What's new in the library? What's new in GenBank? Let PubCrawler tell you

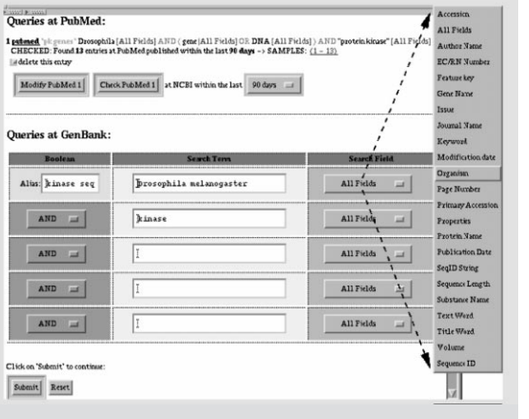
The scientific literature is growing so quickly that many scientists no longer have time to scan the latest issues of all the journals that are relevant to their interests. Thanks to online services like NCBI's Entrez, which has been described in TIG previously[1], it has become possible to search huge libraries for specific articles without leaving the desk. Entrez provides free access to several scientific databases, including PubMed, the world's largest database of biomedical literature. PubMed currently holds the abstracts of approximately nine million scientific journal articles, including the complete contents of MEDLINE. New articles in any research field can be expected nearly at a daily rate, so staying abreast of the current state of science requires frequent electronic library searches. Interesting documents can be overlooked if searches are not made regularly, but carrying out searches can be uninteresting and laborious, particularly at times of day when traffic on the Internet is slow. The repetitive querying of online databases can easily be automated by computer. The PubCrawler WWW service is an automated update alerting service for users of NCBI's PubMed (literature) and GenBank (DNA sequence) databases. PubCrawler carries out personalized searches at NCBI at regular intervals (e.g. daily), keeps track of what records have been seen previously, and produces a WWW page listing the latest hits that match the user's interests. This article describes several features of this service and how to access it.

Automated update delivery

PubCrawler is intended for scientists who want to be informed of the latest publications in their field of interest, as soon as they appear in PubMed. Its journal-monitoring function is similar to that of services that search commercial databases, such as Current Contents® (Institute for Scientific Information, Philadelphia, PA), or SciSearch® at LANL (Los Alamos National Laboratory). However, PubCrawler has the twin advantages of being free, and of being able to monitor new DNA sequences in GenBank as well. Often, a new sequence appears in GenBank months before the paper describing it is published. PubCrawler searches the annotation text of GenBank entries, not the sequence data itself. This can be complemented by services performing BLAST searches[2] against DNA and protein databases like SIB's Swiss-Shop (http://www.expasy.ch/swiss-shop/), the Sequence Alerting System in Peer Bork's lab at EMBL (http://www.bork.embl-heidelberg.de/Alerting/), or NCBI's XREFdb (http://www.ncbi.nlm.nih.gov/XREFdb/).

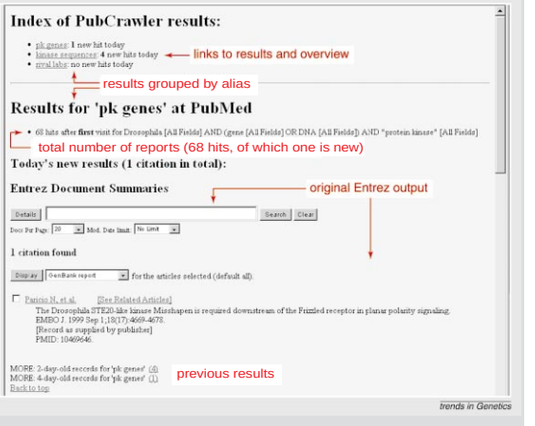
A new PubCrawler user must first create a search profile, consisting of one or more Entrez queries. For example, someone interested in fruitfly protein kinase genes could set up a profile that searches for papers in PubMed whose abstracts contain the words '*Drosophila*', and 'gene' or 'DNA', and the phrase 'protein kinase'. A second query in the profile might search for papers with particular author names (e.g. rival fruitfly protein kinase labs). A third query might scan GenBank for new sequences where the organism is *Drosophila melanogaster* and the annotation contains the word kinase. Any number of queries can be combined into the search profile. A WWW site, the PubCrawler Configurator, has been set up to help with building and editing search profiles (Fig. 1). This allows users to check the syntax of their queries, and to get a feel for how many database entries might match each query in the profile. For PubCrawler to work effectively, the search profile should be neither too broad nor too specific, so that a manageable number of hits are returned each day. Because the search profile can include an unlimited number of queries, and is stored and can be edited at any time using the Configurator WWW page, users can build more-detailed and comprehensive search profiles than they would by occasional use of Entrez. When setting up a search profile, the user chooses how often the searches are to be run (for example, daily on Monday–Friday). They can then browse their results every day at the PubCrawler WWW site, or can have the results e-mailed to them as an HTML document (viewable with mail programs such as Netscape Mail or Microsoft Outlook Express) or as plain text (although this loses the usefulness of having hypertext links to NCBI). All personal information and profile data remains confidential and is password protected.

FIGURE 1. Snapshot of PubCrawler Configurator



An example of a query configured for PubMed is shown on top. This query searches for papers about *Drosophila* protein kinase genes and has the alias name 'pk genes'. Colour codes (not shown here) help structuring the queries (green, alias name; black, search terms; blue, search field; red, boolean operators). Queries can be tested by checking them at NCBI immediately (13 results were found, accessible through a hypertext link). The lower section shows the process of setting up a query for GenBank.

FIGURE 2. Sample PubCrawler results file



PubCrawler incorporates Entrez' original output into a web page and wraps it up with summary information and hypertext links.

The sample results page (Fig. 2) illustrates some of PubCrawler's useful features. The output for each user is a single HTML page, readable with any WWW browser. Each day's new results are presented exactly the way they were received from the NCBI site, with hypertext links (e.g. to view complete abstract or sequence information at NCBI for any of the documents found). A quick index at the top shows at first glance how many documents were found for each search topic. PubCrawler keeps track of all the articles that have been presented, so every time it runs, only new hits matching the personal search profile (i.e. those not previously seen by PubCrawler) will be presented. This avoids the 'have I read this before?' feeling common to absent-minded academics. The results page provides access to older results up to a few days old (the default is seven days, but that can be adjusted) via hypertext links, allowing people to catch up on missed days.

Options

The PubCrawler Configurator allows users to customize their searches in many ways. The variable parameters include the maximum number of documents to retrieve, their maximum age, and how many titles to show on the results page (with the others being accessed by hypertext links from the results page; see Fig. 2). It is possible to specify the frequency at which to run PubCrawler, and a preferred time of day to start it (within NCBI's off-peak hours, 0100–1300 hours GMT). A powerful option is the ability to combine several different queries into a single 'alias'. The hits from all searches in the same alias are merged together and presented as a single result. This is useful because it might be necessary to write several queries to try to cover a scientific topic completely (as in the *Drosophila* protein kinase example above), but the sets of hits returned by these queries might partly overlap. It does not matter if the same database record is hit by two searches that are part of the same alias, because when the hits are merged the record will only be shown once in the final output.

Availability

The PubCrawler WWW service (http://www.pubcrawler.ie) is offered without charge. Its homepage provides a link to the PubCrawler Configurator for setting up personal search profiles. Additionally, the PubCrawler program is freely available for stand-alone installation on a PC, Macintosh or Unix system. The program is written in Perl, which is available for every operating system at no cost. Detailed downloading and installation instructions are available at the same WWW site.

Karsten Hokamp
khokamp@tcd.ie

Ken Wolfe
khwolfe@tcd.ie

Department of Genetics, University of Dublin, Trinity College, Dublin 2, Ireland.

References

1 McEntyre, J. (1998) Linking Up With Entrez. *Trends Genet.* 14, 39–40

2 Altschul, S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410

Heart development, state of the art and the new century

Heart Development
edited by Richard P. Harvey and Nadia Rosenthal
Academic Press, 1998. $159.95 hbk (530 pages) ISBN 0 123 29860 1

This is a beautiful book with outstanding graphics and photographic reproductions. It's almost worth having it just for the pictures. The book covers the main events of cardiovascular development, from early differentiation, to morphogenesis, left–right asymmetry and vascular development. Several chapters are dedicated to gene families that play a role in heart development. The authors deal with topics in depth and there is only very limited, but inevitable, overlap between chapters. Tucked at the end, almost so as to avoid cross-contamination with the other topics, are three chapters illustrating some examples of human cardiac development defects. As up-to-date as a book about a fast-moving field can be, this compendium is overall clearly written and should be useful to both neophytes and experts. An introductory chapter, providing a view of the whole picture, with indication of where the various types of heart defects in humans and animal models are thought to fit, would have been helpful.

Human genetics data provide the most convincing evidence that the mutation of many genes might affect the normal cardiovascular development. A quick search of the Online Mendelian Inheritance in Man database (www3.ncbi.nlm.nih.gov/Omim/) will reveal hundreds of relevant entries. It is also clear that a given heart defect can be associated with different diseases or mutations of different genes. The clinical diagnosis of a heart defect is a picture of the end-product of an abnormal process, and is the result of the reaction and/or adaptation (including hemodynamic factors) of the organism to the primary embryological defect. Such reactions and/or adaptations are presumably under genetic control and influenced by the environment. The action of these factors will have variable results, as illustrated by the incomplete penetrance and variable expressivity of the heart phenotype in many human diseases and mouse mutants. One might think about heart development as the archetypal biomedical problem that should be addressed across species and using many different experimental strategies. This is the kind of problem that should benefit greatly from genomics and the wealth of emerging tools enabling efficient multigene analyses. Examples of these are gene-array expression studies, large-scale mutagenesis screenings and chromosome engineering.

Disappointingly, genetics and genomics do not play a dominant role in this book. The lion's share belongs to classic molecular biology, with a powerful touch of embryology. And yet, it comes out clearly that this development is so genetically complex that understanding it one protein at a time is an improbable task. Gene–gene interactions and gene–environment interactions are critical issues that will keep the heart-development community busy for many years to come.

Those who were waiting for a book that would take this field into the next century might need to wait longer. Nevertheless, I see this as a beautiful summary of the extraordinary work accomplished in the latter part of this century. Scientists growing in the age of global genetic information should take this field from here and develop it using the new tools at our disposal.

Antonio Baldini
baldini@bcm.tmc.edu

Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030-3411, USA.