Diagen: A Model-driven Framework for Integrating Bioinformatic Tools

Maria José Villanueva, Francisco Valverde, Ana Levín, and Oscar Pastor

Centro de Investigación en Métodos de Producción de Software
Universitat Politècnica de València
Camino de Vera S/N 46022, Valencia, Spain
{mvillanueva, fvalverde, alevin, opastor}@pros.upv.es

Abstract. Nowadays, the diagnosis of disease based on genomic information is feasible by searching genetic variations on DNA sequences. However, geneticists struggle with bioinformatic tools that are supposed to simplify DNA sequence analysis. As a universal tool to support every requirement is far from be implemented, geneticists themselves must solve the data exchange among several tools. Due to the fact that there are no standards to support this integration task, it must be managed in every analysis. This paper proposes addressing this integration by means of a model-driven framework. The Diagen framework is a software implementation based on conceptual modeling principles that formalizes data exchange and simplifies bioinformatic tool integration. First, we analyze how conceptual modeling can be used to deal with data exchange among tools. And then, as a proof of concept, the presented framework is used to search for variations on the BRCA2 gene using real DNA samples and a set of specific bioinformatic tools.

Keywords: Model-Driven Development, Tool Integration, DNA sequence analysis

1 Introduction

Recent genetic discoveries have opened the door to personalized disease diagnosis based on DNA sequence analysis. Nowadays, it is possible to predict the risk of getting a certain disease by searching for specific genetic variations on the DNA sequence [1].

Geneticists perform DNA sequence analysis aided by bioinformatic tools. Even though these tools are functional and useful for reducing time and complexity, none of them completely fulfill all the geneticists' requirements [2]. As a consequence, geneticists are forced to use several tools in order to gather all the functionality and, eventually, accomplish the complete DNA sequence analysis.

One important issue regarding these tools is that data exchange among them is required. The problem lies in the fact that each of these tools is isolated and uses its own data format to report the computed information. For this reason, data exchange among tools is a non-trivial task that geneticists must address in each analysis according to the following procedure: 1) Export data from the source tool; 2) Understand the semantics of the tool-specific data format; 3) Perform a translation into the target tool format; and finally, 4) Import the data into the target tool.

As geneticists usually lack Software Engineering knowledge, most of them perform this task manually or develop programming scripts. Although these specific scripts are useful in solving minor problems, they are far from being compliant with good practices of Software Engineering. The implemented scripts to support data exchange are often coupled solutions that integrate only two specific tools. In the end, these solutions cannot be reused and compromise the geneticists flexibility for using other tools.

As a solution, this paper proposes the application of conceptual modeling to develop a model-driven framework that formalizes data exchange and simplifies tool integration. In order to provide a high quality solution, this work has been developed in the context of a collaboration with geneticists from the Genomic Medicine Institute (IMEGEN). As a proof of concept, the proposed framework integrates several tools that are used by IMEGEN geneticists in their daily routine to search for genetic variations using real DNA samples of the BRCA2 gene (a gene related to Breast Cancer).

The paper is organized as follows: Section 2 presents a brief summary of other proposed solutions to solve the tool integration problems in DNA sequence analysis. Section 3 explains the proposed model-driven framework for integrating bioinformatic tools. Section 4 presents how the framework is used for disease diagnosis support using samples of the gene BRCA2 and a set of bioinformatic tools. And finally, section 5 presents the conclusions and future work.

2 Related Work

Several works have attempt to overcome current DNA sequence analysis tool issues. These proposals follow two different approaches.

Several sequence file formats for expressing bioinformatic tools results have emerged. Examples of these formats are: 1) Variant calling formats, such as the Variant Call Format (VCF) proposed for the 1000 Genomes Project [3]; 2) Alignment results formats, such as the Sequence Alignment/Map Format (SAM) [4], which provides a compressed textual representation, and the Genome Variation Format (GVF) [5], which provides a textual format using the Sequence Ontology [6].

All these formats have been defined for the purpose of providing interoperability among different DNA sequence analysis tools. The implementation of decoupled data exchange mechanisms is feasible using any of the above examples as a standard format. However, their main drawbacks are the complexity of each textual format and the mandatory implementation of a low- level mechanism to extract the data. As a consequence, none of them have become a widely applied standard and are only used in the research context where they have been proposed.

Several bioinformatic development frameworks have also been implemented. Some examples of these frameworks are Biojava [7], BioPython [8], or BioPerl [9]. These frameworks provide an API that supports common functionality for DNA analysis tasks. Additionally, they provide several format conversion operations to transform file formats among different tools.

These frameworks have been defined to provide geneticists with the freedom to implement their personalized tools. However, the geneticists still have to worry about low-level programming details and integration issues.

3 An Integrative Framework for Bioinformatics

This work presents a model-driven framework for the integration of DNA sequence analysis tools and retrieval of genetic information. Diagen is classified as a model-driven framework because each of its components (classes, data entities, operations) is a projection of the Conceptual Schema of the Human Genome (CSHG) [10]. The CSGH is a conceptual model created with the collaboration of geneticists, where biological concepts related to the human genome have been precisely addressed and defined. The framework uses this conceptual model to support the following DNA sequence analysis tasks (Figure1):

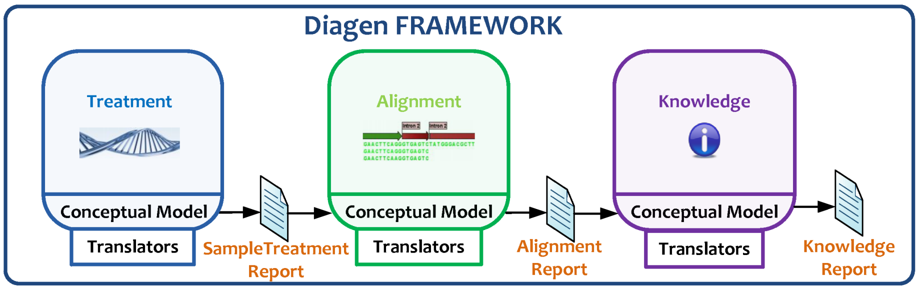

Fig. 1. General View of the Framework

1. Sequence Treatment: A DNA sequence is rebuilt from the fragments generated by the sequencing machines.
2. Sequence Alignment: A DNA sequence is aligned to a reference sequence in order to determine the differences between them.
3. Variation Knowledge: Using data gathered in genomic databases, each sequence difference that is related to a disease is reported.

Data exchange among tools is a difficult task because there is a great variety of formats to express the different results. Taking into account that data exchange is required when a tool calculates data that another tool requires, it can be assumed that both tools must share a set of common concepts. Therefore, it is possible to define a conceptual model that represents those shared concepts and establishes well-defined boundaries and vocabularies.

Diagen establishes the common context to guide data exchange among tools that define a conceptual model for each task transition:

- The Sample Treatment Report conceptual model (Figure 2) defines all the concepts related to the reconstructed sequence in the sequence treatment task (T1) to be analyzed in the sequence alignment task (T2).

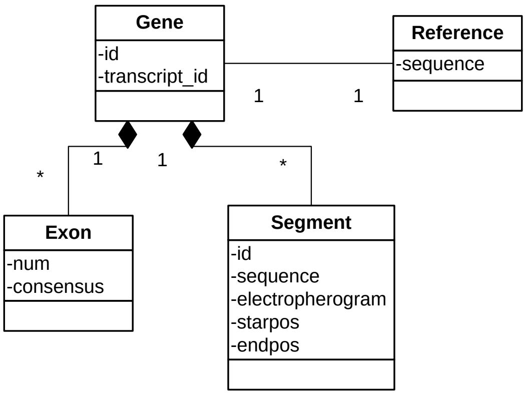

Fig. 2. Sample Treatment Report Conceptual Model

- The Alignment Report conceptual model (introduced in [11]) defines all the concepts related to the differences found in the sequence alignment task (T2) to be characterized in the variation knowledge task (T3).
- The Knowledge Report conceptual model (Figure 3) defines all the concepts related to the characterized variations to be used for other task (for example, a diagnosis report creation task).

Data exchange among tools that perform these tasks usually requires the implementation of a translation mechanism to understand each other. In that case, data expressed in a concrete format needs to be translated into a different format. However, the use Diagen avoids these coupled implementations because a tool to be integrated in the framework only needs a translator that expresses its outputs in terms of the underlying conceptual model. This translator is easier to implement since it only requires establishing the relationships between the output and the conceptual model.

Each task that is supported by the framework has been implemented to be independent from the others, and, therefore, it can be used separately. Thanks to this modularity, it is possible, for example, to use the alignment task in another environment. In this case, the input data should be provided in terms of the input conceptual model (Sample Treatment Report) and the output report should be read in terms of the output conceptual model (Alignment Report).

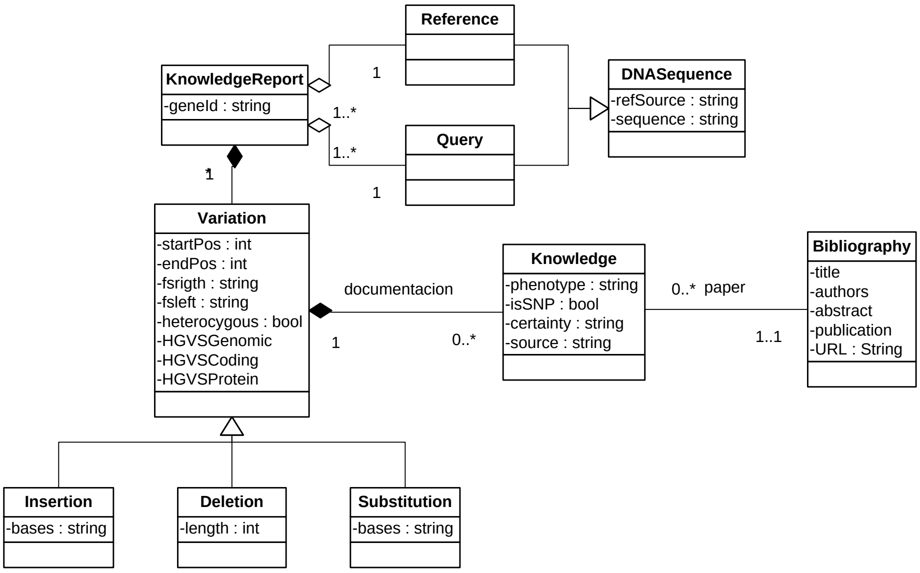

Fig. 3. Variation Knowledge Conceptual Model

The Diagen framework has been implemented using the Java language. Additionally, each conceptual model involved in data exchange has software correspondence with a set of Java classes and a XML representation. In order to manage both representations (Java and XML) JAXB (Java Architecture for XML bindings) [12] has been used. This is a specific API that allows Java objects to be parsed in a XML data and vice-versa.

4 Using Diagen for Disease Diagnosis Support of the BRCA2 Gene

As a proof of concept, the framework has been used to develop a prototype for disease diagnosis support of Breast Cancer. This specific framework configuration integrates several bioinformatic tools that are used daily by the geneticists of IMEGEN.

Recently, the framework (Figure 4) has been applied to integrate:

1. Sequence treatment task: The Sequencher tool [13] is used to rebuild the samples provided by a sequencing machine.
2. Sequence Alignment task: The implementation of the algorithm BLAST from NCBI [14] is used to search for differences in the sequence. There is also an integrated tool that is based on the Smith-Waterman Algorithm (SW Tool) and a tool that looks for known-variations in the sequence by aligning flanking sequences (Flanking Tool).

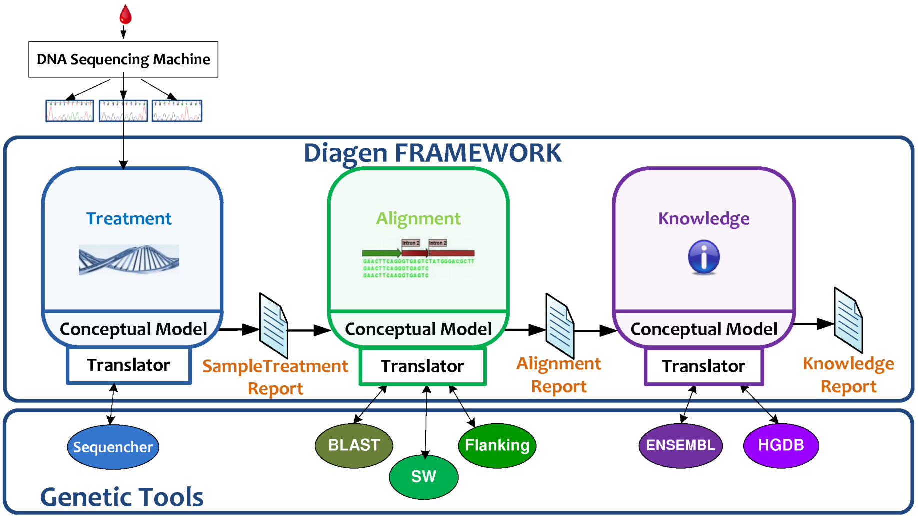

Fig. 4. IMEGEN configuration of the framework

3. Variation Knowledge task: Variation characterization is performed manually by geneticists searching in several databases. However, this framework provides two mechanisms for genetic knowledge data retrieval. The first mechanism obtains some data from the ENSEMBL database [15]. The second mechanism retrieves genetic information from the HGBD database [16] based on the Conceptual Schema of the Human Genome (CSHG) [10].

The prototype supports the three defined tasks needed to perform a DNA sequence analysis. As a result, it retrieves a personalized report containing the genetic variations and the potential diseases of the individual.

The main advantages of the framework are: 1) A decrease in the execution time, 2) A reduction in the efforts needed for data exchange among tools; and 3) The elimination of the need to search for variation data in the huge set of databases spread around the Web.

The prototype has been tested with real samples of the gene BRCA2 (Table 1). The test was carried out analyzing the BRCA2 gene sample from ten different patients (P1-P10). For each patient, the table shows the number of variations characterized by IMEGEN, the number of variations characterized by Diagen, and the accuracy that Diagen offers compared with the IMEGEN manual process. IMEGEN performs the analysis in approximately four hours (depending on the success achieved while searching for a difference in the genetic repositories).

The preliminary test showed that Diagen offers the results almost instantly and with an accuracy rate of between 60-90%. It is also important to emphasize that the variations that were not characterized by Diagen were always the same variations (7 variations in total) that appeared repeatedly in all the analyses.

Table 1. Preliminary BRCA2 tests

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Characterized Var. IMEGEN	7	10	8	8	8	13	9	10	9	8
Characterized Var. Diagen	6	6	7	6	5	8	6	7	6	5
Accuracy rate %	86	60	88	75	63	62	67	70	67	63

5 Conclusions and Future Work

This work proposes a model-driven framework that is based on a well-defined conceptual model of the human genome in order to address DNA sequence analysis. As a proof of concept, the Diagen framework is configured for the development of a disease diagnosis support and is tested by means of real DNA samples of the BRCA2 gene.

We have realized that the tools available actually accomplish some of the geneticists' goals. The problem lies in the fact that geneticists' activities, specifically in the DNA analysis domain, lack standard methodologies, well-defined tasks, fixed vocabularies, and unified knowledge sources. As a consequence, the execution of a DNA sequence analysis cannot be performed efficiently or without geneticists' intervention.

The solution to these problems is not to reinvent new DNA sequence analysis tools but to integrate the most suitable tools according to geneticists' needs. The presented framework applies conceptual modeling to integrate different bioinformatic tools and to provide a common context to exchange data with each other. The main advantage of the presented framework, over other integration approaches is that Diagen is a high-level abstraction framework that provides concise and significant tasks to geneticists instead of low-level tasks. Moreover, with this framework, geneticists can perform a DNA sequence analysis and forget about the data formats of different tools.

As genetics is a very innovative field that is constantly evolving with new discoveries, all concepts must be well-defined without ambiguity. Thanks to the conceptualization of the DNA sequence analysis tasks, all the involved concepts are precisely formalized. As a consequence, it is easier to adapt the tasks to changes or to support new concepts.

The preliminary results are promising, but there is room for improvement. The low accuracy detected is because the missed variations were not described in the integrated sources. As these sources are constantly improving, it is expected that future versions will solve these issues.

As future work, the framework will be extended to support other bioinformatic tasks. The main goal of this extension is to design a complete framework that supports other genetic functionality besides DNA sequence variation analysis. Additionally, the next step is to apply the service-oriented paradigm to provide a more flexible development environment. With this approach, geneticists could select only the required functionality, defined as services, and easily create a personalized tool.

Acknowledgments Thanks to the Instituto de Medicina Genómica (IMEGEN, http://www.imegen.es) for its support and the provision of data. This work has been developed with the support of MICINN under the FPU grant AP2010-1985 and the project PROS-Req (TIN2010-19130-C02-02), and co-financed with ERDF.

References

1. Margaret A. Hamburg and Francis S. Collins. The Path to Personalized Medicine. *New England Journal of Medicine*, vol. 363(4), pp. 301–304, (2010)
2. Nicole Rusk. Focus on Next-Generation Sequencing Data Analysis. *Nature Methods*, vol. 6(11s), pp. S1, (2009)
3. Siva Nayanah. 1000 Genomes Project. *Nat Biotech*, vol. 26(3), pp. 256–256, (2008)
4. Heng Li et al. The Sequence Alignment/Map Format and SAMtools. *Bioinformatics*, vol. 25(16), pp. 2078–2079, (2009)
5. Martin G. Reese et al. A Standard Variation File Format for Human Genome Sequences. *Genome biology*, vol. 11(8), pp. R88+, (2010)
6. Karen Eilbeck, Suzanna Lewis, Christopher Mungall, Mark Yandell, Lincoln Stein, Richard Durbin, and Michael Ashburner. The Sequence Ontology: A Tool for the Unification of Genome Annotations. *Genome Biology*, vol. 6(5), pp. R44, (2005)
7. R. C. G. Holland et al. BioJava: An Open-Source Framework for Bioinformatics. *Bioinformatics*, vol. 24(18), pp. 2096–2097, (2008)
8. Peter J. A. Cock et al. Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics. *Bioinformatics*, vol. 25(11), pp. 1422–1423, (2009)
9. Jason E. Stajich et al. The Bioperl Toolkit: Perl Modules for the Life Sciences. *Genome Research*, 12(10), pp. 1611–1618, 2002.
10. Oscar Pastor, Ana Levin, Matilde Celma, Juan Casamayor, Aremy Virrueta, and Luis Eraso. Model-Based Engineering Applied to the Interpretation of the Human Genome. In Roland Kaschek and Lois Delcambre, (eds.) *The Evolution of Conceptual Modeling*, LNCS, vol. 6520, pp. 306–330. Springer, Heidelberg (2011)
11. Maria Jose. Villanueva, Francisco. Valverde, and Oscar Pastor. Applying Conceptual Modeling to Alignment Tools: One Step towards the Automation of DNA Sequence Analysis. *BIOINFORMATICS 2011*, (2011)
12. E. Ort and B. Mehta. Java Architecture for XML Binding (JAXB). Technical Report Sun Developer Network, (2003)
13. P Curtis C Bromberg, H Cash and CJ Goebel. *Sequencher, Gene Codes Corporation.* Ann Arbor, Michigan, 1995.
14. NCBI BLAST (Basic Local Alignment Search Tool). Availble in http://blast.ncbi.nlm.nih.gov.
15. Hubbard, T. et al. The ENSEMBL Genome Database Project. *Nucleic Acids Research*, vol. 30(1), pp. 38–41, (2002)
16. Oscar Pastor et al. Enforcing Conceptual Modeling to Improve the Understanding of Human Genome. *Research challenges in information Science (RCIS)*, pp. 85-92, (2010)