TECHVIEW
SOFTWARE

Sequence Server Samurai

Pavel Simakov

InforMax Software Solution for Bio-Medicine (SSBM) belongs to a new generation of molecular biology and genetics software products that use the newest computing technologies—relational databases, client-server architecture, and Java framework. The novel software design of SSBM augments considerably the features usually encountered in stand-alone applications running on desktop computers. Not so long ago, the best bioinformatics software could have been identified by simply counting all available features and comparing the speed of similarity searches and sequence alignments. Today, products like SSBM illustrate that, in addition to traditional sequence analysis features, contemporary software dealing with genomic data must solve the problem of database management and provide a secure environment for the collaborative work of many scientists.

Software Solution For Bio-Medicine
InforMax, Inc.
Bethesda, MD.
$250.
800-357-3114
www.informaxinc.com

SSBM belongs to a family of software products for bioinformatics and molecular biology research and extends the functionality of the well-known InforMax Vector NTI Suite. Software products for bioinformatics usually contain computational tools to allow researchers to easily manipulate and study nucleic acid and protein sequences in detail. In its present design, the SSBM package consists of three separate, networked software modules: (i) database files, which contain diverse sequence, annotation, and analysis data; (ii) the server software application, which executes client requests, performs searches, and prepares documents and reports; and (iii) the client software application, which presents the user with tools for managing sequences, interpreting search results, and conducting sequence analyses.

The SSBM database uses the relational database engine Oracle (from Oracle Corporation), which permits storage of a tremendous amount of genomic information while performing fast and reliable searches. The ability to use Oracle is very attractive, not only for its excellent performance, but also because of its widespread use by major pharmaceutical manufacturers.

The SSBM server application is at the center of all of the sequence analysis methods and similarity searches. It performs database similarity searches, conducts multiple sequence alignments, prepares sequence annotations to be viewed by the user, and so forth. SSBM's design allows advanced users to modify the program by adding novel analysis methods directly into the SSBM server application framework. For example, SSBM allows the system administrator to introduce new sequence analysis methods and new user interface elements.

There are certain considerations in purchasing and setting up a client-server application such as SSBM that desktop computer users normally wouldn't have to consider. For example, most desktop bioinformatics software products take the same time to perform an alignment on a Monday afternoon as on a Friday morning. However, a client-server system, such as SSBM, will behave differently. Users will need to do their homework and consider the total number of clients, complexity of an average task performed, network traffic, and size of the database before ordering the product. It is even more important to know how performance may be improved by boosting the computational power of the server with more processors, memory, or disk space. Potential customers should design an activity profile containing an estimated number of users and a list of typical searches, alignments, and analyses to be performed with SSBM. Such an activity profile, in conjunction with InforMax's benchmark tables (which list an average waiting time for a user operation performed on the server in its standard configuration), will assist in planning the server hardware configuration and estimating the total price of the SSBM installation. Additional benchmarks will estimate the effect of a number of connected users on overall system performance.

At the time of this review, benchmark tables were not yet available from InforMax, so potential SSBM customers will have to rely on a subjective measurement of performance. Users of SSBM would benefit from having reference benchmarks reported in the product literature and software manuals.

The SSBM server runs cooperatively with the host company Web server. Most users, however, will interact only with the SSBM client application on a PC via a Web browser. A user launches SSBM simply by navigating to a page on the InforMax Web site and logging in. Once authenticated, the user continues to work in the newly opened browser window, which is brought to life by a powerful Java program called an applet. User actions available through the applet (such as selecting a menu item, choosing a sequence from the list, or pushing a button) are communicated to the SSBM server to obtain more data or to invoke a server function.

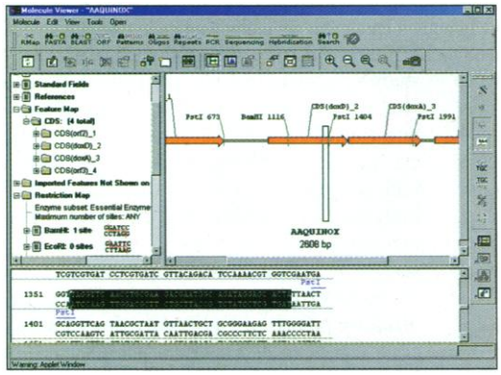

Fig. 1. Molecule Viewer window. Nucleotide and amino acid sequences contain a large number of theoretically predicted and experimentally determined pieces of information. The Molecule Viewer applet summarizes computed and available experimental sequence region properties (top left), creates a map of annotated sequence regions (top right), and allows a detailed view of every sequence residue (bottom; double-stranded DNA is shown).

SSBM combines several different types of data in an extensible work space. With the program, users are finally relieved of the painful task of tracking numerous sequence and report files on different computers and floppy disks. The SSBM environment stores everything in one central location, "the database." The first and most obvious advantage of this arrangement is the ability to store collections of nucleic acid and amino acid sequences. Users can manually enter sequences into the database. The system administrator can deposit entire databases into it, such as SWISS-PROT, GenBank, or PDB. An ordinary SSBM user has no need to know where specific database files are or how to build an index file. "The database" does all of these low-level tasks automatically. All types of reports created within the SSBM work space are stored in the database if the user chooses to save them.

Another advantage of the central sequence database is that any type of sequence search or report created in the program has ready access to the relevant sequence or sequences and analyses. Finally, the structure of the database is kept flexible enough to permit user-definable fields. There are virtu-

The author is at Outplay Consulting, 12 St. Clair Avenue, Post Office Box 69032, Toronto, Ontario, M4T 1K0, Canada. E-mail: psimakov@outplay.com

ally unlimited possibilities to use such fields for storing notes, personal and departmental information, as well as for organization-wide registration codes and labels. Currently, only text and numeric data can be stored in the user-definable fields, however.

SSBM uses a tight security scheme to control access to the database. The system administrator creates new user groups. One member of the group, designated as group manager, is responsible for adding or deleting users for that group. The system administrator always can see and modify any item in the database. All SSBM sequence data, annotations, analyses, and reports have an assigned owner. The SSBM server secures data belonging to an individual user and to a group of users. Any scientist, after creating a new sequence, has an option to save the sequence in a private folder, in the folder shared by his or her research group, or the sequence can be published to everyone with server access. No other scientist will be able to access sequence data, annotations, and analysis for any sequences marked as "private." If the sequence is published to a group of scientists, any group member will be able to make changes or delete a sequence. Everyone in the group might use the same set of sequences for similarity searches, for statistical studies, for temporary annotations, or for information exchange.

The SSBM work space is built around browser, Internet, and Java technology, so it inherits the remarkable convenience and interactivity of desktop applications. The SSBM work space incorporates the Microsoft Office look and feel, including toolbars, pop-up and pull-down menus, fonts of multiple styles and colors, zoom levels, presentation of facts in the tree view, charts and drawings, and color printing from the presentation window. The SSBM work space is claimed by InforMax to be one of the largest user interface developments ever undertaken in Java.

Depending on the situation, users will encounter different types of SSBM view windows including Subset View, Molecule View, Search Result View, and Alignment View. The Subset View shows a list of nucleotide or amino acid sequences containing the word "gene" in the description field. A subset is automatically constructed to represent search results, or sequences can be manually added to create a custom subset. Most essential sequence information is presented in the Subset View.

All sequence annotations can be viewed simultaneously in the Molecule View (Fig. 1), where the sequence "AAQUINOX" is presented. The top left corner of the window contains computed sequence properties and the feature map retrieved from the database. The top right corner contains a cartoon of the "AAQUINOX" sequence with different sequence features marked with separate arrows. The nucleotide sequence itself is shown at the bottom of the window, where sequence residues corresponding to a selected feature are automatically highlighted. If the Molecule View window contains an amino acid sequence, the sequence feature map can be replaced with a hydrophobicity or other sequence profile plot.

Results of similarity searches (Fig. 2) or multiple alignments are presented in the analysis-dependent window. For example, the list of best-similarity hits is automatically extracted from the Basic Local Alignment Search Tool (BLAST) similarity search results file and is presented in easy-to-read graphical form, making it possible to organize and analyze huge amounts of information.

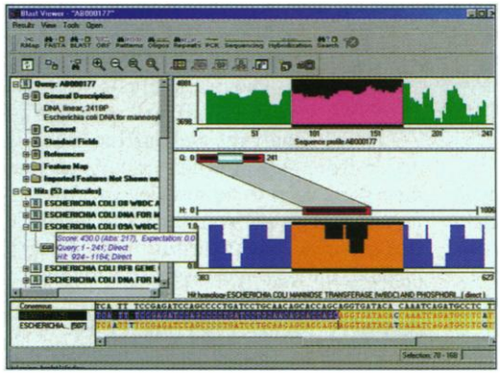

Fig. 2. BLAST Viewer window. SSBM includes convenient tools to graphically interpret the statistical significance of the BLAST search and to compare two or more sequences. A single window conveniently summarizes the similarity hits (top left), statistical sequence profiles (top right), and sequence alignment (bottom).

Users who have the InforMax Vector NTI suite in addition to SSBM can take advantage of instant data transfer between the Vector NTI work space and the sequence currently presented in the molecule view. When data transfer was performed from SSBM into Vector NTI, all sequence annotations were safely preserved. Reverse transfer of the sequence from Vector NTI into SSBM works equally well using a special Web form. Users can specify a private or public database subset for storing new sequences.

SSBM supports, and can be shipped with, any of the publicly available nucleotide or amino acid sequence databases (GenBank, SWISS-PROT, PDB, EMBL, and TREMBL). Customized integration to existing proprietary databases is offered as well. SSBM successfully recognizes virtually all sequence annotations stored in the above databases including feature maps, residue ambiguities, references, species, keywords, organisms, and database cross-references. When new versions of the databases become available, the SSBM system administrator can import the entire database content into SSBM, thereby replacing old data. New index files for BLAST and other searches are built automatically at the time of importing. SSBM also supports incremental updates.

Among the many features of SSBM, its versatile search capabilities are most notable. From the SSBM work space, users can perform the widely used BLAST and FASTA similarity, BLOCKS profile, and PROSITE pattern searches with just a click of the mouse. Text searches can be conducted across any combination of standard (description, author, species, and so forth) or user-defined database fields. Most important, retrieved search hits are always conveniently organized in subsets, and search statistics can be analyzed with the help of specialized viewers.

Anyone concerned about the accuracy of the sequence analyses will appreciate that the majority of the SSBM analysis methods were taken directly from InforMax Vector NTI Suite, which has been in use for several years. In other cases, either original computer programs (as distributed by authors) were used [for example, BLAST, distributed by the National Center for Biotechnology Information] or the original algorithms from the literature were re-created by InforMax and optimized.

In summary, SSBM is a client-server application that is well designed and provides the power and functionality necessary to serve a diverse set of enterprises. It is an important evolutionary step for InforMax that may give its Vector NTI product a significant advantage over other desktop applications.

The minimum system requirements are a Sun Solaris 2.5.1+, SGI Irix 6.4.1, Dec Alpha Digital Unix 4.0 D, 1+ gigabytes (GB) of memory, 4+ CPUs, HTTP server (tested with Apache), and Oracle 7+. The SSBM can be shipped with all of the publicly available databases, which require approximately 70 GB of server disk space for a complete installation. The SSBM Java Client 1.0 can be started from any Java 1.0.2–compliant Web browser: Internet Explorer 4.0+ or Netscape Navigator 3.0+ and Communicator 4.0+ on PC running Windows 95+, NT 4.0+, or Macintosh Power PC running MacOS 7.61 or 8.0+. The virtual reality CosmoPlayer 2.1 plug-in or RasMol are required for viewing PDB files in three dimensions.