017.

[12] A. Lakshman and P. Malik. Cassandra: A decentralized structured storage system. *SIGOPS Oper. Syst. Rev.*, 44(2):35–40, April 2010.

[13] S. Maiella, A. Rath, C. Angin, F. Mousson, and O. Kremp. [orphanet and its consortium: where to find expert-validated information on rare diseases]. *Revue neurologique*, 169(Suppl 1):S3–8, 2013.
[14] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A Byers-Hung. Big data: The next frontier for innovation, competition, and productivity. Technical report, McKinsey Global Institute, June 2011.
[15] Ministry for Primary Industries. COSMIC, the Catalogue Of Somatic Mutations In Cancer. http://cancer.sanger.ac.uk/cosmic. Online; v80, released 13-Feb-17.
[16] T. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997.
[17] P.D. Stenson, M. Mort, E.V. Ball, K. Evans, M. Hayden, S. Heywood, M. Hussain, A.D. Phillips, and D.N. Cooper. The human gene mutation database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Human Genetics*, pages 1–13, 2017.
[18] G. Stoesser, W. Baker, and A. Broek. The embl nucleotide sequence database. *Nucleic Acids Research*, 30:21–26, 2002.
[19] T. A. M. C. Thantriwatte and C. I. Keppetiyagama. NoSQL query processing system for wireless ad-hoc and sensor networks. In *2011 International Conference on Advances in ICT for Emerging Regions (ICTer)*, pages 78–82, Sept 2011.
[20] B. Wang, L. Ruowang, and W. Perrizo. *Big Data Analytics in Bioinformatics and Healthcare*. IGI Global, Hershey, PA, USA, 1st edition, 2014.
[21] R. Wullianallur and V. Raghupathi. Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*, 2(1):1–3, 2014.
[22] B. Zenger. Can big data solve healthcares big problems? *Health Byte*, 2012.

Department of Computer Science, Faculty of Mathematics and Computer Science, Babeş-Bolyai University, Cluj-Napoca, Romania
E-mail address: albert.silvana@cs.ubbcluj.ro