The RNA WikiProject: Community annotation of RNA families

JENNIFER DAUB,[1] PAUL P. GARDNER,[1] JOHN TATE,[1] DANIEL RAMSKÖLD,[2] MAGNUS MANSKE,[1] WILLIAM G. SCOTT,[3] ZASHA WEINBERG,[4] SAM GRIFFITHS-JONES,[5] and ALEX BATEMAN[1]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, CB10 1SA, United Kingdom
[2]Royal Institute of Technology, SE-105 71 Stockholm, Sweden
[3]Department of Chemistry and Biochemistry, University of California at Santa Cruz, Santa Cruz, California 95064, USA
[4]Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, Connecticut 06520-8103, USA
[5]Faculty of Life Sciences, Michael Smith Building, Oxford Road, Manchester, M13 9PT. United Kingdom

ABSTRACT

The online encyclopedia Wikipedia has become one of the most important online references in the world and has a substantial and growing scientific content. A search of Google with many RNA-related keywords identifies a Wikipedia article as the top hit. We believe that the RNA community has an important and timely opportunity to maximize the content and quality of RNA information in Wikipedia. To this end, we have formed the RNA WikiProject (http://en.wikipedia.org/wiki/Wikipedia:WikiProject_RNA) as part of the larger Molecular and Cellular Biology WikiProject. We have created over 600 new Wikipedia articles describing families of noncoding RNAs based on the Rfam database, and invite the community to update, edit, and correct these articles. The Rfam database now redistributes this Wikipedia content as the primary textual annotation of its RNA families. Users can, therefore, for the first time, directly edit the content of one of the major RNA databases. We believe that this Wikipedia/Rfam link acts as a functioning model for incorporating community annotation into molecular biology databases.

Keywords: RNA annotation; RNA databases; Wikipedia

INTRODUCTION

"Wikis" are the current online collaborative writing tool of choice, allowing multiple users to edit concurrently. Wikipedia is the most successful and largest online wiki, with a robust and simple interface. It has a huge audience that doubles as a community of contributing authors. At the time of writing, the Wikipedia site is the seventh most visited site on the internet and attracts ~10% of all internet users (according to www.alexa.com). Many Google searches return a Wikipedia article amongst the top hits. For example, using the keyword RNA in a Google search currently returns the Wikipedia RNA article as the top hit, followed by the website of this esteemed journal. We believe that the RNA community has an important and timely opportunity to ensure that the RNA articles in Wikipedia are as useful as possible and ensure their scientific accuracy.

Our aim in this project was to provide articles about families of noncoding RNAs to Wikipedia. To start this process we used the RNA family annotations from the Rfam database. The resulting Wikipedia articles have now superceded the Rfam articles and are imported into Rfam on a daily basis. Users of Rfam can now directly edit Wikipedia articles, and hence update the Rfam entries. We are now interested in broadening the scope of the RNA-related articles in Wikipedia and call upon readers to become involved in this effort. We believe that this project greatly improves the quality of RNA information in both Wikipedia and Rfam, and demonstrates a viable model for community annotation of molecular biology databases.

WIKIPEDIA AND RFAM

Part of the success of the Wikipedia project is its ability to create communities of authors who focus on creating high-quality articles. There is a growing body of evidence that Wikipedia, despite inexpert editors and the treacheries of vandalism, is of comparable quality to traditional encyclopedias (Giles 2005). The WikiProject Molecular and Cellular Biology (MCB project) is a Wikipedia-user community focused on organizing and improving articles relating to molecular and cellular biology. A small number of Wikipedia entries relevant to RNA structure and

Reprint requests to: Paul P. Gardner, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, CB10 1SA, UK; e-mail: pg5@sanger.ac.uk; fax: 44-7788-120708.

Article published online ahead of print. Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.1200508.

function existed prior to this effort. These entries were representative of a few well-known RNA families (e.g., transfer RNA) or were generalized pages for different types of RNA (e.g., riboswitches, *cis*-regulatory elements, small nucleolar RNA). We have therefore created a daughter project of MCB called WikiProject RNA, with the aim of increasing Wikipedia's RNA-related content and improving existing articles.

The Rfam database (Griffiths-Jones et al. 2003, 2005) specializes in generating hand-curated multiple sequence alignments and secondary structure annotation for families of small structured nonprotein-coding RNAs. Rfam currently contains 607 families, including well-studied examples such as transfer RNA and ribosomal RNA, as well as microRNAs, small nucleolar RNA classes, and RNA structures found in untranslated regions of protein-coding messages.

In Rfam, each family has a short text describing the biological function and salient features of the RNA. The text is typically written by an Rfam curator, who attempts to review and condense the relevant literature into a description, with references to relevant sources for evidence of function, alignments, and structure data. These family annotations provide important contextual information for researchers and links to other resources. However, the primary focus of Rfam is on alignments, secondary structure, and genome annotation; the textual descriptions have not been routinely updated. As a result, these annotations are rarely comprehensive and can quickly fall behind the current state of the research field. In order to better utilize the wealth of knowledge in the research community, Rfam has decided to draw annotation from Wikipedia.

CREATING NEW WIKIPEDIA ENTRIES FROM RFAM

In collaboration with members of the MCB project, we created over 600 new Wikipedia starter articles from existing Rfam family information. This dramatically increased the coverage of known RNAs in Wikipedia. To do this, each Rfam article was exported into a new Wikipedia "stub" page. Stub articles are recognized in the Wikipedia community as starter articles in need of expansion. Each of these articles was then "wikified" to comply with Wikipedia article and policy conventions. Most importantly, this includes removal of Rfam-specific information and linking key words or terms to other Wikipedia articles. We used existing Wikipedia protocols to remove redundant pages by merging and deleting stubs where needed. Categories were added to make the articles easier to find from within Wikipedia. Finally, each Wikipedia entry was linked back to the Rfam database, allowing users to access information that would not be included in Wikipedia. In their current state, the majority of the contributed stub articles comprise "wikified" family text, literature references, secondary structure images, and a link to the Rfam family page.

Rfam now downloads all relevant, updated Wikipedia articles on a daily basis, using the MediaWiki API (http://www.mediawiki.org/wiki/API) and integrates them into the appropriate Rfam family page. The Wikipedia article replaces the old static text and is displayed alongside the Rfam-specific information (alignments, covariance models, phylogenetic trees, and methods). A link is provided from each family page that directs users back to Wikipedia, so that they may edit the article. All contributions to these pages are monitored. It is important to note that the areas in which Rfam specializes (multiple sequence alignments and covariance models) are not open to Wikipedia annotation and remain curated by Rfam's specialist annotators.

CURRENT STATUS OF THE PROJECT

From initiation of the project in June 2007 until August 2008, ~213 users have contributed to the RNA Wikipedia pages. The majority of these annotations are user contributions correcting formatting, errors, and links. Only 13 of these editors are automated scripts ("bots"). Approximately one-third of the manual users have provided scientific contributions by expanding content, adding links, references, or new images. Any acts of obvious vandalism have been reverted by bots or the Wikipedia community, usually within minutes of the vandalism taking place.

As far as we are aware, the introduction of community annotation to Rfam via Wikipedia is the first successful project of this nature. All of the family annotations are now greatly improved, as the previously static text has been replaced with an open access article that is linked and crossed referenced to other Wikipedia pages and information resources. Furthermore, we have provided a mechanism by which each family annotation may be rapidly improved as the research field progresses. A good example of the potential of this project is the significant improvement to the article on Hammerhead ribozymes, which is now encyclopaedic in scope (http://en.wikipedia.org/wiki/Hammerhead_ribozyme).

We expect that the Rfam family classification and the RNA WikiProject pages will diverge as existing pages are merged or made redundant, and new pages appear. These structural differences can easily be dealt with by allowing one-to-many and many-to-one links between Rfam families and Wikipedia pages.

THE RNA WIKIPROJECT NEEDS YOU!

We welcome contributions from the RNA community to RNA WikiProject and all RNA entries in Wikipedia. Many of the articles are still lacking up-to-date scientific content, and could be improved with biological function, figures, references, and translations into non-English languages. Articles on RNA families can be found via Wikipedia itself (http://www.wikipedia.org/), Rfam (http://rfam.sanger.ac.uk [UK], http://rfam.janelia.org/ [US]), or through the

RNA WikiProject (http://en.wikipedia.org/wiki/Wikipedia:WikiProject_RNA).

NOTE ADDED TO PROOF

Since submission of this article, three wiki-based biological database projects have been published: WikiProteins (Mons et al. 2008), GeneWiki (Huss et al. 2008), and Proteopedia (Hodis et al. 2008).

ACKNOWLEDGMENTS

We gratefully acknowledge the contributions to the project by users WillowW, Tim Vickers, and DO11.10. J.D., P.P.G., J.T., M.M., and A.B. are funded by the Wellcome Trust. S.G.-J. is funded by the University of Manchester. Z.W. is funded by Howard Hughes Medical Institute support to Ronald R. Breaker.

Received June 2, 2008; accepted August 29, 2008.

REFERENCES

Giles, J. 2005. Internet encyclopedias go head to head. *Nature* **438:** 900–901.

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., and Eddy, S.R. 2003. Rfam: An RNA family database. *Nucleic Acids Res.* **31:** 439–441.

Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S.R., and Bateman, A. 2005. Rfam: Annotating noncoding RNAs in complete genomes. *Nucleic Acids Res.* **33:** D121–D124.

Hodis, E., Prilusky, J., Martz, E., Silman, I., Moult, J., and Sussman, J.L. 2008. Proteopedia—A scientific "wiki" bridging the rift between 3D structure and function of biomacromolecules. *Genome Biol.* **9:** R121. doi: 10.1186/gb-2008-9-8-r121.

Huss 3rd., J.W., Orozco, C., Goodale, J., Wu, C., Batalov, S., Vickers, T.J., Valafar, F., and Su, A.I. 2008. A gene wiki for community annotation of gene function. *PLoS Biol.* **6:** e175. doi: 10.1371/journal.pbio.0060175.

Mons, B., Ashburner, M., Chichester, C., van Mulligan, E., Weeber, M., den Dunnen, J., van Ommen, G.J., Musen, M., Cockerill, M., Hermyakob, H., et al. 2008. Calling on a million minds for community annotation in WikiProteins. *Genome Biol.* **9:** R89. doi: 10.1186/gb-2008-9-5-r89.