Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

DBCollHIV: a database system for collaborative HIV analysis in Brazil

Luciano V. Araújo[1], Marcelo A. Soares[2], Suelene M. Oliveira[3], Pedro Chequer[3], Amilcar Tanuri[2], Ester C. Sabino[4] and João E. Ferreira[1]

[1]Departamento de Ciência da Computação, Universidade de São Paulo, São Paulo, SP, Brasil
[2]Departamento de Genética, Universidade do Rio de Janeiro, Rio de Janeiro, RJ, Brasil
[3]Programa Nacional de DST/AIDS, Ministério Brasileiro de Saúde Pública, Brasília, DF, Brasil
[4]Hemocentro de São Paulo, São Paulo, SP, Brasil
Corresponding author: L.V. Araújo
E-mail: luciano@ime.usp.br

Genet. Mol. Res. 5 (1): 203-215 (2006)
Received January 10, 2006
Accepted February 17, 2006
Published March 31, 2006

ABSTRACT. We developed a database system for collaborative HIV analysis (DBCollHIV) in Brazil. The main purpose of our DBCollHIV project was to develop an HIV-integrated database system with analytical bioinformatics tools that would support the needs of Brazilian research groups for data storage and sequence analysis. Whenever authorized by the principal investigator, this system also allows the integration of data from different studies and/or the release of the data to the general public. The development of a database that combines sequences associated with clinical/epidemiological data is difficult without the active support of interdisciplinary investigators. A functional database that securely stores data and helps the investigator to manipulate their sequences before publication would be an attractive tool for investigators depositing their data and collaborating with other groups. DBCollHIV

allows investigators to manipulate their own datasets, as well as integrating molecular and clinical HIV data, in an innovative fashion.

Key words: HIV/AIDS, Relational database, HIV database, Resistance and subtype algorithms, Collaborative HIV database, Bioinformatics tools

INTRODUCTION

Understanding the genetic diversity of HIV-1 and its biological consequences is important for designing effective control strategies. The improvement of sequencing technology has greatly increased the capacity for generating sequence data. With the widespread use of antiretroviral compounds against HIV, virus drug resistance has also become an important issue. Genotype testing for HIV drug resistance has proved beneficial for treatment planning, and it is now considered a standard of care procedure for individuals failing antiretroviral treatment (Shafer, 2002). In Brazil, a network of laboratories (RENAGENO) has been organized to perform genotyping on patients failing therapy, and around 5,000 sequences are expected to be generated yearly. Other cohort studies that combine clinical data with viral genome sequences are under way. Databases that manage sequences together with annotation are available for HIV (http://hiv-web.lanl.gov; http://hivdb.stanford.edu) (Kuiken et al., 2003) and can be accessed through web interfaces. These databases are extremely useful for extracting reference sequence alignments. Nevertheless, access to sequence information and corresponding clinical data on patients is still very limited. Furthermore, an important feature for the investigator is to be able to manipulate his/her raw data before publication. At the moment, no freely available database can manage sequences and HIV clinical data integrated with analysis tools prior to publication. The development of a database system that is able to handle and search sequences produced locally and to integrate them with epidemiological and clinical data and bioinformatics tools, would provide a major advantage for research groups in developing countries that do not have the necessary resources to develop their own systems.

We present the database system for collaborative HIV analysis (DBCollHIV), which was developed to manage sequence, clinical, epidemiological, and treatment data generated by ongoing HIV studies in Brazil (http://clinmaldb.usp.br/dbcollhiv).

MATERIAL AND METHODS

Presently, scientists face the challenge of integrating biological data stored over the years in dynamic heterogeneous structures. Many genomic and molecular biology studies have applied a number of computational tools, generating very large databases with different input and output formats, without common standards. The advent of digital archives has created a science and engineering data avalanche. The Genome Project (www.ornl.gov/TechResources/Human_Genome/home.html) and the CERN Large Hadron Collider (http://lhc-new-homepage.web.cern.ch/lhc-new-homepage) are two examples of very large scientific data sets.

There are two conventional approaches for solving this problem. One uses only a relational database to store the data sets (Thakar et al., 2003). Another uses an object-oriented database to integrate and store these same data sets (Cornell et al., 2003).

An important contribution of this type of research is to integrate the heterogeneous HIV analysis applications with the clinical database. In this research, an analytical application is considered an implementation of a computational algorithm that executes some semantic operation on biological data.

A way to allow the development of flexible systems for integrating heterogeneous applications is through construction of modular systems. In these systems, each unit of computation chooses the modules that suit it, being able to choose different versions of a module for each function. Considering the phase of a system project divided into a functional project and a database project, the modularity is treated by the functional project of the system, generating modules composed of transactions that will operate directly in the database. However, the database project follows, considering a global data outline to be implemented through the use of a data repository common to all the modules of the system.

In the case of an application composed of several subsystems, the existence of a single data repository for all these subsystems means that, independent of which parts of the database are to be accessed for each subsystem, the data are stored in a monolithic way. This causes autonomy loss on the part of the subsystems that are involved. To attain autonomy of data administration of the subsystems that compose an application, it is also necessary modularize the database, ensuring that each generated module will have its own data repository, which will only contain the data that are relevant for its transactions.

The basic idea of the modularization of databases is the breakdown of the global application data outline into subschema. The subschema intersection characterizes the sharing of portions of the global data outline, which is also contemplated by the guidelines proposed for the database modularization project. The development of DBCollHIV followed the results of the research made by Ferreira and Busichia (1999) and Barrera et al. (2004).

The proposed architecture includes the advantages of the client-server and distributed architectures, and supports the fundamental requirements of heterogeneous data and a flexible database. This architecture allows the database to be incrementally extended by adding new modules. The primary database is presently composed of four modules: three for data storage (Patient, Virus, and Disease) and one module for data query (User access control). DBCollHIV has three kinds of inter-operability and consulting procedures: database queries (by SQL commands or graphical interfaces), retrieval functions (by C++ and Java language) and web access.

The main advantage of a DBCollHIV system is the automation of analysis tasks for which clinical and molecular data are requested. A file generated from one step can be used in the next step, even if it has some structural incompatibility, solved by an XML file translation (Achard et al., 2001).

A generic DBCollHIV system integration is based on the Genflow approach (Oikawa et al., 2004). In this approach, the applications P_i (i = 1,2, ... n; where *n* is the number of applications installed and available in the operating system) are placed in sequential order, according to their task. This property is defined when P_i is installed in the environment, according to its algorithm running parameters. It defines precedence relations among other applications.

Each P_i application is associated with an execution step E_k (k = 1,2, ..., m; where *m* is the number of tasks for a particular workflow). The chains are built in series and run using input/output files.

All applications follow the precedence order and present semantic compatibility of their input and output files; they are called chained. In Figure 1, we can see some kinds of chaining applications, such as P_1 and P_3, P_2 and P_4, P_3 and P_4.

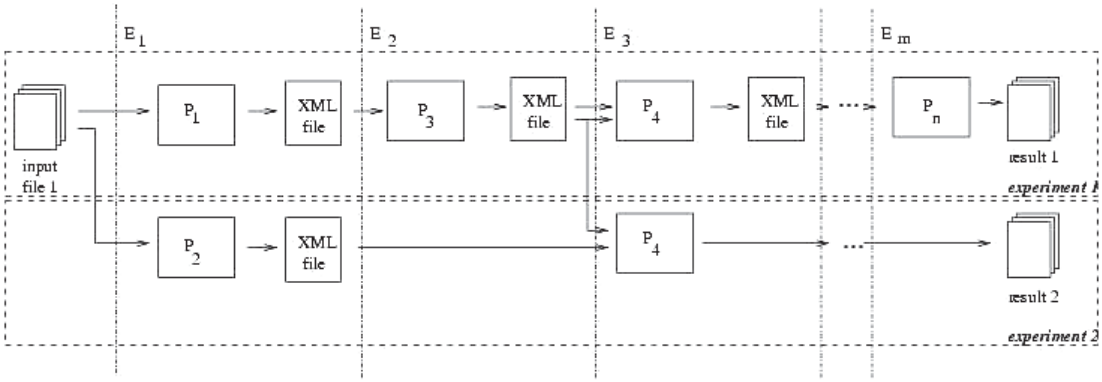

Figure 1. Scheme of integration used by DBCollHIV.

DBCollHIV has two working environments: public and private. In the private environment, the principal investigators and the co-investigators are registered and obtain individual logins to access the system. The data are organized into independent studies; the principal investigators define the studies and the role of the other investigators in accessing each of them through a specific interface. Within this scenario, the investigator relies on the private environment for storing and consulting his/her own data. Integration with other studies can be automatically done whenever convenient and properly authorized. The public DBCollHIV environment supports the access of the data that the principal investigators have released for the general public.

Additionally, the DBCollHIV was developed using free software, such as Apache Web Server, Fast CGI, Postgress database, Perl, Ruby, and Ruby on the rails. The DBCollHIV can also exchange data with other databases and with bioinformatics tools, using XML, Fasta and text files.

The DBCollHIV is available at http://clinmaldb.usp.br/dbcollhiv, where the user can access the demonstration area to evaluate it in a demonstration environment. DBCollHIV facilitates the possibility that researchers interested in HIV share the same tools and clinical database systems, while preserving security, through a control access model, of their private research data. The main idea is to reduce the efforts for developing new tools, data storing and new versions. Consequently, the focus of DBCollHIV is on a single-server site to promote collaboration. DBCollHIV proposely does not stimulate the creation of many DBCollHIV server sites. However, if users would like to construct their own DBCollHIV server site, they can obtain orientation on local installation from the DBCollHIV website.

RESULTS

The results are shown through two sections: Database System and Tools of Bioinformatics.

Description of the database system for collaborative HIV analysis

DBCollHIV was projected for storing information on the clinical evolution of HIV patients. A patient can have a set of several samples. Each sample can be used for distinct exams and for obtaining distinct virus sequences. In cases where the study does not have data covering the four registered areas, data can be partially stored. For instance, sequence data can be linked to a patient even if the sample data are not available. The interaction of the different modules is depicted in Figure 2.

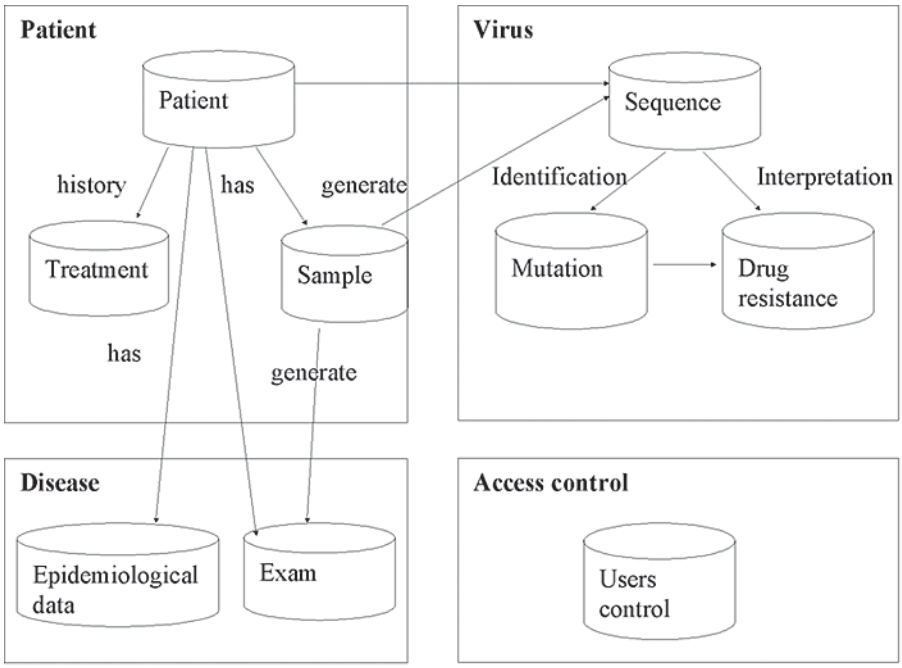

Figure 2. Modules of HIV DBCollHIV and their interactions.

The primary data of each module are described below:

- Patient - Project identification, Alternative identification, Site of origin (Country, Region and State), Gender, Birth date, Date of HIV serodiagnosis, Date of last seronegative result, Detuned test date, Detuned test result, Detuned test window length, indication of AIDS case, Co-infections.
- Sample - Sample label, Sample date, Freezer number, Location in freezer, Box num-

ber, Sample type, Body compartment, Drug exposure, Therapeutical fail.

- Exam - Exam label, Exam date, Exam type, Absolute value, Leukocytes, Whole white cell count, Percent value.
- Treatment - Start and end date of an antiretroviral treatment, Reason for treatment change, Drugs used.
- Sequence - Sequence name, Sample label, Genomic region, Subtype, Accession number, Sequence size, Sequence type, Sequence in FASTA format.

Within the data query interfaces, the investigator can choose which clinical variables and sequence he/she wants to obtain from the database in a user-friendly window (Figure 3). Clinical data can be selected in a text format to be exported for subsequent analysis in a statistical package. Sequences can be retrieved in FASTA format.

Bioinformatics tools

Tools were developed to support rapid sequence analyses. These tools are described below.

PCR contamination tools

It is very important for each laboratory to evaluate whether there was any error in the process of obtaining sequences, such as PCR carryover or mixed samples (Learn et al., 1996). A tool was developed (Figure 4) so that each time a sequence is loaded into DBCollHIV, the software pairwise aligns it to all previous sequences from that laboratory and defines the similarity rate of each sequence pair. A pair of sequences with a similarity higher than a user-defined threshold (e.g., 99%) suggests that PCR contamination has occurred, and this is pointed out by the software (Araújo et al., 2004) (http://clinmaldb.usp.br:8083/hiv/contaminacao2/contaminacao.html).

Drug resistance tool

Rules for the interpretation of drug resistance were created by the Brazilian Ministry of Health RENAGENO Expert Committee (http://www.aids.gov.br/final/tratamento/politicas/projeto_renageno.htm), and they were used to develop an algorithm to automate the process (Brindeiro et al., 2004). This tool is integrated into DBCollHIV, so that reverse transcriptase or protease sequences can be automatically analyzed (Figure 5). An output file is created that allows the researcher to release the results to the physician. The results obtained are saved in the database as mutations that differ from the HXB2 HIV-1 reference strain. In addition, predicted resistance to each available antiretroviral drug is shown. If algorithm rules are changed during the course of the study, an upgraded version of the program is released, allowing the investigators to easily reanalyze their entire data set (http://clinmaldb.usp.br:8083/hiv/resistencia/resistencia.html).

Automated HIV circulating recombinant form tool

This tool uses RIP (Siepel et al., 1995) for large-scale analyses of sequence recombi-

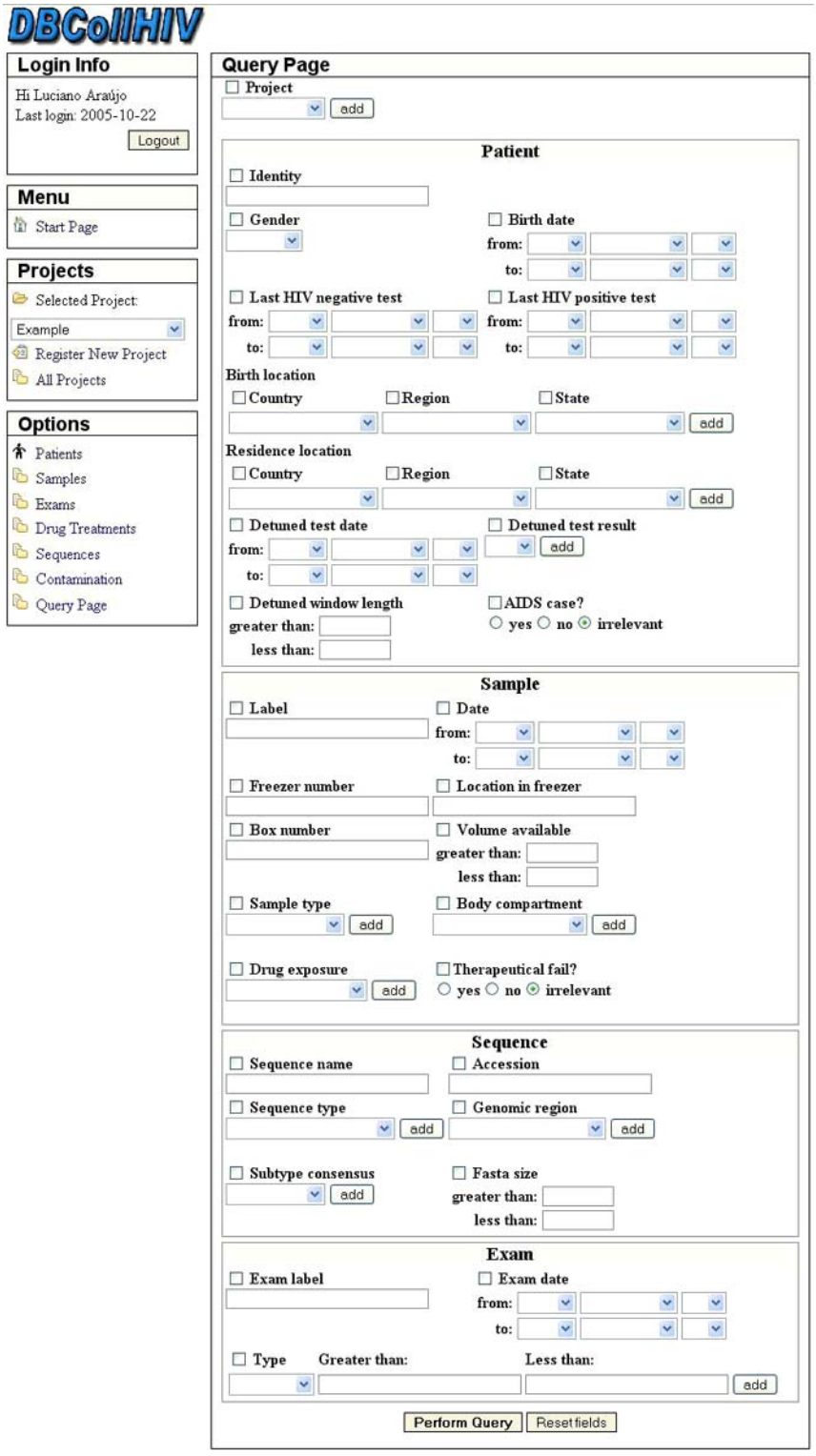

Figure 3. Interface for HIV data queries in DBCollHIV.

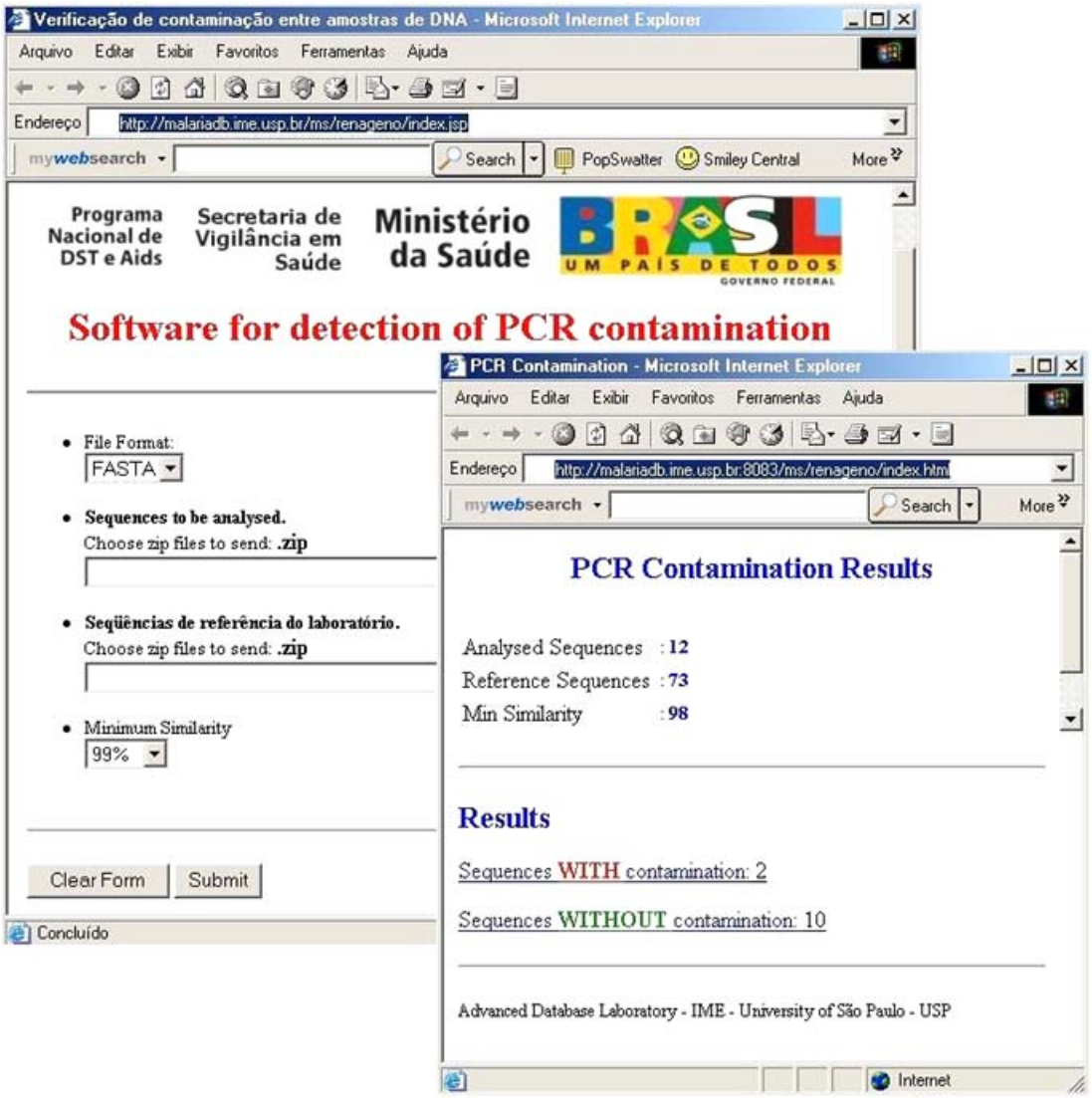

Figure 4. PCR contamination tool integrated into DBCollHIV.

nant forms (Figure 6). These sequences are stored in DBCollHIV, and their analytical results are used by the Automated HIV subtyping tool (http://clinmaldb.usp.br:8083/hiv/crf/crf.html).

Automated HIV subtyping tool

Two different programs, Blast (Altschul et al., 1999) and RIP (Siepel et al., 1995), are used to automatically subtype HIV strains (Figure 7). The investigator can visualize the results obtained by each program and decide which sequences require a more detailed analysis for HIV subtype assignment (http://clinmaldb.usp.br:8083/hiv/blastRip/blastRip.html).

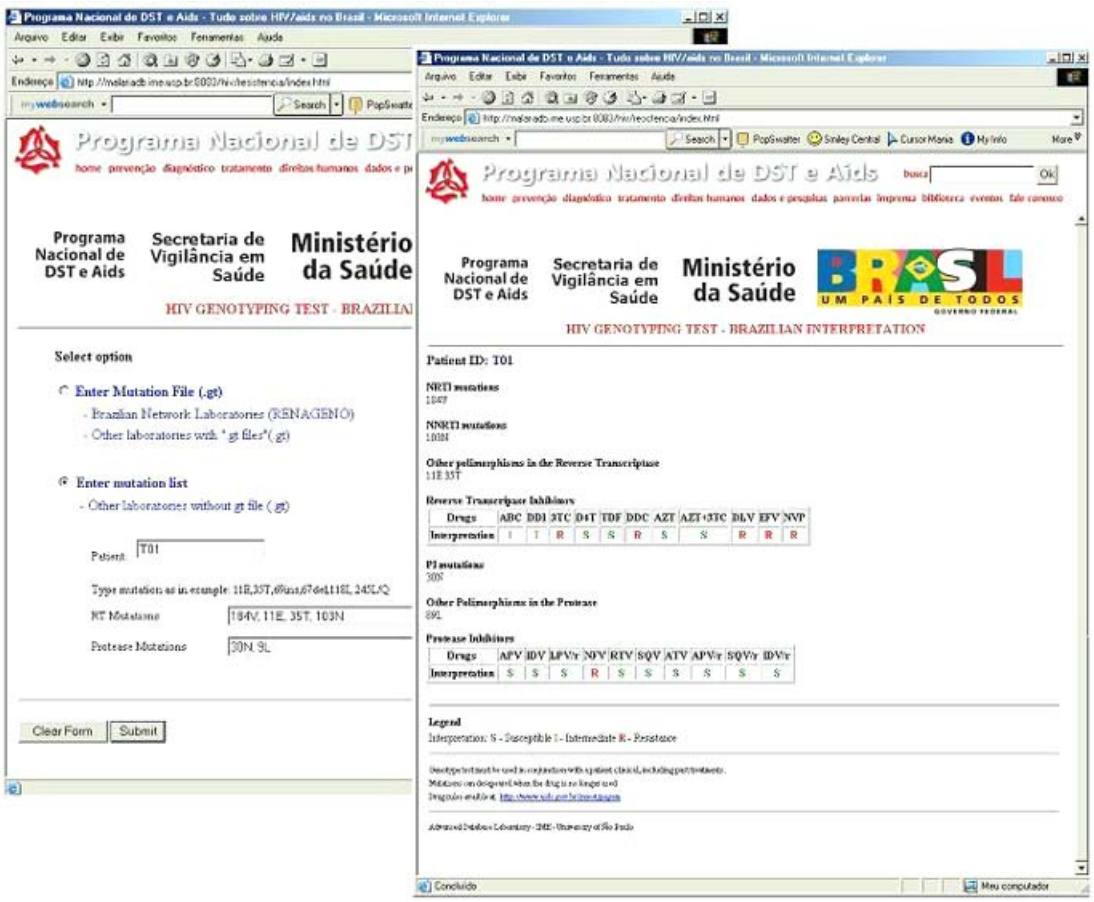

Figure 5. HIV drug resistance interpretation tool integrated into DBCollHIV.

Example of DBCollHIV results

The DBCollHIV database was used to identify subtype and drug resistance pattern in two sets of samples from the Blood Center of São Paulo (Barreto et al., 2006) and from the São Paulo State AIDS Program (Kalmar et al., 2005). We show here an example of an automated analysis performed on a subset of the samples sequenced at the São Paulo Blood Center for the Brazilian Network of HIV Drug Resistance (RENAGENO).

One hundred and sixty-eight samples from individuals under treatment in Campinas and surrounding cities in São Paulo State were selected and submitted to drug resistance and subtype analysis. Among the 168 samples, 160 (95%) were automatically subtyped (135 subtype B, 9 subtype F, 2 subtype C, and 14 circulating recombinant forms, subtype BF). The remaining eight samples needed manual revision for subtype definition. The samples were also analyzed for drug resistance according to the Brazilian algorithm. Figures 8 and 9 show the proportion of samples resistant to protease and reverse transcriptase inhibitors. This demonstrates the capacity to analyze samples according to a specific epidemiological characteristic, which in this example are samples collected from one region of the country.

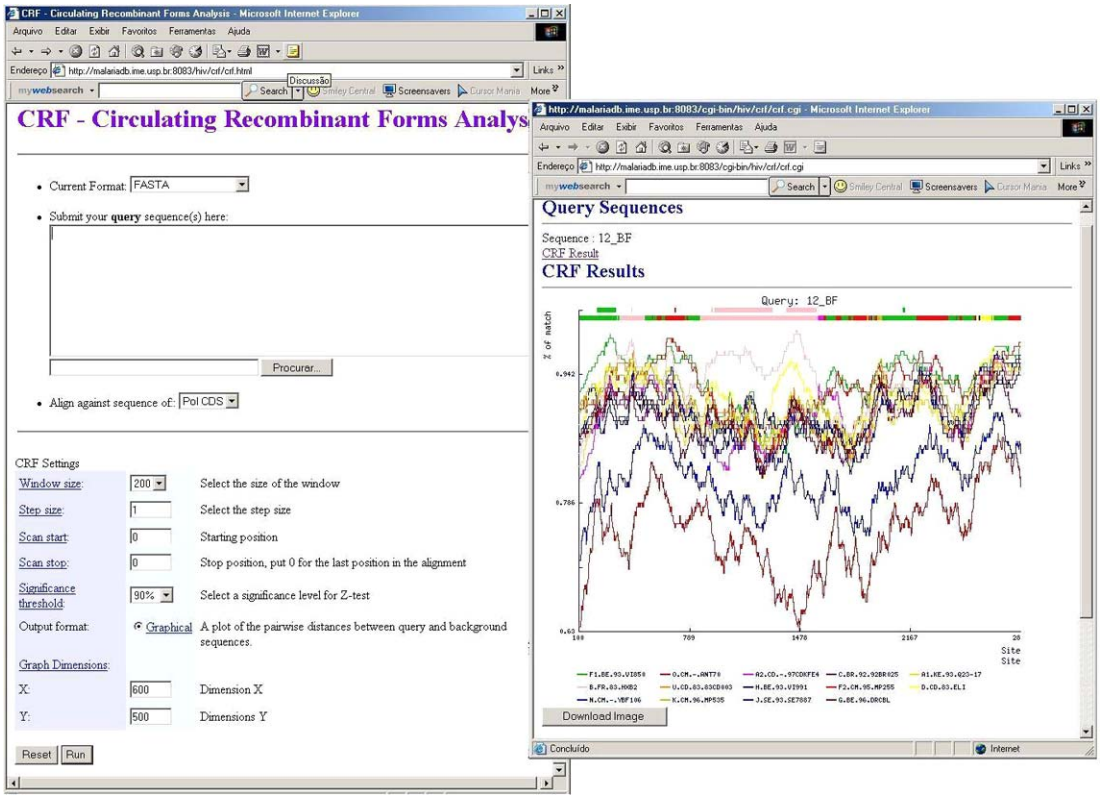

Figure 6. Circulating recombinant form tool integrated into DBCollHIV.

DISCUSSION

Any database that will be useful for research purposes should not be limited to fixed data entries and data processing. It is always necessary to add new types of data and tools to the database. Modularization is an important aspect that can overcome this problem, and it was a key element in the development of DBCollHIV. This database is different from conventional data-modeling cases, in which the object modeling of the target objects is known, such as people, products, invoices, etc. For example, when it is desired to understand the behavior of a virus in its host, a very large experimental data set is generated, including sequence and epidemiological data, clinical exams, results of analysis, etc. After the primary data are stored, it is possible to support many types of data selection and data analysis tools.

DBCollHIV was developed to support HIV research groups in Brazil that do not have the expertise or the necessary funding to develop their own database systems. It will help them to store their data adequately. Since the pattern of the primary data is common to all groups using the system, the integration of sequences from different studies in a laboratory or from different groups can easily be achieved. This will help speed the accessing of sequences for specific requests. One example is the HIV subtype F clade, which includes 10% of the strains in Brazil (Brindeiro et al., 2003). If one were interested in understanding the mutation pattern in subtype F strains that are associated with a specific protease inhibitor, it would be necessary to

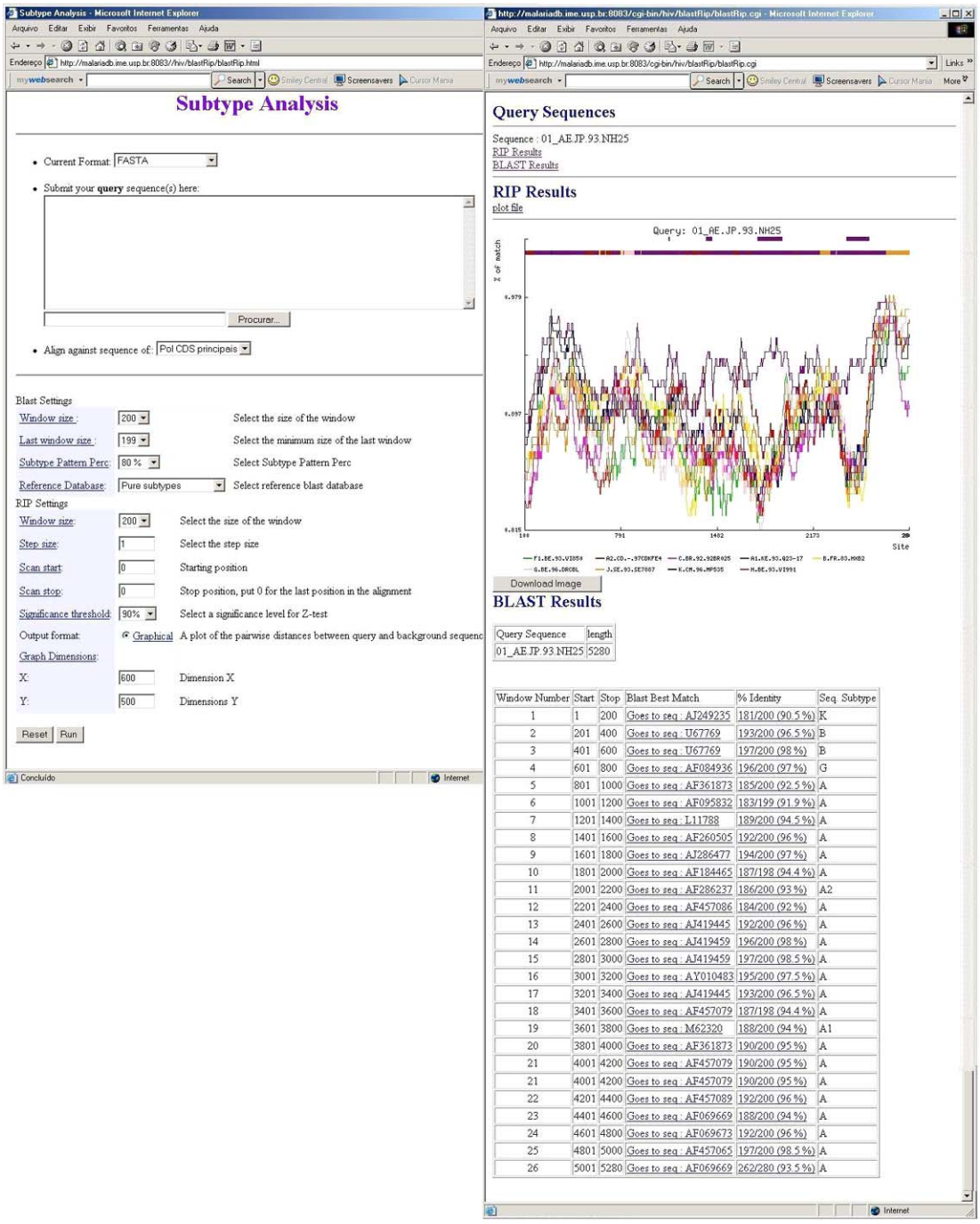

Figure 7. HIV subtype tool integrated into DBCollHIV.

sequence a large set of samples. In a cooperative integrated database, subtype F sequences annotated with drug history data could be easily obtained with DBCollHIV.

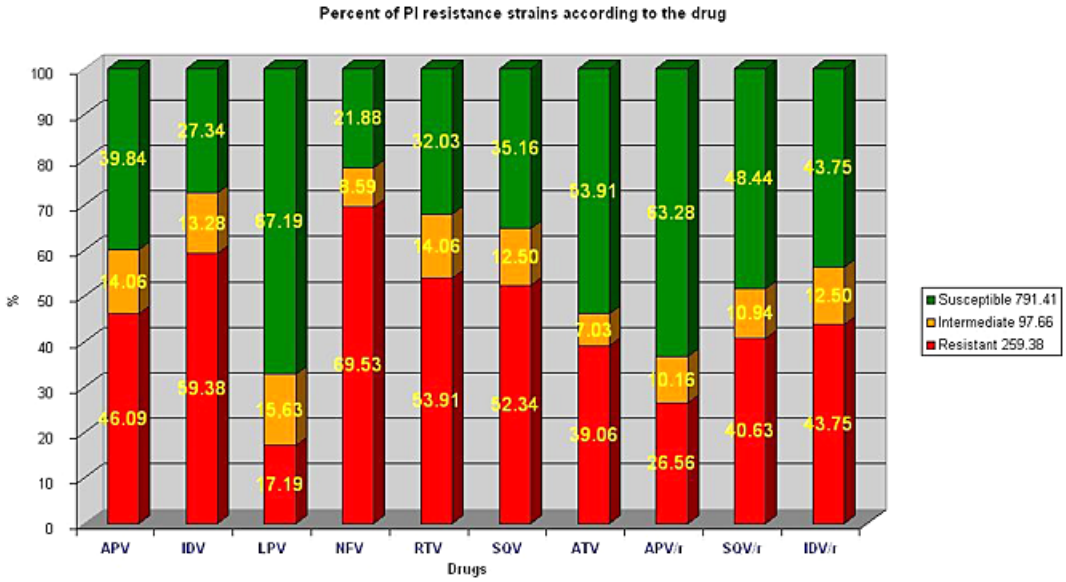

Figure 8. Percent of samples resistant to protease inhibitors (PI).

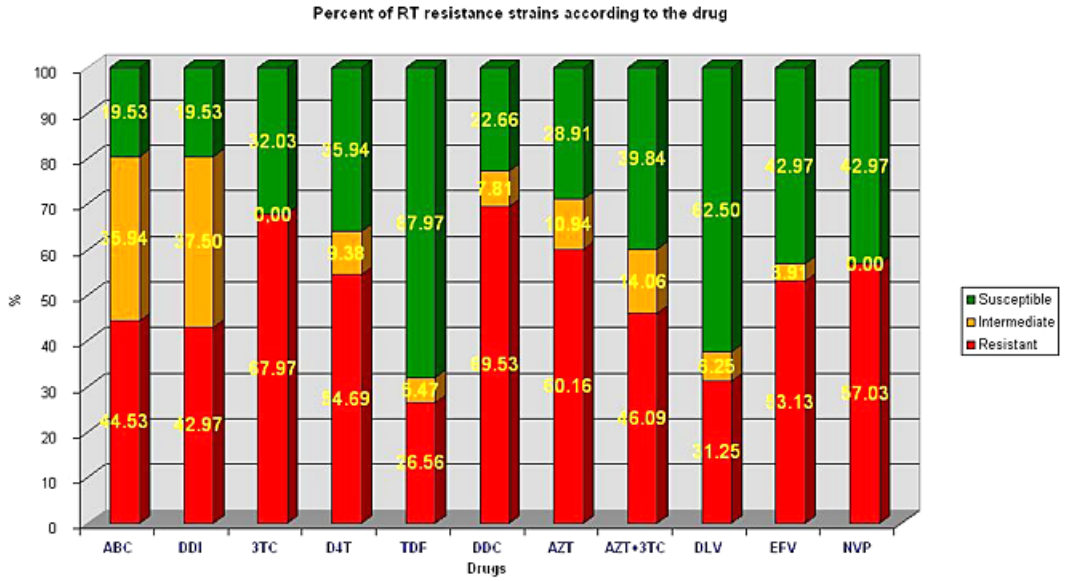

Figure 9. Percent of samples resistant to reverse transcriptase (RT) inhibitors.

Differently from competing with well-established public sequence databases, the aim of this particular database is to allow researchers to manipulate their own datasets locally, and to develop tools that automatically transfer clinical and epidemiological data from our system to other public databases. We believe that DBCollHIV will fill an important gap, in particular in resource-limited settings, such as in Brazil.

ACKNOWLEDGMENTS

We acknowledge Dr. Bette Korber and Dr. Brian Gaschen for supporting the development of the software interfaces to determine HIV subtype and to detect sequence contamination. We also thank Dr. Rodrigo M. Brindeiro and Dr. Ricardo S. Diaz of the Brazilian Genotyping Committee for providing the HIV drug resistance interpretation rules used in the algorithm, and we thank Ronie Uliana and Adriano Dadario for helping in the development of this database system. Research supported by the Brazilian Ministry of Health AIDS/STD Program (grants No. CFA869/02 and CFA167/03) and by the São Paulo State Science Foundation (grant No. CAGE 99/073900).

REFERENCES

Achard F, Vaysseix G and Barillot E (2001). XML, bioinformatics and data integration. *Bioinformatics* 17: 115-125.

Altschul SF, Gish W, Miller W, Myers EW, et al. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215: 403-410.

Araújo LV, Sabino EC, Brindeiro RM and Tanuri A (2004). Bioinformatic tools for HIV-1 sequences developed for the Brazilian STD/AIDS program network for genotype testing. In: MedGenMed. XV International AIDS Conference, Thailand, Vol. 11, p. MoPeB3125.

Barrera J, Cesar Jr RM, Ferreira JE and Gubitoso MD (2004). An environment for knowledge discovery in biology. *Comput. Biol. Med.* 34: 427-447.

Barreto CC, Nishyia A, Araujo LV, Ferreira JE, et al. (2006). Trends in antiretroviral drug resistance and clade distributions among HIV-1-infected blood donors in São Paulo, Brazil. *J. Acquir. Immune Defic. Syndr.* 41: 338-341.

Brindeiro RM, Diaz RS, Sabino EC, Morgado MG, et al. (2003). Brazilian Network for HIV drug resistance surveillance (HIV-BResNet): a survey of chronically infected individuals. *AIDS* 17: 1063-1069.

Brindeiro RM, Diaz RS, Sabino EC, Araújo LV, et al. (2004). Implementation of the quality control program (QC) for HIV-1 resistance genotyping testing network - RENAGENO of Brazilian STD/AIDS program. In: MedGenMed. The XV International AIDS Conference, Thailand, Vol. 11, p. MoPeB3126.

Cornell M, Paton NW, Hedeler C, Kirby P, et al. (2003). GIMS: an integrated data storage and analysis environment for genomic and functional data. *Yeast* 20: 1291-1306.

Ferreira JE and Busichia G (1999). Database modularization design for the construction of flexible information systems. Proceedings IEEE for the IDEAS99, Montreal, Canada, pp. 415-422.

Kalmar EMN, Chen S, Ferreira S, Barreto CC, et al. (2005). Drug resistance among HIV patients who discontinued ARV treatment in Brazil. 3rd International AIDS Society Conference on HIV Pathogenesis and Treatment, Rio de Janeiro, Brazil, p. WePe4.4.C13.

Kuiken C, Korber B and Shafer RW (2003). HIV sequence databases. *AIDS Rev.* 5: 52-61.

Learn Jr GH, Korber BT, Foley B, Hahn BH, et al. (1996). Maintaining the integrity of human immunodeficiency virus sequence databases. *J. Virol.* 70: 5720-5730.

Oikawa MK, Broinizi ME, Dermagos A, Armelin HA, et al. (2004). Genflow: generic flow for integration, management and analysis of molecular biology data. *GMR* 27: 690-697.

Shafer RW (2002). Genotypic testing for human immunodeficiency virus type 1 drug resistance. *Clin. Microbiol. Rev.* 15: 247-277.

Siepel AC, Halpern AL, MacKen C and Korber BT (1995). A computer program designed to screen rapidly for HIV type 1 intersubtype recombinant sequences. *AIDS Res. Hum. Retroviruses* 11: 1413-1416.

Thakar A, Szalay A, Cern PK and Gray J (2003). Migrating a multiterabyte archive from object to relational databases. In IEEE (Instit. Electr. Electron. Eng.) Trans. Biomed. Eng., pp. 16-29.