VSEQ

Documentation for release v0.4.0

Andrea Laguillo Gómez[1], Enrique Calvo Alcocer[1], Jesús Vázquez Cobos[1]

CONTENTS

1 Introduction 1
2 Getting Started 1
3 Configuration 2
4 Tools 3
4.1 Vseq Viewer . . . 3
4.2 Vseq Explorer . . . 5
5 Examples 5

1 INTRODUCTION

Vseq is a set of tools for mass spectrometry (MS) spectra visualization, quality control, peptide searching, and more, developed at the Proteomics Unit of the Spanish National Center for Cardiovascular Research (CNIC).

Contact: **andrea.laguillo@cnic.es**

2 GETTING STARTED

The latest release of Vseq, as well as any previous versions, can be downloaded from the **CNIC Proteomics GitHub page**.

Python 3 is required to run Vseq. The necessary Python packages can be installed using the included `requirements.txt` file and `pip`:

```
$ pip install -r requirements.txt
```

Once all packages are installed, the scripts can be run directly from the command line:

```
$ python Vseq.py
$ python VseqExplorer.py
```

When calling these scripts from a different directory, simply include the full path to the file: `$ python /path/to/Vseq.py`

For more information about available command line arguments, please check each script's section in this document or run the scripts with the `-h` or `--help` argument.

[1] *Cardiovascular Proteomics, CNIC (National Center for Cardiovascular Research), Madrid, Spain*

3 CONFIGURATION

A single `Vseq.ini` file contains all the parameters for both the Viewer and Explorer tools. This file is divided in five sections:

PARAMETERS Default values for each parameter are shown between parentheses.

- `min_dm` (3) Minimum ΔMass value in Dalton required to consider that a peptide is modified.
- `min_ions_matched` (2) Minimum number of matched ions required in a candidate identification.
- `precursor_tolerance` (0.1) Tolerance in Thomson to match the precursor peptide.
- `fragment_tolerance` (15) Tolerance in parts-per-million (ppm) to match the fragment peptides.
- `best_n` (5) Number of candidates to make Vseq plots for.
- `sort_by` (hyperscore) Criterion to sort candidates by. The available options are: ions_matched, e_score, product, v_score, hyperscore. here, 'product' refers to the product of ions_matched and e_score.
- `vseq_threshold` (10) Result plots will only be created for candidates whose value for the chosen sorting option is equal or greater than this value.
- `ppm_plot` (30) Maximum ppm error of fragments that will be included in the interpreted V-plot.
- `parallelize` (candidate) Criterion to parallelize by. The available options are: candidate, sequence, both.
- `int_scanrange` (6) ± Number of MS1 scans to integrate. A value of 6 includes the target scan, 6 scans before it and 6 scans after it, for a total of 13.
- `int_mzrange` (2) ± M/Z range to calculate integration for, centered in the identified peptide's M/Z value.
- `int_binwidth` (0.001) Bin width to use when integrating.
- `poisson_threshold` (0.8) Calculate the set of theoretical peaks that cover at least this percentage of the isotopic envelope.

AMINOACIDS A list of the molecular weights of amino acids in Dalton. By default, the 20 eukaryotic proteinogenic amino acids are included. The list can be expanded by the user, adding each new aminoacid in a new line. A unique letter should be used for each amino acid.

```
A =  71.037114    R = 156.101111    N = 114.042927
D = 115.026943    C = 103.009185    E = 129.042593
Q = 128.058578    G =  57.021464    H = 137.058912
I = 113.084064    L = 113.084064    K = 128.094963
M = 131.040485    F = 147.068414    P =  97.052764
S =  87.032028    T = 101.047679    V =  99.068414
W = 186.079313    Y = 163.063329    U = 150.953630
O = 132.089878    Z = 129.042594    Custom amino acids
```

FIXED MODIFICATIONS A list of molecular weights of modifications in Dalton, assigned to an amino acid or the N-terminal and C-terminal positions. By default, no fixed modificstions are included.

```
A = 0    R = 0    N = 0    D = 0    C = 0    E = 0
Q = 0    G = 0    H = 0    I = 0    L = 0    K = 0
M = 0    F = 0    P = 0    S = 0    T = 0    V =  0
W = 0    Y = 0    Nt = 0    Ct = 0    U = 0    O = 0
Z = 0    Modifications for custom amino acids
```

Modifications in this list will be added to every peptide sequence to be searched, if applicable. Fixed modifications specific to a peptide can be added to the sequence provided in the input table, between brackets and directly after the modified aminoacid. For example, the following peptide contains a fixed modification of 361.20 Da at position 6 (lysine):

```
QAANTK[361.20]SAANL
```

MASSES Other molecular weight values in Dalton that are necessary for Vseq calculations.

```
m_proton   =  1.007276
m_hydrogen =  1.007825
m_oxygen   = 15.994915
```

LOGGING Options for logging.

- `create_log` (1) Create a log file. Enabled by default, set to 0 to disable.
- `create_ini` (0) Create a copy of the `Vseq.ini` file in the input directory. Disabled by default, set to 1 to enable. Specifying custom parameters through the command line will always create a copy of the `Vseq.ini` file.

4 TOOLS

Vseq is composed of two main tools, Vseq Viewer and Vseq Explorer. A third tool, Scan Integrator, can optionally be called from Vseq Viewer and will create an additional results plot.

4.1 Vseq Viewer

Vseq Viewer takes a list of identifications from a previous MS database search and creates a set of plots that can be used to check the quality of the identification.

INPUT Input files and options are provided to Vseq Viewer via command line arguments:

- `-i`, `--infile` Required A table of identifications. Must be a tab-separated text file, containing the following columns:
  - `FirstScan`: Scan number.
  - `Charge`: Charge of the peptide.
  - `MH`: Mass of the peptide in Dalton.
  - `Sequence`: Sequence of the peptide. May contain modifications (mass in Dalton) between brackets, directly after the modified aminoacid.
  - `RetentionTime`: Only shown in the "SCAN INFO" table, not used for any calculations. Can be left blank.
  - `msdataDir`: Directory containing the MS/MS data files
  - `outDir`: Directory where the output will be saved. Will be created if it does not exist.

  - `Raw`: Name of the MS/MS data file. MGF and mzML formats are recognized. mzML format is required if performing scan integration.
  - `DeltaMassLabel`: If any fixed modifications were included in the sequence, the name(s) must be specified in this column.

- `-c`, `--config` Optional A Vseq.ini file containing the parameters described in the Configuration section. If not specified, the default file will be used.
- `-e`, `--error` Optional The value provided through this argument will override the `fragment_tolerance` value set in the Configuration file.
- `-d`, `--deltamass` Optional The value provided through this argument will override the `min_dm` value set in the Configuration file.
- `-n`, `--integrate` Optional Calls the Scan Integrator tool. The integration can only be performed with mzML files.
- `-w`, `--n_workers` Optional Vseq Viewer makes use of paralellization. This argument sets the maximum number of threads to use. The default value is 4.

OUTPUT Vseq Viewer will create a number of output files, depending on the options used. These files will be saved to the directory specified in the `outDir` column from the input table.

The default output file (Figure 1), which will always be created, is a pdf containing several figures: a table with the scan information, a ppm-error vs intensity plot, a V-plot, an intepreted V-plot, a sequence coverage representation and an M/Z vs relative intensity plot.

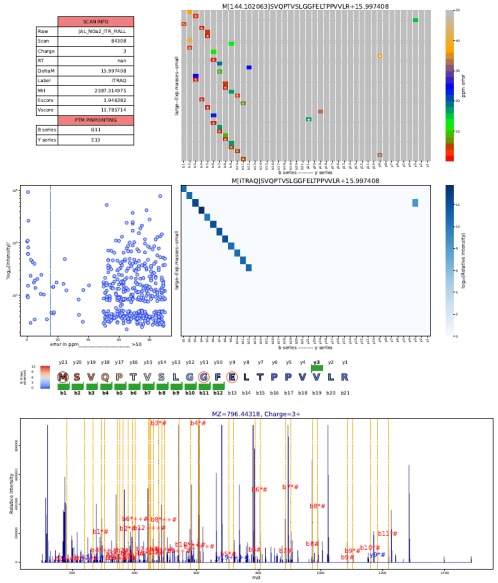

Figure 1: An example of a Vseq Viewer plot

If the `create_log` parameter from the Configuration file was active, a log file will be written. It contains a register of the amount of time taken to perform each task as well as any warning or errors that may have occurred.

Additionally, if Scan Integrator was called, another pdf file will be generated. It contains the raw integration and the apex picking (Figure 2).

INTERPRETATION The first figure is a table containing information about the identification: the Raw file it was obtained from, the scan number, the peptide charge,

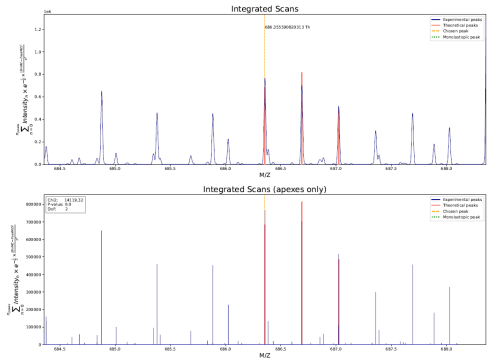

Figure 2: An example of a Scan Integrator plot

the retention time (if specified in the input table), the ΔMass calculated by Vseq (in Dalton), the fixed modifications (as labels), the MH, the E-score, the V-score and the Hyperscore. If the ΔMass is greater than the `min_dm` threshold established in the Configuration file, this table will also contain PTM pinpointing for series b and y.

4.2 Vseq Explorer

[Work in progress.]

5 EXAMPLES

[Work in progress.]