The PSI Extended FASTA Format (PEFF)

Pierre-Alain Binz[1], Jim Shofstahl[2]; Juan Antonio Vizcaíno[3], Harald Barsnes[4]; Robert Chalkley[5]; Gerben Menschaert[6], Emanuele Alpi[3]; Karl Clauser[7]; Jimmy K Eng[8]; Lydie Lane[9]; Sean Seymour[10]; Gerhard Mayer[11]; Martin Eisenacher[11]; Yasset Perez-Riverol[2], Eugene Kapp[12]; Zhi Sun[13], Luis Mendoza[13], Peter R. Baker[5]; and Eric Deutsch[13]

[1] CHUV Centre Hospitalier Universitaire Vaudois, Lausanne, Switzerland; [2] Thermo Fisher Scientific, San Jose, California; [3] European Bioinformatics Institute (EMBL-EBI), Hinxton, United Kingdom; [4] University of Bergen, Bergen, Norway; [5] UCSF, San Francisco, CA; [6] Ghent University, Gent, Belgium; [7] Broad Institute of MIT and Harvard, Cambridge; [8] University of Washington, Seattle, WA; [9] SIB Swiss Institute of Bioinformatics, Geneva, Switzerland; [10] Seymour Data Science, San Francisco, California; [11] Ruhr University Bochum, Bochum, Germany; [12] University of Melbourne, Melbourne, Australia; [13] Institute for Systems Biology, Seattle, USA

Introduction

The Human Proteome Organization (HUPO) Proteomics Standards Initiative (PSI) has just finalized the PSI Extended FASTA Format (PEFF). It is based on the widely-used FASTA format for encoding the protein sequence databases used by most tandem mass spectrometry (MS/MS) search engines, but adds a uniform mechanism for encoding substantially more metadata about the sequence collection as a whole as well as individual entries, including support for encoding known sequence variants, PTMs, and proteoforms. PEFF is defined by a full specification document, controlled vocabulary terms, a set of example files, software libraries, and a file validator [1].

Benefits of using PEFF

- Consistent interpretation of sequence annotations by all supporting tools.
- Enables ability to search for known amino acid substitutions.
- Enables ability to search for certain PTMs at known positions without needing to search for all via search parameters, increasing sensitivity, search speed and reducing false positives. The current neXtProt PEFF database includes a wide set of modifications that are typically not searched for.
- Enables display in search results of the known and considered variants/PTMs, and the known and detected variants/PTMs.
- It can be used to fully describe individual proteoforms.
- Includes metadata about databases in the file.

Specification

The PEFF format defines a file header section and individual sequence entries section. The file header includes information such as the database name, source, version, etc. PEFF sequence entries are similar to standard FASTA sequence entries but with extended description lines.

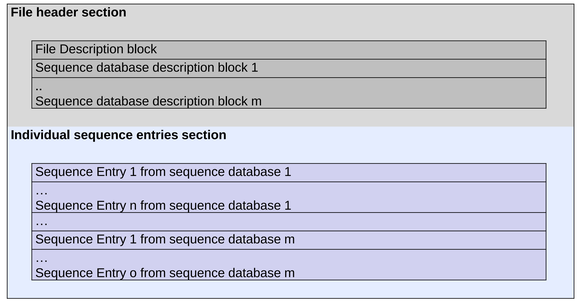

Figure 2. Graphical representation of the PEFF file structure

The **file description block** declares the format version and optionally general comments

As a PEFF file may include multiple sequence databases, metadata about each of them are described in the corresponding **sequence database description blocks**. These include DB name, version, source, sequence type, among others, all encoded using the following format:

 # key=value

Where the *key* elements are controlled vocabulary (CV) terms and the *value* term contains one or more items whose format is defined for each key in the CV.

```
# DbName=neXtProt-extract
# Prefix=nxp
# DbDescription=extract of neXtProt with manual modifications
# GeneralComment= A selection of protein entries
# Decoy=false
# DbVersion=2019-01-11
# DbSource=http://www.nextprot.org
# NumberOfEntries=62
# SequenceType=AA
# //
```

Figure 2. Example of a sequence database description block. This block can be repeated if the file contains multiple databases (such as forward and decoy)

In the **individual sequence entries section**, each of the sequence entries have a description line and a sequence line.

The PEFF description lines are in the following format

>Prefix:DbUniqueId \key=value \key=value …

Where the *key* elements are controlled vocabulary (CV) terms and the *value* term contains one or more items whose format is defined for each key in the CV.

In addition to the mandatory unique entry identifier, information such as protein name, sequence length, sequence variants, PTMs, sequence maturation or post-translational processing and taxonomy can be described.

```
>nxp:NX_Q13217-1 \DbUniqueId=NX_Q13217-1 \PName=DnaJ homolog subfamily C member 3 isoform Iso 1 \GName=DNAJC3
\NcbiTaxId=9606 \TaxName=Homo Sapiens \Length=504 \SV=156 \EV=471 \PE=1 \ModResPsi=(274|MOD:00046|O-phospho-L-
serine)(290|MOD:00064|N6-acetyl-L-lysine) \ModRes=(248||Disulfide)(258||Disulfide)(313||Disulfide)
(329||Disulfide) \VariantSimple=(4|S)(34|S)(35|T)(36|N)(141|*)(194|Q)(194|*)(350|C)(438|N)(441|E)(477|S)
(479|P)(495|E)(495|R)(502|R) \VariantComplex=(17|17|FL)(220|220|)(374|374|)(500|500|K*) \Processed=(1|31|signal
peptide)(32|504|mature protein)
MVAEGSVTSRLGSVFPFLLVLVDLQVEGAECGVNASVEKHLELGKKLLAAGQLADALSQFHAAVDGDPDNYIAYYRRATVFLAMGKSKAALPDLTKVIQL
KMDFTAARLQRGHLLLKQGKLDEAEDDFKKVLKSNPSENEKEAQSQLIKSDEMQRLRSQALNAFGSGDYTAAIAFLDKILEVCVWDAELRELRAECFIK
EGEPRKAISDLKAASKLKNDNTEAFYKISTLYYQLGDHELSLSEVREGLKLDQDHKRCFAHYKQVKKLNKLIESAEELIRDGRYTDATSKYESVMKTEPS
IAEYTVRSKERICHCFSKDEKPVEAIRVCSEVLQMEPDNVNALKDRAEALIEEMYDEAIQDYETAQEHNENDQQIREGLEKAQRLLKQSQKRDYYKILG
VKRNAKKQEIIKAYRKLALQWHPDNFQNEEEKKKAEKKFIEISAAAKEVLSDPEMRKKFDDGEDPLDAESQQGGGGNPFRSWNSWQGFNPFSSGFPFRFK
FSFN
```

Figure 3. Example (simplified) entry for a 504 amino-acid protein from neXtProt with PTMs, SAAVs, other variants, and post-translational processing information.

In the sequence entry example above, the "\ModResPsi" key indicates PTMs encoded using the PSI-MOD CV:

(274|MOD:00046|O-phospho-L-serine) S at position 274 can be phosphorylated

(290|MOD:00064|N6-acetyl-L-lysine) K at position 290 can be acetylated

"\VariantSimple" encodes single amino acid variants (SAAVs) (substitutions):

(4|S) P at position 4 can be swapped to an S

(34|S) N at position 34 can be swapped to an S, etc.

While the represented annotation features such as SAAV or PTMs are to be interpreted as "possible" or non compulsory, the format also supports description of specific proteoforms, where a combination of these features are to be considered as mandatory.

Please refer to the full specification document at **http://psidev.info/peff** to see additional details on the format including key elements not shown in the above example.

Next Steps

Encourage more widespread implementations of PEFF in sequence database producers and consumers.

More information

Web site:

psidev.info/peff

Preprint:

Implementations

Many resources and software packages already implement PEFF and more are on the way. Below is a listing of current implementations:

PEFF Producers

Resource Name	Description of Support
neXtProt	Exports all sequences in database along with curated PTMs and SAAVs
UniProt	Exports requested entries with SAAVs via Proteins API
Pyteomics 4.0	Implements a PEFF class with reader and writer in Python
Proteomics::PEFF	Implements a PEFF class with reader, writer, validator in Perl
Proteoformer	Determines RIBO-seq derived proteoforms and can write results in PEFF

PEFF Consumers

Resource Name	Description of Support
Comet	Sequence search engine can use PEFF and search for PTMs and SAAVs
phpMs	Supports the use and viewing of PEFF files
Pyteomics 4.0	Implements a PEFF class with reader and writer in Python
Proteomics::PEFF	Implements a PEFF class with reader, writer, validator in Perl
ProteinPilot	Supports searching of PEFF databases since V5.0 (2014)
Protein Prospector	Sequence search engine can use PEFF and search for PTMs
Online Validator	Accepts PEFF upload and validates that the file is compliant with latest specification
ProteoMapper	Supports mapping peptide sequences to a PEFF file, taking into account SAAVs
TPP	Full support for processing results searching and viewing SAAVs and PTMs (Poster MP 417)

References

1. Binz PA, Shofstahl J, Vizcaíno JA, Barsnes H, Chalkley RJ, Menschaert G, Alpi E, Clauser K, Eng JK, Lane L, Seymour SL, Sánchez LFH, Mayer G, Eisenacher M, Perez-Riverol Y, Kapp EA, Mendoza L, Baker PR, Collins A, Van Den Bossche T, Deutsch EW, Proteomics Standards Initiative Extended FASTA Format, J Proteome Res. 2019 May 23. doi: 10.1021/acs.jproteome.9b00064

Acknowledgements

This work was supported by the National Institutes of Health NIGMS grants R24 GM127667 and R01 GM087221 and NIBIB grant U54 ES017885.