Published in final edited form as:

Mitochondrial DNA. 2010 December ; 21(6): 191–197. doi:10.3109/19401736.2010.528756.

Optimization of primer specific filter metrics for the assessment of mitochondrial DNA sequence data

PAMELA C. CURTIS¹, JENNIFER L. THOMAS¹, NICOLE R. PHILLIPS¹, and RHONDA K. ROBY^{1,2}

PAMELA C. CURTIS: Pamela.Curtis@unthsc.edu; JENNIFER L. THOMAS: Jennifer.Thomas@unthsc.edu; NICOLE R. PHILLIPS: Nicole.Phillips@unthsc.edu; RHONDA K. ROBY: Rhonda.Roby@unthsc.edu

¹Department of Forensic & Investigative Genetics, University of North Texas Health Science Center, Fort Worth, TX, USA

²Institute of Investigative Genetics, University of North Texas Health Science Center, Fort Worth, TX, USA

Abstract

Filter metrics are used as a quick assessment of sequence trace files in order to sort data into different categories, i.e. High Quality, Review, and Low Quality, without human intervention. The filter metrics consist of two numerical parameters for sequence quality assessment: trace score (TS) and contiguous read length (CRL). Primer specific settings for the TS and CRL were established using a calibration dataset of 2817 traces and validated using a concordance dataset of 5617 traces. Prior to optimization, 57% of the traces required manual review before import into a sequence analysis program, whereas after optimization only 28% of the traces required manual review. After optimization of primer specific filter metrics for mitochondrial DNA sequence data, an overall reduction of review of trace files translates into increased throughput of data analysis and decreased time required for manual review.

Keywords

Filter metrics; expert systems; trace score; contiguous read length; quality assessment

Introduction

Filter metrics are used as a quick assessment of sequence trace files in order to sort data into different categories, i.e. High Quality, Review, and Low Quality, without human intervention (Roby et al. 2009b). Currently, in forensic DNA testing, software tools that use expert system logic are being used to help review single source nuclear DNA data and to reduce the backlog of convicted offender data for upload into the national DNA database (Roby 2008a). An expert system for nuclear DNA is defined by the forensic community as a software program or set of software programs that identifies peaks/bands, assigns alleles, ensures data meet laboratory defined criteria, describes rationale behind decisions, and makes no incorrect allele calls without human intervention (Roby and Christen 2007). An expert system applies “If..., then...” statements to make decisions on the quality of data and automates allele calling (Engelmore, Feigenbaum 1993; Hunt 1986). The software must

Correspondence: Rhonda K. Roby, Ph.D., M.P.H., Department of Forensic & Investigative Genetics, Institute of Investigative Genetics, University of North Texas Health Science Center, 3500 Camp Bowie Boulevard, Fort Worth, TX, USA. Fax: +1 817-735-5016. Rhonda.Roby@unthsc.edu.

Declaration of interest The authors report no conflicts of interest.

provide justification for each decision (Roby, Tincher 2010). Examples of laboratory-defined criteria used for short tandem repeat (STR) analysis in GeneMapper™ ID (Applied Biosystems, Inc. [ABI]) include allele number, peak height ratio, and off-scale data (Applied Biosystems 2003). The software uses these criteria to quickly signal, or fire a rule, regarding data quality. The rule firings expedite data interpretation through the use of shapes and colors displayed in the user interface of the software. If a sample yields good quality data and meets all the laboratory-defined thresholds, a green square is displayed for each parameter. If data do not meet a specific laboratory defined threshold, a rule is fired drawing a scientist's attention to that particular sample or locus via a yellow triangle. A red octagon signifies that a locus or sample failed.

Expert systems have the potential to streamline data analysis and reduce backlogs within laboratories (Roby and Jones 2005; Perlin et al. 2001). An expert system for sequence analysis should reduce the amount of time a scientist spends reviewing sequence data and, therefore, should increase sample throughput. A proposed definition for a sequence analysis expert system is a software program or set of software programs that identifies peaks, assigns bases, ensures data meet laboratory defined criteria, describes rationale behind decisions, reviews sequence data prior to use in a contigs, reviews the quality of each base, skips to positions of bases with low quality, and searches sequence data for unusual patterns (Roby et al. 2010). Expert systems may also reduce the potential for human error, as the process is automated, consistent, and accurate. Implementation of expert systems within a laboratory reduces analysis time; therefore, freeing the scientist for other duties. No complete expert system for sequence data analysis is currently available. This paper presents an existing software program, Sequence Scanner Software v1.0 (Applied Biosystems, Inc. [ABI]), which has rule firings that can assist scientists in the initial review of sequence data.

Software programs are utilized by scientists to build contigs, align trace files, and analyze mitochondrial DNA (mtDNA) sequence data. Prior to analysis of the sequence data, Sequence Scanner Software v1.0 (available from <http://www.appliedbiosystems.com>) can be used for a quick quality assessment of sequence data. Sequence Scanner Software is a downloadable software program that allows the scientist to display, edit, trim, export, and generate quality assessments of ABI BigDye® Terminator sequencing .ab1 files generated by the suite of ABI PRISM® capillary instrumentation. Within this software program, the scientist can set expert system-like rules and rule firings such as quality value (QV), window size, trace score (TS), and contiguous read length (CRL). QV is a value assigned to each nucleotide (base); the calculation for QV is $-10\log_{10}P_e$, where P_e is the probability of error. A value of 1 to 60 may be entered for this parameter. The window size is used to calculate the CRL and refers to the first and last stretch of bases with an average QV greater than the laboratory-defined threshold, thus indicating the beginning and end of a CRL. The window size may be set at a value of 5 to 999. The TS is calculated after trimming the sequence; it is the average of the quality values of all the bases. The CRL is measured by the stretch of bases with a QV greater than or equal to the laboratory defined threshold. Quality assessments can be made quickly with filter metrics defined by the laboratory as High Quality, Review, and Low Quality. With optimized filter metrics, the scientist can quickly assess the quality of sequence files. Data of high quality ("pass") and low quality ("fail") have been characterized within our laboratory and are color-coded with defined thresholds and flagged as green or red, respectively. A yellow flag assigned to a trace file signifies that the data do not fall in the high quality window nor the low quality window and that the scientist should review the data and determine its use in a contig.

Four parameters in Sequence Scanner Software are used as the filter metrics to quickly assess sequence quality (Roby 2008b). Two of these parameters, QV and window size, are held constant while two parameters, TS and CRL, are variable. Window size is held constant

at 20 bases and QV is held constant at 20 for this study. For example, if a base has a QV = 20, this means it has a $P_e = 1\%$, indicating that there is a 1% chance of being the wrong base call (See Table 1). The scientist can set thresholds for TS with color coordination. Prior to optimization of the primer specific filter metrics, a preliminary evaluation was performed to define these settings. For high quality data that requires no human intervention, high quality is defined as 35 to 100 for TS. For a low quality sequence the TS is defined as zero to 20. The review range is defined as 21 to 34. The second parameter assessed is the CRL. The software uses the quality value of a single base and adjacent bases that make up a specified window size to calculate the CRL. The CRL also allows the scientist to set thresholds with color coordination. Prior to optimization of the filter metrics, high quality sequence was set to be 401 or greater. The low quality sequence ranged from zero to 200. The CRL of sequence traces requiring review was set at 201 to 400. Prior to this study, these filter metric settings were applied to all sequence data regardless of the individual primer's sequencing read length or location of sequencing primer.

By optimizing the filter metrics defined for each primer, increased throughput of data analysis can be achieved. Further, the scientist can accurately assess the quality of the sequence data without launching or viewing each trace file to determine if data will be used in the sample's contig. Launching and viewing each trace file is time-consuming. The graphic viewer in most software programs allows a display of approximately 50 bases and requires the scientist to scroll through the sequence data. With optimized filter metrics, the scientist no longer needs to launch and view each trace file in order to ascertain the quality of sequence data; (s)he only needs to review those sequence traces flagged as "review".

Materials and Methods

Laboratory processing

The mtDNA sequence data from a population database were used to optimize the filter metrics for this study (Roby et al. 2009a). DNA from 1000 male buccal swabs was extracted using the DNA IQ™ System (Promega Corporation) on the Freedom EVO® 100 (Tecan Group Ltd., Männedorf, Switzerland) (Plopper et al. 2006). A single amplification of a 1.1kb fragment was performed to generate sequence that encompasses both hypervariable region 1 (HV1) and hypervariable region 2 (HV2) of the mitochondrial genome (see Figure 1). This single large amplicon was generated using primers R1 (forward) and R2 (reverse). Amplification setup was performed on the MiniPrep 75 Sample Processor (Tecan Group Ltd.). The large amplicon was sequenced using BigDye® Terminator v1.1 Cycle Sequencing Kit (Applied Biosystems, Inc. [ABI]). Cycle sequencing was performed with eight sequencing primers to obtain coverage of the entire amplicon (see Figure 1) (Roby et al. 2008). The PCR products were analyzed via capillary electrophoresis on the ABI PRISM® 3130xl Genetic Analyzer. Prior to aligning and analyzing sequence data, filter metrics (TS = 20, 34 and CRL = 200, 400) were used to assess the sequence trace quality. For this study, we reviewed each sequence trace and the corresponding sequence quality for each file whenever yellow "review" flags were fired. The quality of the "review" sequences was manually assessed by the following review criteria: baseline noise, signal intensity, read length and anomalies (e.g. heteroplasmy and homopolymeric stretches). The sequence trace quality was annotated by us on the printed Quality Control (QC) Reports, a feature of Sequence Scanner Software. This review was required in order for us to build contigs and to optimize filter metrics.

Optimization of filter metrics

Using the mtDNA sequence trace files and our original notations, optimization of primer specific filter metrics was performed. Optimization is the process of (1) customizing the

filter metrics to improve the accuracy and effectiveness of the filter; (2) verifying that the rule firings are consistent with the human decision-making process; and, (3) confirming that the software performs these tasks consistently. The QC Reports generated by Sequence Scanner Software were exported and opened in Microsoft® Excel. Our comments were manually entered into the Excel spreadsheet. A passing or failing status was assigned to each sequence trace based on review criteria. In order to calibrate the software, a dataset of 2817 sequence trace files was used. Calibration is the process of modifying the filter metrics and determining if the new settings allow the samples to parse into the appropriate categories, i.e. High Quality, Review, and Low Quality (Roby and Christen 2007; Butler 2006). A Microsoft® Access database was constructed using the data contained in the Excel spreadsheet for each of the eight primers to allow for quick querying of potential filter metric settings. Concordance was performed to demonstrate that the new filter metrics provided a better assessment of the data than the previous values. A total of 5617 trace files was designated for validation and concordance of the optimized settings. Additional Microsoft® Access databases were constructed for each primer to verify the proposed primer specific filter metric settings.

Results

Specific filter metrics were defined for each of the eight primers used. After the calibration study, the new filter metrics were applied to the corresponding sequence files. Figure 2 displays the percentage of trace files requiring review prior to optimization and after optimization. As can be seen on this graph, considerable timesavings can be achieved by using filter metrics that are optimized per primer (see Figure 2). Since these filter metrics were initially defined using the R1 and R2 primers, which have a long potential read length, optimization of primer specific filter metrics shows less improvement than other primers. Primer specific filter metrics for B4, C2, and C1 demonstrate a considerable decrease in the number of trace files that would require manual review.

Primer B4 produces a short sequencing fragment of mtDNA (see Figure 1). With optimized filter metrics for Primer B4, 72% of the trace files did not require manual review; these trace files automatically passed. Table 2 specifies the breakdown of filter metrics prior to optimization and after optimization for a total of 352 trace files. Prior to optimization, none of the trace files fit into the high quality category because the CRL threshold was set too high (i.e. CRL = 400) for this short sequencing fragment of approximately 250 bases. After optimization with a CRL set at 200, 222 of the trace files fit into this category and did not require any review prior to use in its contig for sequence analysis. Prior to optimization, 318 trace files were flagged yellow requiring manual review. After optimization, only 97 of the trace files required manual review. The low quality thresholds were well defined prior to optimization (see Table 2). Figure 3 displays a scatter plot of all passing and failing trace files, green and red respectively, according to our annotations. The green box represents the passing filter metrics: passing threshold for TS and passing threshold for CRL. High quality data should fall in the green box. The red box represents the failing filter metrics: failing threshold for TS and failing threshold for CRL. Low quality data should fall in the red box. All data in the gray area are subjected to manual review. Figure 3a displays the trace files for Primer B4 prior to primer specific optimization and Figure 3b displays the trace files for the B4 primer after optimization. No trace files fall into the green “passing” box prior to primer specific optimization (see Figure 3a). After primer specific filter metrics are applied, fewer trace files require manual review (see Figure 3b). As illustrated, implementation of primer specific filter metrics provides an accurate representation of sequence quality and reduces the amount of time a scientist spends reviewing trace files.

Each of the sequencing primers used has various read length possibilities (see Figure 1). Using the revised Cambridge Reference Sequence (rCRS), as a reference standard, the number of bases 3' to the primer binding location was counted (see Figure 1) (Andrews et al. 1999). Primer R1, for example, has a maximum potential read length of 1183 bases; however, if a sequence trace contains an HV1 homopolymeric stretch, the read length is shortened to approximately 253 bases. If a sequence trace does not contain a homopolymeric stretch in HV1 but does contain a length heteroplasmy in HV2, Primer R1 could sequence through HV1 and into HV2 for approximately 941 bases until it reaches the length heteroplasmy in HV2. When sequencing the complementary strand in the reverse direction, Primer R2 has a maximum potential read length of 1183 bases. If a sequence trace contains a length heteroplasmy in HV2, the read length for Primer R2 stops at approximately 230 bases. If Primer R2 is able to sequence through HV2 and into HV1 but stops at approximately 921 bases, an educated assumption can be made that the sequence contains a homopolymeric stretch in HV1. Primer B1 only sequences HV1 and has a maximum potential read length of approximately 460 bases. However, if a B1 sequence trace has a high TS value and a CRL of approximately 198 bases, that trace file most likely contains a homopolymeric stretch in HV1. Primer C1 only sequences HV2 and has a maximum potential read length of approximately 497 bases. However, if a C1 sequence trace has a high TS value and a CRL of approximately 255, it can be assumed that sequence trace most likely contains a length heteroplasmy in HV2 (see Figure 1).

Optimized filter metrics for the eight sequencing primers used in our laboratory can be found in Table 3. Prior to optimization, data from all eight sequencing primers were assessed with TS settings of 20, 34 and CRL settings of 200, 400. Following primer specific optimization, the TS and CRL settings allow for data to be parsed more consistently based on the primer used for sequencing. Examples of low quality, review, and high quality data for each of the eight sequencing primers can be accessed in the Trace Archive database online (<http://www.ncbi.nlm.nih.gov/Traces/home>). The TI numbers are as follows: 2281021664, 2281021665, 2281021666, 2281021667, 2281021668, 2281021669, 2281021670, 2281021671, 2281021672, 2281021673, 2281021674, 2281021675, 2281021676, 2281021677, 2281021678, and 2281021679.

Discussion

We have shown that filter metrics are an important tool applied to sequence trace files. By optimizing filter metrics to specific sequencing primers, there was an overall decrease in the number of sequence trace files requiring review. Prior to optimization, 57% of the sequence traces were reviewed. After optimization, only 28% of the sequence traces require manual review. Using defined filter metrics for each primer translates into considerable timesavings. Implementing optimized primer specific filter metrics yields an estimated timesavings of approximately 50% prior to building a contig. While humans are prone to error and interruptions, a software program is not and can continuously provide consistent, objective measurements when the software logic is accurate. Increased laboratory throughput has been achieved with optimized filter metrics with a decrease in analysis times and an increase in consistent assessment of trace files. Future software developments could further automate sequence analysis.

Acknowledgments

The authors alone are responsible for the content and writing of the paper. The research conducted in this paper was in partial fulfillment of Pamela Musslewhite's (Curtis) master's thesis entitled, "Optimization of Filter Metrics for Mitochondrial DNA Sequence Analysis", August 2009. Support for this project was partially funded by NIJ Cooperative Agreement 2008-DN-BX-K192, Forensic DNA Unit Efficiency Improvement, FY 2008.

References

- Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nature Genetics*. 1999; 23:147. [PubMed: 10508508]
- Applied Biosystems. GeneMapper™ *ID* Software version 3.1 Human Identification Analysis User Guide. 2003
- Butler J. Debunking some urban legends surrounding validation within the forensic DNA community. *Profiles in DNA*. 2006; 9:3–6.
- Engelmore, RS.; Feigenbaum, E. Expert systems and artificial intelligence. 1993. [Online] (Published May 1993) Available at: http://www.wtec.org/loyola/kb/c1_s1.htm
- Hunt, VD. Artificial intelligence & expert systems sourcebook. New York: Chapman and Hall; 1986.
- Perlin, MW.; Coffman, D.; Crouse, C.; Konotop, F.; Ban, JD. Automated STR data analysis: validation studies. Twelfth International Symposium on Human Identification-2001; 29 November 2001; Madison, WI. Promega Corporation. 2001.
- Plopper, F.; Roby, R.; Planz, J.; Eisenberg, A. High throughput processing of family reference samples for missing persons programs: the use of robotics in extraction and amplification setup for STR and mtDNA analysis. Seventeenth International Symposium on Human Identification-2006; 9-12 October 2006; Nashville, TN. Promega Corporation. 2006.
- Roby RK. Expert systems help labs process DNA samples. *National Institute of Justice Journal*. 2008a; 260:16–19.
- Roby, RK. Doctoral Dissertation. Granada, Spain: University of Granada; 2008b. High throughput mitochondrial DNA analysis: optimization of sequence chemistry, characterization of local dye terminator sequencing frames, and tools for the development of an expert system.
- Roby, RK.; Capt, C.; Macurdy, KM.; Planz, JV.; Lorente, JA.; Eisenberg, AJ. New tools for mitochondrial DNA sequencing and analysis at the University of North Texas Center for Human Identification Laboratory. *Proceedings of the American Academy of Forensic Sciences*; 18-22 February 2008; Washington, D.C.. 2008.
- Roby RK, Christen AD. Validating expert systems: examples with the FSS-i³™ Expert System Software. *Profiles in DNA*. 2007; 10:13–15.
- Roby RK, Gonzalez SD, Phillips NR, Planz JV, Thomas JL, Pantoja Astudillo JA, Ge J, Aguirre Morales E, Eisenberg AJ, Chakraborty R, Bustos P, Budowle B. Autosomal STR allele frequencies and Y-STR and mtDNA haplotypes in Chilean sample populations. *Forensic Science International: Genetics Supplement Series*. 2009a; 2:532–533.
- Roby, RK.; Jones, JP. Evaluating expert systems for forensic DNA laboratories. 2005. [Online] (Published 2005) Available at: http://www3.appliedbiosystems.com/cms/groups/applied_markets_marketing/documents/generaldocuments/cms_042230.pdf
- Roby, RK.; Phillips, NR.; Thomas, JL.; Keppler, R.; Eisenberg, AJ. Quality assessment and alert messaging software for raw mitochondrial DNA sequence data. *Proceedings of the American Academy of Forensic Sciences*; 22-27 February 2010; Seattle, WA. 2010.
- Roby, RK.; Tincher, BM. Expert systems: high throughput analysis of single source samples for forensic DNA databasing. Huntington, WV: United States Department of Justice; In press. [Report]
- Roby, RK.; Thomas, JL.; Phillips, NR.; Gonzalez, SD.; Planz, JV.; Eisenberg, AJ. High-throughput processing of mitochondrial DNA analysis using robotics. *Proceedings of the American Academy of Forensic Sciences*; 16-21 February 2009; Denver, CO. 2009b.

Figure 1.

Control Region of mtDNA with amplification and cycle sequencing primers. Minimally forensic laboratories attempt to obtain sequence information from positions 16024 to 16365 (HV1) and positions 73 to 340 (HV2) in the control region for identification purposes. The blue region represents HV1 and the gray region represents HV2. The green diagonal represents the homopolymeric stretch commonly observed within HV1 and the purple diagonal represents the length heteroplasmy commonly observed within HV2. The white area is the extra information obtained by performing a single large amplification. Black arrows indicate forward primers and orange arrows indicate reverse primers. Amplification for the 1.1kb amplicon is performed using *Primer R1 and *Primer R2 of the displacement loop (D-loop) of the mitochondrial genome.

Figure 2.

Percentage of trace files requiring review per primer. The yellow bars represent the number of traces requiring manual review prior to filter metric optimization. The blue bars represent the number of traces requiring review after primer specific optimization. As can be seen, Primer B4 required more than 90% of the trace files to be manually reviewed prior to optimization. After optimization, less than 30% of the trace files require manual review.

Figure 3.

Scatterplots of Primer B4 filter metrics before and after optimization. The green line and the yellow line on the horizontal axes represent the upper and lower thresholds, respectively, for the TS. The green line and the yellow line on the vertical axes represent the upper and lower thresholds, respectively, for the CRL. Any trace files plotted below both yellow lines in the red box indicate low quality data and automatically fail. Any trace files plotted above both green lines in the green box indicate high quality data and automatically pass. Each trace file plotted in the middle region requires manual review. (a) Primer B4 trace files prior to primer specific optimization; no Primer B4 trace files fit into the high quality category because the thresholds were set too high. (b) Same trace files after primer-specific optimization; high quality trace files fit into the green box and fewer trace files required manual review after primer specific optimization of the filter metrics.

Table 1

Quality value and associated probability of error. Sequence Scanner Software uses the quality per peak, evaluates the overlap of fluorescent signals, and measures a Gaussian fit to determine a peak's quality value. Quality value is a number assigned to each base; the calculation for $QV = -10\log_{10}P_e$, where P_e is the probability of error. As shown in the table, if a base is assigned a QV of 20, there is a 1% chance of that base call being incorrect. According to Sequence Scanner, "high quality pure bases are generally assigned a QV between 20 and 50. Hence, high quality values indicate a low P_e . For the optimization of primer specific filter metrics, the QV threshold was held constant at 20. Table reproduced from information provided in Sequence Scanner Software v1.0 (Applied Biosystems, Inc.).

QV	P_e	QV	P_e	QV	P_e	QV	P_e	QV	P_e
1	79%	11	7.9%	21	0.79%	31	0.079%	50	0.001%
2	63%	12	6.3%	22	0.63%	32	0.063%	60	0.0001%
3	50%	13	5.0%	23	0.50%	33	0.050%	70	0.00001%
4	39%	14	3.9%	24	0.39%	34	0.039%	80	0.000001%
5	31%	15	3.1%	25	0.31%	35	0.031%	90	0.0000001%
6	25%	16	2.5%	26	0.25%	36	0.025%	99	0.000000012%
7	20%	17	2.0%	27	0.20%	37	0.020%		
8	15%	18	1.5%	28	0.15%	38	0.015%		
9	12%	19	1.2%	29	0.12%	39	0.012%		
10	10%	20	1.0%	30	0.10%	40	0.010%		

Table 2

Filter metric Assessment for Primer B4. Prior to optimization, zero trace files had a high quality filter metric (high quality TS and high quality CRL). After optimization, 222 trace files fit into the high quality category. Prior to optimization, 318 trace files require manual review and after primer specific filter metric settings were applied, only 97 trace files required manual review.

Assessment	Total no. of trace files <i>before</i> optimization	Total no. of trace files <i>after</i> optimization
High Quality	0	222
Review	318	97
Low Quality	34	33
TOTAL	352	352

Primer-specific optimized filter metrics. The values below are the primer specific filter metrics for each of the eight sequencing primers defined by our laboratory's internal validation. These values should be used as initial settings for a laboratory. Internal validation should be performed by individual laboratories to define its laboratory-specific settings.

Table 3

Primer	Trace Score				Contiguous Read Length			
	Low Quality	Review	High Quality		Low Quality	Review	High Quality	
R1	0-20	21-29	30-100		0-200	201-250	≥251	
B1	0-20	21-27	28-100		0-150	151-210	≥211	
C1	0-20	21-24	25-100		0-150	151-250	≥251	
R2	0-20	21-24	25-100		0-110	111-250	≥251	
A4	0-20	21-24	25-100		0-200	201-250	≥251	
B4	0-20	21-24	25-100		0-150	151-200	≥201	
C2	0-20	21-24	25-100		0-100	101-150	≥151	
D2	0-20	21-24	25-100		0-100	101-150	≥151	