A *Phaeodactylum* literature database for interactive annotation of content

Alessandra A. Gallina^{a,b}, Mark Layer^a, Zachary A. King^b, Jennifer Levering^b, Bernhard Ø
Palsson^b, Karsten Zengler^{c*}, Graham Peers^{a*}

^aBiology Department, Colorado State University, 80523, Fort Collins, CO.

^bDepartment of Bioengineering, University of California, San Diego, La Jolla, CA 92093

^cDepartment of Pediatrics, University of California San Diego, San Diego, 92093, CA

*Corresponding authors at:

G.P.: Biology Department, Colorado State University, 80523, Fort Collins, CO.

E-mail address: graham.peers@colostate.edu

K.Z.: Department of Pediatrics, University of California San Diego, San Diego, 92093, CA

E-mail address: kzengler@ucsd.edu

ABSTRACT

Phaeodactylumtricornutum is a widely used model organism for studying diatom biology. We created a *P. tricornutum*-specific literature database, or bibliome, that can be interactively annotated by users to improve the relevance and accuracy of search queries. A bibliome represents the current knowledge about a specific organism or biological process and is the foundation for genome-scale computation models. The bibliome of *P. tricornutum* was assembled from literature searches of '*Phaeodactylumtricornutum*' (as topic) in two different existing databases, the Web of Knowledge and PubMed resulting in 2,497 independent articles. We manually curated the bibliome for relevance to biochemical, physiological, and molecular biology studies. Literature that described the use of *P. tricornutum* as feed source and ecotoxicology studies were excluded from the bibliome. This process resulted in a final list of 1,124 entries – 20 % of which received additional classification in the form of notes or keywords. We present the bibliome in an HTML format that is easily searched and that can be customized according to the interests of individual researchers. This HTMLtool can also easily be adapted for other algal species or specific research topics.

Keywords

Phaeodactylumtricornutum; database; literature;bibliome

1. Introduction

Marine diatoms are among the most successful eukaryotic photosynthetic organisms on Earth, and they are responsible for about 20% of primary productivity [1]. Besides their well-known ecological importance, in recent years there has been an increasing attention towards the use of diatoms as a source of biofuels and value-added products [2-4]. Due to their unique evolutionary history, diatoms' genomes and metabolism combine features derived from highly divergent organisms, including bacteria as well as red and green algae [5-7]. *Phaeodactylum tricornutum* represents a model pennate diatom which has been shown to possess unique metabolic traits, including a metazoan-like ornithine-urea cycle [8] and an Entner–Doudoroff glycolytic pathway [9]. Despite its wide use as a model organism, there is currently a lack of a *P. tricornutum*-dedicated knowledgebase that could serve as resource for the broader scientific community. In particular, there is no comprehensive literature database, or bibliome, focusing specifically on *P. tricornutum* and thus enabling fast access to relevant information.

A bibliome is a curated and annotated collection of knowledge about a specific organism or biological process and it is the foundation for the construction of genome-scale models [10]. The curator's selection of literature and the addition of user-generated keywords or notes facilitate the process of information retrieval. This adds value over a standard literature database which often returns large amounts of non-relevant literature [11]. This approach stresses the importance of human-based knowledge for interpreting the literature and it also can contribute to improved efficiency of data mining algorithms [12, 13].

Four examples of curated bibliomes include 1) the Comparative Toxicogenomics Database [14, 15], which is a literature-based database about how exposure to chemicals affects human health; 2) the upstream open reading frame (uORF) literature database [16], which is a literature database

on eukaryotic uORF biology; 3) a curated database of the collected articles that use constraint-based models for biological predictions [17]; and 4) the Mouse Genome Informatics (MGI) biocuration workflow [18], where manual curation of scientific articles is coupled with automated text mining to identify pertinent articles about mouse genetics and genome biology.

Creating a bibliome is essential for the development of a genome-scale metabolic reconstruction. It streamlines the manual annotation and verification of metabolic pathways by providing a literature resource that is focused on biochemistry and physiology [19, 20]. We created a bibliome of *P. tricornutum*-specific literature to facilitate a model that describes its metabolism [21]. The result is a comprehensive literature database enriched in biochemical, physiological, and molecular biology studies on *P. tricornutum* literature. The database is available with the Supplemental Data of this paper, and it can be interactively annotated by individual users. *P. tricornutum* is widely used as the model pennate diatom since the publication of its complete genome in 2008 [5] and this tool will be valuable for the scientific community working on diatoms. This tool and approach could easily be applied for the creation of other topic-specific literature databases, especially in the area of algal research.

2. Materials and Methods

2.1. Scientific literature manual curation

The three most commonly used scientific literature databases, each with its own advantages and disadvantages, are PubMed, Web of Science (WoS), and Google Scholar [22-24]. We decided not to include a literature search in Google Scholar given its previously reported inaccuracy and poor coverage of scientific works [25, 26]. The workflow for bibliome construction is shown in

Figure 1. In November 2013 the search term '*Phaeodactylum tricornutum*' was queried as 'topic' in both the Web of Knowledge (WoK) database (now WoS) and the PubMed database. The search retrieved 2,471 entries in the WoK, and 418 entries in PubMed. The entries were downloaded and merged, resulting in 2,497 unique articles. We continued to update the bibliome for articles published up until December 31st 2015.

We then culled articles from this list that did not contain biochemistry, physiology, or genomics related information. Entries in which *P. tricornutum* was used as feed source, ecotoxicology studies, and studies that did not contain *P. tricornutum*-specific information were excluded from the database. We made exceptions if the work contained 'omics' (i.e. genomic, metabolomics and transcriptomics) or physiological data. Additionally, we decided to include key papers that address scientific topics relevant to diatom research and articles considered to be relevant for production of biofuels and value-added compounds.

About 20% of the entries received additional classification either in the form of notes or user-assigned keywords. The choice of additional keywords was based on relevance to biochemical, physiological, and molecular biology studies for application to the genome-scale model of metabolism. Where applicable, authors' keywords and abstracts were retrieved and integrated into the database.

2.2. Web-based interface development

We built the HTML-based Bibliome Viewer to release our *P. tricornutum* bibliome to the public. The Bibliome Viewer is an HTML file which contains the data of the bibliome and includes HTML and JavaScript code to enable viewing, editing, and saving of the bibliome. The entries for the bibliome for *P. tricornutum* were extracted from the spreadsheet that had been

used for manual curation. The Bibliome Viewer requires the web browser Google Chrome to run, and on most systems the Bibliome Viewer can be started by dragging its HTML file into an open Chrome window. Despite this, the Bibliome Viewer is entirely local to the machine the file exists on and does not require an active Internet connection. We note that the Bibliome Viewer is not meant to replace reference management tools such as Endnote or Mendeley.

3. Results and Discussion

A search of '*Phaeodactylumtricornutum*' in the WoS returned 2,471 hits on November 2013, where the same search performed in PubMed only returned 418 entries. This low number of entries is very likely dependent on the fact that the PubMed database specializes in the field of medicine, since it was derived from 'Medline', a database originally launched in 1971 by the National Library of Medicine [22]. Subsequent searches in the WoS failed in returning the most recent published articles and pre-print articles and this failure mode has been previously discussed being a drawback of the WoS [22]. These are relevant points to take into consideration when performing literature searches especially in the field of algal or plant research, since the use of different databases can influence the results and the information retrieved may not be complete.

We manually curated for *P. tricornutum* content by initially evaluating the title and by reading the abstract of each article. If it was still unclear whether the publication contained relevant information or not, the full text was accessed. The articles that were excluded from the original 2,471 hits included papers that did not have any information on *P. tricornutum*, studies that used *P. tricornutum* as a feed source and ecotoxicological studies (unless they had omics or physiological data). Interestingly, some papers retrieved from WoS did not have any information

on *P. tricornutum* whatsoever. This result may be due to the practice of Thomson Reuters editors assigning “keyword plus” identifiers to WoS entries based off of title alone. Manual curation is required to remove these erroneous hits.

Keywords are the metadata that describe the data and findings of a specific document[11]. It is particularly important that the choice of keywords that are added to a literature database is based on the relevance of the articles’ content. In our *P. tricornutum* literature database we originally intended to identify studies with biochemical, physiological, and molecular biology data for application to genome-scale model reconstruction. Therefore we added additional keywords such as “biomass”, “transcriptomic”, “uptake”, and “lipid” to about 20% of the articles in the database.

The Bibliome Viewer contains numerous features to facilitate user annotation and article searching and it is accessed in the Supplemental Data associated with this manuscript. It already contains articles from 1897 to 2015. Users have the option to add blank entries and the tool accepts new references as comma-separated (CSV) and tab-delimited files imported from PubMed and WoS, respectively, as well as BibTex citations (see Supplemental Data for a user guide). All user-made changes to the bibliome are saved locally as a new timestamped copy of the Bibliome Viewer. Other features include the option to delete entries, sort by field, and add keywords and notes to each entry. Additionally, the Bibliome Viewer allows users to export their bibliome, or selected entries only, as a CSV, BibTex or RIS file. A search function also allows to refine searches by year and keywords of interest. Entries also have a link to a digital object identifier (DOI) if available. The small file size (2 MB HTML file), no software installation requirement and simplicity of the code make the Bibliome Viewer ideal for distributing a reference collection to a wide range of users on various systems with differing levels of technical

ability.

A unique feature of the Bibliome Viewer is that it has a completely customizable interface available through the html and javascript embedded directly on the page. Users can add new functions, such as data integration, to their bibliome. This could include a wide range of data from physiological parameters that facilitate data mining to NCBI gene accessions links. This customizable aspect of the Bibliome Viewer is not currently available in the most commonly used reference management tools.

We envision this tool being used by the *P. tricornutum* scientific community as an interactive resource to build targeted bibliomes. This tool makes scientific research on *P. tricornutum*'s biochemistry and physiology more accessible to new researchers and more experienced research groups. The Bibliome Viewer can easily be adopted into individual research groups for custom annotation and future additions. The tool and approach can easily be implemented by researchers with interests in other organisms and biological processes.

Acknowledgements

This research was supported by the Department of Energy, Office of Science, Biological and Environmental Sciences, with grants DOE-DE-SC0008593 to B.P., and DOE-DE-SC0008595 to G.P.

References

- [1] E.V. Armbrust, The life of diatoms in the world's oceans, *Nature*, 459 (2009) 185-192.
- [2] M. Hildebrand, A.K. Davis, S.R. Smith, J.C. Traller, R. Abbriano, The place of diatoms in the biofuels industry, *Biofuels*, 3 (2012) 221-240.
- [3] O. Levitan, J. Dinamarca, G. Hochman, P.G. Falkowski, Diatoms: a fossil fuel of the future, *Trends Biotechnol.*, 32 (2014) 117-124.
- [4] W.Q. Fu, K. Wichuk, S. Brynjolfsson, Developing diatoms for value-added products: challenges and opportunities, *Nat. Biotechnol.*, 32 (2015) 547-551.
- [5] C. Bowler, A.E. Allen, J.H. Badger, J. Grimwood, K. Jabbari, A. Kuo, U. Maheswari, C. Martens, F. Maumus, R.P. Otiilar, E. Rayko, A. Salamov, K. Vandepoele, B. Beszteri, A. Gruber, M. Heijde, M. Katinka, T. Mock, K. Valentin, F. Verret, J.A. Berges, C. Brownlee, J.P. Cadoret, A. Chiovitti, C.J. Choi, S. Coesel, A. De Martino, J.C. Detter, C. Durkin, A. Falciatore, J. Fournet, M. Haruta, M.J.J. Huysman, B.D. Jenkins, K. Jiroutova, R.E. Jorgensen, Y. Joubert, A. Kaplan, N. Kroger, P.G. Kroth, J. La Roche, E. Lindquist, M. Lommer, V. Martin-Jezequel, P.J. Lopez, S. Lucas, M. Mangogna, K. McGinnis, L.K. Medlin, A. Montsant, M.P. Oudot-Le Secq, C. Napoli, M. Obornik, M.S. Parker, J.L. Petit, B.M. Porcel, N. Poulsen, M. Robison, L. Rychlewski, T.A. Ryneerson, J. Schmutz, H. Shapiro, M. Siaut, M. Stanley, M.R. Sussman, A.R. Taylor, A. Vardi, P. von Dassow, W. Vyverman, A. Willis, L.S. Wyrwicz, D.S. Rokhsar, J. Weissenbach, E.V. Armbrust, B.R. Green, Y. Van De Peer, I.V. Grigoriev, The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes, *Nature*, 456 (2008) 239-244.
- [6] P. Deschamps, D. Moreira, Reevaluating the Green Contribution to Diatom Genomes, *Genome Biol. Evol.*, 4 (2012) 795-800.

192 [7] G. Finazzi, H. Moreau, C. Bowler, Genomic insights into photosynthesis in eukaryotic
 193 phytoplankton, *Trends Plant Sci.*, 15 (2010) 565-572.

194 [8] A.E. Allen, C.L. Dupont, M. Obornik, A. Horak, A. Nunes-Nesi, J.P. McCrow, H. Zheng,
 195 D.A. Johnson, H.H. Hu, A.R. Fernie, C. Bowler, Evolution and metabolic significance of the
 196 urea cycle in photosynthetic diatoms, *Nature*, 473 (2011) 203-207.

197 [9] M. Fabris, M. Matthijs, S. Rombauts, W. Vyverman, A. Goossens, G.J.E. Baart, The
 198 metabolic blueprint of *Phaeodactylum tricornutum* reveals a eukaryotic Entner-Doudoroff
 199 glycolytic pathway, *Plant J.*, 70 (2012) 1004-1014.

200 [10] N.C. Duarte, S.A. Becker, N. Jamshidi, I. Thiele, M.L. Mo, T.D. Vo, R. Srivas, B.O.
 201 Palsson, Global reconstruction of the human metabolic network based on genomic and bibliomic
 202 data, *Proc. Natl. Acad. Sci. USA*, 104 (2007) 1777-1782.

203 [11] L. Grivell, Seek and you shall find?, *EMBO Rep.*, 7 (2006) 10-13.

204 [12] L. Grivell, Mining the bibliome: searching for a needle in a haystack? New computing tools
 205 are needed to effectively scan the growing amount of scientific literature for useful information,
 206 *EMBO Rep.*, 3 (2002) 200-203.

207 [13] J.F. Allen, In silico veritas - Data-mining and automated discovery: the truth is in there,
 208 *EMBO Rep.*, 2 (2001) 542-544.

209 [14] T.C. Wiegers, A.P. Davis, K.B. Cohen, L. Hirschman, C.J. Mattingly, Text mining and
 210 manual curation of chemical-gene-disease networks for the Comparative Toxicogenomics
 211 Database (CTD), *BMC Bioinformatics*, 10 (2009) 12.

212 [15] A.P. Davis, T.C. Wiegers, C.G. Murphy, C.J. Mattingly, The curation paradigm and
 213 application tool used for manual curation of the scientific literature at the Comparative
 214 Toxicogenomics Database, *Database*, (2011) 12.

215 [16] K. Wethmar, A. Barbosa-Silva, M.A. Andrade-Navarro, A. Leutz, uORFdb-a
 216 comprehensive literature database on eukaryotic uORF biology, *Nucleic Acids Res.*, 42 (2014)
 217 D60-D67.

218 [17] A. Bordbar, J.M. Monk, Z.A. King, B.O. Palsson, Constraint-based models predict
 219 metabolic and associated cellular functions, *Nat. Rev. Genet.*, 15 (2014) 107-120.

220 [18] K.G. Dowell, M.S. McAndrews-Hill, D.P. Hill, H.J. Drabkin, J.A. Blake, Integrating text
 221 mining into the MGI biocuration workflow, *Database*, (2009) 11.

222 [19] A.M. Feist, M.J. Herrgard, I. Thiele, J.L. Reed, B.O. Palsson, Reconstruction of
 223 biochemical networks in microorganisms, *Nat. Rev. Microbiol.*, 7 (2009) 129-143.

224 [20] I. Thiele, B.O. Palsson, A protocol for generating a high-quality genome-scale metabolic
 225 reconstruction, *Nat. Protoc.*, 5 (2010) 93-121.

226 [21] J.Levering, J. Broddrick, C.L. Dupont, G. Peers, K. Beerli, J. Mayers, A.A. Gallina, A.E.
 227 Allen, B.O. Palsson, K. Zengler, Genome-Scale Model Reveals Metabolic Basis of Biomass
 228 Partitioning in a Model Diatom. *PLoS ONE* 11(5) (2016): e0155038,
 229 doi:10.1371/journal.pone.0155038

230 [22] M.E. Falagas, E.I. Pitsouni, G.A. Malietzis, G. Pappas, Comparison of PubMed, Scopus,
 231 Web of Science, and Google Scholar: strengths and weaknesses, *FASEB J.*, 22 (2008) 338-342.

232 [23] J.C.F. de Winter, A.A. Zadpoor, D. Dodou, The expansion of Google Scholar versus Web of
 233 Science: a longitudinal study, *Scientometrics*, 98 (2014) 1547-1565.

234 [24] C. Hightower, C. Caldwell, Shifting Sands: Science Researchers on Google Scholar, Web of
 235 Science, and PubMed, with Implications for Library Collection Budgets, *Issues in Science and*
 236 *Technology Librarianship*, (2010).

237 [25] P. Jacso, Google Scholar: the pros and the cons, *Online Inform. Rev.*, 29 (2005) 208-214.

238 [26] P. Jacso, Metadata mega mess in Google Scholar, *Online Inform. Rev.*, 34 (2010) 175-191.

239

240

241

242 Figure 1. Workflow for the construction of the *P. tricornutum* bibliome.

243

244