Reference Type: Journal Article

Authors: Bengtsson, Johan; Eriksson, K. Martin; Hartmann, Martin; Wang, Zheng; Shenoy, Belle Damodara; Grelet, Gwen-Aëlle; Abarenkov, Kessy; Petri, Anna; Rosenblad, Magnus Alm and Nilsson, R. Henrik.

Primary Title: Metaxa: a software tool for automated detection and discrimination among ribosomal small subunit (12S/16S/18S) sequences of archaea, bacteria, eukaryotes, mitochondria, and chloroplasts in metagenomes and environmental sequencing datasets

Journal Name: Antonie van Leeuwenhoek

Cover Date: 2011-10-01

Publisher: Springer Netherlands
ISSN: 0003-6072
Subject: Biomedical and Life Sciences

Start Page: 471
End Page: 475
Volume: 100
Issue: 3

Url: http://dx.doi.org/10.1007/s10482-011-9598-6
Doi: 10.1007/s10482-011-9598-6

Short Communication

Metaxa: A software tool for automated detection and discrimination among ribosomal small subunit (12S/16S/18S) sequences of archaea, bacteria, eukaryotes, mitochondria, and chloroplasts in metagenomes and environmental sequencing datasets

Johan Bengtsson[1,2,*], K. Martin Eriksson[1], Martin Hartmann[3], Zheng Wang[4], Belle Damodara Shenoy[5], Gwen-Aëlle Grelet[6], Kessy Abarenkov[7], Anna Petri[1], Magnus Alm Rosenblad[2], R. Henrik Nilsson[1,7]

[1] Department of Plant and Environmental Sciences, University of Gothenburg, Box 461, 405 30 Göteborg, Sweden

[2] Department of Cell and Molecular Biology, University of Gothenburg, Medicinaregatan 9C, Box 462 SE 405 30 Göteborg, Sweden

[3] Department of Microbiology and Immunology, University of British Columbia, Life Sciences Centre, 4504-2350 Health Sciences Mall, Vancouver, BC, V6T 1Z3 Canada

[4] Department of Ecology and Evolutionary Biology, Yale University, POB 208106, 165 Prospect Street, New Haven, CT 06520-8106, USA

[5] Microbial Type Culture Collection and Gene Bank, Institute of Microbial Technology (CSIR-IMTECH), Sector 39A, Chandigarh 160036, India

[6] Landcare Research, P.O. Box 40, Lincoln 7640, New Zealand

[7] Department of Botany, Institute of Ecology and Earth Sciences, University of Tartu, 40 Lai St., 51005 Tartu, Estonia

Abstract

The ribosomal small subunit (SSU) rRNA gene has emerged as an important genetic marker for taxonomic identification in environmental sequencing datasets. In addition to being present in the nucleus of eukaryotes and the core genome of prokaryotes, the gene is also found in the mitochondria of eukaryotes and in the chloroplasts of photosynthetic eukaryotes. These three sets of genes are conceptually paralogous and should in most situations not be aligned and analyzed jointly. To identify the origin of SSU sequences in complex sequence datasets has hitherto been a time-consuming and largely manual undertaking, but the present study introduces Metaxa (http://microbiology.se/software/metaxa/), an automated software tool to extract full-length and partial SSU sequences from larger sequence datasets and assign

them to an archaeal, bacterial, nuclear eukaryote, mitochondrial, or chloroplast origin. Using data from reference databases and from full-length organelle and organism genomes, we show that Metaxa detects and scores SSU sequences for origin with very low proportions of false positives and negatives. We believe that this tool will be useful in microbial and evolutionary ecology as well as in metagenomics.

Keywords Metagenomics ; microbial communities ; rRNA libraries ; phylogenetic assignment

* Corresponding author. Email: johan@microbiology.se Tel. +46-31-786 2911

Body

Recent methodological advancements in the fields of high-throughput DNA amplification and sequencing have opened new windows on many research questions in the life sciences (Shendure & Ji 2008). One such area is metagenomics, where the total DNA found at any sample site is sequenced and analyzed through, e.g., massively parallel ("454") pyrosequencing (Margulies et al. 2005) or Illumina sequencing (Bentley 2006). This makes a wide range of functional and ecological inferences pertaining to the roles and capacities of the underlying species community possible (Trevors & Masson 2010; Wooley et al. 2010). Common to these pursuits is usually the need, or desire, to also examine the taxonomic composition of the community recovered. This is typically achieved through similarity searches of the ribosomal 12S/16S/18S small subunit (SSU) sequences of the query dataset against nucleotide sequence databases such as GenBank (Benson et al. 2009), SILVA (Preusse et al. 2007), and RDP (Cole et al. 2009).

The process of identifying and annotating sequences with respect to taxonomic affiliation is not trivial and often requires both manual intervention and some degree of familiarity with the lineages recovered (Christen 2008; Kang et al. 2010; Nilsson et al. 2011). Furthermore, as the complexity of the samples and sample sites increase, so does that of the sequence identification process. The SSU, in addition to being present in the nucleus of eukaryotes and the core genome of prokaryotes, is also found in the mitochondria of eukaryotes and in the chloroplasts of photosynthetic eukaryotes. In the two last cases the gene has independent endosymbiotic origins. As a consequence, these different SSU rRNAs should normally not be incorporated into, e.g., joint multiple alignments for taxonomic identification, phylogenetic analysis, or ecological inferences. Thus, if the metagenome under scrutiny contains prokaryotes as well as eukaryotes, there are many situations where the distinct classes of SSU sequences need to be delimited and extracted for separate analysis. This is a time-consuming and largely manual exercise that is further complicated by the considerable proportion of incorrectly identified or otherwise poorly annotated reference entries in the public sequence databases (Bidartondo et al. 2008; Ryberg et al. 2009). The present study offers a remedy, however, in the form of an open source MacOS X/Linux/UNIX software tool – Metaxa – for automated detection and discrimination among ribosomal SSU sequences from archaea, bacteria, eukaryotes, mitochondria, and chloroplasts in large datasets (Online Resource 1; http://microbiology.se/software/metaxa/). The source code is written in Perl and takes advantage of multiple processor cores if available. Internet access is not needed to run the software.

Metaxa has a two-step analysis procedure, where each step may be run separately as needed: it first extracts all SSU sequences from the dataset and then subjects only the SSU sequences to detailed analysis, thus bypassing the need to spend further time on sequences that are not SSU in the first place. It expects query sequences of any number in the FASTA format (Pearson & Lipman 1988). By default, Metaxa starts by examining the query dataset for the presence of SSU sequences of any of the five origins. This is accomplished through HMMER 3.0 (Eddy 1998), the archaeal, bacterial, and eukaryote hidden Markov models (HMMs) of V-Xtractor 2.0 (Hartmann et al. 2010), and a set of newly generated HMMs for the mitochondrial and chloroplast SSU (Online Resource 1). Following the V-Xtractor recommendations, we built the new HMMs from conserved, ~50 basepair (bp.) sequence segments distributed across the full length of the SSUs; an average of 11 HMMs were made for each origin. The first step finds SSU sequences ranging from full length down to about 100-200 bp. and assigns them to a tentative origin based on the HMM (e.g., bacterial) that produced the best match to the sequence in question. Some regions of the SSU are however highly conserved across the organelles and lineages of the tree of life such that HMMs computed for several different origins could potentially produce nearly equally good matches to those regions, cautioning against a final decision already at this stage. Instead, the second step uses the extracted SSU entries in BLAST-based sequence similarity searches (Altschul et al. 1997) against local filtered copies of the manually curated prokaryote, eukaryote, mitochondrial, and chloroplast SSU entries of the GreenGenes (DeSantis et al. 2006), SILVA, CRW (Cannone et al. 2002), and MitoZoa (Lupi et al. 2010) databases. By default, the five best BLAST matches of each query are examined for origin (archaea, bacteria, eukaryote, mitochondria, or chloroplast). The origin of the best BLAST match is given a score of 5; the origin of the second best match a score of 4; that of the third a score of 3; the fourth 2; and the fifth 1. In addition, the origin determined by HMMER is given a score of 5 in order to make the HMMER step influential but not decisive. The score is then summed up for each origin. If the origin with the highest score is the same as the origin suggested by HMMER, the query sequence is assigned to that origin. If the origin with the highest score is different from that suggested by HMMER, the sequence is still assigned to the origin with the highest score but marked as in potential need of further scrutiny. Cases in which the score among origins are tied are treated in the same way, except that the corresponding sequences are classified as “uncertain” and that a multiple alignment is computed in MAFFT (Katoh & Toh 2008) for the query sequence together with its five best BLAST matches to facilitate manual examination and interpretation. This dual approach where both HMMER and the best

BLAST matches influence the final decision minimizes the effect of single incorrectly annotated, or otherwise problematic, reference sequences, whose presence would distort efforts based on BLAST alone. The second step concludes by writing a separate FASTA file for each origin found (e.g., queryfile.SSU_bacteria.fasta), with each such file containing all query sequences of the origin in question. In addition, a detailed log file is generated.

To evaluate the efficacy of Metaxa, we downloaded the 262,032 SSU sequences of the non-redundant SILVA 102 release that were annotated to origin (bacterial, archaeal, nuclear eukaryote, mitochondrial, or chloroplast). We required that each sequence should produce matches to at least two HMMs for the sequence to be classified as an SSU sequence. Metaxa identified more than 99.95% of the sequences to the correct origin (130 out of 262,032 (0.05%) sequences were classified to a different origin than that given by SILVA; Online Resource 2), although a slight drop in accuracy was noted for the mitochondrial sequences (18 out of 434 (4.15%) mitochondrial sequences were assigned to a different origin than that given by SILVA). We furthermore collected 100 random chloroplast SSU sequences from the full-length chloroplast genomes of cpBase (http://chloroplast.ocean.washington.edu/); 100 random SSU sequences from the full-length mitochondrial genomes of GOBASE (O'Brien et al. 2009) and MitoZoa; 100 SSU sequences from the full-length bacterial genomes of UCSC Archaeal Genome Browser (Schneider et al. 2006); 80 SSU sequences from the 80 public full-length archaeal genomes of UCSC Archaeal Genome Browser; and 100 eukaryote SSU sequences from the non-redundant SILVA 104 release. The sequence corpus was run in eight versions through Metaxa to mimic read lengths ranging from those obtained through traditional Sanger sequencing down to those obtained from present pyrosequencing technology and below: the full length, 1250 bp., 1000 bp., 750 bp., 500 bp., 300 bp., 200 bp., and 100 bp. Each length *n* was run in two versions: the first *n* basepairs of the SSU, and a random segment of *n* basepairs along the SSU. The first case simulates traditional, targeted PCR whereas the second case simulates metagenomic data. All 480 sequences were correctly identified to their respective origin in the full-length dataset (Table 1). At pyrosequencing read lengths of 500 bp., the percentage of correct assignments for both versions was at or above 99% for archaea and bacteria; at or above 98% for nuclear eukaryote and chloroplasts; and at or above 94% for mitochondria (Table 1; Online Resource 3). To evaluate the susceptibility of Metaxa to false positives, we generated five 5-million-sequence datasets of random nucleotide data of the lengths 1250 bp., 1000 bp., 750 bp., 500 bp., and 300 bp. in the EMBOSS 6.2.0 suite (Rice et al. 2000). As above, these datasets were run with the requirement that a sequence must produce matches against at least two HMMs to

be classified as an SSU sequence. Three of these 25 million sequences (0.00012%) were incorrectly identified as SSU sequences, suggesting a considerable robustness against false-positive matches (Online Resource 4).

We view the proportion of incorrect assignment to origin – on average well below 0.5% for sequences longer than 750 bp. - as acceptable given the complex evolutionary history of the SSU as reflected across the organelles and lineages of the tree of life. Although expertly curated and further filtered in this study, the BLAST databases employed by the software are likely to contain a small proportion of taxonomically misidentified or otherwise anomalous entries (Hartmann et al. 2011), which would add some degree of noise to the present effort. Lineages that hold basal positions within the five origins – such as those close to the mitochondria/alphaproteobacteria or the chloroplast/cyanobacteria ancestor demarcations – are probably more likely to be incorrectly assigned to origin than lineages deeply nested within the respective clades. Sequences from previously undiscovered or sparsely sampled lineages are similarly subject to a higher risk of misclassification. In light of these observations, we recommend that the entries on whose origin a final decision could not be reached should be examined manually. The software outputs ample information – including multiple alignments in the case of sequences of uncertain assignment – to assist such scrutiny. The focus on shorter sequences of the metagenomics type, as well as the ability to sort those sequences into the five different origins targeted, set Metaxa apart from RNAmmer (Langesen et al. 2007), which is a HMM-based software resource for detection of rRNA genes in full genome sequences. When compared for performance on the data underlying Table 1, Metaxa outperformed RNAmmer in terms of accuracy and speed on all sets of SSU sequences examined. In addition, Metaxa was able to satisfactory address sequences shorter than 1000 bp. as well as sequences of mitochondrial and chloroplast origin, both of which are out of reach for RNAmmer (Online Resource 5).

The time needed by Metaxa to analyse a dataset scales linearly with the number of SSU sequences, such that a doubling of the number of SSU sequences will, on average, double the runtime. Non-SSU sequences do not add much to the runtime, such that a one-million-sequence pyrosequencing metagenome with 0.5% SSU sequences will be processed in under two hours. The test corpus of 262,032 true-positive SILVA SSU sequences took 34 hours to run on a twelve-core 2.0 GHz Linux machine, and the five-million, 1250 bp. sequence dataset of true negatives took 19 hours. Since Metaxa loads the query sequences sequentially, there is no restriction on the number of query sequences. For the same reason,

Metaxa does not require large amounts of computer memory; at no point during the execution of the 262,032 true-positive SSU dataset was more than ~250 Mb of memory needed.

In conclusion, Metaxa detects SSU entries in larger bodies of sequences – such as metagenomes and environmental sequencing datasets - and assigns them to origin with a negligible proportion of false positives and negatives and at a relatively high speed. To rely solely on BLAST for the same purpose, in contrast, would be many times slower and less precise, and would require significant manual intervention. Metaxa is freely available under the GNU GLP v. 3 software licence (Online Resource 1; http://microbiology.se/software/metaxa/), and it is written in a way that makes integration into existing software pipelines for analysis of environmental sequences straightforward. We believe it may increase accuracy in the annotation and analysis of metagenomes and similar datasets, whose sizes tend to defy most attempts at manual processing and examination.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgements

The Frontiers in Biodiversity Research Centre of Excellence (University of Tartu) and the Platform in Ecotoxicology – From Gene to Ocean (University of Gothenburg) are gratefully acknowledged for their support.

References

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389-3402.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2009) GenBank. Nucleic Acids Res 37:D26-D31.

Bentley DR (2006) Whole genome re-sequencing. Curr Opin Genet Dev 16:545-552.

Bidartondo MI, Bruns TD, Blackwell M et al. (2008) Preserving accuracy in GenBank. Science 319:1616.

Cannone JJ, Subramanian S, Schnare MN 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.

Cole JR, Wang Q, Cardenas E et al. (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res 37:D141-D145.

Christen R (2008) Global sequencing: A review of current molecular data and new methods available to assess microbial diversity. Microbes and Environ 23:253-268.

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 72:5069-5072.

Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14:755-763.

Hartmann M, Howes CG, Abarenkov K, Mohn WW, Nilsson RH (2010) V-Xtractor: An open-source, high-throughput software tool to identify and extract hypervariable regions of small subunit (16S/18S) ribosomal RNA gene sequences. J Microbiol Meth 83:250-253.

Hartmann M, Howes CG, Veldre V et al. (2011) V-RevComp: Automated high-throughput detection of reverse complementary 16S ribosomal RNA gene sequences in large environmental and taxonomic datasets. FEMS Microbiol Lett (in press, DOI: 10.1111/j.1574-6968.2011.02274.x).

Kang S, Mansfield MA, Park B, Geiser DM, Ivors KL, Coffey MD, Grünwald NJ, Martin FN, Lévesque CA, Blair J (2010) The promise and pitfalls of sequence-based identification of plant-pathogenic fungi and oomycetes. Phytopathology 100:732-737.

Katoh K, Toh H (2008) Recent developments in the MAFFT multiple sequence alignment program. Brief Bioinform 9:286-298.

Langesen K, Hallin P, Rødland EA, Stærfeldt H-H, Rognes T, Ussery DW (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res 35:3100-3108.

Lupi R, D'Onorio de Meo P, Picardi E, D'Antonio M, Paoletti D, Castrignanò T, Pesolec G, Gissi C (2010) MitoZoa: A curated mitochondrial genome database of metazoans for comparative genomics studies. Mitochondrion 10:192-199.

Margulies M, Egholm M, Altman WE et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376-380.

Nilsson RH, Tedersoo L, Lindahl BD et al. (2011) Towards standardization of the description and publication of next-generation sequencing datasets of fungal communities. New Phytol (in press, DOI: 10.1111/j.1469-8137.2011.03755.x).

O'Brien EA, Zhang Y, Wang E, Marie V, Badejoko W, Lang BF, Burger G (2009) GOBASE - an organelle genome database. Nucleic Acids Res 37:D946-950.

Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. P Natl Acad Sci USA 85:2444-2448.

Preusse EC, Quast C, Knittel K, Fuchs B, Ludwig W, Peplies J, Glöckner FO (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res 35:7188-7196.

Rice P, Longden I, Bleasby A (2000) EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet 16:276-277

Ryberg M, Kristiansson E, Sjökvist E, Nilsson RH (2009) An outlook on the fungal internal transcribed spacer sequences in GenBank and the introduction of a web-based tool for the exploration of fungal diversity. New Phytol 181:471-477.

Schneider KL, Pollard KS, Baertsch R, Pohl A, Lowe TM (2006) The UCSC Archaeal Genome Browser. Nucleic Acid Res 34:D407-D410.

Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotechnol 26:1135-1145.

Trevors JT, Masson L (2010) DNA technologies: what's next applied to microbiology research? Antonie Leeuwenhoek 98:249-262.

Wooley JC, Godzik A, Friedberg I (2010) A primer on metagenomics. PLoS Comput Biol 6:2.

Online Resource 1. The software package together with its documentation, reference sequences, and a test dataset (including, for illustratory purposes, ten sequences from each of the five origins plus ten non-SSU sequences). In addition the user will have to install NCBI-BLAST, HMMER, and MAFFT; detailed installation instructions are provided in the documentation.

Online Resource 2. The summary of the analysis of the non-redundant SILVA 102 release.

Online Resource 3. The 480 SSU sequences retrieved from full-length organelle and organism genomes, and the results of their analysis.

Online Resource 4. The results of the analysis of the five 5-million random sequence datasets.

Online Resource 5. Comparison of detection and classification performance of RNAmmer and Metaxa on the same sequence set of 80 archaeal, 100 bacterial, 100 eukaryote, 100 chloroplast, and 100 mitochondrial SSU sequences extracted from genome sequences. Only classification of archaeal, bacterial, and eukaryotic entries were considered since RNAmmer does not readily support chloroplast or mitochondrial sequences; entries of the latter two origins were however kept in the input dataset to serve as potential decoys. The top three squares (left to right) represent the three origins that RNAmmer sorted the query sequences into: archaeal (left), bacterial (middle), and eukaryotic (right). Inside each square, the true origins of those sequences are specified. Ideally, thus, a square should only contain sequences of one origin – namely the one given in bold – and in a quantity matching the number of input sequences for that origin (above). The bottom three squares (left to right) represent the three origins that Metaxa sorted the query sequences into: archaeal (left), bacterial (middle), and eukaryotic (right). The table shows that RNAmmer is not suitable for detecting SSU sequences in fragmentary data such as metagenomes. In addition, RNAmmer does not readily handle mitochondrial and chloroplast SSU, which Metaxa identify and score to origin with high accuracy, even down to short read lengths.

Legends

Table 1. Number of correctly assigned entries of the 480 reference SSU sequence dataset as reported for the five different origins (80 sequences from archaea, 100 from bacteria, 100 from eukaryotes (nuclear), 100 from chloroplasts, and 100 from mitochondria). Results are shown for the eight sequence lengths examined. Each length *n* was run in two versions: the first *n* basepairs of the SSU, and a random segment of *n* basepairs along the SSU. The first case simulates traditional, targeted PCR whereas the second case simulates metagenomic data.

Table 1.

Sequence length	Archaea (80)		Bacteria (100)		Eukaryota (100)		Chloroplast (100)		Mitochondria (100)	
	Random	5′ end	Random	5′ end	Random	5′ end	Random	5′ end	Random	5′ end
Full-length	80	80	100	100	100	100	100	100	100	100
1 250 bp	80	80	100	100	100	100	100	100	99	100
1 000 bp	80	80	100	100	100	100	100	100	98	100
750 bp	79	80	100	100	100	100	99	100	98	99
500 bp	79	80	100	100	100	99	100	98	94	95
300 bp	78	80	100	100	96	94	99	99	91	91
200 bp	79	80	100	100	86	87	98	99	82	85
100 bp	66	74	90	95	62	84	98	95	63	77