BIOINFORMATICS

Vol. 20 no. 17 2004, pages 3005–3012
doi:10.1093/bioinformatics/bth346

Quantitative DNA methylation analysis based on four-dye trace data from direct sequencing of PCR amplificates

Jörn Lewin[1,*], Armin O. Schmitt[2], Péter Adorján[1], Thomas Hildmann[1] and Christian Piepenbrock[1]

[1]Epigenomics AG, Kleine Präsidentenstrasse 1, 10178 Berlin, Germany and [2]Institut für Nutztierwissenschaften, Humboldt-Universität zu Berlin, Invalidenstrasse 42, 10115 Berlin, Germany

Received on March 24, 2004; revised on May 6, 2004; accepted on May 11, 2004
Advance Access publication July 9, 2004

ABSTRACT

Motivation: Methylation of cytosines in DNA plays an important role in the regulation of gene expression, and the analysis of methylation patterns is fundamental for the understanding of cell differentiation, aging processes, diseases and cancer development. Such analysis has been limited, because technologies for detailed and efficient high-throughput studies have not been available. We have developed a novel quantitative methylation analysis algorithm and workflow based on direct DNA sequencing of PCR products from bisulfite-treated DNA with high-throughput sequencing machines. This technology is a prerequisite for success of the Human Epigenome Project, the first large genome-wide sequencing study for DNA methylation in many different tissues. Methylation in tissue samples which are compositions of different cells is a quantitative information represented by cytosine/thymine proportions after bisulfite conversion of unmethylated cytosines to uracil and PCR. Calculation of quantitative methylation information from base proportions represented by different dye signals in four-dye sequencing trace files needs a specific algorithm handling imbalanced and overscaled signals, incomplete conversion, quality problems and basecaller artifacts.

Results: The algorithm we developed has several key properties: it analyzes trace files from PCR products of bisulfite-treated DNA sequenced directly on ABI machines; it yields quantitative methylation measurements for individual cytosine positions after alignment with genomic reference sequences, signal normalization and estimation of effectiveness of bisulfite treatment; it works in a fully automated pipeline including data quality monitoring; it is efficient and avoids the usual cost of multiple sequencing runs on subclones to estimate DNA methylation. The power of our new algorithm is demonstrated with data from two test systems based on mixtures with known base compositions and defined methylation. In addition, the applicability is proven by identifying CpGs that are differentially methylated in real tissue samples.

Contact: joern.lewin@epigenomics.com

INTRODUCTION

DNA methylation is a chemical modification of the base cytosine to 5′-methyl cytosine. In DNA of vertebrates it occurs in the context of cytosines followed by guanine, so-called CpGs. CpG methylation in human DNA is a tissue specific layer of information that is involved in the regulation of gene expression, genomic imprinting (Reik *et al*., 2001) and cell differentiation (Ehrlich, 2003). Methylation profiles undergo changes in tumorgenesis (Jones, 2002) and allow differentiation of DNA from healthy versus malignant tissue samples (Adorjan *et al*., 2002). Differential methylation patterns are likely to have a large relevance for understanding disease and diagnostic applications.

The main goal of the Human Epigenome Project (Human Epigenome Consortium *et al*., 2003, http://www.epigenome.org/) is to characterize the methylation signatures of different tissue types genome wide. This approach requires methods that easily detect methylation patterns by automated high-throughput technologies.

Different methods for methylation measurement are described in Dahl and Guldberg (2003), Siegmund and Laird (2002). One major group of technologies is based on methylation sensitive enzymatic restriction of the DNA. The other group of technologies is based on bisulfite conversion of unmethylated cytosines (Olek *et al*., 1996). The Bisulfite treatment of DNA leads to a chemical conversion of unmethylated cytosine to uracil. Methylation of cytosines blocks this reaction. In most cases, the PCR is used to amplify regions of interest within the bisulfite converted DNA template whereby positions converted into uracil appear as thymine in the product. Typically, a tissue sample contains a mixture of different cells; therefore, a proper description of methylation

*To whom correspondence should be addressed.

at a certain CpG requires quantification of the proportion of the methylated templates at the investigated CpG. This proportion is referred to as the methylation rate of the CpG. After the bisulfite conversion and the PCR, the methylation rate at a CpG can be determined by assessing the proportion of remaining cytosine relative to the thymine. This can be done, e.g. by hybridization to oligomer probes on DNA chips (Adorjan *et al*., 2002) or by DNA sequencing (Frommer *et al*., 1992). Commonly used sequencing methods include the sequencing of a representative number of subclones of the PCR product or direct PCR sequencing by running independent sequencing reactions for cytosine and thymine using the same dye in different lanes of a sequencing gel (Paul and Clark, 1996). These sequencing methods are expensive and labor intensive. In the Human Epigenome Project, direct PCR sequencing on standard sequencing machines is used to achieve the required throughput in a cost effective way. This technology produces four-dye electropherogram data. The possibility to use such data for quantitative analyses of base compositions within pooled DNA was recently demonstrated for one single nucleotide polymorphism (SNP) (Qiu *et al*., 2003). The same principle is used here for the measurement of methylation in bisulfite-treated DNA product.

Quantitative analysis by direct sequencing of the PCR products from bisulfite-treated DNA implicates several novel challenges: poor signal quality compared to genomic sequencing, overscaled cytosine signals and basecaller artifacts. In combination with the overscaled signals incomplete bisulfite conversion, which is a general problem of all bisulfite-based methylation detection methods, influences signal proportions in the trace significantly. It was therefore necessary to develop a specific algorithm that allows the use of four-dye sequencing trace files to gain quantitative methylation information. This newly developed data analysis method allows the use of established high-throughput sequencing technology for methylation studies. In this paper, we first present the algorithms used for methylation rate estimations based on trace file data originating from direct sequencing of the PCR products from bisulfite-treated DNA. We then assess the two main steps of our algorithm with real data from two experiments and show that they improve the accuracy of the methylation estimation. Finally, we provide a single example based on data from the Human Epigenome Project pilot study to demonstrate the scientific use of the algorithms with real tissue samples.

ALGORITHM

The algorithm we present uses four-dye electropherogram data preprocessed by the base caller of the sequencing machine manufacturer, e.g. Applied Biosystems '.abi' files or the well-described '.scf' files (Dear and Staden, 1992). The data processing includes the following steps: (i) entropy-based clipping, (ii) signal detection, (iii) alignment, (iv) trace correction, (v) alignment-based clipping, (vi) signal normalization, (vii) compensation of incomplete conversion and (viii) methylation estimation (Fig. 1). A scheme of the data and the influence of the algorithmic steps (ii), (iii), (iv) and (vi) is given in Figure 2. Here, we present the algorithms for forward sequencing that aims at the estimation of the proportion of cytosine to thymine at the positions of interest. Traces that originate from reverse sequencing and show guanine and adenine signals at corresponding positions can be analyzed by the same algorithm building the reverse complement of the trace files.

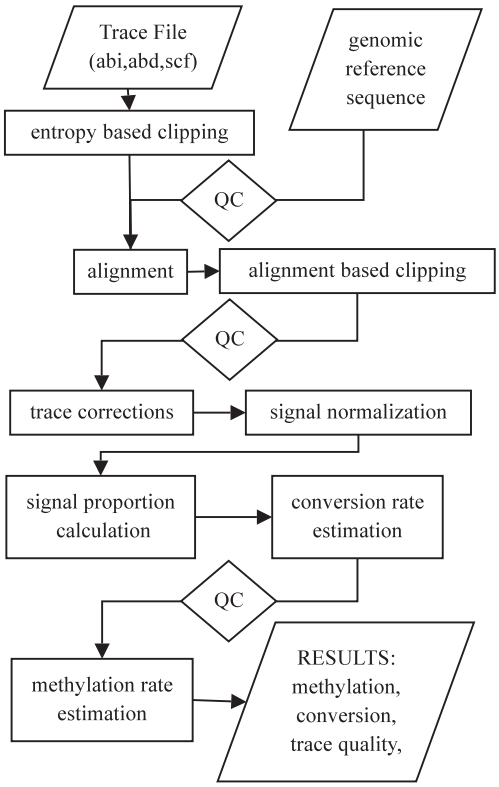

Fig. 1. Flow chart of all data processing steps of the methylation estimation algorithm. Detailed description of the single steps is given in the text. Between all data processing steps quality control (QC) is performed. The analysis of a single trace file is aborted if the file itself is corrupted or if the genomic reference sequence is missing or if the length of good quality sequence, as determined by the clipping procedure, is below a certain threshold (default is 50 bases) or if the bisulfite conversion rates are below a minimum threshold (default is 65%).

(i) *Entropy-based clipping:* We observed that base callers often generate reads that contain long stretches of called bases with up-scaled background signals after the end of an amplificate. These artifacts are detected by using the normalized Shannon entropy

$$H = -\sum_{b\in\{A,C,G,T\}} \left(\frac{S_b}{\sum_{B\in\{A,C,G,T\}} S_B} \log_4 \frac{S_b}{\sum_{B\in\{A,C,G,T\}} S_B} \right) \quad (1)$$

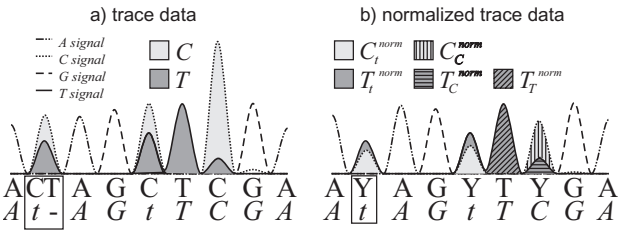

Fig. 2. Schematic representation of a trace file electropherogram obtained by bisulfite PCR sequencing (**a**) before and (**b**) after signal normalization. The upper sequences below the trace curves in (a) represent the sequence called by the standard basecaller and in (b) the peak mixture represented using IUPAC code (*Y* denotes *C* and/or *T*). The sequences at the bottom show the aligned reference sequence whereby *t* are genomic cytosine positions that are not in CpG context, and expected to be unmethylated and therefore completely convertible. Trace curves are shown for all the four bases. For every base position in the reference sequence four base intensities $B^{\text{int}}; B \in \{A, C, G, T\}$ are calculated as the area under the trace curve segment that belongs to the base position (only C^{int} and T^{int} shown in a)). Normalized base intensities for cytosine ($C_b^{\text{norm}}; b \in \{t, T, C\}$) and thymine ($T_b^{\text{norm}}$) seen in (b) are used to estimate the bisulfite conversion rate (base intensities at *t* positions) and the methylation rate at each CpG (base intensities at *C* positions).

$0 \leq H \leq 1$ of the four trace curves S_b, $b \in \{A, C, G, T\}$ in a sliding window of 200 data points. Flanking sequence stretches with an entropy larger than 0.8 are removed.

(ii) *Signal detection:* For each base position in the trace file, we compute corresponding intensities $B^{\text{int}}; B \in \{A, C, G, T\}$ that estimate the base proportions in the molecular mixture. As an appropriate measure we have chosen the areas under the trace corresponding to the respective base for each position in the sequence. By default, the trace segment between neighboring local minima is used for the signal area estimation. If no local minima are present, then the boundaries of the trace segment are estimated as the midpoint between two neighboring inflection points.

(iii) *Alignment:* The base intensities estimated in the previous step are then mapped to an underlying reference sequence, available as genomic sequence from database sources and bisulfite converted *in silico*. The a priori availability of the genomic sequence is a prerequisite for our application. To describe an expected bisulfite converted reference sequence, the commonly used genomic alphabet (*A*,*C*,*G*,*T*) is extended by one letter, the lower case *t*, to distinguish a thymine derived from uracil by bisulfite conversion from a thymine that was present already in the genomic sequence. Cytosines in a CpG context in the reference sequence correspond to positions where we want to quantify unknown methylation, and are therefore still denoted by *C*. For the sake of clarity in the notation, these positions should be distinguishable from *t*, where the sequenced DNA is never methylated and therefore, expected to have a complete conversion by the bisulfite treatment. We use the Smith–Waterman algorithm (Smith and Waterman, 1981; Barton, 1993) for optimal local alignments allowing for gaps to align the called sequence of the trace file with the a priori known reference sequence. Alignment of *t* and *C* in the reference with *C* or *T* in the trace are treated as matches.

The bisulfite-treated DNA contains long stretches of *T* signal. In some cases, this is misinterpreted by base callers by inserting too many *T*s into the called sequence. Accounting for this special situation, we have introduced an additional type of gap cost to guarantee proper mapping of CpGs. Assigning costs for gaps between C and G in the reference sequence forces the alignment of CpGs as one functional block to avoid their mismapping. An example of this is given below: general costs for all gaps (g) are −19 and higher than costs for mismatches (−9) (Barton, 1993). For gaps inserted between C and G in the reference sequence special additional gap costs(sg) of −20 raise the total costs to −39, a punishment outnumbered only by two gaps (−40) which in most cases leads to CpGs treated as one unit that cannot be split, just misaligned.

```
trace      ATTTTTTTGA          ATTTTTTTGA
reference  ATTTTTC-GA          ATTTTT-CGA
           cost(g+sg)=-39      cost(g)=-19
```

(iv) *Trace correction:* Standard base callers expect one homogeneous DNA population to be sequenced, therefore some of them occasionally interpret mixed C and T base intensities at a single position of the reference sequence as two adjacent bases, mostly if there is a small offset of one or two data points between C and T signals. In contrast to standard sequencing, in our experiments we expect signal mixtures from different DNA populations. It follows that the separation of overlaying intensities belonging to one position into two bases by the base caller has to be corrected. We identify the separated base intensities by searching adjacent T and C positions in the called sequence from which one is aligned with *t* or *C* and the other is introducing a gap into the reference sequence. These base pairs in the called sequence are then fused into a single base.

(v) *Alignment-based clipping:* The quality of trace files from PCR product sequencing, especially of amplificates from bisulfite-treated template containing different molecule populations, is lower than sequences from a homogeneous clone template. Alignment quality as a natural measure to assess sequencing quality is used to identify areas of poor quality. Flanking regions of the sequence are clipped such that the remaining inner part has <10% alignment error to the reference sequence.

(vi) *Signal normalization:* We found that cytosine trace curves often are overscaled in direct bisulfite sequencing traces[1]. Base proportion calculation based on trace curves with different baseline intensities would lead to misleading

[1]We speculate that this overscaling is a result of the standard basecaller software compensating for the low frequency of *C* signals.

results. Therefore, we normalize the trace curves prior to calculating the proportions of base intensities to determine bisulphite conversion and methylation rate. The normalized base intensities are denoted by B_b^{norm} ; $B \in \{A, C, G, T\}$; $b \in \{C, t, T\}$ that fulfill constraints (2) and (3) based on average base intensities.

$$\overline{T_T^{\text{norm}}} \equiv \overline{T_C^{\text{norm}} + C_C^{\text{norm}}}. \quad (2)$$

$$\overline{T_T^{\text{norm}}} \equiv \overline{T_t^{\text{norm}} + C_t^{\text{norm}}}. \quad (3)$$

Normalization of C^{int} is performed by multiplication of a global factor F_C.

$$C_b^{\text{norm}} = F_C C_b^{\text{int}}, b \in \{C, t, A, G, T\} \quad (4)$$

Based on the data we use different strategies for normalization. If there are at least three C positions with $C_C^{\text{int}} > T_C^{\text{int}}$ normalization is based on data from these positions [Equation (5) following from Equation (2)]. Otherwise normalization is based on all t positions [Equation (6) following from Equation (3)]. In rare cases when all cytosines were unmethylated and converted completely ($C_C^{\text{int}} = 0$) normalization of the cytosine trace curve is impossible and unnecessary.

$$F_C = \frac{\overline{T_T^{\text{int}}} - \overline{T_C^{\text{int}}}}{\overline{C_C^{\text{int}}}}. \quad (5)$$

$$F_C = \frac{\overline{T_T^{\text{int}}} - \overline{T_t^{\text{int}}}}{\overline{C_t^{\text{int}}}}. \quad (6)$$

(vii) Compensation of incomplete conversion.

(viii) *Methylation estimation:* Cytosine base intensity at CpG positions can arise from two sources: from a population of methylated cytosines in the sample DNA and from an incomplete conversion reaction. It follows that the bisulfite conversion rate has to be first estimated to obtain a correct estimation of the methylation rate in the sample DNA. For an individual t the conversion rate R is estimated by

$$R = \frac{T_t^{\text{norm}}}{T_t^{\text{norm}} + C_t^{\text{norm}}}. \quad (7)$$

Local R_{loc} and global conversion rates R_{glob} can be determined by averaging over R of individual bases within defined ranges. Then the methylation rate $M, 0 \leq M \leq 1$, at a certain CpG can be estimated by using the following simple linear relationship

$$T_C^{\text{norm}} = R_{\text{glob}}(1 - M)(T_C^{\text{norm}} + C_C^{\text{norm}}). \quad (8)$$

The equation describes the fact that T base intensity at a C position T_C^{norm} is expected to arise from the unmethylated portion of the sample DNA that is bisulfite converted by rate R. Furthermore, the sum of the base intensities $T_C^{\text{norm}} + C_C^{\text{norm}}$ is assumed to be proportional to the total of cytosines in the sample DNA. It follows that the methylation rate then can be estimated by incorporating a correction for the incomplete bisulfite conversion

$$M = 1 - \frac{T_C^{\text{norm}}}{(C_C^{\text{norm}} + T_C^{\text{norm}})R_{\text{glob}}}. \quad (9)$$

Signal variance, artifacts or errors in the normalization might lead to negative methylation estimation which is set to 0.

EXPERIMENTAL SETUP

We performed three series of experiments to assess the analytical performance of our algorithm. We have investigated (i) the estimation of cytosine/thymine signal proportions, (ii) the estimation of methylation rates and (iii) the detection of differential methylation using real tissue samples.

Test system with known cytosine/thymine proportions

To test how accurate we can measure base proportions in four dye trace data and if our normalization algorithm improves measurements, we created an artificial test system with known cytosine/thymine proportions. A 669 bp long fragment in the promoter region of the gene G6e was amplified by the PCR after bisulfite treatment of the template DNA. The bisulfite reaction was setup such that the conversion of cytosines were not perfect. The PCR product was subcloned into pCR2.1-Topo vector (invitogen). The 96 clones were sequenced. Out of the 96 clones, 3 showing differences at the most positions of genomic cytosine were chosen. The plasmid concentrations of the three stocks were adjusted to the same level. To gain different cytosine/thymine base compositions volumes were mixed in all six permutations of the proportions 1:2:4. These mixtures contain molecules with cytosine and thymine at the original genomic cytosine positions with expected cytosine/(cytosine + thymine) ratios from 0 to 1 in 1/7 steps. Sense strands of the clone mixtures were sequenced five times using the kit 1.1 on the ABI PRISM 310 (Fig. 3a). Trace files were analyzed by using the ABI basecaller software 310POP4. Our algorithm was then used to estimate base compositions at each original genomic cytosine position. Estimated values were binned by their expected cytosine/(cytosine + thymine) ratios to assess their distributions and the mean absolute errors.

Test system with known methylation rates

To test our algorithm on data from DNA with defined methylation status, unmethylated human genomic DNA (Molecular Staging) was divided into two equal volumes. The DNA in one of the volumes was enzymatically methylated with methylase SssI (NEB) following the manufacturer's protocol. Volumes of methylated and unmethylated DNA were mixed in 20% steps from 0 to 100%. The PCR for 60 amplificates was performed on a Tetrat MJ-research PTC-225. For cycle sequencing, the forward PCR primer was used with ABI kit 1.1 and run on the ABI 3730 DNA analyzer (Fig 3b).

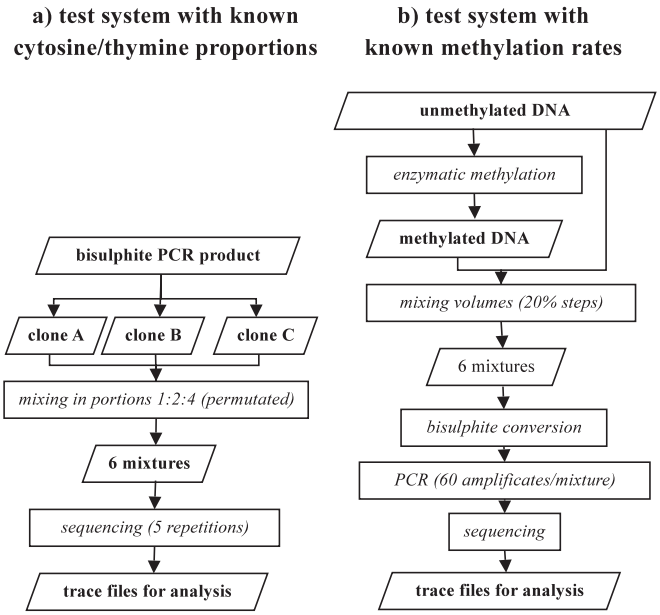

Fig. 3. Experimental setup of (**a**) a test system with known cytosine/thymine proportions (**b**) a test system with known methylation rates. Steps that are potential sources for variances or biases in the test systems like mixing steps, incomplete enzymatic methylation, PCR bias and variance, incomplete bisulfite conversion and variance in the sequencing procedure are typeset in italics.

Trace files were called with ABI's basecaller 3730POP7. Our algorithm was then used to estimate the methylation rates at each CpG position. Methylation rates were binned together by their expected methylation rates and variances and mean absolute errors were assessed.

Methylation in tissue samples

Methylation was estimated using trace files from direct PCR of bisulfite-treated DNA from healthy tissue samples. The data are a subset of trace files from the Human Epigenome Project pilot study by (Human Epigenome Consortium *et al.*, 2003).

RESULTS AND DISCUSSION

Test system with known cytosine/thymine proportions

To assess the effect of our signal normalization step, we used our algorithms on data from the test system with known cytosine/(cytosine + thymine) ratios. Figure 4a and b show the distribution of the estimated ratios against the expected ratios in the test system without and with normalization, respectively. The results demonstrate that the normalization step decreases the mean absolute error (represented by the dashed line on the figures) approximately to the half. Sequencing several subclones from a PCR product is an alternative method to measure the cytosine:thymine ratios in bisulfite-treated DNA. The measurement error of this method depends mainly on the number of subclones that is sequenced. We benchmarked our

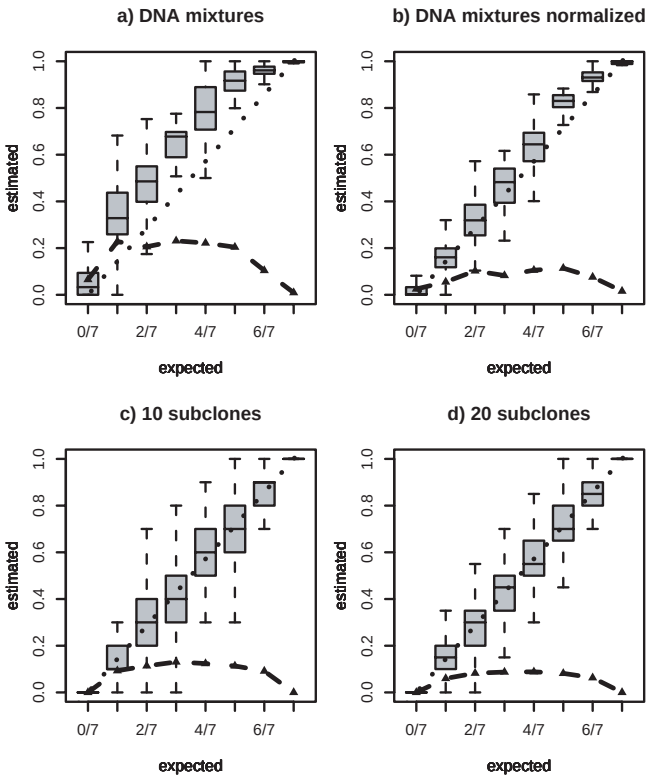

Fig. 4. (**a**) and (**b**) Quantitative measurements of C signal proportions in data from single sequencing runs of six clone mixtures with expected C/(C+T) ratios from 0 to 1 in 1/7 steps. The boxplots show the distribution of the estimated values obtained by our algorithm without normalization and with normalization, respectively. The estimates are plotted against the expected ratios (1039 data points total which means a measurement success rate of 89% given 6 mixtures, 5 repetitions and 39 positions). Dashed graphs show the means of absolute errors. (**c**), (**d**) Simulated data for representations of mixed DNA in 0 to 1 in 1/7 steps by 10 and 20 subclones based on a binomial distribution.

direct sequencing method with the subcloning method. We calculated the smallest theoretical measurement error inherent to the subcloning method by simulating the subsampling of 10 and 20 subclones based on binomial distributions with a certain C:T ratio. Figure 4c and d and Table 1 show that errors in our estimates are comparable with those that could be obtained by sequencing 20 subclones of a PCR product. From this we can conclude that direct sequencing of the bisulfite-treated DNA is a viable alternative to the subclone sequencing if only the mixture rates are the subject of interest.

Test system with known methylation rates

Second, we have evaluated the performance of our algorithm by using the test system with known methylation rates. Figure 5d shows the distribution of the estimated methylation rates against the expected methylation rates in the test system using the complete algorithm. For the estimation of methylation rates, normalization and the correction for

Table 1. Comparison of mean SD and absolute errors of C/(C+T) signal proportions as estimated in our test system with known cytosine/thymine proportions and simulated representation by subclones

	Mean SD	Mean absolute error
Signal proportions	0.110	0.130
Normalized signal proportions	0.077	0.055
10 subclones	0.100	0.083
20 subclones	0.072	0.058

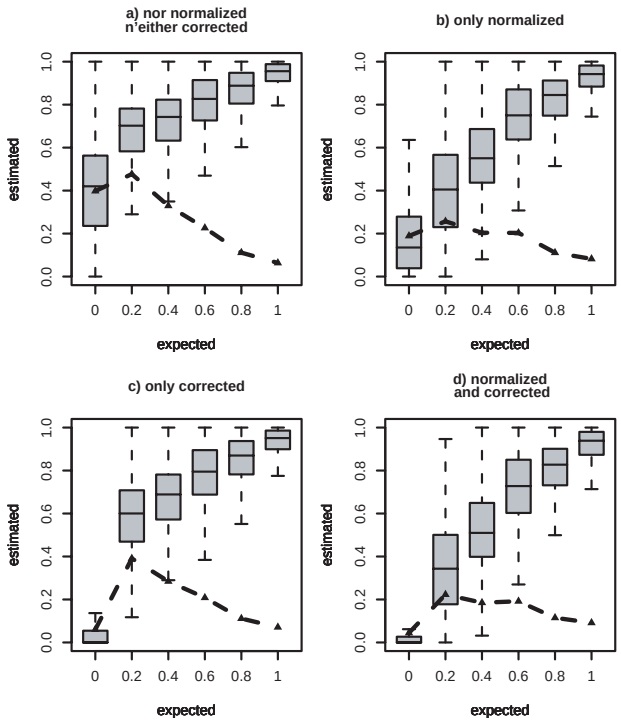

Fig. 5. Estimation of methylation in the test system with known methylation rates. The boxplots show the distribution of the estimated methylation rates as a function of the expected methylation and the mean absolute error (dashed line). Each box includes data from CpGs of all 60 amplificates measured at the expected methylation rate.

bisulfite conversion rate play an important role. This is demonstrated on Figure 5a–c and Table 2. If any of these steps is omitted from the data analysis, then the mean absolute error (dashed line) increases significantly. The normalization has a major impact mainly on low methylation rates, where the absence of C signals leads to an overscaling of the C trace.

The methylation rates estimated in this experiment do not show as accurate correlation with the expected rates as was obtained in the previous C:T proportion experiments, where a mixture of subclones was used as a test system. One possible explanation for this is that the real methylation rate in the mixtures of methylated and unmethylated DNA deviates from the expected methylation rate. Systematic biases in the real values of all 60 covered regions can arise from incomplete enzymatic methylation of the DNA or from amplificate specific biases in the PCR itself.

Table 2. Test system with known methylation rates: mean absolute errors of methylation estimations in 60 amplificates with and without signal normalization and conversion rate correction

Method	Mean absolute error
Raw data	0.27
Normalized	0.17
Corrected	0.19
Normalized and corrected	0.14

Table 3. Test system with known methylation rates: accuracy of sorting paired methylation estimates at identical CpGs in 60 amplificates after normalization and conversion rate correction

Expected rate	0.2	0.4	0.6	0.8	1
0	91	98	97	99	99
0.2		90	98	99	99
0.4			96	97	98
0.6				79	88
0.8					89

Systematic biases in the test system would lead to deviations from expected values and to a higher variance in the complete data but still allow to detect relative differences in the methylation rates at individual CpG positions. To evaluate the capability of our method to detect differential methylation, we paired data from templates with different methylation values for each CpG. Table 3 lists the accuracy of classification of higher versus lower methylated CpGs in the test system. The accuracies for detecting differential methylation in neighboring methylation rates with 20% steps are compared with those that were obtained without normalization and correction for incomplete bisulfite conversion (Table 4). The performance clearly improves by using the normalization and the conversion rate correction steps.

Despite the overlap of the distributions of the estimated methylation values (cf. Fig. 5d), we can conclude that the detection of differential methylation is highly accurate. This is in accordance with our hypothesis of having amplificate specific systematic biases in our reference test system.

We have evaluated threshold parameters for quality control. More stringent parameters do not improve the results significantly but lead to lower measurement success rates. For example, raising the threshold for bisulfite conversion from 65 to 80% reduces the mean absolute error by 0.2% and raises the accuracy by 1.3% but reduces the number of accessible positions by 16%.

Table 4. Test system with known methylation rates: accuracy of sorting paired methylation estimates at identical CpGs in 60 amplificates with 20% difference with and without using the normalization and correction for incomplete bisulfite conversion

Comparison	Correct sorting (%) Raw	Norm/corr.
0/0.2	84	91
0.2/0.4	71	90
0.4/0.6	86	96
0.6/0.8	77	79
0.8/1	89	89

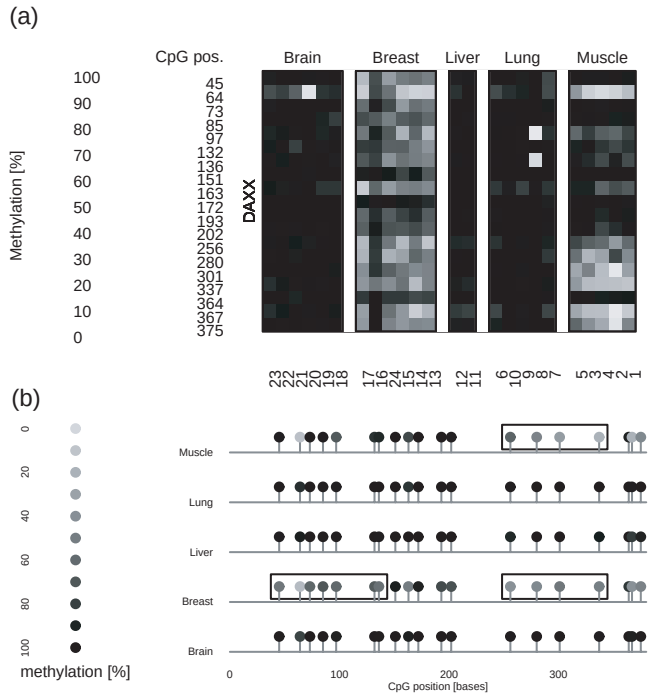

Fig. 6. (**a**) Methylation profiles of the intragenic region of gene DAXX obtained from DNA samples extracted from brain, breast, liver, lung and muscle tissue samples. The gray shading represents the different methylation rates as indicated on the scale bar. Samples from different individuals are arranged in columns, while each row represents one CpG within the amplificate. (**b**) Physical distances and average methylation of CpGs within the tissues. Blocks of differentially co-methylated CpGs are highlighted.

Methylation in tissue samples

Trace files produced in the The Human Epigenome Project by Human Epigenome Consortium *et al*. (2003) are processed by the algorithm presented here. We present a dataset to demonstrate the capability of our method to detect differential methylation in real tissue samples. Figure 6 shows the methylation profiles obtained in brain, breast, liver, lung and muscle samples by bisulfite sequencing the intragenic region of the gene DAXX. The plot shows clear differential methylation in blocks of co-methylated CpGs that distinguishes breast from muscle and both from all other tissues.

CONCLUSION

We have presented an algorithm to estimate methylation from trace files generated by direct PCR sequencing of bisulfite-treated DNA. The Results obtained by reference test systems show that direct PCR sequencing is a viable alternative to estimating methylation rates by sequencing subclones from the PCR product. Furthermore, we have demonstrated that our method can detect differences in methylation rates of 20% highly accurately. Applying our algorithms to bisulfite sequencing data of DNA obtained from healthy tissue samples illustrated that by the aid of the method, CpGs with differential methylation rates between different tissue types can be identified.

The algorithm allows to run big DNA methylation studies like the Human Epigenome Project based on direct sequencing in high-throughput facilities. It will help to gain information about differential methylation in many tissue types and increase our understanding of the epigenetic layer in the complex system of gene expression, cell differentiation and tumorgenesis.

ACKNOWLEDGEMENTS

We acknowledge Matthias Schuster, Christoph König and Bülent Genç for experimental planning and Erik Leu for his great laboratory work that provided most data used in this paper. We thank Kurt Berlin, Stephan Beck and Karen Novik for the establishment of the Human Epigenome Project. The first author also wants to thank Jörn Walter for his kind support of his PhD thesis. Some data shown in this paper was provided by the Human Epigenome Project pilot study supported by the EU (QLRT1999-30417).

REFERENCES

Adorjan,P., Distler,J., Lipscher,E., Model,F., Muller,J., Pelet,C., Braun,A., Florl,A.R., Gutig,D., Grabs,G. *et al*. (2002) Tumour class prediction and discovery by microarray-based DNA methylation analysis.*Nucleic Acids Res.*, **30**, e21.

Barton,G.J. (1993) An efficient algorithm to locate all locally optimal alignments between two sequences allowing for gaps. *Comput. Appl. Biosci.*, **9**, 729–734.

Dahl,C. and Guldberg,P. (2003) DNA methylation analysis techniques. *Biogerontology*, **4**, 233–250.

Dear,S. and Staden,R. (1992) A standard file format for data from DNA sequencing instruments. *DNA Seq.*, **3**, 107–110.

Ehrlich,M. (2003) Expression of various genes is controlled by DNA methylation during mammalian development. *J. Cell. Biochem.*, **88**, 899–910.

Frommer,M., McDonald,L.E., Millar,D.S., Collis,C.M., Watt,F., Grigg,G.W., Molloy,P.L. and Paul,C.L. (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands.*Proc. Natl Acad. Sci., USA*, **89**, 1827–1831.

Human Epigenome Consortium, Epigenomics AG, The Welcome Trust Sanger Institute and Centre National de Genotypage (2003) Human Epigenome Project.

Jones,P.A. (2002) DNA methylation and cancer. *Oncogene*, **21**, 5358–5360.

Olek,A., Oswald,J. and Walter,J. (1996) A modified and improved method for bisulphite based cytosine methylation analysis. *Nucleic Acids Res.*, **24**, 5064–5066.

Paul,C.L. and Clark,S.J. (1996) Cytosine methylation: quantitation by automated genomic sequencing and GENESCAN analysis. *BioTechniques*, **21**, 126–133.

Qiu,P., Soder,G.J., Sanfiorenzo,V.J., Wang,L., Greene,J.R., Fritz,M.A. and Cai,X.Y. (2003) Quantification of single nucleotide polymorphisms by automated DNA sequencing. *Biochem. Biophys. Res. Commun.*, **309**, 331–338.

Reik,W., Dean,W. and Walter,J. (2001) Epigenetic reprogramming in mammalian development. *Science*, **293**, 1089–1093.

Siegmund,K.D. and Laird,P.W. (2002) Analysis of complex methylation data.*Methods*, **27**, 170–178.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.