Published online 25 November 2008 *Nucleic Acids Research, 2009, Vol. 37, Database issue* ***D1–D4***
doi:10.1093/nar/gkn942

Nucleic Acids Research annual Database Issue and the NAR online Molecular Biology Database Collection in 2009

Michael Y. Galperin[1,*] and Guy R. Cochrane[2]

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA and [2]EMBL–European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received November 3, 2008; Revised November 5, 2008; Accepted November 6, 2008

ABSTRACT

The current issue of *Nucleic Acids Research* includes descriptions of 179 databases, of which 95 are new. These databases (along with several molecular biology databases described in other journals) have been included in the *Nucleic Acids Research* online Molecular Biology Database Collection, bringing the total number of databases in the collection to 1170. In this introductory comment, we briefly describe some of these new databases and review the principles guiding the selection of databases for inclusion in the *Nucleic Acids Research* annual Database Issue and the *Nucleic Acids Research* online Molecular Biology Database Collection. The complete database list and summaries are available online at the *Nucleic Acids Research* web site (http://nar.oxfordjournals.org/).

THE 2009 DATABASE ISSUE

The 2009 *Nucleic Acids Research* (NAR) annual Database Issue is the 16th in a series that started in July 1993 with 24 database papers. This current issue comprises 179 papers describing 95 new databases and 84 status updates on databases that were previously described in NAR or other journals. These databases (along with further molecular biology databases that have been described in other journals) have been included into the NAR online Molecular Biology Database Collection (http://www.oxfordjournals.org/nar/database/a/), bringing the total number of databases in the collection to 1170 (16 obsolete databases have been removed from the list). The list of countries represented in the online collection has also increased through the inclusion of the first Argentinean database, TcSNP (http://snps.tcruzi.org), a database of genetic variation in *Trypanosoma cruzi* (1).

On several occasions, we have included in the Database Issue two or more databases that have similar coverage. This issue features, for example, three different databases of tRNA sequences identified in the genomes of various organisms. Two of the papers describe recent updates to the Genomic tRNA Database (GtRNAdb, http://gtrnadb.ucsc.edu/), maintained at Todd Lowe's lab at the University of California—Santa Cruz (2), and to the compilation of tRNA sequences, originally created by Mathias Sprinzl at the University of Bayreuth and currently maintained as the Transfer RNA database [tRNAdb, http://trnadb.bioinf.uni-leipzig.de (3)] by a consortium that includes three more groups at the Universities of Leipzig, Marburg and Strasbourg. The third paper (4) describes a new database, tRNA Gene DataBase Curated by Experts (tRNADB-CE, http://trna.nagahama-i-bio.ac.jp), compiled by Takashi Abe and colleagues at the Nagahama Institute of Bio-Science and Technology in Shiga Prefecture, Japan. The Japanese team report that they have found as much as 4% discordance in tRNA predictions from three different programs and provide manual reconciliation of these results. In addition, the *rrn*DB database (http://ribosome.mmg.msu.edu/rrndb/), maintained by Thomas Schmidt and colleagues at Michigan State University (5), lists the numbers of rRNA and tRNA genes in various prokaryotic genomes. In our opinion, the availability of these databases ensures friendly competition, helps ensure accurate information and benefits the user by providing an unbiased assessment of tRNA predictions in any given organism.

Likewise, the current Database Issue features two different databases of predicted microbial operons. One paper offers an update on the popular OperonDB database (http://www.cbcb.umd.edu/), originally created in 2001 by Steven Salzberg and colleagues (6) at

*To whom correspondence should be addressed. Tel: +1 301 435 5910; Fax: +1 301 435 7793; Email: nardatabase@gmail.com

The Institute for Genomic Research (TIGR) and currently maintained by Salzberg's group at the University of Maryland (7). The other details the Database of prOkaryotic OpeRons [DOOR, http://csbl1.bmb.uga.edu/OperonDB/ (8)], which has been created by Ying Xu and colleagues (9) at the University of Georgia in Athens and utilizes an alternative algorithm for operon prediction. Together with the operon prediction data in the DOE's MicrobesOnline database [http://www.microbesonline.org/operons/ (10)], which relies on yet another prediction algorithm (11), these databases provide three different sets of predictions for the same genes in the same organisms and give the user an opportunity to compare sets and make an informed choice on the prediction that can be trusted.

The importance of studying genomes of pathogens causing emerging and re-emerging diseases prompted inclusion in this issue of such databases as GiardiaDB, PlasmoDB, TrichDB and VectorBase (12–14), products of the Bioinformatics Resource Centers, supported by the US National Institutes of Health, National Institute of Allergy and Infectious Diseases (http://www3.niaid.nih.gov/research/resources/brc/).

On some occasions, the number of papers dedicated to the same topic had to be limited. This year, for example, there were nine submissions of papers describing new databases dealing with microRNAs, not to mention the updates to the popular Rfam and TarBase databases (15,16) and a further seven microRNA databases already included in the NAR Database Collection. Four of these submissions have been accepted (17–20), based largely on volume of manually curated data and convenience for naïve users, but several otherwise viable databases have had to be rejected.

Several databases that have been featured in the previous release of the NAR Database Collection have been removed from the list. One such casualty was the once popular Genome DataBase (GDB) featured in a number of NAR publications (21,22). After its initial success, this database struggled to find its niche, was moved from the Johns Hopkins University in Baltimore, Maryland, to the Hospital for Sick Children in Toronto, Canada, back to Johns Hopkins and finally came to rest in Research Triangle Institute (RTI International) in Research Triangle Park, North Carolina (23), where its operation closed down in 2008 after control of the project reverted to Johns Hopkins. Two other popular databases, *e*MOTIF (24) and HSSP (25), no longer support browsing, although their content remains available for download.

CRITERIA FOR INCLUSION

This Database Issue has been produced by a new team. After five very successful years at the helm, Alex Bateman retired as editor of the NAR Database Issue. Michael Galperin, who was previously responsible for the NAR Database Collection, became the new editor and Guy Cochrane came on board as curator of the NAR Database Collection. We are committed to continuing the policies of Alex Bateman and previous NAR database editors (Richard J. Roberts, Christian Burks and Andreas D. Baxevanis) that brought the NAR annual Database Issue and the NAR Database Collection to their current prominence.

Given the unique position of the NAR Database Issue as a premier forum for the publication of molecular biology databases and the ever-increasing influx of proposed submissions, we feel that it would be useful to reiterate the guiding principles that we use in the selection of databases for inclusion in the NAR Database Issue and the Database Collection. First, coverage is by no means exhaustive; the NAR Database Issue and Database Collection were never intended to represent 'all' molecular biology databases, or even all of those that were publicly available. Rather, the NAR Database Issue features thoroughly curated databases that are expected to be of interest to a wide variety of biologists, primarily bench scientists. The key criteria for selection are the general utility of the database to the scientific community, comprehensiveness of coverage and degree of value added (usually in the form of manual curation) in the production of the database. We are primarily interested in web-accessible databases that offer carefully curated data that are not available elsewhere. Data warehouses, portals, cross-platform search tools and visualization tools are more suitable for such journals as *Bioinformatics*, *BMC Bioinformatics* or *Database: The Journal of Biological Databases and Curation* (http://www.oxfordjournals.org/our_journals/databa/), recently launched by our publisher, Oxford University Press. We will consider, however, data portals that add value to the user by providing a convenient one-stop source of disparate data not available elsewhere and supplement this with convenient search tools and easy-to-use visualization. We would generally avoid accepting databases on gene expression, as the underlying data must be submitted to ArrayExpress (26) and/or GEO (27). Similarly, we would avoid accepting new EST databases, particularly those dealing with individual species, as these data have a home in the DDBJ, GenBank and European Nucleotide Archive databases.

Another important issue is consideration of so-called 'boutique' databases, covering relatively narrow topics. The key judgement here is whether or not the database in question is likely to be useful to those beyond specialists in the field that it covers and could serve as a useful introduction for the general public or scientists unfamiliar with the field. As an example, one of the microRNA databases mentioned above, miR2Disease [http://mlg.hit.edu.cn:8080/miR2Disease (18)] created by Yadong Wang and colleagues at Harbin Institute of Technology in Harbin, China, and Indiana University in Indianapolis, received high marks from the reviewers for linking two important areas and for its potential to introduce pathologists and other clinicians to the world of microRNAs. On the other hand, a large number of interesting plant databases has had to be rejected because the databases were designed to serve only very limited user groups. Plant databases have typically only been accepted when they appear to offer the potential to be of interest to scientists studying general biological problems, such as regulation of gene expression, protein–protein interactions, comparative genomics, and other subjects with universal appeal.

In addition to the scientific quality of a database and its general utility to the scientific community, reviewers are also asked to evaluate whether the database is well curated and is likely to be maintained for a long period of time. Submission of a paper to the NAR Database Issue implies a commitment to maintaining the database on the part of the senior author and the host institution. Once the database paper has been published, graduation of a particular student or a postdoc is not considered a valid reason to discontinue maintaining the database or to move it out of the public domain. Should this happen, the respective senior authors (and in some cases, their host institutions) will be prevented from publishing new papers in the NAR Database Issue.

Another important requirement is for a database not to have been described elsewhere. Authors' desire to popularize their work sometimes results in the simultaneous submission of two different descriptions of a database to NAR and a more specialized journal, respectively. While this may seem trivial, we consider it a matter of principle that NAR papers should be unique. In certain rare cases, upon the request of authors, we may consider including in the NAR Database Issue a paper that was published elsewhere fewer than two years earlier. For example, owing to the importance of the IUBMB Enzyme Classification to a wide variety of biologists, a description of the ExplorEnz database (http://www.enzyme-database.org) has been included in the 2009 NAR Database Issue only a year after the database was first introduced in another publication (28,29). In most cases, however, such duplicate submissions will be rejected, sometimes in the later stages of the review process. This year, this has happened to several otherwise viable databases. Rejection of papers from the Database Issue does not necessarily disqualify these databases from inclusion into the NAR online collection but reduces the chances.

Because the key criterion for inclusion is usefulness of the database to the community, the NAR Database Issue sometimes features unorthodox databases that the editors deem valuable, even if they do not fit standard expectations. For example, the database of highly similar Medline citations (*Déjà vu*, http://spore.swmed.edu/dejavu/), already mentioned in the previous comment (30), has been included into the Database Issue (31), as, in addition to its primary goal, it provides the useful service of allowing the users to search for experts in certain areas, the most appropriate journals in which to publish their work and who potential reviewers may be. Another such unorthodox database in the current issue is BodyParts3D [http://lifesciencedb.jp/ag/bp3d/ (32)], a database of morphological and geometrical knowledge in human anatomy and a visualization tool for 3D reconstruction of the human body that, among other applications, will have huge utility in the mapping of gene-expression data onto tissues.

To simplify the review process, all submissions to the Database Issue are pre-screened by the editor, Dr Michael Galperin (nardatabase@gmail.com). In 2008, the rejection rate of this pre-screening was lower than 50%, which resulted in an unusually high numbers of potential papers and of papers that were ultimately rejected based on reviewers' comments. In future, we will employ stricter criteria for pre-screening, such that submissions will be invited only from those databases with the appropriate commitments to longevity and sustained value to users that have a realistic chance of surviving review.

For update papers, inclusion criteria are even stricter. Only updates from the most popular databases, such as GenBank, the European Nucleotide Archive, DDBJ and UniProt, are published every year. From all other databases, updates can be submitted every other year, but only when there are significant new developments that warrant the publication. The decision on publication of any update paper will be made on a case-by-case basis, considering the importance of the database for the community, the amount of new material, improvements in data presentation and other measures.

The NAR online Molecular Database Collection gets most of its content from the publications in annual NAR Database Issues and database papers published in *Bioinformatics*, our sister journal. As a result, it is a very selective list of databases that have gone through scrupulous peer review. The database list is annually vetted for continuity and obsolete databases are purged from the collection. We strive to maintain the NAR online Molecular Biology Database Collection as a curated list that features the best publicly available molecular biology databases.

ACKNOWLEDGEMENTS

We thank Sir Richard Roberts and Alex Bateman for many helpful comments; Patricia Anderson, Martine Bernardes-Silva, Karen Otto and Gail Welsh for excellent editorial assistance, and the Oxford University Press team lead by Claire Bird and Radha Dutia for their patience and great effort in compiling this issue.

FUNDING

Intramural Research Program of the US National Institutes of Health (to M.Y.G.); European Molecular Biology Laboratory (to G.R.C.). The Open Access publication charges for this article were waived by Oxford University Press.

Conflict of interest statement. The authors' opinions do not necessarily reflect the views of their respective institutions.

REFERENCES

1. Ackermann,A.A., Carmona,S.J. and Agüero,F. (2009) TcSNP: a database of genetic variation in *Trypanosoma cruzi*. *Nucleic Acids Res.*, **37**, D544–D549.
2. Chan,P.P. and Lowe,T.M. (2009) GtRNAdb: A database of transfer RNA genes detected in genome sequence. *Nucleic Acids Res.*, **37**, D93–D97.
3. Jühling,F., Mörl,M., Hartmann,R.K., Sprinzl,M., Stadler,P.F. and Pütz,J. (2009) tRNAdb 2009: compilation of tRNA sequences and tRNA genes. *Nucleic Acids Res.*, **37**, D159–D162.
4. Abe,T., Ikemura,T., Ohara,Y., Uehara,H., Kinouchi,M., Kanaya,S., Yamada,Y., Muto,A. and Inokuchi,H. (2009)

tRNADB-CE: tRNA gene database curated manually by experts. *Nucleic Acids Res.*, **37**, D163–D168.
5. Lee,Z.M., Bussema,C. 3rd and Schmidt,T.M. (2009) rrnDB: documenting the number of rRNA and tRNA genes in bacteria and archaea. *Nucleic Acids Res.*, **37**, D489–D493.
6. Ermolaeva,M.D., White,O. and Salzberg,S.L. (2001) Prediction of operons in microbial genomes. *Nucleic Acids Res.*, **29**, 1216–1221.
7. Pertea,M., Ayanbule,K., Smedinghoff,M. and Salzberg,S.L. (2009) OperonDB: a comprehensive database of predicted operons in microbial genomes. *Nucleic Acids Res.*, **37**, D479–D482.
8. Mao,F., Dam,P., Chou,J., Olman,V. and Xu,Y. (2009) DOOR: a database for prokaryotic operons. *Nucleic Acids Res.*, **37**, D459–D463.
9. Dam,P., Olman,V., Harris,K., Su,Z. and Xu,Y. (2007) Operon prediction using both genome-specific and general genomic information. *Nucleic Acids Res.*, **35**, 288–298.
10. Alm,E.J., Huang,K.H., Price,M.N., Koche,R.P., Keller,K., Dubchak,I.L. and Arkin,A.P. (2005) The MicrobesOnline Web site for comparative genomics. *Genome Res.*, **15**, 1015–1022.
11. Price,M.N., Huang,K.H., Alm,E.J. and Arkin,A.P. (2005) A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res.*, **33**, 880–892.
12. Aurrecoechea,C., Brestelli,J., Brunk,B.P., Dommer,J., Fischer,S., Gajria,B., Gao,X., Gingle,A., Grant,G., Harb,O.S. *et al.* (2009) PlasmoDB: a functional genomic database for malaria parasites. *Nucleic Acids Res.*, **37**, D539–D543.
13. Aurrecoechea,C., Brestelli,J., Brunk,B.P., Carlton,J.M., Dommer,J., Fischer,S., Gajria,B., Gao,X., Gingle,A., Grant,G. *et al.* (2009) GiardiaDB and TrichDB: integrated genomic resources for the eukaryotic protist pathogens *Giardia lamblia* and *Trichomonas vaginalis*. *Nucleic Acids Res.*, **37**, D526–D530.
14. Lawson,D., Arensburger,P., Atkinson,P., Besansky,N.J., Bruggner,R.V., Butler,R., Campbell,K.S., Christophides,G.K., Christley,S., Dialynas,E. *et al.* (2007) VectorBase: a home for invertebrate vectors of human pathogens. *Nucleic Acids Res.*, **35**, D503–D505.
15. Gardner,P.P., Daub,J., Tate,J.G., Nawrocki,E.P., Kolbe,D.L., Lindgreen,S., Wilkinson,A.C., Finn,R.D., Griffiths-Jones,S., Eddy,S.R. *et al.* (2009) Rfam: updates to the RNA families database. *Nucleic Acids Res.*, **37**, D136–D140.
16. Papadopoulos,G.L., Reczko,M., Simossis,V.A., Sethupathy,P. and Hatzigeorgiou,A.G. (2009) The database of experimentally supported targets: a functional update of TarBase. *Nucleic Acids Res.*, **37**, D155–D158.
17. Gerlach,D., Kriventseva,E.V., Rahman,N., Vejnar,C.E. and Zdobnov,E.M. (2009) miROrtho: computational survey of microRNA genes. *Nucleic Acids Res.*, **37**, D111–D117.
18. Jiang,Q., Wang,Y., Hao,Y., Juan,L., Teng,M., Zhang,X., Li,M., Wang,G. and Liu,Y. (2009) miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res.*, **37**, D98–D104.
19. Taccioli,C., Fabbri,E., Visone,R., Volinia,S., Calin,G.A., Fong,L.Y., Gambari,R., Bottoni,A., Acunzo,M., Hagan,J. *et al.* (2009) UCbase & miRfunc: a database of ultraconserved sequences and microRNA function. *Nucleic Acids Res.*, **37**, D41–D48.
20. Xiao,F., Zuo,Z., Cai,G., Kang,S., Gao,X. and Li,T. (2009) miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res.*, **37**, D105–D110.
21. Pearson,P.L. (1991) The genome data base (GDB)—a human gene mapping repository. *Nucleic Acids Res.*, **19**, 2237–2239.
22. Letovsky,S.I., Cottingham,R.W., Porter,C.J. and Li,P.W. (1998) GDB: the Human Genome Database. *Nucleic Acids Res.*, **26**, 94–99.
23. Bonetta,L. (2001) Sackings leave gene database floundering. *Nature*, **414**, 384.
24. Huang,J.Y. and Brutlag,D.L. (2001) The EMOTIF database. *Nucleic Acids Res.*, **29**, 202–204.
25. Sander,C. and Schneider,R. (1993) The HSSP data base of protein structure-sequence alignments. *Nucleic Acids Res.*, **21**, 3105–3109.
26. Parkinson,H., Kapushesky,M., Shojatalab,M., Abeygunawardena,N., Coulson,R., Farne,A., Holloway,E., Kolesnykov,N., Lilja,P., Lukk,M. *et al.* (2007) ArrayExpress—a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res.*, **35**, D747–D750.
27. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Rudnev,D., Evangelista,C., Kim,I.F., Soboleva,A., Tomashevsky,M., Marshall,K.A. *et al.* (2009) NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res.*, **37**, D885–D890.
28. McDonald,A.G., Boyce,S., Moss,G.P., Dixon,H.B. and Tipton,K.F. (2007) ExplorEnz: a MySQL database of the IUBMB enzyme nomenclature. *BMC Biochem.*, **8**, 14.
29. McDonald,A.G., Boyce,S. and Tipton,K.F. (2009) ExplorEnz: the primary source of the IUBMB enzyme list. *Nucleic Acids Res.*, **37**, D593–D597.
30. Galperin,M.Y. (2007) The Molecular Biology Database Collection: 2007 update. *Nucleic Acids Res.*, **35**, D3–D4.
31. Errami,M., Sun,Z., Long,T.C., George,A.C. and Garner,H.R. (2009) Deja vu: a database of highly similar citations in the scientific literature. *Nucleic Acids Res.*, **37**, D921–D924.
32. Mitsuhashi,N., Fujieda,K., Tamura,T., Kawamoto,S., Takagi,T. and Okubo,K. (2009) BodyParts3D: 3D structure database for anatomical concepts. *Nucleic Acids Res.*, **37**, D782–D785.