HIVbase: a PC/Windows-based software offering storage and querying power for locally held HIV-1 genetic, experimental and clinical data

Susanna Lamers¹, Scott Beason¹, Luke Dunlap¹, Robert Compton¹ and Marco Salemi²,*

¹Gene Johnson Inc., St Augustine, FL, USA and ²Department of Ecology and Evolutionary Biology, University of California at Irvine, Irvine, CA, USA

Received on August 20, 2003; accepted on September 17, 2003
Advance Access Publication January 22, 2004

ABSTRACT

Background: Human immunodeficiency virus (HIV) research involves ongoing, repetitious sequencing of the HIV genome and the massive accumulation of associated investigational data. As a result, the storage of annotated DNA and/or protein sequences, as well as information retrieval, have become increasingly difficult tasks, with scientists extracting less information from their collected data than they should.

Objectives: Our objective was to design and develop a software package to aid researchers in the storage, analysis and exploration of their HIV-associated data.

Results: HIVbase contains familiar, easy-to-use interfaces and functionality for integrating many types of disparate data. The software contains tools that allow for the mass import of raw genetic data, eliminate repetitious sequence translations, have the ability to identify automatically and store HIV regions of interest from nucleic acid or protein sequences, allow for the export of data in commonly used analysis-ready formats, and for unique querying approaches.

Contact: msalemi68@hotmail.com

INTRODUCTION

Because of the high mutation and evolutionary rate of human immunodeficiency virus (HIV) (Mansky, 1996; Korber et al., 1998), and the impact that this variation has on treatment and studies concerning the basic epidemiology of the virus (Sansom and Wlodawer, 2001), HIV research requires repetitious sequencing of viral genomes (Leitner et al., 1993). It is, therefore, not uncommon for the typical HIV molecular biologist or molecular evolutionist to struggle with increasingly large datasets. Such information overload, coupled with inadequate data capture, analysis and correlation, may cause scientists to overlook vital molecular processes that would otherwise lead to discovery.

Researchers usually design and build ad hoc applications in the form of spreadsheets and databases to handle their large datasets, although they often lack the necessary programming skills to implement a highly efficient data storage/retrieval system on their local computer. Excel and Access are relational data management systems (RDMS) and highly sophisticated tools, yet they clearly do not fulfill the requirements of the modern HIV molecular investigator.

Several public HIV databases, such as the Los Alamos Sequence Database (http://hiv-web.lanl.gov), exist for published annotated sequences, but they cannot help organize data from any specific laboratory at the local level, and they do not possess user-specific functionality. For these reasons, we decided to develop a software solution for local computers, with user-friendly storage, annotation and querying capabilities that would help researchers effectively manage DNA/amino acid sequences, and related genetic and clinical data.

SYSTEM AND METHODS

We decided to use Microsoft Windows as the backbone for HIVbase, because this operating system is commonly implemented in most molecular biology research laboratories, and because the non-expert user is frequently familiar with its interface.

The software was created employing contemporary Microsoft development tools. The data foundation was built with Microsoft SQL 2000 Desktop Engine (MSDE) and core SQL Server technology. MSDE is a reliable storage engine and query processor for desktop extensions of enterprise applications and allows for almost unlimited querying ability. MSDE can be built into many applications and redistributed royalty-free with other Microsoft development tools.

*To whom correspondence should be addressed.
HIVbase: a local database software for HIV-1 data

Fig. 1. HIVbase Data Querying tool. The top left window lists the aspects of the data that the user wants to view in the resulting query. In the top right, the conditions of the query are set. In this example, the user wants to know or view the sequence name, domain name, region charge, if the motif GPGQ exists, what amino acid occurs at position 9, how many N-T exist (where the period indicates any character), if co-receptor usage information has been entered for the sequence, and amino acid data. The query is further defined by the conditions: domain name equals V3 and region charge must be greater than three. Results from the query appear in the bottom window. The query results can be further organized and re-ordered using familiar MS-like tools.

Microsoft .NET was used to build the Windows HIVbase software environment. HIVbase was written in C#; however, .NET can incorporate many computer languages, including some of the older languages like Fortran. This will be useful as we incorporate other pre-defined and established algorithms into our HIVbase foundation. The .NET framework provided us with many advantages, including: (1) reliability, (2) enhanced performance, (3) improved productivity, (4) powerful security, (5) the ability to integrate with existing software, (6) ease of deployment, (7) native XML support and (8) flexible data access.

A functional beta version of the software is available to researchers at http://68.16.158.182/genejohnson/website/install/hivbase.zip

HIVbase

HIVbase contains three major features to help researchers manage their data: (1) a Data Entry tool, (2) a Annotation Definition tool and (3) a Data Query tool.

The Data Entry tool accepts genetic sequences as individual plain text files or in FASTA format. The user can import nucleic acid or amino acid sequences. When nucleic acid sequences are imported, they are sent through a translation system that will automatically identify HIV regions of interest as well as all major HIV proteins. The system translate nucleotide sequences in all possible six frames and identify HIV proteins by using conserved regions at the 5′ and 3′ end of each open reading frame as anchors. When amino acid sequences are entered, the HIV proteins are also identified using the same anchors. The nucleic acid coding regions and proteins are related to the raw data within the system. At this stage, user-defined annotations can be attached to the new sequences.

The Annotation Definition tool allows users to define annotations as text, an integer, a true/false statement or a decimal. Values for each annotation can also be defined. For example, if an user defines the annotation ‘protease inhibitors’, they can further set the value as ‘restricted text’. The user then, can generate a list of protease inhibitors to choose from. During data entry, this list of protease inhibitors appears in a drop down menu. This reduces data entry time and incorrect data entries. The ability to define annotations with great precision, allows for complex searching ability using our query tool. For example, if an annotation is set as an ‘integer’, the user can query datasets based on conditions such as ‘equal to’, ‘greater than’, and ‘less than’.

As illustrated in Figure 1, the Data Query tool provides an efficient way to analyze and retrieve organized data. Annotated sequences within HIVbase can be retrieved by combining in a search different characteristics (annotations) with the logical operators. In other words, users can retrieve many combinations of data based on their own pre-defined annotation lists. In addition to existing annotation values, the
user can also view the results of special ‘functions’ run against their sequence data. These functions include: counts of patterns, if a certain genetic motif exists, or which amino acid or nucleic acid occurs at a specific position. Results from a query session can be retrieved and formatted as a FASTA file or Excel spreadsheet for subsequent analysis. Also, when using the query tool, users can edit sequence data and associated annotations.

DISCUSSION
HIVbase is a user-friendly interface allowing efficient HIV molecular data storing and annotation on any PC/Windows-based local computer. As shown in the examples above, the querying tool allows users to formulate questions that can be applied towards thousands of internally held sequences. Large amounts of genetic data can be retrieved based on specific genetic characteristics or annotations, and unwieldy raw data transformed into formats that facilitate diagnosis and discovery. In addition, future versions of the software will incorporate new analysis algorithms that can be applied to the data held within HIVbase. Our next goals include the implementation of specific tools for DNA and protein multiple alignments, phylogenetic analysis, automatic HIV subtyping, analysis of recombinant viruses, epitopes and glycosylation sites mapping, and RNA and protein secondary structure prediction.

ACKNOWLEDGEMENTS
We thank Dr. Maureen Goodenow for her support throughout this project. Dr. Michael McGrath for comments and suggestions, and Dr. Walter Fitch for critiquing the manuscript. S.L. was supported through NIH grants R01 HD32259 and AI47723. M.S. was supported by a postdoctoral grant from the Flemish Funds voor Wetenschappelijk Onderzoek. Gene Johnson Inc. was supported by SBIR grants R43 AI56904-01 (NIH) and 0349669 (NSF). We would also like to thank Annie Ferran, John P. Lamers, and Linda S. Dunlap for their immeasurable contributions and encouragement during the formation of Gene Johnson Inc.

REFERENCES