BIOINFORMATICS APPLICATIONS NOTE
Vol. 17 no. 12 2001
Pages 1230–1231

Sequence type analysis and recombinational tests (START)

K. A. Jolley*, E. J. Feil, M.-S. Chan and M. C. J. Maiden

Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK

Received on April 23, 2001; revised and accepted on July 19, 2001

ABSTRACT

Summary: The 32-bit Windows application START is implemented using Visual Basic and C^{++} and performs analyses to aid in the investigation of bacterial population structure using multilocus sequence data. These analyses include data summary, lineage assignment, and tests for recombination and selection.

Availability: START is available at http://outbreak.ceid.ox.ac.uk/software.htm.

Contact: keith.jolley@ceid.ox.ac.uk

Multilocus Sequence Typing (MLST) is a nucleotide sequence-based typing method that indexes the variation present in bacterial housekeeping genes, where most of the variation is selectively neutral (Maiden *et al.*, 1998). Internal fragments of seven housekeeping genes, approximately 450–500 bp in length, are sequenced and novel alleles are assigned with arbitrary numbers sequentially to provide an allelic profile of seven integers that defines the Sequence Type (ST) of each isolate. The technique is designed primarily for global or long-term epidemiology and surveillance, and has the advantage over other typing methods, such as genetic fingerprinting, of electronic portability and unambiguous characterization of isolates. MLST schemes have been developed for a range of bacterial pathogens and databases for these organisms can be interrogated at the MLST web-site (http://www.mlst.net/) thus facilitating rapid comparisons of isolates typed using the method. A further advantage of MLST is that it provides large quantities of data that may be analyzed by a number of evolutionary approaches to yield insights into the structure of bacterial populations and the selective pressures which act upon them.

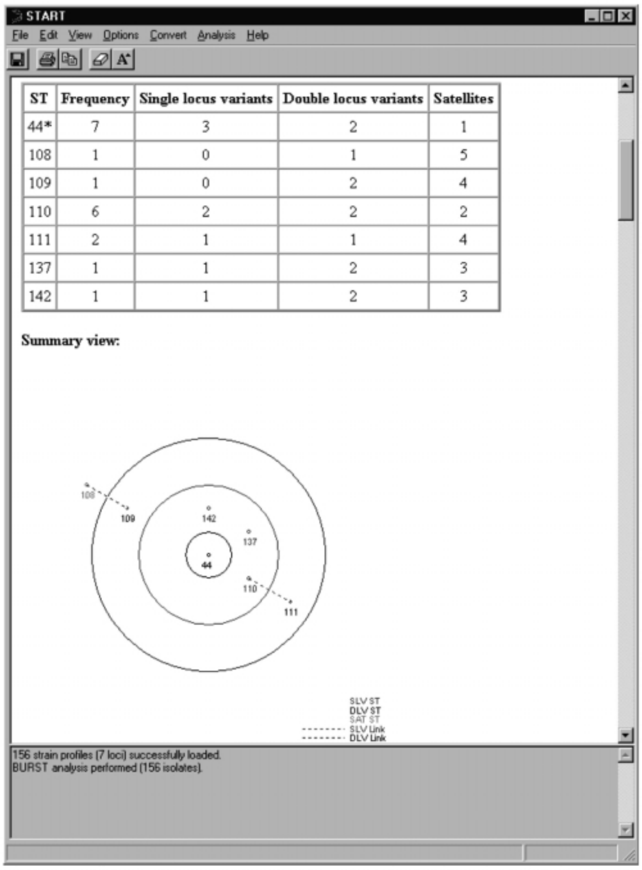
With the increasing availability of MLST data, the need for software to describe and analyze datasets has become apparent. Sequence Type Analysis and Recombinational Tests (START) was written to address this need through the inclusion of multiple analytical techniques in an easy-to-use and intuitive interface for Windows 95/98/NT/2000 operating systems.

Techniques available within the START program are divided into four categories: data summary, lineage assignment, tests for recombination and tests for selection. Two input files are required for many of the tests—allelic profiles, consisting of isolate identifiers and allele numbers, and allele sequences. Profile data can be entered into the program directly from the keyboard, by pasting from the clipboard or by loading a tab-delimited text file while allele sequences need to be in FASTA format. The program utilizes an embedded web-browser for output, enabling easy formatting of tables and inclusion of diagrams generated by the lineage assignment algorithms using HTML, as well as printing and saving of results. Graphical output from analyses is produced in the form of Windows Metafiles (WMF) embedded within the page and these may be saved for manipulation within a graphics package. With all tests it is possible to select subsets of isolates to analyze.

There are five data summary functions available within START: allele and profile frequency functions display the relative abundance of each allele sequence or ST within the dataset; the polymorphism frequency function produces a gene sequence and table showing the positions of all polymorphic sites within the dataset and where these are unique highlight the corresponding isolate name and/or allele number; and the codon usage and GC content functions produce appropriate frequency tables broken down by locus.

To aid in the assignment of STs to lineages, BURST (Feil, in preparation) and UPGMA methods are implemented along with a function to create a distance matrix. BURST is a clustering algorithm designed for use on microbial MLST data which examines the relationships within clonal complexes where isolates are grouped based on the number of locus differences within their profiles. A putative founder genotype may be identified based on its number of single- and double-locus variants and a summary graphical representation displayed. Figure 1 is part of the output obtained from the analysis of 156 MLST profiles using the housekeeping genes *abcZ*, *adk*, *aroE*, *fumC*, *gdh*, *pdhC* and *pgm*, from a carriage population of *Neisseria meningitidis* (Jolley *et al.*, 2000). This

*To whom correspondence should be addressed.



ST	Frequency	Single locus variants	Double locus variants	Satellites
44*	7	3	2	1
108	1	0	1	5
109	1	0	2	4
110	6	2	2	2
111	2	1	1	4
137	1	1	2	3
142	1	1	2	3

Summary view:

```
156 strain profiles (7 loci) successfully loaded.
BURST analysis performed (156 isolates).
```

Fig. 1. BURST analysis in START showing one of the clonal groupings obtained from a carriage sample of *N.meningitidis*. The group comprises 19 isolates with seven unique STs centred around ST-44. The three STs within the inner ring of the diagram are single-locus variants of ST-44, while those in the outer ring are double-locus variants.

shows one of twelve clonal complexes identified by the algorithm, grouped around a recognized hyper-invasive genotype, ST-44, and the inter-relationships within the complex. These functions can also be used to estimate recombinational parameters (Feil *et al.*, 2001).

START includes a number of tests which can be used to investigate the extent and significance of recombination. These are the Sawyer's Runs Test (Sawyer, 1989), the Maximum Chi-Squared (χ^2) Test (Maynard-Smith, 1992), the Homoplasy Test (Maynard-Smith and Smith, 1998) and the Index of Association (I_A) (Maynard-Smith *et al.*, 1993).

The ratio of non-synonymous (d_N) to synonymous (d_S) substitutions per nucleotide site is an indicator of the kind of selective pressure acting on a gene as a whole. START uses the method of Nei and Gojobori (1986) to estimate these parameters, providing values for each locus in the dataset.

The package integrates these methods for analysis of MLST datasets and includes full on-line help and example data.

ACKNOWLEDGEMENTS

K.A.J., E.J.F. and M-S.C. are supported by the Wellcome Trust. We thank John Maynard-Smith for the code used in the I_A and Homoplasy Tests. MM is a Wellcome Trust Senior Fellow in Biodiversity Research.

REFERENCES

Feil,E.J., Holmes,E.C., Bessen,D.E., Chan,M.S., Day,N.P., Enright,M.C., Goldstein,R., Hood,D.W., Kalia,A., Moore,C.E., Zhou,J. and Spratt,B.G. (2001) Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. *Proc. Natl Acad. Sci. USA*, **98**, 182–187.

Jolley,K.A., Kalmusova,J., Feil,E.J., Gupta,S., Musilek,M., Kriz,P. and Maiden,M.C.J. (2000) Carried meningococci in the Czech Republic: a diverse recombining population. *J. Clin. Microbiol.*, **38**, 4492–4498.

Maiden,M.C.J., Bygraves,J.A., Feil,E., Morelli,G., Russell,J.E., Urwin,R., Zhang,Q., Zhou,J., Zurth,K., Caugant,D.A., Feavers,I.M., Achtman,M. and Spratt,B.G. (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl Acad. Sci. USA*, **95**, 3140–3145.

Maynard-Smith,J. (1992) Analysing the mosaic structure of genes. *J. Mol. Evol.*, **34**, 126–129.

Maynard-Smith,J. and Smith,N.H. (1998) Detecting recombination from gene trees. *Mol. Biol. Evol.*, **15**, 590–599.

Maynard-Smith,J., Smith,N.H., O'Rourke,M. and Spratt,B.G. (1993) How clonal are bacteria? *Proc. Natl Acad. Sci. USA*, **90**, 4384–4388.

Nei,M. and Gojobori,T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.*, **3**, 418–426.

Sawyer,S. (1989) Statistical tests for detecting gene conversion. *Mol. Biol. Evol.*, **6**, 526–538.