RUHR-UNIVERSITÄT BOCHUM

Medical Faculty
Medical Proteome Center
Medical Bioinformatics / Functional Proteomics

Do you consider "feature-peptides"? Enabling search-engines to search for already annotated feature-information from UniProtKB

Dominik Lux, Prof. Dr. Katrin Marcus, Prof. Dr. Martin Eisenacher, Dr. Julian Uszkoreit
{dominik.lux, katrin.marcus, martin.eisenacher, julian.uszkoreit}@rub.de

Abstract

In today's identification workflows, it is common to select FASTA-databases containing mostly canonical and isoform-protein-sequences. The workflows then in-silico digest the database to mimic the digestion-enzyme used in the sample preparation and try to match the resulting in-silico peptides with measured spectra with so-called search engines to generate peptide-spectrum-matches (PSMs). However, in the biological domain, sequences are not as uniform as illustrated in commonly used FASTA-databases, not covering variational sequences in species (except for isoforms) and/or mutations that may occur from diseases or are specific to an individual. Furthermore, biological processes continue which could yield already cleaved signal-/pro-peptides or other peptides, like Abeta 40/42. Although UniProtKB already provides so-called feature-information about molecule processing, variational and further sequence-annotations, the search-engines, on the other hand, cannot identify such sequences, due to missing entries in the FASTA-database. Here, we propose a workflow, which generates FASTA-databases in a sophisticated manner while considering various feature-information provided by the UniProtKB. In this workflow, we first use so-called protein-graphs to represent all possible sequence outcomes. Due to the exponentially growing search space, we further implemented a traversal algorithm as a second step, which takes the MS2-precursors from a dataset and exports only peptides fitting to the MS2-precursor (while considering post-translational-modifications) into a peptide-FASTA-database. The resulting peptide-FASTA-database then can be used instead of the original FASTA-database as a drop-in-replacement in common identification workflows, as long as the search engine is able to search without digestion. We applied this workflow on the dataset PXD007555, containing measured CSF from 12 individuals, and show the applicability of generating such peptide-FASTA-databases. Further, we used Comet and Percolator with the generated FASTA-database to showcase the possibility that search-engines can use such FASTA-databases. We highlight interesting identification results and show that around 5% of PSMs in PXD007555 originate from peptides resulting from feature-information, demonstrating that such "feature-peptides" are indeed identifiable. Though further investigation is still needed, this approach shows a promising step to further increase the identification-ratio in CSF-Samples.

Materials and Methods

To implement the workflow, we use Nextflow [1]. An overview and brief description of the workflow can be found in **Figure 1**. This workflow requires two input files: A list of MGF-files of an dataset and the database of a set of proteins (e.g. species) in the SP-EMBL-format. The SP-EMBL-database is similar to a FASTA-database, where instead of FASTA-entries, a text-format containing the canonical sequence as well as feature-information of proteins is downloaded directly from UniProtKB [4].

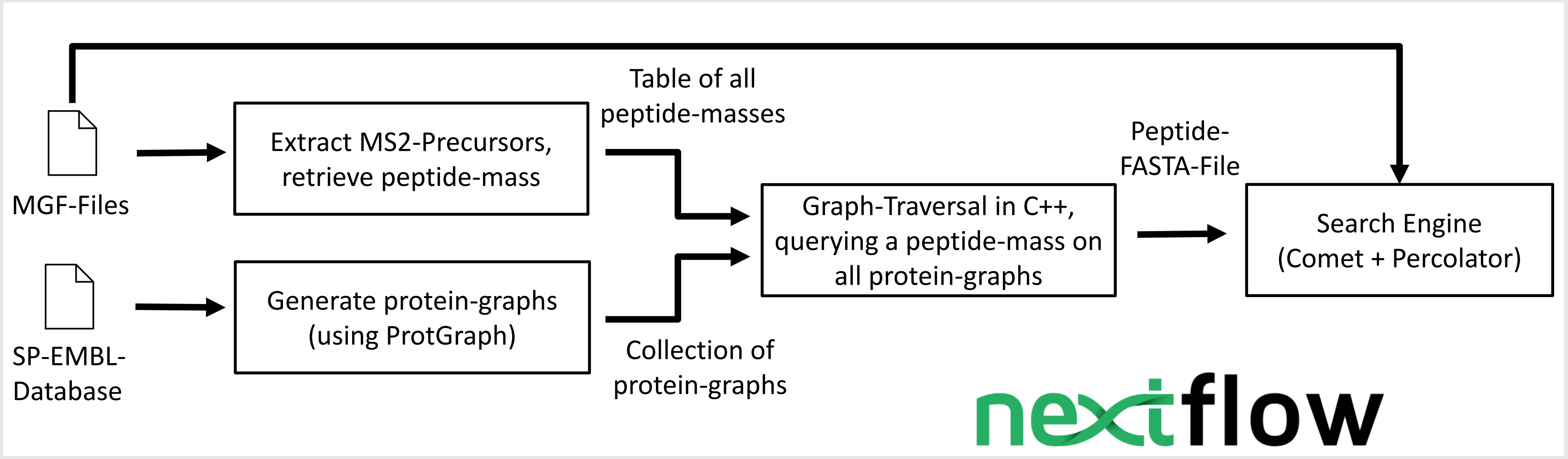
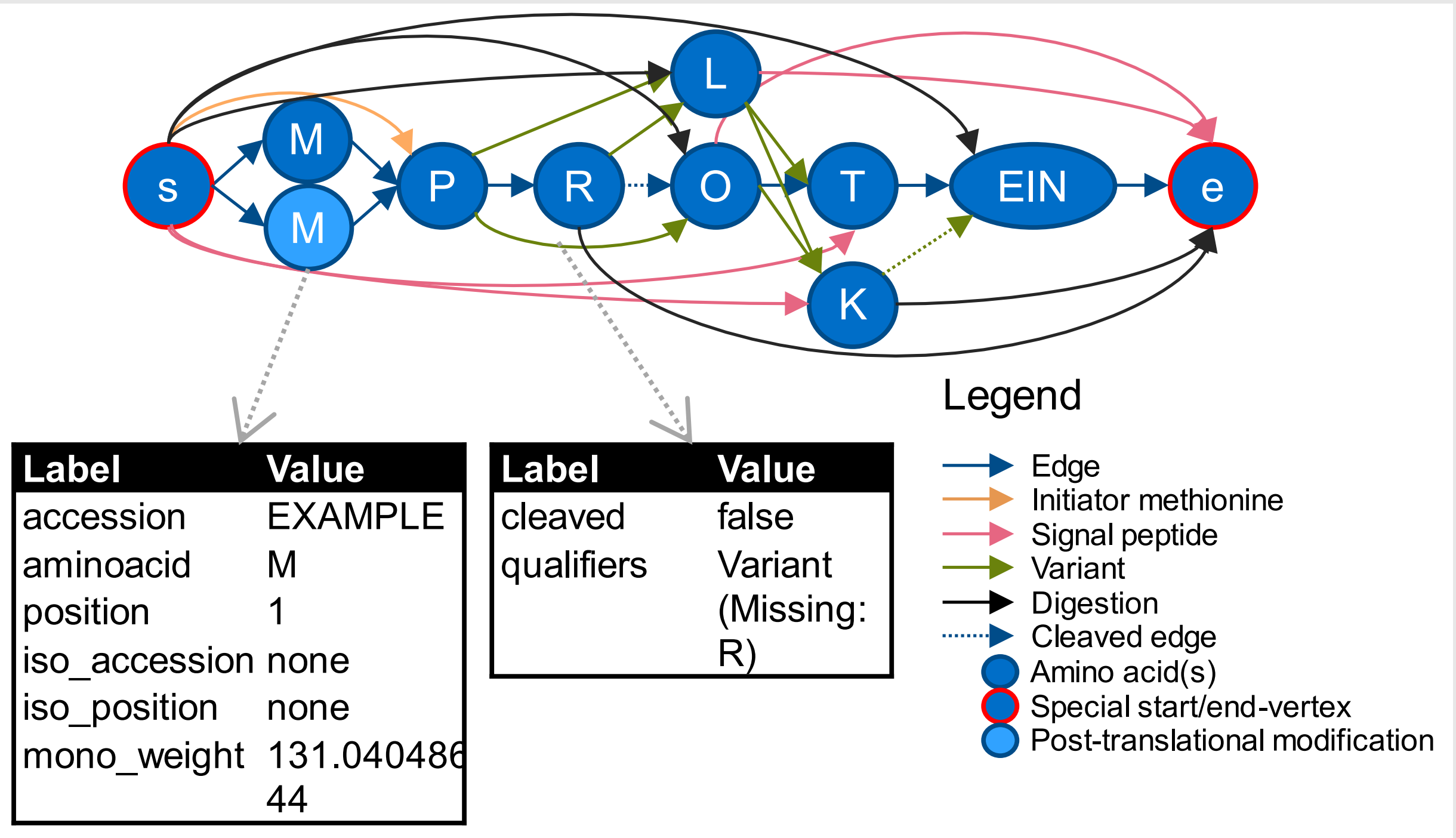

MGF-Files
Extract MS2-Precursors, retrieve peptide-mass
Table of all peptide-masses
SP-EMBL-Database
Generate protein-graphs (using ProtGraph)
Collection of protein-graphs
Graph-Traversal in C++, querying a peptide-mass on all protein-graphs
Peptide-FASTA-File
Search Engine (Comet + Percolator)
nextflow

Figure 1: *Implemented workflow in Nextflow [1]. It requires a list of MGF-files as input (dataset), and an SP-EMBL-database of a set of proteins (e.g. from one or multiple species). From the MGF-Files, we extract MS2-precursor and their corresponding charges and the original peptide-mass in Dalton is calculated and saved in a table. With ProtGraph [2], we generate protein-graphs with user-selected features (and PTMs). The graph-traversal-implementation, queries peptides in protein-graphs using the peptide-mass (while considering the mass-tolerance). The resulting peptide-FASTA with the MGF-Files can be then used with an search-engine.*

The SP-EMBL-Database can be used directly with ProtGraph [2] to generate protein-graphs, with user-selectable annotations. Currently, ProtGraph allows specific feature-information using SP-EMBL as well as digestion and fixed/variable post-translational-modifications (PTMs). An example can be found in **Figure 2**. In addition, we pre-compute and save pattern-databases (PDBs) [3] for each individual node and a specific topological order for each protein- graph via ProtGraph. The second step then uses the peptide-masses to retrieve peptides from protein-graphs. Here, the algorithm iterates over the saved topological order. Starting from node s, we save all possible partial paths fitting to the queried mass and by utilizing PDBs, we can filter out these paths early where the queried mass is not achievable. Additionally, we implemented an upper limit of applied variants and peptide-mass. Paths fitting to the queried mass and starting from node s and ending in node e are then exported as an FASTA-entry to be used by search engines with the corresponding MGF-files.



Label	Value
accession	EXAMPLE
aminoacid	M
position	1
iso_accession	none
iso_position	none
mono_weight	131.040486 44

Label	Value
cleaved	false
qualifiers	Variant (Missing: R)

Figure 2: *Example protein-graph generated by ProtGraph. It contains the following feature-information: 1 SIGNAL (pink), 3 VARIANTs (green), 1 INIT_MET, which is skipped and 1 PTMs (Oxidation of M). This protein-graph is also digested (black, miscleavages are denoted by dotted lines) and an excerpt of node and edge-attributes is illustrated.*

Parameters

We applied this workflow on the dataset PXD007555. We used the homo sapiens proteome (July 2022), which contains around 81 000 proteins to generate protein-graphs. All parsable features (except MUTAGEN), were used excluding the VARIANT feature for the proteins P04637 (P53) and P68871 (HBB), with digestion set to Trypsin and intervals per node to 32. For FASTA-generation, search and protein-graph-generation we used the same search-settings/PTMs as provided in PXD007555 and searched with Comet [5] and Percolator [6]. We also set a limit of up to 5 variants and 4500 Da per peptide. As a cut-off value we used a FDR of 1% using the q-value. We compared the results to the identification results, where we used the same human proteome as a protein-FASTA-database directly.

Results

The generation of the protein-graphs needed in total less than 8 minutes, with the pre-computed topological order and PDBs. Generating the FASTA-file on the other hand needed 14 hours to complete, due to the huge search space. We observed, that the graph-traversal requires more time the higher the queried peptide mass is. In most protein-graphs, this is negligible, due to their simple structure, yielding results in a few milliseconds. The runtime therefore is dominated by the most complex protein-graphs, being the reason why the VARIANT feature was removed for the proteins P04637 (P53) and P68871 (HBB). VARIANT-features (and also other aminoacid-replacing-features) increase the number of peptides (the search space) in a protein-graph exponentially. **Figure 3** illustrates their running time.

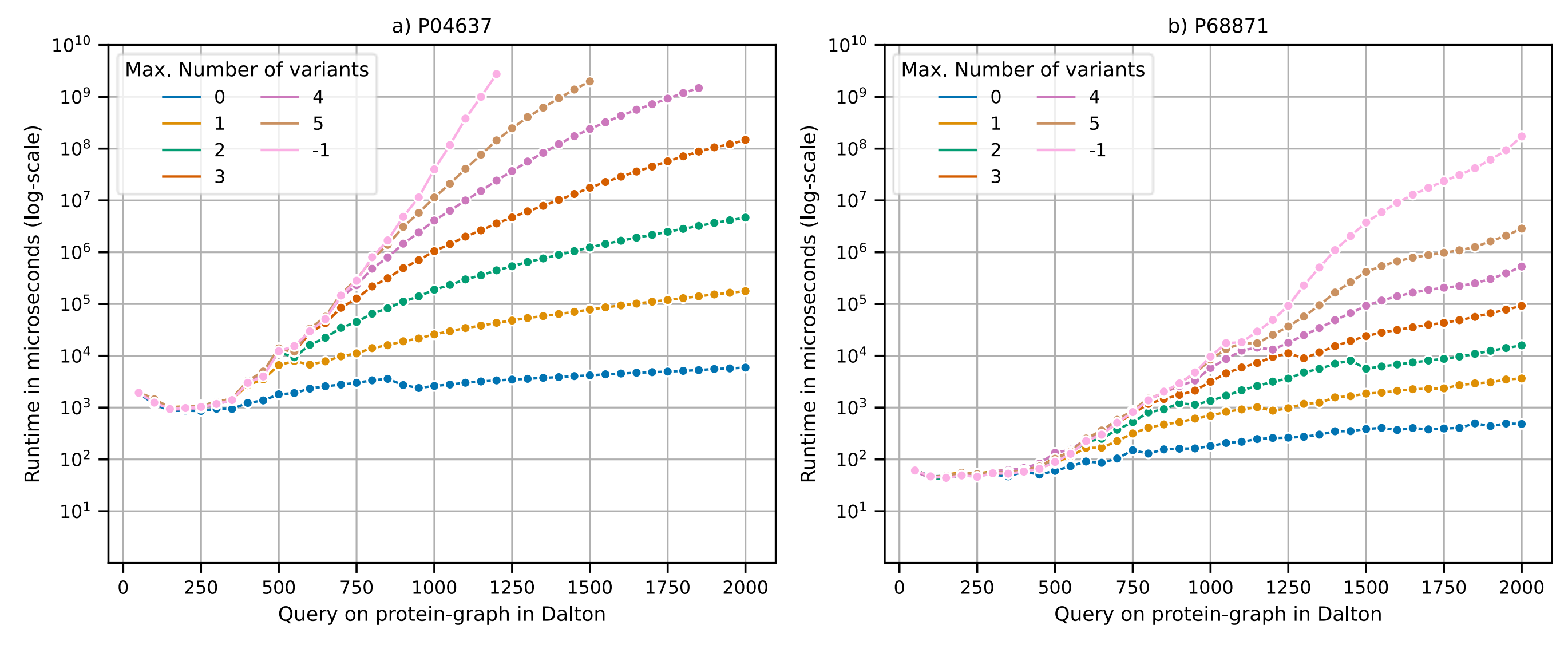


Figure 3: *Runtime of the graph-traversal algorithm in milliseconds for the complex protein-graphs from a) P04637 and b) P66871. The runtime increases, the higher the queried mass is, due to more potential paths which need to be checked. In other words, the search space needs to be further explored. Limiting the number of variants per peptide significantly influences the traversal runtime.*

The resulting FASTA-database from this workflow contains 20 million peptide-entries and is compared to the number of peptides in the digested protein-FASTA-database 6.5 times larger. The identification with the peptide-FASTA-database needed in total 1 hour and identified 9163 more peptide-spectrum-matches (PSMs) compared to a search without features. **Figure 4** shows the number of PSMs after an FDR-cut-off of 1%. From those PSMs, 21 077 (around 5%) originate only from peptides with annotated feature-information. **Figure 5** provides an overview. **Figure 6** showcases two identified unique peptides.

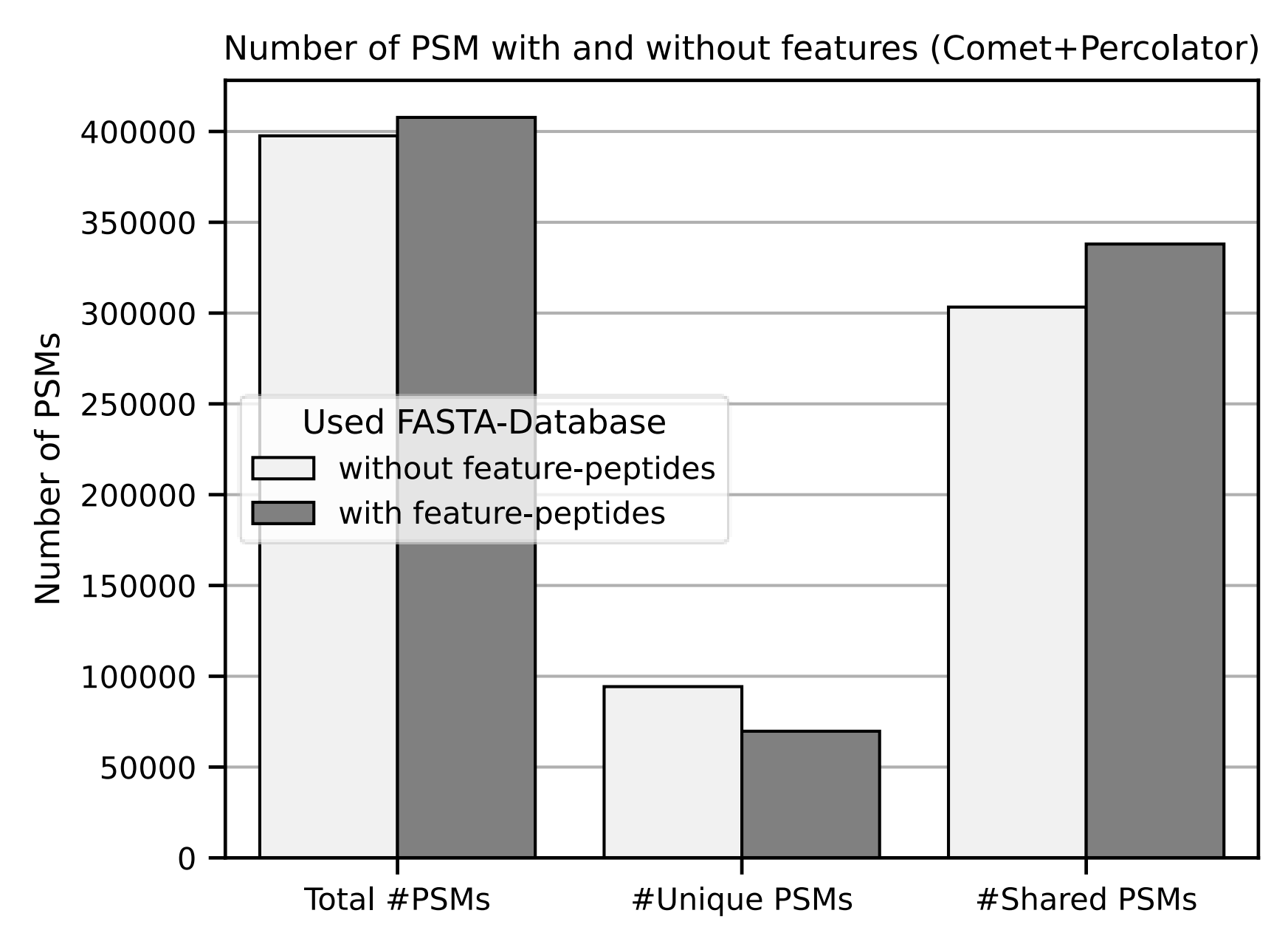


Figure 4: *Number of PSMs in total and divided by PSMs, where the peptide is unique in the FASTA-database or shared. These are the identification results from Comet and Percolator. It is expected that the number of unique PSMs drops (while the number of shared PSMs increases), due to the larger FASTA-database, when considering feature-information. The total number of PSMs slightly increased, with the FASTA-database containing feature-peptides.*

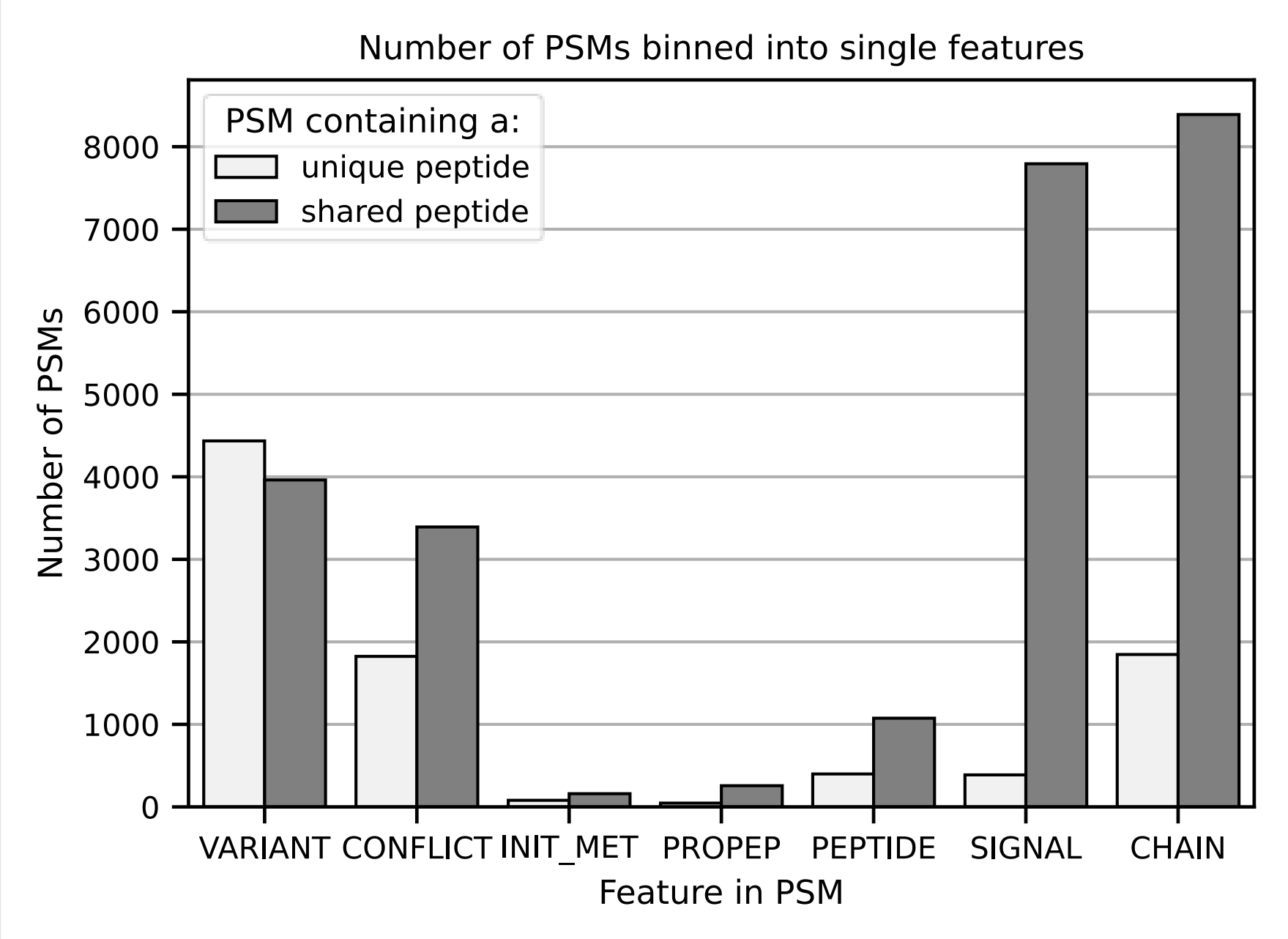


Figure 5: *PSMs binned by single features. It is noticeable that a high amount of VARIANT and CONFLICT-features were identified in CSF, which describe variations from the canonical sequence. There is also a high amount of found PSMs with the feature SIGNAL and CHAIN (and PEPTIDE). These features indicate specific cleavage positions on the sequence, which can occur. Note: PSMs are overcounted since they can contain multiple features at once.*

Showcase of two identified unique peptides originating from feature-information provided by UniProtKB:

a)

FASTA-Entry:

```
>P13521(549:566,mssclvg:1,PEPTIDE[527:566])
EHLNQGSSQETDKLAPVS
```

Gene: SCG2
Protein: Secretogranin-2
# found PSMs: 71

→ Non-Tryptic
→ Original peptide: PRO_0000432735
VPGQGSSEDDLQEEEQIEQAIK**EHLNQGSSQETDKLAPVS**
→ Manserin, neuropeptide

b)

FASTA-Entry:

```
>P02766(69:90,mssclvg:1,VARMOD[1:1,NTERM:42.010565],VARIANT[70:70,S -> R,
VAR_007566],VARIANT[72:72,S -> P, VAR_007567],VARIANT[74:74,E -> G, VAR_007568],VARIANT[75:75,L
-> P, VAR_007569],VARIANT[78:78,L -> H, VAR_007570])
TREPGGPHGHTTEEEFVEGIYK
```

Gene: TTR
Protein: Transthyretin
# found PSMs: 109

→ 5 variants at once in peptide
→ "Famous" mutations, referencing 9 publications
→ All variants under the disease: "Amyloidosis, transtheretin-related"
→ Mostly in heart, kidney, nervous system, …

Figure 6: *Showcase of two identified unique peptides, originating from feature-information. a) displays the protein P13521 (SCG2), where a neuropeptide was found. In a normal search, this peptide would not be identified, due to the non-tryptic cleavage at the C-terminus (in* ***bold****). b) highlights an unique peptide from P02766 (TTR), with 5 variants applied at once (in* ***bold****). Interestingly all variants are annotated under the same disease in this protein. Both highlighted peptides are reasonable to find in CSF.*

Discussion

We showed that the generation of a peptide-FASTA-database with features is possible for the proteins in homo sapiens with a few restrictions. In our experience, the generation of a peptide-FASTA-database, using this workflow, with other species is most likely possible without any restrictions, since those are not as well annotated as homo sapiens. We also show that the generated FASTA-database can be used with Comet and Percolator and that the number of PSMs increased slightly for this dataset. We can be certain that the PSMs with annotated feature-information can only originate from spectra, which have a better match to a peptide originating from a feature than to a canonical tryptic peptide, since we only enrich the canonical tryptic peptides with feature-peptides.

Although we identified feature-peptides, the analysis and further processing of the identification results of this workflow becomes difficult to interpret, due to a custom FASTA-header-format provided by ProtGraph and references within to specific features which need to be manually looked up in UniProtKB. As a next step, we plan to increase the explainability of such identification results and want to extend this workflow to also report quantification information.

References:

[1] Di Tommaso, P., Chatzou, M., Floden, E. et al. Nextflow enables reproducible computational workflows. Nat Biotechnol 35, 316–319 (2017). https://doi.org/10.1038/nbt.3820
[2] Dominik Lux. (2022). mpc-bioinformatics/ProtGraph: Release of ProtGraph v0.3.8 (0.3.8). Zenodo. https://doi.org/10.5281/zenodo.7255518
[3] Schmidt, Tim, Lukas Kuhn, Bob Price, Johan De Kleer, and Rong Zhou. "A depth-first approach to target-value search." In Symposium on Combinatorial Search (SOCS-09). 2009.
[4] The UniProt Consortium, UniProt: the universal protein knowledgebase in 2021, Nucleic Acids Research, Volume 49, Issue D1, 8 January 2021, Pages D480–D489, https://doi.org/10.1093/nar/gkaa1100
[5] Eng, J.K., Hoopmann, M.R., Jahan, T.A. et al. A Deeper Look into Comet—Implementation and Features. J. Am. Soc. Mass Spectrom. 26, 1865–1874 (2015). https://doi.org/10.1007/s13361-015-1179-x
[6] The, M., MacCoss, M.J., Noble, W.S. et al. Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. J. Am. Soc. Mass Spectrom. 27, 1719–1727 (2016). https://doi.org/10.1007/s13361-016-1460-7