PhoStar v1.0

User Manual

February 2017
Sebastian Dorl, Stephan Winkler, Karl Mechtler, and Viktoria Dorfer

1. Installation of PhoStar

In order to install PhoStar please download and extract necessary files:

1. Download the PhoStar .zip archive from http://bioinformatics.fh-hagenberg.at/site/index.php?id=85
2. Right click on the downloaded .zip file and select the menu item "Properties" in the context menu
3. If visible, click "Unblock" at the bottom right of the Properties dialog
4. Click "OK" to close the Properties dialog
5. Extract the .zip file. Note that the model file is highly compressed. If you experience any errors during the extraction make sure that your archiver software is up to date (e.g. WinRAR 5.0+ or 7zip 9.0+).

PhoStar is now ready for use!

2. Functionality of PhoStar

PhoStar splits input files of MS/MS fragment ion peptide spectra into two parts: the spectra that are likely to have originated from phosphorylated peptides and those who are not. For this purpose, the tool calculates a list of numerical features for each spectrum and evaluates the results using a random forest model (.model file in the installation archive). The results are two separate files containing the split input spectra indicated by the suffixes "_phos" and "_nonphos". Additionally, all calculated features and the final classification scores for each spectrum are stored as tab separated values.

3. Usage of PhoStar

PhoStar is run from the command line using the following syntax:

```
PhoStar.exe [input file .MGF|.MSP] [model file .MODEL] [fragment ion tolerance]
[unit PPM|DA] [picked peaks per 100 m/z] [input split threshold 0...1]
```

input file

Full name of the file containing all input spectra to be processed. The file must be in .mgf or .msp format. An example.mgf is provided with the installation.

model file

Full name of the file containing the random forest model (.model file provided with the installation). The .model file contains a serialized random forest object created with ALGLIB (Version 3.10.0.0). For additional information, see section 4. Model description.

fragment ion tolerance

The value of the fragment ion tolerance used for matching peaks during the feature calculation of the input spectra.

unit

The unit used for the fragment ion tolerance. Allowed inputs are "PPM" or "DA".

picked peaks per 100 m/z

PhoStar performs peak picking before calculating features for the fragment ions. This parameter controls the peak picking depth as the number of highest peaks retained per 100 m/z window in the spectrum.

input split threshold

The threshold value that sets the boundary between phospho spectra and non-phospho spectra. The spectra are evaluated using the classification score, which varies between 0 and 1. A high score means that the spectrum is likely to have originated from a phosphorylated peptide. All spectra with a score above the threshold will be written to the "inputfilename_phos" output file and all others to the "inputfilename_nonphos" output file. If this parameter is set to exactly 0 or 1 then no output files for the spectra will be created. Use this option if you are only interested in the features and score values.

PhoStar always creates a text file output "inputfilename.txt" containing the calculated features and classification scores for all spectra as tab-separated values. Furthermore, running the program writes a log file "PhoStar_date.log" to the current directory.

A valid command line call to run PhoStar on the example file would look as follows:

```
PhoStar.exe example.mgf phospho_ensemble.model 10 ppm 10 0.5
```

4. Model description

The random forest model included in the installation was created using ALGLIB (Version 3.10.0.0) *dfserialize()*. The model was trained using a large number of high-confidence MS/MS spectra obtained from publicly available experimental data in the ProteomeXchange repository. The input for the model is a vector of 61 features (60 plus a placeholder for the classification score). Any serialized ALGLIB random forest with the same dimensions can possibly replace the model file. A detailed description of the 60 numerical features can be found in Table 1.

5. License

PhoStar is protected through copyright and a license agreement. For details please read the available license agreement in the installation archive.

The ALGLIB numerical analysis library is protected through copyright and a license agreement. For details please read the available license file in the installation archive.

6. Contact

For any further questions, bug reports, or ideas please contact Sebastian Dorl (sebastian.dorl@fh-hagenberg.at), Stephan Winkler (stephan.winkler@fh-hagenberg.at), Karl Mechtler (karl.mechtler@imp.ac.at), or Viktoria Dorfer (viktoria.dorfer@fh-hagenberg.at).

This research project is a collaboration of the Protein Chemistry Group at the Research Institute of Molecular Pathology, Vienna, the Johannes Kepler University, Linz, and the Bioinformatics Research Group at the University of Applied Sciences Upper Austria, Campus Hagenberg. The University of Applied Sciences Upper Austria and the state Upper Austria funded the research.

Table 1: Description of PhoStar output columns. **Bold** columns mark the numerical features.

Name	Description
file	Original file name if it can be parsed from spectrum title or comments
scan	Scan number if it can be parsed from spectrum title or comments
rtime	Retention time if it can be parsed from spectrum title or comments
precursor	Precursor ion mass in m/z
charge	Precursor ion charge state
A	Number of fragment peak pairs with an m/z difference matching alanine
R	Number of fragment peak pairs with an m/z difference matching arginine
N	Number of fragment peak pairs with an m/z difference matching asparagine
D	Number of fragment peak pairs with an m/z difference matching aspartic acid
E	Number of fragment peak pairs with an m/z difference matching glutamic acid
Q	Number of fragment peak pairs with an m/z difference matching glutamine
G	Number of fragment peak pairs with an m/z difference matching glycine
H	Number of fragment peak pairs with an m/z difference matching histidine
K	Number of fragment peak pairs with an m/z difference matching lysine
M	Number of fragment peak pairs with an m/z difference matching methionine
F	Number of fragment peak pairs with an m/z difference matching phenylalanine
P	Number of fragment peak pairs with an m/z difference matching proline
S	Number of fragment peak pairs with an m/z difference matching serine
T	Number of fragment peak pairs with an m/z difference matching threonine
W	Number of fragment peak pairs with an m/z difference matching tryptophane
Y	Number of fragment peak pairs with an m/z difference matching tyrosine
V	Number of fragment peak pairs with an m/z difference matching valine
J	Number of fragment peak pairs with an m/z difference matching either leucine or isoleucine
Z	Number of fragment peak pairs with an m/z difference matching carboxymethylated cysteine
preNL H20	Normalized intensity of peaks for a precursor neutral loss of water
preNL H2Ox2	Normalized intensity of peaks for a precursor neutral loss of two water
preNL NH3	Normalized intensity of peaks for a precursor neutral loss of ammonia
preNL NH3x2	Normalized intensity of peaks for a precursor neutral loss of two ammonia
preNL H2O+NH3	Normalized intensity of peaks for a precursor neutral loss of water and ammonia
preNL H3PO4	Normalized intensity of peaks for a precursor neutral loss of phosphoric acid
preNL H3PO4+H2O	Normalized intensity of peaks for a precursor neutral loss of phosphoric acid and water
preNL H3PO4+H2Ox2	Normalized intensity of peaks for a precursor neutral loss of phosphoric acid and two water
preNL H3PO4+NH3	Normalized intensity of peaks for a precursor neutral loss of phosphoric acid and ammonia
preNL H3PO4+NH3x2	Normalized intensity of peaks for a precursor neutral loss of phosphoric acid two ammonia
preNL H3PO4+H2O+NH3	Normalized intensity of peaks for a precursor neutral loss of phosphoric acid, ammonia, and water
preNL H3PO4x2	Normalized intensity of peaks for a precursor neutral loss of two phosphoric acid
preNL HPO3	Normalized intensity of peaks for a precursor neutral loss of phosphite
Im(pTyr)	Normalized intensity of the phosphotyrosine immonium ion
pSer	Number of fragment peak pairs with an m/z difference matching phosphorylated serine
pThr	Number of fragment peak pairs with an m/z difference matching phosphorylated threonine
pTyr	Number of fragment peak pairs with an m/z difference matching phosphorylated tyrosine

fragNL H3PO4	Number of fragment peak pairs with an m/z difference matching a fragment neutral loss of phosphoric acid
fragNL HPO3	Number of fragment peak pairs with an m/z difference matching a fragment neutral loss of phosphite
fragNL H3PO4+H2O	Number of fragment peak pairs with an m/z difference matching a fragment neutral loss of phosphoric acid and water
fragNL H3PO4+NH3	Number of fragment peak pairs with an m/z difference matching a fragment neutral loss of phosphoric acid and ammonia
fragNL H3PO4+H2Ox2	Number of fragment peak pairs with an m/z difference matching a fragment neutral loss of phosphoric acid and two water
fragNL H3PO4+NH3x2	Number of fragment peak pairs with an m/z difference matching a fragment neutral loss of phosphoric acid and two ammonia
fragNL H3PO4+H2O+NH3	Number of fragment peak pairs with an m/z difference matching a fragment neutral loss of phosphoric acid, ammonia, and water
total precursor NL phos	Sum of normalized intensities of all peaks that match a precursor neutral loss involving H3PO4 or H3PO (considering isotopes with 0, 1, and 2 C13 atoms)
ratio fragment NL phos	Sum of intensity of all peaks that correspond to any fragment neutral loss involving H3PO4 or HPO3, divided by the total intensity of picked peaks. Fragment ion matching considers charge states of +1 and +2, as well as isotopes with 0, 1, and 2 C13 atoms. The matching of fragment ion pairs also uses theoretically calculated complementary peptide ions.
ratio phos fragment peaks	Ratio of all picked fragment ion peaks that are involved in at least one peak pair matching a fragment neutral loss involving H3PO4 or HPO3
mean phos peak hits	The average amount of peak pairs, matching a fragment neutral loss involving H3PO4 or HPO3, that each fragment peak is part of
0 hit peaks	Number of picked fragment ion peaks that are part of 0 peak pairs matching a fragment neutral loss involving H3PO4 or HPO3
1 hit peaks	Number of picked fragment ion peaks that are part of 1 peak pairs matching a fragment neutral loss involving H3PO4 or HPO3
2 hit peaks	Number of picked fragment ion peaks that are part of 2 peak pairs matching a fragment neutral loss involving H3PO4 or HPO3
3 hit peaks	Number of picked fragment ion peaks that are part of 3 peak pairs matching a fragment neutral loss involving H3PO4 or HPO3
4 hit peaks	Number of picked fragment ion peaks that are part of 4 peak pairs matching a fragment neutral loss involving H3PO4 or HPO3
5 hit peaks	Number of picked fragment ion peaks that are part of 5 peak pairs matching a fragment neutral loss involving H3PO4 or HPO3
6 hit peaks	Number of picked fragment ion peaks that are part of 6 peak pairs matching a fragment neutral loss involving H3PO4 or HPO3
7 hit peaks	Number of picked fragment ion peaks that are part of 7 peak pairs matching a fragment neutral loss involving H3PO4 or HPO3
8 hit peaks	Number of picked fragment ion peaks that are part of 8 peak pairs matching a fragment neutral loss involving H3PO4 or HPO3
9 hit peaks	Number of picked fragment ion peaks that are part of 9 peak pairs matching a fragment neutral loss involving H3PO4 or HPO3
10+ hit peaks	Number of picked fragment ion peaks that are part of 10+ peak pairs matching a fragment neutral loss involving H3PO4 or HPO3
PHOS_S	Set to “1” if the spectrum title or comments contain a flag for phosphorylation of serine
PHOS_T	Set to “1” if the spectrum title or comments contain a flag for phosphorylation of threonine
PHOS_Y	Set to “1” if the spectrum title or comments contain a flag for phosphorylation of tyrosine
PHOS_X	Set to “1” if the spectrum title or comments contain a flag for any amino acid phosphorylation modification
ESTIMATE	PhoStar random forest model classification score