General guidelines for validation of decoy models for HRM/DIA/SWATH as exemplified using Spectronaut

Oliver M. Bernhardt; Roland M. Bruderer; Tejas Gandhi; Saša M. Miladinović; Magdalena Bober; Tobias Ehrenberger; Oliver Rinner; Lukas Reiter
Biognosys, Zurich, Switzerland

Introduction

Data-independent acquisition (DIA) combined with targeted data analysis is a powerful proteomics technology for quantitative protein profiling across many samples or conditions. In order to efficiently analyze the large amount of data generated in DIA, statistical approaches to control the error rates based on representative target-decoy models is vital.

Our goal was to develop a "fair" decoy model that closely represents false peptide identifications to guarantee an accurate false discovery rate (FDR) estimation. Here we suggest standard procedure to verify a novel decoy model.

Methods

To validate the decoy models, a set of 5000 positive and 5000 negative control peptides was generated. The positive control was generated from a human HEK-293 spectral library and filtered for the 5000 most abundant peptides. The negative control was generated similarly from cumulative *E. coli* shotgun runs. A peptide is defined as a unique modified sequence plus charge combination.

Protein extracts of the human HEK-293 cell line were spiked with the HRM Kit and measured in six replicates on a Thermo Scientific Q Exactive mass spectrometer with a 2h gradient in HRM mode and analyzed in Spectronaut. To perform a fair analysis the human HEK-293 library was combined with the E. coli library such that the analysis software did not know about the positive and negative control. This combined library was then tested with three common decoy approaches.

A null distribution was estimated based on the Cscore distribution of the decoy peptides to calculate p-values for all target peptides similar to mProphet [1]. The p-values are then used to estimate the FDR [2].

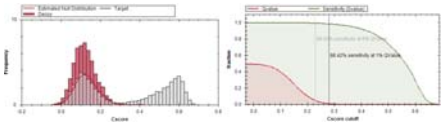

Figure 1. Left: Cscore distribution of scrambled decoys (red) and the combined human HEK-293 and *E. coli* spectral library (grey) as visualized in Spectronaut. The null distribution is estimated on the decoys using kernel density estimation and scaled on the estimated null count (red line). Right: Estimated sensitivity and FDR (Qvalue) at a certain Cscore cutoff.

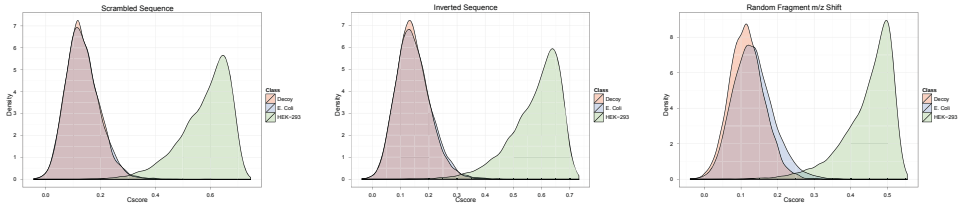

Figure 2. Cscore distributions of the human HEK-293 assays, *E. coli* assays and the three decoy models. The scrambled as well as the inverse sequence model both show no significant difference between the decoy distribution and the distribution of truly absent *E. coli* peptides (p-value > 0.05 using a two-sample K-S test). The Cscore distribution of the random fragment m/z shift model shows a significant difference when compared to the negative control (p-value = $2.2e^{-16}$).

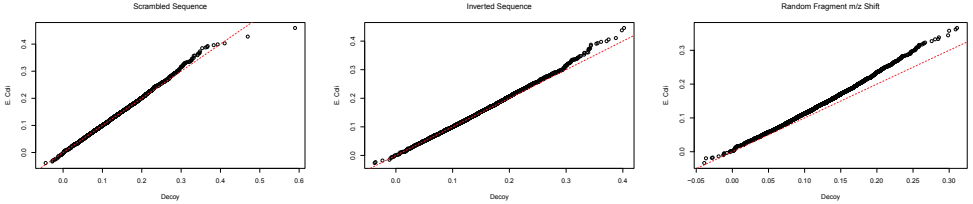

Figure 3. Goodness of fit analysis using Cscore quantile-quantile plots of the three different decoy models (Scrambled, Inverted and Fragment m/z Shift) compared to the negative control set (*E. coli* assays). Both the scrambled and the inverted sequence decoy model show a high coefficient of determination to the 45° line (indicated as the dashed line in these plots) of 99.7% and 99.5% respectively. The random fragment m/z shift model shows a systematic error from the 45° line resulting in a coefficient of determination of only 89.0%.

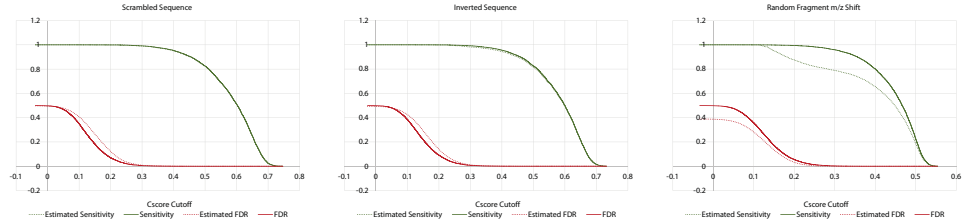

Figure 4. Comparison of the estimated FDR and sensitivity to known classes. Both, scrambled and inverted sequence, lead to very accurate estimates of the FDR and sensitivity. In both cases the estimated FDR was conservative (0.6% true FDR at 1% estimated FDR cutoff in both cases), slightly overestimating the true number of false identifications. The estimated FDR based on the random fragment m/z shift model on the other hand is underestimating both, the true number of false identifications and the sensitivity of the method. This results in a higher number of false identifications at a given FDR cutoff (2.3% true FDR at 1% estimated FDR cutoff). Dashed lines indicate estimations as calculated in Spectronaut, solid lines indicate estimations based on the known classes (human HEK-293 and *E. coli* assays).

	Identifications of Full HEK-293 Spectral Library (47,835 Peptides)			
Replicate	***E. Coli***	**Scrambled (ns)**	**Inverted (ns)**	**m/z Shift (***)**
R01	37,234	37,502	37,488	40,183
R02	40,144	39,567	40,479	42,763
R03	37,817	38,545	38,280	40,916
R04	39,045	38,987	39,140	41,934
R05	39,778	39,228	40,326	42,162
R06	38,287	40,056	39,505	41,762
Average	**38,717**	**38,981**	**39,203**	**41,618**

Table 1. Number of identifications for an FDR of 1% using different decoy strategies in a full scale analysis of 6 technical replicates (2h gradient). Both the scrambled as well as the inverted sequence decoy model did not show a significantly different number of identifications if compared to an analysis using truly absent peptides as decoys (*E. Coli*). The random fragment m/z shift model on the other hand did show a significant difference (*** p-value = 0.0006).

Conclusion

Given a representative decoy model, accurate FDRs can be estimated for DIA experiments. Two of the three tested models where able to generate decoy distributions that do not show a significant difference to the *E. coli* negative control set (p-value > 0.05).

The third model, although resulting in the highest number of identifications, did not reflect the distribution of truly absent peptides. This lead to a wrong FDR estimation and an increased number of false identifications. Conclusively, the number of identifications alone should not be used as a qualifier to determine the validity of a certain target-decoy model. Instead, a thorough analysis comparing the decoys with a control set of known false signals should be performed.

As implemented in Spectronaut, the scrambled sequence model appears to produce the most reliable results of the three compared decoy strategies.

All runs and libraries as well as the software used for this analysis can be downloaded at www.spectronaut.org.

References

[1] Reiter, L., et al., "mProphet: automated data processing and statistical validation for large-scale SRM experiments" Nature methods 8.5 (2011): 430-435

[2] Storey, J. "A direct approach to false discovery rates." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64.3 (2002): 479-498.