Systematic Assessment of Entrapment Approaches in DIA Proteomics FDR Validation

Monika Pepelnjak[1], Oliver M. Bernhardt[1], Veronique Laforte[1], Tejas Gandhi[1] and Lukas Reiter[1]
[1]Biognosys AG, Schlieren, Switzerland

ABSTRACT PRESENTATION NUMBER: PV.02.111

BIOGNOSYS
NEXT GENERATION PROTEOMICS

INTRODUCTION

Target decoy-based analysis is the universal approach for identification in bottom-up discovery proteomics. A modern data analysis software typically uses different scores and complex data analysis paradigms. It is important to ensure that the false positive rate is controlled as specified. Such a validation is typically done with an entrapment analysis. The core concept of an entrapment analysis is that the software is blind to the difference between the real targets and entrapment targets facilitating the estimation of the entrapment FDR rate (eFDR) of the software based on the rate of identification of entrapment proteins. While being intuitive, overlooking the complexity of generating an entrapment space can lead to wrong interpretation of the results, that can either be too conservative or too optimistic FDR evaluation.

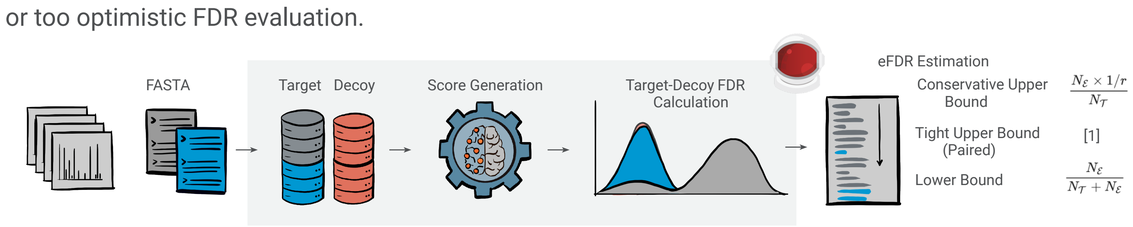

Figure 1 FDR evaluation with entrapment FASTA. Raw data is searched with target and entrapment FASTA file. Within the software of choice, targets and decoys are generated from FASTA files, the scores for individual identifications are calculated using machine learning algorithms and FDR is calculated using target-decoy model. The software then provides a list of identifications based on the provided FDR filters. The entrapment proteins on the target list are used for evaluation of Software eFDR. The upper and lower eFDR are calculated as indicated with the formulas. Ne = N of entrapment proteins, Nt = N of target proteins, R = ratio of target to entrapment FASTA.

METHODS

• Software version: Spectronaut® 20.2

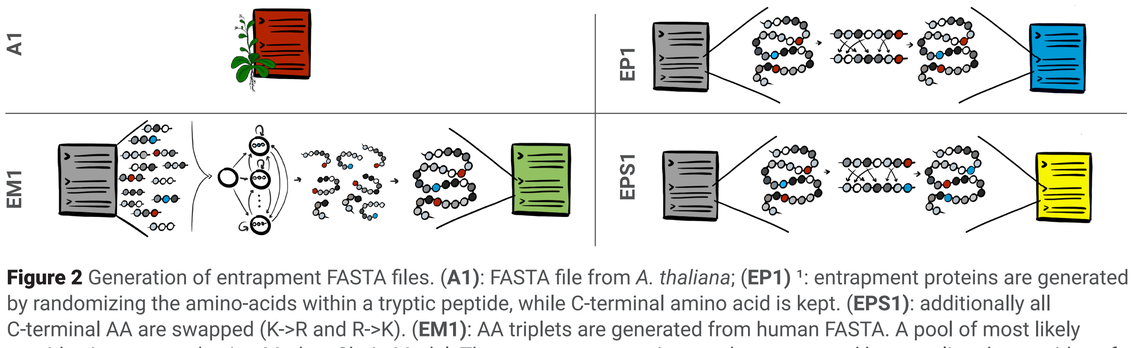

Figure 2 Generation of entrapment FASTA files. (**A1**): FASTA file from *A. thaliana*; (**EP1**) [1]: entrapment proteins are generated by randomizing the amino-acids within a tryptic peptide, while C-terminal amino acid is kept. (**EPS1**): additionally all C-terminal AA are swapped (K->R and R->K). (**EM1**): AA triplets are generated from human FASTA. A pool of most likely peptides is generated using Markov Chain Model. The entrapment proteins are then generated by sampling the peptides of the same length and the same number of peptides per protein for each given target protein, generating a 1:1 entrapment to human.

IDENTIFYING THE BIASES IN DIFFERENT ENTRAPMENT FASTA FILES

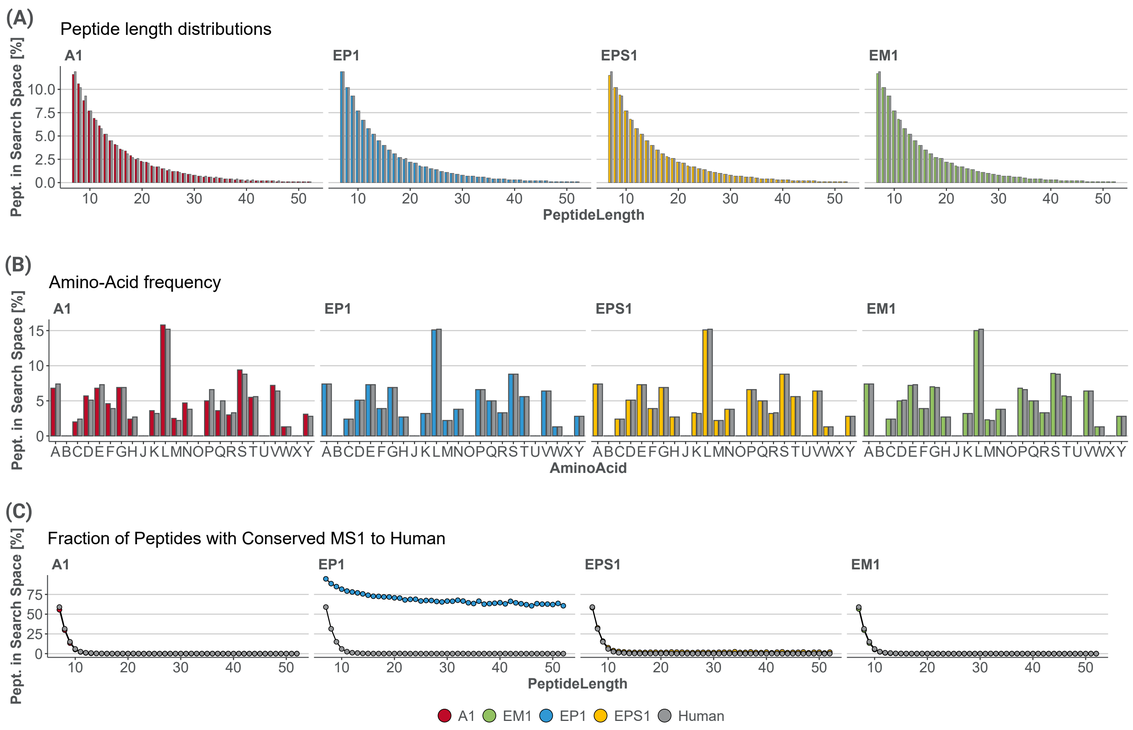

Figure 3 Identifying biases in entrapment FASTA files. All entrapment FASTA files show similar distribution to human when checking peptide length distributions (**A**) or Amino-Acid frequency (**B**). However, there is a bias when looking at the fraction of peptides that share MS1 signal with at least one peptide from human FASTA file. While A1, EPS1 and EM1 follow the profile of human FASTA, EP1 has unproportionally high frequency of peptides with conserved MS1 signal.

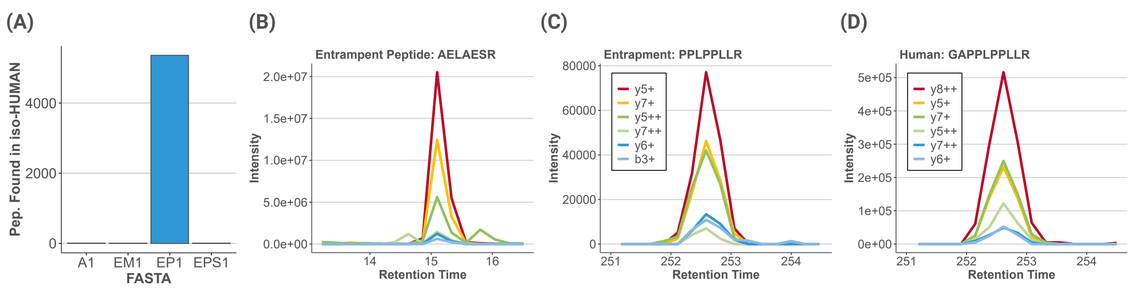

Figure 4 Importance of correct FASTA cleanup. (**A**) Number of peptides in the entrapment search space that can be found in isoform human FASTA, either as fully tryptic peptide or unspecific peptide. (**B**) Example XIC for entrapment peptide that is not found in human FASTA, but can be found in isoform of a protein. (**C**) Example of an entrapment peptide, that can be found in human FASTA as a half tryptic peptide. It show a true identification that is caused by in-source fragmentation of human peptide (**D**).

RESULTS

ESTIMATED ENTRAPMENT FDR

Figure 5 Estimated eFDR for a Thermo dataset (4 runs) at different Protein FDR filters for different FASTA files. The two FASTA files without detected biases, EPS1 and EM1, look much more similar to A1 FASTA and show valid FDR control.

EVALUATION OF DIFFERENT FDR SETTINGS

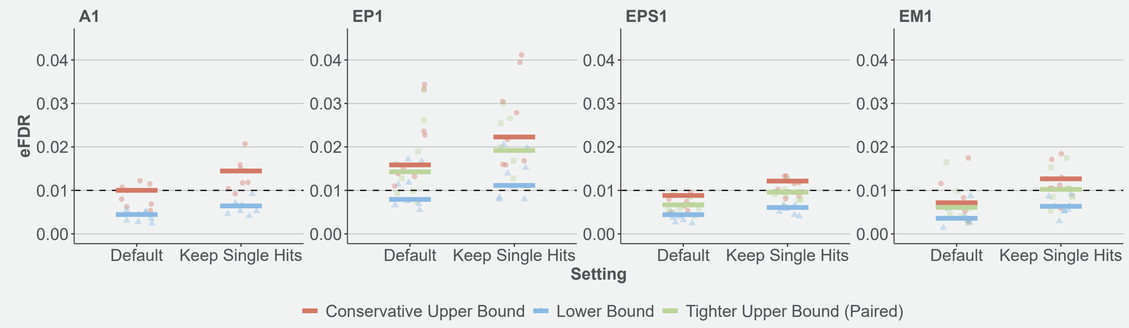

Figure 6 FDR evaluation of SN20.2 with stratified single hit protein FDR (Default) vs Keep Single Hit Proteins setting. Each dot represents a dataset (N = 8) with lower or upper bound estimation. The line represents the median for all the datasets. Notably, EP1 FASTA file consistently leads to higher eFDR estimation compared to other synthetic FASTA files or A1. The default setting leads to significantly lower eFDR. Based on the evaluation with FASTA files without confirmed bias we can conclude that Spectronaut 20 Default setting is often too conservative. Such conclusion can only be made when a FASTA file without a bias is used for eFDR estimation.

ANALYSIS WITH MULTIPLE ENTRAPMENT FASTA FILES

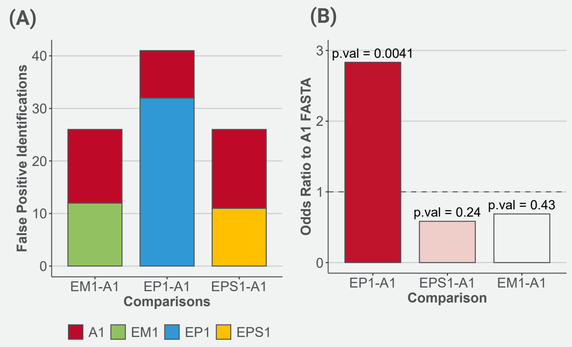

Figure 7 eFDR evaluation with combined FASTA files. The search was performed with the combination of A1 and one synthetic entrapment FASTA file. If neither of the FASTA files contained a bias, we would expect a random distribution of FP hits between the FASTA files. With the combination with EM1 and EPS1, the FPs more frequently map to A1 FASTA file. The bias is much more apparent with EP1, where majority of the FPs map to EP1 (**A**). The bias was confirmed with a Fisher's exact test, showing that the FP is almost 3x as likely to fall to EP1 FASTA file compared to A1 (p.val = 0.0041). There is no significant bias for EPS1 and EM1 (**B**).

[1] Wen, B., Freestone, J., Riffle, M. et al. Assessment of false discovery rate control in tandem mass spectrometry analysis using entrapment. Nat Methods 22, 1454–1463 (2025).

CONCLUSIONS

- **Evaluating the biases in the entrapment FASTA is crucial in entrapment FDR estimations (Figure 3).**
- **Besides eliminating the biases, proper entrapment FASTA cleanup is required (Figure 4).**
- **When an entrapment FASTA with a bias is used for FDR evaluation, the calculated eFDR can either be too pessimistic or too optimistic, depending on the bias. This leads to wrong conclusions about Software FDR control (Figure 5).**
- **Proper eFDR evaluation allows us to evaluate different Spectronaut settings, showing that for most datasets the default settings are too conservative (Figure 6).**
- **While synthetic entrapment FASTAs have the benefit of scalability and allow us to calculate the Tighter Upper Bound, they can easily introduce unwanted biases. Therefore, each FDR evaluation test should also be performed with a natural species, such as *A. thaliana* as a control.**

CONTACT

For project consultation and contract research services, please contact:

Monika Pepelnjak, PhD
Senior Data Scientist II

E: monika.pepelnjak@biognosys.com