MUTATIONAL DATA LOADING ROUTINES FOR HUMAN GENOME DATABASES
The BRCA1 Case

Matthijs van der Kroon, Ignacio Lereu Ramirez, Ana M. Levin, Óscar Pastor
Centro de Investigación en Métodos de Producción de Software, Universidad Politécnica de Valencia
Camino de Vera s/n, 46022 Valencia, Spain

Sjaak Brinkkemper
Department of Information and Computing Sciences Utrecht University, Utrecht, The Netherlands

Keywords: BRCA1, Conceptual model, Data integration, Human genome.

Abstract: The last decades a large amount of research has been done in the genomics domain which has and is generating terabytes, if not exabytes, of information stored globally in a very fragmented way. Different databases use different ways of storing the same data, resulting in undesired redundancy and restrained information transfer. Adding to this, keeping the existing databases consistent and data integrity maintained is mainly left to human intervention which in turn is very costly, both in time and money as well as error prone. Identifying a fixed conceptual dictionary in the form of a conceptual model thus seems crucial. This paper presents an effort to integrating the mutational data from the established genomic data source HGMD into a conceptual model driven database HGDB, thereby providing useful lessons to improve the already existing conceptual model of the human genome.

1 INTRODUCTION

Looking from an information system point of view, the human genome is an extremely complex system in which exists a lot of ambiguity. For example, basic concepts of what exactly defines a gene are still not explicitly described by the domain. Biology largely depends on domain experts interpreting data, in order for knowledge to appear. Combining the lack of proper data structure and the very large amounts of data generated, a clear problem emerges. How can domain experts dedicate their limited time to the right pieces of information if these are buried in noise? Computers excel at processing large amounts of data, and thus a logical step would be to apply this excellence to the present day problem in genetics, sifting the noise from potentially useful information. For this process to take place, a conceptual modeling approach is essential: it allows for an adequate representation of the domain. Present day solutions that pretend to do exactly this (i.e. ontologies) usually provide controlled vocabularies instead of fixing a conceptual gamut.

A proper conceptual model is expected to provide a clear data structure, enabling efficient and effective access to genomic data, thereby offering ways of reusing previously researched data by pharmaceutic, medical and research institutes as mentioned by (Pastor, 2008). Also, the paradigm shift implicated by considering the genome as a complex information system is expected to allow for exciting new views. To present day, most bioinformatics research is located in the solution space, by attempting to interpret the data that comes out of ’the black box’. For instance by applying powerful sequence alignment tools like BLAST and BLAT. Another point of view is offered by (Pastor, 2008), whose efforts are directed at tracing and understanding the processes effectively leading to these data. Essentially, seen from an informatics point of view, finding the source-code, by analyzing the object-code, of what may very well be the most sophisticated software ever to be analyzed: life itself.

In section 2 earlier work will be discussed. For a detailed description of the conceptual model of the Human Genome (CSHG) often referred to in this

van der Kroon M., Lereu Ramirez I., M. Levin A., Pastor Ó. and Brinkkemper S..
MUTATIONAL DATA LOADING ROUTINES FOR HUMAN GENOME DATABASES - The BRCA1 Case.
DOI: 10.5220/0003130902660269
In *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms* (BIOINFORMATICS-2011), pages 266-269
ISBN: 978-989-8425-36-2

work please consult (Pastor et al., 2010a) and (van der Kroon et al., 2009). Section 3 will discuss the results of the extraction of data from the external sources, listing the encountered problems and resulting adjustments to the conceptual model. Ultimately, in section 4 conclusions will be drawn, along with suggestions for further research.

2 RELATED WORK

Other solutions to the ambiguity problems associated to the genetics domain include ontologies (Ashburner et al., 2000). To understand why ontologies alone do not fulfill the job of obtaining a full understanding of any given domain, some background information is necessary. Conceptual definitions exist on two levels: conceptually and semantically. The semantic aspect refers to instances of concepts; e.g. the BRCA2 gene, which is an instance of the abstract ”*gene*” concept. A problem here for example means ambiguity about naming conventions, for instance the BRCA2 gene is also known as: ”Fancd1” and ”RAB163”. The conceptual aspect is more abstract and handles questions like ”What is a gene?”. It is our strong belief that for the proper and complete understanding of any given domain, both are vital.

An information systems approach to this specific problem space is not entirely new. (Okayama et al., 1998) describes the conceptual schema of a DNA database using an extended entity-relationship model. (Paton et al., 2000) advanced on these efforts by presenting a first attempt in conceptually modeling the *S. cerevisiae* genome by proposing a collection of conceptual data models for genomic data. Among these conceptual models are a basic schema diagram for genomic data, a protein-protein interaction model, a model for transcriptome data and a schema for allele modeling.

Whereas (Paton et al., 2000) provides a broader view by presenting conceptual models for describing both genome sequences and related functional data sets, (Pastor, 2008) converged on the basic schema diagram for genomic data adapting it to the human genome and eventually produced a database, the human genome database (HGDB) corresponding to this model and following the standard rules of logical design. This database is now in the prototype phase and the first 2 genes, NF1 and BRCA1, have been partially loaded. (Pastor et al., 2009) describes the evolution HGDB went through during the process of conceptually mapping HGDB and HGMD to each other. (Pastor, 2008) describes the evolution of the model more in general and provides a descriptive overview of how the model came to be, and from where it evolved to what it is now.

3 RESULTS

(Pastor et al., 2010b) reports a study of comparing the HGMD to the CSHG, in order to identify a conceptual mapping between the two. It is this mapping that is followed in this document, and the following section will report the encountered problems for actually loading the information from the HGMD into the HGDB for the BRCA1 gene. Roughly the problems can be separated in two categories; intrinsic data properties and data representation. Verifiably incorrect, inconsistent or incomplete data (tuples) are examples of these encountered mishaps with the actual data, or intrinsic data properties. Difficulties associated to the process of extracting the data from the external source and ambiguous descriptions of mutation properties are typical examples of data representation problems. Naturally, the division between the two categories is not strict and thus some overlap exists, it is however useful to keep in mind that intrinsic data property problems tend to affect the entire genetics domain, while the data representation difficulties are restricted to HGMD.

3.1 Data Loading Problems

HGMD distinguishes 10 mutation types: Missense/nonsense, Splicing, Regulatory, Small Deletions, Small Insertions, Small Indels, Gross Deletions, Gross Insertions, Complex Rearrangements and Repeat Variations. Roughly all the types can be mapped to the *Variation* and *Precise* concepts of the CSHG, except for the Gross Deletions, Gross Insertions, Complex Rearrangements and Repeat Variations. These latter are described in a very unstructured manner, almost natural language, and are thus considered impossible to process automatically. The CSHG facilitates these tuples as *Imprecise*, which merely stores a description of the mutation.

3.1.1 Intrinsic Data Properties

In some cases the HGMD mutational data lacks entries. For instance, the splicing mutations overview provided by HGMD mentions 5 mutations in intron 22, while (Panguluri et al., 1999) states at least 2 other mutations; IVS22+67(T>C) and IVS22+8 (T>A). Three concrete examples of this problem were encountered, all three in Splicing mutations. However, this particular type of problem is very difficult to detect, since finding them involves rereading the articles HGMD provides which is hard to automate. Thus, although only three concrete occurrences of this problem have been encountered, it is likely more exist.

Splicing mutation CS961492 describes a C>T mutation, as a possible phenotype HGMD indicates Breast cancer. However, having the read the corresponding article (Langston et al., 1996), not once breast cancer is mentioned in combination with this mutation. The article does mention the mutation as being affiliated with men suffering from prostate cancer. Thus, deducing from the rather limited information made available by HGMD on this specific mutation, it is concluded HGMD made an error during data entry.

Splicing mutations CS063247 and CS011027 should be located near intron 4. According to the splice junctions overview HGMD provides, there exists no intron 4, nor an exon 4. However, literature explains the ambiguity as a result of misidentification of an inserted Alu element (Smith et al., 1996).

3.1.2 Data Representation

Some data is provided in natural language. For instance the fact that the first two BRCA1 exons are alternative non-coding exons is only mentioned in the header of the Splice Junctions overview. Adding to this, in Small Deletions (2 instances) and in Small Insertions (3 instances) some mutations are located through mouse-over tags, the information communicated by these tags is highly unstructured to a degree that we might call it natural language as well. Also, in the case of imprecise mutations (Gross Deletions, Gross Insertions, Complex Rearrangements and Repeat Variations), the greater part of the information presented by HGMD is in natural language, impeding an automated approach severely in the affected cases.

In some cases, the HGMD database uses different ways of locating mutations, within the same type of mutations. For instance, Small Insertion mutations CI030168, CI962219 and CI022582 happen in non-coding areas of the gene, just like the Small Deletions mutations CD991644 and CD994433. Since HGMD generally uses a cDNA codon referenced way of locating these types of mutations, and given that non-coding sequences simply not exist in the cDNA, HGMD locates these earlier mentioned mutations in a different way. In the case of Small Insertions, HGMD provides a Splice Junction reference, very much like the method used to locate Splicing mutations. In this case the CI030168, CI962219 and CI022582 mutations are located at IVS20+21, IVS20+48 and IVS20+64 respectively. So IVS20 indicates the intron number, where +21 indicates the offset, however since no acceptor/ donor information is provided, it is unclear from which side of the intron the offset should be referenced. In the case of Small Deletion mutations CD991644 and CD994433 at first sight, no indication of how to locate them is provided. However, this information is provided through mouse-over tags in the Splice Junctions referenced form, described earlier. CD991644 is thus located by ”I7E8-24, aka IVS7 -15 del10” and CD994433 is located by ”I12+34 / polymorphism ?”. This problem was encountered 3 times in Small Insertions and 2 times in Small Deletions, making a total of 5 occurrences.

In splicing mutations, HGMD uses a different way of locating mutations. Here mutations are located by referring to splice-junctions. An offset is given, to indicate the amount of nucleotides between the indicated splice junction and the actual mutation. In a so-called splicing mutations overview HGMD then provides a sample sequence for each intron/exon-junction contained in the gene. This method of locating mutations is used primarily in splicing mutations (80 instances), but in some exceptional cases HGMD also uses this notation to provide locational data for other types of mutations. For instance, In Small Deletions (2 instances) and in Small Insertions (3 instances).

In the HGMD phenotype ambiguity exist, i.e. mutations may or may not result in a certain phenotype. This is indicated by a question mark following the supposed phenotype. However, no probability scores are stated and a mutation without a (noticeable) phenotype is considered to be a variation with neutral effect. Since variations and mutations are considered to be two different concepts in the conceptual model of the human genome, this poses problems with loading the database correctly. 94 instances of this problem have been identified: missense/nonsense mutations account for the most instances (73), splicing mutations contains another 16, small deletion mutations 2 and small insertion mutations account for 3 instances.

4 CONCLUSIONS

In this document we have confirmed the primary reason of existence for conceptual modeling techniques. The HGMD is considered an extremely useful source of data about genetic mutations in the field. For being curated, it is also considered to be highly reliable. However, this document shows that a lot remains to be wished for. The apparent lack of a thorough conceptual modeling approach seems to bear it’s traces on the service. Every tuple in the HGMD is supposed to represent a genetic variation, known to be associated to disease. This quite rigorous definition becomes endangered in cases where indicated variations ’might’ be associated to disease, as indicated in the HGMD by the question mark. Indeed, a variation that is not associated to disease should not be considered a mutation and thus not enter the dataset as is. The CSHG handles these cases nicely by providing the *neutral polymorphism* dimension, for the *Variation* concept. Another point of improvement is the lack of a proper way of facilitating the various reference sequence in common use by research papers. For illustration, a certain mutation might be located in position 131 in reference sequence X, but correspond to position 125 in reference sequence Y. The HGMD provides it’s own cDNA sequence, from which it locates the majority of it’s mutations. However this cDNA sequence is ’based’ on an NCBI sequence, and can thus differ from it.

For an optimal use of the data provided by HGMD, the above means an expert in many cases still needs to evaluate and interpret the data. This is expensive in both time and money. Aligning the HGMD set of mutations to the NCBI reference sequence, that is considered to be the ’golden standard’ thus seems a logic step. Concretely, we suggest two major changes to the HGMD: (i) facilitate a more elaborate way of handling associated phenotype, perhaps link directly to the Online Mendelian Inheritance in Man (OMIM) database. And (ii) add a new column, in which the reference sequence indicated by the source paper is also stored. This will allow for a much easier, and more efficient use of the HGMD data set. Considering data is acquired manually from the papers, adding this element of extracted data seems to be relatively low cost.

When we look at the HGMD we can not help but notice that although very useful, a lot is still to be wished for from an information systems point of view. It is our strong belief that the only way of accurately representing any data, and perhaps genetic data in particular, can only be done by means of careful analysis of the domain. The CSHG aims to do exactly this, by applying a conceptual modeling approach.

REFERENCES

Ashburner, M., Ball, C., and Blake, J. (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–30.

Langston, A., Stanford, J., Wicklund, K., Thompson, J., Blazej, R., and Ostrander, E. (1996). Germ-line brca1 mutations in selected men with prostate cancer. *American Journal of Human Genetics*, 58:881–885.

Okayama, T., Tamura, T., Gojobori, T., Tateno, Y., Ikeo, K., Miyazaki, S., Fukami-Kobayashi, K., and Sugawara, H. (1998). Formal design and implementation of an improved ddbj dna database with a new schema and object-oriented library. *Bioinformatics*, 14(6):472.

Panguluri, R., Dunston, G., Brody, L., Modali, R., Utley, K., Adams-Campbell, L., Day, A., and Whitfield-Broome, C. (1999). Brca1 mutations in african americans. *Human Genetics*, 105(1-2):28–31.

Pastor, O. (2008). Conceptual modeling meets the human genome. In *Conceptual modeling - ER 2008*, volume 5231 of *Lecture Notes in Computer Science*, pages 1–11. Springer-Verlag Berling Heidelberg.

Pastor, O., Levin, A., Casamayor, J., Celma, M., Virrueta, A., and Eraso, L. (2009). *The Evolution of Conceptual Modeling*, chapter Model driven-based engineering applied to the interpretation of the human genome. Springer-Verlag.

Pastor, O., Levin, A., Celma, M., Casamayor, J., Schattka, L. E., Villanueva, M., and Perez-Alonso, M. (2010a). *Proceedings of the IVth Int. Conference on Research Challenges in Information Science*, chapter Enforcing Conceptual Modeling to Improve the Understanding of Human Genome. IEEE Press.

Pastor, O., Pastor, M., and Burriel, V. (2010b). Conceptual modeling of human genome mutations: a dichotomy between what we have and what we should have. In *Proceedings of Bioinformatics 2010*, pages 160–166. BIOSTEC Bioinformatics.

Paton, N., Khan, S., Hayes, A., Moussouni, F., Brass, A., Eilbeck, K., Goble, C., Hubbard, S., and Oliver, S. (2000). Conceptual modeling of genomic information. *Bioinformatics*, 16(6):548–557.

Smith, T., Lee, M., Jerome, N., McEuen, M., Taylor, M., Hood, L., and King, M. (1996). Complete genomic sequence and analysis of 117 kb of human dna containing the gene brca1. *Genome Research*, 6:1029–1049.

van der Kroon, M., Ramirez, I. L., Levin, A., Pastor, O., and Brinkkemper, S. (2009). Mutational data loading routines for human genome databases: the brca1 case. Report UUCS2009020, Utrecht University.