Accessing Genomic Data through XML-based Remote Procedure Calls

Alberto Riva, Ph.D., Isaac S. Kohane, M.D., Ph.D.
Children’s Hospital Informatics Program
Children’s Hospital, Boston MA 02115
Alberto.Riva@TCH.Harvard.edu, Isaac.Kohane@Harvard.edu

Abstract

As the amount of data in public genomic databases grows, interoperability among them is becoming an increasingly critical feature. The ability for automated systems to mine and integrate data will be crucial to extracting knowledge from sources of data whose volume far exceeds the capabilities of human researchers. The currently dominant paradigm of presenting information as Web pages and using hyperlinks to describe relationships between pieces of information favors usability, but makes interoperability and automated data exchange more difficult. In this paper we describe how SNPper, a web-based system for the retrieval and analysis of Single Nucleotide Polymorphisms (SNPs), was augmented with a Remote Procedure Call interface, allowing client applications to query our program for SNP data and to receive the response as an XML document. Data represented in this form can be easily parsed by the requesting program, and thus reused for other applications. In this paper we describe the implementation of the interface and we show examples of its usage in a number of existing applications.

1 Introduction

One of the most notable byproducts of the Human Genome project and related efforts [1] is the exponentially fast growth of the amount of data stored in public biomedical databases. Information sources such as Genbank, LocusLink, PubMed and the Human Genome Browser are increasingly becoming essential tools for biological research, and the amount of data they contain far exceeds our ability to turn it into knowledge without the aid of automated tools. The challenge that the field of bioinformatics is now facing is the creation of automated knowledge discovery tools able to leverage on the available data in an effective way. A first step toward this aim can be accomplished by intelligently interconnecting different databases in order to discover new relationships between the data they contain. An example of this is represented by SNPper [2], a program developed by the authors to retrieve and display data on human genome polymorphisms, by integrating it with information coming from the Draft Human Genome site.

As is the case with most other bioinformatics databases, SNPper uses web pages as the presentation medium. This solution offers considerable benefits in terms of development, deployment and usability: web-based systems are only installed in a single location, can be easily upgraded with minimal disruptions to the users, and do not require special-purpose client software. On the other hand, HTML pages are well-suited to present information to human users, but not to distribute it to other applications. Relationships between different pieces of data are usually represented as hyperlinks, that do not contain enough semantic information to be useful to automated tools. Moreover, extracting data from HTML pages is relatively hard, since they do not obey to any standard format and can change without notice in significant ways. We believe that improved interoperability between sources of genomic information could be obtained by providing existing systems with the ability to communicate the data they own to other applications.

In this paper we describe the design, implementation and use of an XML-based interface to SNPper. The interface implements a Remote Pro-
procedure Call (RPC) system, allowing external applications to retrieve SNP data in a standardized, machine-readable format encoded as an XML document. In the next section we describe the SNPer system in greater detail, showing examples of its use. We then present the RPC interface by describing the calls it implements and the format of the corresponding XML documents, and we show examples of its use in real-world applications. We conclude with an evaluation of this tool and a comparison with similar features in other bioinformatics applications.

2 SNP analysis using SNPer

Single-Nucleotide Polymorphisms (SNPs) are the most common form of variation in the human genome. A SNP represents a single base change (e.g., A to G) at a specific genomic location. SNPs arise randomly during DNA replication, and are passed from one generation to the next. It is estimated that SNPs occur on the average once every 1,000 bases, although their frequency can vary considerably from one region of the genome to another.

SNPs are a very powerful tool in the hands of geneticists, since they are more abundant and easier to detect and to characterize than the other polymorphisms (deletions, microsatellites, etc.). While the majority of SNPs lie in non-coding regions of the genome, some of them fall in or close to a gene, and can produce variations in the amino acid sequence, in gene expression levels, in splicing patterns, etc. Overall, SNPs account for 90% of the inter-individual variability, and for as many as 100,000 amino acid differences [3]. SNPs are also responsible for a variety of genetic diseases, either alone or in combination with other genetic and environmental factors. For example, polymorphisms for gene APOE (Apolipoprotein E) are known to play a major role in the development of Alzheimer’s disease, as well as in some cardiovascular disorders [4, 5].

Even when a SNP produces no discernible effect per se, it may still be useful as a marker in the genetic sequence. Association studies look for a correlation between the presence of a phenotype and the alleles of one or more of a set of appropriately chosen SNPs, whose locations on the chromosomes are known with a sufficient degree of certainty [6]. If a high correlation can be established with one of the markers, this provides an indication that the location of the genetic cause of the phenotype is close to the marker [7], and may thereby lead to identifying the gene or genes involved in the underlying biological process. It is clear that the success of these studies depends on the availability of a sufficiently dense map of SNPs whose position is known and whose actual existence has been validated through repeated experiments [8]. A large effort is currently being undertaken to discover and map the highest possible number of human SNPs. SNPs are discovered by comparing overlapping genetic clones from different individuals, or by in silico matching of known overlapping sequences [9]. A number of public databases have been set up in order to store SNP data, the most important ones being the dbSNP Polymorphism Repository2 and the SNP Consortium3 [10]. These sites provide information about the nature of each polymorphism, its chromosomal location, the surrounding DNA region, the genetic clones it was found on, the date and location of its discovery.

SNPer is a web-based application to search for human SNPs in public databases, that relies on a local database built by combining information from dbSNP and goldenPath. Its main purpose

Figure 1: The SNPer page showing information about a gene. The top portion of the page contains general information about a gene; the remaining two boxes show the gene structure (number and position of its exons) and the SNPs belonging to that gene.
is to allow users to create sets of SNPs exhibiting some desired property by querying the database, and to efficiently manipulate such sets in various automated ways. A common way of using SNPper is to query for a gene by name; the program will then locate all the SNPs lying over it and present them as a "SNP set" together with general information about the gene. The SNP set can be displayed in a variety of formats (graphically, overlaid on the DNA sequence, as amino acid variations, etc.), can be refined or filtered (for example, by selecting only SNPs that fall on the gene exons, or only those that were validated by multiple observations), and can be saved or exported in multiple formats.

SNPper was built using a development environment for dynamical web-based applications. Each "page" displayed by the application is the result of running a function whose output is HTML code. These functions, in turn, rely on a lower layer that executes the required queries by connecting to a local relational database. The clean separation between the database layer and the presentation layer allowed us to uncouple them and to provide an alternative access to the former, as described in the next section.

3 The RPC Interface

The purpose of the RPC interface is to expose the set of queries implemented by SNPper to other applications. Instead of accessing the desired information through web pages, clients may submit a request encoded as a special-purpose URL to the web server that hosts SNPper, and receive an XML document containing equivalent information as the response. A query is represented as a URL having the following format:

/rpcserv/name?cmd=command&args...

where name is the name that will be assigned to the retrieved XML document, command is the name of the call being invoked, and args... represents the arguments to the call. For example, the following call is used to retrieve information about the MC4R (melanocortin-4 receptor) gene:

/rpcserv/MC4R.xml?cmd=geneinfo&name=MC4R

The above URL can be included as a link in a web page, or can be sent directly to the server as an HTTP request by a client program. In both cases, the server replies with the XML document shown in Figure 2. The whole document is enclosed in a GENEINFO tag, and is composed of several section. The first one shows the gene name, chromosome, strand and product. The next two sections show the position of the gene transcript and coding sequence as absolute genomic locations, using the coordinates provided by goldenPath. Next comes a section with Genbank accession numbers, followed by links to other databases (LocusLink and OMIM in this case). Finally, the number of SNPs known to lie on this gene is shown, and the document is closed.

We now list the current set of calls implemented by the SNPper RPC system, with the arguments and a brief explanation for each of them.

genelayout (name) - Returns the position of the gene specified by name and of all its exons, as absolute genomic locations.

geneinfo (name) - Return general information about the specified gene. Described in text.

snpinfo (id) - Returns information about an in-

Figure 2: The XML document showing general information about a gene.
Figure 3: The XML document showing information about an individual SNP.

dividual SNP. The data provided includes dbSNP and TSC identifiers, genomic location and alleles. For SNPs lying on genes, the document also indicates the gene name, the position of the SNP in the coding sequence and the amino acid change it causes, if any. Figure 3 shows an example relative to a SNP in the MC4R gene.

genesisinrange (chr start end) - Returns the list of all genes in the genomic region between start and end on the specified chromosome. The data provided for each gene includes its name, product, transcript location and number of SNPs.

snpsinrange (chr start end) - Returns the set of all the SNPs in the genomic region between start and end on the specified chromosome. The information provided for each SNP is the same that is returned by snpinfo.

genessnps (name) - Returns information on the set of all SNPs lying over the specified gene. The information provided for each SNP is the same that is returned by snpinfo.

4 Usage examples

The simplest application of the XML-based interface in SNPper is represented by the ability to download XML representations of most data objects displayed by the interface. In Figure 1, for example, each of the three subwindows in the page that displays information on a gene contains a link called “XMLXport”. The URL each of those links points to is the RPC call to retrieve the same data; the links can therefore be used to download and save XML representations of the data the user is viewing.

The main advantage provided by the above described RPC system is that it makes it possible for other applications to exploit the power of the sophisticated data mining and analysis methods implemented by SNPper. External sites that need to access SNP data can do so in a transparent way, and are guaranteed to always obtain the most up to date and correct information. In the remainder of this section we describe two applications that take advantage of this feature.

The Innate Immunity PGA project is a large-scale association study funded by the NHLBI, whose purpose is to determine the genetic causes of asthma and other inflammatory diseases. As part of this project, that represented one of the motivations for the development of SNPper, more than fifty genes are being sequenced for SNP discovery purposes. The newly discovered SNPs are stored in the SNPper database, but information about them needs to be displayed on the project web site4. The site uses the XML-based RPC interface for this purpose, querying SNPper for this data, parsing the resulting XML documents and reformatting them appropriately as HTML pages for viewing. The advantage of this solution is that the data is only stored in one location, and that the information presented on the site is guaranteed always to be up to date.

Figure 4 shows the output of another data integration application focused on SNP data. Given a

4http://innateimmunity.net/
gene name, the application retrieves variation data from a number of different databases, and presents it in a single page. Currently shown information is retrieved from PubMed, the Human Gene Mutation Database\(^5\), OMIM and SNPper. While integrating the first three databases requires parsing HTML pages and extracting the relevant portions, SNPper data is retrieved using the RPC interface. This makes the system faster, more reliable and immune to changes in the user interface.

In the first seven months after it was made available, the RPC interface has received more than 5,900 requests from 285 distinct IP addresses. While the majority of requests came from anonymous users, 135 registered users out of a total of over 700 made use of this facility.

5 Discussion and conclusions

XML is increasingly being used as the data format of choice for the exchange of data between applications. A natural extension of this concept is the idea of using XML to communicate the results of computations, such as database queries or more complex data manipulations. Our RPC system was inspired by XML-RPC\(^6\), one of the most successful efforts in this direction. Our system employs a different syntax, procedure calls are represented in a simpler way (for example, we do not specify the types of the argument) and the requests are sent using the HTTP GET methods instead of POST. Also note that, for simplicity, our XML documents do not yet conform to a standard DTD.

To our knowledge, SNPper (available at http://bio.chip.org/biotools) is the only public bioinformatic application using an XML-based RPC interface to export data in real-time. We believe that a wider adoption of the techniques described in this paper could lead to a much improved interoperability between biomedical data repositories, leading to the faster development of powerful data mining and automated knowledge discovery tools.

Acknowledgments

We wish to thank Ross Lazarus for his help in designing and testing the tool described in this paper. This work was supported by grant U01-HL66795 from the National Heart, Lung and Blood Institute of NIH.

References


