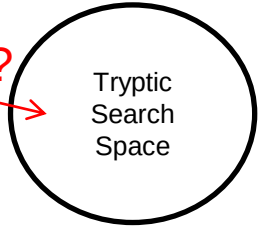
RUHR-UNIVERSITÄT BOCHUM

REPRESENTING PROTEINS AND PEPTIDES WITH VARIATIONAL FEATURE INFORMATION IN GRAPHS USING PROTGRAPH

Dominik Lux, Prof. Dr. Katrin Marcus, PD Dr. Martin Eisenacher, Dr. Julian Uszkoreit
Ruhr-University Bochum, Medical Faculty, Medical Proteome Center, Medical Bioinformatics
{dominik.lux, katrin.marcus, martin.eisenacher, julian.uszkoreit}@rub.de

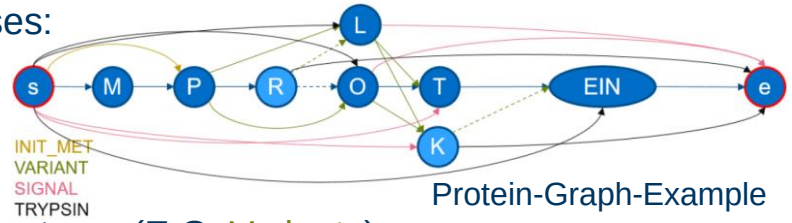
We want to look at the size of the tryptic search space (upper bound) of different protein-databases ?



Overview and Motivation

- We use ProtGraph to generate protein-graphs from protein-entries containing feature information like “Signal Peptides”, “Variants” and more.
  - Protein-graphs can also be “Digested” in silico, which then contain all possible peptides in a compact graph representation.
  - We use these graphs to calculate the upper bound of peptides represented in them.

- In this presentation / poster we are especially interested to explore the search space on different protein-databases while including / excluding feature information
  - Here, we investigated the following protein-databases:
    - E. Coli / Human (/ UniProt)



Protein-Graph-Example

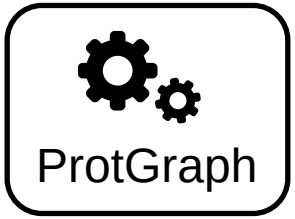
- Why could this be important/beneficial?
  - Check if it is feasible to generate FASTA-files with features (E.G. Variants)
  - Illustrate the complexity of the search space

ProtGraph is available on:
Github: https://github.com/mpc-bioinformatics/ProtGraph
PyPI: pip install protgraph
BioConda: conda install -c bioconda protgraph

ProtGraph in a Nutshell

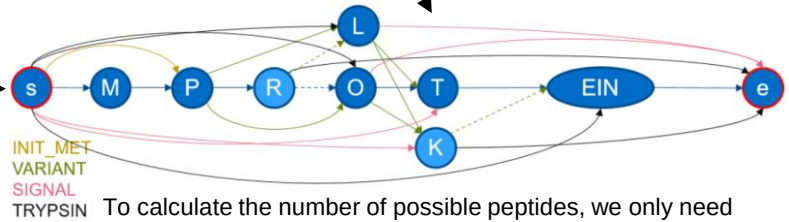
Instead of using FASTA-entries as an input, we use specifically the SwissProt-EMBL-format. It contains the canonical sequence of a protein and additionally contains feature information like variants, isoforms and more.

SwissProt-EMBL



Adding **N** SwissProt-EMBL-entries into ProtGraph will generate **N** many graphs

Generated graphs are always DAGs!



To calculate the number of possible peptides, we only need the following:

- Topological Order (retrievable since all graphs are DAGs)
- Dynamic Programming

Algorithm (simplified):
To retrieve the number of possible peptides the number of possible paths between s and e is retrieved instead. This can be calculated by summing the number of possible paths from node s to a node **x** until e is reached, while going through the topological order.

Caution:

- This is an upper limit! Graphs may contain the same peptide multiple times and therefore the number of peptides may be smaller.
- The number of possible peptides contains peptides with up to “infinite” many miscleavages!

- We extended ProtGraph to include:
  - → Mutagens / Conflicts
  - → Aminoacid-Ambiguity-Replacements
- In total ProtGraph has 512 possibilities on how to generate the corresponding protein-graphs

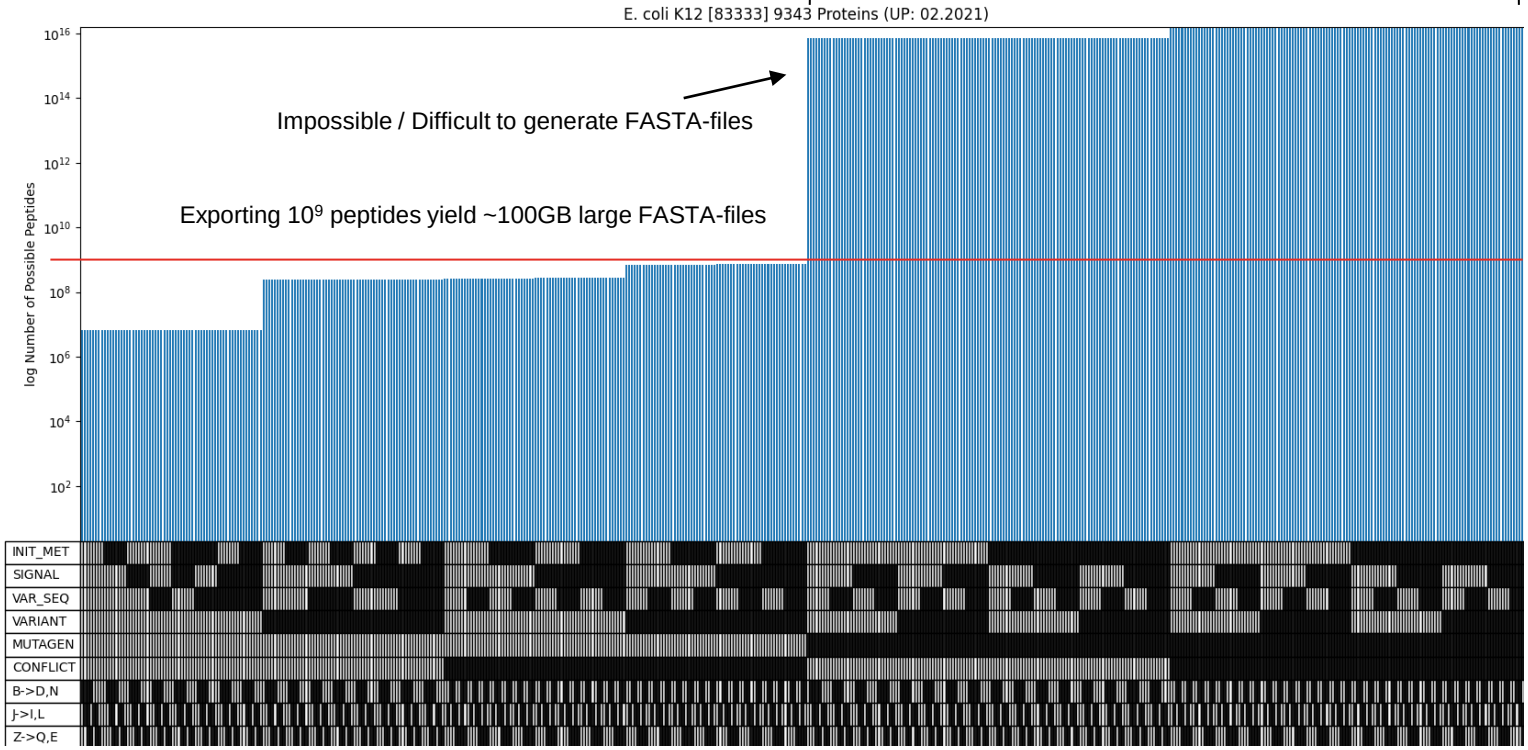
→ Look into the search space with all possibilities!

E. Coli K12, 9343 Proteins

Generating this figure needed in total ~2 hours

- Each vertical bar represents a combination of applied or not applied features
- Bars are sorted in ascending order
- Size of search space is very similar while including specific features[1]
- Exporting FASTA-files using all features is not feasible for E. Coli K12

[1]Only change the search space slightly



RUHR UNIVERSITÄT BOCHUM RUB

Applied Feature
Not Applied Feature

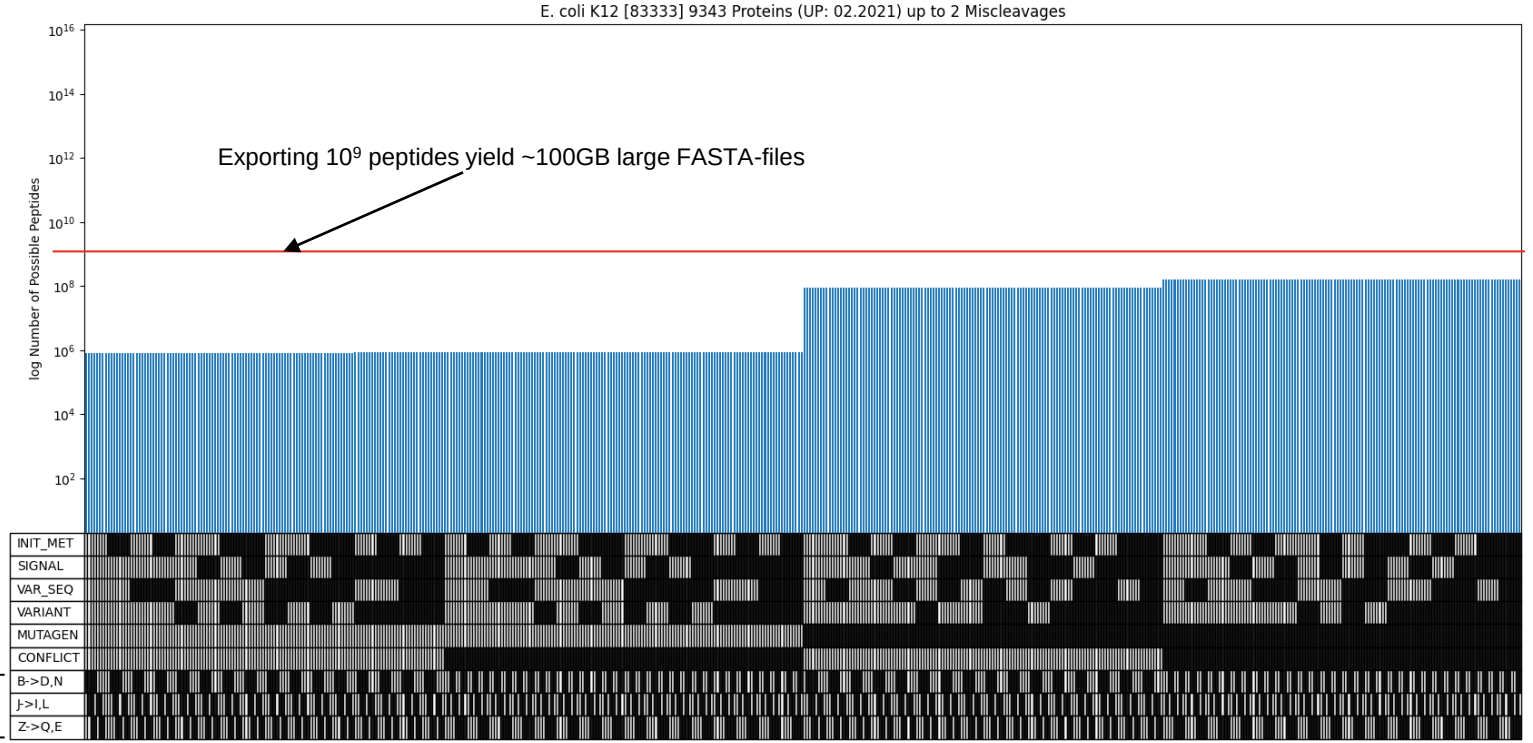
E. Coli K12, 9343 Proteins

Generating this figure needed in total ~2 hours

“Large” increase while using mutagens

- Looking at the number of possible peptides with up to 2 miscleavages decreases the search space
- The export is now feasible if only peptides with up to 2 miscleavages are considered

→ A generated FASTA-file of E. Coli K12, containing all possible peptides while considering all 9 features could be used by a search engine



Only change the search space slightly

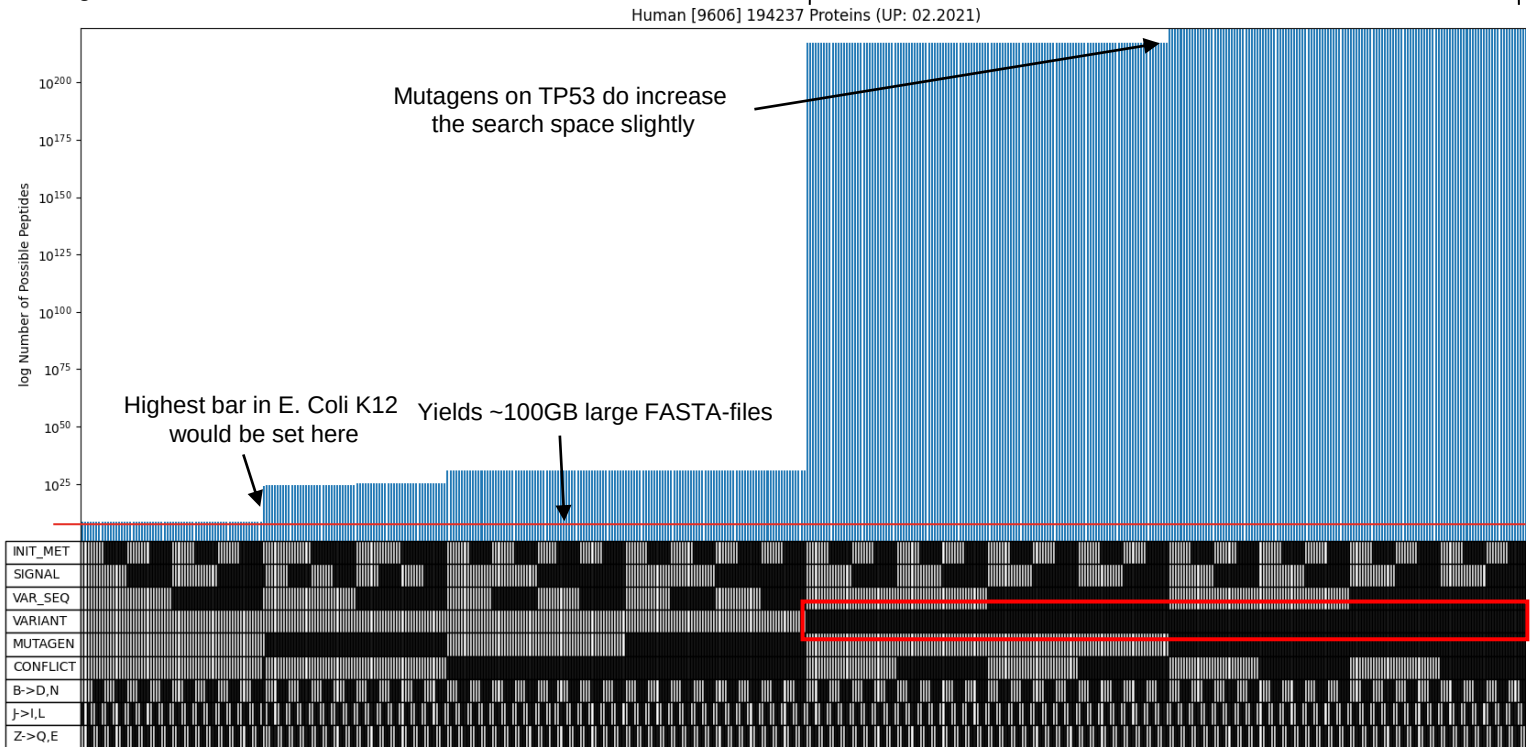
RUHR UNIVERSITÄT BOCHUM RUB

Human,194 237 Proteins

This figure needed ~5 days to be generated

- This figure differs drastically to the previous figures of E. Coli K12
- The search space increases here drastically, at some feature.
- The main factor for the large search space increase:
  - Variants
- Interestingly, the search space is dominated by a single protein:
  - P04637
  - TP53
- Reason: TP53 has the highest number of variants (~1000)

Only change the search space slightly

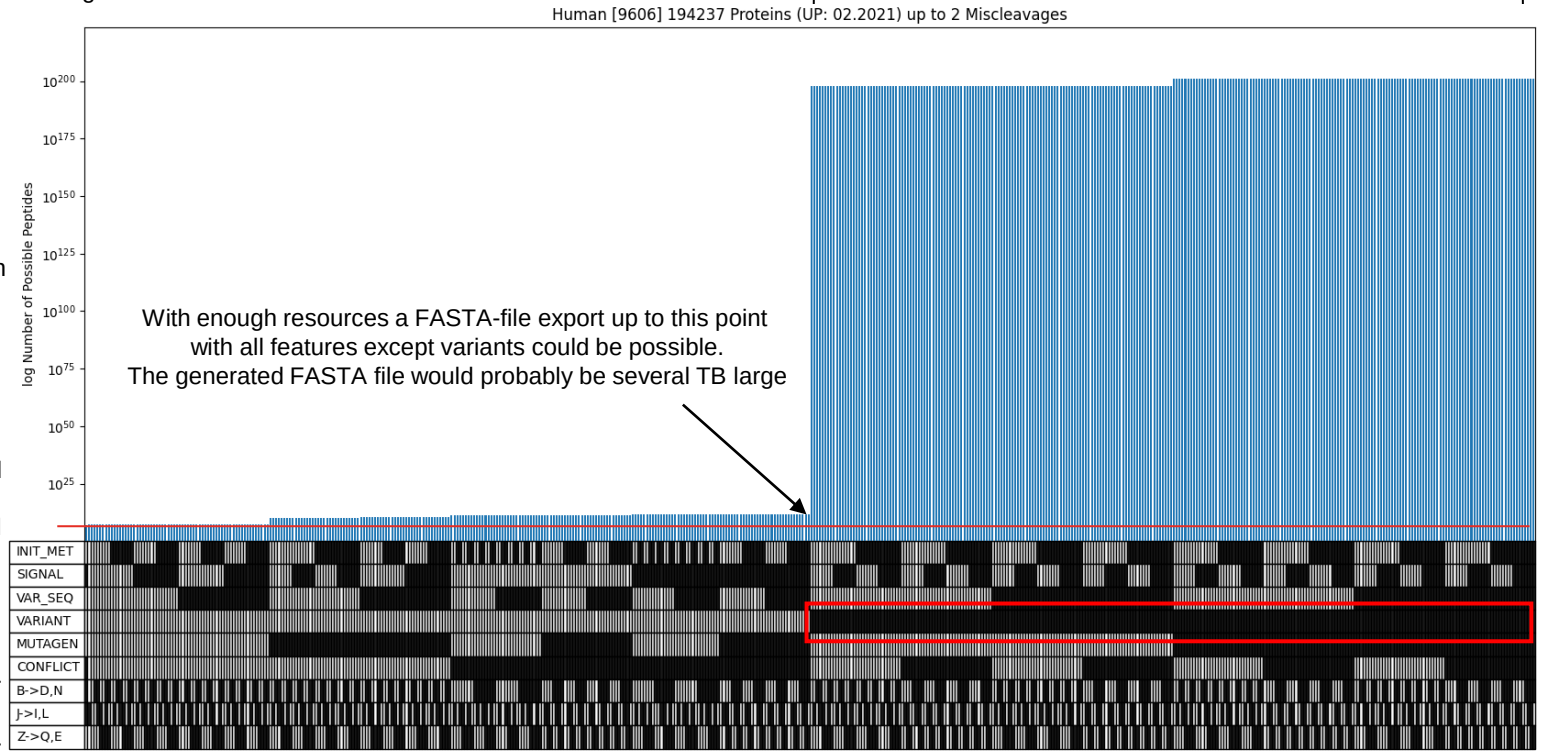


Applied Feature
Not Applied Feature

Human,194 237 Proteins

This figure needed ~5 days to be generated

Search space “explodes” while using variants



- Looking into the search space of possible peptides with up to 2 miscleavages yields a similar large search space
- The search space still explodes while using variants
- A FASTA-file export with all features in the human protein-database would still be not feasible

Only change the search space slightly

RUHR UNIVERSITÄT BOCHUM RUB

Applied Feature
Not Applied Feature

Complete UniProt, ~220M Proteins

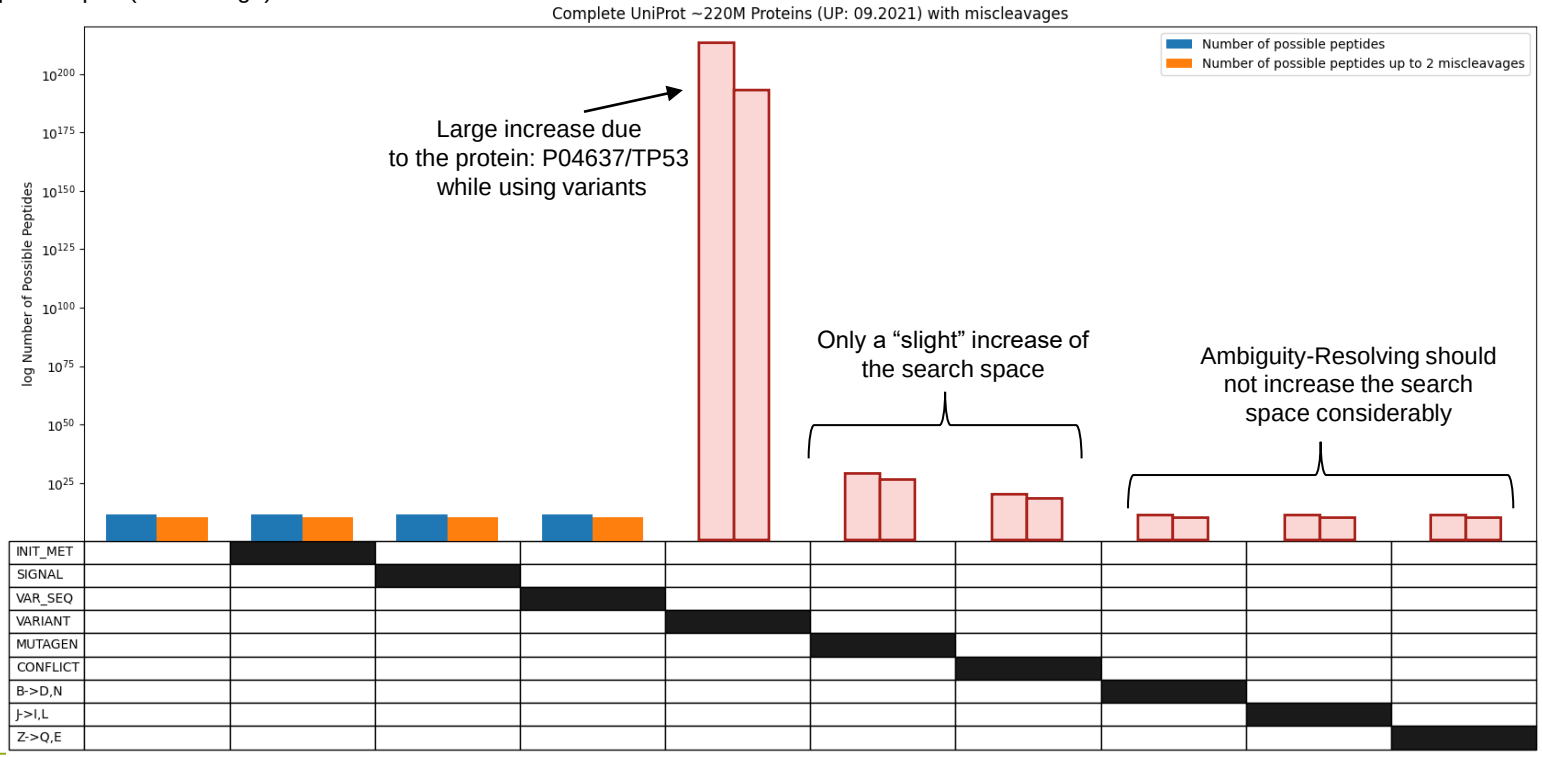
Calculation needed ~26h per bar-pair (blue/orange)

- Figure contains incomplete results.
- Assumptions of the search space of not calculated bars still can be made:

- Blue/Orange:
  - Calculated
- Red:
  - Assumptions

- P04637/TP53 would still dominate the search space, since it is the protein with the highest number of variants in complete UniProt.

- Mutagens/Conflicts may increase the search space slightly. This increase should be considerably less compared to variants



Applied Feature
Not Applied Feature

Complete UniProt, ~220M Proteins (updated slides)

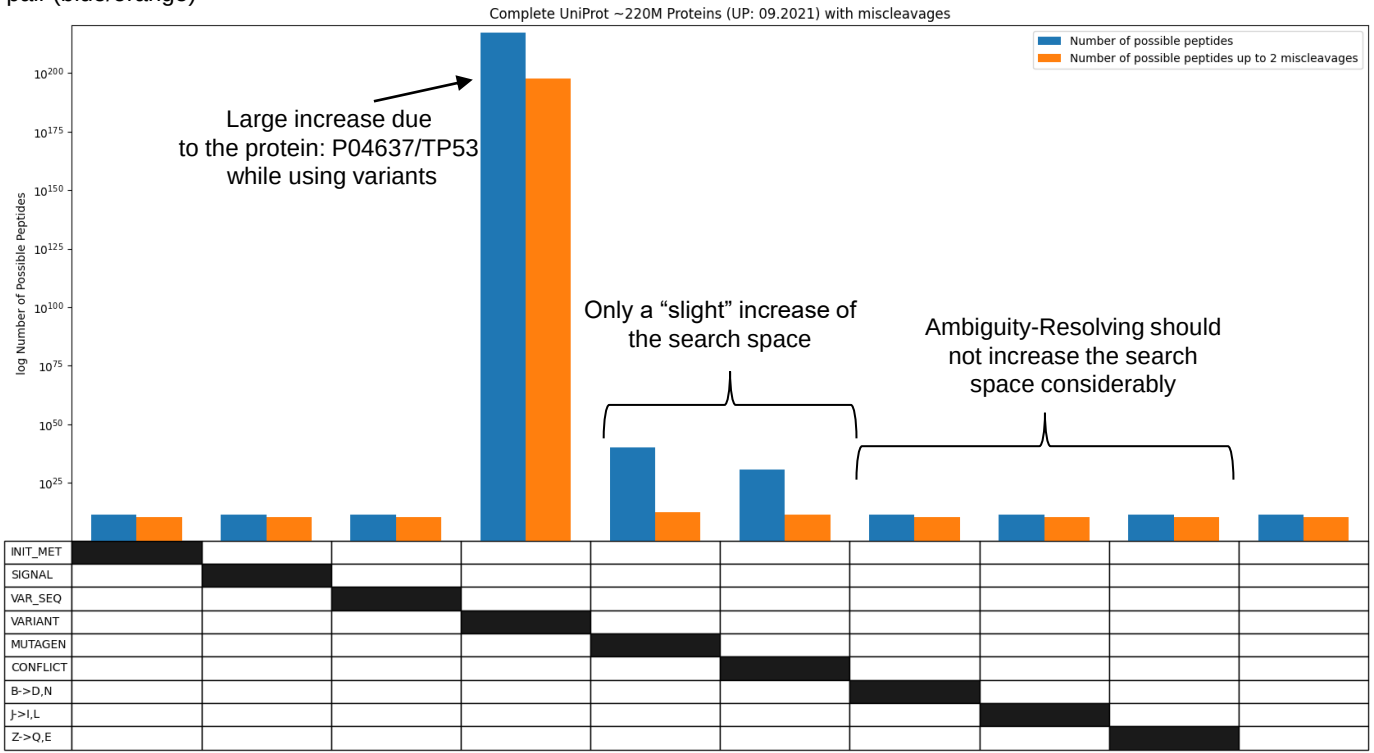
Calculation needed ~26h per bar-pair (blue/orange)

- Figure contains incomplete results.
- Assumptions of the search space of not calculated bars still can be made:

- Blue/Orange:
  - Calculated
- Red:
  - Assumptions

- P04637/TP53 would still dominate the search space, since it is the protein with the highest number of variants in complete UniProt.

- Mutagens/Conflicts may increase the search space slightly. This increase should be considerably less compared to variants



RUHR UNIVERSITÄT BOCHUM
RUB

Conclusion

- → Graphs can contain huge amounts of peptides while being small in size
  - → DAGs cannot contain infinite many peptides. They are limited to contain up to $2^{\#Nodes - 2}$ many peptides

- → The number of variants increase the search space to unmanageable amounts of peptides
  - → FASTA-Exports may not be feasible in such cases!
  - → However, FASTA-Exports may be possible with small datasets (E.G. E. coli K12)

- → These insights of the search space can be used to add more peptides to MacPepDB [1]
  - Resolving ambiguity / signal peptides / isoforms / initiator methionine (/ conflicts / mutagens) can be included safely without exploding in size

- → These statistics could be used protein-wise to design (limits on) algorithms on protein-graphs
  - This could be useful for extracting peptides directly from graphs
    - And to apply sophisticated algorithms only for complex proteins

[1] https://doi.org/10.1021/acs.jproteome.0c00967

Acknowledgement

Medical Proteome Center:

- PD Dr. Martin Eisenacher
- Dr. Julian Uszkoreit
- Dr. Michael Turewicz
- Dr. Markus Stepath
- Dirk Winkelhardt
- Daniel Kleefisch
- Karin Schork
- Sai Spoorti Ramesh

Thank you for your interest!

F953N-19

FoRUM

SPONSORED BY THE

Federal Ministry of Education and Research

FKZ 031 A 534A

de.NBI

GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

PRODI

mpc MEDIZINISCHES PROTEOM CENTER

RUHR UNIVERSITÄT BOCHUM

RUB