D234–D236 *Nucleic Acids Research, 2010, Vol. 38, Database issue* *Published online 21 October 2009*
doi:10.1093/nar/gkp874

REBASE—a database for DNA restriction and modification: enzymes, genes and genomes

Richard J. Roberts*, Tamas Vincze, Janos Posfai and Dana Macelis

New England Biolabs, Inc., 240 County Road, Ipswich, MA 01938, USA

Received September 14, 2009; Revised September 29, 2009; Accepted September 30, 2009

ABSTRACT

REBASE is a comprehensive database of information about restriction enzymes, DNA methyltransferases and related proteins involved in the biological process of restriction–modification (R–M). It contains fully referenced information about recognition and cleavage sites, isoschizomers, neoschizomers, commercial availability, methylation sensitivity, crystal and sequence data. Experimentally characterized homing endonucleases are also included. The fastest growing segment of REBASE contains the putative R–M systems found in the sequence databases. Comprehensive descriptions of the R–M content of all fully sequenced genomes are available including summary schematics. The contents of REBASE may be browsed from the web (http://rebase.neb.com) and selected compilations can be downloaded by ftp (ftp.neb.com). Additionally, monthly updates can be requested via email.

OVERVIEW

The previous description of REBASE in the 2007 NAR Database Issue (1) described 3805 biochemically or genetically characterized restriction–modification (R–M) systems and included an analysis of approximately 400 bacterial and archaeal genomes that had been deposited in the RefSeq Database of GenBank (2,3). Analysis of the available sequence information in GenBank led to the prediction of 2709 restriction enzyme (R) genes and 4485 DNA methyltransferase (M) genes. These numbers have now risen to 4990 R genes and 8080 M genes of which 3511 R and 5497 M genes have arisen from the 1050 completely sequenced bacterial and archaeal genomes. These putative R–M system genes are given systematic names according to the agreed upon nomenclature rules (4). The names all carry the suffix 'P' to indicate their putative status. In many cases, the recognition specificity of these systems can be assigned with some degree of confidence because of their similarity to biochemically well-characterized enzymes.

The REBASE web site (http://rebase.neb.com) summarizes all information known about every restriction enzyme and any associated proteins. This includes the recognition sequences, cleavage sites, source, commercial availability, sequence data, crystal structure information, isoschizomers and methylation sensitivity. Within the reference section of REBASE, links are maintained to the full text of all papers whenever they are readily available on the web. Also, there is extensive reciprocal cross-referencing between REBASE and NCBI, including links to GenBank and PubMed and NCBI's LinkOut utility. Links to other major databases such as UniProt (5), PDB (6) and Pfam (7) are also maintained. There are currently 3945 biochemically or genetically characterized restriction enzymes in REBASE and of the 3834 Type II restriction enzymes, 299 distinct specificities are known. Six hundred and forty one restriction enzymes are commercially available, including 235 distinct specificities.

As shown in Figure 1, the rate of discovery of new putative restriction and modification genes is rising rapidly. In contrast, the rate at which candidates are being characterized biochemically has actually dropped to the level it was three decades ago. Nevertheless, because of the large number of sequenced examples of biochemically characterized restriction systems, the putative recognition sequences of predicted restriction enzymes and DNA methyltransferases can be inferred. Currently, all new sequences entering GenBank are checked using data mining techniques for the presence of R–M systems and, following extensive manual checking, the resulting inferences are all included within REBASE where they are clearly marked as predictions. When analyzing DNA sequence data, it is the DNA methyltransferase genes that are the more reliable indicators of an R–M system and the presence, proper order and characteristic spacing of well-conserved motifs that are used to suggest likely candidates.

It should be noted that at the present time it is not possible to distinguish DNA methyltransferases reliably enough to be completely confident in the assignments.

*To whom correspondence should be addressed. Tel: +1 978 380 7405; Fax: +1 978 380 7406; Email: roberts@neb.com

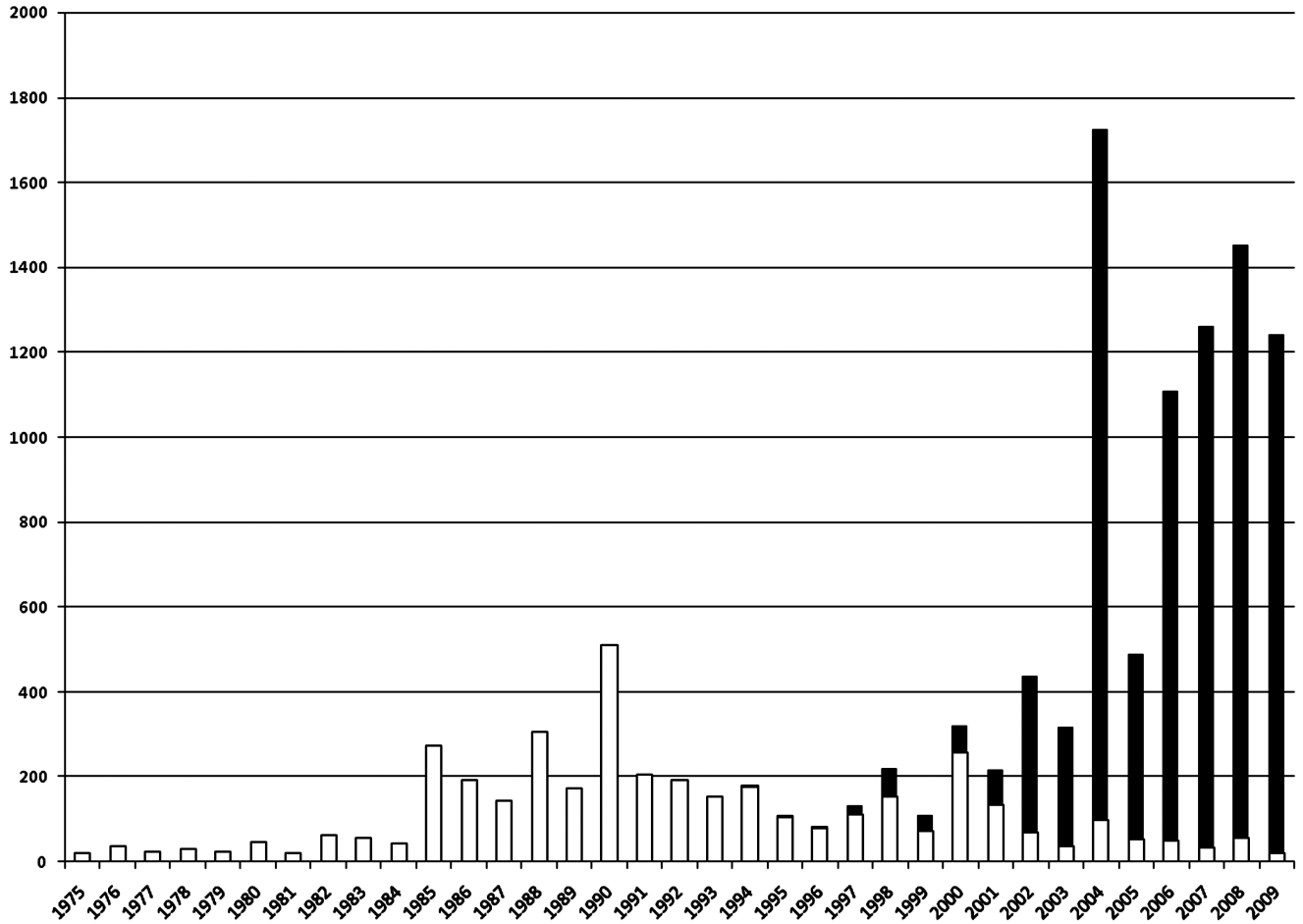

Figure 1. The graph shows the numbers of R–M systems entering REBASE since its inception in 1975. The open bars show systems that have been characterized either biochemically or genetically. The black bars show the increasing accumulation of potential R–M systems that have been found by bioinformatic analysis of sequences in GenBank. The surge in 2004 represents the addition of metagenomic sequences from the Sargasso Sea collecting expedition (9).

Some RNA and protein methyltransferases can sometimes be confused for DNA methyltransferases as is widely reflected by the annotations found in GenBank files. In general, REBASE takes a liberal approach and includes all likely candidates until it becomes clear that non-DNA methyltransferases have been included erroneously and then these are culled from the database. The more widely divergent genes that encode the restriction enzymes always reside close to the genes for their cognate methyltransferases, but often they cannot be recognized directly because they are a rapidly evolving set of genes and frequently lack any sequence similarity to any other genes in GenBank. However, other methods can sometimes be used to infer their presence such as the analysis of shotgun sequence data from which missing clones can be inferred to be caused by the presence of active restriction enzyme genes (8).

Given the wealth of experimental data, both published and unpublished, contained within REBASE, it can be an especially valuable resource during the annotation of bacterial and archaeal genomes. With the plethora of restriction systems that occur in all sequenced microbial genomes, annotators are encouraged to use the resources of the REBASE database or to contact the REBASE staff if help is needed. Custom analyses of unpublished genome sequence data are carried out upon request.

From the REBASE web site users have a variety of resources available that facilitate the analysis of sequence information including tools for analyzing sequences (REBASE tools) that allow restriction enzyme recognition sites to be found in submitted sequences (NEBcutter) and an implementation of BLAST to allow searching against all sequences in REBASE. Specialty lists of sequence data (REBASE lists) such as all known Type II restriction enzyme genes, all known Type I specificity subunit genes, etc., are available for download.

The coming year will see some major additions to REBASE in terms of new sequence acquisitions, such as the inclusion of all metagenomic sequence data (only partially analyzed to date) and a tool to permit users to perform their own analysis of newly sequenced genomes.

ACKNOWLEDGEMENTS

Special thanks are due to the many individuals who have so kindly contributed their unpublished results for

inclusion in this compilation and to the REBASE users who continue to guide our efforts with their helpful comments. We are especially grateful to Karen Otto for administrative help.

FUNDING

National Library of Medicine (LM04971); New England Biolabs, Inc. Funding for open access charge: New England Biolabs; National Institutes of Health grant.

Conflict of interest statement. None declared.

REFERENCES

1. Roberts,R.J., Vincze,T., Posfai,J. and Macelis,D. (2007) REBASE—enzymes and genes for DNA restriction and modification. *Nucleic Acids Res.*, **35**, D269–D270.
2. Benson,D.A., Karsch-Mizrachi,I., Lipmann,D.J., Ostell,J. and Sayers,E.W. (2009) GenBank. *Nucleic Acids Res.*, **37**, D26–D31.
3. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
4. Roberts,R.J., Belfort,M., Bestor,T., Bhagwat,A.S., Bickle,T.A., Bitinaite,J., Blumenthal,R.M., Degtyarev,S.K., Dryden,D.T.F., Dybvig,K. *et al.* (2003) A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Res.*, **31**, 1805–1812.
5. The UniProt Consortium. (2009) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **37**, D169–D174.
6. Kouranov,A., Xie,L., de la Cruz,J., Chen,L., Westbrook,J., Bourne,P.E. and Berman,H.M. (2006) The RCSB PDB information portal for structural genomics. *Nucleic Acids Res.*, **34**, D302–D305.
7. Finn,R.D., Tate,J., Mistry,J., Coggill,P.C., Sammut,S.J., Hotz,H.R., Ceric,G., Forslund,K., Eddy,S.R., Sonnhammer,E.L. *et al.* (2008) The Pfam protein families database. *Nucleic Acids Res.*, **36**, D281–D288.
8. Zheng,Y., Posfai,J., Morgan,R.D., Vincze,T. and Roberts,R.J. (2009) Using shotgun sequence data to find active restriction enzyme genes. *Nucleic Acids Res.*, **37**, e1.
9. Venter,J.C., Remington,K., Heidelberg,J.F., Halpern,A.L., Rusch,D., Eisen,J.A., Wu,D., Paulsen,I., Nelson,K.E., Nelson,W. *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66–74.