

Classification of DNA methyltransferases with Profile Hidden Markov Models

Christian Rausch,¹ Alexander Thielen,² Daniel H. Huson³

Keywords: automatic classification of proteins, profile hidden Markov models, HMM, DNA methyltransferases, amino acid sequence motifs, automatic prediction by similarity, REBASE

1 Introduction.

We present a strategy and results of an automatic classification of DNA methyltransferases into their types and subtypes using profile hidden Markov models (HMMs) [1, 2] for each conserved sequence motif. The profiles were built based on hand-curated multiple sequence alignments of all currently in [5] available annotated DNA amino-methyltransferases.

2 Biological Background and Previous Work.

DNA methyltransferases are enzymes that specifically recognize a target sequence (typically four to six bases long) on a DNA double strand. They transfer a methyl group from S-adenosyl-L-methionine to a specific position on a specific base of their target sequence. These enzymes can be divided into two major classes, according to the position methylated. Either the methyl group is transferred onto a pyrimidine ring carbon yielding C5-methylcytosine (5mC) or onto exocyclic amino nitrogens, forming either N6-methyladenine (N6mA) or N4-methylcytosine (N4mC). With the help of structure-guided analysis T. Malone et al. [3] were able to detect nine conserved motifs, corresponding to motifs I to VIII and X previously defined in C5-cytosine methyltransferases. Based on the sequential order of these motifs it was possible to divide the amino methyltransferases into three groups (α , β , and γ), see Fig. 1.

3 Materials and Methods, Program, and Results.

We have compiled an up-to-date set of annotated methyltransferases based on REBASE data [5] and aligned the sequences with ClustalW [6] and T-Coffee[4] and manually checked that the motifs given by T. Malone et al. [3] aligned correctly. For each motif of each subset we have built profile hidden Markov models (HMM) [2, 1] with the program `hmmbuild` and `hmmcalibrate` of the HMMER package [2, 1]. We developed a Java program that detects these motifs in a given unknown methyltransferase and gives a prediction for the type and subtype based on the scores and order of the detected motifs. With the help of this program we are able to give predictions for unannotated methyltransferases in the protein sequence databases and to detect erroneous annotations in REBASE [5].

¹Center for Bioinformatics Tübingen (ZBIT), Tübingen University, Germany.
E-mail: rausch@informatik.uni-tuebingen.de

²Center for Bioinformatics Tübingen (ZBIT), Tübingen University, Germany.
E-mail: thielen@informatik.uni-tuebingen.de

³Center for Bioinformatics Tübingen (ZBIT), Tübingen University, Germany.
E-mail: huson@informatik.uni-tuebingen.de

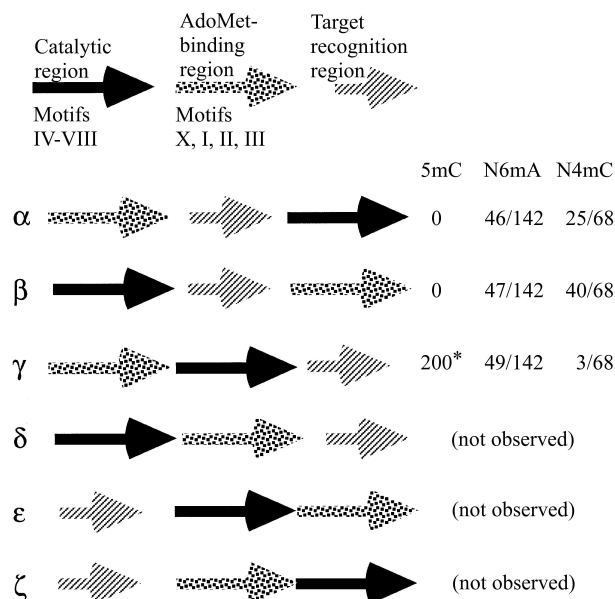


Figure 1: Possible arrangements of the three major regions found in DNA methyltransferases. The number of the currently annotated instances in REBASE is given. The asterisk (*) indicates that 5mC methyl transferases differ because they have the motif X near the C-terminus. (Graphic redrawn from [3].)

4 Availability of the Program.

The program will be freely usable through the webinterface at:
<http://www-ab.informatik.uni-tuebingen.de/software>.

References

- [1] Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. *Biological Sequence Analysis*. Cambridge University Press, 1999.
- [2] S. R. Eddy. Hmmer: Profile hidden markov models for biological sequence analysis. <http://hmmer.wustl.edu>, 2005.
- [3] T. Malone, R. M. Blumenthal, and X. Cheng. Structure-guided analysis reveals nine sequence motifs conserved among DNA amino-methyltransferases, and suggests a catalytic mechanism for these enzymes. *Nucleic Acids Research*, 253:618–632, 1995.
- [4] C. Notredame, D. G. Higgins, and J. Heringa. T-coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, 302(1):205–17, Sep 2000.
- [5] R. J. Robert, T. Vincze, J. Posfai, and D. Macelis. REBASE—restriction enzymes and DNA methyltransferases. *Nucleic Acids Research*, 33(Database Issue):D230–D232, Jan 2005.
- [6] J. D. Thompson, D. G. Higgins, and T. J. Gibson. improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22):4673–4680, Nov 1994.