Tracker: continuous HMMER and BLAST searching

Madelaine Marchin, Paul T. Kelly and Jianwen Fang*

Department of Molecular Biosciences, University of Kansas, Lawrence, KS 66045, USA

Received on April 21, 2004; revised on July 28, 2004; accepted on August 31, 2004

Advance Access publication September 3, 2004

ABSTRACT

Summary: Tracker is a web-based email alert system for monitoring protein database searches using HMMER and Blast-P, nucleotide searches using Blast-N and literature searches of the PubMed database. Users submit searches via a web-based interface. Searches are saved and run against updated databases to alert users about new information. If there are new results from the saved searches, users will be notified by email and will then be able to access results and link to additional information on the NCBI website. Tracker supports Boolean AND/OR operations on HMMER and BLASTP result sets to allow users to broaden or narrow protein searches. Availability: The server is located at http://jay.bioinformatics.ku.edu/tracker/index.html. A distribution package including detailed installation procedure is freely available from http://jay.bioinformatics.ku.edu/download/tracker/
Contact: jwfang@ku.edu

Email alert services allow users to efficiently research and obtain information on a variety of topics with little effort. We have developed a new comprehensive email alert service. Protein domains and sequences and nucleotide sequences can now be monitored more closely. A search is entered once and performed nightly on new information entered into local and external databases. If new results are found for the search, users are notified via email and then access the results in the form of a generated web page. A few sequence search alert systems already exist. For example, Swiss-shop (http://www.expasy.org/swiss-shop/) provides sequence/pattern or keyword-based searches on the current non-cumulative weekly additions in Swiss-Prot protein database. To the best of our knowledge, Tracker is the first web-based email alert system that integrates HMMER and BLASTP protein searches, as well as Boolean combinations of search results. These features allow users to broaden or narrow their searches as needed and retrieve protein sequences of interests. In addition, Tracker includes nucleotide sequence searches and literature searches. The system relieves users from the burden of repeating searches in order to keep up with the exponentially growing amount of information found in biological sequence and literature databases.

Tracker is implemented with open source software and consists of a web interface, Perl scripts and a local relational database. The site and database are located on a Linux server. Users register and manage searches through the web interface created in HTML and using Common Gateway Interface (CGI) programs written in Perl. The user information and search information are stored in a MySQL database. Perl scripts and shell scripts are used to download databases, perform search commands, process results and email users as needed. User registration is required and user login information is stored during each user session to avoid repeated logins. Searches can be modified or deleted by users. Search results are accessed through email alert links or directly on the website. Tracker also integrates a help page that explains how to use the system, contains troubleshooting tips, and has information on search refinement.

To monitor new protein database results, users can choose to use HMMER or BLASTP approaches or combinations thereof. HMMER (Eddy, 1998, http://hmmer.wustl.edu/) is a package for building and searching with profile hidden Markov models (profile HMMs). For HMMER searches, users are able to enter a sequence and find matching protein domains from the Pfam database (Bateman et al., 2000). Users then select which domains are of interest and track all new database entries that also match their chosen domains. BLASTP (Altschul et al., 1990) sequence similarity searches are also available. These two methods of protein searches can be combined in a Boolean fashion to allow for broader or narrower search results. Users can choose to receive all results from the HMMER and BLASTP searches (logical OR between result sets) or receive only results that are present in both search result sets (logical AND between result sets). In addition, if users want to track multiple protein domains, they can choose to receive results from any of the domains and receive all BLASTP results [(HMMER OR HMMER) OR BLASTP] or receive results from any of the domains only if they are also present in the BLASTP results [(HMMER OR HMMER) AND BLASTP]. These Boolean combination features allow users to broaden or narrow searches as needed.
and these search combinations can be edited on the site along with other features of the search. Tracker searches protein databases including entries in the NCBI’s ‘month’ database (including all new or revised GenBank CDS translation, PDB, Swiss-Prot, PIR and PRF released in the last 30 days). Every 30 days, if a protein search has generated new results, users are sent an email with a link that generates an HTML page displaying search results. Users are provided with results that include alignment, E-value, score and description, along with a link to more information for each result sequence on the NCBI site.

Nucleotide sequence searches are performed using BLASTN (Altschul et al., 1990). Nucleotide sequences searched are from the NCBI’s ‘month’ database, which includes all new or revised GenBank, EMBL, DDBJ and PDB sequences released in the last 30 days. Every 30 days, if a search generates any new results, users receive an email with a link to all the new sequence database entries that align with their search sequence. Again, users are provided with alignment, E-value, score, description and a link to NCBI in the result pages.

In addition to the protein and nucleotide sequence search options, Tracker also monitors PubMed literature searches. Similar systems have already been developed: PubCrawler, (http://www.pubcrawler.ie/), Bio-mail (http://biomain.sourceforge.net/) and Amedeo (http://www.amedeo.com/). In Tracker, queries may be written in PubMed search syntax so that users may generate queries at PubMed and then transfer their queries to Tracker to continually monitor queries at their chosen frequency and be notified of results via email. Literature searches can also be edited or deleted at the site.

Tracker is a free software that was developed and tested in a Dell PC running LINUX RedHat 9. It was also successfully installed in RedHat 7, RedHat AS3 and a quad AMD64 system running SuSe Enterprise Server 8. Source code, database schema and installation and help files are available at http://jay.bioinformatics.ku.edu/download/tracker/. A number of further improvements are planned. For example, more domain and motif databases will be incorporated into searches. Batch submission of protein and DNA sequences will also be supported to accommodate the needs of sequencing centers.

ACKNOWLEDGEMENTS
This work was supported by NIH grant number P20 RR16475 from the BRIN Program of the National Center for Research Resources and NSF EPSCoR for Kansas. We thank Dr Gerry Lushington and Dr Igor Kuznetsov for discussions and feedback on this work.

REFERENCES