CyMATE
Cytosine Methylation Analysis Tool for Everyone

CyMATE:
Cytosine Methylation Analysis Tool for Everyone

CyMATE is a software module for quick and extensive analyses of epigenetic properties, i.e., methylation patterns of DNA sequences. It handles bisulfite-converted samples ('clones') and evaluates the significance of polymorphic occurrences of patterns with respect to a reference sequence ('master'). CyMATE provides for both text and graphical output.

CyMATE is implemented in the programming language Python and uses the Cairo Graphics library to create PDF graphics. The module represents is compatible with the `vdiff` package, which is an equally fast and reliable application for detailed analysis of polymorphic sites in molecular sequence data (DNA and protein).

CyMATE analysis services are available via a web-based interface. It provides for a quick and comprehensive evaluation of considerably large data sets.

Keywords

Cytosine Methylation Analysis; Plant Epigenetics; Bisulfite Sequencing;

A detailed example of how CyMATE can be used is given in

> **Hetzl, J, Förster, A, Raidl, G, and Mittelsten Scheid, O**. "*CyMATE: a new tool for methylation analysis of plant genomic DNA after bisulfite sequencing*", 2007 (in preparation)

Contents

CyMATE
User Guide

Availability

The web interface for using CyMATE is available at http://www.gmi.oeaw.ac.at/CyMATE/.

The use of CyMATE is free of charge. To use CyMATE services, you have to specify your email address. This email address will be used to send you the results from CyMATE.

Input file submission is possible in common sequence and alignment formats. You may submit only one alignment file per request, but you can repeat the analysis and post multiple requests for different input files and analysis parameters. The number of posts is not restricted in terms of a fair use policy. Should you choose to use CyMATE results in a publication, please cite

> **Hetzl, J, Förster, A, Raidl, G, and Mittelsten Scheid, O**. "*CyMATE: a new tool for methylation analysis of plant genomic DNA after bisulfite sequencing*", 2007 (in preparation)

The results that will be sent to you per email include

- ➡ converted alignment file for repeated analysis
  (AFA; if not submitted)
- ➡ summary of analysis
  (TXT; if you are using Microsoft Windows, open with Microsoft Word).
- ➡ graphical representation of results
  (PDF)
- ➡ message (and conversion, if applicable) log file
  (TXT; if you are using Microsoft Windows, open with Microsoft Word)

All result files will be named automatically, using the same file name as for the specified input file, and the appropriate file extension.

Input files

Files must contain prealigned sequence data. Ambiguous base and gap characters are allowed, but are not treated as specific characters, thus are not included in the calculations.

Files can be submitted in either sequential (standard FASTA, indicated by `.fa`, `.fas`, or `.fasta` file extensions) or interleaved format (CLUSTAL, indicated by `.aln` file extension), as well as in CyMATE native AFA (`.afa` file extension) format.

CyMATE requires the sequences in the input files to be of equal length, i.e., sequence files must be submitted in prealigned form. Gaps and ambiguous bases are allowed. For details, see section **Input File Format**.

Bug reports and comments

You can submit comments and bug reports encountered during the use of CyMATE.

Please send your reports to the author (jennifer.hetzl@gmi.oeaw.ac.at), specifying the problem as detailed as possible. Please include your reply-to address and the respective log file to complement your report.

Input File Format

CyMATE only works on a set of aligned sequences, i.e., calculations are based on multiple sequence alignments of DNA sequences. This is required to correctly reference sequence positions, as CyMATE handles sequence information column-wise, not row-wise. CyMATE requires the following order of sequences:

1st sequence	the **master** sequence
2nd sequence	a **clone** sequence
3rd sequence	a **clone** sequence
...	...
nth sequence	a **clone** sequence

resulting in a total of one master sequence, and (*n* - 1) clone sequences to make up a total of *n* sequences.

Currently, the master sequence **must** come first, i.e., the master sequence must be the initial sequence in the alignment (file). Sequence alignment tools allow to specify a particular input order, so please confirm that the reference sequence always precedes all sample sequences to ensure correct results.

The author recommends either to use ClustalW or ClustalX [1,2] for creating sequence alignments, or to save the multiple alignment created with other tools in either FASTA or CLUSTAL format.

CyMATE reqires the sequences in the input files to be of equal length, i.e., sequence files must be submitted in prealigned form.
Alternatively, single sequences, e.g., homologous gene sequences, must be of equal length, and starting at the same position relative to a reference point to ensure correct results. The file extension for such concatenated sequences is `.seq` .

Gaps and ambiguous base characters are allowed.

---

[1] http://www.ebi.ac.uk/clustalw/

[2] **Higgins D., Thompson J., Gibson T., Thompson J.D., Higgins D.G., Gibson T.J. (1994)** *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice*. Nucleic Acids Res. 22:4673-4680.

The AFA file format

The AFA (‘almost fasta’, `.afa`) file format is the internal format CyMATE is processing. It has been derived from the common standard FASTA format and provides for a sequential representation of prealigned sequences, i.e., sequences of equal length. The CyMATE native AFA file format provides for a matrix-like representation of aligned sequences, with sequence labels delimited from molecular sequences.

The AFA file format requires ASCII text in a 2-column table structure, with colums separated by whitespace (one or multiple) characters or tabulators, and rows separated by a newline character. Content in the fields must not contain spaces or tabulators. The first column is reserved to hold the sequence label names, and the second column to hold the sequence data.

Prealigned sequences may contain leading, trailing, or interspersed gap characters. However, gap characters, namely dashes (‘–’), and ambiguous base characters are not treated as specific characters: these occurrences will be excluded from the calculations.

Example

```
MyMasterSeq          AAAACATATGCCGTAAAA
MyFirstCloneSeq      AAAACA---GCTGTAAAA
MySecondCloneSeq     AAAATAAATGTCGTAA--
MyThirdCloneSeq      --AATATATGCTGTAAAA
```

File conversion

Usage of CyMATE is possible after converting the input files to AFA. When submitting aligment files in another format than AFA, automatic file conversion is triggered, and for subsequent requests, an AFA file is appended to the list of result files that will be sent to you.

Prealigned sequence files provided in common sequence and alignment formats are recognized and converted. Sequence files either in sequential (e.g., FASTA) or interleaved format (e.g., CLUSTAL) will be read and converted to a compatible format ('almost fasta', AFA) for subsequent use with CyMATE. The conversion script recognizes files specified by the corresponding file type , i.e., extension.

Supported input file formats are:

- ➡ clustal: CLUSTAL alignment files (interleaved format)
- ➡ fasta: FASTA alignment files (sequential format)

Currently, no other file formats are specifically supported. Closely related file formats to `.fa`(`s`(`ta`)) and `.aln` may be read, but are not necessarily converted successfully.

Output Modes and Options

Methylation analysis mode

In methylation analysis mode, CyMATE evaluates the location, variant, and frequency of occurrence of methylation patterns in bisulfite-treated sample sequences ('clones') with respect to a master sequence.

In contrast to `vdiff`'s standard mode, only directional C/T transitions are considered as relevant positions: in standard mode, all bidirectional transitions and transversions are included into the analysis. If CyMATE finds that the number of relevant positions for variable columns is different than the number and location of methylation sites, an error message is triggered. Only when the two lists are identical, it can be assumed that a clean sequence set without artefacts has been submitted. The results from the analysis will not be affected from this consistency check, but a respective message will be included to the text result file.

Methylation patterns

A methylation pattern consists of two elements: the initiator element ('Y'), followed by two bases. Methylation patterns are classified according to these successor bases:

	1st base	2nd base	3rd base	Pattern class
Is pattern?	**'C' or 'T' ('Y')**	any base (='N')	any base (='N')	none
1st 'sieve'	'C' or 'T' ('Y')	**'G'**	any base (='N')	'YGN'
2nd 'sieve'	'C' or 'T' ('Y')	any base but 'G' (='H')	**'G'**	'YHG'
'Catch Basin'	'C' or 'T' ('Y')	any base but 'G' (='H')	any base but 'G' (='H')	**'YHH'**

The rule for classifying patterns at each level is indicated in bold face. Methylation patterns are allowed to overlap, e.g., 'CCA' at position 1, and 'CAG' at position 2 in a sequence 'CCAGTTTTTT'.
For each pattern class, the occurrences of patterns is analyzed separately, considering

- ➡ **Location:**
  the sequence position (index) where a pattern starts is returned
- ➡ **Variant:**
  the methylation state (methylated/unmethylated 'C') is returned
- ➡ **Frequency:**
  the methylation frequency of 'C' instances is returned

Remember that the pattern length is always three, so the last pattern that can be found starts at the (n - 2)nd position, when n is the length of the sequence.

CyMATE is not affected by false positive patterns or other sequence artefacts, since it identifies methylation sites with initiator elements 'C' in the master sequence only, not in clone sequences, and returns the results from evaluation of the clone sequence patterns with respect to the master sequence.

Sequences are allowed to have gap characters and ambiguous base characters. Naturally, these instances can **not** be handled as specific sequence information or specific positions. Those occurrences are excluded from per-column and per-sequence calculation of methylation frequencies, and graphical output ('empty' fields in the pattern matrix representation). Note that the comparison to other 'full information' columns or sequences may hence be biased, due to missing/ambiguous sequence information.

CyMATE methylation analysis generates two output files:

- ➡ the text output (TXT) files summarizes the findings about methylation sites, adn methylation frequencies
- ➡ the additional graphics file (PDF) visualizes the results for better perception of local and global frequency distribution. Graphical output is available for the entire sequence lenght or restricted regions within the sequence.

As in standard analysis and in query mode, input data must be prealigned sequences. However, to ensure correct processing in methylation analysis mode, the master sequence **must** be the first sequence in the file.

Options in methylation analysis mode

Several options are available in methylation analysis mode to customize your query. It is possible to combine the options.

Offset

Specify a positive (downstream from a reference point; the reference point is "to the left") or negative (upstream of a reference point; the reference point is "to the right") positional offset to use reference indices, e.g., relative to the transcription start of a gene.

Example: Sequence indices can be re-defined to match the position in the original sequence context, e.g.,

- ➡ a negative offset is applicable for promoter sites
- ➡ a positive offset is applicable for transcription factor binding sites

with repect to the transcription start.

Zoom-in

Restrict graphical output to 40 consecutive bases (not patterns), for a to-scale overview of a particular region of interest. The text output is not affected by this option. Currently, the given start position for zoom-in **must** be a methylation site, i.e., the position index must be one of the indices given in the “O” part of the test results.

Example: Zoom-in is recommended for positions with prior knowledge about methylation sites in a particular subregion of the sequence, e.g., transcription factor binding sites.

Extensions

`vdiff` is provided as the running environment for CyMATE. In the following, a summary of those relevant functions of the `vdiff` package is given that have proven useful for a more detailed analysis of the sequence alignment. These functions are also provided via the web interface. Should you choose to use `vdiff` results in a publication, please cite

> **Hetzl, J.** "*In silico analysis of canine and feline parvovirus evolution*", Master Thesis. University of Vienna, 2004.

Standard mode

In standard mode, `vdiff` reads prealigned sequences, and searches for variable [3], i.e., non-conserved positions. These positions are further considered as *relevant position*s and subsequently analyzed for character frequency distributions. `vdiff` analyzes the distribution of base or amino acid characters for each relevant position. The standard text output includes

- ➡ total number (and raw indices) of all variable positions
- ➡ total number and distribution of different characters (i.e., a list of taxa sharing identical positions) for all variable positions
- ➡ summary of specific positions for each taxon

The text output has the same structure as in methylation analysis mode with CyMATE. Currently, no graphics file is included in the results form standard mode.

In standard mode, two basic operations are considered:

Raw indices: Without specifying an upstream ("-n") or downstream offset ("n"; see section **Options in standard mode**), the first index is always designated position "1". Without using a specific positional offset, indices are referred to as "raw indices".

Group analysis: If specified, `vdiff` performs group analysis for relevant columns. Based on split information for columns, `vdiff` evaluates the significance of the same split occurring at other relevant positions in the same data set. For details, see section **Options in standard mode**.

Options in standard mode

Several options are available in standard analysis mode to customize your analysis. It is possible to combine the options.

---

[3] Variable positions identified by `vdiff` are not necessarily parsimony informative sites.

Offset (Integer)

Specify a positive (downstream to a reference point; reference point it “to the left”) or negative (upstream to a reference point; reference point is “to the right”) positional offset to use reference indices, e.g., relative to the transcription start of a gene.

Example: Sequence indices can be re-defined to match the position in the original sequence context, e.g.,

- ➡ a negative offset is applicable for promoter sites
- ➡ a positive offset is applicable for transcription factor binding sites

with repect to the transcription start.

Group-wise analysis (Boolean)

Select additional or exclusive group analysis to evaluate the significance of splits created by the character distribution at relevant positions in your data set. Splits are analyzed in order of their frequency of occurrence.

Query mode

In query mode, analyses are either processed by position or by name. Positional analysis allows to specify one or multiple positions to limit the results to one or several columns. Queries by name only prints the results for one particular sequence, i.e., indicates the characters found in this sequence at relevant positions.

Query mode results include

- ➡ total number (and raw indices) of all variable positions
- ➡ query results.

The text output has the same structure as in methylation analysis mode with CyMATE. Currently, no graphics file is included in the results form standard mode.

Two basic query types are handles in query mode:

Position-specific query: Limits the output to one or more particular sequence positions, considering all sequences in the set. Consideration of the sequence context is possible.

Sequence-specific query: Returns the specific, i.e., unique positions for one sequence in which the sequence differs from all other sequences. A threshold can be set to determine the maximum subset size for comparison with the sequence set.

Options in query mode

Several options are available in query mode to customize your analysis. It is possible to combine some of the options.

Offset (Integer)

Specify a positive (downstream to a reference point; reference point is “to the left”) or negative (upstream to a reference point; reference point is “to the right”) positional offset to use reference indices, e.g., relative to the transcription start of a gene.

Example: Sequence indices can be re-defined to match the position in the original sequence context, e.g.,

- ➡ a negative offset is applicable for promoter sites
- ➡ a positive offset is applicable for transcription factor binding sites

with repect to the transcription start.

Sequence context (Boolean)

For a positional query, select contextual analysis to include the following two positions to the position(s) specified in positional mode. **Not available for sequence-specific queries.**

Threshold (Integer)

For a sequence-specific query, set an additional threshold value to determine the maximum size of the subset in which the given sequence is found to differ from the rest of the sequence set. **Not available for positional queries.**

Combination (Boolean)

Combine queries by name and by position to use the full power of the query mode: re-use the results from a query by name, i.e., the specific positions for a sequence, for a positional query. **Not available for positional queries.**

Appendix: Summary of modes and options

CyMATE and `vdiff` allow to specify parameters that are best suited to solve your problem, and the modularity of the packages and modules ensure interoperability between modes. Currently, the following modes are explicitly supported:

- ➡ Standard analysis mode:
  perform variable site analysis in this mode
- ➡ Query mode:
  restrict your standard analysis to specific positions or sequences
- ➡ Methylation analysis mode:
  perform your bisulfite analysis of methylation sites in this mode

Each mode offers different options to perform specific and individual analyses. It is possible and encouraged to re-use results from one mode as input for another to fully benefit from CyMATE's and `vdiff`'s features.

Analysis mode → ↓ Available options	**Standard**	**Query**	**Methylation**
Group analysis	**+**		
Offset	**+**	**+**	**+**
Query by name		**+**	
Re-use position		**+**	
Subset threshold		**+**	
Query by position		**+**	
Sequence context		**+**	
Text output	**+**	**+**	**+**
Graphical output			**+**
Zoom-in			**+**