Quebecor - On-Ramp graphic goes here run visible portion of black circle in red - left align with text

Internet On-Ramp

Be Alert: Free Sequence Alerting Services

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

It has always been a challenge to keep up with all of the new data flowing into biological databases. Now, as the stream of DNA sequences from the Human Genome Project has increased to a torrent, the chances have drastically increased that something new has rendered obsolete the database searches that you did last week. You can spend lots of time obsessively repeating your searches or look for an automated service that will perform regular searches and send you notification when relevant new sequences have entered the database. There are also some free services that will check for new additions to the PubMed®/MEDLINE® database and notify you when relevant new journal articles appear.

Several free automated "sequence alerting" services exist on the Web in various forms. These can serve as a method to attract scientists to a multifunctional Web site, or merely be an altruistic contribution from a bioinformatics group. There are two basic types of sequence alerting services: those that perform similarity searches based on a user's query sequence and those that perform keyword searches of title/annotation information. In addition, various services are designed to monitor different databases, such as the **SwissShop** service (**http://www.expasy.ch/swiss-shop**), which only tracks updates to the SwissProt database.

A crucial aspect of an alerting service is that it only sends you NEW finds from the database, not merely repeating a search at some interval and sending you a complete listing of all matches, which would lead to huge amounts of time wasted sorting through the same sequences in every report.

The European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany operates a **Sequence Alerting System** (**http://www.bork.embl-heidelberg.de/Alerting/**). The system searches each day for sequences similar to your query sequence and will inform you by e-mail if it has detected a new match. However, the system only searches protein databases (e.g., TREMBL, PDB, SwissProt and PIR), so it will not detect new ESTs or genomic sequences that contain sequences that have not been automatically detected and submitted to a translated protein database such as TREMBL.

The ExPASy server at the University of Geneva (home of the SwissProt database) runs the alerting service called **SwissShop**. New protein sequences that are added to the SwissProt database can be scanned on a weekly basis for a user-defined query. The service provides for both keyword searches and sequence-based searches (using either the BLAST similarity algorithm or pattern searches use exact matching of PROSITE type patterns).

GeneNet (**http://brac.postech.ac.kr/eng/**) is a meta-search system for the analysis of sequence similarity via the Web that is developed and maintained by the Biological Research Information Center, Pohang University, Korea. For protein sequences, searches are performed on GenBank®, PDB, BLOCKS and KEGG; for DNA sequences, only GenBank is searched. The summarized results are provided by e-mail. Users can get full results on a Web page by following an embedded HTML link. Once an initial analysis is made, users can choose to have the search repeated on a monthly basis and the results sent by e-mail.

PubCrawler (**http://www.pubcrawler.ie/**) is a free alerting service that scans daily updates to the NCBI MEDLINE (PubMed) and GenBank databases, so it can alert users both to new sequences and new journal articles. PubCrawler can keep scientists informed of the current contents of MEDLINE and GenBank by listing new database entries that match their research interests. "It goes to the library. You go to the pub."™

PubCrawler maintains a database of previous matches for each user, so you are only alerted when new sequences and/or journal articles appear that you have not seen before. Full results can be sent by e-mail notification when new results are available, or the user can just retrieve them directly on the PubCrawler Web site whenever he or she wishes to check them.

PubCrawler was developed in the Department of Genetics, Trinity College, Dublin, Ireland. The PubCrawler software (UNIX®, Microsoft® Windows™ and Macintosh® versions) is also available for free to install on any computer to manage an update service for one user or an entire institution.

BioNavigator (**http://www.bionavigator.com**) is a commercial Web-based bioinformatics service operated by the eBioinformatics company (Eveleigh, NSW, Australia), which provides a free notification service for BLAST and keyword searches. BLAST searches of GenBank and text/keyword searches are free and can be set to be repeated weekly using the "Notify" function. Keyword searches use the "TextSearch" program, which utilizes the SRS database engine developed by Lion Bioscience and allows the user to search GenBank, BLOCKS, Pfam, PDB, Prosite, SwissProt and TREMBL. Registration for the BioNavigator service is also free and does not require the user to enter billing information. All results are stored on the Web in individual user's project areas.

DoubleTwist (**http://doubletwist.com**) is a commercial service focused on genomics. It provides free daily sequence updates via monitor agents for up to 10

query sequences (DNA or protein). DoubleTwist uses multiple similarity searching algorithms to compare the user's query sequences against about 20 public and proprietary databases including DNA, protein, ESTs, structures, motifs, domains, patented sequences, etc. When new matching sequences enter these databases, DoubleTwist notifies you by e-mail. The service offers e-mail alerts, but the actual results are retrieved via the Web. For a monthly subscription fee, additional query sequences can be searched and additional databases are available such as DoubleTwist's proprietary computationally generated clustered EST and annotated human genome databases.

The Biotechnology Computing Facility at the University of Arizona has created an automated **Entrez Query Tool** (**http://bcf.arl.arizona.edu/query/**) that makes regular searches of the Entrez database at NCBI and notifies users when new items are found. The user can customize an account for creating and managing individual queries. In each query, the search term is specified, along with the database (PubMed/MEDLINE, Protein, Nucleotide, Structure and Genome), frequency of the query (Daily, Weekly or Monthly) with dates for starting and terminating the automated search. If new results are found for a query, then the information is e-mailed to the user. The service is free and open to the public.

BioMail (**http://www.biomail.org** and **http://biomail.sourceforge.net/biomail/**) is a free Web application that regularly (weekly by default) searches for articles that have recently appeared in the PubMed/MEDLINE database using user-customized search terms. It then e-mails lists of the found articles to the user. The HTML-formatted e-mails generated by BioMail can be used to view selected references in MEDLINE format (which is compatible with most reference manager programs). BioMail is developed and maintained at SUNY Stony Brook. BioMail source code was written in Perl for Linux; it is free software released under GNU GPL license.

JADE (Journal Articles Delivered Electronically) (**http://www.biodigital.org/jade/**) is another automated MEDLINE search agent developed and maintained by The National Center for Emergency Medicine Informatics. JADE makes weekly searches of MEDLINE with user-specified search terms and sends the results by e-mail. JADE also has a "Personal Archive" feature that allows each user to store citations to articles found in the searches in a personal database. Individual articles can also be added to the archive using the MEDLINE UID.

DBWatcher (**http://www-igbmc.u-strasbg.fr/BioInfo/LocalDoc/DBWatcher/**) is a free UNIX program that makes automated BLAST searches. It is developed and distributed by the Institut de Genetique et de Biologie Moleculaire et Cellulaire, Strasbourg, France. This is NOT a Web-based service, but rather a program intended to be installed on a UNIX computer to serve a number of scientists. The program is extremely customizable both in terms of query terms and databases searched (GenBank's BLAST databases or a local BLAST formatted database). Searches can be set to execute at any interval.

Only novel similarities are reported, thus saving the time of browsing through bulky results files. Results are sent by e-mail to one or several addresses, making DBWatcher particularly suitable for collaborative database surveys and for users who do not have access to locally maintained databases.

There is also a vast array of journal article/table of contents-based alerting services. Some will e-mail you the table of contents when a new issue of a specific journal is released. Others will scan titles (or abstracts) for keyword matches from a group of journals and notify you by e-mail when a relevant article is published. Many of these services are offered by specific publishers for their own journals or by document delivery services that have distribution agreements with a number of publishers.

By making use of one or more of these alerting tools, researchers can stay on top of newly discovered sequences and new journal articles relevant to their research with no need to constantly repeat the same database searches. However, for those who prefer to repeat searches manually, rather than receive automatic updates by e-mail, the NCBI has recently added a new feature called the **Cubby** to their Entrez Web site (**http://www.ncbi.nlm.nih.gov:80/entrez/cubby/login.fcgi?call=so.SignOn..Login**).

The Cubby is a password-protected personal storage space for Entrez queries (free registration is required). Stored search information in the Cubby includes the search name, date and time last updated, database searched, search terms and field limits—but not actual sequences or journal citations. Up to 100 different search queries can be stored. This can be very useful because it is difficult to remember from month to month the exact combination of keywords and field limits that were used to find a useful set of sequences or journal articles. A stored search can be updated with a single click, which will retrieve only those matching database records that have been added to the database since this search was last performed. Alternatively, an entire search can be repeated to bring up all matching records regardless of whether they have been seen before. Stored searches can also be modified before repeating them. Use these Internet tools to power your research with the flow of genome information rather than being swamped by it.