D2–D4 *Nucleic Acids Research, 2008, Vol. 36, Database issue* *Published online 19 November 2007*
doi:10.1093/nar/gkm1037

The Molecular Biology Database Collection: 2008 update

Michael Y. Galperin*

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

Received October 28, 2007; Accepted October 30, 2007

ABSTRACT

The *Nucleic Acids Research* online Molecular Biology Database Collection is a public repository that lists more than 1000 databases described in this and previous *Nucleic Acids Research* annual database issues, as well as a selection of molecular biology databases described in other journals. All databases included in this Collection are freely available to the public. The 2008 update includes 1078 databases, 110 more than the previous one. The links to more than 80 databases have been updated and 25 obsolete databases have been removed from the list. The complete database list and summaries are available online at the *Nucleic Acids Research* web site, http://nar.oxfordjournals.org/.

COMMENTARY

The current issue of the *Nucleic Acids Research* features 100 new databases and 82 updates for databases that were previously described in NAR or other journals. Inclusion of these databases into the *Nucleic Acids Research* online Molecular Biology Database Collection (http://www.oxfordjournals.org/nar/database/a/, see Supplementary Table S1) brought the total count to more than 1000 databases for the first time in the 12-year history of this list.

Several more databases, recently described in *Bioinformatics* and other journals, were added to the list. The database collection now includes databases from two more countries, Chile and Tunisia [AlterORF (1), a database of alternate open reading frames in prokaryotic genomes, http://www.cienciavida.cl/alterorf/, no. 1044 in the *NAR* Database Collection, and BACTIBASE (2), a database of bacteriocins, natural antimicrobial peptides, http://www.pfba-lab.org/bactibase, no. 1167, respectively]. A year after the only Norwegian database, tinyGRAP, was dropped from the list, there is again a database from Norway, SuperCAT [http://mlstoslo.uio.no/, no. 1135, (3)]. Some of the new databases come with remarkably creative names, such as eggNOG (4), a database of evolutionary genealogy of genes: Non-supervised Orthologous Groups (http://eggnog.embl.de, no. 1068), GEISHA (5), a database of *Gallus* expression *in situ* hybridization analysis (http://geisha.arizona.edu, no. 1173), PhosPh*At* (6), a database of protein phosphorylation sites in *Arabidopsis thaliana* (http://www.plantenergy.uwa.edu.au/applications/phosphat/, no. 1112), and STITCH (7), a search tool for interactions of chemicals (http://stitch.embl.de/, no. 1134). It was interesting to see researchers in different countries coming up with databases covering very similar areas. Examples include MethyCancer [http://methycancer.genomics.org.cn/, no. 1096, (8)] and PubMeth [http://matrix.ugent.be/pubmeth/, no. 1125, (9)] that both examine the links between DNA methylation levels and cancerogenesis; microRNA.org [http://www.microrna.org/, no. 1097, (10)] and miRGator [http://genome.ewha.ac.kr/miRGator/miRGator.html, no. 1098, (11)] that both deal with prediction of microRNA targets and gene expression data; Greglist [http://tubic.tju.edu.cn/greglist/, no. 1082, (12)] and QuadBase, [http://quadbase.igib.res.in/, no. 1126, (13)] that both search for G-quadruplex motifs in the promoters of potentially G-quadruplex regulated genes. Independent creation of these databases shows the importance of the respective topics and ensures friendly competition that should keep these databases in a good shape. Another pair of databases deals with the genomes of pea aphid *Acyrthosiphon pisum* [AphidBase, http://www.aphidbase.com/, no 1152, (14)] and its bacterial endosymbiont *Buchnera* sp. APS [BuchneraBase, http://www.buchnera.org/, no. 1153, (15)]. The combination of the two should provide further insights in the mechanisms of this interesting symbiotic relationship. One of the most remarkable new databases is Déjà vu (http://spore.swmed.edu/dejavu/, no. 1171), a database that uses the eTBLAST tool, recently described in the *NAR* Web Server issue (16), to find highly similar abstracts in bibliographic databases, including MEDLINE. Most of

*To whom correspondence should be addressed. Tel: +1 301 435 5910 Fax: ;+1 301 435 7793; Email: galperin@ncbi.nlm.nih.gov

the citations retrieved that way appear to be relatively benign duplicate publications of the same data by the same authors (including my own comments to the 2006 and 2007 releases of the *NAR* Database Collection). Some highly similar publications, however, come from different authors and look extremely suspicious.

Almost two dozen databases that have been featured in the previous release of the NAR database collection but are no longer maintained have been dropped from the list. One of such casualties was the popular listing of ongoing microbial genome sequencing projects on the TIGR web site (former no. 99). However, this task is carried out by much more comprehensive listings in the GOLD database [http://www.genomesonline.org/, no. 75, (17)] and Entrez Genomes web site [no. 458, (18)]. The recently created diArk database [http://www.diark.org/diark/, no. 1172, (19)] lists eukaryotic genome sequencing projects. Several more databases have been superseded by more advanced or more comprehensive databases. For example, the famous Families of Structurally Similar Proteins (FSSP) database [no. 469, (20)] was superseded by the Dali database [http://ekhidna.biocenter.helsinki.fi/dali/start, no. 442, (21)], while the BayGenomics database [http://baygenomics.ucsf.edu/, no 416, (22)] was superseded by the International Gene Trap Consortium database [http://www.genetrap.org/, no 827, (23)]. Two EBI database projects, the Alternative Splicing Database [ASD, (24)], and the Alternative Transcript Diversity Database [ATD, (25)] have been combined into a single Alternative Splicing and Transcript Diversity database (ASTD, no. 28). In addition, two important and widely used databases had to be removed from the list because they limited access only to the registered users: the Human Gene Mutation Database (HGMD®, http://www.hgmd.cf.ac.uk/, no. 133) and Unified Medical Language System® (UMLS, http://umlsks.nlm.nih.gov/, no. 317).

As a result of all these changes, the current release of the *NAR* online Molecular Biology Database Collection includes 1078 databases, 110 more than the previous one. It is probably useful to reiterate that this listing is by no means exhaustive; it was never intended to represent *all* molecular biology databases, or even all publicly available ones. Rather, the *NAR* Collection is a very selective list, chosen among numerous molecular biology-related databases available all over the world. Most of the databases in the list come from the annual *NAR* database issues and therefore reflect the choice of the editors of these issues, based on scrupulous peer review. Additional databases are selected from publications in other journals, such as *Bioinformatics* or *BMC Bioinformatics*. Finally, this collection includes a relatively small number of databases that have been submitted to a *NAR* database issue, received some positive reviews, but were not accepted for publication because of the highly specialized data and/or narrow target audience. All these databases are subsequently vetted for continuity and those that are not being maintained or updated are being gradually removed from the collection. Summing up, NAR online Molecular Biology Database Collection can itself be viewed as a curated database: each of its entries has been looked at, evaluated and deemed useful for the community.

ACKNOWLEDGEMENTS

I thank Rich Roberts, Alex Bateman and my colleagues at NCBI for helpful comments. This work was supported by the Intramural Research Program of the US National Institutes of Health at the National Library of Medicine. 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. Pedroso,I., Rivera,G., Lazo,F., Chacón,M., Ossandón,F., Veloso,F. and Holmes,D.S. (2008) AlterORF: a database of alternate open reading frames. *Nucleic Acids Res*., **36**, in press (gkm886).
2. Hammami,R., Zouhir,A., Ben Hamida,J. and Fliss,I. (2007) BACTIBASE: a new web-accessible database for bacteriocin characterization. *BMC Microbiol*, **7**, 89.
3. Tourasse,N. and Kolstø,A.-B. (2008) SuperCAT: a supertree database for combined and integrative Multilocus Sequence Typing analysis of the *Bacillus cereus* group of bacteria (including *B. cereus*, *B. anthracis*, and *B. thuringiensis*). *Nucleic Acids Res*., **36**, in press (gkm877).
4. Jensen,L.J., Julien,P., Kuhn,M., von Mering,C., Muller,J., Doerks,T. and Bork,P. (2008) eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res*., **36**, in press (gkm796).
5. Darnell,D.K., Kaur,S., Stanislaw,S., Davey,S., Konieczka,J.H., Yatskievych,T.A. and Antin,P.B. (2007) GEISHA: an in situ hybridization gene expression resource for the chicken embryo. *Cytogenet. Genome Res*., **117**, 30–35.
6. Heazlewood,J., Durek,P., Hummel,J., Selbig,J., Weckwerth,W., Walther,D. and Schulze,W.X. (2008) PhosPhAt: A database of phosphorylation sites in *Arabidopsis thaliana* and a plant-specific phosphorylation site predictor. *Nucleic Acids Res*., **36**, in press (gkm812).
7. Jensen,L.J., Kuhn,M., von Mering,C., Campillos,M. and Bork,P. (2008) STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Res*., **36**, in press (gkm795).
8. He,X., Chang,S., Zhang,J., Zhao,Q., Xiang,H., Kusonmano,K., Yang,L., Sun,Z.S., Yang,H. and Wang,J. (2008) MethyCancer: the database of human DNA methylation and cancer. *Nucleic Acids Res*., **36**, in press (gkm730).
9. Ongenaert,M., Van Neste,L., De Meyer,T., Menschaert,G., Bekaert,S. and Van Criekinge,W. (2008) PubMeth: a cancer methylation database combining text-mining and expert annotation. *Nucleic Acids Res*., **36**, in press (gkm788).
10. Betel,D., Wilson,M., Gabow,A. and Sander,C. (2008) The microRNA.org resource: targets and expression. *Nucleic Acids Res*., **36**, in press (gkm995).
11. Nam,S., Kim,B., Shin,S. and Lee,S. (2008) miRGator: an integrated system for functional annotation of microRNAs. *Nucleic Acids Res*., **36**, in press (gkm829).
12. Zhang,R., Lin,Y. and Zhang,C.T. (2008) Greglist: a database listing potential G-quadruplex regulated genes. *Nucleic Acids Res*., **36**, in press (gkm787).
13. Yadav,V.K., Abraham,J.K., Mani,P., Kulshrestha,R. and Chowdhury,S. (2008) QuadBase: genome-wide database of G4 DNA occurrence and conservation in human, chimpanzee, mouse and rat promoters and 146 microbes. *Nucleic Acids Res*., **36**, in press (gkm781).
14. Gauthier,J.P., Legeai,F., Zasadzinski,A., Rispe,C. and Tagu,D. (2007) AphidBase: a database for aphid genomic resources. *Bioinformatics*, **23**, 783–784.
15. Prickett,M.D., Page,M., Douglas,A.E. and Thomas,G.H. (2006) BuchneraBASE: a post-genomic resource for Buchnera sp. APS. *Bioinformatics*, **22**, 641–642.
16. Errami,M., Wren,J.D., Hicks,J.M. and Garner,H.R. (2007) eTBLAST: a web server to identify expert reviewers,

appropriate journals and similar publications. *Nucleic Acids Res.*, **35**, W12–W15.
17. Liolios,K., Tavernarakis,N., Hugenholtz,P. and Kyrpides,N.C. (2006) The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. *Nucleic Acids Res.*, **34**, D332–D334.
18. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R. *et al.* (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **35**, D5–D12.
19. Odronitz,F., Hellkamp,M. and Kollmar,M. (2007) diArk–a resource for eukaryotic genome research. *BMC Genomics*, **8**, 103.
20. Holm,L. and Sander,C. (1994) The FSSP database of structurally aligned protein fold families. *Nucleic Acids Res.*, **22**, 3600–3609.
21. Holm,L. and Sander,C. (1998) Touring protein fold space with Dali/FSSP. *Nucleic Acids Res.*, **26**, 316–319.
22. Stryke,D., Kawamoto,M., Huang,C.C., Johns,S.J., King,L.A., Harper,C.A., Meng,E.C., Lee,R.E., Yee,A. *et al.* (2003) BayGenomics: a resource of insertional mutations in mouse embryonic stem cells. *Nucleic Acids Res.*, **31**, 278–281.
23. Nord,A.S., Chang,P.J., Conklin,B.R., Cox,A.V., Harper,C.A., Hicks,G.G., Huang,C.C., Johns,S.J., Kawamoto,M. *et al.* (2006) The International Gene Trap Consortium Website: a portal to all publicly available gene trap cell lines in mouse. *Nucleic Acids Res.*, **34**, D642–D648.
24. Stamm,S., Riethoven,J.J., Le Texier,V., Gopalakrishnan,C., Kumanduri,V., Tang,Y., Barbosa-Morais,N.L. and Thanaraj,T.A. (2006) ASD: a bioinformatics resource on alternative splicing. *Nucleic Acids Res.*, **34**, D46–D55.
25. Le Texier,V., Riethoven,J.J., Kumanduri,V., Gopalakrishnan,C., Lopez,F., Gautheret,D. and Thanaraj,T.A. (2006) AltTrans: transcript pattern variants annotated for both alternative splicing and alternative polyadenylation. *BMC Bioinformatics*, **7**, 169.