D118–D121 *Nucleic Acids Research, 2009, Vol. 37, Database issue* *Published online 14 October 2008*
doi:10.1093/nar/gkn710

MODOMICS: a database of RNA modification pathways. 2008 update

Anna Czerwoniec[1], Stanislaw Dunin-Horkawicz[2], Elzbieta Purta[3,4], Katarzyna H. Kaminska[3], Joanna M. Kasprzak[1], Janusz M. Bujnicki[1,3], Henri Grosjean[5] and Kristian Rother[3,*]

[1]Bioinformatics Laboratory, Institute of Molecular Biology and Biotechnology, Adam Mickiewicz University, Umultowska 89, PL-61-614 Poznan, Poland, [2]Max Planck Institute for Developmental Biology, Department 1, Protein Evolution Spemannstr. 35, 72076 Tuebingen, Germany, [3]Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology, Ks. Trojdena 4, PL-02-190 Warsaw, Poland, [4]Institute of Biochemistry and Biophysics PAS, Pawinskiego 5a, 02-106 Warsaw and [5]IGM, Univ Paris-Sud, UMR 8621, Orsay, F 91405, France

Received September 15, 2008; Accepted September 29, 2008

ABSTRACT

MODOMICS, a database devoted to the systems biology of RNA modification, has been subjected to substantial improvements. It provides comprehensive information on the chemical structure of modified nucleosides, pathways of their biosynthesis, sequences of RNAs containing these modifications and RNA-modifying enzymes. MODOMICS also provides cross-references to other databases and to literature. In addition to the previously available manually curated tRNA sequences from a few model organisms, we have now included additional tRNAs and rRNAs, and all RNAs with 3D structures in the Nucleic Acid Database, in which modified nucleosides are present. In total, 3460 modified bases in RNA sequences of different organisms have been annotated. New RNA-modifying enzymes have been also added. The current collection of enzymes includes mainly proteins for the model organisms *Escherichia coli* and *Saccharomyces cerevisiae*, and is currently being expanded to include proteins from other organisms, in particular Archaea and *Homo sapiens*. For enzymes with known structures, links are provided to the corresponding Protein Data Bank entries, while for many others homology models have been created. Many new options for database searching and querying have been included. MODOMICS can be accessed at http://genesilico.pl/modomics.

INTRODUCTION

Numerous chemical changes in RNA nucleotides are introduced by enzymatic modifications in the process of RNA maturation. The location and distribution of various types of modification vary greatly between different RNA molecules, organisms and organelles. Recent discoveries document that this field has developed rapidly. Modifications have been found to occur in microRNA (1). New modifications (i.e. new chemical structures) have been found (2), and biosynthesis pathways of known modifications have been elucidated (3). In many cases, the biochemical and physiological roles of modifications have been found, e.g. in the decoding process for modifications in tRNA (4). Finally, numerous new RNA-modifying enzymes have been identified, including a number of rRNA methyltransferases (5–7).

To adequately represent this rapid accumulation of knowledge, we have added both to the variety and volume of data in the MODOMICS database. The most significant improvements are addition of modifications in rRNA with their positions, updates of the according modification enzymes and pathways, 3D structures of modifications and structures of many modification enzymes, including a collection of homology models for enzymes with no experimental structure available. The data has been formalized to a higher extent, resulting in development of an ontology for modified bases, and flat file parsers that make the data more accessible for batch download and further analyses.

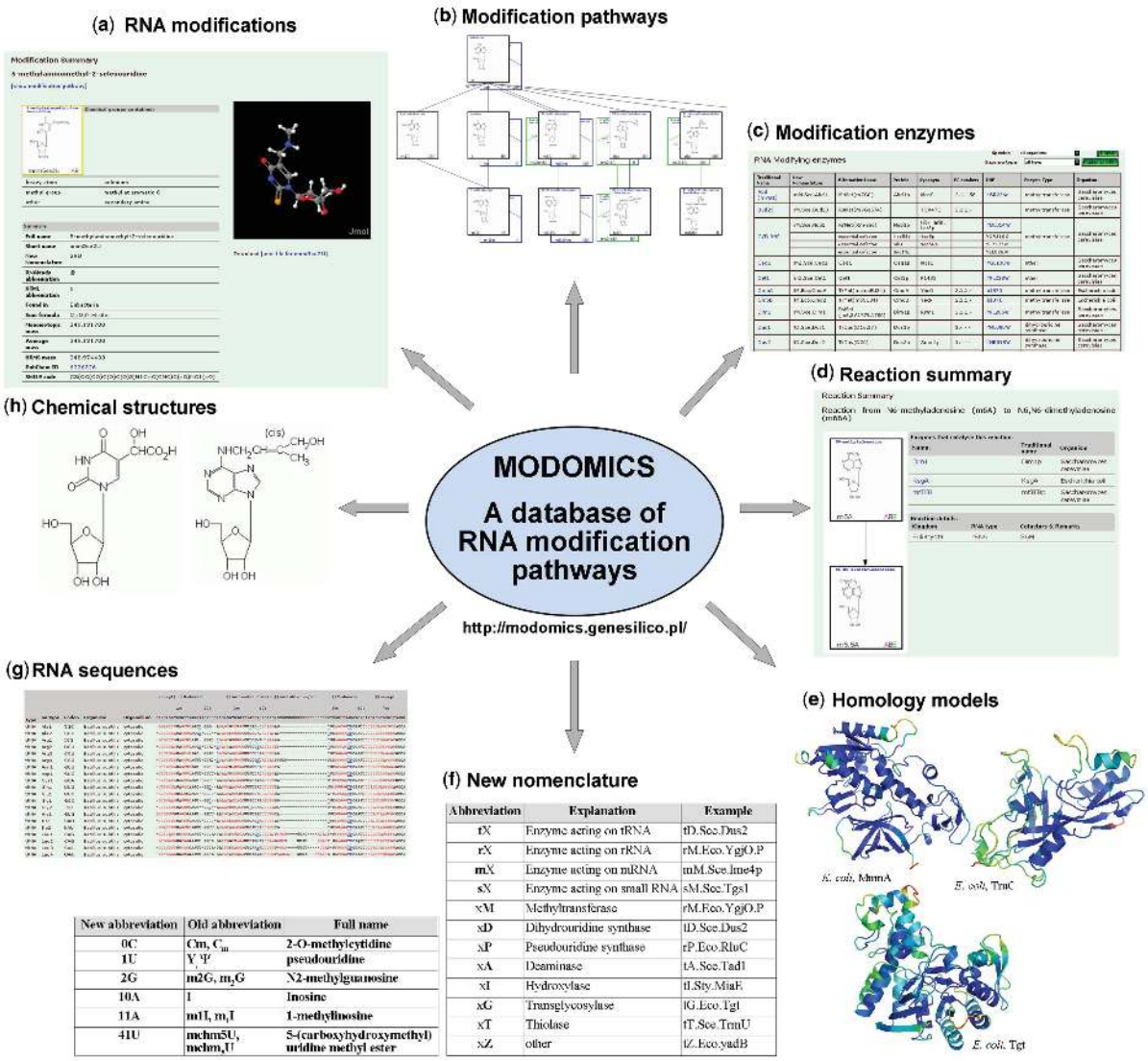
*To whom correspondence should be addressed. Tel: +48-22 597 0752; Fax: +48 22 597 0715; Email: krother@genesilico.pl
Correspondence may also be addressed to Henri Grosjean. Tel: +33(0)1-6915 4637; Fax: +33(0)1-6915 4629; Email: henri.grosjean@igmots.u-psud.fr; Janusz M. Bujnicki. Tel: +48-22 597 0750; Fax: +48 22 597 0715; Email: iamb@genesilico.pl

DATABASE CONTENT

The MODOMICS database (http://modomics.genesilico.pl) was developed to house and distribute collections of RNA modification pathways, chemical structures of modified nucleosides, RNA sequences containing these modifications and enzymes responsible for individual reactions. MODOMICS was created as a single resource to organize and present all these data in a convenient and easily understandable way. An overview of the data stored in MODOMICS is given in Figure 1.

MODIFICATIONS

At present, MODOMICS contains 119 different modifications that have been identified in RNA molecules. A typical database entry for a modified nucleoside presents basic chemical properties, the phylogenetic distribution (with respect to Domains of Life), and the type of RNA where the modification is found. The list of modifications can be browsed by their names, the standard bases from which they originate, and the chemical groups they contain. The available details contain full and short names, the sum formula, and—to facilitate MS analyses of modified RNA—their monoisotopic and average masses. The chemical structures of modified nucleosides are represented by 1D SMILE codes, 2D structure plots and 3D structures in the Protein Data Bank format displayed interactively on the website by a Jmol applet. Reactions linking a modified nucleoside to its precursor(s) are annotated separately. Each reaction is specified by a chemical type (methylation, aminoacylation, etc.), and the



New abbreviation	Old abbreviation	Full name
0C	C_m, C_m	2-O-methylcytidine
1U	Y, Ψ	pseudouridine
2G	m2G, m_2G	N2-methylguanosine
10A	I	Inosine
11A	m1I, m_1I	1-methylinosine
41U	mchm5U, mchm,U	5-(carboxyhydroxymethyl) uridine methyl ester

Abbreviation	Explanation	Example
tX	Enzyme acting on tRNA	tD.Sce.Dus2
rX	Enzyme acting on rRNA	rM.Eco.YgjO.P
mX	Enzyme acting on mRNA	mM.Sce.Ime4p
sX	Enzyme acting on small RNA	sM.Sce.Tgs1
xM	Methyltransferase	rM.Eco.YgjO.P
xD	Dihydrouridine synthase	tD.Sce.Dus2
xP	Pseudouridine synthase	tP.Eco.RluC
xA	Deaminase	tA.Sce.Tad1
xI	Hydroxylase	tI.Sty.MiaE
xG	Transglycosylase	tG.Eco.Tgt
xT	Thiolase	tT.Sce.TrmU
xZ	other	tZ.Eco.yadB

Figure 1. Contents of the MODOMICS database. (**a**) Detailed report on the modification mnm5se2U. (**b**) A fragment of the uridine modification pathway. In addition to A, C, G and U, a separate graph for queuosine is available. (**c**) Six of 124 modification enzymes, as seen on the web. (**d**) The methylation of m^6A (6A according to the new nomenclature) to the hypermodified base $m^{6,6}A$ (7A). (**e**) Three out of 28 model structures for modification enzymes: *E. coli* MnmA; tT.Eco.MnmA according to the new nomenclature); *E. coli* TruC (tP.Eco.TruC); *E. coli* Tgt (tG.Eco.Tgt)). (**f**) New nomenclature for modified bases and modification enzymes. (**g**) Alignment of tRNAs with modified positions indicated. (**h**) Chemical structures of the chm^5U (36U) and i^6A (19A) modifications.

organism and RNA classes that are known to serve as substrates. Because a single nucleotide can be modified more than once—leading to hypermodified nucleosides—the reactions form complex pathways. In MODOMICS, pathways originating from a particular nucleoside can be viewed as a whole, or the users can 'zoom' onto subpathways involving a particular modification.

RNA MODIFYING ENZYMES

MODOMICS contains information about more than 100 different enzymes and co-factors determined experimentally as well as those predicted based on various predictions. The current collection of enzymes includes mainly proteins for model organisms *Escherichia coli* and *Saccharomyces cerevisiae*, and is currently being expanded to include proteins from Archaea and *Homo sapiens*. It must be emphasized that amino acid sequences are known for most, but not all enzymes that have been characterized biochemically or whose existence is predicted based on the knowledge of a reaction product. For some known or putative biochemical activities, the corresponding genes or open reading frames (ORFs) have been only predicted, e.g. with bioinformatics methods, and such cases are indicated in the MODOMICS database. The catalogue of enzymes can be browsed by organism and type of reaction (deamination, pseudouridine formation, etc.). Database entries concerning individual enzymes contain name(s) and synonyms and information about the catalyzed reaction. Each enzyme is also linked to literature articles describing its characterization. If sequences are available, the name of the corresponding ORF and accession numbers for the Swiss-Prot (8) and NCBI (9) databases are provided. Enzymes with experimentally determined structures are linked to appropriate entries in the Protein Data Bank (10). We also provide homology models for some of the modification enzymes for which the 3D structures have not been solved experimentally. Most of the models have been taken from the literature, others have been constructed based on a hybrid protocol comprising fold recognition, comparative modeling and *de novo* modeling (11). Models can be viewed with a Jmol applet or downloaded in the PDB format from the MODOMICS website.

RNA SEQUENCES

MODOMICS provides RNA sequences with modifications. For large families of homologous RNAs multiple sequence alignments are available. The original release of MODOMICS contained only a small set of tRNA sequences curated by hand. Recently, we have expanded our database to include further tRNA and tDNA sequences from the Bayreuth database (12), with modifications curated manually, and rRNA sequences from the Comparative RNA Website (13). For rRNA sequences, we used data from the The Small Subunit rRNA Modification Database (14) and from published analyses concerning modifications in both ribosomal subunits. In both cases, only these sequences were included, where the presence of modification had been confirmed experimentally. Altogether, MODOMICS contains over 200 cytoplasmatic and mitochondrial tRNA sequences, 32 small ribosomal subunit sequences and 27 large ribosomal subunit sequences. Sequences are visualized with all modifications highlighted linked to the according modification record. The secondary structure based on RFAM is indicated for each alignment. The alignments can be downloaded in plain text format.

Nomenclature

The names of modified bases have been a matter of controversy. The IUPAC-conform names have been used with many variations; they however lead to problems when used in plain text sequences, PubMed abstract and other machine-centric data formats, as the formatting such as superscripts and subscripts is easily lost or confused. A system of one-letter abbreviations has turned out too limited when the number of modifications exceeded the number of available ASCII characters. To address this problem, a new simple nomenclature system for modifications has been proposed in the course of a round-table discussion at the conference in Aussois (Rich Roberts, Stephen Douthwaite, Adrian Ferre D'Amare, Jef Rozenski, Saulius Klimasauskas, Xiaodong Cheng, Tim Bestor, K.R., J.M.B. and H.G., details to be published elsewhere). In the new system, a numerical prefix describing a particular modification is added to a letter describing the original unmodified nucleoside. Thereby, a unique identifier is assigned for each entity (Figure 1f). In the case of RNA modifications this numbering scheme that is accompanied by a list of synonymous identifiers. Analogous to modifications, a unified nomenclature system for modification enzymes has been proposed, analogous to the one previously coined for DNA restriction-modification enzymes (15). Each enzyme name consists of three parts divided by dots (optionally a suffix can be added). The first part defines the type of enzyme and the target nucleic acid. The second part defines the source organism. The third contains an abbreviation for the enzyme. The optional suffix 'P' is included to discriminate 'putative' or 'predicted' enzymes from genuine enzymes with experimentally determined functions.

Future prospects

The total number of confirmed modifications and RNA modifying enzymes in both cases exceeds over 120 and 140, respectively. There is an overwhelming amount of experimental information available. Still, there are many modified positions in well-characterized RNA molecules, for which the according enzymes are not known. Moreover, new modified nucleosides are still being discovered, even for such well-studied molecules like rRNA. Thus, characterization of RNA modification pathways appears to be a moving target. In the future, MODOMICS will become a part of a RNA systems biology network. The most recent developments presented in this article have been coordinated with the RNA Ontology Consortium (16). MODOMICS will also develop towards more intense utilization of general pathway resources such

as KEGG (17) or enzyme databases such as BRENDA (18), in order to allow both a generic view on the systems biology of RNA, and detailed information on its components.

AVAILABILITY

The data are accessible freely for research purposes at the http://genesilico.pl/modomics. Most of the data is available for download in plain text formats. Modified nucleosides are also available as structure files and images. Images of pathways are available from the web page. Program code for parsing the plain text formats is available on request.

ACKNOWLEDGEMENTS

We would like to thank Rich Roberts, Stephen Douthwaite, Adrian Ferre D'Amare, Jef Rozenski, Saulius Klimasauskas, Xiaodong Cheng and Tim Bestor for their contribution and intense discussion of the modification and enzyme nomenclature system.

FUNDING

Polish Ministry of Science and Higher Education (PhD grant N301 010 31/0219 to A.C.); Marie Curie 6th EU-6FP Research Training Network 'DNA Enzymes' (MRTNCT-2005-019566 to K.R.); EU-6FP Network of Excellence 'EURASNET' (LSHG-CT-2005-518238 to J.M.B.). Funding to pay the Open Access publication charges for this paper has been waived by Oxford University Press. NAR Editorial Board members are entitled to one free paper per year in recognition of their work on behalf of the journal.

Conflict of interest statement. None declared.

REFERENCES

1. Yu,B., Yang,Z., Li,J., Minakhina,S., Yang,M., Padgett,R.W., Steward,R. and Chen,X. (2005) Methylation as a crucial step in plant microRNA biogenesis. *Science*, **307**, 932–935.
2. Guymon,R., Pomerantz,S.C., Ison,J.N., Crain,P.F. and McCloskey,J.A. (2007) Post-transcriptional modifications in the small subunit ribosomal RNA from Thermotoga maritima, including presence of a novel modified cytidine. *RNA*, **13**, 396–403.
3. Nasvall,S.J., Chen,P. and Bjork,G.R. (2007) The wobble hypothesis revisited: uridine-5-oxyacetic acid is critical for reading of G-ending codons. *RNA*, **13**, 2151–2164.
4. Johansson,M.J., Esberg,A., Huang,B., Bjork,G.R. and Bystrom,A.S. (2008) Eukaryotic wobble uridine modifications promote a functionally redundant decoding system. *Mol. Cell Biol.*, **28**, 3301–3312.
5. Sergiev,P.V., Bogdanov,A.A. and Dontsova,O.A. (2007) Ribosomal RNA guanine-(N2)-methyltransferases and their targets. *Nucleic Acids Res.*, **35**, 2295–2301.
6. Purta,E., Kaminska,K.H., Kasprzak,J.M., Bujnicki,J.M. and Douthwaite,S. (2008) YbeA is the m^3Psi methyltransferase RlmH that targets nucleotide 1915 in 23S rRNA. *RNA*, **14**, 2234–44.
7. Purta,E., O'Connor,M., Bujnicki,J.M. and Douthwaite,S. (2008) YccW is the m^5C methyltransferase specific for 23S rRNA nucleotide 1962. *J. Mol. Biol.*, Aug 29, [Epub ahead of print] doi:10.1016/j.jmb.2008.08.061.
8. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bradley,P., Bork,P., Bucher,P., Cerutti,L. *et al.* (2005) InterPro, progress and status in 2005. *Nucleic Acids Res.*, **33(Database Issue)**, D201–D205.
9. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S., Helmberg,W. *et al.* (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **33**, D39–D45.
10. Deshpande,N., Addess,K.J., Bluhm,W.F., Merino-Ott,J.C., Townsend-Merino,W., Zhang,Q., Knezevich,C., Xie,L., Chen,L., Feng,Z. *et al.* (2005) The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res.*, **33(Database Issue)**, D233–D237.
11. Kolinski,A. and Bujnicki,J.M. (2005) Generalized protein structure prediction based on combination of fold-recognition with de novo folding and evaluation of models. *Proteins*, **61(Suppl 7)**, 84–90.
12. Sprinzl,M. and Vassilenko,K.S. (2005) Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.*, **33**, D139–D140.
13. Cannone,J.J., Subramanian,S., Schnare,M.N., Collett,J.R., D'Souza,L.M., Du,Y., Feng,B., Lin,N., Madabusi,L.V., Muller,K.M. *et al.* (2002) The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics*, **3**, 2.
14. McCloskey,J.A. and Rozenski,J. (2005) The Small Subunit rRNA Modification Database. *Nucleic Acids Res.*, **33**, D135–D138.
15. Roberts,R.J., Belfort,M., Bestor,T., Bhagwat,A.S., Bickle,T.A., Bitinaite,J., Blumenthal,R.M., Degtyarev,S., Dryden,D.T., Dybvig,K. *et al.* (2003) A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Res.*, **31**, 1805–1812.
16. Leontis,N.B., Altman,R.B., Berman,H.M., Brenner,S.E., Brown,J.W., Engelke,D.R., Harvey,S.C., Holbrook,S.R., Jossinet,F., Lewis,S.E. *et al.* (2006) The RNA Ontology Consortium: an open invitation to the RNA community. *RNA*, **12**, 533–541.
17. Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
18. Barthelmes,J., Ebeling,C., Chang,A., Schomburg,I. and Schomburg,D. (2007) BRENDA, AMENDA and FRENDA: the enzyme information system in 2007. *Nucleic Acids Res.*, **35**, D511–D514.