

Internet On-Ramp

Get Your Bioinformatics on the Web!

Stuart M. Brown, NYU School of Medicine
(browns02@med.nyu.edu)

A new kind of Web-based bioinformatics service is emerging to aid bench scientists. Integrated suites of programs are being hosted on remote servers and being offered as a package for a membership fee, or for free with the option to purchase added value services. The companies providing these Web tools are following a new Internet business strategy known as the Application Service Provider (ASP) model.

The basic idea of an ASP is that it is more efficient to have complex and expensive software hosted centrally on the well-maintained computers of an expert service provider rather than to buy the software and run it on one's own computer. From the perspective of a lab manager or PI, this is not so different from the central computing centers that most large universities and pharmaceutical research divisions provide for their scientists. However, these new bioinformatics ASPs offer this level of professional service to individual investigators and small labs that do not have access to high-quality central computing facilities. An ASP can be considered an extreme form of outsourcing of computer technology—where the vendor not only provides the computer hardware, software and system maintenance, but the actual computers that are located off-site as well.

Bioinformatics software plays a similar central role at a pharmaceutical company or biomedical research institution as does enterprise software for a bank or large retail company. Forrester Research has estimated that large companies spend about 80% of the initial cost of their major enterprise software applications each year on maintenance. This large investment also means that it is hard to change these applications quickly. In the ASP model, the costs and hassles of maintenance and upgrades are the responsibility of the ASP vendor, not the scientist. If new or better software becomes available, the ASP vendor can add it, or the scientist can move to another vendor that offers better service. Another advantage of an ASP is that it is possible for users to pay for access to software as they use it, and to have access to more different types of software that offer richer functionality than they might be able to afford on their own.

The availability of bioinformatics tools on the Web is nothing new—NCBI has been providing BLAST searches on the Web for over five years. The primary advantages of these new integrated packages offered by the ASP vendors are a permanent place to store your data and the results of your analysis and a wide selection of programs available from a single Web interface. There are also various automated tools that can regularly search new sequence data with a set of your own query sequences or automatically perform a series of analyses on a batch of data files. Some of the ASP vendors also offer various forms of technical support and user training, but this tends to be a more costly value-added service.

Fee for Service Bioinformatics ASPs

The Genetics Computing Group (GCG) is a prime mover in the area of bioinformatics ASPs with its new SeqWeb™ interface to the Wisconsin Package of DNA analysis programs.

GCG has recently entered a strategic partnership with Viaken Systems, Inc. to provide secure Internet access to SeqWeb™ as well as GCG's new SeqStore™ data management product, so that labs can have a central database to store sequence data, primers, plasmids, genotypes and data from gene expression microarrays. The target for these services are labs at smaller universities that do not provide central molecular biology computing resources, and small companies. However, Viaken may be able to offer better service and technical support than some larger university "computing centers" that often provide an out-of-the-box GCG installation and out-of-date local databases with minimal technical support from a systems manager with no knowledge of molecular biology or bioinformatics.

The University of Sidney has spun off a private company called eBioinformatics, Inc., based on technology developed as part of the Australian National Genomic Information Service (ANGIS) to provide commercial bioinformatics services to academic and corporate scientists. eBioinformatics offers access to its own comprehensive suite of Web-based bioinformatics tools called BioNavigator™ with a usage-based fee for time used and amounts of data stored. Registration on BioNavigator provides access to GCG SeqWeb as well as eBioinformatics' own proprietary software and databases. Subscribers are offered a comprehensive system that includes storage of files, software, databases, documentation, training and support. BioNavigator tools include PCR primer design, restriction mapping, protein secondary structure prediction, database similarity searching (and scheduled automatic searches), pairwise and multiple alignment and construction of phylogenetic trees. BioNavigator provides "Protocols" that are step-by-step tutorials that guide users through common bioinformatics procedures such as making multiple alignments or running a similarity search. BioNavigator also provides the feature "SMARTlink™" that features an intelligent application service link to any IMAGE clone that is identified in a bioinformatics operation, such as a BLAST search of GenBank®, and an immediate connection to GenomeSystems Inc. for acquisition of the clone.

The Canadian Bioinformatics Resource provides Web and Telnet access to GCG and other bioinformatics applications for a small yearly fee. However, accounts are only available to researchers who are located in Canada and are affiliated with a university, hospital, government department or nonprofit organization. This service is a replacement for the sale of outsourced GCG accounts, which was formerly offered by the bioinformatics company, Base4, Inc.

Hyseq Inc. has set up its own version of a Web-based bioinformatics/e-commerce tool called GeneSolutions™. GeneSolutions offers pay-as-you-go access to Hyseq's own proprietary databases of over 12 million DNA sequences (primarily ESTs). The Hyseq databases include many rarely expressed genes, extended sequence of cDNAs that are found in public databases and sequences of the same gene from multiple individuals that may aid in the characterization of polymorphisms. The GeneSolutions Web site allows users free access to store query sequences and to run scheduled searches against new data, but there are fees to see the sequences discovered by the searches and much larger fees to obtain the clones from Hyseq. Hyseq has also recorded tissue-specific expression profiles for thousands of genes based on actual counts of the numbers of molecules found in RNA extracted from approximately 50 libraries of tissue

Internet On-Ramp

BioBit: URLs for Bioinformatics ASPs and Integrated Web Sites

Fee for Service ASPs

Viaken Systems, Inc. (provides GCG SeqWeb and SeqStore): <http://www.viaken.com>

eBioinformatics, Inc. (BioNavigator and GCG SeqWeb): <http://www.ebioinformatics.com>

Canadian Bioinformatics Resource: <http://www.cbr.nrc.ca/newdocs/home/>

HySeq GeneSolutions: <http://www.genesolutions.com>

Free ASPs

DoubleTwist, Inc.: <http://www.doubletwist.com>

NCSA Biology Workbench: <http://biology.ncsa.uiuc.edu/>

UK Human Genome Mapping Project Resource Centre: <http://www.hgmp.mrc.ac.uk/>

Curagen CuraTools at the GeneScape Portal: <http://curatools.curagen.com>

Integrated Bioinformatics Web Sites

BCM Search Launcher: <http://www.hgsc.bcm.tmc.edu/SearchLauncher>

The Biologists Search Palette: <http://www.biozentrum.uni-wuerzburg.de/biolinks/search.html>

Emmanuel Skoufos' Gene Discovery Page: <http://bioinformatics.weizmann.ac.il/gdp/gdp.html>

sources. Tissue-expression data may be added to a sequence profile for yet another fee.

"Free" ASPs and Integrated Bioinformatics Web Sites

DoubleTwist, Inc. (formerly Pangea Systems of Oakland, CA, USA) is in the process of creating a Web-based bioinformatics tool called DoubleTwist™. DoubleTwist is a "free integrated Web portal for online genetic research" that will include automated "research agents" to simplify common bioinformatics operations. Users will be able to store their own sequences on the site, set up automated searches and store search results. The site will make use of proprietary databases such as DoubleTwist's EST consensus sequences, the ProNet™ database of protein interactions from Myriad Genetics and a set of 50 000 proprietary full-length cDNA sequences from AlphaGene. A researcher using DoubleTwist who finds a match between a query sequence and an AlphaGene clone will be able to license or purchase the clone. DoubleTwist will also provide "contextual" e-commerce links from life science vendors such as Clontech Laboratories that will offer products for sale that target a researcher's specific needs. In a sense, DoubleTwist is a portal Web site for biologists, offering them free bioinformatics tools to attract them and then sell them items such as clones and reagents. Other portal sites such as the BioMedNet HMS Beagle (www.biomednet.com/hmsbeagle) have targeted biologists by providing interesting magazine-like content or science-related news headlines (www.biotaq.com).

The Computational Biology Group at the National Center for Supercomputing Applications (NCSA), University of Illinois, has developed a free Web-based program called The Biology Workbench, a computing environment that integrates many

standard protein and nucleic acid sequence databases and a wide variety of sequence analysis programs into a single interface. The WorkBench allows users to perform database searches, multiple alignments, phylogenetic analyses, protein secondary structure prediction and motif searches, all without worrying about file formats or often-confusing command line options and flags. Sequences, projects and results are saved in a password-protected area for each user, and encryption is used to transmit data between the NCSA computer and the Web browser on the user's computer.

The UK Human Genome Mapping Project Resource Centre offers free access (for bona fide academic researchers) to a comprehensive range of programs and databases to aid genomic and proteomic research. Online tools include GCG, Staden, GDE, PHYLIP and a bewildering array of other software. They have also developed their own graphical Web-based interfaces for a number of common molecular biology tasks including DNA sequence assembly (using Phred and Phrap), BLAST searches, identification of genes in genomic sequences, protein identification, multiple alignment, phylogenetic analyses and genetic linkage analyses.

CuraGen Corporation of New Haven, CT, USA has created a suite of genomics tools, which it calls CuraTools™, available for free on a Web site called the GeneScape Portal™. CuraTools is an integrated suite of bioinformatics tools that allows users to easily perform analyses on collections of DNA and protein sequences. Users can upload their own sequence data that can be edited and stored indefinitely in a password-protected personal file storage area on the CuraGen Web site. Using a simple Web interface, users can rapidly analyze their data and associate it with many other forms of genomic data following similarity searches, motif analyses, structure-function predictions, etc. Results from the various analyses are displayed with the aid of specifically tailored viewers and interpreters. Results can be saved together with personal data sets.

CuraTools includes BLAST similarity searches, ClustalW multiple alignment, PCR primer design, DNA sequence assembly with CAP2, DNA restriction analysis and ORF prediction, various protein analyses, protein domain analyses with the PROSITE, BLOCKS, PRODOM, PRINTS, SBASE and Pfam databases and phylogenetic tree analyses calculated with the PHYLIP neighbor-joining algorithm. Most of these tools are implemented by sending queries out to a wide variety of other Web servers and capturing the results that are reformatted into a consistent "report" format. While not as powerful as GCG, the CuraTools Web site offers a simpler interface and better integration of results than other sites that provide a single point of access to aggregations of Web tools such as the BCM Search Launcher, the Biologists Search Palette and Emmanuel Skoufos' Gene Discovery Page. In addition, the ability to store personal sequences and search results makes CuraTools by far the best free integrated bioinformatics Web site.

There are also risks associated with outsourcing crucial bioinformatics software. If a business or lab does not own the software applications they use, they may find it difficult to maintain a depth of knowledge about them, which could lead to a risky level of dependence on the ASP. There is also a vulnerability inherent in depending on someone else to maintain a computer system—slowdowns or failures at the ASP could delay critical research. A relationship with an ASP for bioinformatics software is like a strategic partnership, a concept that is quite familiar to many biotechnology and pharmaceutical companies.