Nucleic Acids Research, 2006, Vol. 34, Database issue ***D3–D5***
doi:10.1093/nar/gkj162

The Molecular Biology Database Collection: 2006 update

Michael Y. Galperin*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Received October 28, 2005; Revised and Accepted November 1, 2005

ABSTRACT

The NAR Molecular Biology Database Collection is a public online resource that contains links to all databases described in this issue of *Nucleic Acids Research*. In addition, this collection lists databases that have been featured in previous issues of NAR, as well as selected other databases that are freely available to the public and may be useful to the molecular biologist. The 2006 update includes 858 databases, 139 more than the previous one. The databases come with brief summaries, many of which have been updated recently. Each database is assigned a stable accession number that does not change if the database moves to a new location and its URL, authors' names or the contact person address are updated. The complete database list and summaries are available online at the *Nucleic Acids Research* website http://nar.oxfordjournals.org/.

COMMENTARY

This is the 13th annual database issue of *Nucleic Acids Research*, and the first one that goes entirely paperless. The list of molecular biology databases keeps getting bigger, and despite negative connotations sometimes connected to this growth, the databases are getting better and even more diverse. The current release of the *Nucleic Acids Research* online Molecular Biology Database Collection (Supplementary Table 1) includes 92 new databases, first described in this issue, and 49 additional new databases, featured in *Bioinformatics*, *BMC Bioinformatics* and other journals. These include first ever databases from Ireland, Portugal and United Arab Emirates (1–3) and a variety of other databases maintained all over the world.

Meanwhile, existing databases show remarkable resilience: out of 719 databases featured in the last year's release (4), only 2 were no longer maintained because their authors graduated, retired or changed focus, and one more has shifted to restricted access. In contrast, three databases, ABCdb, EID and KDBI, that were considered dead last year and had been crossed off the list, have now been resurrected. In each case, their authors have moved to new work places and were able to resume maintenance of their databases. As promised last year, their accession numbers have not been re-used and these databases are now listed under the same entry numbers, 157, 32 and 138, respectively, that they had in previous releases. These numbers can be used to gain access to updated summaries of these databases on the NAR website, e.g. http://www.oxfordjournals.org/nar/database/summary/157. Similarly, PUMA2 (5), which replaced the WIT2 database, has kept its number 118 in list.

After 12 years of database issues and 8 years of the accompanying web supplement, it was interesting to check if they are really having an impact. In other words, how many people really care about them and use them? To evaluate the impact of the NAR database issues, I have used a tool that, despite all complaints and caveats, is commonly utilized for evaluating research productivity, namely the *Science Citation Index*® produced by the Institute for Scientific Information. If databases are put on the web for the benefit of the research community, the frequency with which people use (and cite) a given database could serve as an indication of whether this database serves a useful purpose. An inspection of the citation figures for the 141 papers published 2 years ago in the 2004 NAR Database Issue (all citation data are as of October 15, 2005) revealed a very encouraging trend. Most of the papers were well—or very well—cited. Only five papers have not been cited at all and the same number of database descriptions — five—have been cited >100 times, becoming, in ISI parlance, instant 'citation classics'. Whatever the caveats, the fact that the paper describing the Pfam domain database [http://www.sanger.ac.uk/Software/Pfam/, NAR Collection entry no. 210, Ref. (6)] has been cited 375 times in <2 years definitely indicates that this database is widely used by the research community. Indeed, comparing a protein sequence against Pfam has become standard practice in sequence analysis, particularly in genome annotation. It is probably no coincidence that the first author of the Pfam paper also serves as the Editor of

*To whom correspondence should be addressed. Tel: +1 301 435 5910; Fax: +1 301 435 7794; Email: galperin@ncbi.nlm.nih.gov

the NAR database issues. In the interest of full disclosure, I have cited this Pfam paper myself eight times since its publication in 2004.

The second best cited database, Gene Ontology (GO) [http://www.geneontology.org/, NAR Collection entry no. 487, Ref. (7)] provides structured, controlled vocabularies and classifications that are also widely used in genome annotation, as well as for a variety of bioinformatics tasks. Other databases in the top five, UniProt [http://www.uniprot.org, NAR Collection entry no. 318, Ref. (8)], SMART [http://smart.embl.de/, NAR Collection entry no. 218, Ref. (9)] and KEGG [http://www.genome.ad.jp/kegg/, NAR Collection entry no. 112, Ref. (10)], are also used by scientists all over the world. It is worth noting that each of these databases allows free downloading of its full content: they work by adding valuable expertise to the sequence data and have nothing to hide.

The databases that form the International Nucleotide Sequence Database Collaboration, NCBI's GenBank, EMBL Nucleotide database and Japanese DDBJ (NAR Collection entries no. 1–3), also attract a respectable number of citations, even though they are usually mentioned in the literature without a formal citation. The same is true for the Protein Data Bank (PDB) (NAR Collection entry no. 276). More databases are probably headed the same way of becoming household names that are not considered to need a citation.

On the other side of the spectrum are the databases that have never been cited in these 2 years, even by their own authors. This does not mean, of course, that these databases do not offer a useful content but one could always suggest a reason why nobody has used this or that database. Usually these databases were too specific in scope and offered content that could be easily found elsewhere. For example, TopoSNP [http://gila-fw.bioengr.uic.edu/snp/toposnp/, NAR Collection entry no. 590, Ref. (11)], maps single nucleotide polymorphisms onto known protein structures, allowing one to trace the location of the affected amino acid residues and correlate it with disease phenotypes. However, most of its data are extracted from OMIM (http://www.ncbi.nlm.nih.gov/omim/, NAR Collection entry no. 143), which is where the user would probably go first. VirGen [http://bioinfo.ernet.in/virgen/virgen.html, NAR Collection entry no. 397, Ref. (12)] is a database of complete genome sequences of plant and animal viruses. However, it often takes a while for the server to produce a response, which contains little information that would not be available in other databases, such as VIPERdb [http://viperdb.scripps.edu/, NAR Collection entry no. 761, Ref. (13)], Viral Bioinformatics Resource Center (http://www.virology.ca/, NAR Collection entry no. 798), VIDA (http://www.biochem.ucl.ac.uk/bsm/virus_database/VIDA.html, NAR Collection entry no. 201) or the NCBI Viral Genomes (http://www.ncbi.nlm.nih.gov/genomes/VIRUSES/viruses.html, NAR Collection entry no. 602). The Ribosomal Protein Gene database (RPG) [http://ribosome.miyazaki-med.ac.jp/, NAR Collection entry no. 573, Ref. (14)] lists ribosomal proteins from just a handful of organisms, offering a tiny fraction of information that is available through Pfam, UniProt, KEGG orthology groups and a variety of other sources. The same problem plagues EyeSite, a database of protein families in the eye [http://eyesite.cryst.bbk.ac.uk/, NAR Collection entry no. 464, Ref. (15)]. Even the terrific graphics on its front page cannot compensate for the fact that researchers interested in eye proteins can get their sequences from UniProt and other sequence databases and their structures from PDB. Finally the Signal Transduction Classification Database (STCDB) [http://bibiserv.techfak.uni-bielefeld.de/stcdb/, NAR Collection entry no. 395, Ref. (16)] offers an interesting approach to the hierarchical classification of eukaryotic signaling proteins. However, so many people use the GO classification that it has become *de facto* standard and nobody is looking for alternative classification schemes. Thus, the fact that this comment will most probably be the first time in 2 years that TopoSNP, VirGen, RPG, EyeSite or STCDB are mentioned in the literature could be a direct consequence of the overwhelming success of other databases. It is an open global marketplace of ideas, tools and approaches; fortunately, nobody goes out of business.

Suggestions for the inclusion of additional databases in this Collection should be directed to the author at galperin@ncbi.nlm.nih.gov.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

I thank Rich Roberts, Alex Bateman and my colleagues at NCBI for helpful comments. This study was supported by the Intramural Research Program of the National Library of Medicine at the US National Institutes of Health. The author's opinions do not necessarily reflect the views of the NCBI, NLM or the National Institutes of Health. The Open Access publication charges for this article were waived by Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES

1. Byrne,K. (2006) Visualizing syntenic relationships among the hemiascomycetes with the Yeast Gene Order Browser. *Nucleic Acids Res.*, **34**, D452–D455.
2. Tadmouri,G.O. (2006) CTGA: the database for genetic disorders in Arab populations. *Nucleic Acids Res.*, **34**, D602–D606.
3. Sá-Correia,I. (2006) The YEASTRACT database: a tool for the analysis of transcription regulatory associations in *Saccharomyces cerevisiae*. *Nucleic Acids Res.*, **34**, D446–D451.
4. Galperin,M.Y. (2005) The Molecular Biology Database Collection: 2005 update. *Nucleic Acids Res.*, **33**, D5–D24.
5. Maltsev,N. (2006) PUMA2—grid-based high-throughput analysis of genomes and metabolic pathways. *Nucleic Acids Res.*, **34**, D369–D372.
6. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
7. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
8. Apweiler,R., Bairoch,A., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.*, **32**, D115–D119.
9. Letunic,I., Copley,R.R., Schmidt,S., Ciccarelli,F.D., Doerks,T., Schultz,J., Ponting,C.P. and Bork,P. (2004) SMART 4.0: towards genomic data integration. *Nucleic Acids Res.*, **32**, D142–D144.

10. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res*., **32**, D277–D280.
11. Stitziel,N.O., Binkowski,T.A., Tseng,Y.Y., Kasif,S. and Liang,J. (2004) topoSNP: a topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association. *Nucleic Acids Res*., **32**, D520–D522.
12. Kulkarni-Kale,U., Bhosle,S., Manjari,G.S. and Kolaskar,A.S. (2004) VirGen: a comprehensive viral genome resource. *Nucleic Acids Res*., **32**, D289–D292.
13. Shepherd,C.M., Borelli,I.A., Lander,G., Natarajan,P., Siddavanahalli,V., Bajaj,C., Johnson,J.E., Brooks,C.L., III and Reddy,V.S. (2006) VIPERdb: a relational database for structural virology. *Nucleic Acids Res*., **34**, D386–D389.
14. Nakao,A., Yoshihama,M. and Kenmochi,N. (2004) RPG: the Ribosomal Protein Gene Database. *Nucleic Acids Res*., **32**, D168–D170.
15. Lee,D.A., Fefeu,S., Edo-Ukeh,A.A., Orengo,C.A. and Slingsby,C. (2004) EyeSite: a semi-automated database of protein families in the eye. *Nucleic Acids Res*., **32**, D148–D152.
16. Chen,M., Lin,S. and Hofestaedt,R. (2004) STCDB: Signal Transduction Classification Database. *Nucleic Acids Res*., **32**, D456–D458.