From Local Laboratory Data to Public Domain Database in Search of Indirect Association of Diseases: AJAX Based Gene Data Search Engine

Nawaz Khan
Computing Science, Middlesex University, London, UK,
n.x.khan@mdx.ac.uk

Ham Long
Computing Science, Middlesex University, London, UK
lamhong2000vn@yahoo.com

Shahedur Rahman
Computing Science, Middlesex University, London, UK
s.rahman@mdx.ac.uk

Tony Stockman
Computer Science, Queen Mary, University of London
tonys@dcs.qmul.ac.uk

Abstract

This paper presents an extensible schema for capturing laboratory gene variance data with its meta-data properties in a semi-structured environment. This paper also focuses on the issues of creating a local and task specific component database which is a subset of global data resources. An XML based genetic disorder component database schema is developed with adequate flexibilities to facilitate searching of gene mutation data. A web based search engine is developed that allows researchers to query a set of gene parameters obtained from local XML schema and subsequently allow them to automatically establish a link with the public domain gene databases. The application applies AJAX (Asynchronous Javascript and XML), a cutting-edge web technology, to carry out the gene data searching function.

1. INTRODUCTION

At present no single, unambiguous, complete and static model exists for storing gene mutation data for the purpose of analysing indirect association of diseases (Geihs, 2002 and Donnelly, 2003). Initially all the databases stored data on the basis of reviewing papers. An expert reads the literature to extract the required information and then interpret the results based on previous published results. But as the volume of data increased, new methods emerged to accelerate this process. One method to speed up journal scanning is the use of automated scanning procedures. Another method is to allow researchers to submit their data directly to the databases. Direct data submission in comparison with journal scanning gives better and faster data transfer result. However, the problem with this method of submission requires data to be published and the researchers are unable to submit the full set of results and photomicrographs because of the limitations of the interfaces that are used for data capturing. For example, Human Gene Mutation Database (HGMD) databases does not deal with image data formats. Moreover, in many genome research, data are needed to be submitted to multiple databases. A coordinated research program of a particular laboratory publishes data through a number of different databases such as GenBank, Protein Data Bank (PDB), Gene Data Bank and others. These laboratories would be unnecessarily burdened if they require to prepare their data for multiple databases and even to take the responsibility for checking the validity of links among the related elements in all the databases**.** Developing a coordinated direct data-submission method for all genome databases is still a major challenge. This paper is proposing to create a component database to address this issue and to submit data in details which will be a component of the whole federated database. The proposed database will be a small-scale, laboratory based and dedicated which will store all the laboratory results. It will also be coordinating with other community databases by providing automatic links which will enable the biological researchers to be able to submit the result with a single electronic transaction. This can then be coordinated automatically with multiple relevant

Twentieth IEEE International Symposium on Computer-Based Medical Systems (CBMS'07)
0-7695-2905-4/07 $20.00 © 2007 IEEE

data sources. This method of data submission can be used for a broad spectrum of the research community, starting from low-volume to high-volume data producers. Identifying and understanding the properties of genome, genome mutation and conducting comparison of genomes and its products depend on the provision of storing, sharing and analysing the complete genomic data sets. For example, mutation rate, mutation frequency, or protein images are fundamental data sets for comparing with wild type proteins and for variance analysis of diseases. This paper proposes a complete schema for storing the complete set of gene mutation data which can facilitate the search for the relevant gene data in order to carry out a variance analysis (Khan and Rahman, 2002). The following section describes the parameters that are used in XML model.

2. MODELS

In order to analyse the genetic variance in human, the parameters, mutation rate, mutation frequency in male and female for different ethnic background are needed to be stored. These will allow establishing an indirect association of diseases in human. The emphasis here is on presenting indirect association of disease data rather than presenting the direct association of disease information which can be accessed from HGMD. The proposed XML data models for depicting indirect association of diseases consist of the following models: Genetic trait data model, mutation type data model, laboratory data model for mutation study and pathological lesions data model. The implementation of conceptual level does not depend either on software or on the platform of implementation. Schemas are developed on the basis of conceptual level. The conceptual diagrams of the gene mutation data presented here will provide facilities for wide range of analysis by applying an appropriate search engine. These models are described here.

Genetic trait data model: A schema for storing the data for inheritance pattern analysis is proposed in Figure 1. The data model stores the following information for inheritance pattern analysis: recombination fraction, frequency of abnormalities and relative dinucleotide mutabilities. In inheritance pattern for Trait analysis has the following subtypes: background, ethnicity, geographical region, number of cases studied, types of genetic events and frequency of abnormalities. It is necessary to store background, ethnicity and geographical regions of cases. It has been shown that genetic expression varies from one geographical region to another geographical region. For example, *dyslipidaemia* gene falls in the same *loci* to which diabetes and obesity genes for many western countries, but the same *loci* might not be the root for this disorder in population from other geographical locations. A trait analysis depends on multiple corresponding reported clinical features that may be reported by different groups. Many *dismorphological* details are needed to correlated with gene mutation data of the component database.

Gene mutation type data model: There are two types of mutations which are mutation due to the production of less synthesised gene products (polypeptides) and mutations due to the production of abnormal gene products. Figure 2 proposes the conceptual schema for classification of gene mutation data. Amount of synthesis of gene products can be affected if transcription or translation is affected. The other reason which might affect the amount of synthesis of gene products is malfunctioning of gene structure or promoter function. Transcription region of mRNA can be affected due to the mutation at transcriptional site, splice junction, *3′ UTR* (3′ untranslated region) mutation or due to the *polyadenylation* at transcription site. On the other hand translation can be affected when there is any mutation at initiation codon, termination codon, nonsense codon or at *5′ UTR*. Abnormal gene product can also be produced when two genes are fused together, or if there is any defective post translational processing. Abnormal gene product can also be produced if gene products are shortened or gene products are elongated. Shortened gene product and elongated gene product might be caused due to deletion or frame shift in a particular gene. Deletion, frameshift and insertion can also affect the gene structure. A gene mutation ontology has been conceptualised in figure 2 in order to capture laboratory based gene mutation data.

Empirical data model: It is necessary to do experiment on restriction fragment length polymorphism (RFLP) to analyse an indirect association of diseases. RFLP experiment produces a number of photographs and data that explains gene position and gene restriction sites. In this experiment results are compared for genotypic variance analysis and to detect polymorphism in human gene that might responsible for diseases. This RFLP technique is used as genetic marker, by which genotype variance can be examined instead of assessing any phenotypic variation. In RFLP experiment each gel electrophoresis is compared with corresponding references. Figure 3 has captured the RFLP data requirements. A conceptual model for storing empirical data for this purpose is proposed in Figure 3.

Pathological lesions data model : Pathological lesions might be caused due to deletion, insertion or inversion. Base pair substitution and abnormalities in specific repeat sequences could also lead to the pathological lesions. Base pair substitution has two subtypes: single base pair, or multiple base pair. A conceptual diagram for pathological lesions data modeling is proposed in Figure 4. The model presented in Figure 4 enables biological researchers to store the pathological lesions data for future clinical references so that a correlation can be established with gene mutation data when any modification in gene occurs due to substitution by a single nucleotide. Multiple factors that are responsible for gene expression change are captured in Pathological lesions data model.

3. From Laboratory Data to Public Domain Data

A search engine application is designed to help researchers query the XML database for the information about Indirect Association of Diseases. This XML database includes four different XML files: Genetic_trait.xml, Gene_mutation_type.xml, Empirical.xml and Pathological.xml. Each of these files contains XML data developed from a gene data modelling schema. The detail of these schemas were described in section 2. Each schema displays the structure of one gene data modeling type such as its child elements of different sub-levels, the attributes, sequence and data type of each child element. The gene variance data are stored in the XML file as a node. Each node is specified by its ID attribute. Each node carries the gene variance data through its sub-level elements.

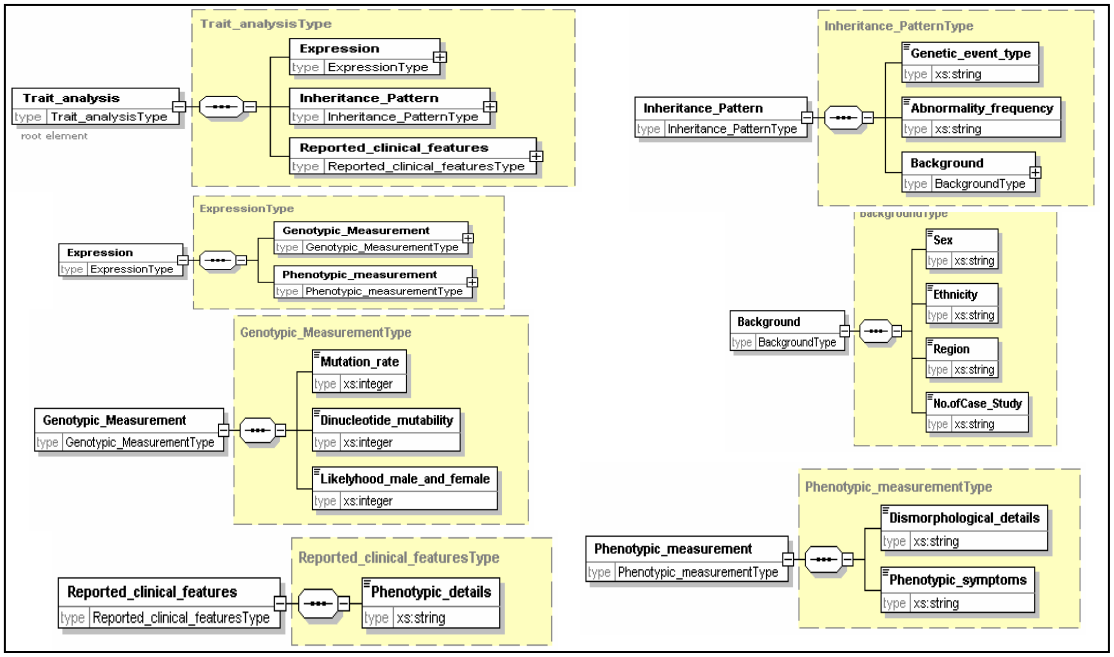

Figure 1. Genetic trait data modeling

This application allows researchers to submit new gene data record to the XML database. The new record will be added to the appropriate XML file as a new node and the XML tree will be extended with the new data. The system will save and update the extended XML file, so the newly added data will be available soon after they are submitted by researchers.

The application was built as a web-based application. It may also be developed within intranet or internet scale according to the need of the research. The application provides researchers with a search engine for making a query against the XML based gene variance database. The request and response are transmitted in the background of the webpage by using AJAX (Asynchronous Javascript And XML) technology. The Search page provides an HTML form for users to interact with the web page. This form has three main elements: two drop down lists on the left and one value box on the right. The first drop down list is the search term list. This list contains the name of root elements and other elements that have child elements in four XML files. The second drop down list is the parameter list. The parameter list displays all the corresponding child elements of the selected element in the search term list. Once the selected value in the search term list is changed, the parameter list will also automatically update with the new corresponding child elements. Every time the selected value in parameter list is changed, the AJAX functionality will be invoked to send an XMLHttp request to a server for processing. The server will take the XML element name and return the values of all the elements having the same name. The Search page gets these values back from server and displays them in the value box. The AJAX function carries out this job without reloading the Search page. Figure 5 shows the XML values in the Value box after user selects a search term of Phenotypic_measurement and a parameter of Phenotypic_symptoms. When a particular search term and parameter are selected, the green box on the right of the webpage shows a list of the corresponding gene data found from the XML database. If these values are gene names such as: APP, ATM or BTK, researchers can query the gene information by clicking on one of the gene names shown in the Value box. The AJAX function sends an XMLHttp request with the selected XML value to the server. The server processes and sends a request of the selected gene name to the HGMD website and retrieves the data of this gene. A data extractor engines captures the set of data and sends it back to the Search page. Figure 5 and 6 show the results when a user clicks on the APP gene.

4. Discussion and Future Work

Implementing the schema in relational or object relational model has drawbacks when it is to be used in the domain of medical data. Sheu *et al.* (2000) highlighted the limitations of using relational model. He pointed out that SQL has limited scope to make query on data, for example, finding any particular data corresponding to any behaviour. Moreover, SQL requires the formal representation of data in table relational form which is not always possible for molecular biology data as data is not homogenised. Conceptualising molecular biology data is a major challenge for computer scientist as it requires capturing heterogeneous data distributed over many locations. So, problems can not be solved by homogenising the data structure as it will increase the problem in several folds. A partial solution of avoiding relational data model is to use object relational model to represent the laboratory data but it has its own drawbacks. It is not possible to describe multiple inheritance and it lacks describing complex data type such as sets, lists, *etc.* which are suitable for molecular biology data. Davidson (1999) highlighted that the schemas within the domain of molecular biology evolve rapidly in response to changing requirements and experimental techniques. Therefore, creating laboratory databases are far more challenging than creating databases for business environment. The laboratory data is irregular and in general they are incomplete. It also has the potential of rapid and unpredictable changes. The molecular biology data have to be web dependent so that it can have an immediate access and able to cope with disparate databases. This makes it almost impossible to fit in any proper schema model like relational or object oriented model. There are a number of database management systems that

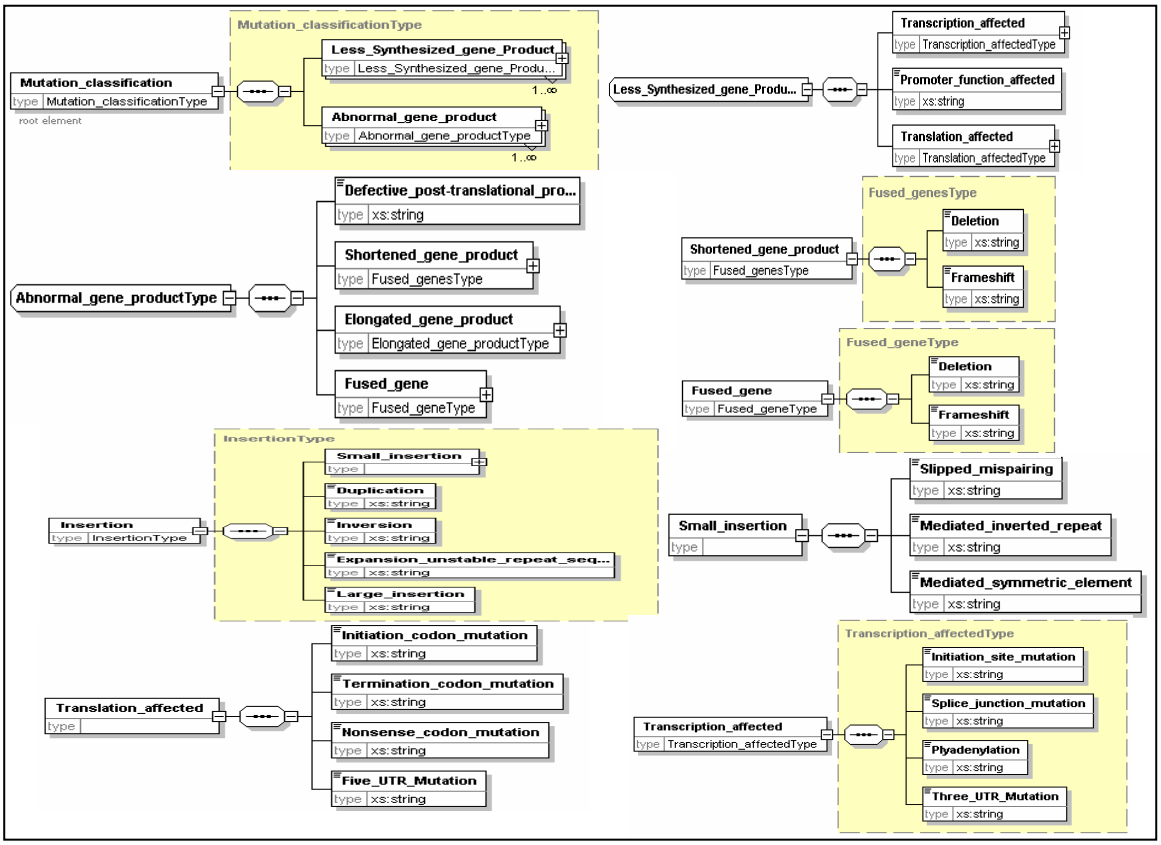

Figure 2. Gene mutation type data modeling

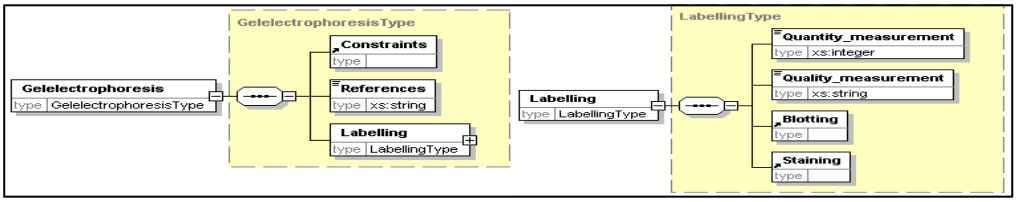

Figure 3. Empirical data modeling

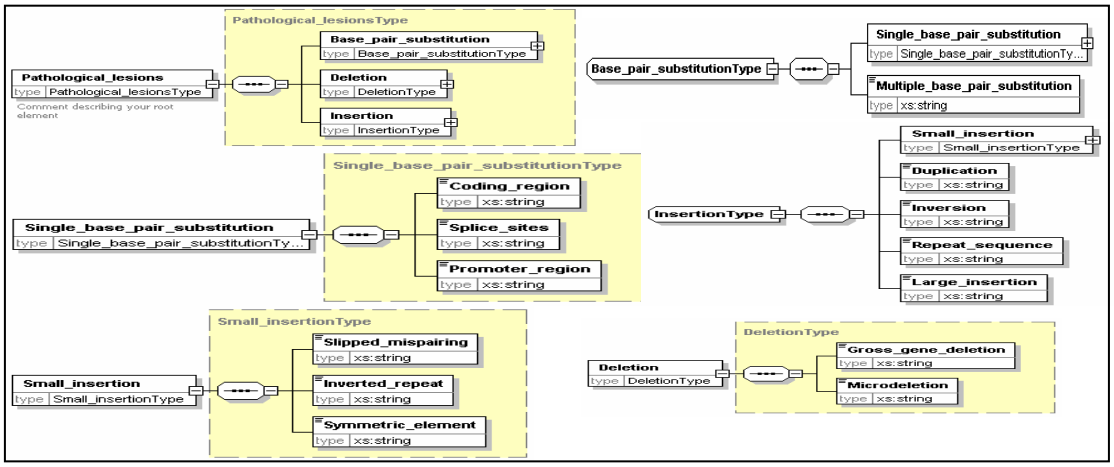

Figure 4. Pathological lesions data modeling

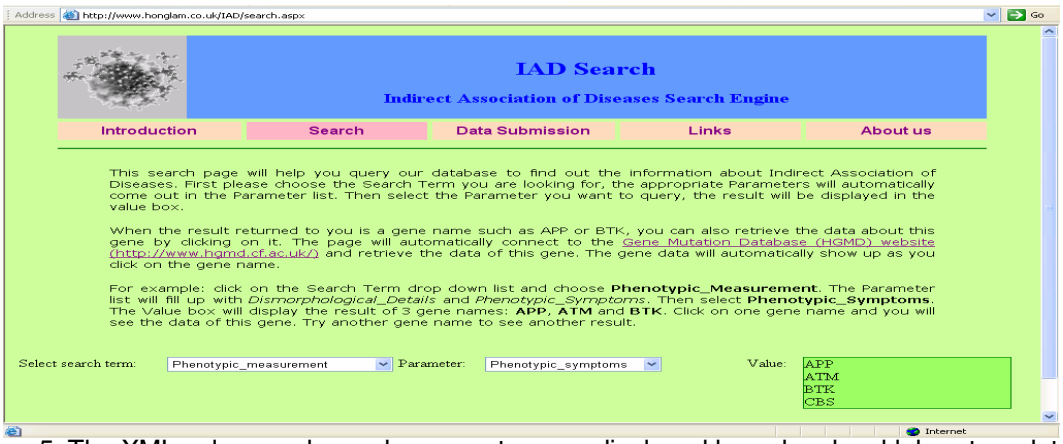

Figure 5: The XML values and search parameters are displayed based on local laboratory data

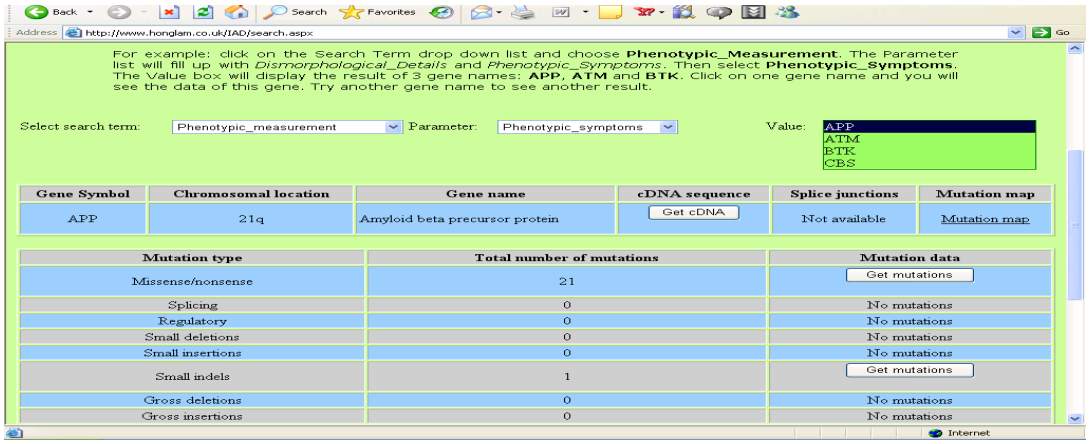

Figure 6: The data for APP gene are retrieved from public database

describe and manage the semi-structured data, *e.g.* LORE (Lightweight Object Repository) and XML. XML is the mostly used database management system for semi-structured data which allows designers to create their own customised tags to provide functionality. Considering these facts, this paper demonstrated how a laboratory based component database can initiate a search for specific gene data residing at different public domain databases. The search facilities will be extended further by providing automated navigations to other public domain medical database that will enable researchers to retrieve more meaningful information that can contribute to identify any indirect association of diseases. The research also aims to integrate a context based database integration to deal with unknown parameters and search terms.

4. References

[1] Davidson, S, Buneman P (et al.) (1999); Integrating biomedical data and analysis packages; In: Letovsky S, (ed). Bioinformatics databases and systems. Kluwer Academic Publisher; pp. 201-212.

[2] Donnelly, B., (2003); Data integration technologies: An unfulfilled revolution in the drug discovery process; Biosilico; vol. 1, no. May; pp 59-63.

[3] Geihs, K (2001); Middleware challenge ahead; In 5th International Workshop The, Internet Challenge: Technology and Applications; Berlin, Germany, Kluwer Publishers; pp. 24-31.

[4] Khan, N and Rahman, S., (2002); A cooperative environment for genetic variance analysis using component database for database integration; 15th IEEE CBMS.. IEEE computer society press; pp. 365.