

ONLINE PROTOCOL FOR:

BEsTRF: a tool for optimal resolution of terminal restriction fragment length polymorphism analysis based on user defined primer-enzyme-sequence databases

Blaž Stres^{1,*}, James M. Tiedje², Boštjan Murovec³

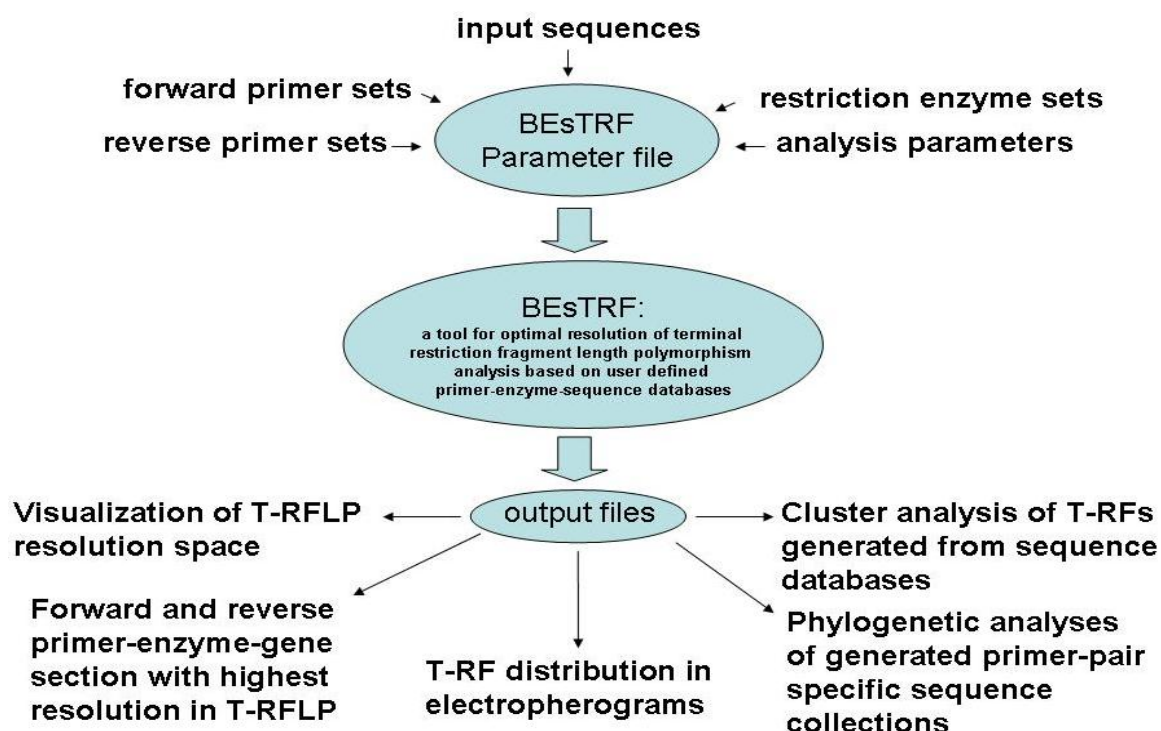
¹Department of Animal Science, Biotechnical Faculty, Chair for Microbiology and Microbial biotechnology, University of Ljubljana, Groblje 3, 1230 Domžale, Slovenia

²Michigan State University, Department of Crop and Soil Sciences and Center for Microbial Ecology, Plant and Soil Science Building 540, MI-48828 East Lansing, USA

³Faculty of Electrical Engineering, Laboratory for computer integrated manufacturing, University of Ljubljana, Tržaška 25, 1000 Ljubljana, Slovenia

Journal: Bioinformatics (please, see the next page for citation information)

The current protocol describes a suitable procedure for using BEsTRF local tool for determining the optimal configuration of T-RFLP. BEsTRF provides an in-depth and controlled environment for an up to date exploration of Primers-Enzymes-Gene sections combinations used in T-RFLP, simulating PCR and fragment generation. User defined sequences database can be processed and the resolution of user specified sets of primers and restriction endonucleases can be analyzed on either forward or reverse terminal fragments thus exploring the vast multidimensional space of forward and reverse primers, restriction endonucleases and sequence detection specificity. In addition, BEsTRF can be used to generate T-RFLP histograms, “virtual clone libraries”, sequences subsamples, retrieved based on primer specificity for downstream phylogenetic analyses. In this protocol, we consider the case of generating optimal configuration for T-RFLP analysis of microbial community using RNA polymerase, beta subunit (*rpoB*) genes as an example. Other worked examples are also presented (please see Worked examples in Supplementary material at http://lie.fe.uni-lj.si/bestrf_examples).



BEsTRF homepage: <http://lie.fe.uni-lj.si/bestrf>

Content:

1.BEsTRF Overview	3
1.1 Choosing input sequences for BEsTRF	3
1.2 Choosing the primer sets	4
1.3 Choosing restriction enzyme sets	5
1.4 Preparing the parameter file for BEsTRF	5
1.5 Running BEsTRF	6
1.6 How BEsTRF works	7
1.7 Computation time	8
2. Output reports and files	9
2.1 Visualization of T-RFLP resolution space	13
2.2 T-RF fragment histogram files and the distribution of T-RFs in theoretical electropherograms	14
2.3. Primer pair histograms and distribution of amplicon lengths (also for teaching purposes)	15
2.3 Cluster analysis of T-RFs: BEsTRF generated data as input files for other downstream programs (also for teaching purposes)	17
2.4 Phylogenetic analyses: BEsTRF generated data as input files for other downstream programs (also for teaching purposes)	18
3. Troubleshooting	21
4. Hints and tips	23
5. Contents of README.txt	24
6. Contents of Readme_first.txt (for bacterial and archaeal examples of Worked examples)	28

Citation:

BEsTRF: a tool for optimal resolution of terminal-restriction fragment length

polymorphism (T-RFLP) analysis based on user defined primer-enzyme-sequence databases

Blaz Stres; James M. Tiedje; Bostjan Murovec

Bioinformatics 2009; doi: 10.1093/bioinformatics/btp254

Free-access link to abstract of the article:

<http://bioinformatics.oxfordjournals.org/cgi/content/abstract/btp254?ikey=qw9v7v9LEcKhXd7&keytype=ref>

1. BEsTRF overview:

1.1 Choosing input sequences for BEsTRF

Choose nucleotide sequences you want to work with. BEsTRF accepts sequences encoded in a widely adopted FASTA (plain *.txt) format. These can be either directly generated from users' specific targets or they can be retrieved from relevant public databases such as Ribosomal Database Project II [{{http://rdp.cme.msu.edu}}](http://rdp.cme.msu.edu), Functional Gene Pipeline / Repository [{{http://flyingcloud.cme.msu.edu/fungene}}](http://flyingcloud.cme.msu.edu/fungene) or NCBI [{{http://www.ncbi.nlm.nih.gov}}](http://www.ncbi.nlm.nih.gov). If you decide to use *rpoB* gene (RNA polymerase, beta subunit) for profiling microbial communities at phylogenetic level instead of 16S rRNA gene, you can do it so by utilizing a set of available *rpoB* sequences [{{http://flyingcloud.cme.msu.edu/fungene}}](http://flyingcloud.cme.msu.edu/fungene). A comprehensive set of sequences results in a primer-enzyme combination that can detect a wide range of sequences and generate T-RFLP profiles of high resolution with numerous distinct peaks.

Sequences should be stored in FASTA formatted plain ASCII files (with usual but not required extensions *.fa or *.txt). BEsTRF can utilize either one file or an arbitrary number of them as a concatenated source of sequences. The current version of BEsTRF was tested with up to 800 000 sequences, but their number is not limited by the program.

Aligned sequence collection looks like this:

```
>NC_007204 DNA-directed RNA polymerase subunit beta [Psychrobacter arcticus 273-4]
atggcatattcttatactgaaaaa-----
-----AA--GCGTATT-----C-----GC-----AAAAGTTTGCTGAATTG
>CP000267 DNA-directed RNA polymerase, beta subunit [Rhodoferax ferrireducens T118]
gtgACAGCAGAAGGGCTTCAAGCCGCTTTGAAG-CGGCTTTCCCGATCATTTAC-AC---AATGGC---TTTG-TCGAGATG-~~~~~AA
AT-ATCTC-----GAGTACAACCTGG-CC
```

whereas the unaligned one looks like this:

```
>NC_007204 DNA-directed RNA polymerase subunit beta [Psychrobacter arcticus 273-4]
atggcatattcttatactgaaaaaAAGCGTATTGCGAAAAGTTTGCTGAATTGCCTACTGTGATGGACATTCCCTATTGTTGTCTATCCAAGTAG
ATTCTTATGAGCAATTTTGCAAGAGCATAAA
>CP000267 DNA-directed RNA polymerase, beta subunit [Rhodoferax ferrireducens T118]
gtgCTGGAAATTCCTTACCTGTTGCAATGCAAAAGGATGCCTACACCGCGTTCTGCAGGCTGATGTTACCCCCAAAAGCGAACAGCAGAAGGGC
TTCAAGCCGCTTTTGAAGCGGCTTTCCCGATCA
```

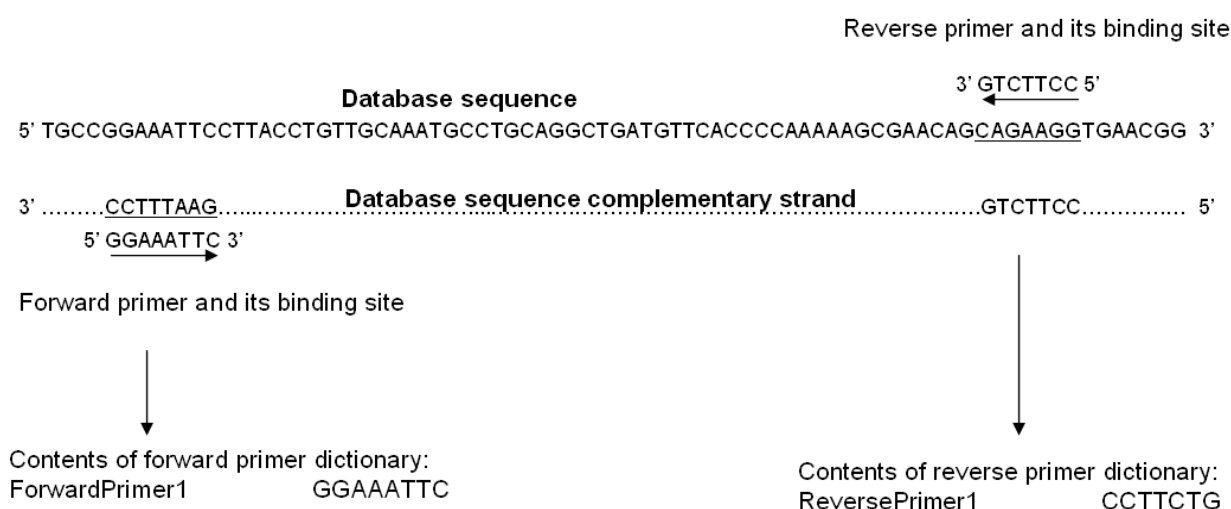
Note: Input sequences are neither required to be aligned nor to be of the same length. However, it is recommended that the 5' orientation is maintained throughout the collection, although BEsTRF can be instructed to automatically reverse sequences.

Note: When aligned sequence databases or user collections are analyzed BEsTRF can handle – or ~ signs for gaps.

1.2 Choosing the primer sets

Numerous primer sets exist in relevant published literature that can be used or refined. In addition, novel primers can be designed using various commercial and freely available tools. BEsTRF was not designed to construct novel primers or probes, but to explore their PCR sampling capacity and T-RFLP resolution.

The selected probes or forward and reverse primers are organized in two separate tab delimited files, called forward and reverse dictionaries (see below fragments from *forward_rpoB.txt* and *reverse_rpoB.txt* dictionaries for an example). Reverse primers must be specified as a reverse complement of matching DNA sequences. For example, the DNA pattern 5' AAACR 3' matches with reverse primer DNA pattern 5' YGTTT 3'. This is essential for proper preparation of forward and reverse primer dictionaries and subsequent correct recognition of primer binding sites as is shown schematically in another example below:



Fragment from forward primer dictionary *f_rpoB.txt*:

rpoBf-6	AGGTCAACTAGTTCAGTATGGACG
rpoB1-f	ATTGACCACTTGGGTAACCGTCG
rpoB1o-f	ATCGATCACTTAGGCAATCGTCG

Fragment from reverse primer dictionary *r_rpoB.txt*:

rpoB2-r	ACGATCACGGGTCAAACCACC
RPOBRa-642	GTTHTGNCDDTTGCATGTT
RPO666Rb	gcttgggtaacctcggag

Note: The current version of BEsTRF was tested to take, as input, up to 50 forward and 50 reverse primers, but their number is not limited by the program. However, the correct primer orientation should be maintained throughout the dictionaries. Please, see also section **Computation time** and the BEsTRF parameter file for additional information.

1.3 Choosing restriction enzyme sets

Numerous restriction enzymes and their derivatives exist, however, only a portion of all available at The Restriction Enzyme Database (<http://rebase.neb.com/rebase/rebase.html>) is used on a regular basis in T-RFLP. Most common ones are four-cutters. The resolution of all enzymes available at Rebase can be explored, while our demo dictionary *4cutters.txt* contains a list of the most frequently used restriction enzymes in T-RFLP analyses. Similarly to specification of primers, the selected restriction enzymes are organized in a tab delimited dictionary, as the following fragment reveals.

Fragment from dictionary *4cutters.txt*:

```
AccII  CG^CG
AciI   C^CGC
AfaI   GT^AC
AspLEI GCG^C
```


1.4 Preparing the parameter file for BEsTRF

Copy and rename template file *BEsTRF_params.txt* (which can be obtained at <http://lie.fe.uni-lj.si/bestrf>) to a new file (e.g. *rpob_params.txt* or *rpobparam1M.txt*) and open it in a plain ASCII text editor to change settings to your preferences. Parameter files contain information that BEsTRF queries for instructions about exploration of T-RFLP space and creation of output files. In addition, this file serves as documentation that can be preserved to keep track of the sessions. The template parameter file contains detailed description of all parameters that BEsTRF accepts.

When configuring parameters numerous decisions with respect to sequence quality, primer degeneracy, enzyme degeneracy, report generation, sequence exploration and others, need to be made by the user. Basically, users enter the name(s) of sequence database(s), forward and reverse primer dictionaries they want to explore, specify enzyme collections and make decisions regarding parameters for guiding the course of analysis as mentioned above. The latter affect the final outcome of a particular analysis. Therefore this step is crucial and deliberately made sequential and recorded as a file. BEsTRF creates separate folders for storing output files that can be named to reflect the parameter file name to keep track of the work in progress.

An example of *rpob* parameter file preparation is illustrated below. The left shaded box presents a fragment from a previously prepared parameter file. For a different sequence collection to be used, simply type in the correct name of your chosen sequence collection after BEsTRF parameter “DNA_File_Names” (like *my_new_rpob_sequences.txt* as in the right shaded box). Arrow points to specified new sequence database name assigned to parameter “DNA_File_Names” for analysis. Thus the original (left) and the modified (right) sections of BEsTRF parameter file should look like this:

<pre>;1. The parameter "DNA_File_Names" specifies file names with DNA ; patterns in FASTA or NumFASTA format (both, aligned or unaligned). ; Note the plural in the name of the parameter. You can specify ; several files to perform a united analysis on their contents. ; I.e. all specified files are regarded as one big file with ; concatenated contents. ; ; NOTE: this parameter is required. If you do not specify at least ; one file name, the program execution will be aborted. ; ;DNA_File_Names rdp_download_129580seqs.fa DNA_File_Names "download_rpoB_Nucleotide_seqs.txt"</pre>	<pre>;1. The parameter "DNA_File_Names" specifies file names with DNA ; patterns in FASTA or NumFASTA format (both, aligned or unaligned). ; Note the plural in the name of the parameter. You can specify ; several files to perform a united analysis on their contents. ; I.e. all specified files are regarded as one big file with ; concatenated contents. ; ; NOTE: this parameter is required. If you do not specify at least ; one file name, the program execution will be aborted. ; ;DNA_File_Names rdp_download_129580seqs.fa DNA_File_Names "my_new_rpob_sequences.txt"</pre>
---	---



Continue and modify all parameters to reflect your preferences and save the file. BEsTRF site also provides the second template file *short_BEsTRF_params.txt*, which merely lists all the parameters, but it omits their comprehensive explanations; it is meant for users already familiar with BEsTRF.

Note: Checkup list before running BEsTRF: Make sure you formatted your sequence, primer and enzyme files appropriately, specified and typed the correct names of your primer and enzyme files into BEsTRF parameter file and set the analysis and output options to your specific needs.

1.5 Running BEsTRF:

The files *README.TXT* and *BEsTRF_params.txt* (as well as this document) contain information about running BEsTRF under Windows or Linux. “BEsTRF.exe” is executable file that can be run under Windows (either by clicking its icon or running it in console window), whereas “bestrf” (not presented in the following figure) is executable file for Linux. In both cases, a parameter file must be specified upon starting the execution (in console window when BEsTRF asks for it (Figure 2) or on a command line after the name of executable file).

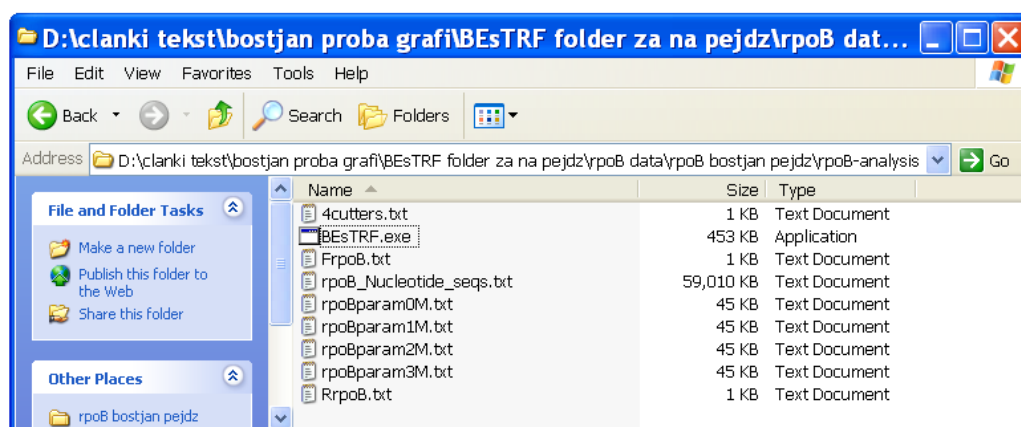


Figure 1: The files needed to run BEsTRF. The file *4cutters.txt* stands for enzyme dictionary containing numerous restriction endonucleases while *RrpoB.txt* and *FrpoB.txt* contain forward and reverse primers, respectively, and are therefore primer dictionaries. The file *rpoB_Nucleotide_seqs.txt* containing target sequences was obtained from Functional Gene Repository / Pipeline (<http://fungene.cme.msu.edu/>). The file *rpoBparam0m.txt* is parameter file of this particular analysis setup.

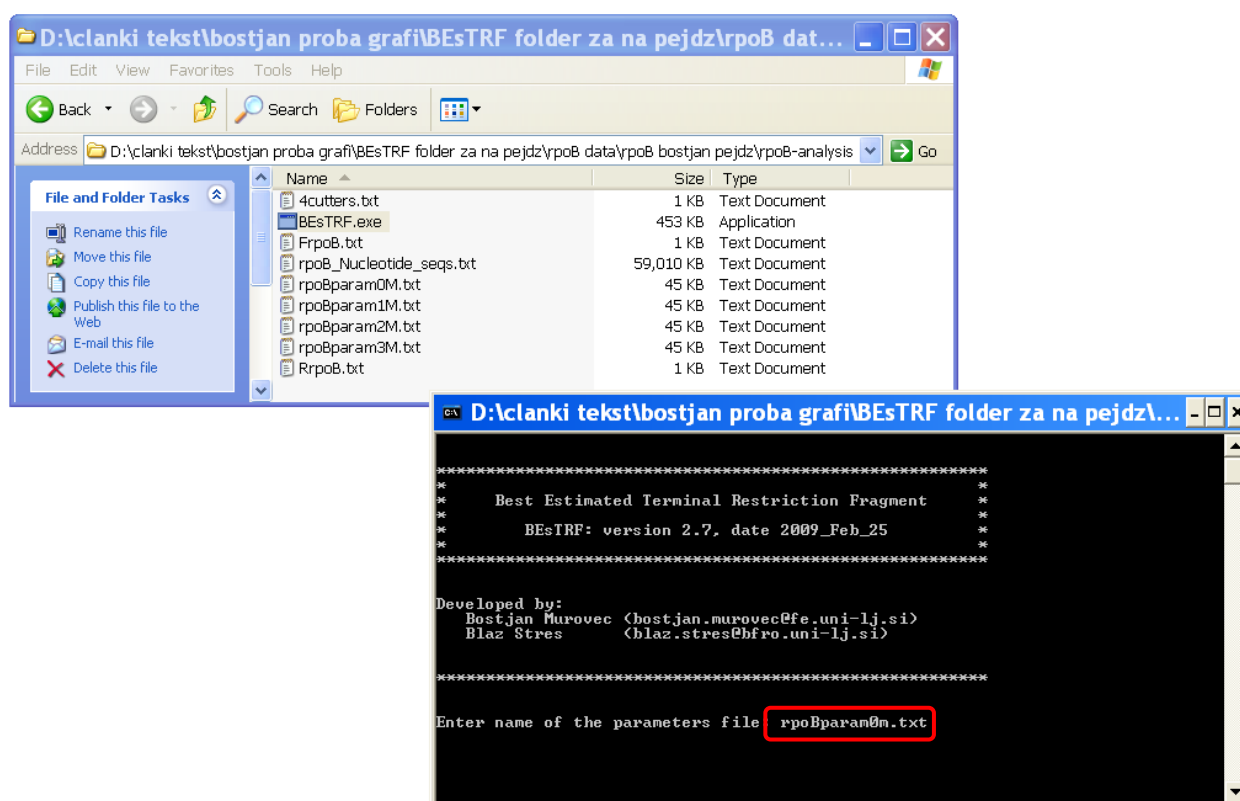


Figure 2: Running BEsTRF with parameter file *rpoBparam1m.tx*.

1.6 How BEsTRF works

Upon execution BEsTRF reads previously described input files. It starts examining user's specified parameter file, where it learns about filenames of other required input files containing sequences database(s), primers and enzymes. The parameter file also prescribes parameters of analysis to be done as well as specifies desired reports to be generated.

The program performs the following steps for each sequence entering an analysis and for each specified forward primer, reverse primer and enzyme combination. During reading from a file, a sequence is optionally examined not to contain any sites with higher degeneration levels than user allows, or else the sequence is ignored and does not take part in an analysis (sequence quality control upon entrance).

The first actual step of an analysis attempts to discover annealing site of a forward primer. Search begins at the start of a sequence and proceeds toward its end. Depending on user's preferences, BEsTRF checks for an exact match between primer and sequence pattern, allows a configured number of mismatches in a plain site-to-site comparison, or utilizes a more elaborate Levenshtein distance [{{http://en.wikipedia.org/wiki/Levenshtein_distance}}](http://en.wikipedia.org/wiki/Levenshtein_distance) for the task. Further, user can specify maximal degeneration level of primer annealing sites. It is important to note that the increased number of allowed ambiguity in primer or restriction enzyme recognition sites promotes the probability of false positive matches. This, however, may or may not be desirable, depending on particular analysis goals.

When annealing site is discovered by means of not utilizing an exact match, BEsTRF can be optionally instructed to examine a certain number of sites beyond the initially discovered annealing position in an attempt to discover position with lesser number of mismatches. Namely, when high level of mismatches is allowed, BEsTRF might discover annealing position certain number of sites before its optimal position due to a loosened matching criterion.

If forward primer annealing position is discovered, the process is repeated for the reverse primer. This time search begins at the end of the sequence and proceeds toward its start, however BEsTRF stops searching when it reaches previously discovered start of forward primer annealing position, by means of which the start of reverse primer annealing position (5' orientation) cannot be located before the start of forward primer one.

If both forward and reverse primer annealing sites are discovered, BEsTRF optionally saves the detected sequence fragment (defined by forward and reverse primer binding sites) into a separate file, thus creating collections of sequences generated by specific primer pairs. Note that the same sequence or its different portions (fragments) can appear in many or all of these files depending on the forward and reverse primers used.

As a sequence quality control step, BEsTRF can count degenerated sites of each sequence fragment between primers (inclusive), and thus reject the sequence from further analysis if degeneracy exceeds a prescribed amount.

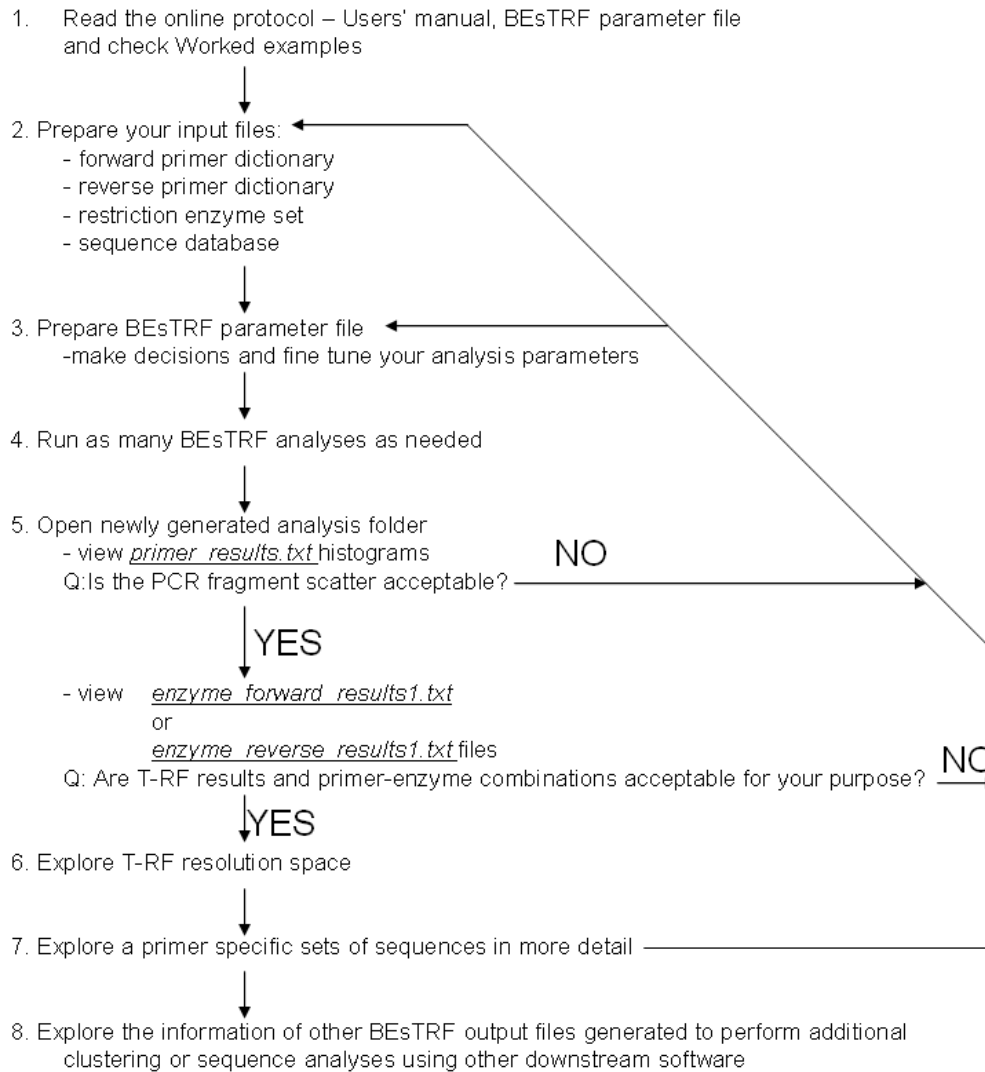
An optional feature of BEsTRF is that it automatically reverses a sequence from 5' orientation to 3' orientation if none of the specified primer pairs bind to it; then the whole analysis repeats. However, spurious results may result from a mixture of 5' and 3' sequence database. For optimal results it is suggested that 5' orientation is maintained.

A resulting fragment length from the start of forward primer annealing site to the end of reverse primer annealing site is added to a respective histogram of a primer pair in question for later generation of analysis reports.

The next step of BEsTRF analysis is discovery of two recognition sites for a particular enzyme for the purpose of generating the forward and reverse terminal-restriction fragments (T-RFs). Search for the first one begins at the start of forward primer annealing position and ends at the end of reverse primer annealing position (in orientation 5'). Search for the second one is done in the opposite direction within the same fragment limits. Enzyme matching criterion options and settings parallel those of primer discovery mechanism.

The possible outcomes of the two searches are two distinctive recognition sites, the same recognition site discovered in either search directions, or no discovered recognition site at all. In the first case BEsTRF adds fragment length from the start of forward primer annealing position to the first cleavage position, to a "forward" histogram of the respective forward primer – reverse primer – enzyme combination, while fragment length from the second cleavage position to the end of reverse primer annealing position, to a "reverse" histogram. If both searches reveal the same recognition site, fragment length from the start of forward primer annealing position to the cleavage position is added to a "forward" histogram, whereas fragment length from the same cleavage position to the end of the reverse annealing position is added to a "reverse" one. In the case of no recognition site, user selects whether nothing or fragment length from the start of forward primer annealing position to the end of reverse primer annealing position is added to the both histograms. Resulting histograms are the basis for generating BEsTRF reports of an analysis.

BEsTRF analysis flowchart



1.7 Computation time

The time required to complete an analysis depends on the following user defined parameters:

- number of sequences
- sequence database asymmetry
- presence/absence of primer annealing sites in sequence database
- number of primer pairs
- number of enzymes
- selection and the extent of Levenshtein distance utilization
- selection of reports to generate (generation of pivot tables (pg 12))

In addition, the following computer's horsepower specifications should be considered:

- processor speed
- amount of RAM memory
- disk performance and space (bacterial database containing genes for 16S rRNA unzips to >20 Gb)

2. Output reports and files

The desired output is determined by configuring BEsTRF parameter file (e.g. *BEsTRF_params.txt*) that is specifically created by the user to keep track of each session's work in progress. After each run, numerous files are generated and stored in a novel directory that is created and named according to user defined specifications in parameter file.

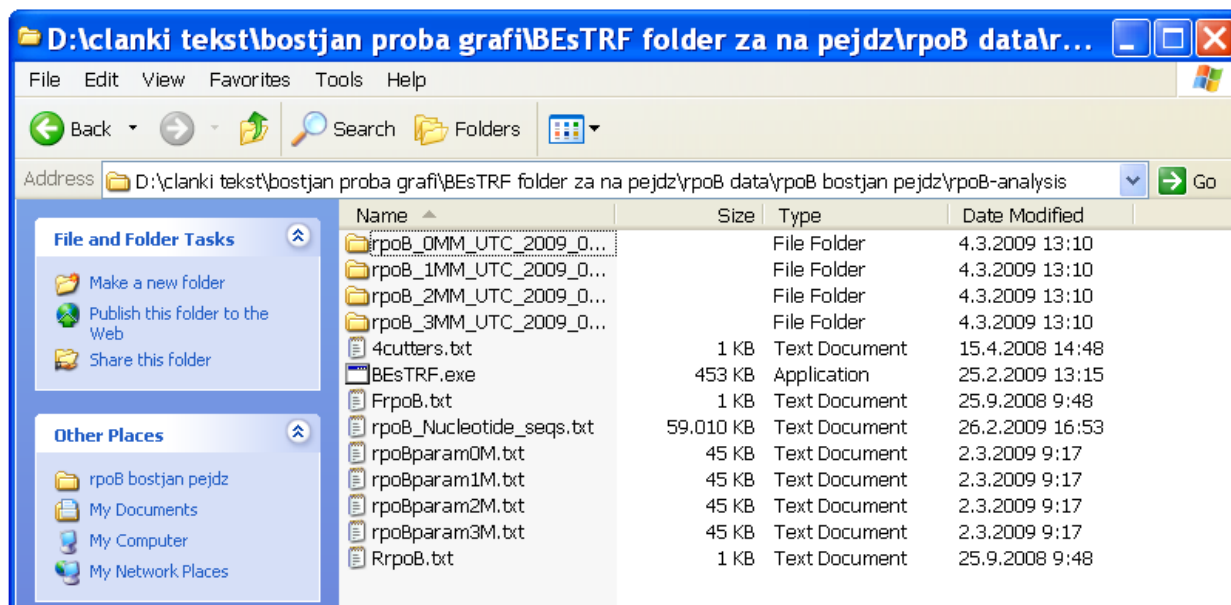


Figure 3: New directories were generated and named according to specifications in various BEsTRF parameter files after each run.

In general, the following files are generated (Figure 4):

- T-RF results for forward primer dictionary primarily sorted by unique T-RFs (i.e. *enzyme_forward_results1.txt*)
- T-RF results for forward primer dictionary primarily sorted by the number of detected sequences (i.e. *enzyme_forward_results2.txt*)
- T-RF results for reverse primer dictionary primarily sorted by unique T-RFs (i.e. *enzyme_reverse_results1.txt*)
- T-RF results for reverse primer dictionary primarily sorted by the number of detected sequences (i.e. *enzyme_reverse_results2.txt*)
- *fwd_fragments.txt* - forward T-RF data generated for all primer pairs with all user defined enzymes organized as a pivot table
- *rev_fragments.txt* - reverse T-RF data generated for all primer pairs and with all user defined enzymes organized as a pivot table
- numerous libraries of sequences detected by each primer pair according to user defined acceptable quality and primer site recognition parameters (i.e. *accepted_RPOB175F_RL2.txt*)
- accepted sequence library according to user defined acceptable quality – (i.e. *accepted.txt*)
- rejected sequence library according to user defined acceptable quality – (i.e. *rejected.txt*)
- rejected sequence library when no primer binding sites were detected – No Primer (i.e. *rejected_NP.txt*);
- rejected sequences library when no enzyme cutting site was detected – No Enzyme (i.e. *rejected_NE.txt*).
- primer pair fragment histograms and distribution of amplicon lengths – *primer_results.txt*

For various users' defined thresholds, options and details that affect BEsTRF analyses and output files, please consider template BEsTRF parameter file (*BEsTRF_params.txt*).

Once the particular newly created folder is opened, the output files and folders become visible.

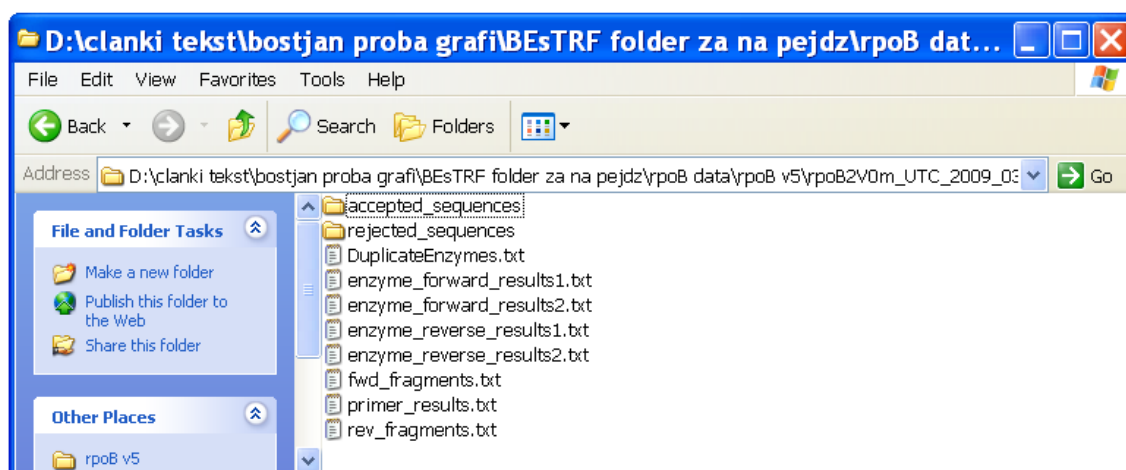


Figure 4: Newly created analysis folders contain two subfolders *accepted_sequences* and *rejected_sequences* that contain the accepted and rejected sequence files as described above. In addition, enzymes with identical cutting sites are listed in new file *DuplicateEnzymes.txt*.

The following files contain the relevant T-RF results:

- *enzyme_forward_results1.txt*: T-RF results for forward primer dictionary primarily sorted by unique T-RFs
- *enzyme_forward_results2.txt*: T-RF results for forward primer dictionary primarily sorted by the number of detected sequences
- *enzyme_reverse_results1.txt*: T-RF results for reverse primer dictionary primarily sorted by unique T-RFs
- *enzyme_reverse_results2.txt*: T-RF results for reverse primer dictionary primarily sorted by the number of detected sequences

These files can be opened, viewed or imported into various text editors such as Notepad, WordPad or Word under Windows, Open Office, joe or gedit under Linux, or spreadsheet and other programs such as Statistica, SAS, Matlab, SPSS, ORIGIN, gnuplot or root. 2D XY or 3D XYZ graphs can be plotted as schematically shown below following the chosen spreadsheet program commands.

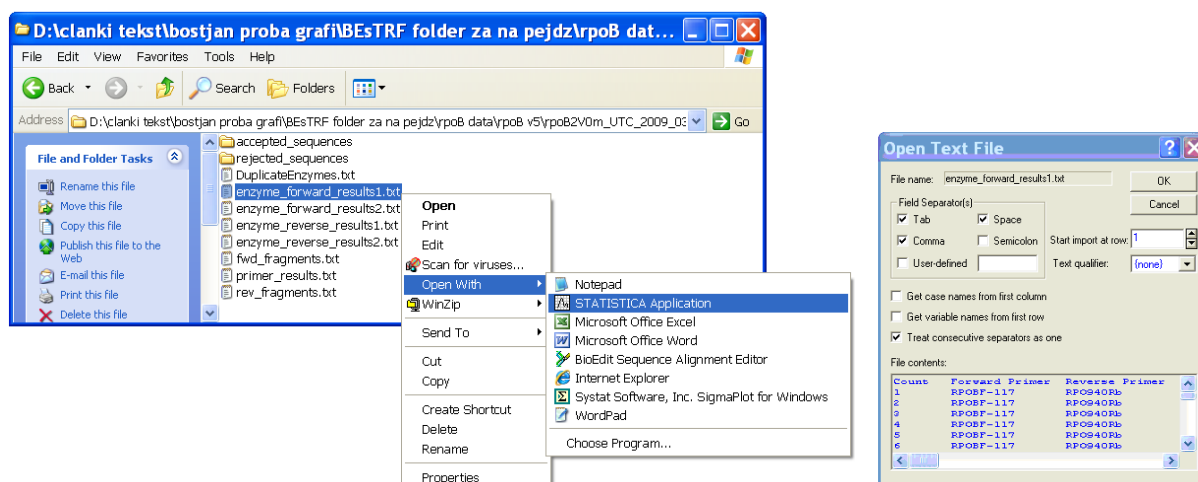


Figure 5: The T-RF results stored in various tab delimited *.txt text files such as *enzyme_forward_results1.txt* can be viewed in a variety of spreadsheet programs.

a)

A	B	C	D	E	F	G	H	I	J	K	L	M
							min PCR amplicon length (bp)	max PCR amplicon length (bp)	max-min PCR amplicon length (bp)	average PCR amplicon length (bp)	std dev. of PCR amplicon length (bp)	min T-RF length (bp)
count	forward primer	reverse primer	combination name	enzyme	DNA matches (N)	unique T-RFs (N)						
1	RL1	RL2	RL1 @ RL2	AspLEI	18	3	369	369	0	369	0	92
2	RL1	RL2	RL1 @ RL2	Glal	18	3	369	369	0	369	0	91
3	RL1	RL2	RL1 @ RL2	Hin6I	18	3	369	369	0	369	0	90
4	RL1	RL2	RL1 @ RL2	SetI	18	3	369	369	0	369	0	17
5	RPOBF-117	RL2	RPOBF-117 @ RL2	AccII	13	3	159	159	0	159	0	50
6	RPOBF-117	RL2	RPOBF-117 @ RL2	Acil	13	3	159	159	0	159	0	22
7	RPOBF-117	RL2	RPOBF-117 @ RL2	CviJJ	13	3	159	159	0	159	0	95
8	RPOBF-117	RL2	RPOBF-117 @ RL2	HpyCH4V	13	3	159	159	0	159	0	58
9	RPOBF-117	RL2	RPOBF-117 @ RL2	Sell	13	3	159	159	0	159	0	48
10	RPOBF-117	RPO940Rb	RPOBF-117 @ RPO940Rb	Acil	6	3	524	524	0	524	0	22
11	RPOBF-117	RPO940Rb	RPOBF-117 @ RPO940Rb	Afal	6	3	524	524	0	524	0	104
12	RPOBF-117	RPO940Rb	RPOBF-117 @ RPO940Rb	Alul	6	3	524	524	0	524	0	212
13	RPOBF-117	RPO940Rb	RPOBF-117 @ RPO940Rb	AspLEI	6	3	524	524	0	524	0	47
14	RPOBF-117	RPO940Rb	RPOBF-117 @ RPO940Rb	BshFI	6	3	524	524	0	524	0	58
15	RPOBF-117	RPO940Rb	RPOBF-117 @ RPO940Rb	BsiSI	6	3	524	524	0	524	0	67
16	RPOBF-117	RPO940Rb	RPOBF-117 @ RPO940Rb	Csp6I	6	3	524	524	0	524	0	103
17	RPOBF-117	RPO940Rb	RPOBF-117 @ RPO940Rb	CviJJ	6	3	524	524	0	524	0	58
18	RPOBF-117	RPO940Rb	RPOBF-117 @ RPO940Rb	EsaBC3I	6	3	524	524	0	524	0	133
19	RPOBF-117	RPO940Rb	RPOBF-117 @ RPO940Rb	Glal	6	3	524	524	0	524	0	46
20	RPOBF-117	RPO940Rb	RPOBF-117 @ RPO940Rb	Hin6I	6	3	524	524	0	524	0	45
21	RPOBF-117	RPO940Rb	RPOBF-117 @ RPO940Rb	HpyCH4V	6	3	524	524	0	524	0	114
22	RPOBF-117	RPO940Rb	RPOBF-117 @ RPO940Rb	PabI	6	3	524	524	0	524	0	105
23	RPOBF-117	RPO940Rb	RPOBF-117 @ RPO940Rb	SetI	6	3	524	524	0	524	0	48
24	RPOBF-117	RPO940Rb	RPOBF-117 @ RPO940Rb	Sse9I	6	3	524	524	0	524	0	32
25	RPOBF-117	RPO940Rb	RPOBF-117 @ RPO940Rb	Sth302II	6	3	524	524	0	524	0	68
26	RPOBF-117	RPO940Rb	RPOBF-117 @ RPO940Rb	TaqI	6	3	524	524	0	524	0	132
27	RL1	RL2	RL1 @ RL2	AccII	18	2	369	369	0	369	0	68
28	RL1	RL2	RL1 @ RL2	Acil	18	2	369	369	0	369	0	66
29	RL1	RL2	RL1 @ RL2	Afal	18	2	369	369	0	369	0	314
30	RL1	RL2	RL1 @ RL2	Alul	18	2	369	369	0	369	0	268
31	RL1	RL2	RL1 @ RL2	Bfal	18	2	369	369	0	369	0	16
32	RL1	RL2	RL1 @ RL2	BshFI	18	2	369	369	0	369	0	118
33	RL1	RL2	RL1 @ RL2	BsiSI	18	2	369	369	0	369	0	277
34	RL1	RL2	RL1 @ RL2	Csp6I	18	2	369	369	0	369	0	313
35	RL1	RL2	RL1 @ RL2	CviAll	18	2	369	369	0	369	0	135
36	RL1	RL2	RL1 @ RL2	CviJJ	18	2	369	369	0	369	0	118
37	RL1	RL2	RL1 @ RL2	Fael	18	2	369	369	0	369	0	138
38	RL1	RL2	RL1 @ RL2	FatI	18	2	369	369	0	369	0	134
39	RL1	RL2	RL1 @ RL2	HpyCH4IV	18	2	369	369	0	369	0	27
40	RL1	RL2	RL1 @ RL2	HpyCH4V	18	2	369	369	0	369	0	172
41	RL1	RL2	RL1 @ RL2	PabI	18	2	369	369	0	369	0	315
42	RL1	RL2	RL1 @ RL2	Sell	18	2	369	369	0	369	0	86

b)

Fragments histogram for Forward_Primer/Reverse_Primer/Enzyme combination RL1/RL2/AspLEI (statistics row: 1)					
Fragment length	Frequency				
92	11				
107	4				
369	3				
Fragments histogram for Forward_Primer/Reverse_Primer/Enzyme combination RL1/RL2/GlaI (statistics row: 2)					
Fragment length	Frequency				
91	11				
106	4				
369	3				
Fragments histogram for Forward_Primer/Reverse_Primer/Enzyme combination RL1/RL2/Hin6I (statistics row: 3)					
Fragment length	Frequency				
90	11				
105	4				
369	3				

11

The next tab delimited text files generated by BEsTRF are named *fwd_fragments.txt* or *rev_fragments.txt* and contain T-RF data generated for all primer pairs using all user defined restriction enzymes organized as a pivot table in two dimensions: columns contain data for each primer pair while horizontal sections are organized according to restriction enzymes used (encircled in Figure 7).

The relative and absolute T-RF fragment abundances are sorted by size in a manner of a pivot table as shown below. Next horizontal section beneath the first one contains data for the second enzyme and so forth (Figure 7), thus resulting in very large files for numerous forward and reverse primer combinations and restriction enzymes.

Please see the following sections **Cluster analysis of T-RFs** and **Phylogenetic analyses** for more details and further use of these files.

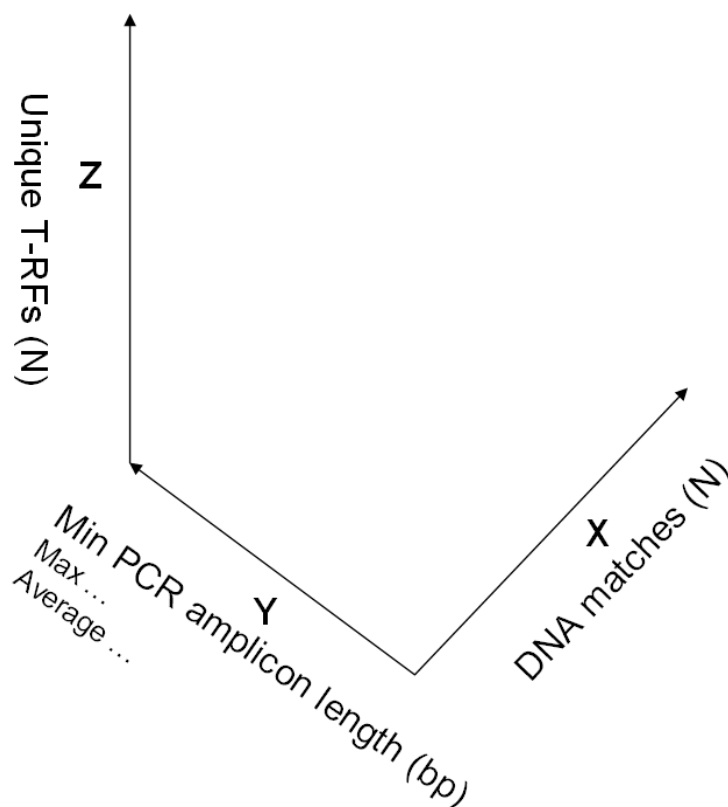
A		B		C		D		E		F		G		H		I															
Primers: RL1 - RL2		Enzyme		Fragment Size		Fragment Relative Abundance (%)		Fragment Abundance (no. of seg.)		Primers: RL1 - RPO66Rb		Enzyme		Fragment Size		Fragment Relative Abundance (%)		Fragment Abundance (no. of seg.)		Primers: RL1 - RPO66Rb		Enzyme		Fragment Size		Fragment Relative Abundance (%)		Fragment Abundance (no. of seg.)			
Enzyme		Accil								Enzyme		Accil								Enzyme		Accil									
				23		0		0		0		23		0		0		0		0		0		0		0		0			
				25		0		0		0		25		0		0		0		0		0		0		0		0			
				26		0		0		0		26		0		0		0		0		0		0		0		0			
				29		4,91803		3		3		29		66,6667		4		4		4		4		4		4		4			
				32		0		0		0		32		0		0		0		0		0		0		0		0			
				34		0		0		0		34		0		0		0		0		0		0		0		0			
				36		0		0		0		36		0		0		0		0		0		0		0		0			
				38		0		0		0		38		0		0		0		0		0		0		0		0			
				39		0		0		0		39		0		0		0		0		0		0		0		0			
				40		0		0		0		40		0		0		0		0		0		0		0		0			
				46		0		0		0		46		0		0		0		0		0		0		0		0			
				49		0		0		0		49		0		0		0		0		0		0		0		0			
				50		0		0		0		50		0		0		0		0		0		0		0		0			
				55		0		0		0		55		0		0		0		0		0		0		0		0			
				56		0		0		0		56		0		0		0		0		0		0		0		0			
				58		0		0		0		58		0		0		0		0		0		0		0		0			
				62		0		0		0		62		0		0		0		0		0		0		0		0			
				68		27,8689		17		17		68		0		0		0		0		0		0		0		0			
				74		0		0		0		74		0		0		0		0		0		0		0		0			
				75		0		0		0		75		0		0		0		0		0		0		0		0			
				77		0		0		0		77		0		0		0		0		0		0		0		0			
				81		0		0		0		81		0		0		0		0		0		0		0		0			
				83		3,27869		2		2		83		0		0		0		0		0		0		0		0			
				86		0		0		0		86		0		0		0		0		0		0		0		0			
				89		0		0		0		89		0		0		0		0		0		0		0		0			
				90		0		0		0		90		0		0		0		0		0		0		0		0			
				92		9,83607		6		6		92		0		0		0		0		0		0		0		0			
				95		0		0		0		95		0		0		0		0		0		0		0		0			
				101		6,55738		4		4		101		0		0		0		0		0		0		0		0			
				107		0		0		0		107		0		0		0		0		0		0		0		0			
				113		0		0		0		113		0		0		0		0		0		0		0		0			
				121		0		0		0		121		0		0		0		0		0		0		0		0			
				122		0		0		0		122		0		0		0		0		0		0		0		0			
				134		0		0		0		134		0		0		0		0		0		0		0		0			
				154		0		0		0		154		0		0		0		0		0		0		0		0			
				157		1,63934		1		1		157		0		0		0		0		0		0		0		0			
				159		0		0		0		159		0		0		0		0		0		0		0		0			
				163		0		0		0		163		0		0		0		0		0		0		0		0			
				166		0		0		0		166		0		0		0		0		0		0		0		0			
				189		0		0		0		189		0		0		0		0		0		0		0		0			
				190		0		0		0		190		0		0		0		0		0		0		0		0			
				191		0		0		0		191		0		0		0		0		0		0		0		0			
				201		0		0		0		201		0		0		0		0		0		0		0		0			
				203		0		0		0		203		0		0		0		0		0		0		0		0			
				215		0		0		0		215		0		0		0		0		0		0		0		0			
				239		0		0		0		239		0		0		0		0		0		0		0		0			
				245		0		0		0		245		0		0		0		0		0		0		0		0			
				248		0		0		0		248		0		0		0		0		0		0		0		0			
				256		0		0		0		256		33,3333		2		2		2		2		2		2		2			
				260		1,63934		1		1		260		0		0		0		0		0		0		0		0			
				265		0		0		0		265		0		0		0		0		0		0		0		0			
				286		0		0		0		286		0		0		0		0		0		0		0		0			
				289		0		0		0		289		0		0		0		0		0		0		0		0			
				291		1,63934		1		1		291		0		0		0		0		0		0		0		0			
				292		0		0		0		292		0		0		0		0		0		0		0		0			
				293		0		0		0		293		0		0		0		0		0		0		0		0			
				294		0		0		0		294		0		0		0		0		0		0		0		0			
				299		0		0		0		299		0		0		0		0		0		0		0		0			
				300		0		0		0		300		0		0		0		0		0		0		0		0			
				311		0		0		0		311		0		0		0		0		0		0		0		0			
				317		0		0		0		317		0		0		0		0		0		0		0		0			
				328		6,55738		4		4		328		0		0		0		0		0		0		0		0			
				332		0		0		0		332		0		0		0		0		0		0		0		0			
				338		0		0		0		338		0		0		0		0		0		0		0		0			
				359		0		0		0		359		0		0		0		0		0		0		0		0			
				363		0		0		0		363		0		0		0		0		0		0		0		0			
				365		0		0		0		365		0		0		0		0		0		0		0		0			
				369		36,0666		22		22		369		0		0		0		0		0		0		0		0			
				380		0		0		0		380		0		0		0		0		0		0		0		0			
				386		0		0		0		386		0		0		0		0		0		0		0		0			
				402		0		0		0		402		0		0		0		0		0		0		0		0			
				404		0		0		0		404		0		0		0		0		0		0		0		0			
				410		0		0		0		410		0		0		0		0		0		0		0		0			
				411		0		0		0		411		0		0		0		0		0		0		0		0			
				446		0		0		0		446		0		0		0		0		0		0		0		0			
				473		0		0		0		473		0		0		0		0		0		0		0		0			
				524		0		0		0		524		0		0		0		0		0		0		0		0			
				530		0		0		0		530		0		0		0		0		0		0		0		0			
				533		0		0		0		533		0		0		0		0		0		0		0		0			
				602		0		0		0		602		0		0		0		0		0		0		0		0			
				630		0		0		0		630		0		0		0		0		0		0		0		0			
				701		0		0		0		701		0		0		0		0		0		0		0		0			
				728		0		0		0		728		0		0		0		0		0		0		0		0			
				734		0		0		0		734		0		0		0		0		0		0		0		0			
				740		0		0		0		740		0		0		0		0		0		0		0		0			
				743		0		0		0		743		0		0		0		0		0		0							

2.1 Visualization of T-RFLP resolution space

In order to visualize the T-RFLP resolution space as determined by available sequences, their length and quality, primer sets used, enzymes and amplicon lengths or other, the data can be plotted using common graphic software as Statistica, SAS, Matlab, SPSS, ORIGIN, R/project, BioPerl, gnuplot, root and others. The files *enzyme_forward_results1.txt* or *enzyme_reverse_results1.txt* should be used for this purpose, where user chooses from the following information (columns) listed sequentially below:

```
Row count
Forward Primer
Reverse Primer
Combination name
Enzyme
* DNA Matches (N)
* Unique T-RFs (N)
* Min PCR amplicon length(bp)
Max PCR amplicon length(bp)
Max-Min PCR amplicon length(bp)
Average PCR amplicon length(bp)
Std. Dev. of PCR amplicon length (bp)
Min T-RF length (bp)
Max T-RF length (bp)
Max-Min T-RF length (bp)
Average T-RF length (bp)
Std. Dev. of T-RF length (bp)
```

To plot 3D xyz graphs (as schematically shown below; Figure 8), user needs to define three columns of data (in this case marked with * above) in the imported files for graphing. Other graphs that may suit your needs better can be generated by marking other columns, or by plotting 2D XY graphs using your preferred software. We used Statistica (StatSoft Inc.) to produce 3D XYZ graphs presented in Figure 8 for various analysis parameters.



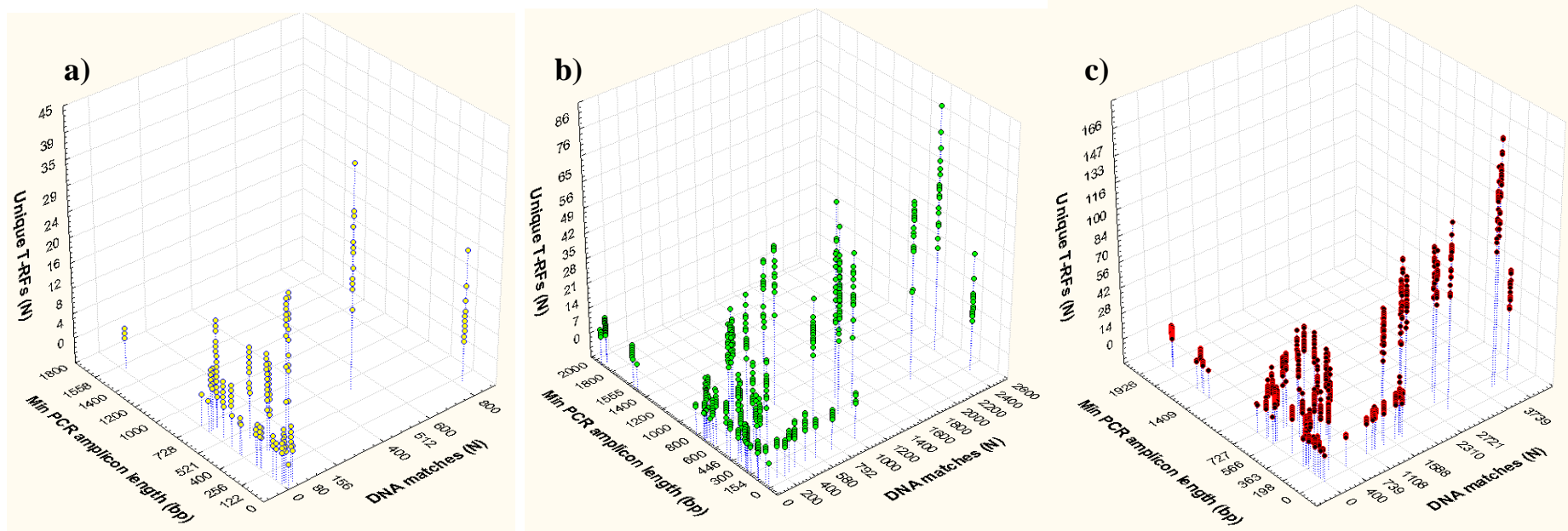


Figure 8: A three dimensional plot of candidate primer and enzyme combinations showing the extensive differences among sampling capacities of various primer combinations from underlying (most often asymmetric) sequence database and a gradient in the stringency of analyses: **(a)** one mismatch was allowed for each primer, but no insertions or deletions (Levenshtein off), maximum primer degeneracy code was 0 (only perfect matches allowed), sequence degeneracy was also 0 (only IUPAC characters (ATGC) were allowed); **(b)** two mismatches were allowed in addition to Levenshtein parameters (1, 1, 1) that allowed for one insertion or deletion plus maximum primer degeneracy was set to four (less than perfect matches allowed) thus enabling more relaxed recognition of primer binding sites; **(c)** an even more relaxed version of analysis where three mismatches were allowed in addition to all parameters of **(b)** resulted in increased number of Unique T-RFs, DNA matches, but also background, as can be viewed in corresponding *primer_results.txt* files. These plots were for instance generated using Statistica (StatSoft Inc.).

2.2 T-RF fragment histograms and distribution of T-RFs in theoretical electropherograms (also for teaching purposes)

The best primer-enzyme fragment histogram (Figure 6a; Figure 9) is the first in hierarchical display of the same output file *enzyme_forward_results1.txt* or *enzyme_reverse_results1.txt* generated by different settings. The comparison/toggling between primer combinations results (Figure 6a) and the corresponding T-RFLP histograms (Figure 6b; Figure 9) is enabled by simple Find/Replace navigation. Yet, you might want to explore other primer-enzyme sets in the output file that better reflect your needs. The length and frequency distribution of T-RF peaks and evenness of the most promising combinations can also be visualized using common graphic software as Statistica, SAS, Mathlab, SPSS, ORIGIN, gnuplot, R/project, BioPerl or other (Figure 10).

Fragments histogram for Forward_Primer/Reverse_Primer/Enzyme combination RPOBF-117/RPO940Rb/AciI (statistics row: 2)		
	Fragment length	Frequency
	22	296
	48	4
	56	6
	81	9
	105	5
	132	2
	201	4
	237	10
	273	2
	284	10
	290	1
	300	2
	335	2
	368	2
	372	2
	400	2
	406	2
	433	1
	495	2
	524	18
	527	2
	533	2

Figure 9: A typical view of fragments histogram summarizing the detected T-RF fragment length distribution.

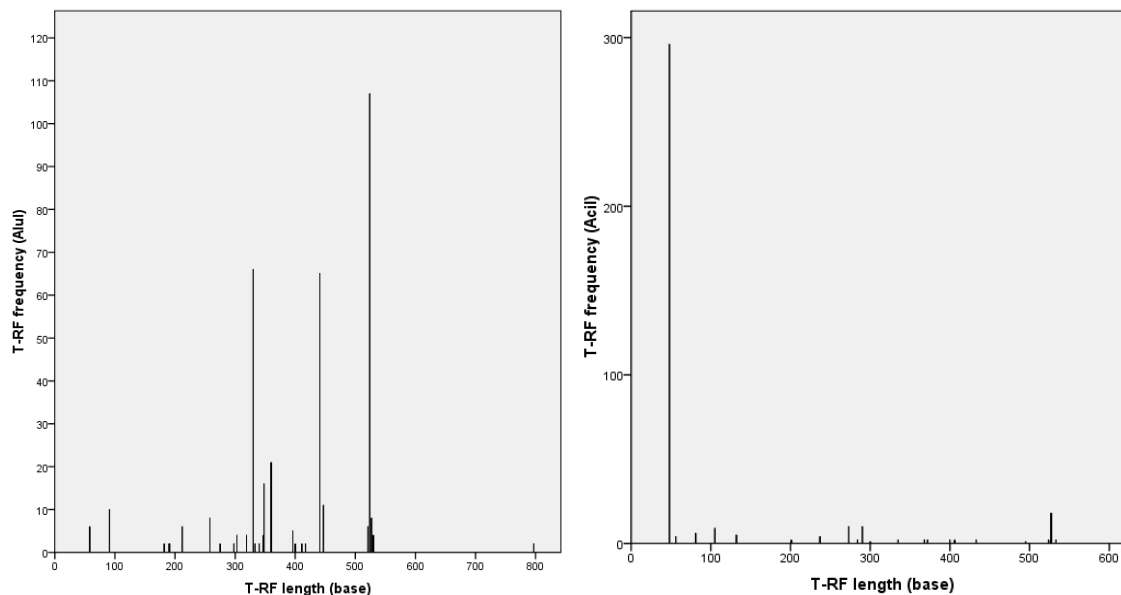


Figure 10: Distinct T-RF distribution can result from the same sequence collections detected by a primer pair depending on the restriction enzyme used. Two typical T-RF combinations (*AluI* and *AciI*) exhibiting differential T-RF frequency and length evenness are presented. You may find this information useful for judging the separation of T-RF peaks generated by various restriction enzymes before analyzing true samples on capillary sequencer.

2.3 Primer pair fragment histograms and distribution of amplicon lengths (also for teaching purposes)

Ideally, all detected fragments by particular forward and reverse primer pair should be of the same length. However, this is only rarely the case. Highly different lengths of the detected sequence fragments can be generated due to sequence internal structure (conserved, variable regions), sequence database entries of varying quality or multiple mismatches allowed in sampling procedure performed by BEsTRF according to user defined options presented in BEsTRF parameter file.

Thus, the lengths of sequence fragments retrieved by each primer combination are mapped, scanned for differences and presented as histograms in *primer_results.txt* files. Note that these histograms contain information regarding PCR fragment lengths of detected sequences as determined by forward and reverse primer combination and do NOT represent T-RFLP histograms.

This information is relevant as a control of sampling procedure performed by BEsTRF and to visualize the fragment length scatter. As you might want to visualize and explore also other primer sets in the output file that better reflect your needs, the following parameters can also be visualized using common graphic software as Statistica, SAS, Mathlab, SPSS, ORIGIN, gnuplot, R/project, BioPerl or other (Figure 11):

a)

count	forward primer	reverse primer	primer combination name	DNA matches (N)	different PCR amplicon lengths (N)	min PCR amplicon length (bp)	max PCR amplicon length (bp)	(max-min) PCR amplicon length (bp)
1	RPOBF-117	RL2	RPOBF-117 @ RL2	829	1	159	159	0
2	RPOBF-117	RPO940Rb	RPOBF-117 @ RPO940Rb	513	13	521	797	276
3	rpoB1698f	RPO940Rb	rpoB1698f @ RPO940Rb	165	7	356	398	42
4	rpoB1698f	rpoB2041r	rpoB1698f @ rpoB2041r	156	7	358	374	16
5	rpoB1-f	RL2	rpoB1-f @ RL2	88	1	363	363	0
6	RL1	RL2	RL1 @ RL2	81	1	369	369	0
7	RPOBF-117	rpoB2041r	RPOBF-117 @ rpoB2041r	80	4	524	677	153
8	RPOBF2-9	RL2	RPOBF2-9 @ RL2	48	1	81	81	0
9	rpoB1-f	rpoB2041r	rpoB1-f @ rpoB2041r	38	1	734	734	0
10	RL1	RPO940Rb	RL1 @ RPO940Rb	37	5	734	746	12
11	RPOBF-117	RPO666Rb	RPOBF-117 @ RPO666Rb	37	1	46	46	0
12	RL1	rpoB2041r	RL1 @ rpoB2041r	21	3	734	797	63
13	RPO216Fbs	rpoB2041r	RPO216Fbs @ rpoB2041r	20	1	66	66	0
14	RPOBF-117	rpoB2-r	RPOBF-117 @ rpoB2-r	20	1	90	90	0
15	RPO310F	RPO940Rb	RPO310F @ RPO940Rb	19	3	631	633	2
16	RPO216Fbs	RPO940Rb	RPO216Fbs @ RPO940Rb	14	2	66	67	1
17	rpoB1-f	RPO940Rb	rpoB1-f @ RPO940Rb	12	2	728	767	39
18	rpoBF6	RL2	rpoBF6 @ RL2	11	2	1,558	1,561	3
19	RPOBF-117	RPOBRa-642	RPOBF-117 @ RPOBRa-642	8	2	531	532	1
20	RPO310F	RPO666Rb	RPO310F @ RPO666Rb	8	1	154	154	0

b)

Fragments histogram for Primer pair RPOBF-117/RL2 (statistics row: 1)			
Fragn	Frequency		
159	829		
Fragments histogram for Primer pair RPOBF-117/RPO940Rb (statistics row: 2)			
Fragn	Frequency		
521	17		
522	2		
523	1		
524	428		
525	3		
526	1		
527	13		
530	36		
533	2		
539	4		
647	2		
650	2		
797	2		
Fragments histogram for Primer pair rpoB1698f/RPO940Rb (statistics row: 3)			
Fragn	Frequency		
356	8		
359	146		
361	1		
362	1		
365	6		
374	2		
398	1		
Fragments histogram for Primer pair rpoB1698f/rpoB2041r (statistics row: 4)			
Fragn	Frequency		
358	2		
359	119		
360	1		
364	2		
365	26		
373	4		
374	2		

Figure 11 continued

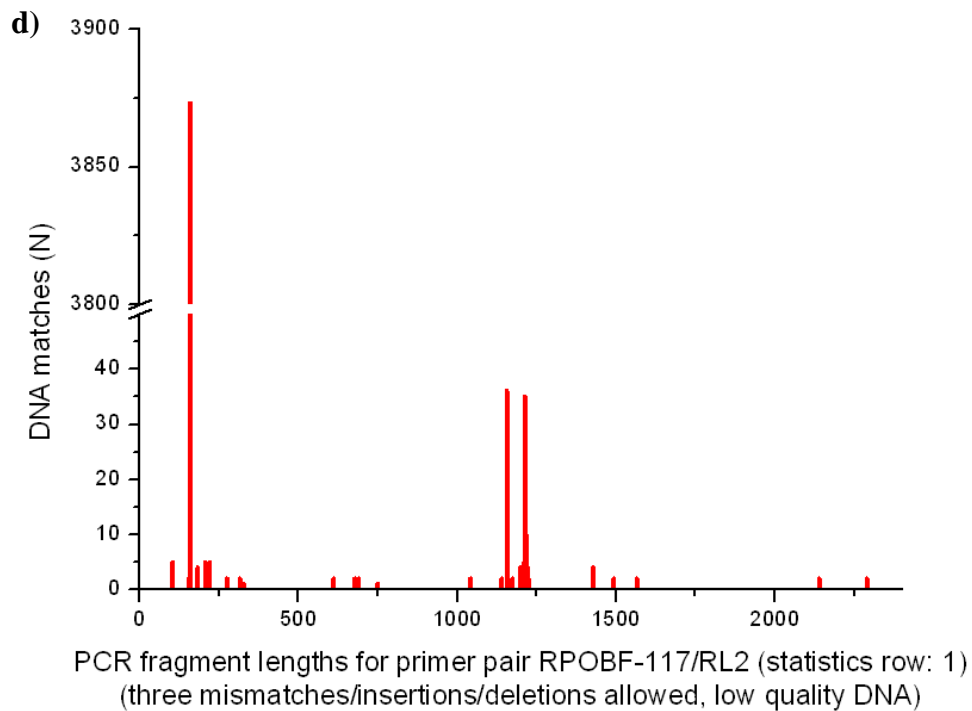
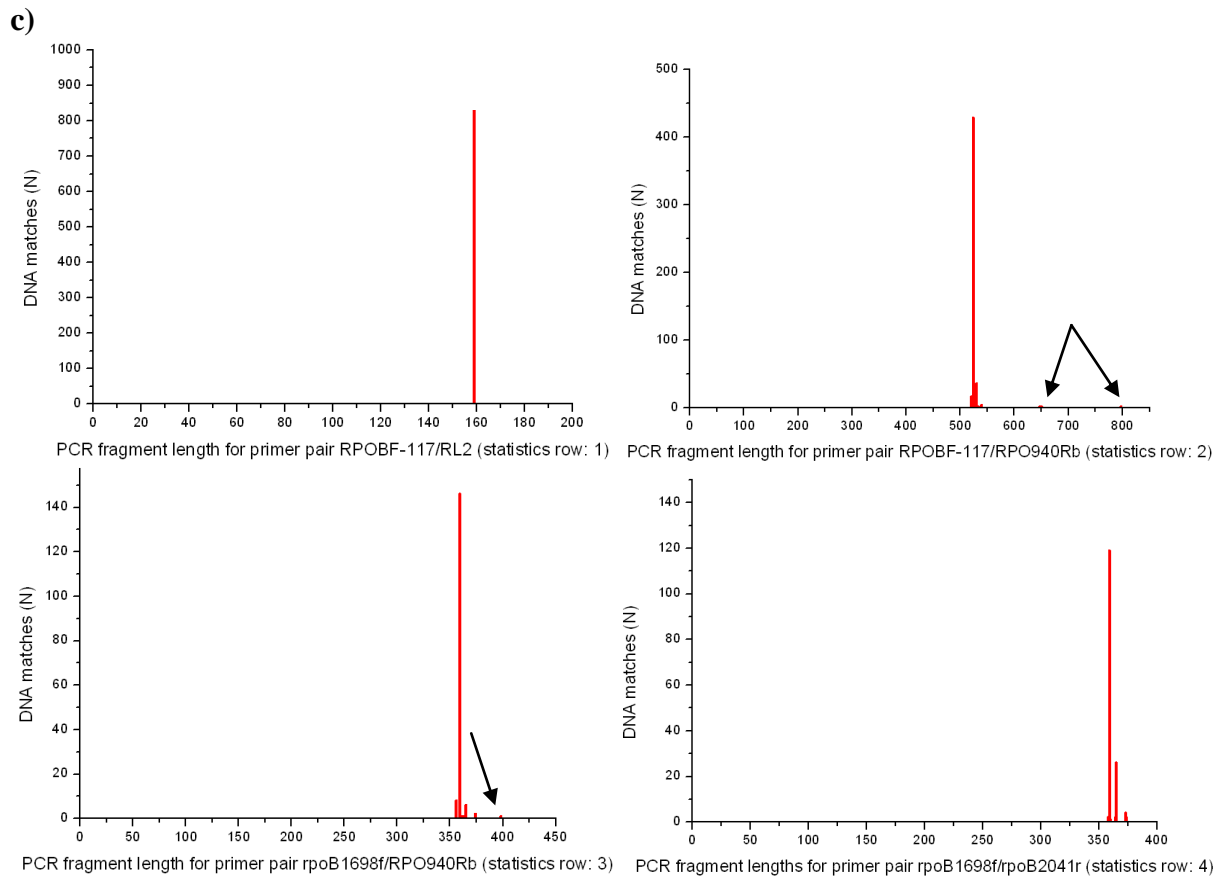


Figure 11: Distinct numbers and lengths of PCR fragments are detected by particular primer combinations (a) thus enabling one to map the PCR fragment length scatter (b, c, d) as a function of increasing sequence quality or decreasing number of allowed mismatches in consecutive analyses. Arrows (c) indicate less abundant fragments of markedly different lengths. Numerous distinct PCR fragment lengths can be detected by highly degenerate sampling procedure (d).

2.4 Cluster analysis of T-RFs: BEsTRF generated data as input files for other downstream programs (also for teaching purposes)

The above mentioned fragments histograms can be generated from various user-generated or theoretically defined sequence collections for cluster analysis either on forward or reverse terminal fragments. The next two results files *fwd_fragments.txt* and *rev_fragments.txt* contain organized fragment histograms generated from detected sequences by the restriction enzymes specified. Histograms generated for each enzyme are concatenated according to restriction enzyme alphabetical order and can therefore serve as input data for comparative tests of molecular fingerprinting approaches on real or simulated data. Numerous programs such as BioNumerics, Statistica, PCord, CANOCO, SAS, Mathlab, SPSS, R/project, BioPerl, root or other can be used for the purpose.

Five practical topics on the use of BEsTRF generated files in T-RFLP cluster analysis worth considering in research or teaching are listed below:

* Topic 1:

Test the effects of various reverse primers paired with only one forward primer, or conversely, test a multitude of forward primers in combination with only one reverse primer on distribution of peaks, their evenness, identify which combinations give the most similar results despite the intrinsic differences in amplicon length, sampling capacity, distribution of enzyme cutting sites, mismatches, primer degeneracy and are therefore more stable, and thus more comparable across various past studies as well as potentially more informative for future studies. You can use the following two results files *fwd_fragments.txt* and *rev_fragments.txt* for such analyses.

* Topic 2:

Identify primer pairs targeting particular gene that generate amplicons of sufficient length and resolution most suitable for simultaneous use in quantitative Real-Time PCR and for population profiling of quantified PCR products using T-RFLP on various genes. You can do this by analyzing files *enzyme_forward_results1.txt* or *enzyme_reverse_results2.txt*.

* Topic 3:

Compute T-RF histograms from defined artificial microbial communities (sequences downloaded from various databases, user defined model communities or research), import the following two results files *fwd_fragments.txt* and *rev_fragments.txt* into statistical packages and analyze them using various thresholds of detection, primer combinations, enzyme sets and compute the reliability of their clustering.

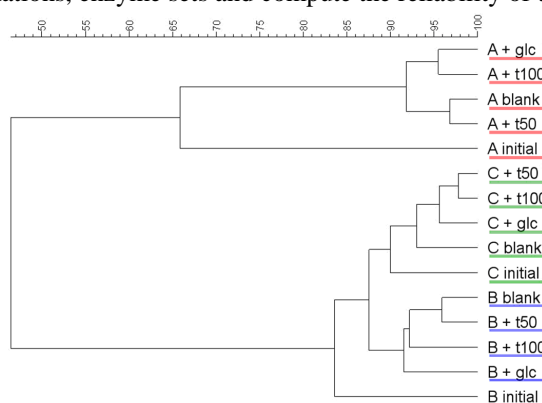


Figure 12: Final dendrogram of T-RFs generated by BioNumerics (Applied Maths, Belgium) from various user-defined theoretical communities. The horizontal axis indicates Pearson's similarity of T-RFs generated from these theoretical communities.

* Topic 4:

Compute the robustness of detection for each of the primer pairs using various degrees of allowed mismatches and other parameters. Perform cluster analyses using the following two results files *fwd_fragments.txt* and *rev_fragments.txt* to elucidate which primer combinations are stable or more prone to progressive loss of specificity as is often the case in PCR reactions using complex environmental DNA as template.

* Topic 5:

Form a simple artificial community or numerous communities from available microbial genomes as an approximation of the complexity present in the environmental DNA. Repeat the analyses and create T-RFs using various target genes, mismatches and other thresholds in order to compare the concordance of T-RF results generated from various complex model communities using various marker genes.

2.5 Phylogenetic analyses: BEsTRF generated data as input files for other downstream programs (also for teaching purposes)

The last sort of BEsTRF results files contain sequences collections generated by the primer pairs used in analysis. The file names are generated from the names of the forward and reverse primer pair (*accepted_forward*primer*name_reverse*primer*name.fa*) and can bear the extension *.fa, *.txt or any other (the contents of the files is the same in all cases) depending on the users' specifications in BEsTRF parameters file. Thus, user specifies whether the results file should bear the *.fa or *.txt extension, like:

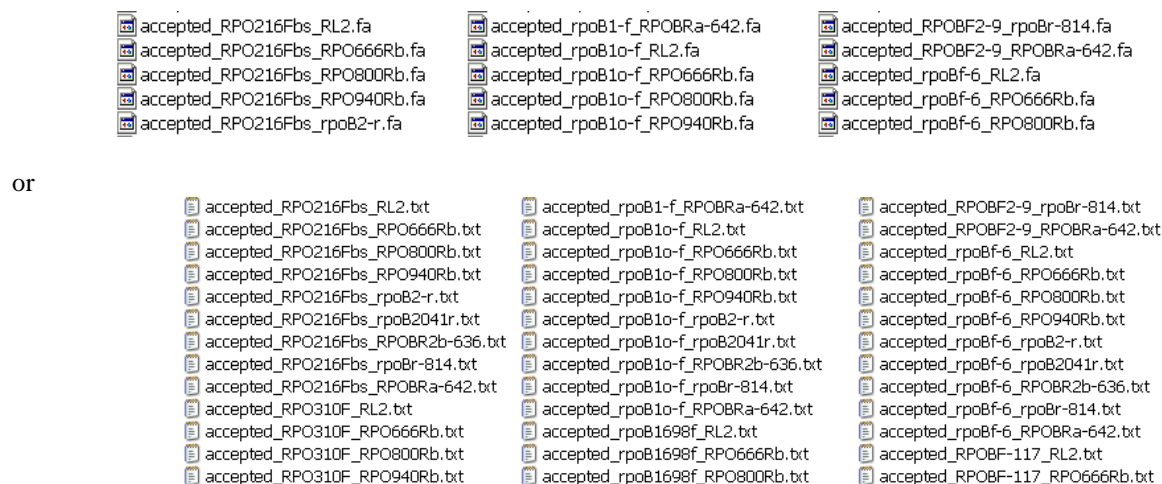


Figure 13: A typical view of numerous files containing sequences recognized by certain primer pairs. In any case *.fa or *.txt extensions can be generated according to user's naming convention, set in BEsTRF parameter file.

These files can serve as input files for programs involved in other downstream analyses for exploration of qualitative and quantitative differences among sequences detected and sampled from databases under different parameters and primer pairs:

- **ARB** (Ludwig *et al.*, 2004),
- **UniFrac** (Lozupone *et al.*, 2006),
- **DOTUR** (Schloss and Handelsman, 2005),
- **Libshuff** (Singleton *et al.*, 2001),
- **TreeClimber** (Schloss and Handelsman, 2006),
- **SONS** (Schloss and Handelsman, 2006),
- **Estimate S** (Colwell, 2005)
- **TRFMA** (Nakano *et al.*, 2006)

References:

Colwell, R. K. 2005. EstimateS: Statistical estimation of species richness and shared species from samples. Version 7.5. User's Guide and application published at: <http://purl.oclc.org/estimates>.

Lozupone, C. *et al.* (2006) UniFrac - An Online Tool for Comparing Microbial Community Diversity in a Phylogenetic Context. *BMC Bioinformatics* **7**, 371-385.

Ludwig, W. *et al.* (2004) ARB: a software environment for sequence data. *Nucleic Acids Res.*, **32**, 1363-1371.

Nakano, Y., *et al.* (2006) TRFMA: a web-based tool for terminal restriction fragment length polymorphism analysis based on molecular weight. *Bioinformatics*, **22**, 1788-1789.

Schloss, P.D., and Handelsman, J. (2005) Introducing DOTUR, a Computer Program for Defining Operational Taxonomic Units and Estimating Species Richness. *Appl. Environ. Microbiol.*, **71**, 1501-1506

Schloss, P.D., and Handelsman, J. (2006) Introducing TreeClimber, a Test To Compare Microbial Community Structures. *Appl. Environ. Microbiol.*, **72**, 2379-2384.

Schloss, P.D., and Handelsman, J. (2006) Introducing SONS, a Tool for Operational Taxonomic Unit-Based Comparisons of Microbial Community Memberships and Structures. *Appl. Environ. Microbiol.*, **72**, 6773-6779.

Singleton D.R. *et al.* (2001) Quantitative Comparisons of 16S rRNA Gene Sequence Libraries from Environmental Samples. *Appl. Environ. Microbiol.*, **67**, 4374-4376.

Five practical topics on usage of BEsTRF generated sequence datasets in downstream phylogenetic analyses are listed below:

*** Topic 1:**

The resulting sequence collections generated by numerous primer pairs or thresholds represent various subsamples of the original community and can thus serve as an additional control of T-RFLP fingerprinting. The statistical similarity of detected primer pair specific sequence collections can be tested using the above mentioned programs to see whether there are any significant differences among the collections due to sampling effort (primer mismatches allowed). In addition, a statistical test can be performed confronting T-RFLP clustering and sequence collection results.

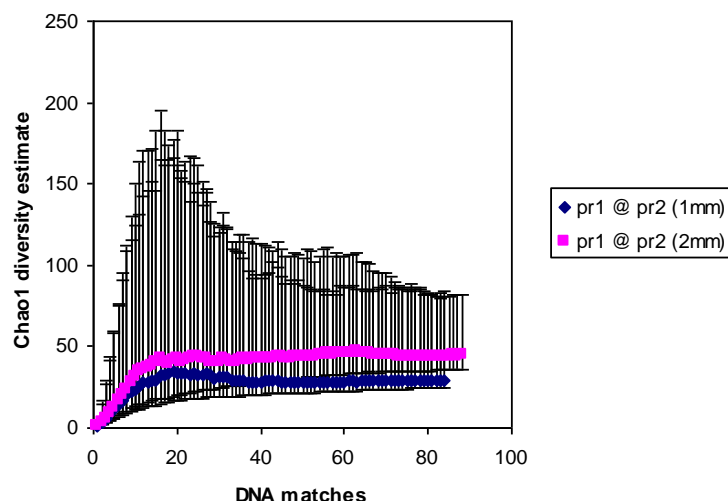


Figure 14: An example of statistical analysis of sequence hits detected by test pr1 @ pr2 primer set when one (1) or two (2) mismatches were allowed. This figure was generated from BEsTRF output sequence datasets detected by the two primers that served as input files for ARB fast distance matrix preparation. DOTUR was used for fast classification of sequences into operational taxonomic units (OTU) based on ARB distance matrix. OTU distribution was tested in EstimateS. The 95% confidence interval overlap indicated no significant difference in performance of this primer pair.

*** Topic 2:**

Sequence collections can be used to identify primer pairs detecting phylogenetically very closely related sequences as their calculated diversity is going to be low. These findings can be contrasted with the results of T-RFLP fingerprinting.

*** Topic 3:**

The number of Operational Taxonomic Units in sequence collections detected by each primer pair can be defined and statistically compared between pairs.

*** Topic 4:**

Various classical parametric or nonparametric ecological indices and estimates of diversity can be computed using sequence collections generated under various analysis parameters and thus reliability of community diversity estimation and prediction can be verified.

*Topic 5:

The researchers and students can get hands on experience about how detection and sampling biases affect the final outcomes of their analyses as shown below.

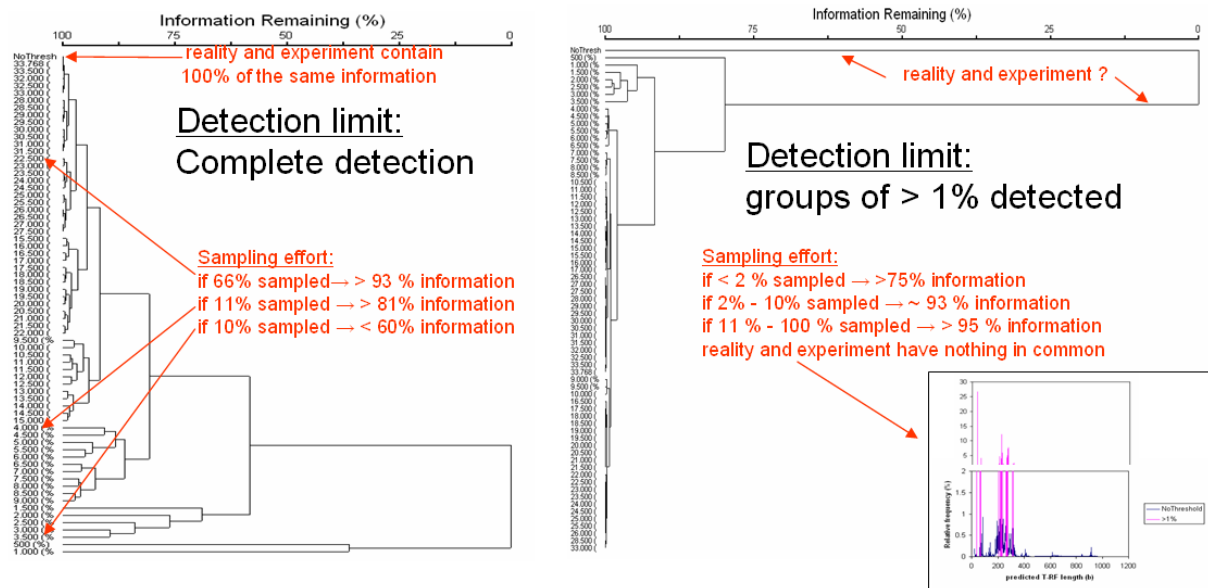


Figure 15: The influence of sampling effort and varying detection limit on the outcome of T-RFLP analyses comparing *in-silico* data with real model microbial community indicating the fraction of shared information between clone libraries of varying size and community T-RFLP. This figure was produced using PCord.

* Topic 6:

Using phylogenetic approaches in combination with BEsTRF a primer combination generating the highest sequence diversity can be identified from all those primer combinations found most suitable for generating optimal T-RFLP resolution.

Note:

*.fa files are plain ASCII (txt) files, which can be viewed or analyzed using plain text editors such as Notepad or WordPad under Windows, Open Office text editors or joe, gedit, Kate or other under Linux.

* Double clicking on a file in file browser does not open the *.fa file, since extension *.fa is not registered with operating system. Open text editor first, browse for your *.fa file and then open it through the text editor menu.

3. Troubleshooting

* If BEsTRF output on screen does not indicate normal flow of execution, please first check the messages in the command window (Figure 2, Figure 15). Errors and warnings due to misspelled file names, inappropriate file formats, identical primer or enzyme names with distinct recognition sites are pointed out to user.

```
D:\clanki tekst\bostjan proba grafi\BEsTRF folder za na pejdz\...
Number of Forward Primers in the dictionary: 11
      RL1      GAIGATATCGATCAYCTDGG
      RPO175F   gcactcatggggtcgaaacat
      RPO216F   atcgctcatcccggttcctg
      RPO216Fbs atcgctSatcccggttcctg
      RPO310F   GAGGGSATGAGICTSCAGGA
      rpoB1-f   ATTGACCACTTGGTAAACCGTGG
      rpoB169f  AACATCGCTTGGTCAAC
      rpoB10-f  ATCGATCACTTAGCAATCGTGG
      RPOBF-117 CACTTATGATGATCAG
      rpoBf-6   AGGTCACACTAGTTCAGTATGGACG
      RPOBF2-9  ACTCGTGAGCGTGCTGGTTTT

Isolated Forward Primers are not specified.
The analysis is going to be done on all Forward Primers in the dictionary.

Number of Reverse Primers in the dictionary: 9
      RL2      TTCUGGCGTTTCAATNGGAC
      RPO666Rb  gcttggtgggtaacctcgagg
      RPO800Rb  tcctggagactcatccgctc
      RPO940Rb  CnTTGCATGTTSGAGCCCAT
      rpoB2-r   ACGATCAGGGGTCAAACCCAC
      rpoB2041r CGTTGCATGTTGGTACCCAT
      rpoBr-814 GTCTACATTTGGCAAGATCGTATC
      RPOBR2b-636 ATCAAAGCCCGGTGAGCAT
      RPOBRa-642 GTHTGNCDTTGCATGTT

Isolated Reverse Primers are not specified.
The analysis is going to be done on all Reverse Primers in the dictionary.

***** WARNING: pattern: CG^CG with name: Bsh1236I already exists with name:
AccII <ignoring second definition>

***** WARNING: pattern: ^AATT with name: TspEI already exists with name: Sse
9I <ignoring second definition>

Number of Enzymes in the dictionary: 30
      AccII     CG^CG
      AclI      C^CGC
      AfaI      GT^AC
      AluI      AG^CT
      AspLEI    GCG^C
      BfaI      C^TAG
      BfuCI     ^GATC
      BshFI     GG^CC
      BsiSI     C^CGG
      BstKTI     GAT^C
      ChaI      GATC^
      Csp6I     G^TAC
      CuiAII    C^ATG
      CuiJI     RG^CY
      DpnI      GA^TC
      EsaBC3I   IC^GA
      FaeI      CATG^
      FstI      ^CATG
      G1aI      GC^GC
      Hin6I     G^CGC
      HpyCH410  A^CGT
      HpyCH40   IG^CA
      MseI      I^TAA
      PabI      GTA^C
      SclI      ^CGCG
      SetI      ASSI^
      Sse9I     ^AATT
      Sth302II  CC^GG
      Tail      ACGT^
      TaqI      I^CGA

Isolated Enzymes are not specified.
The analysis is going to be done on all Enzymes in the dictionary.

Opening DNA file: rpob.txt... OK.
DNA pattern 2643 in file rpob.txt is not valid.
Processed DNA sequences in file:          9163
Total so far processed DNA sequences: 9162

Generating report: Forward_Enzyme_Fragments... done

Generating report: Reverse_Enzyme_Fragments... done

Press any key to exit...
```

Figure 15: A typical BEsTRF output information on screen.

* If calculations of T-RFs does not work or results are wrong, please check input files according to the following tips:

1. Are parameters in BEsTRF parameter file set correctly and according to your needs?
2. Is the right sequence database chosen and specified in BEsTRF parameter file?
3. Are the right primer and enzyme dictionaries chosen? Have you checked their 5' orientation?
4. Are the correct output reports and files specified?

Correct potentially discovered misspecifications and restart BEsTRF.

* If there are no errors reported and large databases are being analyzed using numerous primers and enzymes, let the program run for a while as it takes a bit of a time to compute results for really large.

*.fa files are plain ASCII (txt) files, which can be viewed or analyzed using plain text editors such as Notepad or WordPad under Windows, Open Office text editors or joe, gedit, Kate or other under Linux.

* Double clicking on a file in file browser does not open the *.fa file, since extension *.fa is not registered with operating system. Open text editor first, browse for your *.fa file and then open it through the text editor menu.

* Input sequences are neither required to be aligned nor to be of the same length. However, it is recommended that the 5' orientation is maintained throughout the collection, although BEsTRF can be instructed to automatically reverse sequences.

* When aligned sequence databases or user collections are analyzed BEsTRF can handle – or ~ signs for gaps.

* The current version of BEsTRF was tested to take, as input, up to 50 forward and 50 reverse primers, but their number is not limited by the program. However, the correct primer orientation should be maintained throughout the dictionaries. Please, see also section **Computing time** and the BEsTRF parameter file for additional information.

* Checkup list before running BEsTRF: Make sure you formatted your sequence, primer and enzyme files appropriately, specified and typed the correct names of your primer and enzyme files into BEsTRF parameter file and set the analysis and output options to your specific needs.

* If you encounter difficulties displaying all information while importing BEsTRF results into spreadsheet program, please consider using a different spreadsheet program that supports larger number of columns or lines than for instance MS Excel does. User can choose between common graphic software as Statistica, SAS, Mathlab, SPSS, ORIGIN, R/project, BioPerl, gnuplot, root and other.

* Please report bugs and wishes to bestrf@lie.fe.uni-lj.si.

* The email bestrf@lie.fe.uni-lj.si is also intended for providing help and feedback, as well as receiving questions and comments regarding BEsTRF.

* Visit BEsTRF web site for updates, potential bug reports and novel worked examples.

4. Hints and tips

* A classification of restriction enzyme subtypes is available at [REBASE](http://rebase.neb.com/rebase/rebase.html) (<http://rebase.neb.com/rebase/rebase.html>) under REBASE Enzyme Sub Types.

* Make sure to check that the enzymes you intend to use with BEsTRF are in fact suitable for T-RFLP.

* There is a multitude of restriction enzymes that mostly belong to a class described as Type IIP enzymes. At present it is not possible to separate out all of those that are not or those that cut at unusual sites or require methylation for cleavage. Thus, such enzymes are included in the commercially available enzyme list even though they may not be optimal for T-RFLP.

* All commercially available Type IIP restriction endonucleases (i.e. not including Type IIA,B, etc.) can be obtained from [REBASE](http://rebase.neb.com/rebase/rebase.html) (<http://rebase.neb.com/rebase/rebase.html>). The easiest way to use a subset of the REBASE list is to open the REBASE plaintext file and remove the lines you do not want to include and save the file as a new enzyme dictionary.

* Please note that there are enzyme isoschizomers (different enzymes with identical cleavage site). Depending on your preferences you can use all of them, choose only some of them, or none. If BEsTRF encounters enzymes with identical restriction sites in user defined restriction enzyme collections, duplicated entries are filtered out and listed in a separate file *DuplicateEnzymes.txt* as described below:

```
+-----+
| Duplicate entries exist in dictionary Enzymes |
+-----+

Duplicate entries do not take part in the analysis since
each of them produces the same results.
Results of fictitious processing of duplicates can be found
in reports under surrogate names according to the following table.

Duplicate name   Name under which it is referred to in reports
-----
Bsh1236I        AccII
BsnI            BshFI
Bsp143I         BfUCI
BspACI          AciI
BspANI          BshFI
BssMI           BfUCI
BstFNI          AccII
BstHHI          AspLEI
BstMBI          BfUCI
BstUI           AccII
BsuRI           BshFI

-----
The above information is summarized in the following alternative way
-----

Pattern AccII has 4 duplicate entries.
The following duplicates are included in reports under name AccII:
Bsh1236I, BstFNI, BstUI, MvnI.

Pattern AciI has 2 duplicate entries.
The following duplicates are included in reports under name AciI:
BspACI, SsiI.

Pattern AfaI has duplicate entry.
Duplicate RsaI is included in reports under name AfaI.

Pattern AspLEI has 3 duplicate entries.
The following duplicates are included in reports under name AspLEI:
BstHHI, CfoI, HhaI.
```

Figure 16 : Description of duplicated entries that are filtered out and alphabetically listed in a separate file *DuplicateEnzymes.txt*.

* Decreasing the stringency of recognition sites (many mismatches allowed; use of sequences containing degenerated code; highly degenerated primers) may produce spurious and in some extreme cases also completely wrong results. The number of acceptable mismatches allowed should thus be carefully explored. BEsTRF enables all these possibilities but you must understand them and be aware of what you are doing to obtain meaningful results.

* Carefully explore the length distribution of your newly generated T-RFs to ensure their equal distribution in the capillary sequencer output chromatograms (electropherograms)

* Publicly available sequence databases are normally asymmetric as they contain sequences of unequal length. In addition, some sequences had their primer sites removed before submission and the quality of the deposited sequences may not be uniform.

5. Contents of README.txt (obtainable at BEsTRF web site and presented here for your convenience)

BEsTRF.exe (Windows) or bestrf (Linux) is an executable file for T-RFLP space exploration which finds a set of primers with highest sampling capacity, generating highest number of peaks after restriction enzyme digestion resulting in highest T-RFLP resolution based on input sequences.

```
*****
;
;
; Running the application
;
*****
;
; How to run the program in MS(R) Windows(TM) and Linux(R)?
;
; A. Download the appropriate ZIP archive.
;
; B. Unpack the ZIP archive into the folder/directory
;    of your choice.
;
; C. Have your FASTA sequences at hand (on your local disk).
;
; D. Prepare BEsTRF parameters file (like this one; see ONLINE PROTOCOL FOR BEsTRF) according to your
;    specifications and desires.
;
;
;
;
; ***** Windows specific (the first way) *****
;
; E. Double click on the EXE file to start the program.
;
; F. When the program asks you for the name of the file with
;    parameters (the one that you have prepared in step "D"),
;    properly satisfy its curiosity.
;
; G. Examine the screen for potential error messages (see below
;    for a realistic scenario).
;
; H. Grab a cup of coffee or some lunch, depending on an imposed
;    workload.
;
; I. Use the results of the analysis in whatever way you like.
;
; ***** End of Windows specific (the first way) *****
;
;
;
; The just described procedure may work satisfactory or not.
; The beauty of the approach is its simplicity. The drawback
; is the fact that it is tough to inspect the output on the
; screen for potential error and warning messages as well as
; to check that the actual values of parameters are in
; accordance with your expectations. So, here is the second
; way to make the program run.
;
;
;
;
;
```



```

; In Linux you usually do not have the possibility of the first
; approach that we described for Windows because graphics
; user interfaces do not let you double click on an executable
; to run it. Instead, you must resort to the analogy of the
; second approach.
;
;
; ***** Linux specific *****
;
; E. Open "Terminal Window" or "Linux Console" into the directory
; with program executable. Do this by selecting the appropriate
; choice in a menu of your GUI "Linux explorer"
; (like Konqueror if you are using KDE(R) desktop environment).
; Note that in Linux it is usually possible by default to open
; the terminal window directly into the directory of
; your choice.
;
; F. Start the program by entering the following command:
;
; ./BEsTRF_version your_parameter_file.txt >screen_output.txt
;
; ...note the sequence of characters "./" at the beginning
; of the line, which you must not forget;
;
; ...again, note the character ">", which you must not forget.
;
; ***** End of Linux specific *****
;
; Everything about screen redirection and unattended executing
; that we mentioned above applies equally to Linux too.
;
; Alternatively, you can copy program executable into some
; directory on the PATH (usually, a suitable choice
; is "/usr/local/bin"), so that you can execute it from any
; directory and without the annoying sequence "./".
;
;

```

6. Contents of README_first.txt (for bacterial and archaeal analyses using files presented in Worked examples)

This example is provided without underlying sequences database(s), which has/have to be downloaded separately from the following links:

1. Ribosomal Database Project II at
<http://rdp.cme.msu.edu/misc/resources.jsp;jsessionid=81D78ABD098F7AF5EA2CC102FFAC0F4C>
2. Greengenes at
<http://greengenes.lbl.gov/Download>
3. ARB - silva at
<http://www.arb-silva.de/download>
4. other potential resources

You can run this analysis yourself by simple modification of one line in BESTRF parameter file(s) of your choice as described below. Other parameters are already set for ten distinct analyses for you. Thus you have ten distinct BESTRF parameter files at hand that differ in the number of allowed primer mismatches (1mm, 2mm, 3mm), insertions and deletions (Levensthein (L111)) as well as recognition of degenerated primer binding sites (4MPD).

Therefore:

»0mm« in parameters' file names mean: no allowed mismatches

»1mm«, »2mm«, »3mm in parameters' file names mean: 1, 2 or 3 allowed mismatches, respectively,

»4MPD« in parameters' file names mean: allow binding of primer to degenerated sites of any (up to level four) degeneracy,

»L111« in parameters' file names indicates usage of Levenshtein distance where mismatches, insertions and deletions all cost one quantum of penalty, and file name part »1mm«, »2mm«, »3mm« indicates total allowed mismatching cost.

Create a new directory on your disc and unzip the downloaded sequence database contents. The downloaded database name should now read like the following, depending on the database release:

release10_7_arch_aligned.fa

Optionally, you can rename the downloaded and unzipped sequence database to a name of your preference.

Open all or selected BESTRF parameter files that are provided for this demonstration in your preferred text (ASCII) editor. The names of these files start with params_....txt . In the first line there is a keyword "DNA_File_Names" after which type in the proper sequences input file name. The modified line of BESTRF parameter file should now read like this:

DNA_File_Names release10_7_arch_aligned.fa

In case you decided to change the name of sequence database to e.g. myBacteria.fa the line should read like this:

```
DNA_File_Names myBacteria.fa
```

Now save the newly modified BESTRF parameter file (under the same name if you intend to run BESTRF through already prepared RunME_xxxxxxx.bat files, or alternatively you can save the file under new name and fire up BESTRF manually).

Run demonstration(s) by double clicking on the appropriate file(s)
RunMe_xxxxxxx.bat

Please read section 1.7 Computation time on page 8 of Users' manual.