SOP_007_NU_3_Minimum_Information_for_Reporting_Proteoforms_October_2017_v1

Richard D. LeDuc

❖ Scope

This document applies to any researcher using NRTDP resources, and represents scholarly best practices for publishing the results of top-down proteomics studies based on the software that we supply. All internal NRTDP researchers are required to comply with these standards, and external users are strongly encouraged to comply. The NRTDP Software Development Team is available to help, should you have any questions regarding this process.

❖ Rationale

This SOP has been enacted to meet the need for a unified understanding of what it means to report the existence of a proteoform, and to attempt to increase the general replicability and utility of top down proteomic studies.

❖ Assumption

Once you publish one or more proteoforms, you and your co-authors assert that you have found physical evidence for the existence of the reported proteoform within a specific biological setting. When you make this assertion, both the proteoforms that are claimed to exist and the files containing the mass spectrometric observations belong in the public domain. Therefore, your proteoforms, the raw data, and the rationale that you used to assert the existence of the proteoforms will be made public.

❖ 7 Definitions

- **Protein Entry:** a gene-specific identification (which can map to 1 or more proteoforms). This is denoted by an accession number to a gene-centric database like UniProtKB (e.g. P01234).
- **Instantaneous q-value:** a metric of confidence for identification that can be applied at three molecular levels: 1) the protein entry level (accession number), 2) the isoform level, or 3) the proteoform level (PFR). This is the false discovery rate (FDR) that the whole study would have if the entity in question was the worst entity to be identified.
- **Proteoform (PFR) Identifier:** an identifier that uniquely maps to a proteoform from a given organism, assigned by the Consortium for Top-Down Proteomics (CTDP) Proteoform Repository.
- **C-Score:** a metric reflecting the quality of proteoform characterization (ranges from 0 to ~1500). This scores how uniquely a given observation matches the best proteoform from a database search.
- **TDPORTAL:** the NRTDP's cloud-based, high-throughput top-down data search solution.
- **TDVIEWER:** a free Windows application that displays the results of a search on the TDPortal.
- ***.tdReport*:** a report file viewable in **TDVIEWER**. (These files are in SQLite database format). For each proteoform returned, the TDViewer also contains the **C-score** and the **PFR identifier**. For simplicity of reporting results, by default the proteoform- and isoform-level results are anchored to those protein accession numbers viewed with a user-specified FDR value. In this manner, each proteoform in the ***.tdReport*** file is assured to map to an accession number that is also in the set of proteins identified.

❖ Required Elements to Report

To adequately report one or more proteoforms, four types of data must be made public. These are: the proteoforms with their required metrics, the raw spectral data files, the database that the proteoforms were searched against, and the search strategies employed. Each of these is considered below:

1. **Proteoforms:** A proteoform is considered reported when it is published in the CTDP Proteoform Repository. Internally, reporting a proteoform requires the following metrics:

   - **Unique PFR or Proteoform Identifier.** These are assigned by the CTDP Proteoform Repository. Anyone using the **TDPortal** to process their data will be given these values automatically in the ***.tdReport*** file. Others are required to work with the NRTDP Software Development Team to acquire these values.
   - **Instantaneous q-value, or other related metric:** Related metrics include p-values, E-values and PCS scores. These values measure the confidence that the proteoform reported is associated with a given protein entry from the search database. This is taken as evidence that the reported proteoform is derived from the gene associated with the protein – in other words, this is a measure of the confidence that the identified proteoform is derived from a given gene. The ***.tdReport*** file provides the instantaneous q-value for each proteoform identified on the Proteoforms tab.
   - **C-Score:** This is a measure of the uniqueness of the proteoform relative to all other related proteoforms. Users of the **TDPortal** will get this value automatically in the ***.tdReport*** file.
   - **Quantification Note:** The above metrics are only for reporting the existence of a proteoform. If you also intend to assert that the proteoform is differentially expressed between two or more cell types, you will need to report additional information – usually a separate instantaneous q-value for the assertion of differential expression, and some measure of relative or absolute difference in abundance.

2. **Raw data:** The mass spectrometric files supporting your claims must be made public. Oftentimes, specific projects or funding agencies require that researchers publish their raw data though specific data repositories. For example, the Clinical Proteomic Tumor Analysis Consortium (CPTAC) requires that raw files be published through the CPTAC Data Portal (https://cptac-data-portal.georgetown.edu/cptacPublic/). Further, data repositories rise and fall in popularity, and researchers are encouraged to use their favorite. In the absence of personal preference, the CTDP has an agreement with Indiana University to use their ScholarWorks data repository for storing raw data. Members of the CTDP can work with the Development Team to use this service. Users should be aware that this process can often take several weeks, so please build lead time into your requests.

3. **Databases Searched:** For someone to understand what you have done, and to be able to determine if they agree with your conclusions, researchers need to be able to access not only your observations, but also the database that you searched those observations against. If you are using the **TDPortal**, this means that you need to make available the UniProt data set that you used, preferably in the .xml format. ProSightPC users need to archive the ProSight Warehouse files that they search against. Usually, these will be included in a bundle with the raw data files.

4. **Search Methodology:** The parameters of the search or searches used are also important. For ProSightPC users, the .xml file found in the Search Trees folder that was used to define the search tree in the experiment will be sufficient. For users of the **TDPortal**, the Standard Search Tree version (e.g. v1.3) should be reported. Currently, this will need to be included in the Methods section of your manuscript, as the **TDPortal** manuscript is under review.

❖ Recommended text

The following text and citations are intended for users of the TDPortal, and are valid as of the writing of this SOP. As better references become available, this SOP will be updated, so double check that you have the most current version of this document. Researchers using other top down resources can modify this text to match the conditions used in their analysis.

Proteoform Identification. For all analyses, a standardized three-pronged search strategy was employed. Using the search modes as defined for ProSight PTM 2.0 [Zamdborg, et al., 2007], the strategy used a narrow absolute mass search (with an MS^1 tolerance of 2.2 Da and 10 ppm tolerance for MS^2), a biomarker search (akin to a no-enzyme type search in peptide data analysis; MS^1 tolerance of 10 ppm and 10 ppm tolerance for MS^2), and a wide absolute mass search (MS^1 tolerance of 200 Da and 10 ppm tolerance for MS^2, with delta M mode activated). All proteoforms had an instantaneous FDR of 1% or less and a C-Score of 40 or higher [LeDuc et al., 2014].

LeDuc, R. D.; Fellers, R. T.; Early, B. P.; Greer, J. B.; Thomas, P. M.; Kelleher, N. L. **The C-score: a Bayesian framework to sharply improve proteoform scoring in high-throughput top down proteomics.** *Journal of Proteome Research* 2014, 13 (7), 3231-3240.

Zamdborg, L.; LeDuc, R. D.; Glowacz, K. J.; Kim, Y.-B.; Viswanathan, V.; Spaulding, I. T.; Early, B. P.; Bluhm, E. J.; Babai, S.; Kelleher, N. L. **ProSight PTM 2.0: improved protein identification and characterization for top down mass spectrometry.** *Nucleic Acids Research* 2007, 35 (Suppl. 2), W701-W706.

❖ Two Important Deadlines

First Deadline: If you are not using the **TDPortal**, you will need to work with the NRTDP Software Development Team to have your proteoforms loaded into the CTDP repository. This will allow you to acquire PFR identifiers to refer to in your manuscript. To do this you must allow **at least three weeks** between when you contact the team and when you need the values.

Second Deadline: For users who intend to use the Indiana University ScholarWorks for archiving their raw data, please be aware that this process can take several weeks. You should contact the NRTDP Software Development Team well before the final publication in order to ensure that DOI's to the data can be added to the manuscript during the final proofs. To acquire a DOI from Indiana University ScholarWorks, you must allow **at least 4 weeks** between when you contact the team, and when you need the values.

❖ **Manuscript Checklist**

- ☐ **Proteoform identifiers assigned and used in the manuscript.** If you are not using the **TDPortal**, have you contacted the NRTDP Software Development Team with enough lead time to allow your data to be published?
- ☐ **Raw data in a repository.** Are you using the CTDP mass storage at Indiana University? If so, have you worked with the NRTDP Software Development team to get the data published? If not, do you have the URLs for the data from the repository of your choice?
- ☐ **Search databases in a repository.** Did you add the databases used in your study to the raw data above?
- ☐ **Searches adequately defined.** Did you use the standard text provided, modified to match exactly what you did?

❖ **Acknowledgement**

This protocol was developed with support from the National Institute of General Medical Sciences P41GM108569 for the National Resource for Translational and Developmental Proteomics (NRTDP) based at Northwestern University.