Proceedings of the 1st Central European Diatom Meeting 2007
Kusber, W.-H. & Jahn, R. (ed.)
Botanic Garden and Botanical Museum Berlin-Dahlem, Freie Universität Berlin
ISBN 978-3-921800-63-8, © BGBM, Berlin 2007.
doi: 10.3372/cediatom.113 (available via http://dx.doi.org/)

Diatoms and DNA barcoding: A pilot study on an environmental sample

Regine Jahn[1], Holger Zetzsche[1], Richard Reinhardt[2] & Birgit Gemeinholzer[1]

[1]Botanischer Garten und Botanisches Museum Berlin-Dahlem, Freie Universität Berlin, Königin-Luise-Str. 6-8, 14195 Berlin, Germany; r.jahn@bgbm.org
[2]Max Planck Institute for Molecular Genetics, Ihnestr. 63-73, 14195 Berlin, Germany

INTRODUCTION

Diatoms live in all types of water bodies from fresh to marine habitats and also in soil and aero-terrestrial ecosystems. Because they are sensitive concerning pollution, acidification, and salinization, they are important bio-indicators and are often used for routine water quality assessments. Since morphological identification is time-consuming and demands specialized in-depth knowledge, diatoms are an ideal model group to establish DNA barcoding methods to provide an easy to use, standardized and fast organism identification tool. So far, little has been published on the topic DNA barcoding in diatoms (Evans et al. 2007; Kaczmarska et al. 2007).

DNA barcoding in general demands a molecular locus, being variable enough to discriminate on species level for the organisms under study and a molecular reference database for comparison. The similarity or divergence of the molecular sequence of an unknown organism to a vouchered reference sequence in the database is used as quality indicator for species identification. DNA barcoding of environmental samples requires DNA extraction from an environmental pooled sample, PCR amplification of a target locus; cloning of the resulting PCR products, sequencing and the analysis.

In a pilot study on a plankton water sample from Lake Tegel in Berlin, Germany, we wanted to test if identification of diatoms by classical identification method via morphology using light microscopy (LM) and by molecular means via DNA sequencing would result in comparable diversity assessments and test the potential for molecular based diatom identification. In the diatom flora of Berlin (Geissler & Kies 2003) 541 taxa are mentioned to occur in Lake Tegel (Geissler et al. 2006). We chose the 18S region to be amplified for our investigation as this marker has most often been used for phylogenetic analyses and therefore the number of comparable diatom sequences in reference databases was highest. 480 sequences were available for comparison (www.algaterra.org and GenBank: http://www.ncbi.nlm.nih.gov/). As the 480 used diatom sequences of the 18S region are from world wide occurrences, only a small number of individuals from Lake Tegel were expected to match precisely on the species level. By selecting this marker for our pilot investigation we have been aware, that it might not be the best for species discrimination but it was chosen to test the feasibility of the new methodology per se.

MATERIAL & METHODS

A one litre water plankton sample was taken from the north-eastern shore of Lake Tegel (Berlin, Germany) on 10 May 2006. The sample was concentrated by centrifugation at 2000 x g for 15 min and the supernatants were discarded. Cell recovery for DNA extraction of water plankton samples was proven to be optimal by centrifugation rather than filtration (Boström et al. 2004). The pellet was split into two parts; one half was used for identification and counting via LM. The second part was conserved at -20°C for molecular analyses.

For morphological identification of the diatom frustules, the first part was oxidized with H_2O_2, rinsed and mounted on slides in Naphrax, deposited at B, digital images of the identified specimens are databased in Jahn & Kusber (2007). From the second part total genomic DNA was isolated using two different DNA extraction methods, the DNeasy Mini Plant Kit (QIAGEN, Germany) and the Dynabeads DNA Direct Universal Kit (Dynal Biotech, Norway), the latter being non-destructive, leaving the frustules for latter verification. Sample pre-treatment for the QIAGEN protocol included pellet re-suspension in 100 µl extraction buffer and frustule destruction by grinding (Retsch MM301, Haan, Germany). For the Dynabeads protocol pellets were re-suspended in 10 µl water. DNA extraction was performed according to the respective DNA isolation protocol from the manufacturer.

The 18S locus was amplified for both extracts separately by using universal diatom-specific primers (Brinkmann et al. 2007, Friedl et al., in prep). PCR amplification was performed with the following protocol (initial denaturation: 5 min at 95°C, 20 cycles: 1min at 94°C, 45s at 50°C, 4 min at 72°C, final extension 10 min 72°C). Each 20 µl PCR consisted of 0.5 u *Taq* (QIAGEN, Germany), dNTPs (0.2 mM each), primers (0.5 µM each), 6 % DMSO and 1X PCR buffer. The adapted PCR protocol is in accordance to Wang & Wang (1996), Qui et al. (2001) and O'Brien et al. (2005) to avoid chimeric sequences and other PCR artefacts by lowering the number of cycles and prolonging the extension time. To obtain sufficient PCR product and minimise amplification bias (products being amplified during the first PCR cycles will always occur most frequently) for each DNA extraction method separately six PCR reactions were carried out and pooled. Pooled products were purified and concentrated to 20 µl using MSB® Spin PCRapace (Invitek, Berlin, Germany). Fragments were cloned using the TOPO TA CloningTM Kit (Invitrogene) transfered into *Escherichia coli* strain DH10B (Invitrogene) via electroporation. Recombinant plasmids were isolated by alkaline lysis and sequencing from both ends was performed (ABI PRISM® 3730 XL, Applied Biosystems). Due to the length of the 18S rDNA of approximately 1800 bp, only parts of the gene sequence were covered (450 bp). Sequences were checked against the molecular reference databases (GenBank: http://www.ncbi.nlm.nih.gov/ and Algaterra: www.algaterra.org) by using the BLAST ® algorithm (Altschul et al. 1990).

RESULTS

The LM investigation resulted in an identification of 62 taxa with about 1000 individuals being counted for the quantitative assessment; parts of these results are presented in Tables 1-4.

Table 1. Exact Matches of Morphological (LM) and Molecular Data (MOL) on the Species Level.

Taxon	**LM**	**MOL**	**Reference**
Achnanthidium minutissimum (Kütz.) Czarn.	x	x	AlgaTerra*
Amphora copulata (Kütz.) Schoeman & R.E.M.Archibald	x	x	AlgaTerra*
Amphora pediculus (Kütz.) Grunow	x	x	AlgaTerra*
Aulacoseira granulata (Ehrenb.) Simonsen	x	x	GenBank
Cymbella cymbiformis C.Agardh	x	x	GenBank
Diatoma tenuis C.Agardh	x	x	GenBank
Encyonema caespitosum Kütz.	x	x	AlgaTerra*
Fragilaria capucina Desmazières	x	x	AlgaTerra*
Fragilaria vaucheriae (Kütz.) J.B.Petersen	x	x	AlgaTerra*
Gomphonema olivaceum (Hornem.) Bréb.	x	x	AlgaTerra*
Melosira varians C.Agardh	x	x	GenBank
Navicula cari Ehrenb.	x	x	AlgaTerra*
Navicula gregaria Donkin	x	x	AlgaTerra*
Navicula tripunctata (O.F.Müll.) Bory	x	x	AlgaTerra*
Nitzschia dissipata (Kütz.) Grunow	x	x	AlgaTerra*
Nitzschia linearis (C.Agardh) W.Sm.	x	x	AlgaTerra*
Staurosira construens Ehrenb.	x	x	GenBank
Tabularia tabulata (C.Agardh) P.Snoeijs	x	x	AlgaTerra*
Ulnaria acus (Kütz.) Aboal	x	x	AlgaTerra*
Ulnaria ulna (Nitzsch) Compère	x	x	AlgaTerra*

*Molecular Data from AlgaTerra is yet unpublished.

The molecular analysis resulted in 349 sequences retrieved for the QIAGEN DNA extraction and 350 sequences from the Dynabead DNA extraction. Combining the results of both methods the 699 retrieved sequences resulted in 62 different "best hit species" when tested against GenBank and/or AlgaTerra (matches above 92 % similarity). The number of discovered taxa lowers to 56 if only matches above 96 % are considered; however, 8 out of 10 of these taxa are characterized by multiple matches. Only two taxa of the molecular species list, *Navicula recens* (Lange-Bert.) Lange-Bert. and *Craticula cuspidata* (Kütz.) D.G.Mann, result from sequence similarity below 96 % and only one match of the reference database.

The different DNA extraction methods resulted in different numbers of taxa identified. 39 taxa (35 considering > 96 % matches only) were discovered based on the QIAGEN extract whereas 48 (42) could be determined using the Dynabeads method.

Table 2. Fair Matches of Morphological (LM) and Molecular Data (MOL) on the Genus Level.

Taxon	**LM**	**MOL**	**Reference**	**comment**
Encyonema cespitosum Kütz.	x	x	AlgaTerra*	
Encyonema minutum (Hilse) D.G.Mann	x		AlgaTerra*	sequence available
Encyonema silesiacum (Bleisch) D.G.Mann	x			sequence missing
Encyonema triangulum (Ehrenb.) Kütz.		x	GenBank	not to be expected
Navicula brockmannii Hust.		x	AlgaTerra*	to be expected
Navicula cari Ehrenb.	x	x	AlgaTerra*	
Navicula gregaria Donkin	x	x	AlgaTerra*	
Navicula menisculus Schumann		x	AlgaTerra*	to be expected
Navicula phyllepta Kütz.		x	GenBank	Identification?
Navicula recens (Lange-Bert.) Lange-Bert.		x	AlgaTerra*	to be expected
Navicula reichardtiana Lange-Bert.	x			sequence missing
Navicula tripunctata (O.F. Müll.) Bory	x	x	AlgaTerra*	
Navicula veneta Kütz.		x	GenBank	to be expected

*Molecular Data from AlgaTerra is yet unpublished.

Table 3. Uneven Matches of Morphological (LM) and Molecular Data (MOL) in Centric Diatoms.

Taxon	**LM**	**MOL**	**Reference**	**comment**
Actinocyclus normanii f. *subsalsus* (Juhl.-Dannf.) Hust.	x			sequence missing
Aulacoseira ambigua (Grunow) Simonsen	x			sequence missing
Aulacoseira granulata (Ehrenb.) Simonsen	x	x	GenBank	
Cyclostephanos dubius (Hust.) Round	x			sequence missing
Cyclotella comta Kütz.	x			sequence missing
Cyclotella meneghiniana Kütz.		x	AlgaTerra*	to be expected
Cyclotella schumannii (Grunow) Håk.	x			sequence missing
Melosira varians C. Agardh	x	x	GenBank	
Stephanodiscus alpina Hust.	x			sequence missing
Stephanodiscus hantzschii Grunow	x		GenBank	sequence available
Stephanodiscus minutulus (Kütz.) Cleve et J.D. Möller		x	AlgaTerra*	to be expected
Stephanodiscus neoastraea Håk. et B. Hickel	x			sequence missing
Thalassiosira lacustris (Grunow) Hasle		x	AlgaTerra*	to be expected
Thalassiosira hendeyi Hasle & Fryxell		x	GenBank	not to be expected
Thalassiosira minima Gaarder		x	GenBank	not to be expected

*Molecular Data from AlgaTerra is yet unpublished.

Comparing the molecular and morphological identification 62 taxa were detected by both methods; however, the similarity is only superficial as the results turned out to be more heterogeneous than first expected: Only 20 exact taxon matches were detected by the molecular method and by LM (Table 1). 35 taxa identified by LM are not yet represented in the reference library and therefore had no matching sequences in GenBank and/or AlgaTerra. 25 taxa were detected by molecular method but were not found by LM; however, their presence in the lake could be expected. Three taxa identified by LM had matching sequences in the reference databases but were not detected by molecular means. Eight taxa were detected by the molecular method but their presence would not be expected because of their autecology (i.e. marine, oligotrophic waters, biogeography) and six taxa were detected by the molecular method but their identification is questionable because reference vouchers are missing in the reference databases (for selected taxa see Tables 2 & 3).

The comparison of quantitative identification by LM and molecular methods is shown in Table 4. LM-based and molecular retrieved abundances differ significantly. The most common five taxa identified by LM add up to 75.5 %, whereas corresponding sequences of these taxa were only retrieved for 4.5 % in the combined molecular analysis.

Table 4. Comparison of Quantitative Data of Taxa Identified by Morphology (LM) versus DNA sequences (MOL) (occurrences >1%).

Taxon	LM %	Taxon	MOL %
Fragilaria vaucheriae (Kütz.) J.B. Petersen	50.0	*Ulnaria ulna* (Nitzsch) Compère	24.6
Staurosira construens Ehrenb.	8.5	*Navicula veneta* Ehrenb.	9.1
Amphora pediculus (Kütz.) Grunow	7.6	*Ulnaria acus* (Kütz.) Aboal	8.8
Karayevia clevei (Grunow) Bukhtiyarova	6.0	*Thalassiosira* sp.	7.9
Cyclotella comta Kütz.	3.4	*Diatoma tenuis* C.Agardh	4.5
Achnanthidium minutissimum (Kütz.) Czarn.	2.0	*Fragilaria vaucheriae* (Kütz.) J.B.Petersen	4.2
Navicula reichardtiana Lange-B.	1.9	*Fragilaria capucina* var. *mesolepta* (Rabenh.) Rabenh.	3.9
Planothidium rostratum (Østrup) Lange-Bert.	1.9	*Cymbella proxima* Reimer	3.1
Melosira varians C. Agardh	1.6	*Stauroneis phoenicenteron* (Nitzsch) Ehrenb.	3.1
Achnanthes conspicua Ant.Mayer	1.2	*Stauroneis kriegeri* R.M.Patrick	3.0
Cocconeis neothumensis Krammer	1.2	*Fragilaria nanana* Lange-Bert.	2.0
Rhoicosphenia abbreviata (C.Agardh) Lange-Bert.	1.2	*Fragilariforma virescens* (Ralfs) D.M.Williams & Round	2.0
Ulnaria ulna (Nitzsch) Compère	1.2	*Gyrosigma limosum* Sterrenburg & G.J.C. Underw.	2.0
Cyclostephanos dubius (Hust.) Round	1.1	*Navicula* sp.	1.9
Diatoma vulgaris Bory	1.1	*Nitzschia linearis* (C.Agardh) W.Sm.	1.9
Nitzschia linearis (C.Agardh) W.Sm.	1.0	*Cymbella* sp.	1.7
Ulnaria acus (Kütz.) Aboal	0.9	*Melosira varians* C.Agardh	1.6
Nitzschia dissipata (Kütz.) Grunow	0.7	*Nitzschia dissipata* (Kütz.) Grunow	1.2
Nitzschia fonticola Grunow	0.5	*Encyonema* cf. *cespitosum* Kütz.	1.1
Cymbella cistula (Ehrenb.) Kirchner	0.5	*Bacillaria paxillifera* (O.F.Müll.) Hendey	1.0
Fragilaria capucina Desmazières	0.5	*Nitzschia acicularis* (Kütz.) W.Sm.	1.0

Abundances among molecular methods (not presented here) show closer similarities such as the five most frequent taxa from the Dynabeads EXTRACTION methods (56.3 % of all database hits) were also common by using the QIAGEN technique (32.0 % of all hits). But even here, strong comparability limitations need to be taken into account as is indicated by the most abundant *Navicula veneta* Kütz. (QIAGEN extract) being 27th on the Dynabeads list or *Stauroneis kriegeri* R.M.Patrick not detected by QIAGEN sequences coming out third on the Dynabeads list.

DISCUSSION

To assess diatom diversity, the molecular identification method - even without prior optimization - seems to work comparatively well, since a similar number of hits were detected by the modern molecular as well as the classical morphological method. Exact matches of morphological and molecular identification were discovered for one third of the sample (Table 1). Therefore, it seems to be promising to optimize DNA barcoding for diatom identification.

Further diatom specific optimization needs to be carried out on two different levels, the refinement of the laboratory protocols and the reference database. Concerning the laboratory optimization, it seems to be too early to propose one DNA extraction method to be superior above the other for identification purposes. The advantage of the Dynabead DNA extraction kit is its non-destructiveness of the frustules for later verification.

Even though the 18S rDNA region was not explicitly selected for DNA barcoding purposes it worked quite well for our analysis; however, a shorter stretch of this gene region might be sufficient for diatom identification. Nevertheless, primer screenings have to be carried out to analyze if there are more suitable regions for DNA barcoding in diatoms.

Preventing PCR artefacts was attempted by pooling separate amplifications, reducing the number of PCR cycles and expanding the extension time. Misleading molecular taxa identification as result from chimeric sequences and other PCR artefacts might in this way have been successfully avoided. This is indicated by high similarity parameters if sequences where tested against the databases; furthermore, for several taxa within the reference database multiple matches could be discovered, especially if the threshold was set to 92 %. Focusing on the two taxa with low sequence similarity and only one database match, *Navicula recens* and *Craticula cuspidata*, none of these species were detected by LM even though both species had proven records in Lake Tegel in past taxonomic surveys (Geissler & Kies 2003).

This pilot study elucidates the potential of DNA barcoding for biodiversity assessment but it also demonstrates the need to expand the reference database to include all genotypes of occurring taxa (with picture of its morphology) of the biogeographical region to be investigated. Approximately half of the taxa were found by LM only, because reference sequences were missing. Six taxa of the closely related diatom genera *Cyclotella, Cyclostephanos, Stephanodiscus* were identified by LM; while only 2 different species of these genera could be detected by molecular methods as sequences for comparison were not available (Table 3). This indicates that many of the necessary sequences for identification are not yet available in reference databases such as GenBank and AlgaTerra. Most of the sequences are only available in AlgaTerra because this database focuses on benthic taxa from fresh waters but this data is yet unpublished.

On the other hand, another half of the taxa were detected by molecular means but not by LM. This could be due to diatoms present in the subset of the molecular but not the LM study and/or to cryptic species not identified or identifyable by LM, such as might be hidden behind the high abundances of *Fragilaria vaucheriae* (Kütz.) J.B. Petersen. By comparison of this data on the genus level, though, it seems to be balanced: taxa whose sequences are missing are compensated by taxa of the same genus that have sequences available and are detected (see Table 2 for *Navicula* taxa) or even by a taxon of the same genus that would not be expected to occur in a German Lake (i.e. *Encyonema triangulum* (Ehrenb.) Kütz.). However, also three *Thalassiosira* taxa were detected via molecular methods, none of them were seen in LM, although one is known to occur at this site (see Table 3); the other two are marine species whose identification is unfortunately not backed by vouchers in GenBank.

Problems to address the quantification of individual taxa of environmental samples seem to be overwhelming at this point. Comparisons of the number of occurrences of LM and molecular identification methods indicate that measuring quantitative composition based on DNA sequences will lead to wrong conclusions and indicate that several severe molecular issues have to be mastered. Potentially, the DNA content and/or the mean cell volume per species should be investigated providing potential indications for quantification. Second, cell recovery from samples as well as cell lyses techniques might have to be optimised to fully extract DNA from different types of diatom frustules. Third, it has to be ensured that the likelihood of PCR primer binding to the correct site is the same for all species. The application of tailed primers might have to be tested. And forth, not only PCR but also cloning into *E. coli* has proportional influences upon the analysis. It is now being tested if the number of clones required to completely cover the diatom diversity of a fresh water sample can be determined.

For now, the traditional LM method for identification of diatoms in a mixed sample is still faster and more reliable for a trained diatomologist.

ACKNOWLEDGEMENTS

The authors are grateful to Michael Kube (MPI) for laboratory assistance and the Forschungskomission of the Freie Universität Berlin for financial support.

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. 1990: Basic local alignment search tool. – Journal of Molecular Biology **215:** 403-410.

Brinkmann, N., Behnke, A., Bruns, S., Mohr, K., Jahn, R. & Friedl, T. 2007: Assessing the diversity of pennate benthic diatoms in calcifying biofilms of hard water creeks. – Pp. 11-14 in: Kusber, W.-H. & Jahn, R. (ed.): Proceedings of the 1st Central-European Diatom Meeting 2007. – Berlin. [CrossRef]

Boström, K. H., Simu, K., Hagström, A. & Riemann, L. 2004: Optimization of DNA extraction for quantitative marine bacterioplankton community analysis. – Limnology and Oceanography: Methods **2:** 356-373.

Evans, K. M., Wortley, A. H. & Mann, D.G. 2007: An Assessment of Potential Diatom "Barcode" Genes (cox1, rbcL, 18S and ITS rDNA) and their Effectiveness in Determining Relationships in *Sellaphora* (Bacillariophyta). – Protist **158:** 349-364. [CrossRef]

Geissler, U. & Kies, L. 2003: Artendiversität und Veränderungen in der Algenflora zweier städtischer Ballungsgebiete Deutschlands: Berlin und Hamburg. – Nova Hedwigia, Beiheft **126:** 1-777.

Geissler, U., Kusber, W.-H. & Jahn, R. 2006: The diatom flora of Berlin (Germany): A spotlight on some documented taxa as a case study on historical biodiversity. – Pp. 91-105 in: Witkowski, A. (ed.): Proceedings of the Eighteenth International Diatom Symposium, Międzyzdroje, Poland. 2nd -7th September 2004. – Bristol.

Jahn, R. & Kusber, W.-H. (ed.) 2007: AlgaTerra Information System [online]. – Botanic Garden and Botanical Museum. Berlin-Dahlem, Freie Universität Berlin. [cited 2007-09-30]. Available from <http://www.algaterra.org>.

Kaczmarska, I., Reid, C. & Moniz, M. 2007: Diatom taxonomy: morphology, molecules and barcodes … – Pp. 69-72 in: Kusber, W.-H. & Jahn, R. (ed.): Proceedings of the 1st Central-European Diatom Meeting 2007. – Berlin. [CrossRef]

O'Brien, H., Parrent, J. L., Jackson, J. A., Moncalvo, J. M., & Vilgalys, R. 2005: Fungal community analysis by large-scale sequencing of environmental samples. – Applied and Environmental Microbiology **71:** 5 544-5550. [CrossRef]

Qui, X., Wu, L., Huang, H., McDonel, P. E., Palumbo, A. V., Tiedje, J. M. & Zhou, J. 2001: Evaluation of PCR-generated chimeras, mutations, and heteroduplexes with 16S rRNA Gene-based Cloning. – Applied and Environmental Microbiology **67(2):** 880-887. [CrossRef]

Wang, G. C. & Wang, Y. 1996: The frequency of chimeric molecules as a consequence of PCR co- amplification of 16S rRNA genes from different bacterial species. – Microbiology **142:** 4522-4529.