Manual for OEZY

OEZY

*O*ptimising *E*n*ZY*me selection
for best performing
restriction fragment length polymorphism (TRFLP)
analysis using ARB

Martin Mühling[1], Rico Beier[2], Dirk Labudde[2]

Tuesday 11th August, 2015

[1] TU Bergakademie Freiberg, Institute of Biological Sciences
Leipziger Str. 29, 09599 Freiberg, Germany

[2] University of Applied Sciences Mittweida
Technikumplatz 17, 09648 Mittweida, Germany

Contents

1 Installation **3**
- 1.1 Operating system . . . 3
- 1.2 Automatic Installation using standard packages . . . 3
- 1.3 Manual Installation . . . 4
- 1.4 Running OEZY from ARB 6 . . . 6

2 Methodological approach **7**
- 2.1 The principle . . . 7
- 2.2 Perl communication interface . . . 9
- 2.3 Search strategies . . . 9
- 2.4 Scoring approaches . . . 11

3 How to use OEZY **13**
- 3.1 Initialization . . . 13
- 3.2 Defining the reference distance matrix . . . 14
- 3.3 Defining the PCR primer positions . . . 15
- 3.4 Defining the search space . . . 18
- 3.5 Defining details on the resulting fragments . . . 19
- 3.6 Defining scoring parameters . . . 19
- 3.7 Defining the strategy to screen the search space . . . 20
- 3.8 Run . . . 21
- 3.9 Results . . . 22

4 Copyrights **24**
- 4.1 Copyright notice for OEZY . . . 24
- 4.2 Copyright notice for software that is integrated into OEZY: Archeopteryx . . . 25

References **26**

This document was created on: Tuesday 11th August, 2015 at 12:18.

1 Installation

1.1 Operating system

OEZY is written in Java. Although both, ARB and Java, are available under UNIX and MAC operating systems, we chose Mint Linux 16, a Debian based distribution (`http://www.linuxmint.com`) for the development of OEZY. It should therefore be noted that OEZY is not limited to Linux. Using OEZY on other non-Debian based Linux distributions, UNIX or MAC systems only necessitates that the commands provided in this manual are adapted to the alternative operating system.

1.2 Automatic Installation using standard packages

Using the propagated Linux Mint operating system OEZY can be automatically installed including all requirements like ARB (Ludwig et al., 2004; Westram et al., 2011), Java and Perl. After downloading the debian installation package `oezy.deb` from our website `http://www.bioforscher.de/oezy` or `http://tu-freiberg.de/fakultaet2/bio/environmental-microbiology` the installation can either be performed using the commands (`gdebi` or `dpkg`) as shown hereafter or by simply double-clicking the downloaded `.deb` file:

```
(sudo) gdebi oezy.deb
 -- OR --
sudo dpkg -i oezy.deb
sudo apt-get install -fy
```

After installation `oezy` can be started in command line as follows:

```
oezy
```

1.3 Manual Installation

For running OEZY a working Java Runtime Environment needs to be installed in version 7 or later. The communication between OEZY and ARB additionally requires a Perl installation including an extension module. Most Linux distributions will come with proper installed packages by default. If this is not the case Java and Perl can be installed directly using the package manager:

```
sudo apt-get install openjdk-7-jre perl libswitch-perl
```

If libswitch-perl is not available, the library can also be installed using `cpan`. Provided that cpan is correctly configured this approach then automatically results in the recognition and use of the library by perl:

```
cpan
-- INSIDE CPAN --
cpan[1]> install Switch
```

The easiest way is simply installing ARB from the package repository of Linux Mint. In this case the installation path `/usr/lib/arb` has to be kept in mind. It has to be used as ARB installation directory in the subsequent configuration of OEZY:

```
sudo apt-get install arb
```

If the package is not available in your system or a newer version is provided as source code from the authors' homepage `www.arb-home.de`, this can be downloaded and compiled from the source as follows. The first step involves downloading and uncompressing the ARB source code package from `http://download.arb-home.de/release/latest/arbsrc.tgz`.

```
wget -N download.arb-home.de/release/latest/arbsrc.tgz
tar xzvf arbsrc.tgz
cd arbsrc_*
```

In order to compile the source files, a set of system packages is required, which can be installed using the standard Linux packaging system. This process is automated by a prepared library installation script in the ARB `SH` subdirectory:

```
sudo SH/arb_installubuntu4arb.sh compile_noOpenGL
```

A copy of the configuration template `config.makefile.template` with the new name `config.makefile` is used to configure the compilation process. Appropriate options like OpenGL usage and the processor architecture of 32/64-bit have to be defined in this context:

```
cp config.makefile.template config.makefile
sed -i -re "s/OPENGL := 1/OPENGL := 0/g" config.makefile
sed -i -re "s/ARB_64 := 0/ARB_64 := 1/g" config.makefile
```

Since the original makefile may not support the installed version of the GCC compiler, this version number needs to be manually added to the list of supported version numbers:

```
GCC=`SOURCE_TOOLS/arb_gcc_version.pl`
sed -i -re "s/ALLOWED_GCC_4xx_VERSIONS=\\\\/
    ALLOWED_GCC_4xx_VERSIONS=$GCC\\\\/g" Makefile
```

After providing the ARB source directory in two system variables, the compilation process can be triggered:

```
export ARBHOME=`pwd`
export PATH=$ARBHOME/bin:$PATH
make all
```

Once the compilation process is finished ARB can be started by calling its main executable:

```
./bin/arb
```

When updating the system's perl interpreter to a new version, it might be neccessary to recompile ARB from source because the installed version of perl and the version used to compile the ARB library have to match.

The last manual installation step involves downloading and unpacking the OEZY zip package from our website `http://www.bioforscher.de/oezy` or `http://tu-freiberg.de/fakultaet2/bio/environmental-microbiology`. The program can be launched by executing:

```
java -jar OEZY.jar
```

1.4 Running OEZY from ARB 6

The latest ARB Version (Version 6) will support user defined menu items. Starting OEZY for the first time will automatically generate such a menu item, which subsequently permits starting OEZY from within ARB. The OEZY menu item is located in the user menu section, which is currently under Tools > Network, but will be relocated to the Tools menu in the next ARB release (see under `http://bugs.arb-home.de/changeset/13598`).

When running OEZY from within ARB, the current ARB session will be detected by OEZY. The "Run ARB" button is therefore replaced by a "Connect to ARB" button, which establishes the connection to the running ARB session. OEZY settings in respect to the ARB installation directory and the database file path are ignored in this case. Nevertheless, the information on the database field containing the sequence information is still used and therefore needs to be set correctly.

2 Methodological approach

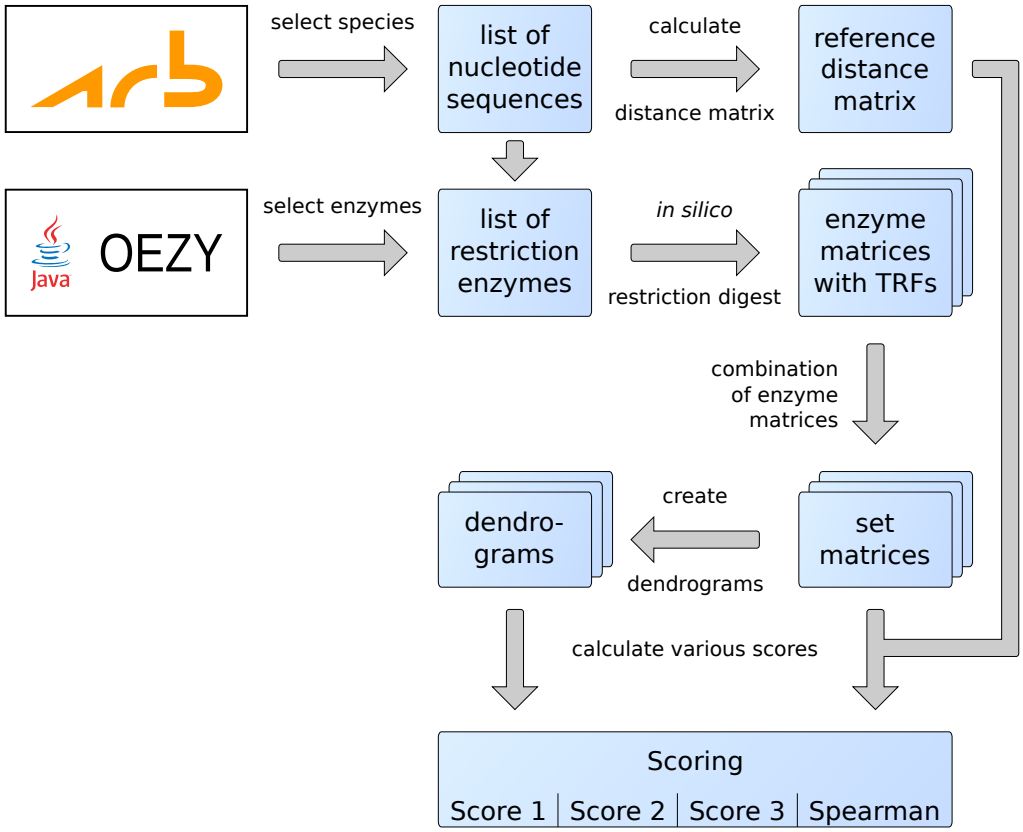
2.1 The principle

OEZY uses the sequences selected within ARB to internally calculate the lengths of the TRFs resulting from the *in silico* restriction analysis with a particular restriction enzyme (RE), either assuming the forward or the reverse primer as being labelled. A particular advantage of OEZY is connected to the option that sequences, which lack sequence information within the fragment that is to be explored for restriction sites, can be included in the *in situ* analysis. Sequence gaps can either be simply due to the missing primer binding site (i.e. the primer sequence was removed from the sequence) or caused by limited sequence information downstream of the primer binding site.

If such a sequence is to be selected for OEZY analyses then the user must also include at least one sequence of a closely related strain (as a minimum: of the same genus) that does not have missing sequence information downstream of the primer binding site since OEZY uses the information of only those sequences for the prediction of the number of nucleotides within the missing sequence parts that derive from strains of the same genus. The identification of the genus name is achieved by comparison of the first word in the database field "full name" within the ARB sequence entries.

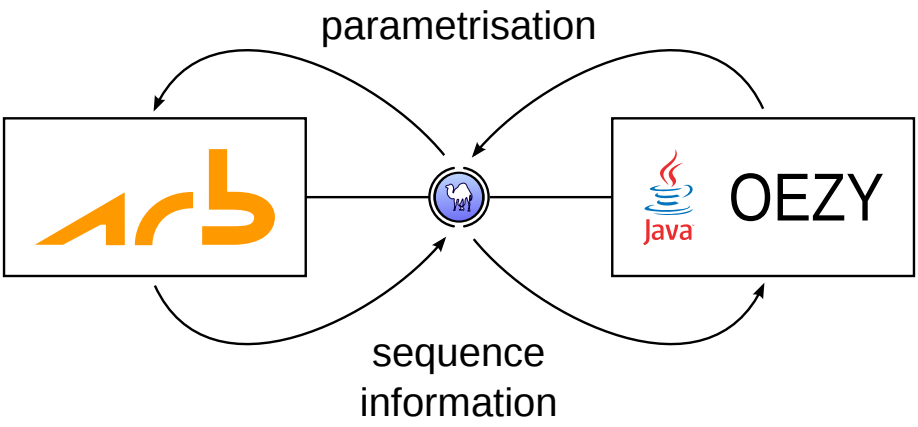
This, however, also means that OEZY will simply provide the value of the terminal fragment of those sequences (with missing sequence information downstream of the "labelled" primer) where such a reference sequence is not included in the analysis. In this context the user has to further bear in mind that the results obtained from OEZY analyses do not include those restriction sites that lie within the missing part of a sequence. Such a scenario also impacts on the calculation of the Spearman rank correlation and of the dendrogram (e.g. such a sequence would be wrongly assigned within the dendrogram).

Subsequently to and based on these *in situ* predictions of the terminal restriction fragments (TRFs) enzyme matrices are generated. An enzyme matrix only contains values 0 and 1 indicating whether two sequences (i.e. strains or species etc.) can (1) or can not (0) be distinguished solely by the size of their terminal fragments. In the next step enzyme matrices of the various REs of a RE set are combined by simple addition to form a so called set matrix. (Please see below – section 2.3 Search strategies – for details on the approach taken to assemble the various RE sets.) Each set matrix is then evaluated against the reference matrix generated earlier from the nucleotide sequences that were selected for restriction analysis (section 2.4 Scoring approaches).



2.2 Perl communication interface

Due to its simplicity and robustness the communication interface that is provided by ARB and accessible through Perl is utilized by OEZY. Primarily designed for accessing ARB functionality from external applications and secondarily used as part of the macro language included in ARB, this interface allows execution of programme commands and access to the database content. In this context, a straightforward Perl script was developed as communication intermediary between Java and ARB. In this context, a Perl script was developed as communication intermediary between Java and ARB. It receives high level commands from the Java application and utilizes the ARB interface to execute available low level commands and transfer back the execution results.



2.3 Search strategies

The first option offers an extensive search, which usually is referred to as "brute force". It is the only option that screens all possible combinations of REs. This method requires most computational time and, therefore, is likely to be only useful in cases where a narrow search space has been defined. If, however, a combination of large RE set size with many REs and sequences (i.e. strains and species) is to be included in the analysis then the two alternative approaches might be preferred by the user to reduce computational time.

BVStep is based on the stepwise approach developed by Clarke & Warwick (1998). In essence, this method sequentially analyses a set of REs whereby starting from a randomly chosen set of REs. Additional REs are added in forward (selection) steps or dropped from in backward (elimination) steps according to whether this addition leads to a higher or lower score (i.e. score 3 or Spearman rank coefficient between the distance matrices based on the resulting TRF pattern and that based on the nucleotide sequences selected for the analysis). If forward or backward steps do not result in improved values for score 3 or the Spearman rank coefficient then the process is terminated. Even though this procedure would not necessarily find a unique set of REs that is optimal to the TRFLP analysis for the given set of nucleotide sequences it still provides a means to screen a fraction of the search space in reasonable time frames for possibly near-optimal solutions. The extent of this shortcoming can be reduced by repeating this process several times (i.e. number of iterations) starting with different randomly selected initial RE sets, thus exploring various regions of the search space (Clarke & Warwick, 1998).

The third methodological approach is based on the BVStep principle but was designed to search a larger fraction of the sequence space with each of the iterations. This is achieved by not only adding or dropping REs from the list according to whether this addition leads to higher or lower scores (score 3, Spearman rank correlation – see above). Instead, the approach taken within BVTolStep saves those results from the corresponding forward and backward steps that are within a tolerance range for subsequent analysis whereby the hierarchy of analysis is based on the best performing enzyme sets. An enzyme set with a score of s_r is tolerated if $s_r > s_m \cdot (1+t) - t$, where s_m is the maximal score achieved in this iteration and t is the given tolerance setting. When choosing an appropriate value for the tolerance limit it must be considered that both Spearman rank coefficient and Score 3 have values between 0 and 1. This, in turn, means that the tolerance must not exceed a reasonable fraction of 1. Furthermore, the user is also asked to define another parameter (search broadness) which is a safeguard against unacceptable computation time due to an unexpected high number of RE sets within the set tolerance limit. This parameter (absolute number) determines the maximal amount of RE sets within the tolerance limit that will be considered for further analysis. Finally, the search space to be analysed by the BVTolStep method can, identical to the BVStep approach, be increased by increasing the number of iterations of the calculations whereby each starts with a randomly chosen RE or set of REs.

Should the search unexpectedly and despite above mentioned precautions exceed an acceptable time period then the user can simply stop the search whereby the best performing enzyme set found up to then will be provided as part of the output.

2.4 Scoring approaches

The primary scoring function used for choosing the optimal enzyme set is Spearman's Rank Correlation applied on a vector representation of the pairwise distance matrix, which is defined as the Pearson correlation coefficient between the ranked variables. The ranking moves the focus from numerical detail towards the order of pairwise distances. The results are further evaluated by three additional scores, which are based on properties of dendrograms created from the distance matrices (that is the reference distance matrix and that based on the performance of the REs within an RE set).

Selecting REs with restriction sites that are absent in a sequence (i.e. strain or species) impacts on the calculation of the Spearman rank coefficient and may also result in the position of the sequence to be wrongly assigned within the dendrogram.

Score 1 is calculated as the ratio of the length of the longest branch clustering identical sequences in each of the two dendrograms to the length of the longest branch in the reference dendrogram. The score has a range between 0 and 1 whereby smaller values indicate a better resolution. The value 1 indicates that the dendrograms differ in the position of the sequence that is most dissimilar to the rest of the sequences used for the analysis, i.e. the outgroup within a dendrogram.

The aim of scores 2 and 3 is the assessment of the level of the overall genetic resolution achieved by a particular set of REs.

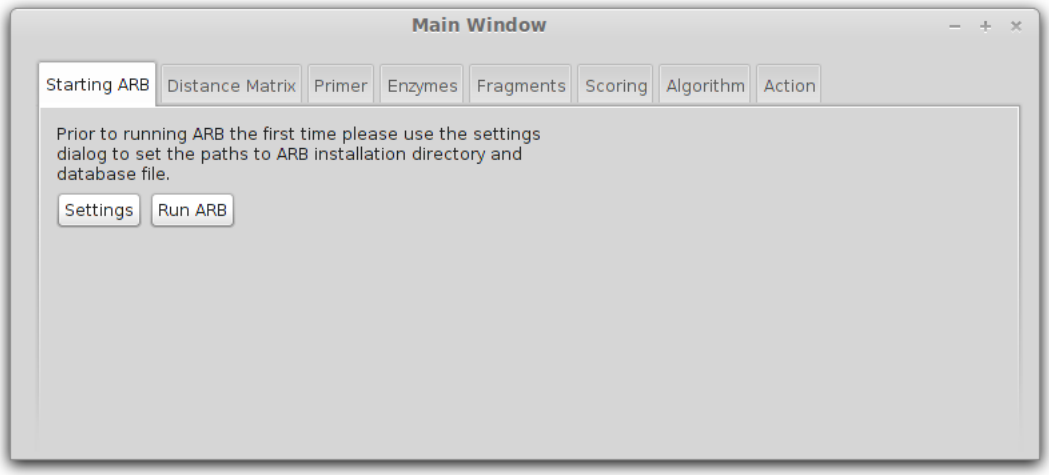
Score 2 is attained by computing the average length of a branch within the dendrogram that is based on the TRF pattern for a specific set of REs. OEZY calculates Score 2 as the ratio of sum of all horizontal branch lengths to the total number of branches. Low values therefore indicate high genetic resolution.

Score 3 also provides a measure of the overall genetic resolution achieved by a specific set of REs and is calculated as the ratio of the number of resolved sequences to the total number of sequences used for the selection of the REs. Score values range from 0 to 1 with 1 indicating that all sequences can be distinguished from each other using the specific set of REs for TRFLP analysis.

3 How to use OEZY

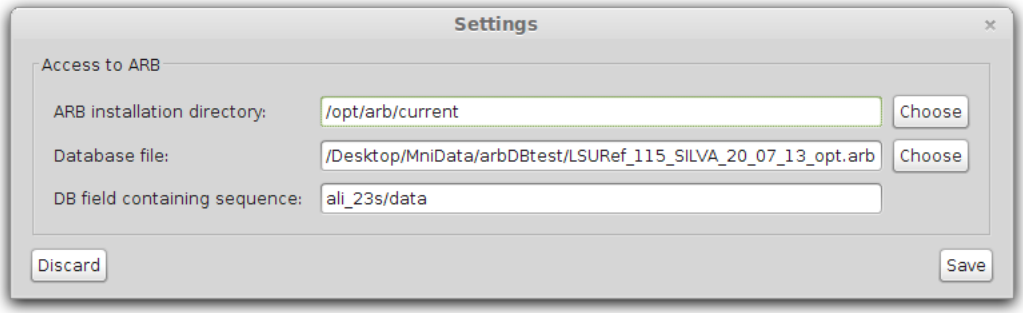
3.1 Initialization

OEZY utilizes ARB to perform terminal restriction fragment (TRF) length predictions, which requires a Perl-based communication channel. Thus, all following tabs are deactivated until ARB has successfully been started and connected to OEZY *via* the *Run ARB* or *Connect to ARB* button.



OEZY requires information on the ARB installation and the location and structure of the sequence database file that will be used to start ARB and to build up the communication channel. This information has to be provided *via* the *Settings* menu. If OEZY is started from within ARB, on the other hand, the ARB installation directory and database file information are inferred from the ARB session that was used to run OEZY rather than from the information provided in the *Settings* menu. Nevertheless, the database field that contains the sequence information is still used and therefore needs to be set correctly.

OEZY attempts to automatically detect the ARB installation directory using the information provided through the system environment variables $PATH and $ARBHOME.

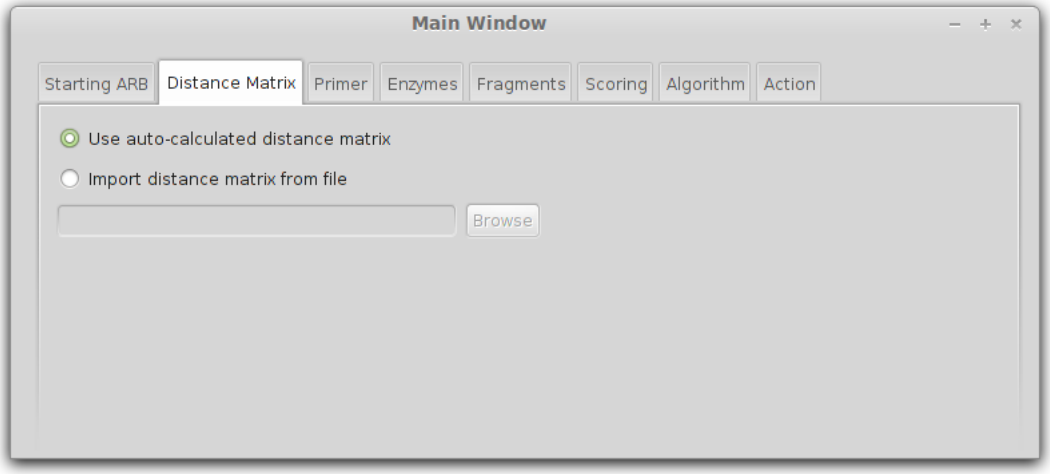


After successfully starting ARB the target sequences for subsequent analyses have to be marked in the ARB database. Please note that the option *Quicksave changes* can be used to store the current selection. Otherwise the selection will not be saved by ARB.

Sequences other than ribosomal RNA sequences can also be added to and aligned within ARB using automated alignment functions. However, the user may prefer to employ other software that runs on more commonly used operating systems (i.e. Windows or Mac). In this context it should be noted that databases of molecular sequences can also be created using any standard alignment software (MEGA, CLUSTAL W). The aligned sequence data can be exported in e.g. FASTA format for subsequent import into ARB. A new database of the aligned sequences can then be produced by simply saving the alignment within ARB.

3.2 Defining the reference distance matrix

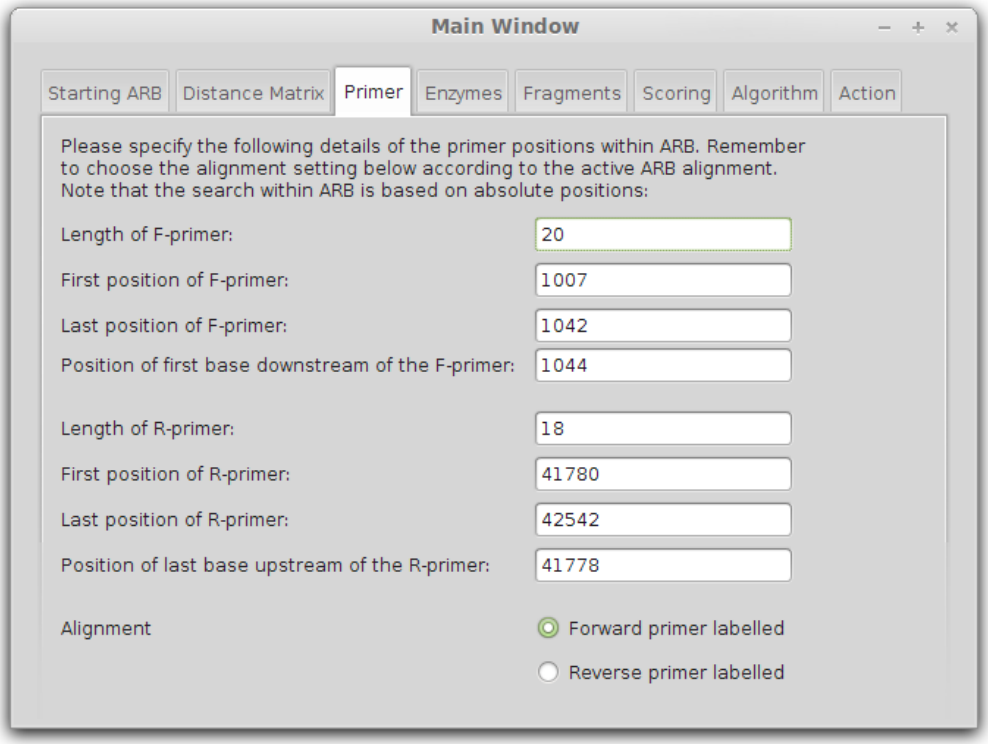
A reference distance matrix is required to evaluate the suitability of the various sets of REs based on their predicted TRFs. The reference matrix is a pairwise distance matrix of all sequences that are selected within ARB.



An automatically calculated distance matrix based on the standard method and using standard parametrization within ARB can be used to provide a fast and simple approach. This option does not require any further action by the user. It is, however, recommended to use the option to import a distance matrix that has been individually calculated from the selected sequences by the user. (The reader is referred to the ARB manual for detailed instructions on how to perform the calculation.)

3.3 Defining the PCR primer positions

The next step involves adding relevant information on the primers, specifically the primer positions and primer lengths. Please note that the search for REs within OEZY is based on absolute positions within the ARB alignment. The user also needs to specify, which of the primers (i.e. forward or reverse) is fluorescently labelled for the TRFLP analysis.



Required parameters:

- First and last positions of primers: place the cursor in ARB directly onto the first and last bases of the primer.

- Length of the primers: number of nucleotides of the primer sequences.

- Position of first base downstream of the forward primer: since absolute positions are used within the ARB alignment, the position of the first base after the forward primer is not necessarily equal to the last of the labeled primer plus 1.

- Position of last base upstream of the reverse primer: since absolute positions are used within the ARB alignment, the position of the last base upstream of the reverse primer is not necessarily equal to the first of the reverse primer minus 1.

- Direction: specifies, which primer is fluorescently labelled for subsequent detection of the TRF

The details on the first base after the forward primer and the last position before the reverse primer is required to fill in the sequence gap due to removal of the primer sequence from the sequences submitted to online databases. These sequences could otherwise not be included in *in silico* analyses and represents a particular strength of OEZY and TRF-CUT. Details on the suitability of 16S rRNA gene targeting primers used for PCR amplification are provided by, for example, Amann et al. (1995) and Mühling et al. (2008).

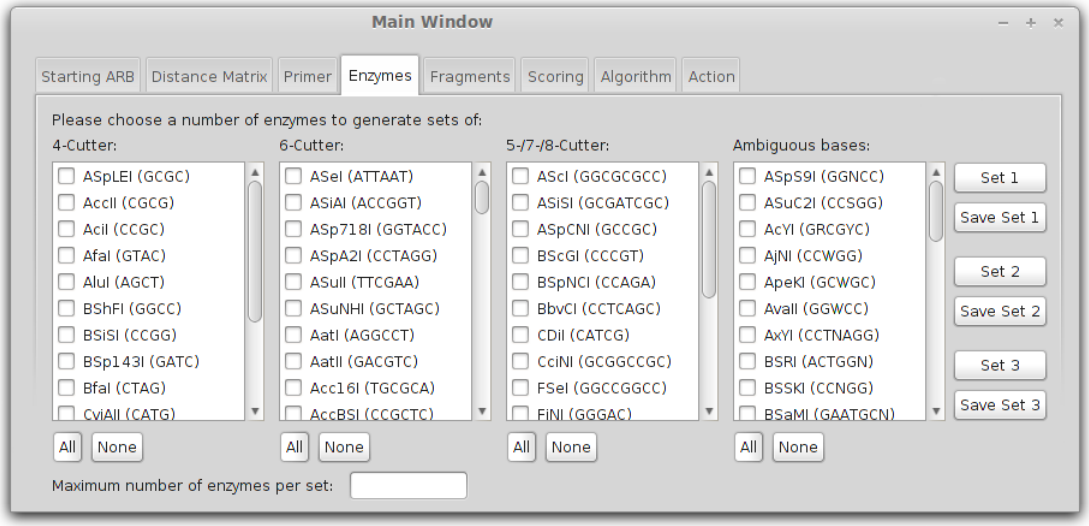
Only the user specified primer binding position is requested by OEZY since in this way all selected sequences, including those with missing sequence information at and downstream of the labelled primer position, can be entered into the analysis. This, however, also means that, firstly, OEZY does not examine whether the primer of interest binds to all marked sequences. If there are doubts concerning primer binding to particular sequences, they can be easily checked by opening the ARB edit window and performing a user specific primer search. Secondly, this caveat also extends to the region downstream of the primer if sequence information is missing there. The approach taken within OEZY is designed to compensate for missing base information when calculating the resulting TRFs. In essence, OEZY identifies those sequences out of the set of sequences selected for the search for REs that belong to the same genus as the sequence lacking sequence information. The information on the presence of bases at positions where no sequence information is available for such a sequence is then included in the calculation for the length of the resulting TRF. However, this approach does not analyse the sequence in terms of base composition. This means that restriction sites within this region or variation in the sequence to that of other sequences of the same genus are not considered by OEZY, but must be manually analysed by the user.

Care must also be taken when "uncultured clone" sequences are considered for the analysis. In this case only sequences without sequence gaps downstream of the primer binding site should be used in the analysis. Alternatively, only such "uncultured clone" sequences that are presumed to derive from the same genus can be included as long as one of those sequences fulfils the above criterion (i.e. OEZY recognises the word "unknown" in those sequences as equivalent to the taxonomic identity).

3.4 Defining the search space

The optimal RE sets are searched within a specific search space that results from the combination of selected REs and nucleotide sequences.

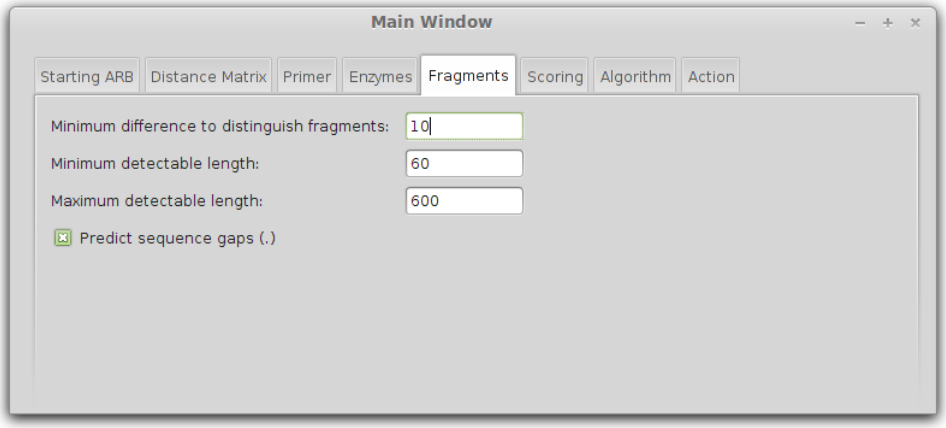
The REs that are to be included in the analysis have to be selected from a catalogue of possible REs, which has been retrieved from the REBASE (`http://rebase.neb.com/rebase/rebase.html`) and, for ease of the selection process, further grouped into 4- and 6-cutters and REs with other recognition sites than those. Isoschizomers have been removed in order to avoid the occurrence of multiple RE sets with the same TRF pattern.



Following the selection of the REs the comma separated list on the right side is instantly updated to give a fast overview on the current selection. Additionally, three preferred enzyme sets can be saved and retrieved for future analyses. The search process will be limited to the maximum number of REs per set that is specified by the user.

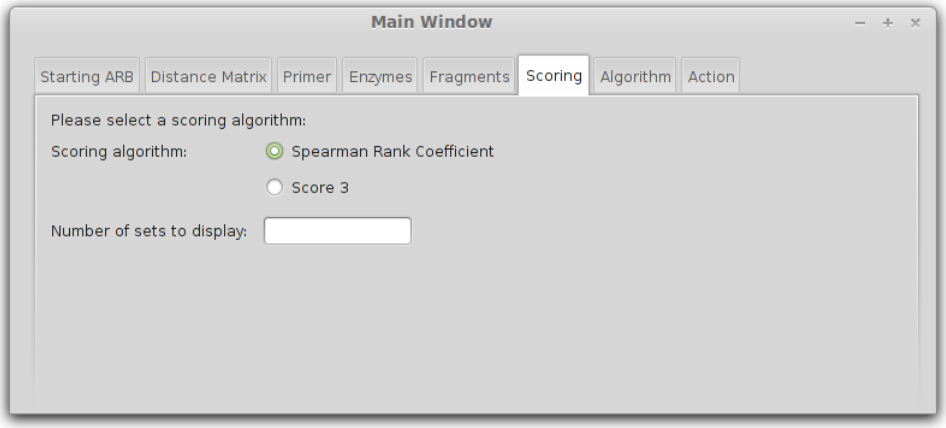
3.5 Defining details on the resulting fragments

The user is requested to define the minimum and maximum fragment sizes that can be detected by the instrumentation available in the user's laboratory and the number of nucleotides that are required for two TRFs to be reliably judged to be different to each other. As sequences may contain unidentified areas (dots) that are ordinarily not treated as bases, a simple prediction based on all selected sequences of the same generic group can improve the calculation results.



3.6 Defining scoring parameters

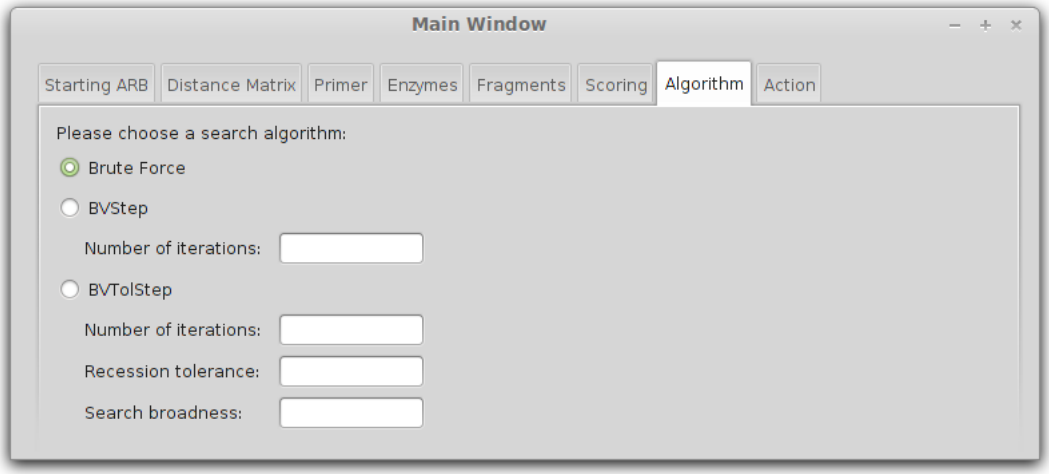
The numerous distance matrices obtained on the basis of predicted fragment length are scored to aid the user in identifying the best performing RE sets. In addition to the scores described above OEZY also uses a statistical approach to evaluate the correlation between the distance matrices based on the nucleotide sequence data (the reference matrix) and the *in silico* determined TRF pattern. The Spearman rank coefficient and Score 3 are currently the only available options within OEZY for this task, though future versions of the programme may include further algorithms. The score resulting from the matrix comparison is then used to identify the optimal enzyme sets.



The user is further requested to specify the number of RE sets to be displayed as part of the output.

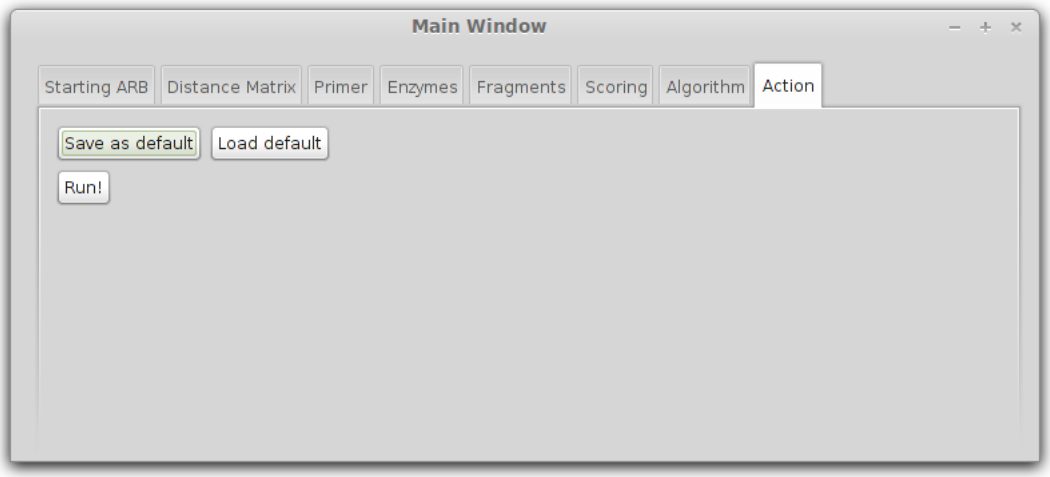
3.7 Defining the strategy to screen the search space

Three options covering both, exact and heuristic approaches, are provided to screen the search space - defined by the selection of REs and nucleotide sequences - for best performing enzyme sets. Please see section 2.3 for more information on search strategies.

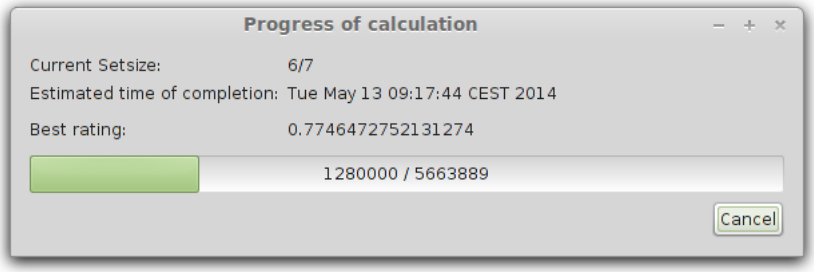


3.8 Run

If settings are to be also used for further analyses then the user may wish to save those as default. Retrieval of those settings is achieved *via* the *Load default* button.



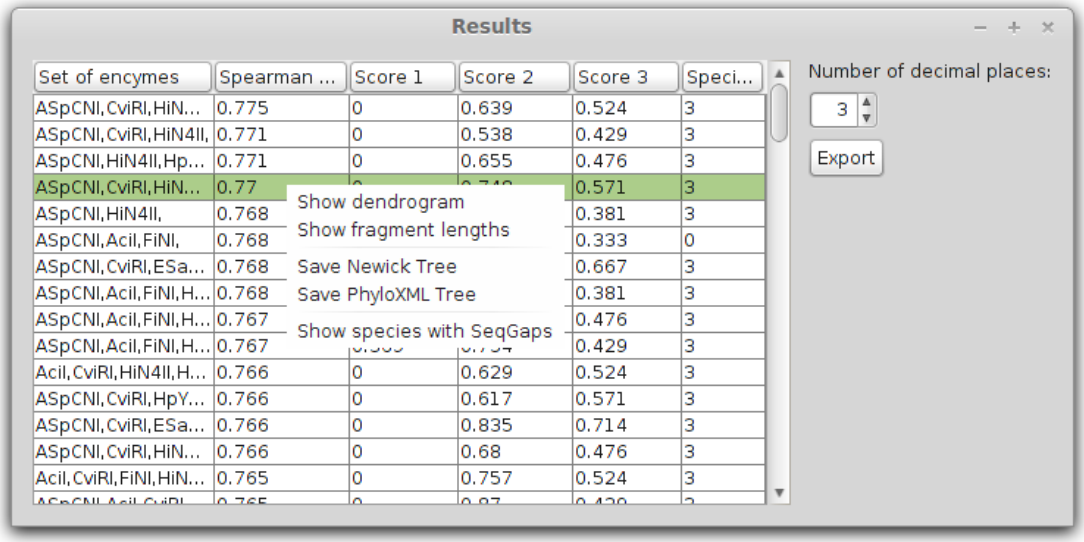
While the search process is running a status window displays the current search progress including an approximated left temporal expenditure:



The search process can be interrupted by the user. In this case the best enzyme sets found up to this time point are shown.

3.9 Results

Following the completion of the computation process the best performing enzyme sets together with the corresponding scores and the Spearman rank coefficient are summarised in a new window ("Results") in form of a table. For more details on the scoring see section 2.4. This information can be exported *via* the Export function as CSV table format for archiving purposes. Additionally, selection of a particular enzyme set from the results table *via* the cursor (left click followed by right click) provides the user with a choice of five options for further details:



1. Show dendrogram: opens a dendrogram using the software tool Archaeopteryx. This dendrogram, which is generated using the unweighted pair group method with arithmetic mean (UPGMA), is based on the distance matrix calculated from the RE set (i.e. the individual set matrices). It is important that the dendrogram is viewed as phylogram (not cladogram) within Archaeopteryx. The complexity of the dendrogram for a given selection of sequences depends on the number of REs within an enzyme set. That is, for one restriction enzyme there are only two branch lengths possible: zero for no difference in TRF length or 1 for a difference (e.g. larger than 10 bp) between two TRFs. In cases where a set of two (or more) restriction enzymes is used the sum of the two (or more) individual distance matrices is used to compute the dendrogram

(see above under section 2.1) resulting in higher complexity. Although visual comparison is convenient, the user is also directed towards the table of TRFs (see below) for confirmation of the results.

2. Show fragment lengths: provides the size (number of nucleotides) of the TRF produced with each of the RE within a RE set. This table can be exported in CSV format for archiving purposes *via* the Save button. Please note that the value -1 for the length of a TRF indicates that the restriction site is not present within the defined window for TRF detection (see above under section 3.5). In these cases the user is strongly adviced to manually check whether these sequences have complete sequence records since the absence of a restriction site may also be caused by lack of sequence information. Manual checks are also strongly recommended in situations where TRFs for some of the selected strains are particularly small (e.g. up to 100 bp) since this indicates restriction sites in regions that are often missing in sequence data (i.e. at the 5'-ends of sequences).

3. Save Newick dendrogram: the dendrogram can be exported in Newick format, which permits archiving and subsequent visualisation *via* a number of alternative tools.

4. Save PhyloXML dendrogram: alternatively, OEZY offers to export the dendrogram in PhyloXML format

5. Show species with sequence gaps: summarises the sequences (name as provided within ARB database field "full name" plus accession number) of sequences with missing sequence information at or downstream of the primer binding site.

4 Copyrights

4.1 Copyright notice for OEZY

Copyright © 2015, Rico Beier and Martin Mühling.
All rights reserved.

The binary distribution together with the manual and the source code of OEZY are freely available for download, use and redistribution in its original or any modified version provided that the following conditions are met:

1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

3. Neither the name of the copyright holder nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

4. Redistributions that integrate or call this program are free of charge and do not generate revenue.

This software is provided by the copyright holders and contributors "as is" and any express or implied warranties, including, but not limited to, the implied warranties of merchantability and fitness for a particular purpose are disclaimed. In no event shall the copyright holder or contributors be liable for any

DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

4.2 Copyright notice for software that is integrated into OEZY: Archeopteryx

This program is free software: you can redistribute it and/or modify it under the terms of the GNU Lesser General Public License as published by the Free Software Foundation, either version 3 of the License, or (at your option) any later version.

This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU Lesser General Public License for more details.

You should have received a copy of the GNU Lesser General Public License along with this program. If not, see `http://www.gnu.org/licenses/`.

References

Amann, R. I., W. Ludwig, and K. H. Schleifer. 1995. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiological Reviews **59**:143–69. URL `http://mmbr.asm.org/content/59/1/143.abstract`.

Clarke, K. R. and R. M. Warwick. 1998. Quantifying structural redundancy in ecological communities. Oecologia **113**:278–289. URL `http://dx.doi.org/10.1007/s004420050379`.

Ludwig, W., O. Strunk, R. Westram, L. Richter, H. Meier, Yadhukumar, A. Buchner, T. Lai, S. Steppi, G. Jobb, W. Förster, I. Brettske, S. Gerber, A. W. Ginhart, O. Gross, S. Grumann, S. Hermann, R. Jost, A. König, T. Liss, R. Lüßmann, M. May, B. Nonhoff, B. Reichel, R. Strehlow, A. Stamatakis, N. Stuckmann, A. Vilbig, M. Lenke, T. Ludwig, A. Bode, and K. Schleifer. 2004. Arb: a software environment for sequence data. Nucleic Acids Research **32**:1363–1371. URL `http://nar.oxfordjournals.org/content/32/4/1363.abstract`.

Mühling, M., J. Woolven-Allen, J. C. Murrell, and I. Joint. 2008. Improved group-specific PCR primers for denaturing gradient gel electrophoresis analysis of the genetic diversity of complex microbial communities. The ISME Journal **2**:379–392.

Westram, R., K. Bader, E. Prüsse, Y. Kumar, H. Meier, F. O. Glöckner, and W. Ludwig. 2011. Arb: a software environment for sequence data. Handbook of molecular microbial ecology I: metagenomics and complementary approaches p. 399–406.