International Congress Series 1261 (2004) 94–96

www.ics-elsevier.com

Evaluation and validation of quality assessment of mitochondrial control region sequence data by Phred

E. Sørensen*, E.M. Rasmussen, B. Eriksen, H.J. Larsen, N. Morling

Department of Forensic Genetics, Institute of Forensic Medicine, University of Copenhagen, 11 Frederik V's Vej, DK-2100 Copenhagen, Denmark

Abstract. This study presents the evaluation and validation of the computer programme Phred for semi-automated quality assessment of sequence data from the mitochondrial hypervariable regions HV1 and HV2. We conclude that Phred is a valuable aid in the quality control of mtDNA sequence data. However, Phred is not well suited for identification of base caller induced deletions. This can be compensated for by repeated sequencing. In addition, Phred does not accurately detect base heteroplasmy. Other types of software can be used to supplement Phred in the identification of heteroplasmy. © 2003 Elsevier B.V. All rights reserved.

Keywords: DNA sequencing; Quality assessment; Phred; mtDNA; Forensic genetics

1. Introduction

Sequencing of the mitochondrial control region for forensic purposes generates a large amount of data and, therefore, a thorough quality control of each sequenced base pair is very time consuming. The computer programme Phred (CodonCode) provides quality values (q-values) for each base call in a sequence. The q-values are logarithmically correlated to error probabilities. The relation between a q-value and the error probability (p) is defined as $q = -10 \times \log_{10}(p)$. Thus, a q-value of 20 corresponds to an error probability of 0.01 for the base call, while a q-value of 30 corresponds to an error probability of 0.001 [1,2]. The alignment program Sequencher (GeneCodes) provides a graphical interface, which presents the q-values above and below a chosen threshold (e.g. q-value 20) as various background colours to the nucleotide sequence. In this way, attention is directed to problematic areas in the sequencing results. Thus, the combination of the Phred and the Sequencher programmes is potentially a powerful tool to ensure that base pairs in a sequence are of sufficient quality. Furthermore, the q-values from Phred

* Corresponding author. Tel.: +45-3532-6110; fax: +45-3532-6120.
E-mail address: erik.soerensen@forensic.ku.dk (E. Sørensen).

doi:10.1016/S0531-5131(03)01507-3

also make it possible to compare the quality of two or more sets of sequencing results. The subject of the current study is to evaluate and validate the accuracy of Phred's error prediction in a forensic genetic venue.

2. Materials and methods

Data from a project involving sequencing of the hypervariable regions HV1 and HV2 from 32 hair shafts (120,419 base pairs) were used for the evaluation. All data originated from electrophoresis on the ABI 310 instrument. Base calling was done by the ABI Sequencing Analysis programme and the data were further analysed by Phred (version 0.000925.c) and then aligned by the Sequencher programme. The true sequences were established by typing each region with both BigDye-primer and -terminator chemistry in both forward and reverse directions. The consensus of at least four typing reactions was considered the correct sequence. The aligned sequences was exported from Sequencher and analysed/grouped by a computer programme (CAF2Table) developed in cooperation with the manufacturers of Phred. After analysis with CAF2Table, the data were imported to MS Excel for the final analysis. The number of actually erroneous base calls, i.e. discrepancies between the actual base call and the consensus of base calls from at least four sequences, was compared to the error rate predicted by Phred.

The data (or segments thereof) were analysed in three different set-ups. In the first set-up, all base pairs were included. When base caller induced deletions occurred (such deletions were revealed by the repeated sequencing), they were given the same q-value as the neighbouring base of least quality. In the second set-up, all bases were included as well, but base caller induced deletions were given a q-value of zero (equal to an error probability of 1). In the third set-up, all sequences with length heteroplasmy were excluded from analysis, and base caller induced deletions were given a q-value of zero.

3. Results

In the first analysis, the correlation between Phred's predictions and the number of observed errors was not satisfactory; e.g. among the 12,379 bases with q-values between 31 and 35, Phred only predicted 5.1 errors while 39 errors were observed (data not shown). In the second analysis, Phred's predictions were better, but still not satisfactory. Here, Phred still predicted 5.1 errors among the 12,361 bases with q-values between 31 and 35, while the observed number of errors was 21 (data not shown). However, in the third analysis, Phred's predictions were in better correlation with the number of observed errors. Table 1 shows a comparison between observed and expected errors grouped according to q-values. In addition, the number of observed errors for each of the sequencing-chemistries and -directions is presented. The table shows that while only a few base caller errors occurred for bases with q-values above 20 sequenced with terminator chemistry, this was not the case for bases of the same quality level sequenced with primer chemistry. This was especially the case for bases sequenced with primer chemistry in the forward direction. A comparison between Phred's predictions and the observed number of base caller errors for each base position in HV1 and HV2 revealed that a highly reproducibly base compression at position 16,053–16,054 was the reason for Phred's inadequate predictions. This compression was seen in sequences that originated from sequencing with primer chemistry

Table 1
Comparison between observed and expected errors grouped according to q-values

q-value	Total			Obs err with different sequencing chemistries			
	N	Obs err	Exp err	PrimFwd	PrimRev	TermFwd	TermRev
0–5	588	587	586.2	288	229	23	47
6–10	1939	243	302.8	139	73	14	17
11–15	1189	66	66.1	11	22	14	19
16–20	2250	56	37.8	21	26	3	6
21–25	1763	17	9.1	12	2	1	2
26–30	2686	6	4.1	4	1	0	1
31–35	7749	2	3.1	2	0	0	0
36–40	15,651	7	2.1	6	1	0	0
41–45	12,474	18	0.5	18	0	0	0
46–50	5337	3	0.1	2	1	0	0
51–65	32,460	0	0.1	0	0	0	0
Total	84,086	1005	1012.0	503	355	55	92

N: number of bases, obs err: observed errors, Exp err: expected errors, PrimFwd: primer chemistry forward, PrimRev: primer chemistry reverse, TermFwd: terminator chemistry forward, TermRev: terminator chemistry reverse.

in the forward direction. In all other positions, with both sequencing chemistries and directions, the q-values were highly accurate (data not shown).

4. Discussion

In the first analysis, base caller induced deletions were given the same q-value as the neighboring base of the lowest quality, while in the second analysis, deletions were given a q-value of zero. This improved Phred's predictions considerably. This indicates that Phred is not accurate when it comes to detecting base caller induced deletions. However, repeated sequencing can compensate for this inaccuracy as it in our experience is unlikely that a base caller induced deletion occurs at the same position using different sequencing chemistries and directions.

In the third analysis, all sequences with length heteroplasmy were removed from the data-set. This also improved Phred's predictions considerable. This result indicate that Phred is not well suited for identification of heteroplasmic bases. This insufficiency could be compensated for by other types of software, e.g. the secondary peak finder in Sequencher.

References

[1] B. Ewing, L. Hillier, M. Wendl, P. Green, BaseCalling of automated sequencher traces using Phred: I. Accuracy assessment, Genome Research 8 (1998) 175–185.

[2] B. Ewing, P. Green, BaseCalling of automated sequencher traces using Phred: II. Error probabilities, Genome Research 8 (1998) 186–194.