Check for updates

SOFTWARE TOOL ARTICLE

ProbeSpec: batch specificity testing and visualization of oligonucleotide probe sets implemented in ARB [version 1; peer review: 2 approved with reservations]

Tim Kahlke^{1,2}, Paavo Jumppanen², Ralf Westram³, Guy C.G. Abell², Levente Bodrossy²

¹Climate Change Cluster, University of Technology Sydney, Broadway, NSW, 2007, Australia

²CSIRO Oceans and Atmosphere, Battery Point, TAS, 2004, Australia

³Max Planck Institute for Marine Microbiology, Bremen, Germany

V1 First published: 06 Dec 2018, 7:1901
<https://doi.org/10.12688/f1000research.16905.1>

Latest published: 06 Dec 2018, 7:1901
<https://doi.org/10.12688/f1000research.16905.1>

Abstract

High-throughput molecular methods such as quantitative polymerase chain reaction (qPCR) and environmental microarrays are cost-effective methods for semi-quantitative assessment of bacterial community structure and the identification of specific target organisms. Both techniques rely on short nucleotide sequences, so-called oligonucleotide probes, which require high specificity to the organisms in question to avoid cross-hybridization with non-target taxa. However, designing oligonucleotide probes for novel taxa or marker genes that show sufficient phylogenetic sensitivity and specificity is often time- and labor-intensive, as each probe has to be *in-silico* tested for its specificity and sensitivity. Here we present ProbeSpec, to our knowledge the first batch sensitivity and specificity estimation and visualization tool for oligonucleotide probes integrated into the widely used ARB software. Using ProbeSpec's interactive "mismatch threshold" and "clade marked threshold" we were able to reduce the development time of highly specific probes for a recently published environmental oligonucleotide microarray from several months to one week.

Keywords

Probe design, qPCR, microarray, Bioinformatics, molecular ecology, microbiology

This article is included in the **Max Planck Society** collection.

Open Peer Review

Approval Status ? ?

	1	2
version 1	?	?
06 Dec 2018	view	view

1. **Michael Dondrup**, University of Bergen, Bergen, Norway

2. **Jizhong Zhou**, University of Oklahoma, Norman, USA

Naijia Xiao, University of Oklahoma, Norman, USA

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Tim Kahlke (tim.kahlke@uts.edu.au)

Author roles: **Kahlke T:** Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Jumppanen P:** Software, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Westram R:** Conceptualization, Software, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Abell GCG:** Conceptualization; **Bodrossy L:** Conceptualization, Funding Acquisition, Supervision, Validation, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by the Environmental Genomics grant from CSIRO Oceans & Atmosphere (R-02412).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2018 Kahlke T *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Kahlke T, Jumppanen P, Westram R *et al.* **ProbeSpec: batch specificity testing and visualization of oligonucleotide probe sets implemented in ARB [version 1; peer review: 2 approved with reservations]** F1000Research 2018, 7:1901 <https://doi.org/10.12688/f1000research.16905.1>

First published: 06 Dec 2018, 7:1901 <https://doi.org/10.12688/f1000research.16905.1>

Introduction

The analysis of the microbial community structure and abundance based on universal conserved marker genes has become a powerful tool for many disciplines in life science with a specific focus on next-generation sequencing technologies^{1,2}. In addition to these qualitative methods technologies such as environmental microarrays and quantitative polymerase chain reaction (qPCR) offer cost-effective and highly reproducible techniques for semi-quantitative estimation of microbial communities. Genetic markers commonly used for microarrays and qPCR are ribosomal RNA (rRNA) genes, e.g. 16S, for bacterial communities^{3,4}, as well as functional genes that determine microbial community structure with regards to specific metabolic functions^{5,6}. Both technologies rely on taxon-specific short nucleotide sequences of the marker gene of interest, so-called oligonucleotide probes (OPs). In qPCR experiments OPs act as the primer to initiate the amplification reaction whereas in microarrays the probe is spotted onto a glass slide and the complementary sequence is hybridized with it.

A major challenge in using both techniques for novel organisms and marker genes, however, is the development of OPs with appropriate levels of taxonomic specificity and sensitivity: especially functional genes show highly variable levels of conservation, not only between sequences of different taxa but also between sequences of closely related organisms. Thus, depending on the experiment, the functional marker and the organisms of interest, hundreds or even thousands of OPs with varying levels of conservation have to be designed and subsequently *in-silico* tested for their phylogenetic specificity and sensitivity. A major bottleneck for this process is the lack of software tools that enable researchers to test multiple potential OPs for their phylogenetic specificity at once.

Here we present ProbeSpec⁷, a user-friendly, interactive probe specificity and sensitivity assessment tool for OPs with batch analysis support. ProbeSpec's functionality is incorporated into the widely used ARB software⁸ which is freely available for non-commercial use (detailed copyright information can be found [here](#) and in the license agreement included in each tarball). To our knowledge, ProbeSpec is the only batch probe specificity assessment tool which provides interactive manipulation of specificity and sensitivity thresholds.

Methods

Class structure

ProbeSpec is implemented in ARB's PROBE_DESIGN class utilizing its prefix tree database server. ProbeSpec's functionality is implemented in the classes *ArbProbe* and *ArbProbeCollection* (abstraction of OP sequences and import/export functionality), *ArbProbeMatchWeighting* (providing weighting matrices for position specific nucleotide substitutions), *ArbMatchResult*, *ArbMatchResultSets* and *ArbMatchResultsManager* (abstraction of OP to PT-Server sequences with given weighting matrices and maximum number of mismatches) and *ArbStringCache* (providing string to disk caching of match string results).

Probe specificity matching

Probe specificity calculations in ProbeSpec are based on the initial mismatch penalties given by a 4×4 substitution matrix for all possible nucleotide substitutions. Additionally, each mismatch penalty is weighted based on the position of a mismatch in the probe: mismatches at the ends of an OP are less likely to affect the binding of complementary sequences than mismatches in the center of a probe. Positional weights are calculated as follows: for a mismatch at position p in a given OP sequence of length l a weight W is calculated with

$$W = e^{Sp^2} \quad (1)$$

where

$$S = \frac{-\ln(10)}{w} \quad (2)$$

and

$$P = \frac{2 * p - l}{l} - b \quad (3)$$

The weight distribution given by (1) follows a bell curve penalizing mismatches at either end of the OP sequence less than mismatches in the center of the sequence. The user defined parameter w in equation (2) controls the spread of the weight distribution; user defined parameter b in equation (3) controls the midpoint and therefore enables the user to increase positional weights on either side of the OP sequence. For default parameters of $w=1$ and $b=0$, positional weights range from a minimum 0.1 for mismatches at the first and last nucleotide in the sequence to a maximum of 1 for mismatches at the center.

Operation

For user interaction with ProbeSpec ARB's general user interface was extended with four new dialog windows: (i) a *Probe Collection* dialog, (ii) a *Probe match with specificity*, (iii) *Match display control* dialog and (iv) a *Tree Marker settings* dialog (Figure 1).

The *Probe match with specificity* is the main entry point of ProbeSpec. It displays all loaded probes which can be edited, imported and exported through the *Probe Collection* dialog. Additionally, the *Probe Collection Dialog* allows the user to change the default settings for substitution penalties and positional weight parameters.

The main GUI of ARB was extended to graphically represent the probe matching results: each probe is represented by a colored vertical bar indicating a match of the OP to the specific phylogenetic group. Incomplete cover of a phylogenetic group is represented by transparency of a bar: the fewer members of a group that are covered by a given probe the higher the transparency of a bar is.

The dialogs *Match Display Control* and *Tree marking settings* enable interactive adjustment of probe match parameters such as

mismatch threshold, group marked and group partially marked threshold.

ARB and the included ProbeSpec functionality can be run on any common PC, laptop or workstation. However, we recommend system specifications of at least 4GB of RAM and a dual-core processor to run ProbeSpec.

Use case

Using ProbeSpec we were able to test the specificity of 345 OP sequences against an ARB database of 20,314 bacterial and archaeal ammonia mono-oxygenase sequences on a Ubuntu Virtual Machine with 4 GB of RAM and one processor allocated in less than 30 minutes. In comparison: sequential specificity testing without ProbeSpec for a recent publication⁹ on the same data set took several days

Initial ARB set-up

For any probe development, ProbeSpec requires a phylogeny of target sequences and organisms that the OPs should match to as well as a list of potential OPs.

For an introduction to sequence analysis using ARB, please refer to the main ARB documentation at <http://www.arb-home.de/documentation.html>. For evaluation purposes a sub-set of the data published in Krausfeldt *et al.* (2017) can be found on [Zenodo](https://zenodo.org/record/1000000)¹⁰. To set up ARB select the provided *nitrifiers_2017_04_for_paper.arb* database file on start of ARB. To be able to run ProbeSpec a *PTServer* has to be created from the database via the *Probes* tab and the *PT_Server Admin* option in the *PT Server Admin* widget. Select the loaded database and click *Build server*. After completion close the progress bar and the *PT Server Admin* widget.

Create a probe collection

Before running a batch specificity test, a probe collection, i.e., a list of probes to be tested, has to be created using the *Probe Collection* window where probes can be added to and removed from a collection: Open the *Probe Match with Specificity* window via the *Probes*-tab in ARB (Figure 1A) and select *Edit* (Figure 1B) to open the *Probe Collection* window (Figure 1C). To open the provided test data set use the *load* button and select the provided *amoA70mers.xpc* probe collection. Additionally, the sequence of new probes can be entered into

Figure 1. ProbeSpec GUI: configuration dialogs and visualisation of matching probes in ARB's main window. (A) Probe Collection dialog. **(B)** Tree marking settings dialog. **(C)** Match Display Control; **(D)** Match Display Control. Coloured vertical bars on the left of the main window represent (partially) matching probes.

the *Target String* text field. To add new OPs to the collection press *Add*. Probe collections can also be in this dialog.

Probe specificity configuration

The *Probe Collection* window can be used to define the specificity measures used by ProbeSpec to identify matching probes. This includes the definition of specific mismatch penalty values as well as the values for bias b and weight w (see subsection Operations in the Methods section for details).

Match probes

After creation of a probe collection and configuration of the match parameters the *Probe Collection* window can be closed and the specificity search can be started by clicking the *Match* button (Figure 1B). A status dialog will appear and show the progress of the search.

Result visualization

The final match results are shown in the ARBs main window: each matching probe is represented by a coloured bar next to the group/clade the probe matches with the given thresholds (Figure 1A). The visualization can be configured using the two dialogs *Match Display Control* (Figure 1B) and *Tree Display settings* (Figure 1C), the latter of which can be accessed via the *Marker Display Settings* button on the *Match Display Control* widget.

Conclusion

Here we present ProbeSpec, to our knowledge, the first tool for batch specificity testing of OP sequences implemented in ARB.

ProbeSpec offers significant time saving for projects developing and testing large oligonucleotide probe datasets for use in technologies such as qPCR and environmental microarrays.

Data availability

For test and validation purposes, a sub-set of the data published in Krausfeldt *et al.* (2017) can be found at Zenodo, DOI: <http://doi.org/10.5281/zenodo.1482958>¹⁰. The dataset includes a phylogeny of archaeal and bacterial *amoA* sequences (*nitrifiers_2017_04_for_paper.arb*) as well as a sub-set of 185 OPs used to create the environmental microarray.

Software availability

ProbeSpec is included in the production version of ARB, available at: <http://download.arb-home.de/special/manual-builds/>.

Archived version of the production version directory: <http://doi.org/10.5281/zenodo.1482947>.

License: ARB License.

Grant information

This work was supported by the Environmental Genomics grant from CSIRO Oceans & Atmosphere (R-02412).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Brown MV, van de Kamp J, Ostrowski M, *et al.*: **Systematic, continental scale temporal monitoring of marine pelagic microbiota by the Australian Marine Microbial Biodiversity Initiative.** *Sci Data*. 2018; 5: 180130.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
2. Bork P, Bowler C, de Vargas C, *et al.*: **Tara Oceans. Tara Oceans studies plankton at planetary scale. Introduction.** *Science*. 2015; 348(6237): 873.
[PubMed Abstract](#) | [Publisher Full Text](#)
3. Lazarevic V, Gaia N, Girard M, *et al.*: **Decontamination of 16S rRNA gene amplicon sequence datasets based on bacterial load assessment by qPCR.** *BMC Microbiol*. 2016; 16: 73.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
4. Figueroa IA, Barnum TP, Somasekhar PY, *et al.*: **Metagenomics-guided analysis of microbial chemolithoautotrophic phosphite oxidation yields evidence of a seventh natural CO₂ fixation pathway.** *Proc Natl Acad Sci U S A*. 2018; 115(1): E92–E101.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
5. Abell GC, Robert SS, Frampton DM, *et al.*: **High-throughput analysis of ammonia oxidiser community composition via a novel, *amoA*-based functional gene array.** *PLoS One*. 2012; 7(12): e51542.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
6. Lee YJ, van Nostrand JD, Tu Q, *et al.*: **The PathoChip, a functional gene array for assessing pathogenic properties of diverse microbial communities.** *ISME J*. 2013; 7(10): 1974–1984.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
7. Kahlke T, Jumppanen P, Westram R, *et al.*: **ARB tarballs 19.10.2018 (Version r17491).** *Zenodo*. 2018.
<http://www.doi.org/10.5281/zenodo.1482947>
8. Ludwig W, Strunk O, Westram R, *et al.*: **ARB: a software environment for sequence data.** *Nucleic Acids Res*. 2004; 32(4): 1363–1371.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
9. Krausfeldt LE, Tang X, van de Kamp J, *et al.*: **Spatial and temporal variability in the nitrogen cyclers of hypereutrophic Lake Taihu.** *FEMS Microbiol Ecol*. 2017; 93(4): fix024.
[PubMed Abstract](#) | [Publisher Full Text](#)
10. Kahlke T, Jumppanen P, Westram R, *et al.*: **ProbeSpec validation data [Data set].** *Zenodo*. 2018.
<http://www.doi.org/10.5281/zenodo.1482958>

Open Peer Review

Current Peer Review Status: ? ?

Version 1

Reviewer Report 14 May 2019

<https://doi.org/10.5256/f1000research.18484.r46526>

© 2019 Zhou J et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Jizhong Zhou

Department of Microbiology and Plant Biology, University of Oklahoma, Norman, OK, 73072, USA

Naijia Xiao

University of Oklahoma, Norman, OK, USA

The authors present a novel tool, ProbeSpec, to test specificity of probes against sequence databases. The tool is presented with enough detail and deserves indexing if the authors successfully address the following questions:

- The authors assert that ProbeSpec's performance is superior than other similar tools in the "Use case" section. Only one recent publication was mentioned.
 1. Are there any additional publications recently with similar settings?
 2. Can the author provided more specific instructions for the reviewer to find where Krausfeldt mentioned their testing "took several days"?
 3. What is the most important improvement ProbeSpec has done to outperform other tools? Can the authors provide a breakdown of the running time of a typical ProbeSpec run to prove it?
- In "Probe specificity matching", the authors proposed a new scheme for positional weights.
 1. The formula is very similar to a formula on page 35 of the supplement material from "Spatial and temporal variability in the nitrogen cyclers of hypereutrophic Lake Taihu¹" (can be found [here](#)). Why use $w=1$ as default, which is different from $w=3$ in the above reference.
 2. Eq. (3). It is strange to use $b!=0$, in which case the weighting function will not be symmetric. What is the rationale to assign different weight to the starting and ending bp? If not necessary, what is the point to introduce parameter b ?
 3. It is a little confusing to use both p and P , while there are plenty of other letters available.

Reference:

1. Krausfeldt LE, Tang X, van de Kamp J, et al.: Spatial and temporal variability in the nitrogen cyclers of hypereutrophic Lake Taihu.¹

References

1. Krausfeldt L, Tang X, van de Kamp J, Gao G, et al.: Spatial and temporal variability in the nitrogen cyclers of hypereutrophic Lake Taihu. *FEMS Microbiology Ecology*. 2017; **93** (4). [Publisher Full Text](#)

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Reviewer Report 22 January 2019

<https://doi.org/10.5256/f1000research.18484.r41599>

© 2019 Dondrup M. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Michael Dondrup

Department of Informatics, Sea Lice Research Centre, University of Bergen, Bergen, Norway

In this manuscript, the authors present a software for testing the specificity of oligonucleotide probe-sets that is integrated into the sequence analysis software ARB with a focus on microbial ecology. The software allows for batch-scoring the specificity of probe-sets using a bell-shaped scoring function based on the distance from the center of the probe, with several user-definable parameters.

The application domain of the new functionality is timely because high-throughput methods that

rely on oligonucleotide probes are becoming more and more common. The article is overall well written, but lacking with respect to the description of the methodology, state of the art, and depiction of software development and code.

For the sake of a comprehensive evaluation, I have tested the ARB software using the latest development build 6.1.rev17491 and the author-provided dataset in a virtual machine under Linux Mint 9.

Major Concerns:

1. The authors present a new scoring function that introduces bell-shaped weights of mismatches depending on the distance to center base of the probe. It is unclear what the motivation for using this function is, and what underlying assumptions are that it is based on, or if has been used in the literature before. I assume it is trying to compensate for some hybridization effects, but these might also vary between technologies. Other publications have discussed thermodynamic parameters instead, see e.g. Kabilov et al.¹. It should also be made more clear, if and how the user can specify parameters, adjust them to the assay or how to use a flat scoring function instead.
2. The authors claim that their software is the first tool for batch testing, which might be true. However, with respect to the new scoring functions, authors should still compare their estimates to other existing tools. Searching for other tools I found probeCheck², which is an online-tool also using ARB as a back-end, but allows display of max 10 results at a time, and does not allow to use custom databases, but the outcomes could be comparable for the same probe set.
3. As ProbeSpec is part of large software package and is only obtainable as a component it is hard to evaluate the contribution by source code. It should be clearly pointed out which files of the ARB distribution contain the source code of ProbeSpec or possibly provide a separate (e.g. git) repository that contains the sources.
4. The functional test completes properly, but I noticed that the probe set is loaded and match parameters could be set under Edit which is a bit confusing (substitution matrix, shape parameters). It should be better explained in the UI that these are matching parameters.

Minor concerns:

1. $S = -\ln(10)$ in Equation (2), should it be $\ln(10)$ for logarithm (L vs. I)
2. Even though it is an aspect of the software which cannot be changed easily, I would like to question the use of ARB as the software platform. The appearance of the user interface is archaic, while this might be dismissed as cosmetics, worse so the build system is arcane too. While I was finally able to compile the latest ARB revision including ProbeSpec from source under Linux Mint, this will be prohibitive for most users, mostly due to undocumented/not up to data dependencies (e.g. boost library never mentioned anywhere) including combined with the lack of a (e.g. autoconf generated) configure script.
3. Therefore, for most users the options are restricted running pre-compiled binaries to either Debian, Linux Mint or a very old Ubuntu (10) release under a virtual machine, something not very timely in the age of Docker and containers. This is a pity because common package managers contain the latest stable ARB release 6.0.6, dating back to August 2016, but not containing ProbeCheck. I would therefore recommend to the authors to: contact the maintainers of ARB to get a new stable release out supporting popular OS', contribute to a

complete and updated dependency list, or provide a containerized version of the whole system.

References

1. Kabilov M, Pyshnyi D: Analytical consideration of the selectivity of oligonucleotide hybridization. *Journal of Biophysical Chemistry*. 2011; **02** (02): 75-91 [Publisher Full Text](#)
2. Loy A, Arnold R, Tischler P, Rattei T, et al.: probeCheck--a central resource for evaluating oligonucleotide probe coverage and specificity. *Environ Microbiol*. 2008; **10** (10): 2894-8 [PubMed Abstract](#) | [Publisher Full Text](#)

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Partly

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

No

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatic, genomics, computational biology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research