GenoProfiler: batch processing of high-throughput capillary fingerprinting data

Frank M. You\textsuperscript{1,†}, Ming-Cheng Luo\textsuperscript{1,†}, Yong Qiang Gu\textsuperscript{2}, Gerard R. Lazo\textsuperscript{2}, Karin Deal\textsuperscript{1}, Jan Dvorak\textsuperscript{1} and Olin D. Anderson\textsuperscript{2,*}

\textsuperscript{1}Department of Plant Sciences, University of California, Davis, CA 95616, USA and \textsuperscript{2}Western Regional Research Center, Agricultural Research Service, US Department of Agriculture, 800 Buchanan Street, Albany, CA 94710, USA

Received on April 28, 2006; revised on August 23, 2006; accepted on September 22, 2006

Advance Access publication October 2, 2006

Associate Editor: Keith A Crandall

\begin{abstract}
Summary: High-throughput content fingerprinting techniques employing capillary electrophoresis place new demands on the editing of fingerprint files for the downstream contig assembly program, FPC. A cross-platform software application, GenoProfiler, was developed for automated editing of sized fingerprinting profiles generated by the ABI Genetic Analyzers. The batch-processing module extracts the sized fragment information directly from the ABI raw trace files, or from data files exported from GeneMapper or other size calling software, removes the background noise and undesired fragments, and generates fragment size files compatible with the FPC software.

Availability: \url{http://wheat.pw.usda.gov/PhysicalMapping/}
Contact: oandersn@pw.usda.gov
\end{abstract}

1 INTRODUCTION

Thousands of bacterial artificial chromosome (BAC) clones need to be fingerprinted in large-scale physical mapping projects in which an high-information content fingerprinting (HICF) technique with automated sample preparation and capillary electrophoresis are employed (Ding \textit{et al.}, 2001; Luo \textit{et al.}, 2003, Schibler \textit{et al.}, 2004; Xu \textit{et al.}, 2005). These high-throughput fingerprinting procedures place a pressing need on the automated editing of fingerprint files for downstream BAC contig assembly (Soderlund \textit{et al.}, 1997, 2000). The editing operations include (1) restriction fragment size calling, (2) distinguishing restriction fragment peaks from background peaks and primer dimer peaks, and eliminating them from the profiles, (3) detecting and eliminating substandard fingerprinting profiles, and (4) detecting and eliminating profiles resulting from cross-contamination during BAC library construction, replication and fingerprinting.

We developed a software package, named GenoProfiler, which provides a batch module for fully automated fingerprint profile editing and a set of utility tools that allow users to carry out various tasks related to the above editing operations for downstream contig assembly. The major utilities include (1) chromatograph (trace) viewer, (2) BAC cross-contamination check, (3) fragment frequency analysis, (4) file management and (5) fingerprint viewer.

The final exported fragment size files can be used for contig assembly by using the FPC software (Soderlund \textit{et al.}, 1997, 2000). GenoProfiler is a cross-platform software package based on the Java environment (Sun Microsystems, Palo Alto, CA, USA). It is designed as a graphical user interface (GUI) based application running on multiple platforms. The proposed algorithms and functional modules facilitate users to set up the optimized parameters based on their own datasets and to achieve the best results. The functionality in this software is unique and is not provided in commercially available software such as GeneScan or GeneMapper (Applied Biosystems, Foster City, CA, USA). The software has been distributed to more than 300 laboratories worldwide, and successfully employed in several physical mapping projects, e.g. \textit{Aegilops tauschii}, barley and soybean. We discuss here the major algorithms and implementations of the batch-processing module. Additional descriptions and features of the software can be found at the software website and in user’s manual.

2 ALGORITHMS AND IMPLEMENTATION OF BATCH PROCESSING

Batch processing of sample data files includes extraction of sized fragments, elimination of background noise and identification of ‘true’ fragment peaks (see definition in the forthcoming Step 2), and exclusion of some undesired fragments, such as vector fragment(s), off-scale fragment(s) and wide peaks resulting from co-migrating fragments. The input includes two different types of data files. The first type of data files is the raw trace files generated by any ABI Genetic Analyzer using Data Collection software version 1.x. The second type of data files are the sized fingerprint data files exported by GeneMapper or other size calling software. The output of the batch-processing is edited fragment size files compatible with the FPC software (Soderlund \textit{et al.}, 1997, 2000). The following are major algorithms and their implementations.

The fingerprint profile, the characteristics of clone fingerprint after size-calling, of each clone extracted from a sample data file contains ‘true’ fragment peaks, background noise, numerous peaks of low fluorescence intensity caused by \textit{Escherichia coli} DNA contamination in BAC DNA preparations and peaks due to occasionally incomplete digestion of BAC DNAs. Background peaks are present in \textasciitilde80% of profiles after size calling (data not shown). In those profiles, they could account for as much as 90% of all peaks present. To deal with this problem, a two-step algorithm was...
designed to determine profile-specific threshold that minimizes the number of background peaks in the edited profiles.

2.1 Step 1: find a background threshold
This step is to eliminate ‘real’ background peaks, which are observed as low density peaks even in negative control (no DNA template), based on the frequency histogram of peak heights (relative fluorescence units, RFU) of a fingerprinting profile (Fig. 1).

1. Given a fingerprint profile in a specific color channel, calculate the frequency histogram of peak heights using a bin width of 20 RFUs and convert the frequencies into percentages.
2. Smooth the histogram to a continuous distribution curve using the Savitzky–Golay smoothing filter (local cubic polynomial fit with five point smoothing scheme) (Press et al., 1988).
3. Since the background peaks are the most frequent peaks in most profiles, they form the global maximum of the smoothed curve (the portion of the curve from 0 to 155 RFUs in Fig. 1). Find the global maximum of the smoothed frequency distribution curve. Find the first local minimum to the right of the global maximum of the curve. All peaks to the left of this minimum (black line in Fig. 1) are background peaks. If the global maximum is ≤5 RFUs and the local minimum is <2 RFUs then the threshold is set to the default value or a user-specified value.
4. Exclude the peaks below the threshold from the profile.

2.2 Step 2: find the ‘true’ fragment threshold
Peak heights form a continuum in a typical fingerprint profile, which necessitates adjustment of the background threshold to minimize peaks due to various artifacts (such as primer–dimers and unspecific amplification products) or the peaks that are too low to be reproducibly above a background threshold. Therefore, after setting the basic background level and removing the background peaks in Step 1, a ‘true’ fragment threshold (called adjusted background threshold in Fig. 1) is obtained in the following manner.

The remaining peaks after background removal are ranked by their height and the heights of the sixth through the tenth highest peaks are averaged. The ‘true’ fragment threshold is defined as the peak average of the five-highest peaks multiplied by a specified ratio. If the ‘true’ fragment threshold is less than or equal to the background threshold, then Step 2 stops. Otherwise the peaks with a peak height less than the ‘true’ fragment threshold will be removed. We elected to use the sixth through the tenth highest peaks after the exclusion of off-scale peaks because these peaks are the most representative of a profile. Users have an option to change this parameter. The ratio is chosen so as to optimize the balance between the false positive and false negative (F+/F−) fragments.

Theoretically, a true fragment in a fingerprint profile is defined as the fragment corresponding to a predicted fragment on the basis of the restriction enzyme patterns based on known nucleotide sequence. Hence, if a fragment in the capillary fingerprinting profile can be matched to a predicted fragment, the fragment is considered to be a true fragment. The percentage of fragments in a profile that do not match predicted fragments is the ‘false positive’ (F+) rate. If a predicted fragment is not found in a fingerprint profile, the fragment is ‘false negative’ and percentage of such fragments is the ‘false negative’ (F−) rate. The goal is to set the editing parameters such that both F+ and F− rates are optimally balanced. Usually, equal number of F+ and F− is acceptable. However, F+ causes more serve problem during contig assembly. In the event that a large number of clone fingerprints are involved, less F+ than F− is favorable. In the A.tauschii physical mapping project, we used the repeated fingerprints of two sequenced BAC clones (Luo et al., 2003) as a training dataset. The ratios of the four color channels were optimized as 0.35 for the blue channel, 0.22 for the green channel, 0.34 for the yellow channel and 0.28 for the red channel. Different values of the ratio reflect variation in the activity of restriction enzymes during DNA digestion resulting in variation in peak heights. A case study of determining the ‘true’ fragment threshold can be found in user’s manual. In practice, users can empirically optimize this ratio for each color channel to match their data. GenoProfiler provides a graphic interface to set the ratios.

3 EVALUATION OF ALGORITHMS
To evaluate the algorithms, fully sequenced Triticum monococcum BAC 115G1 was repeatedly fingerprinted (Luo et al., 2003). A sample of 611 fingerprint profiles was used to estimate the F+ and F− rates. The means and standard deviations of F+ and F− rates were 6.8 ± 1.33% and 6.8 ± 1.57%, respectively. The number of sized fragments were averagely 109 ± 2.3 with the coefficient of variation of 2.1%, and the profiles shared 90.7 ± 4.9%
fragments, The 611 edited fingerprints were then used as an input in contig assembly with FPC, to determine at which level of Sulston score (Soderlund et al., 1997, 2000) they will fail to assemble as a single stack. Tolerance of 0.4 bp was used. Even at a Sulston score of $1 \times 10^{-9}$, the clones were assembled into a single stack. These results showed acceptable accuracy and high reproducibility of the editing process based on these algorithms. We also compared the two-step algorithm with the Step 2 only algorithm. The results showed that the Step 2 only algorithm could not entirely remove background noise from some fingerprint profiles (data not shown).

GenoProfiler can process any number of sample files with a minimum of memory required (RAM 256 MB). The processing time depends on the number of samples. For processing of 100 000 BAC fingerprints, it took ~4.5 h to process raw data off the ABI3100/3700 sequencers or 5 min to process sizing files exported by GeneMapper with a 3.2 GHz CPU and 1.0 GB RAM computer.

ACKNOWLEDGEMENTS

This publication is based upon work supported by the National Science Foundation grant no. DBI-0077766 and is in part associated with the efforts of the United States Department of Agriculture, Agricultural Research Service (Current Research Information System CRIS No. 5325-21000-011-00D).

Conflict of Interest: none declared

REFERENCES


