Mycorrhiza (2007) 17:133–136
DOI 10.1007/s00572-006-0102-5

TECHNICAL NOTE

# FragMatch—a program for the analysis of DNA fragment data

T. A. Saari · S. K. Saari · C. D. Campbell ·
I. J Alexander · I. C. Anderson

Received: 4 September 2006 / Accepted: 14 December 2006 / Published online: 17 January 2007

**Abstract** FragMatch is a user-friendly Java-supported program that automates the identification of taxa present in mixed samples by comparing community DNA fragment data against a database of reference patterns for known species. The program has a user-friendly Windows interface and was primarily designed for the analysis of fragment data derived from terminal restriction fragment length polymorphism analysis of ectomycorrhizal fungal communities, but may be adapted for other applications such as microsatellite analyses. The program uses a simple algorithm to check for the presence of reference fragments within sample files that can be directly imported, and the results appear in a clear summary table that also details the parameters that were used for the analysis. This program is significantly more flexible than earlier programs designed for matching RFLP patterns as it allows default or user-defined parameters to be used in the analysis and has an unlimited database size in terms of both the number of reference species/individuals and the number of diagnostic fragments per database entry. Although the program has been developed with mycorrhizal fungi in mind, it can be used to analyse any DNA fragment data regardless of biological origin. FragMatch, along with a full description and users guide, is freely available to download from the Aberdeen Mycorrhiza Group web page (http://www.aberdeenmycorrhizas.com).

**Keywords** T-RFLP · Microsatellite · Ectomycorrhiza · Fungal communities · Species identification

## Introduction

In recent years, the importance of investigating the below-ground diversity of ectomycorrhizal (EcM) communities has been recognised (Horton and Bruns 2001), and many field-based studies now involve analysis of the ectomycorrhizas themselves or their external mycelium rather than relying on above-ground fruiting structures. The application of molecular community profiling techniques has been crucial in advancing our understanding of the diversity and ecology of EcM mycelial communities in the field (e.g., Dickie et al. 2002; Genney et al. 2006). At present, terminal restriction fragment length polymorphism (T-RFLP) is emerging as the method of choice for such analyses from bulk root tip and soil samples (Anderson 2006). However, the identification of individual species in a mixed EcM community using this approach relies on the initial development of a T-RFLP database containing terminal fragment sizes generated for reference sporocarps or EcM root tips. The importance of developing a reference database containing fungi from the field site where the research takes place has been recently highlighted (Avis et al. 2006).

Whereas T-RFLP is fast and relatively inexpensive, the data analysis can be extremely time consuming, particularly given that sampling designs are getting more ambitious as large sample numbers are required to ensure that data are representative and ecologically relevant (Taylor 2002; Lilleskov et al. 2004). Therefore, to make T-RFLP more

T. A. Saari
Helsinki University of Technology,
P.O. Box 1000, 02015 Espoo, Finland

S. K. Saari · C. D. Campbell · I. C. Anderson (✉)
The Macaulay Institute,
Craigiebuckler, Aberdeen AB15 8QH, UK
e-mail: i.anderson@macaulay.ac.uk

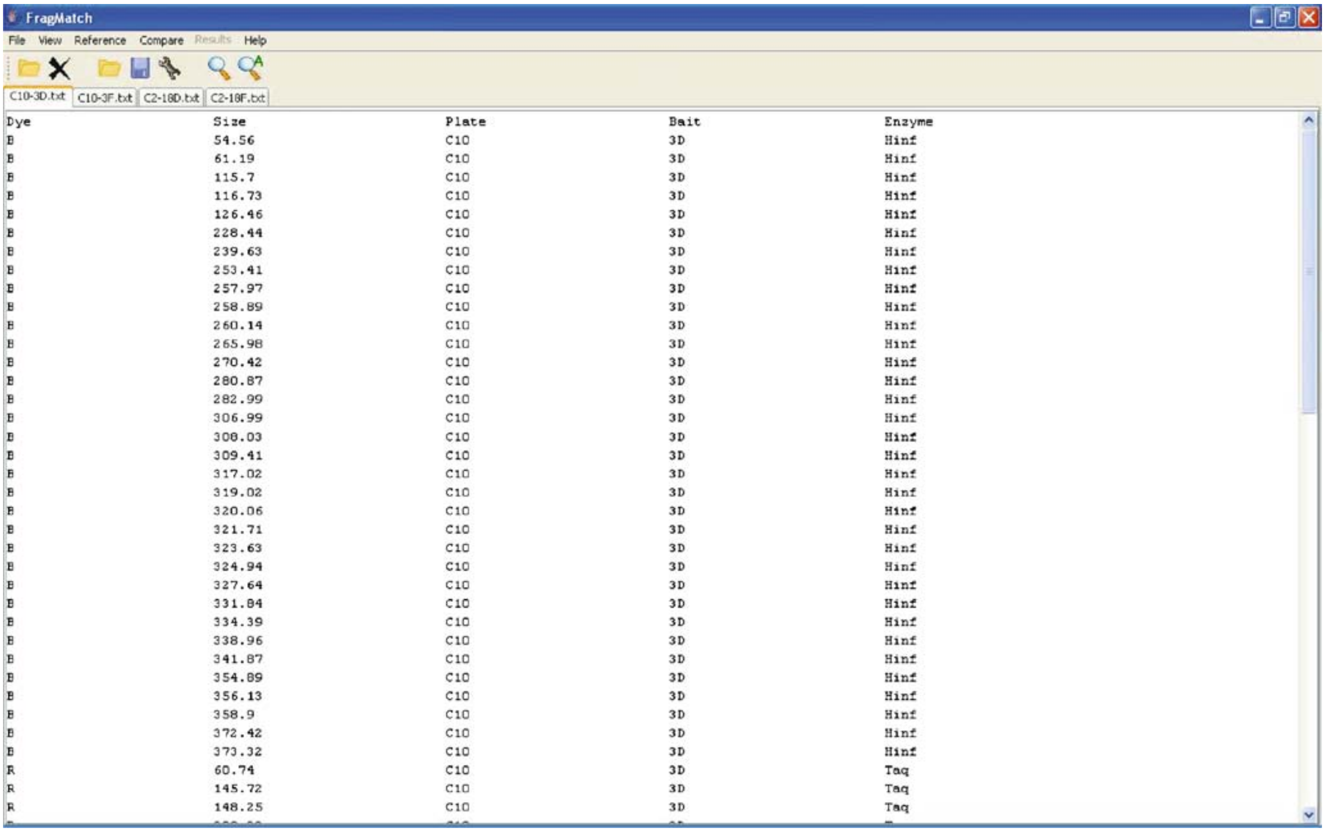
S. K. Saari · I. J. Alexander
School of Biological Sciences, Plant and Soil Science,
University of Aberdeen,
St. Machar Drive,
Aberdeen AB24 3UU, UK

viable for large field studies, the data analysis process needs to be automated. The first steps at automating the identification of EcM species from community T-RFLP profiles were made by Dickie et al. (2002) with their T-RFLP analysis matching program (TRAMP) program based on Visual Basic for Applications (VBA) macros in Excel (Microsoft, Redmond, WA, USA). Dickie et al. (2003) also developed another VBA-macro-based Excel spreadsheet program, the good-enough RFLP matcher (GERM), for identifying EcM fungi based on RFLP patterns. Although these programs produce reliable results, they are restricted by the number of enzymes/primers that can be used in any analysis and the number of fragment values that can be entered per enzyme/primer. In addition, only one mixed community sample can be analysed and compared against the reference database at any one time. These restrictions make the programs more difficult to use, particularly when the reference data contain multiple restriction enzymes per database entry.

In this short paper, we present a user-friendly independent Java-supported program, FragMatch, designed for use with Windows with a graphical user interface with Windows "look and feel", but that is equally operable on Unix and Macintosh systems. Operation of the program requires a Java2 version 1.5 runtime environment (freely available from Sun Microsystems) and a PC with 20 MB available system memory. The program is routinely used in our group to locate known EcM terminal restriction fragments (TRFs) in the fungal community T-RFLP profiles and has also been used to analyse microsatellite data that contain several fragment values per primer (e.g., Saari et al. 2005). The program, along with a full description and a user guide, is freely available to download from the Aberdeen Mycorrhiza Group web page (http://www.aberdeenmycorrhizas.com).

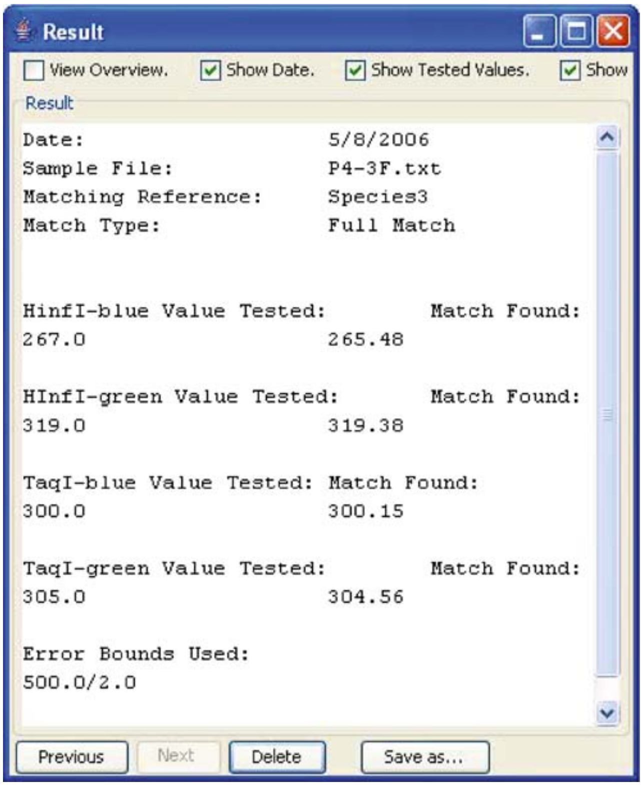
## Materials and methods

The program is designed to take sample files in tab-delimited text (.txt) format (Fig. 1) that can be directly exported from sequence analysis software (e.g., ABI GeneMapper software, Applied Biosystems, Warrington, UK). Care must be taken to ensure that all information relating to one sample is contained within a single file. For example, fragment data from different primers/enzymes must be found within the same input file and data from different primers/enzymes must be clearly labelled. If required, sample files can be modified in either Notepad or Excel before performing the analysis. The sample files may contain any amount of additional output data columns, such as peak height or peak area, as the data columns to be



| Dye | Size | Plate | Bait | Enzyme |
|---|---|---|---|---|
| B | 54.56 | C10 | 3D | Hinf |
| B | 61.19 | C10 | 3D | Hinf |
| B | 115.7 | C10 | 3D | Hinf |
| B | 116.73 | C10 | 3D | Hinf |
| B | 126.46 | C10 | 3D | Hinf |
| B | 228.44 | C10 | 3D | Hinf |
| B | 239.63 | C10 | 3D | Hinf |
| B | 253.41 | C10 | 3D | Hinf |
| B | 257.97 | C10 | 3D | Hinf |
| B | 258.89 | C10 | 3D | Hinf |
| B | 260.14 | C10 | 3D | Hinf |
| B | 265.98 | C10 | 3D | Hinf |
| B | 270.42 | C10 | 3D | Hinf |
| B | 280.87 | C10 | 3D | Hinf |
| B | 282.99 | C10 | 3D | Hinf |
| B | 306.99 | C10 | 3D | Hinf |
| B | 308.03 | C10 | 3D | Hinf |
| B | 309.41 | C10 | 3D | Hinf |
| B | 317.02 | C10 | 3D | Hinf |
| B | 319.02 | C10 | 3D | Hinf |
| B | 320.06 | C10 | 3D | Hinf |
| B | 321.71 | C10 | 3D | Hinf |
| B | 323.63 | C10 | 3D | Hinf |
| B | 324.94 | C10 | 3D | Hinf |
| B | 327.64 | C10 | 3D | Hinf |
| B | 331.84 | C10 | 3D | Hinf |
| B | 334.39 | C10 | 3D | Hinf |
| B | 338.96 | C10 | 3D | Hinf |
| B | 341.87 | C10 | 3D | Hinf |
| B | 354.89 | C10 | 3D | Hinf |
| B | 356.13 | C10 | 3D | Hinf |
| B | 358.9 | C10 | 3D | Hinf |
| B | 372.42 | C10 | 3D | Hinf |
| B | 373.32 | C10 | 3D | Hinf |
| R | 60.74 | C10 | 3D | Taq |
| R | 145.72 | C10 | 3D | Taq |
| R | 148.25 | C10 | 3D | Taq |

**Fig. 1** Main program window showing four sample input files as tabs



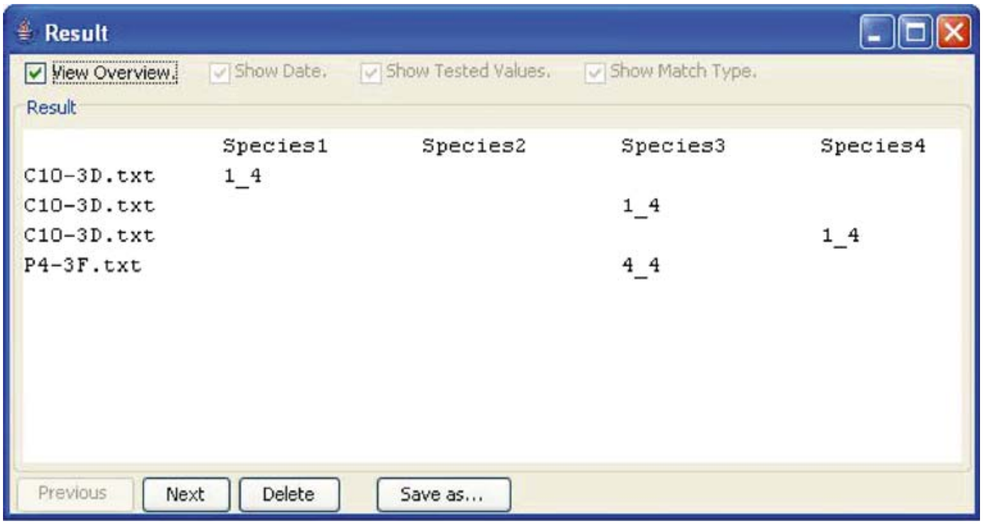
☐ View Overview. ☑ Show Date. ☑ Show Tested Values. ☑ Show

```
Date:                      5/8/2006
Sample File:               P4-3F.txt
Matching Reference:        Species3
Match Type:                Full Match


HinfI-blue Value Tested:           Match Found:
267.0                      265.48

HInfI-green Value Tested:          Match Found:
319.0                      319.38

TaqI-blue Value Tested:  Match Found:
300.0                      300.15

TaqI-green Value Tested:           Match Found:
305.0                      304.56

Error Bounds Used:
500.0/2.0
```

**Fig. 2** Detailed results window showing an example of the output for a full match to reference database entry Species3 in the sample input file P4-3F.txt

used in the analysis are defined by the user. In addition, the number and names of primers/enzymes are user-defined, and there is no limit to the number of primers/enzymes within reference units of a database. Each primer/enzyme may contain either none, one, or several reference values. Apart from the computer memory capacity, there are no limits to the number of reference units per database. Nevertheless, all reference units within a single database must contain data for the same set of primers/enzymes.

The program uses a simple algorithm to match unknown fragment patterns to a set of reference fragment patterns resulting from T-RFLP, microsatellite PCR, or a similar analysis that produces distinctive and reproducible DNA fragment patterns. The iterative calculations used by the algorithm are outlined as follows: {min=[(current reference value)−(reference error bound)], max=[(current reference value)+(reference error bound)], whereas (sample has values or match found) {if (min≤current sample value≤max) {*match found* break loop}}}. "Reference set"=a database of known fragments for, e.g., species, individuals, etc.; "reference unit"=a known individual/species within the reference set, "reference value"=a DNA fragment value generated by a particular enzyme/primer within the full fingerprint for the reference unit, "sample value"=a DNA fragment value generated by a particular enzyme/primer within the sample file, "error bound"=user-defined variation in the sizing of the DNA fragment values (in base pairs). In summary, the algorithm starts with the first reference value within the first reference unit in the active reference set and checks for a match through all sample values within a sample file, and if a match (within the error bounds) is detected or if no match is found when all sample values are checked, the next reference value within the reference unit is checked. The manufacturers of DNA sequencers usually list the accuracy of fragment sizing (in base pairs) for any given fragment size class, and this information, or user-defined error margins, may be directly used in determining reference error bounds. The algorithm keeps track of how many full matches (i.e., detection of all reference values) to reference units were found, and it also notifies whether there were any partial matches to reference units (i.e., detection of only a subset of the reference values for a database entry). Once the algorithm has run through all primers/enzymes within a reference unit and the analysis is complete, a results window is displayed, detailing all



☑ View Overview. ☑ Show Date. ☑ Show Tested Values. ☑ Show Match Type.

```
               Species1       Species2       Species3       Species4
C10-3D.txt     1_4
C10-3D.txt                                   1_4
C10-3D.txt                                                  1_4
P4-3F.txt                                    4_4
```

**Fig. 3** Overview results window showing results for sample files P4-3F.txt and C10-3D.txt. A full match for sample P4-3F.txt to Species3 is reported as 4_4, demonstrating that all four reference fragments for Species3 were found within the sample file. Because the analysis parameters were set to also show partial matches, 1_4 is shown for Species1, 3, and 4 for sample file C10-3D.txt, demonstrating that one out of four possible fragments was matched from the sample file to each of these reference units

reference unit values used in the analysis along with the corresponding matched sample values (Fig. 2). Apart from this detailed output window, the results can also be viewed in an overview table where all the analysed sample files appear stacked in a column to the left, with all the reference units in subsequent separate columns to the right (Fig. 3). These columns contain a summary of how many fragments from the sample file matched to the reference unit out of all fragments contained in that reference unit, e.g., “3_4” means that a match was found within the input sample file for three out of four possible reference values. Both types of result displays can be saved as text files and subsequently viewed or modified in Excel or Notepad. The user can define how many, if any, partial matches are reported by the program. An unlimited number of sample files may be analysed together using the “compare-all” function.

## Discussion

Although the idea behind FragMatch is similar to that in TRAMP (Dickie et al. 2002), FragMatch is designed to be more user friendly and more flexible so that the analysis parameters can be adapted for different types of data according to the users specifications. There are no limits to the number of enzymes/primers used or the number of fragment values per enzyme/primer. The ability to include any number of enzymes per database unit is an important feature, as it is common practice to perform restriction digests with up to three or four enzymes per sample for robust identification of EcM fungi (e.g., Avis et al. 2006; Burke et al. 2005). This flexibility is even more important for microsatellite analyses where the diagnostic identification of genetic individuals of EcM fungi can require the analysis of numerous loci (e.g., Kretzer et al. 2003; Wu et al. 2005). The number of reference units in a database is also unlimited and a large number of samples may be analysed in a single run using the “compare all” function, thus, speeding up the analysis. These features are a clear improvement on the TRAMP and GERM programs (Dickie et al. 2002, 2003) and are designed to permit the rapid and flexible analysis of large data sets.

When using FragMatch for identifying EcM species by matching community T-RFLP data against a T-RFLP reference database, the accuracy of the matches are obviously only as good as the accuracy of the reference data in the database. Therefore, the quality of the reference data is extremely important. In addition, the usual precautions that have to be taken when identifying EcM species based on restriction fragments, including controlling for intraspecific variation, should be taken into account when using FragMatch. FragMatch output files record the accuracy of matches for all reference values within each reference unit so the user can check the data for poor matches and identify samples that may need to be re-run. This gives more confidence in the data, and the resultant output files can be easily sorted to prepare frequency distribution graphs of species abundance in the sample set, for example.

Techniques that require the investigator to match DNA fragments produced on a gel or a sequencer against a reference database are becoming increasingly used in mycorrhizal community and population ecology research. FragMatch has been specifically designed as a user-friendly, flexible program to aid in the analysis of such data sets and to overcome some of the limitations of currently available programs.

**Acknowledgment** The authors would like to thank Helsinki University of Technology for providing funding to T. A. Saari. The Macaulay Institute receives funding from the Scottish Executive Environment and Rural Affairs Department. Caroline Reiff, David Genney, and Pamela Parkin are thanked for their help with testing FragMatch and for their useful comments.

## References

Anderson IC (2006) Molecular ecology of ectomycorrhizal fungal communities: new frontiers. In: Cooper JE, Rao JR (eds) Molecular approaches to soil, rhizosphere and plant microorganism analysis. CABI, Wallingford, UK, pp. 183–197

Avis PG, Dickie IA, Mueller GM (2006) A “dirty” business: testing the limitations of terminal restriction fragment length polymorphism (TRFLP) analysis of soil fungi. Mol Ecol 15:873–882

Burke DJ, Martin KJ, Rygiewiczc PT, Topa MA (2005) Ectomycorrhizal fungi identification in single and pooled root samples: terminal restriction fragment length polymorphism (TRFLP) and morphotyping compared. Soil Biol Biochem 37:1683–1694

Dickie IA, Xu B, Koide RT (2002) Vertical niche differentiation of ectomycorrhizal hyphae in soil as shown by T-RFLP analysis. New Phytol 156:527–535

Dickie IA, Avis PG, McLaughlin DJ, Reich PB (2003) Good-enough RFLP matcher (GERM) program. Mycorrhiza 13:171–172

Genney DR, Anderson IC, Alexander IJ (2006) Fine-scale distribution of pine ectomycorrhizas and their extramatrical mycelium. New Phytol 170:381–390

Horton TR, Bruns TD (2001) The molecular revolution in ectomycorrhizal ecology: peeking into the black-box. Mol Ecol 10:1855–1871

Kretzer AM, Dunham S, Molina R, Spatafora JW (2003) Microsatellite markers reveal the below ground distribution of genets in two species of *Rhizopogon* forming tuberculate ectomycorrhizas on Douglas fir. New Phytol 161:313–320

Lilleskov EA, Bruns TD, Horton TR, Taylor DL, Grogan P (2004) Detection of forest stand-level spatial structure in ectomycorrhizal fungal communities. FEMS Microbiol Ecol 49:319–332

Saari SK, Campbell CD, Rusell J, Alexander IJ, Anderson IC (2005) Pine microsatellite markers allow roots and ectomycorrhizas to be linked to individual trees. New Phytol 165:295–304

Taylor AFS (2002) Fungal diversity in ectomycorrhizal communities: sampling effort and species detection. Plant Soil 244:19–28

Wu B, Nara K, Hogetsu T (2005) Genetic structure of *Cenococcum geophilum* populations in primary successional volcanic deserts on Mount Fuji as revealed by microsatellite markers. New Phytol 165:285–293