A guide to spolTools

24 July 2008

This document is provided as a reference for using spolTools [1]. There are three sections in this document: Data sets and formatting, Programs and a Glossary.

1 Data sets and formatting

The repository contains data sets from spoligotyping [2].

Searching the repository

Searching through the repository can be done using either of the two search pages: by spoligotypes or by publication. Each of these search pages have search fields and display options. Hovering the mouse over the search fields shows a black pop-up box with a description of these fields. Selecting tickboxes of the display options on the right of these search pages selects the fields to be displayed.

Sorting results

The results page from searching the repository will return a list of data sets, filtered using the search options you have chosen. This list can be sorted by a specific field by clicking the heading name in the column. A button labelled '**Get Details**' beside a data set brings up the data set display.

Data set display

The data set is displayed with summary statistics. Buttons linking to the rich spoligotype format (RSF), spoligotype patterns display, and the spolTools programs (DESTUS [3] and spoligoforests[4])are also provided. Descriptions of these programs are in Section 2.

Converting a data set between formats

There are different formatting conventions for spoligotypes [5]. The **Convert formats** page allows conversion between the binary, hex, octal and gapformats. Go to **Run Programs** > **Convert formats**.

For example, type a row in the text box such as:

```
1018 : 3, 9, 16, 19, 32, 39-43 : 20
```

where 1018 is the label of the spoligotype, the number list "3, 9, 16, 19, 32, 39-43" is its spoligotype pattern in gap format, and 20 is the number of isolates with this pattern in the data set.

Note that you can paste multiple rows (of the same format) in this text box. Click **Show all formats** to display the data set in all the formats. From here you may choose to proceed with creating the RSF of this data set.

Creating RSF of a data set

The RSF of a data set can be prepared in the **Create RSF** page.

Submitting a data set

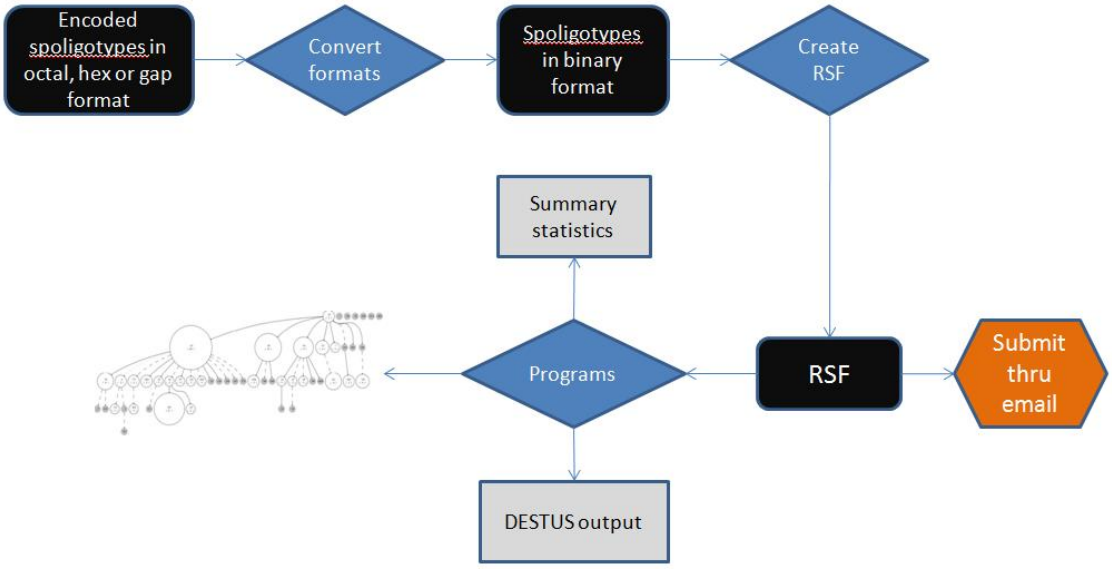
The spolTools repository is manually curated. To submit a data set for inclusion, perform the following steps.

1. Prepare your data set in binary format. If your data set is in gap, octal or hex format, use the **Convert formats** page to convert the spoligotypes to the required binary format.
2. Generate the RSF for your data set in the **Create RSF** page.
3. Save the RSF as a text file.
4. Email the curator at spolTools@unsw.edu.au. In this email, indicate your contact details and other relevant information. Attach your RSF file to this email.

The curators of spolTools will contact you to confirm your submission to the repository.

2 Programs

Programs in spolTools can also be run on data sets not in the repository. Using the programs requires a data set in RSF (See previous section on Creating RSF of a data set).



Programs results page

Like the data set display when searching through the spolTools repository, the results from spolTools programs consists of summary statistics and buttons for running DESTUS and generating spoligoforests.

Spoligoforests

Spoligoforests [4] can be generated using either of two different layout methods. The hierarchical method shows a 'ranked' layout where spoligotypes that are inferred to be derived from another spoligotype are placed below the inferred parent. The output of the Fruchterman-Reingold method resembles a 'burst' layout. Solid edges are relationships between spoligotypes that are found to be unique in the data set, i.e., a spoligotype is inferred to arise from only ONE specific parent spoligotype. Edges that are broken lines are chosen among multiple edges according to a deletion model [4]. Dotted lines have probability measure less than 0.5, while dashed lines have probability measure greater than or equal to 0.5. A spoligoforest often consists of disjoint 'trees'.

DESTUS

A method to determine emerging strains within a spoligotype data set is implemented by the program DESTUS [3]. The output includes emerging genotypes detected using three different techniques for correcting for multiple statistical tests. These techniques are listed from the most sensitive to the least sensitive test. Storey tests for q-values that are below a specified threshold (default is 0.8). Benjamini-Hochberg also considers a cut-off q-value (default is 0.8). Dunn-Sidak requires an input significance level (default is 0.01).

Retrieving jobs

Results from previously submitted jobs can be retrieved in the **Programs page** by entering the Job ID number.

3 Glossary

DESTUS Detecting Emerging Strains of Tuberculosis by Using Spoligotypes, is a statistical tool to identify emerging strains in a data set of tuberculosis isolates.

gap format a formatting of a spoligotype that lists the numbered positions in the spoligotype where there are gaps, i.e. absent spacers.

RSF Rich Spoligotype Format, is a text file formatting of a spoligotype data set that allows parsing for the different programs in spolTools. RSF also allows easy exchange of information as it summarises details of a data set.

spoligoforests a visualisation tool for relationships between spoligotypes. The construction of this graph is based on a model of evolution of spoligotypes that considers deletions of spacer sequences as irreversible.

spoligotyping Spacer oligonucleotide typing, a genotyping method for *M. tuberculosis* that exploits polymorphism of the direct repeat (DR) region.

References

[1] Tang, C., Reyes, J., Luciani, F., Francis, A., and Tanaka, M. *To appear* .

[2] Kamerbeek, J., Schouls, L., Kolk, A., van Agterveld, M., van Soolingen, D., Kuijper, S., Bunschoten, A., Molhuizen, H., Shaw, R., Goyal, M., and van Embden, J. *J Clin Microbiol* **35**(4), 907–914 April (1997).

[3] Tanaka, M. M. and Francis, A. R. *Proc Natl Acad Sci U S A* **103**, 15266–15271 (2006).

[4] Reyes, J. F., Francis, A. R., and Tanaka, M. M. *Unpublished manuscript* .

[5] Dale, J., Brittain, D., Cataldi, A., Cousins, D., Crawford, J., Driscoll, J., Heersma, H., Lillebaek, T., Quitugua, T., Rastogi, N., Skuce, R., Sola, C., Van Soolingen, D., and Vincent, V. *Int J Tuberc Lung Dis* **5**, 216–9 (2001).