MethTools 2.0 - a webtool for the generation and analysis of DNA methylation patterns from bisulfite sequencing raw data

Jean-Luc Maspimby & Christoph Grunau (*)

UMR 5244 CNRS-EPHE-UPVD

Biologie et Ecologie tropicale et méditerranéenne

Université. Perpignan

58 Av. Paul Alduy

66860 Perpignan Cedex, France

Tel: 33 (0)4 68 66 21 80

Fax: 33 (0)4 68 66 22 81

(*) corresponding author

Abstract

DNA methylation is an important carrier of epigenetic information. Today, the majority of DNA methylation analyses rely on the bisulfite conversion technique. Treatment of genomic DNA with bisulfite and subsequent PCR of the region of interest delivers PCR products in which originally unmethylated cytosines occur as thymines and methylated cytosines as cytosines. Subcloning and sequencing of the PCR products can be used to deduce DNA methylation patterns at single-base resolution (bisulfite genomic sequencing). Here we describe the second generation of a web-server service that compares the sequence data of the PCR products with the genomic sequence, generates graphical representations of methylation patterns, methylation density and methylated consenus sequences, and performs a number of numerical analysis. The on-line tool is accessible directly at http://methdb.igh.cnrs.fr/methtools/, or via the "Links" section of the DNA methylation database (http://www.methdb.net).

Keywords

Epigenetics, DNA methylation, bisulfite treatment, webserver tool, remote data analysis, information theory

Introduction

Bisulfite genomic sequencing [1] is now a common tool for the analysis of DNA methylation. The method relies on the conversion of unmethylated cytosine residues in the genomic DNA into deoxyuracil by a treatment with concentrated sodium-bisulfite solution. In a second step, the genomic region of interest is amplified by PCR. PCR products are subcloned and sequenced. In the PCR products, unmethylated cytosines are displayed as thymines and methylated cytosines as cytosines. By comparison with the original sequence, the position of methylated and unmethylated cytosines, their distribution, average methylation and other parameters can be deduced.

Manual comparison of the sequences derived from bisulfite-treated DNA is error-prone and laborious. Six years ago we had therefore developed a suite of software tools for the analysis of such data [2]. The programs were made available for download, and in addition a web-server was established that allowed for remote analysis. The results were returned by e-mail. This service underwent silent updates in the past, but will not be maintained in the current form. In response to requests from users who for several reasons (in particular spam filters) could no longer use our MethTools in the previous form, we developed an improved version of the MethTools that is entirely web-server based. In addition to this conceptual change we added new features in graphical representation of the data and numerical analysis.

Description

Input

The submission form contains two fields: one for the email address of the user, and one for the sequence raw-data in fasta format [3]. The email address of the user is requested in case of unforeseeable problems with data submission. The input format of the sequence raw data has not changed for this new version of MethTools: it consists of a text-file containing

concatenated sequences in fasta format. The first sequence must be the unconverted genomic sequence. All following sequences correspond to PCR products obtained after bisulfite treatment of the genomic sequence and PCR of the region of interest. Their sequence must be aligned to the unconverted genomic sequence, and must be of the same length. Gaps can be represented by "-" and unknown bases by "n". The original genomic sequence must not contain gaps. Virtually all alignment programs can generate this type of fasta file. In our experience, PCR products from bisulfite treated DNA require some manual inspection and alignment with software that allow for accessing the trace data (e.g. the Staden package (https://sourceforge.net/projects/staden) or Sequencher (Gene Codes, Ann Arbor, Michigan)). All sequence data must be sent as a flat text file. Depending on the operating system and web-browser, it might be necessary to add the suffix ".txt" to the file name. An example input file can be downloaded from our web-page. Check-boxes allow for (de)selecting of particular analysis tools, and a number of different output formats for graphics can be chosen (gif, png, and svg). Gif and png are pixel formats that are adapted to on-screen display. Scalable Vector Graphics (svg) is a resolution-independent vector format that is suitable for representation of prints and posters, but finds also entry in many web-pages and can be used for mobile devices (http://www.w3.org/Graphics/SVG/).

Output

MethTools compares the sequences of the PCR products with the unconverted genomic sequence, and generates fasta files in which T in the PCR products that align to C in the genomic sequence are represented as unmethylated cytosines, and C (PCR) that align with C (genomic) as 5-Methyl-cytosines (5mC). In case of conflict, the sequence of the PCR product is chosen. The IUB Nomenclature Committee does not provide a code for 5mC or other modified bases (http://www.chem.qmul.ac.uk/iubmb/misc/naseq.html#502). We have therefore decided to represent 5mC as upper-case C and all other nucleotides in lower case.

These sequence files can be downloaded. They are also suitable for submission to the DNA methylation database MethDB [4].

Based on the results of this comparison, graphical representations are generated. Here, the principal novelty to the previous version of MethTools is that we offer a rich choice of different graphics formats including the pixel formats gif and png, and the vector formats svg, postscript and pdf. The latter are resolution-independent and can be used to create publication-ready figures. All graphic files can be edited with suitable editors (e.g. Adobe Illustrator for pdf, or the free and platform-independent software Inkscape for svg). If the option "table with information content and LOGO" is chosen, a so-called logo is generated that shows the methylated consensus sequence based on information theory [5].

MethTools 2.0 allows one to perform numerical analyses of the submitted raw data. It counts 5mC and unmethylated C for each individual PCR clone (i.e. the methylation patterns of individual cells) and calculates average methylation for each cell and for the entire data set. This feature is useful for generation of density distribution, variance analysis and the statistical analysis of differences between samples. Methylation profiles are generated with the average methylation at each C site in the sequence. Both files can be imported into spreadsheet software. If requested, an error report is generated that shows the number of unexpected base exchanges, which allows for evaluation of sequence quality.

All result files can be downloaded one-by-one, or as a set as compressed archive in zip format. Downloads are available for 24 hours after submission.

Discussion

The MethTools belong to a small but frequently used number of free web-tools specifically designed for the growing scientific community that works on DNA methylation. For the preparation of experiments, MethPrimer [6] (http://www.urogene.org/methprimer) and Snake-

Charmer (http://insilico.ehu.es/restriction/two_seq/snake_charmer.html) facilitate primer design for bisulfite PCR and the choice of restriction enzymes for combined bisulfite restriction analysis (COBRA). Our web-tool was developed to accelerate data analysis and visualization of the experimental results. To our knowledge, the only other software tool that can be used for this purpose is the BiQ Analyzer [7] (http://biq-analyzer.bioinf.mpi-inf.mpg.de/). Here, an alternative concept was developed: the software is java-based and must be installed locally, and a single fixed output format is offered. The MethTools are particularly useful to users who do not wish to install additional software and who prefer a flexible output. In order to make the web-tool available to a large scientific public, we have chosen to use a conservative user interface that works with all browsers that support the HTML 4.0 protocol. Future development will be done according to user requests. The underlying CGI scripts are organized in modules with standardized inputs, and further modules can easily be added under the same user interface.

References

[1] Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW, Molloy PL and Paul CL. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc Natl Acad Sci U S A 1992; 89: 1827-1831.

[2] Grunau C, Schattevoy R, Mache N and Rosenthal A. MethTools--a toolbox to visualize and analyze DNA methylation data. Nucleic Acids Res 2000; 28: 1053-1058.

[3] Pearson WR and Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A 1988; 85: 2444-2448.

[4] Grunau C, Renault E, Rosenthal A and Roizes G. MethDB--a public database for DNA methylation data. Nucleic Acids Res 2001; 29: 270-274.

[5] Schneider TD and Stephens RM. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res 1990; 18: 6097-6100.

[6] Li LC and Dahiya R. MethPrimer: designing primers for methylation. Bioinformatics 2002; 18: 1427-1431.

[7] Bock C, Reither S, Mikeska T, Paulsen M, Walter J and Lengauer T. BiQ Analyzer: visualization and quality control for DNA methylation data from bisulfite sequencing. Bioinformatics 2005; 21: 4067-4068.