# MoDec

MoDec is a motif deconvolution software that finds the motifs and corresponding binding core offsets desribing the data in a list of peptides. It is described in the publication (available here):

Racle, J., et al. Robust prediction of HLA class II epitopes by deep motif deconvolution of immunopeptidomes. *Nat. Biotechnol.* 37, 1283–1286 (2019).

## Installation

1. Download MoDec-1.2.zip file and move it to a directory of your choice, where you have writing permissions.

2. Unzip MoDec-1.2.zip package.

3. (Optional) In order to make an html report of the runs from MoDec and to draw the logos from each motifs, you need R and some R-packages:

    1. Rscript needs to be in your path.
    2. You need to install the R-package *ggseqlogo*, based on the fork found at https://github.com/GfellerLab/ggseqlogo. For this, from within R, type the command: `devtools::install_github("gfellerlab/ggseqlogo")`
    3. You also need the R-package *htmlTable* that can be found at https://cran.r-project.org/web/packages/htmlTable/index.html. Command to install from within R: `install.packages("htmlTable")`
    4. You then finally need to open the file *make_logo_report.R*, locate the line starting with `.libPaths` in this file and replace the `"PATH/TO/R/LIBRARIES"` by the path where you installed these R-packages.

4. To test your installation, make sure you are in *MoDec-1.2* directory and run the following command, depending on your operating system:

    - Mac OS: `./MoDec -i test/testData.txt -o test/test_out --Kmax 5 --MHC2 --nruns 2 --makeReport`
    - Unix: `./MoDec_unix -i test/testData.txt -o test/test_out --Kmax 5 --MHC2 --nruns 2 --makeReport`
    - Windows: `MoDec.exe -i test/testData.txt -o test/test_out --Kmax 5 --MHC2 --nruns 2 --makeReport`

    Running the software can take few minutes depending on data size and number of motifs. If you didn't do the optional step (3) above, the `--makeReport` option shouldn't be used as MoDec won't be able to build this report.

    Your folder *test_out* should contain the same elements than *out_compare* folder (with few differences about the timing and file location indicated within the results files; the motifs found in *test_out/test_out_report.html* and *out_compare/test_out_report.html* should be highly similar but there can be small differences due to the random initial conditions that rely on different random number

generators in the various operating systems - note that also the order of the motifs can differ due to these different random number generators).

The *testData.txt* file corresponds to HLA-II peptidomics data from meningioma tissues obtained in our study (sample *3869-GA*, with 5987 unique peptides).

5. (Optional) To run MoDec from anywhere on your computer, make an alias of MoDec executable (see above for which one depending on operating system) or add it in your path.

# Running

## Command

```
MoDec -i input_file -o output_name [additional options]
```

- Depending on your operating system, use Modec, MoDec_unix or MoDex.exe as indicated in the installation instructions.
- Do not use spaces in your file or directory names.
- Do not use other special characters (e.g., *, ?, %, &,...) in file or directory names.

## Required arguments

- Input file (command `-i` or `--input`): File listing all the peptides with one peptide per line. It can also be a fasta file (lines starting with ">" are skipped).

- Output name (command `-o` or `--output`): The name of the output for the current run (including the path where to save it). MoDec will create a folder of this name and save all results in this folder.

## Optional arguments

- `--Kmin 1` and/or `--Kmax 6`: To run MoDec searching for motifs between the *Kmin* value (default is 1) and *Kmax* value (default is 6).

- `-L 9` or `--Lmotif 9`: To search for motifs of the given length (default is 9). A value of 0 makes that we use a motif length equal to the longest peptide from the input data (and all peptides shorter that the longest are removed in this case).

- `--MHC2` or `--MHCII` or `--MHC-II`: This is an option recommended when working on MHC-II / HLA-II peptidomics data, to override some default parameters with those recommended in this case (it is equivalent to using the options `--pepLmin 12 --specInits`). When the options MHC2 and pepLmin are given, the pepLmin option value will have preference.

- `--pepLmin 12`: To remove from the input the peptides shorter than this length (default is equal to the motif length). **For HLA-II peptidomics, we recommend using a value of 12, to remove potential HLA-I contaminants often present in the data**.

- `-r 20` or `--nruns 20`: To make this number of runs for the deconvolution with each selected number of motifs (default is 20).

- `-b 1` or `--nruns_out 1`: To output the results from this number of the best runs (by default we just output the results from the best run but can also output more). When this is used, the standard *report.html* still contains only the best results while another *..._report_multiRuns.html* file is created showing the results from the various runs.

- `--specInits`: To perform the computations based on multiple types of initial condition biases (e.g. performing some runs with a hydrophobic preference at P1). **This is recommended for HLA-II peptidomics** to increase the chances of finding the best motifs representing the data. When this is used, MoDec will run in fact 5 times more runs than what is given by the option `--nruns` (this is why the result summary files and *report.html* indicate a *n_runs_input* and *n_runs_real* value).

- `--makeReport`: To make an html report containing the logos found in the best run from each number of motifs (need to have R installed and *ggseqlogo* as described in step 3 above).

- `-y 3` or `--logoYmax 3`: Only useful with `--makeReport`. To define the max y value for the plots of the logos (default is 3).

- `--no_flat_mot`: When used, MoDec will not include a flat motif representing contaminant peptides.

- `--out_fullResp`, `--out_w_ks`: To output some additional files giving the full responsibilities of each peptides and the full weight of the motifs.

- `--no_bestPepResp`: When used, we don't output the file containing the best responsibility from each peptide (by default this file is outputted).

- `--outAdd`: Normally, MoDec will not run if the folder given in the `--output` argument already existed. When this option is used, MoDec will still run. **This option is to use with care**: previous results might get overwritten as it is using same file names, and, if different parameters are used for the different runs, the report html file will only indicate one of the sets of parameters used, without warning. This option is useful when launching MoDec for different numbers of motifs (`Kmin`, `Kmax` arguments): in such a case the results won't be overwritten as the output has a different name for each number of motifs searched for.

- `--seed 38230811`: To give an initial seed for the random number generator used by MoDec (default value is 38230811). When using the seed value 0, MoDec will try using a fully random seed or use the current time if such fully random device isn't available.

- `--log`: To make that MoDec outputs its logs to a log file created in the result folder instead of outputting it to the terminal.

- `-k 9` or `--kmer 9`: The kmer size used when comparing input peptides to give weights on each peptide based on the similarity between each of the input peptides. Use a value of 0 if you do not want to give any weight on peptides (default is a value equal to *Lmotif*, which is recommanded for HLA-II peptidomics where many peptides are highly similar, coming from the same protein but often trimmed at other positions).

- `-a ACGT` or `--alphabet ACGT`: The letters allowed in the alphabet. This is used if we want to run MoDec on other letters than the 20 standard amino acids, for example to apply it to nucleotides. By default it corresponds to *ACDEFGHIKLMNPQRSTVWY* but if using the 20 standard AAs, it is better not

to use this option because for the moment it would disable some specific settings that are only available in the standard case.

- `-u X` or `--unk_aa X`: Defines the letter used to describe an unknown amino acid in the sequences (by default it is -). This should be a single letter, if more than one letter is given, only the first one is used. The unk_aa are treated differently in the computation.

- `-c chemistry` or `--col_scheme chemistry`: Defines the color scheme used by ggseqlogo if making an html report of the results. Available schemes depend on ggseqlogo version, but should include *auto* (default value), *chemistry*, *chemistry2*, *hydrophobicity*, *nucleotide*, *nucleotide2*, *base_pairing*, *clustalx*, *taylor* and *modified*.

- `--theta_norm 1`: Tells which type of background frequency normaization to use for the theta (it tells which type of *f* to use in equation (1) from our paper). For the moment, these values are available:

    - 0: no normalization, all AA have an equal weight;
    - 1: estimate of background frequencies in HLA-II peptidomics data (default value, recommended for HLA-II data);
    - 2: AA frequencies from the human proteome;
    - 3: AA frequencies computed based on the current dataset - i.e. frequencies from AA found in all peptides from the dataset.
    - 4: estimate of background frequencies in MHC-II peptidomics data from mice. Note that when the *alphabet* is given, the only possible values are 0 and 3.

- `--flat_freq 2`: Tell which type of frequency to use for the flat motif (corresponds to the *h_i* frequency used for *theta^0_l,i* in equation (1) from our paper). The possible values are the same as for *theta_norm* above (default value is *2*). Note that when the *alphabet* is given, the only possible values are 0 and 3.

- `-S 0`, `--Salign 0`: A switch telling how to count to motif offset *alignment* S, giving more weights to motifs found at similar locations along the sequences:

    - 0: consider centered motifs independent of peptide size;
    - 1: count from the N-terminal of the sequences;
    - 2: count from the C-terminal of the sequences;
    - 3: do not consider offset preferences, i.e. motifs found anywhere along sequence will have same weight.

## Results returned and additional information

- MoDec creates a folder containing various files and subfolders with the results.

- The files *Summary/res_K.._out.txt* gives various information, like the parameters used in the current run, the final log-likelihood or a comparison between the motifs (PWM values) found in the various run (of this number of motifs - this comparison is found in the last lines, values correspond to Kullback-Leibler divergences between each motif from the best set of motifs and the set of motifs from each other run).

- The outputted PWMs (and corresponding logos in the report) correspond to the $\theta^{k}_{l,i}$ from equation (1) of our paper.

- The peptide responsibilities are returned in the *Responsibilities/bestPepResp_....txt* files. This gives for each peptide its two best responsibility values indicating to which motif and offset this corresponds as well as the corresponding core binding sequence.
  With the option `--out_fullResp` it is possible to view the complete matrix of the peptide responsiblities (values from each peptide towards each motif and towards each binding core offset).

- The optimal number of motifs is difficult to automatically determine and existing measures provide unsatisfactory results. We use the Akaike Information Criterion (AIC) as a guide but we encourage the users to manually look at the motifs and determine the optimal number (AIC information is found in the html report and in *Summary/nMotScores....txt*).

- In the html report, the numbers indicated below each logo correspond to the number of peptides associated to this motif (assigning here each peptide to the motif towards which this peptide had its best responsibility value). The number in parenthesis next to it is the fraction of peptides assigned to this motif.

- The PWMs and logos of the flat motifs are also returned by MoDec.

## Latest version

Latest version of MoDec is available at https://github.com/GfellerLab/MoDec/releases.

## License

MoDec can be used freely by academic groups for non-commercial purposes (see the license file). The product is provided free of charge, and, therefore, on an "as is" basis, without warranty of any kind.

**FOR-PROFIT USERS**: If you plan to use MoDec (version 1.2) or any data provided with the script in any for-profit application, you are required to obtain a separate license. To do so, please contact nbulgin@lcr.org at the Ludwig Institute for Cancer Research Ltd.

## Contact information

For scientific questions, please contact Julien Racle (julien.racle@unil.ch) or David Gfeller (david.gfeller@unil.ch).

For license-related questions, please contact Nadette Bulgin (nbulgin@lcr.org).

## How to cite

To cite MoDec, please refer to:

Racle, J., et al. Robust prediction of HLA class II epitopes by deep motif deconvolution of immunopeptidomes. *Nat. Biotechnol.* 37, 1283–1286 (2019).