International Journal of Data Mining, Modelling and Management. 2013; 5(3): 193-209

SNP locator: a candidate SNP selection tool

José A. Seoane, Vanessa Aguiar-Pulido, Alba Cabarcos, Sonsoles Quintela, Juan Ramon Rabuñal and Julián Dorado

Information and Communications Technologies Department, School of Computer Science. Campus de Elviña s/n, University of A Coruña, 15071, A Coruña, Spain

Abstract
In this work, a data integration approach using a federated model based on a service oriented architecture (SOA) is presented. The BioMOBY middleware was used to implement each service which is part of the integration process. As an example of usage of this architecture, a web tool for candidate SNP selection has been developed. Thus, several BioMOBY services have been created as the model layer of the web application. Each data source has a wrapper which communicates with the federated model, that is, the BioMOBY model, and this model is the one that interacts with the client.

Keywords:
Data integration; Federated data integration; SNP; Service oriented architecture; SOA; BioMOBY; Association studies; Webtool; Biomedical data integration; Federated database systems

Biographical notes:
José A. Seoane received his MSc in Computer Science from the University of A Coruña, Spain in 2008. In 2010 he finished his MS in Statistical Learning and Data Mining from the Spanish National University of Distance Learning. He is a scientific/technical personnel of the same university since 2004. He has participated in several research projects related to biomedicine and artificial intelligence. He is a member of the Galician Bioinformatics Network. His work lines are focused on data integration, artificial neural networks, evolutionary computation, data mining, data integration and bioinformatics.

Vanessa Aguiar-Pulido received her MS in Computer Science at the University of A Coruña, Spain in 2008. She is currently a PhD student in Information and Communication Technologies at the Faculty of Informatics of the University of A Coruña. She has received different research grants from different administrations for three years. She is member of the Galician Bioinformatics Network. Her main research interests are data mining, evolutionary computation and bioinformatics.

Alba Cabarcos received her MS in Computer Science at the University of A Coruña, Spain, in 2009. She is a scientific/technical personnel of the same university, and her work lines are focused on biomedical informatics and text mining.

Sonsoles Quintela studied at the University of A Coruña, Spain, finishing in 2006 her Bachelor's degree in Technical Engineering in Computer Management, and in 2009 her MS in Computer Science. During her studies, she performed two final year projects, both in the field of biomedicine. One of them involved microarray image analysis, and the other was related to single-nucleotide polymorphisms in human genetics. In 2005 and 2007 she received student scholarships as assistant in the Department of Information and Communication Technologies from the University of A Coruña.

Juan Ramon Rabuñal obtained his MS in Computer Science in 1999, PhD in Computer Science in 2002, and PhD in Civil Engineering in 2008, all at the University of A Coruña. Nowadays, he shares his time between his lecturer position at the Faculty of Computer Science and the direction of the Centre of Technological Innovations in Construction and Civil Engineering. He has also headed several research projects. His main

research interests are artificial neural networks, genetic programming, genetic algorithms and artificial intelligence in civil engineering.

Julian Dorado is an Associate Professor in the Faculty of Computer Science in the University of a Coruña. He finished his graduate studies in Computer Science in 1994. In 1999, he obtained his PhD, with a special mention of European Doctor. In 2004, he has finished his graduate in Biology. He has worked as a teacher of the university for more than eight years. He has published many books and papers on several journals and international conferences. He is at present working on bioinformatics, evolution

1 Introduction

The analysis of the differences existing in DNA sequences is a great source of information to identify genes that may have an influence on the development of a disease or on regular biological processes, such as growth or reproduction. In disease development studies, information about genetic variations is critical to understand which influence genes may have and how genetic and functional variations are related. Moreover, reaction to therapy or drugs can also be affected by genetic differences.

Nowadays, among the existing variations at genes, the most studied ones are single-nucleotide polymorphisms (SNPs), as it is the simplest way, and also the most frequent, of genetic variability. They are responsible for 80% of the variations between two individuals (Carlson, 2008). Thus, in certain disease studies, among the SNPs located at a gene, candidate SNPs which could be associated to a disease can be determined.

The amount of available genetic information has increased a lot with online databases, which are frequently updated and provide access to a great amount of contemporary information. These services allow researchers to determine certain interesting characteristics of genes (for example, details about their sequence, location or expression pattern) in a faster way. The main problem is that there isn't a standard to represent biological information in these databases. Thus, the task of integrating all of this information is not trivial.

To perform this integration, a middleware which allows the different laboratories to share their data and analysis algorithms is needed. This type of infrastructure should be implemented as a web application, allowing distributed data discovery, exploration, analysis and integration (Foster and Grossman, 2003). Data integration tasks from heterogeneous data sources is a previous step to data analysis in all data mining tasks which involve heterogeneous and distributed data sources. There already exist generic approaches which use workflows and web services to perform integration previous to analysis tasks (Perez et al., 2007; Olejnik et al., 2009; Congiusta et al., 2008; Diamantini and Potena, 2008).

Finally, in disease development studies there is a need to prioritise the list of SNPs considered, as the number of them can be of hundreds or thousands, in order to reduce costs. Studying these candidate SNPs involves many databases and a great amount of data to be managed, as well as analysing the necessary samples for studying the whole data. Therefore, all of this is very costly. For this reason, in this work, a web-based tool that allows obtaining a set of candidate SNPs for a specific disease, following criteria specified by the expert who uses the tool, is presented. This tool will use information retrieved from different genetic databases, integrating and filtering it according to specific criteria, to finally return a reduced set of ordered SNPs. Thus, this will make the task of searching relevant SNPs easier and faster and will decrease the cost resulting from studying thousands of SNPs.

2 Biological background

An association study tries to identify which variations in the genome predispose to develop a certain disease. Therefore, the objective is to obtain information about which genetic variants are involved in this predisposition.

Among the existing genetic variants, SNPs are the most studied ones. A SNP (den Dunnen and Antonarakis, 2000) is a single nucleotide site where two (of four) different nucleotides occur in a high percentage of the population, that is, at least in 1%. These variants can influence the development of a disease and are used in genomics to compare regions of the genome related to the disease studied.

In general, there is a large number of SNPs which may be related to a disease. Thus, experts in genomics must determine the minimum number of them which are significantly relevant in its development.

2.1 Variations in the human genome

A surprising fact about the human genome is that if two DNA sequences from two different human beings are compared, very few differences may be found. Only 0.1% of the genome bases make a person different from another one. A variation occurs when the order of the bases in a DNA sequence changes. Variations may involve a change in one or more bases, although not all of them have an effect on humans. The consequences of a change in the DNA depend on two factors: which part of the genome is modified and the exact nature of the modification.

Most variations do not have a known effect, as they occur in non-coding DNA regions. Furthermore, there are some variations that even though they occur in coding regions, no visible effect is observed. All of these variations are called silent variations. Some variations which occur in coding regions are harmless. For example, they can determine the colour of the eyes, the height... There also exist some which do not have negative effects because they don't affect the function of the protein produced.

However, there exists a group of variations in coding regions which have negative effects and may cause diseases, as the changes in the genome alter important proteins. Finally, there are some genetic variations which have 'latent' effects. This type of variation increases the risk of developing a disease, but only after an exposition to certain agents present in the environment.

2.2 Single nucleotide polymorphisms (SNPs)

SNPs are distributed all over the genome and can be found both in coding and non-coding regions. They may cause silent, harmless, negative or latent effects. There can be billions of SNPs in each human genome. Thus, the large amount of SNPs and the fact that it is easy to measure them makes this type of variation very significant. Depending on the effect the SNP has, it can be classified as (Brown, 2002):

- Synonymous SNP: the mutation produced is silent. The mutated gene encodes the same protein as the original gene.

- Non-synonymous SNP: the protein encoded by the mutated gene has changed in one amino acid. Most of the time there isn't a significant effect in the biological activity of the protein because most proteins can tolerate some changes in their amino acids without causing visible effects. However, if the change occurs in certain amino acids, the patient may develop some symptoms of a specific disease.

- Intronic SNPs: those that appear in non-coding regions, so they are not translated into a protein.

Another important characteristic of a SNP is whether it is a tagSNP (Smith, 2008) or not:

- There is a great correlation among SNPs located at the same region. Thus, it is possible to use a SNP (the tagSNP) to predict the presence of other SNPs. This allows reducing the number of SNPs that have to be analysed in order to study genetic variants associated to specific diseases.

2.3 Association studies and SNPs

Association studies of the genome allow scientists to identify genes related to a certain human disease. This method searches, in the genome, for those SNPs that appear more frequently in patients with a certain disease than in patients which did not develop that disease. Each study can search for hundreds or thousands of SNPs at the same time. Researchers use this type of data to mark genes that could contribute to increase the risk of a person to develop a disease. With this type of study, SNPs related to diseases such as diabetes, Parkinson or Chron have been identified.

3 Related work

There exists a great amount of works related to the prioritisation of SNPs (Bhatti et al., 2006). Subsequently, the most relevant ones are described.

MutDB (Mooney and Altman, 2003) is a web portal that allows determining which SNPs may be related to the alteration of the function of its associated protein. To determine this, multiple sequence alignment is mainly used, as well as the protein structure.

SnpSelector (Xu et al., 2005) is a web system which allows selecting SNPs from a list of genes or a specific genomic region. Once all the SNPs have been retrieved, the system prioritises them based on whether they are tagSNPs or not, their allelic frequencies, their function and their regulatory potential. The system output is a spreadsheet. All of the data is stored in a local database and retrieved from Hapmap, SNP Consortium, JSNP, Affymetrix and Perlegen.

The SNP Function Portal (Wang et al., 2006) offers the possibility to obtain the potential biological impact of a series of SNPs and the relationships between genes and markers using different databases. These relationships are obtained from SNPs that have already been classified in six categories (genomic elements, transcription regulation, protein function, pathways, diseases or population genetics).

Function Analysis and Selection Tool for Single Nucleotide Polymorphisms (FASTSNP) (Yuan et al., 2006) is a web server that allows identifying and prioritising SNPs which are relevant in relation to phenotypic risks and functional effects. This web server follows an always-update approach based on the usage of wrappers over external databases. Thus, it does not store SNPs or related data in an own database. Once updated SNPs have been obtained, the system prioritises them based on five levels depending on whether the SNP is at a coding or untranslated region, whether it is synonymous or non synonymous, its position at the gene, allelic frequency and haplotypic information.

SNP@Domain (Han et al., 2006) is a tool, available as a web interface, which can identify SNPs within human proteome domains. SNPs were annotated from dbSNP with protein structure-based domains, as well as sequence-based domains.

PupaSuite (Conde et al., 2006) is a web tool for the selection of SNPs with potential phenotypic effect specifically oriented to help in the design of large-scale genotyping projects. The input of the program can be a set of genes or chromosomal regions which would correspond with two common types of analysis: genes probably related to a disease because they are functionally related or genes present in a chromosomal region linked to a disease. The features considered for choosing the appropriate SNPs were the following: transcription factor binding sites from the trasnfac database, intron/exon border consensus sequences, exonic splicing enhancers, triplex-forming oligonucleotide target sequences, SNPs in exons causing amino acid change, pmut predictions, selective strengths and SNP effect predictions.

Single Nucleotide Polymorphism Annotation Platform (SNAP) (Li et al., 2007) is a server designed to analyse single genes and relationships between genes based on SNPs identifier or accession code. Several databases like Enseml, Uniprot, Pfam, CBS-DAS, BIND and KEGG were integrated.

SNPBrowser (De La Vega et al., 2006) is a tool created to assist the selection of SNPs for linkage disequilibrium studies. This stand-alone software is based on two paradigms: the selection of evenly spaces on the physical or metric linkage disequilibrium maps and the selection of non-redundant subsets of haplotype tagging SNPs.

PolyDoms (Jegga et al., 2007) provides a database from which the user can select a list of candidate SNPs to be evaluated in experimental or epidemiological studies for impact on protein functions and disease risk association.

QuickSNP (Grover et al., 2007) is a web server that allows the user to select SNPs for association studies. This web server follows a gene-centric approach to tagSNP selection, accepting multiple genes as input. It also allows rejecting too closely spaced SNPs and finally calculates the cost of the whole genotyping study.

The Structure SNP (StSNP) web server (Uzun et al., 2007) integrates data related with non-synonymous SNPs. Key functional and structural information along with known pathways the protein is involved in, have all been linked together to provide users some advantages when compared to other current resources. It provides the sequence, structure and pathway information, as well as graphical displays of the nsSNPs, and loads models of the proteins described.

The Funcional Single Nucleotide Polymorphism (F-SNP) database (Lee and Shatkay, 2008, 2009) integrates information about functional effects of SNPs. These effects are predicted and indicated at the splicing, transcriptional, translational and pos-translational level. It integrates data from several databases and bioinformatics tools. This tool provides a set of potential disease-causing SNPs for association studies.

Potentially functional SNPs (pfSNP) (Wang et al., 2011) is a web portal that aims to identify the potential functional significance of SNPs, based on previously published reports, inferred potential functionally from genetic approaches and sequential motifs.

Varietas (Paananen et al., 2010) is a web-based database portal that has been designed to aid researchers to easily retrieve information on a set of variations (SNPs or CNVs) related with genes or genomic elements. The retrieved information can be exported using a web browser or downloaded into a text file. It also offers links to external resources such as Pubmed, dbSNP, SNPedia or Ensembl. This approach uses a local database where all data are integrated, updated periodically.

Finally, Spot (Saccone et al., 2010) is a web system which integrates biological databases and allows prioritising SNPs for subsequent GWAs analysis. This system uses the ‘genomic information network’ (GIN) prioritisation described in Saccone et al. (2008).

Many tools analysed in this section allow selecting SNPs using as input a gene or a region, but none of them allow performing the selection and prioritisation of SNPs using as input a disease. However, once candidate SNPs have been obtained, certain tools allow retrieving more data related to these SNPs in order to improve filtering and prioritisation. Finally, the fact that this tool has been developed as a set of BioMOBY modules of free-use allows that each of these parts can be incorporated to other tools, improving their potentialities. Unfortunately, none of the tools presented include this feature.

4 Candidate SNP selection tool

One objective of the work presented was to develop a tool that allows ordering SNPs which are related to a specific disease, following certain criteria specified by a researcher and, thus, obtain a set of candidate SNPs, retrieving and integrating data from different biological databases also specified by the researcher. Each model of the system was developed following the BioMOBY standard for bioinformatics web services.

For this purpose, it is necessary to provide the researcher with a single interface that, on one hand, allows him/her not having to look for information in several databases and, on the other hand, gives him/her the possibility to establish criteria that will be used to order the set of candidate SNPs according to some specific characteristics. The tool should also be open for the inclusion of new biological data sources, permitting also the inclusion of new functionalities, and its implementation should be reusable and offer the possibility to be shared with other researchers.

These features will allow reducing the time consumed by the queries that obtain SNPs from several databases and the number of SNPs that must be studied, reducing as well the costs of the association study.

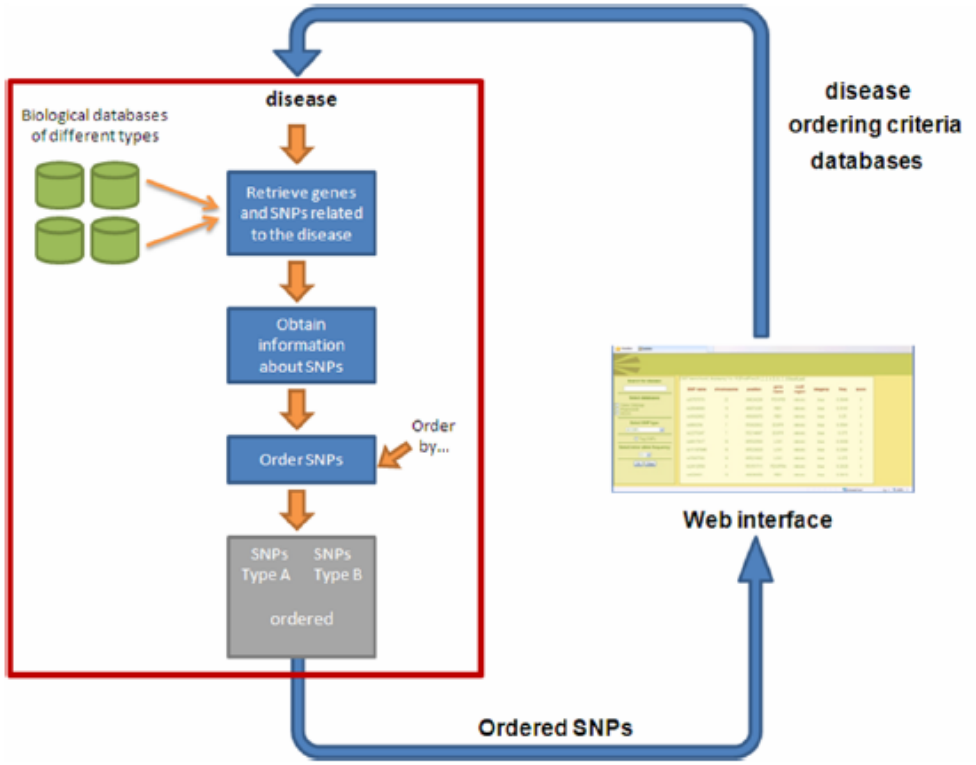

Figure 1 General workflow of the web-based tool (see online version for colours)

4.1 Tool description

As mentioned before, a tool with the previously described features has been developed as an example of usage of the integration approach proposed. This tool provides the researcher with a web interface in which he/she introduces a disease and certain characteristics that will be used to order the candidate SNPs returned as output, as well as the databases from which to retrieve the information.

Taking into account the disease query, the tool retrieves data from the specified biological databases and obtains SNPs or genes (depending on the database queried) which are related to the query, ordering them by relevance, constructing a set of candidate SNPs. Once the set of SNPs or genes has been obtained, data about SNPs at those genes (or genes obtained previously) are retrieved. Finally, SNPs are ordered following the criteria established by the researcher, related to the characteristics gathered about the SNPs. After this, the collection of genes is displayed to the researcher using the web interface. All of this process is shown in Figure 1.

4.2 Federated approach architecture

The federated data integration architecture was considered as the most adequate to solve this problem. This architecture uses a common data model for the whole system and each data source is adapted to the model using the wrapper pattern. Each database or web service provides data using its own data model. Thus, this type of architecture locates a wrapper over each data source, adapting the output of the databases or the web services to the common model. Furthermore, it provides a uniform way of accessing data from all of the data sources.

4.3 BioMOBY

The Service Oriented Architecture (SOA) uses web services to provide support to the software requirements of the users. A very important characteristic of SOA is that it provides interoperability and extensibility between operations. With this architecture, services are very loosely coupled and are highly interoperable, as well as portable. Therefore, it would be interesting to take advantage of the characteristics of this architecture in this work.

To introduce the characteristics of SOA to work with biological data, the BioMOBY platform was used. BioMOBY is a registry of web services used in bioinformatics. It allows interoperability between biological data hosts and analytical services by annotating the services with terms taken from standard ontologies. It also provides an architecture to search for and distribute biological data using web services and, nowadays, can be considered a ‘de facto’ standard for searching for and integrating these data.

BioMOBY allows also creating independent services that deal with biological data and can be shared with different organisations and used in different processes, integrating them to achieve more complex processing. Furthermore, it offers a common data model to represent the information, regardless of the source. For example, if we have gene data from different biological databases, BioMOBY allows registering gene as a data type with several fields which have been considered important, such as an identifier, a name, the position in the gene and its description. If another user needs to use it, he/she can use the gene data type in his/her service and can also add it as part of a data type information registered by him/her.

4.4 Usage of BioMOBY in the tool presented

The previously described characteristics led to the use of BioMOBY. This platform allows not only working with information from different sources and unifying its format, but also creating interoperable and independent services that will allow the inclusion of new functionalities easily. Furthermore, BioMOBY is also a good choice to implement the federated architecture, as this platform provides its own data model to which data obtained from different data sources can be translated.

As an implementation of the SOA architecture, BioMOBY has a centralised registry of services and data types named Moby Central, service providers and BioMOBY clients.

The Moby Central registry is shown in Figure 2, as well as public biological databases from which the implemented services retrieve the data they need.

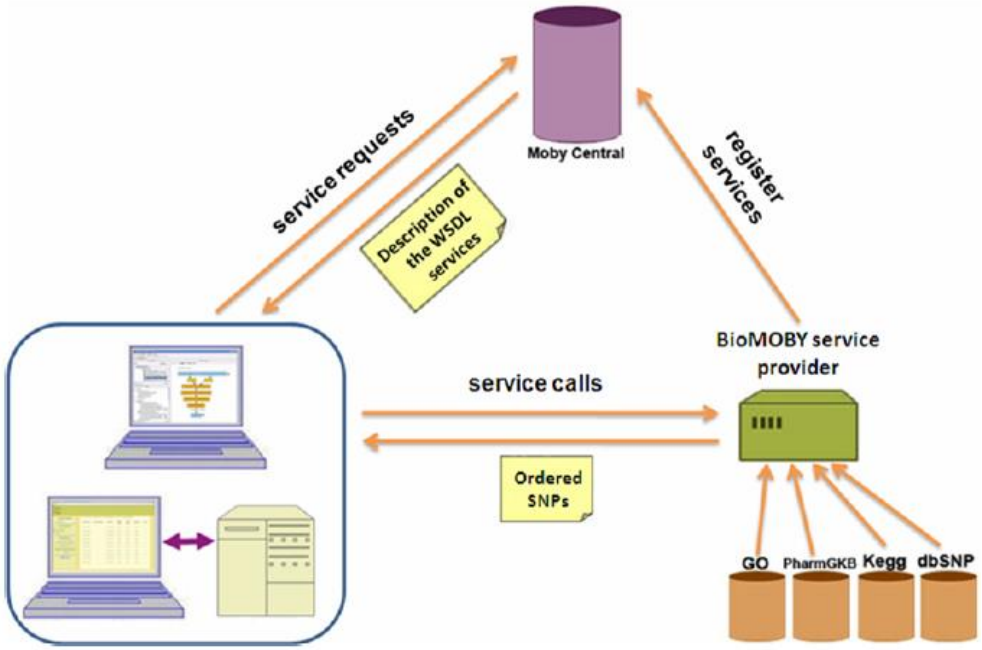

Figure 2 Web-based tool and BioMOBY (see online version for colours)

With these elements, the message exchange in the tool works as follows:

- Firstly, the BioMOBY registry provider implemented for the tool registers the services in Moby Central, indicating its inputs, outputs and the data types provided.

- Once the services have been registered in Moby Central, they can be used by any researcher. These services retrieve data from biological databases to process them.

- After the user uses the web interface or any BioMOBY client uses the implemented services, the implemented client queries Moby Central for the services needed. Subsequently, Moby Central returns a description of the service interface.

- After this, the client can make the execution of the service automatically in order to finally obtain the SNPs ordered.

In order to allow a researcher to use the tool regardless of his/her location and platform, it was integrated in a J2EE web application. Using this web interface, researchers can retrieve SNPs related to a disease from specific databases ordered following the criteria they want.

A temporal database was also included to avoid having to reprocess all the information and perform again the same queries to all the biological databases in case the researcher would like to order the SNPs following different criteria than before. However, the implemented services can also be used without the temporal database, using as input a data collection and returning another one.

4.5 Tool architecture

This section describes how each database can be accessed and how it is used and, after that, a figure showing the data integration approach followed is included. Finally, the different web services implemented following the BioMOBY standards are described.

4.5.1 Data integration model

Four databases are accessed by the web services implemented in this work: GO, KEGG, PharmGkb and dbSNP. Subsequently, it is explained how information is retrieved from the different databases used.

- information from KEGG and PharmGkb is retrieved using web services
- information from GO is retrieved directly accessing the database, as well as using web services.
- information from dbSNP is retrieved through Ensembl either using web services or directly accessing the database.

Thus, in this work all the information is retrieved from external databases or web services, following a federated approach. This approach has an important advantage: data recovered are always updated, since the information sources are responsible for these updates.

The main drawbacks are availability and performance problems. Firstly, if one source is not available temporarily, it will not be possible to retrieve the information it contains. In the tool presented, this problem does not happen very often, since the selection of the sources has been carefully considered. All sources included are widely known and used, and they do not usually have availability problems. Despite this fact, the candidate SNP selection tool handles the errors that can appear so that errors in one information source do not suspend the whole process of information retrieval. Thus, if one source is not available temporarily, results recovered will not include information from that source, but it will obtain data from the other ones. With regard to performance, these temporary fails do not reduce it in a significant way. The tool will try to connect to the source (and this will consume some time), but if it is detected that the source is not available, the retrieval process will continue normally, processing other sources.

Another issue of the approach selected is response time, since time needed is bigger when sources are accessed remotely. The alternative to this option would consist in having a local copy of the information repositories, but this has an important drawback: data retrieved are not always updated. Having in mind that the option selected is the first one, response time of the services implemented depend basically on the time consumed by the sources during the information retrieval process. In general, services that access databases directly are faster than those that use web services.

Data are retrieved from different databases so there may be different types of conflict (semantic, format, etc.). In order to solve these conflicts, each module of BioMOBY transforms the data using a wrapper-mediator architecture. Thus, data following the original model of the data source is transformed to a common model, that is, the BioMOBY model. For each data source there exists a specific wrapper. Once this transformation has been done, the BioMOBY model is used during the interaction of the models, as well as during external access. However, it can be extended to other specific applications. Thus, this model is common to the whole framework.

Figure 3 shows the data integration model used in this work. For each database, there exists a wrapper which communicates with the federated model, that is, the BioMOBY model. Thus, the client does not know that the system is retrieving data from four different databases.

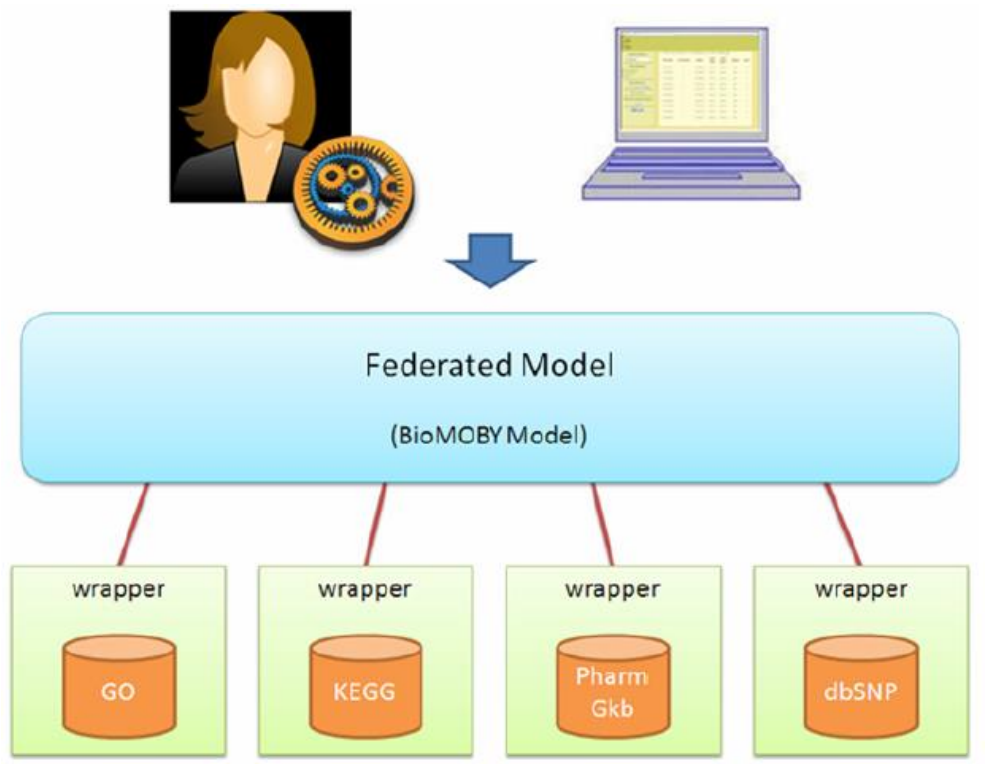

Figure 3 Federated model architecture (see online version for colours)

4.5.2 Web services

Below, the different web services implemented in this work are described. These services are provided as part of the candidate SNP selection tool, as well as BioMOBY web services.

- Services which retrieve information from a disease:
  - *PharmGKB_SNPsFromDisease* – retrieves the SNPs related to a disease from the PharmGKB database.
  - *GO_GenesFromDisease* – retrieves the genes related to a disease from the gene ontology database.
  - *KEGG_GenesFromDisease* – retrieves the genes related to a disease from the Kyoto Encyclopaedia of Genes and Genomes database.

The first web service receives as input a disease and returns a collection of disease-related SNPs, while the second and the third services return genes instead of SNPs.

- Filters:
    - *FilterSNPsByCodingRegion* – retrieves a score for each SNP of a collection of SNPs, by filtering them according to the type of coding region where they are located.
    - *FilterTagSNPs* – retrieves a score for each SNP of a collection of SNPs, by filtering them according to if they are tag SNPs or not.
    - *FilterSNPsByAlleleFreq* – retrieves a score for each SNP of a collection of SNPs, by filtering them according to their allele frequency.

These web services receive as input a collection of SNPs related to some specific genes and return a collection of rated SNPs according to different criteria.

- Other services:
    - *SNPsFromGenes* – retrieves the SNPs related to a collection of genes, from the dbSNP database. This web service returns a collection of SNPs.
    - *GetOrderedSNPs* – takes a collection of SNPs and returns it ordered according to the scores obtained for each SNP.
    - *GetBasicGeneDataFromGenBankIDs* – retrieves a collection of basic data (gene name, identifier and description) from a collection of gene identifiers of GenBank (NIH genetic sequence database, an annotated collection of all publicly available DNA sequences).

4.6 Tool operation

In Figure 4, the elements of the tool are shown in an execution. In this figure, there is one user, a web client that will call the BioMOBY client and a temporal database, in which intermediate data is stored to improve efficiency.

The tool works as follows:

- Firstly, the user introduces a disease using the web interface of the tool, indicating whether he/she wants to retrieve data from all of the available biological databases, and also indicates which criteria have to be followed when ordering the SNPs.
- After that, the web application calls the BioMOBY client, providing the disease and the ordering criteria. Subsequently, the BioMOBY client calls the services.
    - The first call is made to the service which retrieves genes from the gene ontology database. This service takes the disease and accesses the gene ontology database using JDBC, searching for the genes related to the given disease and returning them as output or inserting them in the temporal database.
    - In a similar way, other services retrieve information from KEGG and PharmGKB.
    - Another service retrieves SNPs from the temporal database or from a collection of SNPs provided as input and, looking for these SNPs in dbSNP, retrieves the required information about them. These data can also be returned as a collection or stored in the temporal database.

- According to the criteria indicated by the researcher, the services that rate the SNPs are called. These services receive as input a collection of SNPs or take them from the temporal database and return a collection of ratings, one for each SNP. They can also directly assign them to each SNP in the temporal database. Finally, the service returns a collection of SNPs ordered by the ratings.

- Thus, the BioMOBY client returns a collection of ordered SNPs to the web application, which shows them to the user.

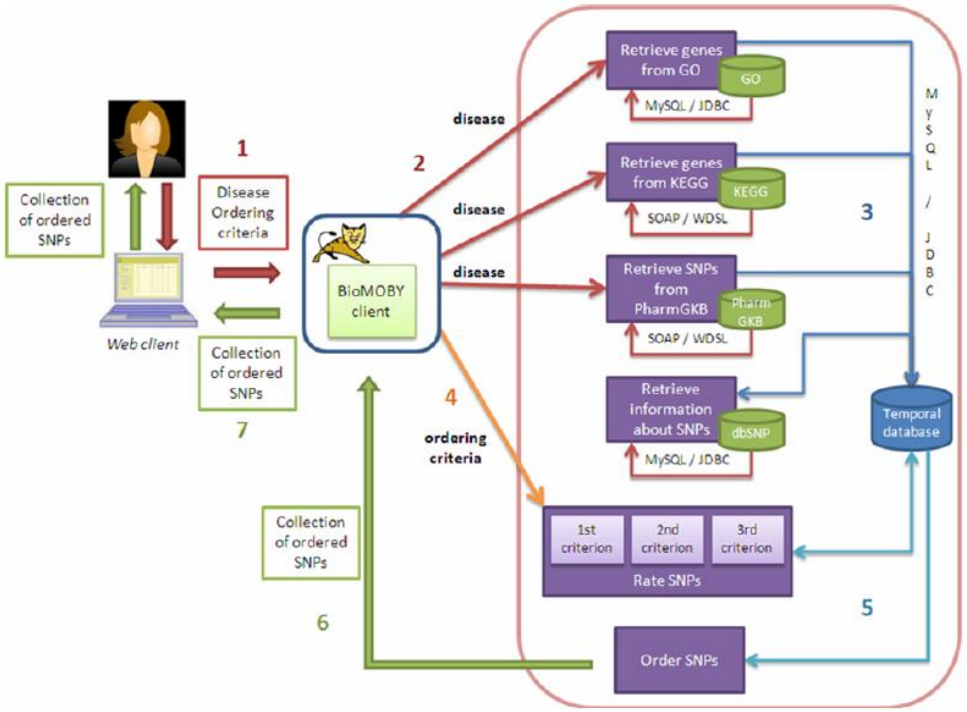

Figure 4 How the web-based tool works (see online version for colours)

All of the implemented services for the tool are registered in the BioMOBY service registry. Thus, they are accessible to any researcher through the Moby Central repository.

Therefore, these services can be used implementing a BioMOBY client that accesses them, as done in the implementation of this tool, or they can be used integrating them in previously existing applications which are BioMOBY clients (for example, Taverna).

Taverna (Hull et al., 2006) is a tool which offers the possibility to design and execute workflows with BioMOBY services, among others. To test the services, Taverna was used as client to access the BioMOBY services in order to check the correct functioning of the services, working all together as well as individually. Figure 5 shows a workflow created for testing the services.

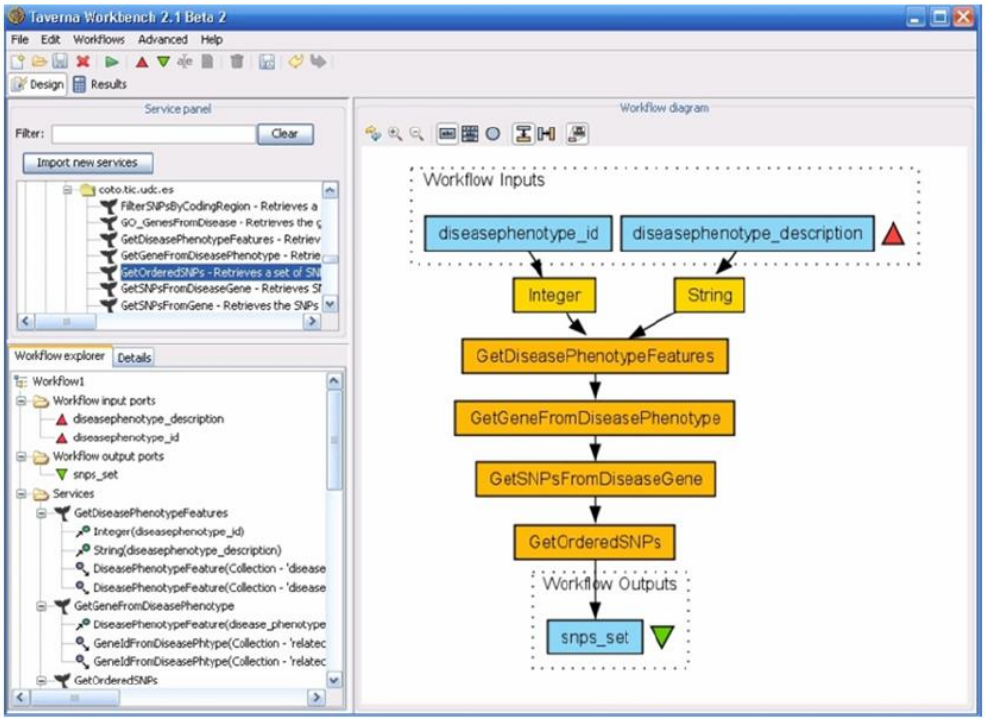

Figure 5 Tests with Taverna (see online version for colours)

5 Example of execution

Here, an example of how the tool presented in this paper works is shown.

The system was tested using the following parameters:

- the disease introduced was ‘glioma’
- the ‘gene ontology’ database was chosen
- ‘all types of SNPs’ was selected as SNP type
- ‘tag SNPs’ was checked.

Results of this query are shown in Figure 6.

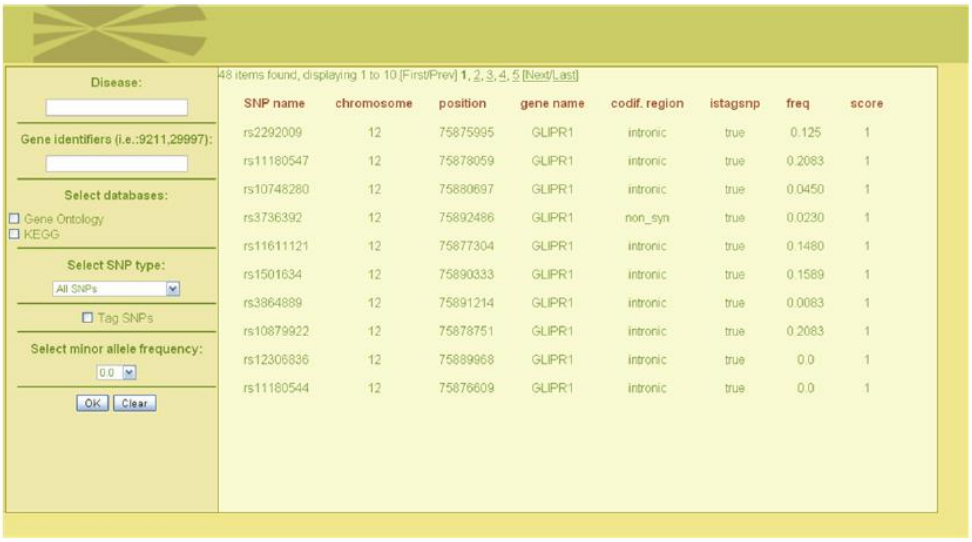

Figure 6 Results (see online version for colours)

6 Conclusions and future work

In this work, in brief, a data integration approach using a federated model based on a SOA was presented. Therefore, the BioMOBY middleware was used to implement the services which are part of the integration process. Information from four different well-known databases was integrated in this federated model.

A tool for retrieving sets of candidate SNPs has been presented as an example of the integration approach proposed. This tool allows:

- obtaining a set of SNPs ordered according to a rating which follows criteria specified by a researcher
- providing access to data from different biological databases.

Both aspects allow reducing economical costs. Furthermore, the second functionality allows also reducing time consuming and effort, as the researcher does not have to study SNPs from different databases, one by one.

A web application, with a simple interface, has been developed with the aim of making the work of researchers easier, gathering data from different data sources in a single application and giving the possibility to include new databases and ordering criteria easily.

BioMOBY was used as a SOA platform to obtain reusable and independent web services. The integration architecture followed the federated data model approach, with the BioMOBY data model as common data model. This way, two objectives were achieved. On one hand, the integration process among the different BioMOBY services created for the SNP selection application is performed nearly automatically, requiring only the implementation of a wrapper which transforms the data from the data source to the BioMOBY model. On the other hand, using the model of BioMOBY allows reusing services developed for other biomedical applications, increasing the power of the application developed, and including new functionalities such as data processing or other data sources already developed as BioMOBY services.

Future work may include adding new relevant biological databases, more filters, or considering other genome variations. Another future line may involve offering the possibility of including a set of genes that may have an impact on the development of a disease. Finally, another possible future work could be to add functionalities to the web application which allow researchers to personalise their searches.

Acknowledgements

José A. Seoane acknowledges the funding support for a research position by Isabel Barreto grant from Xunta de Galicia (Spain). This work is supported by the following projects: Galician Network for Colorectal Cancer Research (REGICC, Ref. 2009/58) from the General Directorate of Research, Development and Innovation of Xunta de Galicia, Ibero-American Network of the Nano-Bio-Info-Cogno Convergent Technologies, Ibero-NBIC Network (209RT-0366) funded by CYTED (Spain), grant Ref. PIO52048 and RD07/0067/0005 funded by the Carlos III Health Institute, 10SIN105004PR funded by Economy and Industry Department of Xunta de Galicia and PHR2.0: Registro Personal de Salud en Web 2.0 (Ref. TSI-020110-2009-53) funded by the Spanish Ministry of Industry, Tourism and Trade.

References

Bhatti, P., Church, D.M., Rutter, J.L., Struewing, J.P. and Sigurdson, A.J. (2006) 'Candidate single nucleotide polymorphism selection using publicly available tools: a guide for epidemiologists', *Am J Epidemiol*, Vol. 164, No. 8, pp.794–804.

Brown, T.A. (2002) *Genomes*, 2nd ed., Wiley-Liss, Oxford.

Carlson, B. (2008) 'SNPs – A shortcut to personalized medicine. Medical applications are where the market's growth is expected', *Genetic Engineering and Biotechnology News*, Vol. 28, No. 12.

Conde, L., Vaquerizas, J.M., Dopazo, H., Arbiza, L., Reumers, J., Rousseau, F., Schymkowitz, J. and Dopazo, J. (2006) 'PupaSuite: finding functional single nucleotide polymorphisms for large-scale genotyping purposes', *Nucleic Acids Res*, Vol. 34, No. 2, pp.621–625.

Congiusta, A., Talia, D. and Trunfio, P. (2008) 'Service-oriented middleware for distributed data mining on the grid', *Journal of Parallel and Distributed Computing*, Vol. 68, No. 1, pp.3–15.

De La Vega, F.M., Isaac, H.I. and Scafe, C.R. (2006) 'A tool for selecting SNPs for association studies based on observed linkage disequilibrium patterns', *Pac Symp Biocomput*, pp.487–498.

den Dunnen, J.T. and Antonarakis, S.E. (2000) 'Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion', *Hum Mutat*, Vol. 15, No. 1, pp.7–12.

Diamantini, C. and Potena, D. (2008) 'Semantic annotation and services for KDD tools sharing and reuse', *ICDMW '08 Proceedings of the 2008 IEEE International Conference on Data Mining Workshops*, pp.761–770.

Foster, I. and Grossman, R.L. (2003) 'Data integration in a bandwith-rich world', *Communications of the ACM – Blueprint for the Future of High-Performance Networking*, Vol. 46, No. 11, pp.50–57.

Grover, D., Woodfield, A.S., Verma, R., Zandi, P.P., Levinson, D.F. and Potash, J.B. (2007) 'QuickSNP: an automated web server for selection of tagSNPs', *Nucleic Acids Res*, Vol. 35, No. 2, pp.115–120.

Han, A., Kang, H.J., Cho, Y., Lee, S., Kim, Y.J. and Gong, S. (2006) 'SNP@Domain: a web resource of single nucleotide polymorphisms (SNPs) within protein domain structures and sequences', *Nucleic Acids Res*, Vol. 34, No. 2, pp.642–646.

Hull, D., Wolstencroft, K., Stevens, R., Goble, C., Pocock, M.R., Li, P. and Oinn, T. (2006) 'Taverna: a tool for building and running workflows of services', *Nucleic Acids Res*, Vol. 34, No. 2, pp.729–732.

Jegga, A.G., Gowrisankar, S., Chen, J. and Aronow, B.J. (2007) 'PolyDoms: a whole genome database for the identification of non-synonymous coding SNPs with the potential to impact disease', *Nucleic Acids Res*, Vol. 35, No. 1, pp.700–706.

Lee, P.H. and Shatkay, H. (2008) 'F-SNP: computationally predicted functional SNPs for disease association studies', *Nucleic Acids Res*, Vol. 36, No. 1, pp.820–824.
Lee, P.H. and Shatkay, H. (2009) 'An integrative scoring system for ranking SNPs by their potential deleterious effects', *Bioinformatics*, Vol. 25, No. 8, pp.1048–1055.
Li, S., Ma, L., Li, H., Vang, S., Hu, Y., Bolund, L. and Wang, J. (2007) 'Snap: an integrated SNP annotation platform', *Nucleic Acids Res*, Vol. 35, No. 1, pp.707–717.
Mooney, S.D. and Altman, R.B. (2003) 'MutDB: annotating human variation with functionally relevant data', *Bioinformatics*, Vol. 19, No. 14, pp.1858–1860.
Olejnik, R., Fortis, T.F. and Toursel, B. (2009) 'Webservices oriented data mining in knowledge architecture', *Future Generation Computer Systems*, Vol. 25, No. 4, pp.436–443.
Paananen, J., Ciszek, R. and Wong, G. (2010) 'Varietas: a functional variation database portal', *Database (Oxford)*, Vol. 2010, baq016.
Perez, M.S., Sanchez, A., Robles, V., Herrero, P. and Peña, J.M. (2007) 'Design and implementation of a data mining grid-aware architecture', *Future Generation Computer Systems*, Vol. 23, No. 1, pp.42–47.
Saccone, S.F., Bolze, R., Thomas, P., Quan, J., Mehta, G., Deelman, E., Tischfield, J.A. and Rice, J.P. (2010) 'SPOT: a web-based tool for using biological databases to prioritize SNPs after a genome-wide association study', *Nucleic Acids Res*, Vol. 38, No. 2, pp.201–209. Saccone, S.F.,
Saccone, N.L., Swan, G.E., Madden, P.A., Goate, A.M., Rice, J.P. and Bierut, L.J. (2008) 'Systematic biological prioritization after a genome-wide association study: an application to nicotine dependence', *Bioinformatics*, Vol. 24, No. 16, pp.1805–1811.
Smith, M. (2008) *Translational Research in Genetics and Genomics*, Oxford University Press, New York.
Uzun, A., Leslin, C.M., Abyzov, A. and Ilyin, V. (2007) 'Structure SNP (StSNP): a web server for mapping and modeling nsSNPs on protein structures with linkage to metabolic pathways', *Nucleic Acids Res*, Vol. 35, No. 2, pp.384–392.
Wang, J., Ronaghi, M., Chong, S.S. and Lee, C.G. (2011) 'pfSNP: an integrated potentially functional SNP resource that facilitates hypotheses generation through knowledge syntheses', *Hum Mutat*, Vol. 32, No. 1, pp.19–24.
Wang, P., Dai, M., Xuan, W., McEachin, R.C., Jackson, A.U., Scott, L.J., Athey, B., Watson, S.J. and Meng, F. (2006) 'SNP function portal: a web database for exploring the function implication of SNP alleles', *Bioinformatics*, Vol. 22, No. 14, pp.523–529.
Xu, H., Gregory, S.G., Hauser, E.R., Stenger, J.E., Pericak-Vance, M.A., Vance, J.M., Zuchner, S. and Hauser, M.A. (2005) 'SNPselector: a web tool for selecting SNPs for genetic association studies', *Bioinformatics*, Vol. 21, No. 22, pp.4181–4186.
Yuan, H.Y., Chiou, J.J., Tseng, W.H., Liu, C.H., Liu, C.K., Lin, Y.J., Wang, H.H., Yao, A., Chen, Y.T. and Hsu, C.N. (2006) 'FASTSNP: an always up-to-date and extendable service for SNP function analysis and prioritization', *Nucleic Acids Res*, Vol. 34, No. 2, pp.635–641.