Citation for published version:

DOI:
10.3897/rio.3.e13589

Publication date:
2017

Link to publication

Publisher Rights
CC BY

University of Bath

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
A machine-compiled microbial supertree from figure-mining thousands of papers

Ross Mounce†, Peter Murray-Rust‡, Matthew A Wills§

† University of Cambridge, Cambridge, United Kingdom
‡ Milner Centre for Evolution, University of Bath, Bath, United Kingdom

Abstract

Background

There is a huge diversity of microbial taxa, the majority of which have yet to be fully characterized or described. Plant, animal and fungal taxa are formally named and described in numerous vehicles. For prokaryotes, by contrast, all new validly described taxa appear in just one repository: the International Journal of Systematics and Evolutionary Microbiology (IJSEM). This is the official journal of record for bacterial names of the International Committee on Systematics of Prokaryotes (ICSP) of the International Union of Microbiological Societies (IUMS). It also covers the systematics of yeasts. This makes IJSEM an excellent candidate against which to test systems for the automated and semi-automated synthesis of published phylogenies.

New information

In this paper we apply computer vision techniques to automatically convert phylogenetic tree figure images from IJSEM back into re-usable, computable, phylogenetic data in the form of Newick strings and NEXML. Furthermore, we go on to use the extracted phylogenetic data to compute a formal phylogenetic MRP supertree synthesis, and we compare this to previous hypotheses of taxon relationships given by NCBI’s standard
taxonomy tree. This is the world's first attempt at automated supertree construction using data exclusively extracted by machines from published figure images. Additionally we reflect on how recent changes to UK copyright law have enabled this project to go ahead without requiring permission from copyright holders, and the related challenges and limitations of doing research on copyright-restricted material.

**Keywords**

Phylogenetics, Supertree, Microbes, Systematics, Computer Vision, Synthesis, Data Re-use, Data Extraction

**Introduction**

A recent study estimated that there are more than 114,000,000 documents in the published scientific literature (Khabsa and Giles 2014). It would be highly desirable to synthesize information scattered across these disparate sources: to join together all the little pieces in order to see the overall 'big picture'. Unfortunately, the manner in which scientific data are published often hinders such a synthesis:

- The data underlying articles are often not published in the first place (Wicherts et al. 2006, Stoltzfus et al. 2012, Drew et al. 2013, Magee et al. 2014, Caetano and Aisenberg 2014)
- Some journals allow data to be "embargoed" and not made available for up to 10 years after the publication of an associated article (Roche et al. 2014)
- When data are published, it can be in a manner that is not machine-readable. Data are frequently obfuscated in pixel-based figures, and even where data are tabulated the formatting is often esoteric. Metadata often lack unambiguous identifiers.
- Articles are frequently published in subscription journals to which potential re-users of information do not have access.
- Published data can disappear over time because there is no sustainable long term archiving (Vines et al. 2014)
- There are frequently copyright-imposed restrictions on the re-use and modification of published content (Hagedorn et al. 2011, Taylor 2016, Egloff et al. 2017)

In this paper, we present the results of our efforts to extract phylogenetic data from images contained in the primary research literature. We acknowledge the many previous efforts to extract phylogenetic information from figured trees, including TreeThief (Rambaut 2000), TreeSnatcher (Laubach and Haeseler 2007), TreeSnatcher 2 (Laubach et al. 2012), TreeRipper (Hughes 2011), and TreeRogue (Matzke 2012).
Scholarly literature as BigData

The international corpus of public scholarly literature now has many of the features and problems of ‘Big Data’. In particular, we highlight four issues:

- **Volume.** About 2.7 million scholarly objects receive CrossRef DOIs, of which about 2.2 million are ‘articles’ of some sort. Assuming, for illustrative purposes, that an article (with supplemental data and images) requires ca 10 MBytes, this equates to in the region of 30 Terabytes. Although a moderate volume of data by some standards (those repositories archiving the results of High Energy Particle [HEP] physics, or the Square Kilometer Array [SKA]) this is still more than most researchers or teams can analyse or utilise. The majority of this information is in Scientific/Technical/Medical disciplines (STM).

- **Velocity.** Assuming 50 working weeks of 40 hours each (2000 hours/year), this volume arrives at 15 GB/hour or 4 Mb / second. This equates to 2.5 million articles per year, or 1000 titles per hour (Ware and Mabe 2015). We are now at the stage where no individual can read even the titles of published scholarship, let alone the abstracts.

- **Variety.** Some big data (e.g. from instruments such as ATLAS [Doglioni 2012] or the SKA [Carilli and Rawlings 2004]) is highly structured with clear schemas. Much of the remainder, however, is unstructured or at best semi-structured. Most scientific publications consist of human readable text with non-semantic document structure and implicit semantics. Moreover this is highly variable: there are over 28,000 STM journals and each has their own approach to semi-structure (Ware and Mabe 2015).

- **Veracity/Validity/Verification.** Can we use machines to establish the degree of trust that we can put in data? More specifically, can we reliably read and interpret the information in the way that the community expects? We are concerned with the semantic definition of the information and its schema-validity (does it conform to an implicit/explicit specification?) and value-validity (are the values captured accurately).

We believe that machines are now essential to enable us to make sense of the stream of published science, and this paper addresses several of the key problems inherent in doing this. We have deliberately selected a subsection of the literature (limited to one journal) to reduce the volume, velocity and variety axes, concentrating primarily on validity. We ask whether high-throughput machine extraction of data from the semistructured scientific literature is possible and valuable.
Phylogenetic Trees

We chose to extract phylogenetic trees and combine them into a supertree: a process that is both tractable and useful. Phylogenetic trees are inferred from the distributions of putatively homologous characters (or traits) of organisms, resulting in one or more variously optimal trees as the output from an analysis. Computing a well-supported phylogenetic tree can entail many tens or hundreds of CPU hours. Because of this expense, trees are usually inferred for a small subset of species or 'leaves' (perhaps 10 to 500) rather than for all of those available. Each tree created and published is therefore a small but important contribution to our understanding of the classification and relationships of taxa.

Methods for synthesising larger trees ('supertrees') from a collection of smaller trees exist, but supertrees are rarely created because of the scarcity of published trees in semantic, reusable forms (Stoltzfus et al. 2012, Drew et al. 2013, Magee et al. 2014). Although formats and ontologies exist (e.g. Newick and NEXML (Vos et al. 2012)), and there are specialised communal databases for archiving these data (TreeBASE Morell 1996; MorphoBank O’Leary and Kaufman 2011) the voluntary take-up is very small: < 4% of all trees are captured (Stoltzfus et al. 2012). Although deposition by authors is always the best solution, authors are often reluctant or lacking in incentive to do this. Demonstrating the utility of accumulated data may help to redress this, and our ancilliary agenda here is to encourage the scientific community to invest in data aggregation.

Phylogenetic trees are usually represented diagrammatically, with a topological object isomorphic to a tree with nodes and edges. There are several major styles, but the commonest is a set of edges, normally orthogonal, meeting at (often undistinguished) nodes where three or more edges meet. The tree is normally 'rooted' (i.e. one node is nominated as the root), but this root is often implicit (at the midpoint of the 'top' edge). The tree is often directional, with the root usually (but not always) either on the left or at the bottom of the diagram, and the 'leaves' or 'tips' opposite the root. The tips (univalent nodes) are almost always labelled with text (e.g. the name of the species or higher taxa constituting the tree). The internal, multivalent nodes are often unlabelled, but may also have annotations such as confidence limits, support measures, pie charts, or the names of higher groups. The edges often have right-angle bends, but may be variously angled or even curved. These differences are usually solely cosmetic, and have no biological or other significance. The diagram is potentially semi-metric: distances along the root-tip direction (we shall use $h$ to denote this as $x$ and $y$ are frequently interchanged) may sometimes represent the similarity between daughter nodes. Distances in the other direction ($w$) are usually meaningless as is the order of tips: only the ancestry of nodes is relevant. Very occasionally, fonts and weights are used to convey information but these nuances have been ignored. The thickness of lines is normally irrelevant, as are the colours of lines, leaf names and other labels.

In some cases, two or more trees are presented in a single diagram. Sometimes they are multiple trees for different leaf sets, using the same $h$-direction. At other times they are multiple trees for the same leaf set and may be oriented tip-to-tip ($h$ and -$h$) to show
similarities and differences. Tips can be annotated with bars or checkerboards in the $w$-direction (e.g. to show clades, geographical distributions of leaves or other data). Further decorations include non-orthogonal arrows and schematic images of species.

In essence, therefore, a tree is described by a collection of nodes with ($h,w$ coordinates), associated labels, a root node, and inter-node edges. Omitting the meaningless $w$, we can represent this in Newick (Lisp-like) or NeXML (XML) formats without loss of essential information other than elements of visual style.

Legal Aspects

The re-use of data from the scientific literature is potentially a major ‘good’, and many policy makers are pushing for liberalisation of access to - and re-use of - published science. 'Text and Data Mining' (or as we prefer, 'Content Mining') is now actively promoted for reform, especially in Europe. Unfortunately electronic documents are formally covered by copyright, which acts as a major barrier to re-use for legally-aware scientists. We also suspect that many scientists knowingly and routinely infringe copyright:

“I hardly know any scientists who don’t violate copyright laws. We just fly below the radar and hope that the publishers don’t notice.” - Anonymous scientist quoted in Van Noorden (2014).

We (RM and PMR), have been involved in the proposed European copyright reform and note:

1. Copyright is absolute and complex. Almost all documents that we have mined are copyrighted, and it is usually unclear what rights the researcher has and what risks they undertake. It is also jurisdiction-dependent and a work may fall under more than one jurisdiction.
2. Two main approaches are: (a) to seek permission from every copyright holder, and if necessary pay a licence fee or (b) to contend that the law allows the present activity under exceptions and precedent. There is relatively little certainty in (b) and the researcher runs the risk of being accused of copyright infringement.
3. Images have often had a special position as they are often regarded by the author or copyright-holder as ‘creative works’ (but see Egloff et al. 2017 for an alternative view which we support).

In this research we have taken route 2(b) and used material to which we have legitimate access. Since we all work in the UK and are funded by UK institutions (including Research Councils) we refer to UK law (but are not formally making a legal case here). The seminal aspects are:

- Facts are uncopyrightable. We contend that much of the information we use (including the hierarchical structure of phylogenetic trees) is factual - a record of work performed by the authors, not capable of creative interpretation or re-expression.
• The UK Copyright reform (2014) allows for copying for mining (data analytics) and other non-commercial research purposes.
• The same reform allows for ‘fair quotation’, which we contend allows us to embed the extracted facts in enough context to make scientific sense.

We offer the output as facts, and assign them to the public domain by using the **CC0 waiver** of Creative Commons.

**Materials and Methods**

Prior work (Mounce 2013) determined that the *International Journal of Systematic and Evolutionary Microbiology* (IJSEM) has a greater number of phylogenetic tree diagrams published in it per annum than any other single journal. Moreover, the style of phylogenetic tree figures in IJSEM is much more consistent between articles than in other systematic and phylogenetic journals. Therefore, IJSEM makes an ideal starting point for automating phylogenetic data extraction from images, as it is both rich, voluminous, and style-consistent in target image data (phylogenetic tree figures). The workflow that we describe in this section is summarised in Fig. 1.

![Figure 1. Overall work flow; from content acquisition to stripping figure images out of the PDF, to image filtering, image analysis and reconversion back into re-usable, machine-readable phylogenetic data.](image)

**Content acquisition:**

Full text articles were systematically downloaded from the IJSEM website as PDF files using the open source command-line program *GNU Wget* version 1.15. An eleven-year span of articles was downloaded, from January 2003 (Volume 53, Issue 1) through to December 2013 (Volume 63, Issue 12) inclusive. From each publication year, Tables of contents (TOCs) and full text PDF links were extracted and subsequently downloaded with *Wget*. No distinction was made between research articles, editorial matter and erratums; all
were downloaded. A total of 5,816 source PDFs were obtained in this manner. PDF filenames were renamed by their unique partial DOI to aid provenance tracking (e.g. ijs.0.004572-0.pdf corresponds to the article which is available for ‘free’ via this URL: http://dx.doi.org/10.1099/ijs.0.004572-0). Electronic supplementary material was neither examined nor downloaded for the purpose of this analysis. At the time these downloads were undertaken, the IJSEM website was managed by Highwire Press. It has since been ported to a new Ingenta platform with a significantly different structure.

**Extraction and isolation of images from their PDF containers:**

The open source command-line program `pdfimages` version 0.25.0 (part of the Poppler library: http://poppler.freedesktop.org/) was used to automate the extraction and isolation of all figure images from each PDF. A total of 8,221 source images were extracted from the 5,816 source PDFs. Each source image was named according to the unique DOI of the PDF it came from, plus a three digit identifier to indicate which image it was in the PDF (e.g. ijs.0.004572-0-000.jpg, ijs.0.004572-0-001.jpg, ijs.0.004572-0-002.jpg, reflecting the sequence in which the images appeared throughout the PDF).

**Selection of phylogenetic tree images:**

All 8,221 images were loaded into Shotwell version 0.17.0, an open-source GUI image management program. One of us (RM) manually tagged all phylogeny-containing figure images, resulting in a selection of 4,336 images that contained a dendrogram of some form or another. This manual process took about 5 hours. This phylogeny selection set was exported (copied) using Shotwell to its own clean directory path, safely away from the non-phylogeny images for further processing. In retrospect, and with the application of some machine learning techniques, we could probably have automated this step too, with a high degree of precision and recall.

**Converting raster images to re-usable phylogenetic data:**

The set of 4,336 images containing a phylogeny (e.g., Fig. 2) were split into nine different batches of up to 500 images each, before further processing with ami-phylo. There was no single method that would universally and reliably extract semantic data from images.

Problems included:

- the quality of the target image
- the types of graphic object to be extracted,
- the natural language in the diagram,
- the error rate in creating the image
- orthogonal sources of information
- complexity
In this project we explored several possible approaches before alighting upon a scheme that worked well.

In general anyone wanting to use diagram mining in science should ask; is the image the original or a copy? Copies (such as photographs, photocopying, scanning) may introduce noise, distortion, contrast, antialiasing, bleeding, holes and line breaks. Fully computer generated images have the benefit of consistency, but sometimes introduce artefacts such as drawing lines twice for emphasis. In this paper we exclusively utilise machine generated diagrams.

We focused upon IJSEM because it is a key microbiological journal, and because papers within it follow the same systematic layout. Each new species to be described was placed within a phylogeny, and almost all papers therefore contained one or more trees. Tree figures in IJSEM are typically created with the same software, and the resultant diagrams are all oriented similarly, have the same font-set and the same semantics. These figures also typically include Genbank accession numbers alongside each taxon (or terminal leaf) so that their identity can be verified. Unlike at many other journals, IJSEM uses a minimum of extraneous graphics in phylogeny figure images, avoiding 'chartjunk' (sensu Tufte 2001) which adds no science but makes mining much harder.

The process of tree and tip data extraction required the following operations:

- Identify all characters
- Aggregate these into 'words' and 'phrases'
- Interpret phrases and check for correctness. The main tool was 'lookup' in 'Genbank' and other resources for taxonomy
- Identify all paths (lines and curves)
- Build these into a tree
- Identify errors

We experimented with several methods including:

- Writing our own OCR software
- Edge detection and fitting lines
- Identifying horizontal and vertical lines

Ultimately, we converged on:

- **Tesseract** (Smith 2007) for optical character recognition
- Our own software for phrase detection
- Zhang-Suen Thinning (Zhang and Suen 1984)
- Recreation into connected graphs
- Image Segmentation (Blanchet and Charbit 2013)

Because the images were strictly binary, there was no need for contrast manipulation, binarization or posterization. However we needed to determine the nodes and edges of the source tree. Fortunately, the only information required was the coordinates of the nodes (either tips with 1 edge or nodes where 3 or more edges met). The edge thickness, texturing, and 'kinks' of internal branches are purely stylistic. Thinning reduces all connectivity to a single pixel edge. We wrote a superthinning algorithm to require every pixel to be either in a node, 2-connected in an edge, or a tip. We included diagonal connectivity in all algorithms.

The phrases (e.g. species binomial names and Genbank accession numbers) were extracted as follows:

- The diagram was binarised, but not thinned.
- The binarised file was submitted to Tesseract. This extracted characters and assembled words from pixels (e.g., 'Pyramidobacter', 'Jonquetela', 'Anthropi'). By computing the bounding boxes we were able to compute inter-word vectors and deduce whether these were sequential on a horizontal line, or vertical (different lines).
- We cross-checked putative Genbank accessions numbers against NCBI's records. This is a very strong check on correctness.
- Taxon labels were attached to the tips. This is an easy operation for humans, but not trivial for machines, since there is often variation in how authors add name annotations.
Once these operations were complete, tip labels and graphs were combined into NeXML format.

**MRP-matrix creation:**

After taxon-name standardisation across the 924 source Newick strings, trees were converted into a matrix-representation with parsimony (MRP) matrix using the open source command-line programme *mrpmatrix* ([https://github.com/smirarab/mrpmatrix](https://github.com/smirarab/mrpmatrix), Mirarab et al. 2014). This process created an MRP matrix of 2,269 unique species by 6,261 parsimony-informative 'group inclusion' characters. The matrix was extremely sparse: 99.4% of the matrix was coded as missing data (?). This sparsity is not unexpected: Thomson and Shaffer (2009) report successfully using a 93% missing data matrix to accurately infer species relationships of Testudines for a matrix of 213 taxa by 10,000 characters.

**Analysis of the MRP matrix using Maximum Parsimony**

The MRP matrix was analysed with the closed source command-line programme *TNT* version 1.1 (Goloboff et al. 2008) using traditional search techniques, specifically: 100 random addition sequences saving up to 1 tree per replication, swapping trees with Tree-Bisection Reconnection (TBR), with a 24-hour timeout command. The strict consensus of all shortest length trees was saved, collapsing unsupported relationships with zero-length branches. This supertree was used for all subsequent comparisons.

To compare our supertree to the NCBI taxonomy tree, a pruned NCBI taxonomy tree with labels exactly matching the 2,269 in our supertree was created using *PhyloT* ([http://phylot.biobyte.de/](http://phylot.biobyte.de/)). Descriptive statistics for both the supertree and the pruned NCBI taxonomy tree, relative to the MRP matrix, were calculated in PAUP* version 4.0b10 (Swofford 2002) including the consistency index (CI; Kluge and Farris 1969), retention index (RI; Farris 1989) for each tree, and the Robinson-Foulds distance (Robinson and Foulds 1981) between the supertree and the pruned NCBI taxonomy tree.

**Open Notebook Science working practices**

When this study was initiated (early 2013), the primary emphases were on methodology and extraction of scientific results. At that time, the UK had not enacted the ‘Hargreaves’ copyright exception allowing mining and fair quotation, and so we could not start with a completely ‘Open’ methodology. Nevertheless, our tools were developed in the expectation that copyright would be liberalised in the UK, and that it would be possible to mine images on a large scale. In 2014-17 the exception was enacted and we were able to partially implement an ‘Open Notebook Science’ (Bradley 2007) approach for the project, where data, intermediate results and discourse were available to the whole world at the time they were created or published. ONS is informed by practices in Free and Open Software (F/LOSS) which have proved to be very valuable in creating, sharing and re-using code. We believe that ONS has features which benefit science in several ways:
• All results are captured and saved immediately. There is no process of ‘writing up’ (revisiting discourse and data from disparate sources which are often difficult to find). ONS saves components in persistent, versioned repositories where data never ‘gets lost’ and where the history of operations can be completely recovered.

• The results are shared within and beyond the team. There is no need to e-mail or otherwise distribute versions of the data, since everyone shares the same pointers/addresses to the data. The systems allow branching (‘forking’) so that experiments (e.g. re-analyses) can be carried out without corrupting the data.

• In some projects, the wider community can communicate with the project and re-use the data, add observations or in some cases even become active contributors to the project.

• Increased quality, often encouraged by having the raw data available to everyone immediately. F/LOSS puts a high value on validation and communally agreed quality. This is often done automatically (‘unit and integration tests’) so that the project knows that code is fit for purpose without having to check history. This is harder for data and discourse, but it is possible to check that all components are present and reviewed by humans or machines.

ONS is relatively new and there are fewer systems available to support it. Our ONS software stack was based on freely usable repositories on Github (https://github.com/ContentMine/ijsem) and Bitbucket (https://bitbucket.org/petermr/ami-plugin), with open communications hosted on a Discourse installation (https://github.com/discourse/discourse; http://discuss.contentmine.org/c/community/phylogeny) as used by rOpenSci (https://discuss.ropensci.org/) amongst other projects. Since we only introduced ONS halfway through the project, we decided to use Git to support our repositories and Disqus to support threaded, searchable discourse.

Data resources

Due to copyright restrictions imposed by the publisher of IJSEM, we do not feel that we can safely share all of the 5,816 source PDFs or the 8,221 figure images we found in those PDFs, that are used or referred-to in this study. However, we do provide a list of the URLs of these 5,816 publications as supplementary material (Suppl. material 1). We note that if all these publications had been published open access under the Creative Commons Attribution License (Hagedorn et al. 2011) we could have provided this source material alongside this publication to make this work more easily reproducible.

Results

The automated image processing was a lossy-process (see Fig. 3). We obtained re-usable, machine-readable data using ami-phylo in a completely automated manner from just 924 of the 4,336 input images (21.3%). There was a complete failure to output any phylogenetic data from 931 of the images (21.5%). Of the 3,405 output phylogenetic data files from ami-
phylo, 997 contained simply 'null;' and 955 were partially complete but contained a warning of 'UNKNOWN'. There were 529 files that contained only partial subtrees containing 3 or fewer leaves (terminals) and these were discarded. This left 924 phylogenetic tree data files containing trees comprising 4 or more leaves (see Fig. 4 and Fig. 5 for example source tree output).

Figure 3.
Number of leaves (terminal taxa) in each of 1614 source tree images (blue) and number of leaves recovered-from each image (orange). The modal number of taxa recovered per image was 12, the median was 13, and the mean was 13.96. The modal number of taxa not recovered from the trees was 2, the median was 5 and the mean was 7.15. The image mining process is lossy since most output tree files did not recover all of the taxa from the source image.

Figure 4.
Output from image analysis of the input tree image in figure 1. All taxa and relationships are correctly reproduced, with branch lengths also preserved with high fidelity. (Note that the vertical ordering of the tips is not meaningful and is arbitrarily created by the display software.)
These 924 Newick format tree files were then concatenated into one file, in a known order to preserve the chain of data provenance. We determined that 48 taxon labels in the 924-source-tree-set represented non-specific or environmental taxa such as "Marine clone", "Peat bog", "Leptotrichia oral", "Human colonic", "Hot spring", "Halophilic bacterium", "Sea urchin" and "Termite gut" These 48 taxa were deleted from the 924 source Newick trees.

Many of the taxon name strings given in each of the 924 source tree Newick strings were either misspelt through OCR errors, were invalid synonymous taxon names relative to modern NCBI taxonomy, or had only a species name (with no genus given). Across the 924 source trees supplied from ami-phylo, there were 1,742 unique taxon name strings encountered that did not initially match existing and valid NCBI binomial names. We used the open source command-line program tre-agrep 0.7.2 (http://packages.ubuntu.com/xenial/tre-agrep) to help semi-automate the re-matching of incorrect names to correct names by comparing each OCR'd taxon name string to valid taxon names in NCBI's taxdump and suggesting the best match. After human-review, it was determined that tre-agrep using Levenshtein edit-distance matching alone correctly suggested the correct name for 1417 out of 1742 (81%) of the non-matching OCR'd name strings. We acknowledge but did not use the Taxamatch algorithm (Rees 2014). The remaining 325 taxon names for which tre-agrep suggested an incorrect name were manually assigned their correct taxonomic name relative to NCBI taxonomy. These mistakes ranged from an edit distance of 2 for "Pichia silvicola" (original OCR string: suggested by tre-agrep to be a match to Helicic silvicola but actually represents Wickerhamomyces silvicola) up to an edit distance of 12 for "Gaetbulimicrobium brevivitae" (original OCR string: suggested by tre-
agrep to be *Methylomicrobium buryatense* but actually represents the taxon we now call *Aquimarina brevivitae*).

**Supertree Analysis Results**

The maximum parsimony analysis of the MRP matrix timed-out after 24-hours, equating to 40 random addition sequence searches. The best (shortest) tree length was 7,834 steps (Fig. 6), and 336,124,385,824 different rearrangements were examined during this search. The consistency index of the supertree to the MRP matrix was 0.780 and the retention index was 0.874. Unsurprisingly the pruned NCBI taxonomy tree did not match the MRP matrix data as well: it had a consistency index of 0.415 and retention index of 0.369.

In terms of tree-to-tree distance measures, the supertree and the pruned NCBI taxonomy tree are clearly different: the Robinson-Foulds distance (Robinson and Foulds 1981) between them is 1,691. A representative example of the kind of tree-to-tree differences encountered is depicted in Fig. 7. We cannot discuss all such differences exhaustively as there are far too many.
Instead we used alternative measures of tree-to-tree distance to complement the Robinson-Foulds distance. The following tree-to-tree comparisons between our supertree and the NCBI taxonomy tree for the same set of 2269 taxa were all calculated with Dendroscope version 3.5.8 (Huson and Scornavacca 2012).

- Tripartition distance (Moret et al. 2004): 867.0
- Hardwired cluster distance (Huson et al. 2010): 895.0
- Softwired distance (Huson et al. 2010): 895.5
- Path multiplicity distance (Cardona et al. 2009a): 868.0

We tried to compute the significance of difference between our supertree and the NCBI taxonomy tree but found that neither PAUP* (Swofford 2002) nor TreeCmp (Bogdanowicz et al. 2012) nor any other software implementation we know of could handle the size and difference of our trees. Data documenting all input (including all 924 machine-readable source trees and the reference 2,269 tip NCBI taxonomy tree) and output files from analyses presented in this subsection are archived at Zenodo (Mounce and Murray-Rust 2017).
Discussion

The PLUTo workflow implements several key advances simultaneously:

Optical Character Recognition combined with ‘Optical Tree Recognition’ so that phylogenetic branch lengths and relationships and tip-label data are recovered from an image, correctly matched-up with tips and output into an immediately re-usable format for further phylogenetic analysis.

This is one of the largest formal supertree syntheses ever created in terms of source trees used and number of tips feeding into the MRP-matrix (see Table 1). Since it only used a quarter of the input images, it will be considerably bigger when the software is developed to process diagrams currently rejected as unprocessable or error-rich. Even though the Open Tree of Life project (Hinchliff et al. 2015) is a synthetic tree, not a formal supertree, its impressive 2.3 million taxon tip coverage derives from just 785 source publications (https://tree.opentreeoflife.org/about/references; accessed 2017/03/29), of which 424 (54%) had data already deposited in TreeBASE. Acquisition of accurate machine-readable source tree data is still clearly the biggest rate limiting factor in phylogenetic syntheses.

<table>
<thead>
<tr>
<th>Taxon-focus</th>
<th>Number of Source Trees</th>
<th>Number of Tips</th>
<th>Year of Publication</th>
<th>Bibliographic Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microbial taxa</td>
<td>924</td>
<td>2269</td>
<td>2017</td>
<td>(this study)</td>
</tr>
<tr>
<td>Teleostei</td>
<td>120</td>
<td>617</td>
<td>2016</td>
<td>Clarke et al. 2016</td>
</tr>
<tr>
<td>Anomura</td>
<td>60</td>
<td>372</td>
<td>2016</td>
<td>Davis et al. 2016</td>
</tr>
<tr>
<td>Marseilleviridae</td>
<td>5</td>
<td>9</td>
<td>2016</td>
<td>Dornas et al. 2016</td>
</tr>
<tr>
<td>Ornithopoda</td>
<td>5</td>
<td>112</td>
<td>2016</td>
<td>Strickson et al. 2016</td>
</tr>
<tr>
<td>Decapoda: Achelata</td>
<td>55</td>
<td>475</td>
<td>2015</td>
<td>Davis et al. 2015</td>
</tr>
<tr>
<td>Birds</td>
<td>1036</td>
<td>6326</td>
<td>2014</td>
<td>Davis and Page 2014</td>
</tr>
<tr>
<td>Lissamphibians</td>
<td>89</td>
<td>319</td>
<td>2013</td>
<td>Marjanović and Laurin 2013</td>
</tr>
</tbody>
</table>

Table 1. **A comparison of the size of our supertree and other published formal supertrees.** This tabulation is not intended to be exhaustive. Supertree studies have been omitted if it was unclear how many source trees contributed to the supertree, or if the supertree study was superseded by a newer and more inclusive study.
<table>
<thead>
<tr>
<th>Taxon</th>
<th>Species</th>
<th>GenBank</th>
<th>Year</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carnivora</td>
<td>188</td>
<td>294</td>
<td>2012</td>
<td>Nyakatura and Bininda-Emonds 2012</td>
</tr>
<tr>
<td>Corals</td>
<td>15</td>
<td>1293</td>
<td>2012</td>
<td>Huang 2012</td>
</tr>
<tr>
<td>Hymenoptera</td>
<td>77</td>
<td>134</td>
<td>2010</td>
<td>Davis et al. 2010</td>
</tr>
<tr>
<td>Dogfish sharks</td>
<td>11</td>
<td>24</td>
<td>2010</td>
<td>Klug and Kriwet 2010</td>
</tr>
<tr>
<td>Galloanserae</td>
<td>400</td>
<td>376</td>
<td>2009</td>
<td>Eo et al. 2009</td>
</tr>
<tr>
<td>Mammalia (not specifi)</td>
<td>5020</td>
<td></td>
<td>2009</td>
<td>Fritz et al. 2009</td>
</tr>
<tr>
<td>Cyprinidae</td>
<td>56</td>
<td>397</td>
<td>2009</td>
<td>Gaubert et al. 2009</td>
</tr>
<tr>
<td>Dinosauria</td>
<td>165</td>
<td>455</td>
<td>2008</td>
<td>Lloyd et al. 2008</td>
</tr>
<tr>
<td>Adepagha</td>
<td>43</td>
<td>309</td>
<td>2008</td>
<td>Beutel et al. 2008</td>
</tr>
<tr>
<td>Drosophilidae</td>
<td>117</td>
<td>624</td>
<td>2008</td>
<td>der Linde and Houle 2008</td>
</tr>
<tr>
<td>Temnospondyli</td>
<td>30</td>
<td>173</td>
<td>2007</td>
<td>Ruta et al. 2007</td>
</tr>
<tr>
<td>Ruminantia</td>
<td>164</td>
<td>197</td>
<td>2005</td>
<td>Fernández and Vrba 2005</td>
</tr>
<tr>
<td>Cetartiodactyla</td>
<td>141</td>
<td>290</td>
<td>2005</td>
<td>Price et al. 2005</td>
</tr>
<tr>
<td>Angiosperms</td>
<td>46</td>
<td>379</td>
<td>2004</td>
<td>Davies et al. 2004</td>
</tr>
</tbody>
</table>

**Post-hoc analyses**

After writing-up most of this paper, one of us (RM) attended a workshop hosted by the authors of the Supertree Toolkit (Davis and Hill 2010) and Supertree Toolkit 2 (Hill and Davis 2014), called "Tools and methods for constructing the Tree of Life" (https://jonxhill.wordpress.com/2016/11/15/tools-and-methods-for-constructing-the-tree-of-life/).

RM learned how to use Supertree Toolkit 2 in a modified manner to do an assessment of the overlap of the 924 source trees that were put into the MRP matrix (Fig. 8). Unfortunately, this analysis demonstrated that the 924 source trees extracted from *IJSEM* do not link-up to form one contiguous, connected island of data (as depicted in the centre). This probably explains why the supertree is so discordant from the NCBI taxonomy tree, in places. We would recommend users of phylogeny interested in microbial phylogeny to use the Open Tree of Life (Hinchliff et al. 2015) or SILVA (Yilmaz et al. 2013) phylogenies, and not this experimental phylogeny. In future analyses, we hope to build this type of overlap analysis into our PLUTO workflow so that "unconnected" trees can be excluded prior to analysis. It is hoped that with the continued accumulation of machine-readable phylogenetic data we will be able to connect up the gaps in our knowledge so that all microbial source trees can be meaningfully used in future iterations of this work.
Acknowledgements

We would like to thank Jon Hill and Katie Davis for providing expert guidance in the usage of STK2 software.

Funding program

BBSRC Tools and Resources Development Fund (TRDF)

Grant title

PLUTo: Phyloinformatic Literature Unlocking Tools. Software for making published phyloinformatic data discoverable, open, and reusable. We are grateful to the BBSRC (grant BB/K015702/1 awarded to MAW and supporting RM) for funding this research.

Hosting institution

The University of Bath
References


Supplementary materials

Suppl. material 1: List of URLs of the 5816 source PDFs used in this research

Authors: Ross Mounce
Data type: URL links
Brief description: A one-per-line UTF-8-encoded plain-text list of URLs of the 5816 source PDFs used in this research
Filename: 5816sourcepublications.txt - Download file (309.52 kb)

Suppl. material 2: NeXML data from figure 5

Authors: Ross Mounce
Data type: NeXML
Brief description: The machine-readable text from the screenshot.
Filename: ijs.0.65226-0-000.pbm.nexml.xml - Download file (4.22 kb)