System for Identification of Mutations

using Mass Spectrometry of Proteome

Miroslav Hruška, Jiří Voller, Petr Džubák, Marián Hajdúch

INSTITUTE OF MOLECULAR AND
TRANSLATIONAL MEDICINE

PALACKÝ
UNIVERSITY
OLOMOUC

Outline

Outline

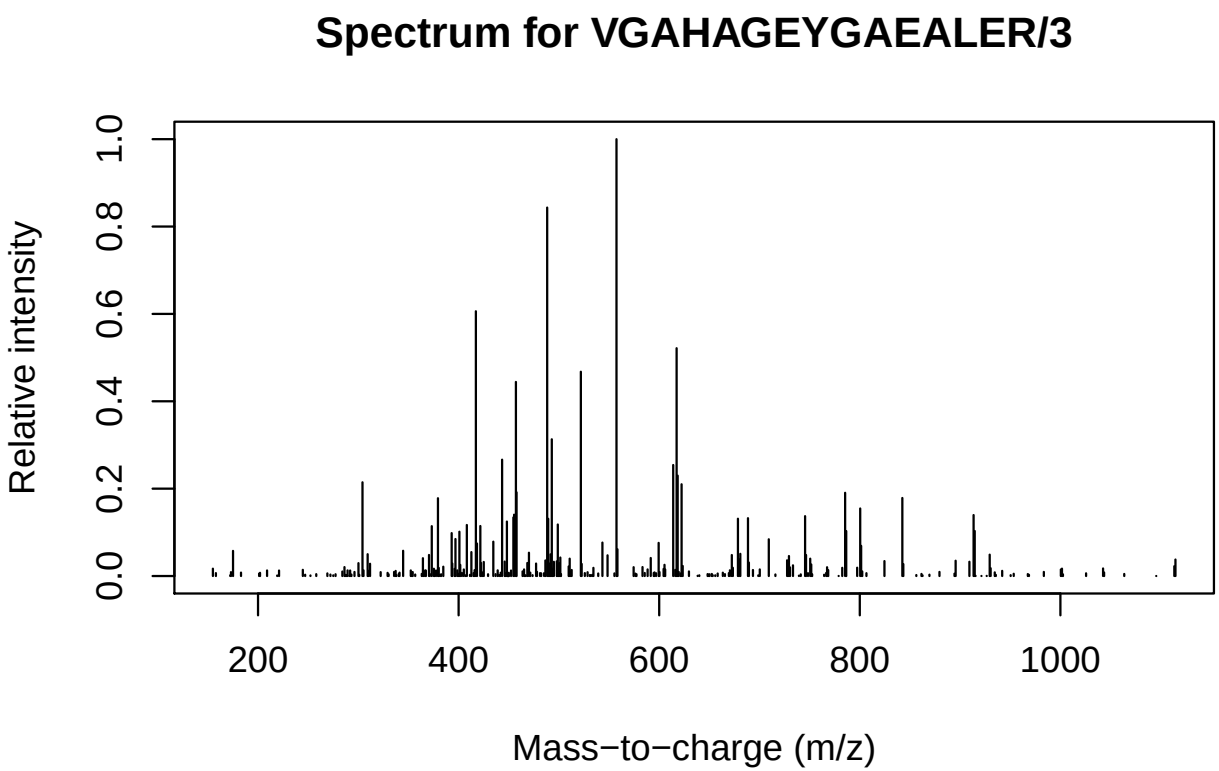
Diseases and alterations

- typical cancer cell carry alterations in up to hundreds of genes
- knowledge of mutation profile helps us to understand which biological processes are altered and select therapy accordingly
- alteration screening is—in high-throughput manner—done at nucleic acid level by SNP chips and NGS sequencing
- our interest: utilization of mass spectrometry for mutation screening

Outline

MS^2 spectrum

The task: Given MS^2 spectrum, determine the molecule, which produced it.[1]



[1] MS^2 spectrum shows fragment ion abundance and depends on fragmentation method.

Peptide database search

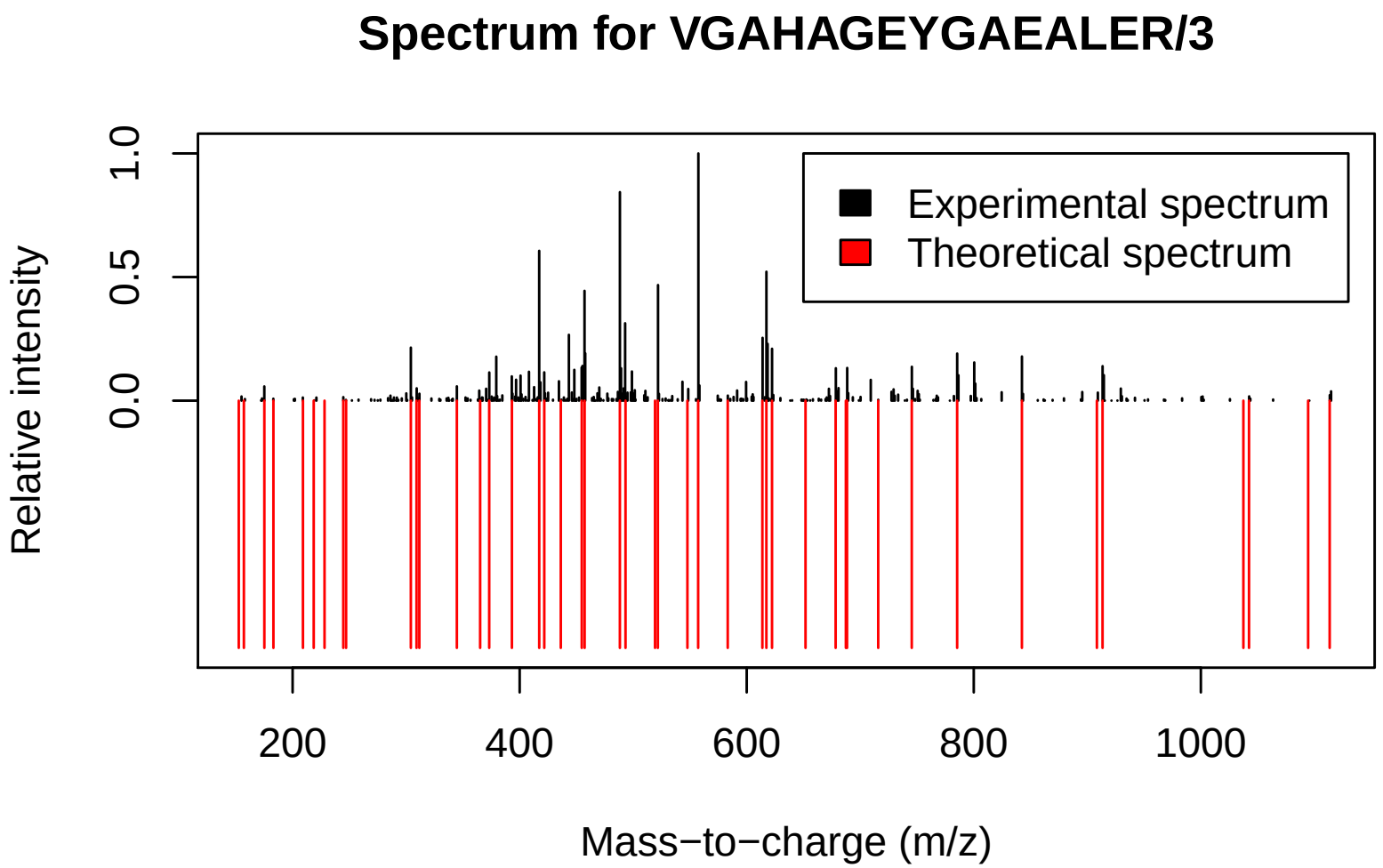
1. load protein sequences
2. create theoretical spectrum for candidate peptides[2]
3. evaluate similarity between theoretical and experimental spectrum

Advantages & Disadvantages

+ straightforward to use with any set of proteins
− does not take naturally into account intensity of peaks

[2] After proteolytic digestion.

Experimental and theoretical spectrum



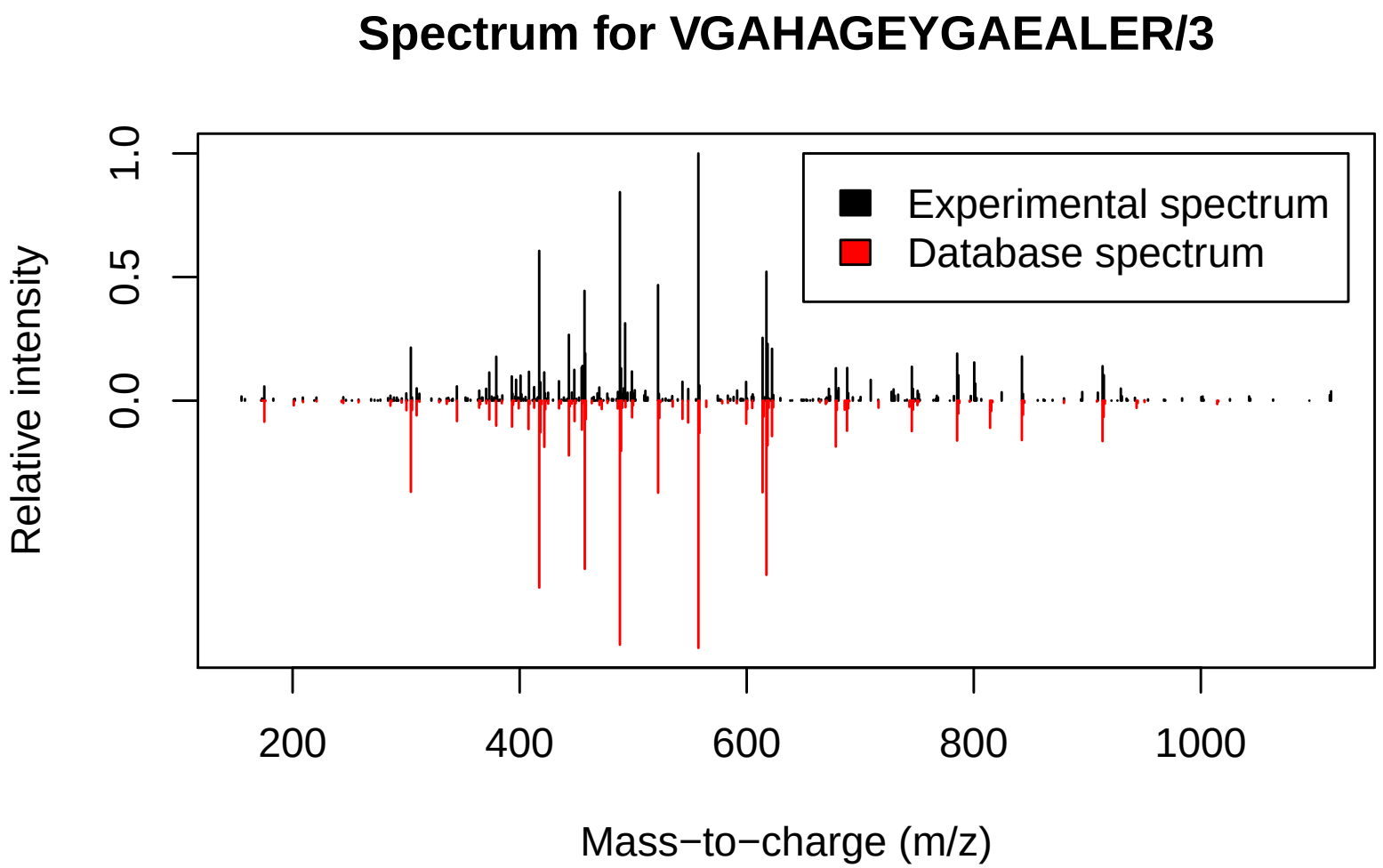
Spectral database search

1. load database of confirmed peptide spectra
2. evaluate similarity between experimental and database spectrum

Advantages & Disadvantages

+ naturally takes into consideration intensity of peaks
+ faster than peptide database search
− only known spectra

Experimental and database spectrum



De-novo sequencing approach

- start from observed peaks
- explain m/z differences for peaks

- complete fragmentation and data of very high quality essential
- used mainly for extraction of tags[3] from spectrum

Advantages & Disadvantages

+ in idealized form: database not needed
+ orthogonal approach with respect to database search
− incomplete fragmentation is very common

[3] Short, fixed-length chains of amino-acids.

Mutation identification methods for proteomics

Available methods:

- de-novo peptide tagging and peptide reconstruction[4]
- error-tolerant peptide database search[5]

Our method:

- peptide database search using recreated proteome[6]

Other possible methods:

- spectral database search with update of corresponding fragment ions[7]

[4] With or without reference database guidance.
[5] Available in MASCOT, X!Tandem.
[6] Actually peptidome.
[7] Potentially coupled with prediction of intensity update.

Outline

Dymka—reliable identification system

Motto: "Reliable identification of peptides from MS^2 spectra."

Integrated with:

- 5 peptide database search engines[8]
- 2 spectral database search engines[9]
- 3 de-novo systems[10]

Other properties:

- cluster-enabled, deployed at IMTM (250+ cores)
- statistical evaluation based on target-decoy approach.

[8] crux (Sequest), MASCOT, MyriMatch, OMSSA, X!Tandem
[9] Pepitome, SpectraST
[10] CompNovoCID, DirecTag, PepNovo

Rationale

- peptide identification systems use different algorithms of evaluation
- crucial property—evaluation of false discovery rate for search systems is possible
- addition of a search engine could not make things worse, i.e.: could not bias results—potential of algorithm for confident identification is evaluated using target-decoy approach

Target-decoy approach

- for use with database systems
- search engines are given decoyed databases
- databases consist of two equal-sized parts
  - target—what we are searching for
  - decoy—what, we know, is not in the analyzed sample
- then each match to decoy part is incorrect
- each score, say s, is associated with q-value
  - the proportion of decoy matches with score $\geq s$

Example of conflicting information

- reliability of match could be established and the conflicting information can be analyzed

Example of conflicting information

- consider a candidate peptide for a spectrum

scan number	peptide	charge	MZ	RT
12311	ALGFENATQALGR	2	674.8461	3192.8735

- its scores and associated q-values across search engines

	SpectraST	Pepitome	MyriMatch	OMSSA	X!Tandem	crux	Mascot
score	0.683	148.642	21.521	NA	NA	723.587	19.42
q-value	0.0	0.0	0.7139	NA	NA	0.0	0.02877

search engine	q-value	interpretation
crux, Pepitome, SpectraST	≤ 0.01	confident match
MASCOT, MyriMatch	> 0.01	non-confident match
OMSSA, X!Tandem	NA	no report for match

Outline

Recreating proteome/peptidome

- to use peptide database search for identification of mutations, we need to generate proteome
- we are not interested in completely mutated proteins, but in a series of proteins as a result of various combinations of alterations
- proteins are not identified as a whole—they're inferred from identified peptides[11]
- we do not have to generate variously altered proteome, which becomes infeasible[12]
- we are actually interested in altered peptidome

[11] By means of proteolytic digests.

[12] It can be considered infeasible when number of alterations $\gtrsim 20$. It is common to have ≥ 50 alterations per protein.

Naïve, combinatorial algorithm

- main use—just for clarification what needs to be solved

Algorithm 1 Naïve enumeration algorithm

```
1: procedure NAÏVE-ENUMERATE(alts, mRNA)
2:     combs ← COMBINATIONS(alts)
3:     for c ∈ combs do
4:         protein ← TRANSLATE(UPDATE(mRNA, c))
5:         peptides ← DIGEST(protein)
6:         APPEND-OUTPUT(peptides)
7:     end for
8: end procedure
```

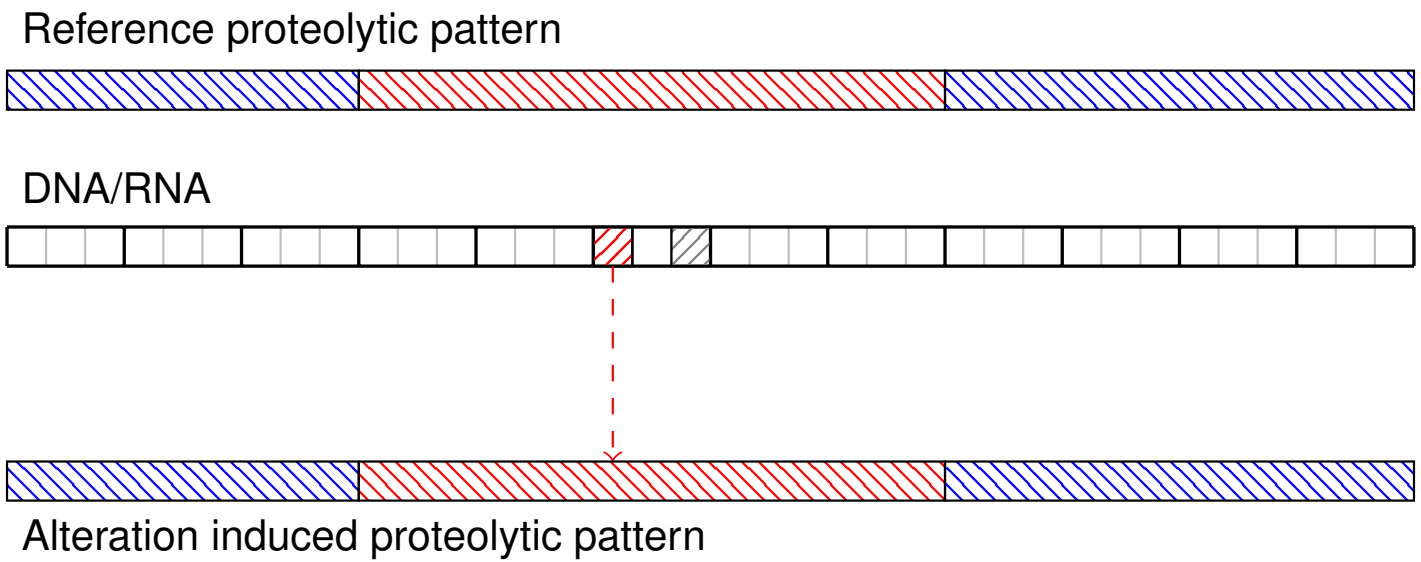
- as was said in previous slide, this algorithm becomes infeasible quickly

Mutation induced difference in pattern

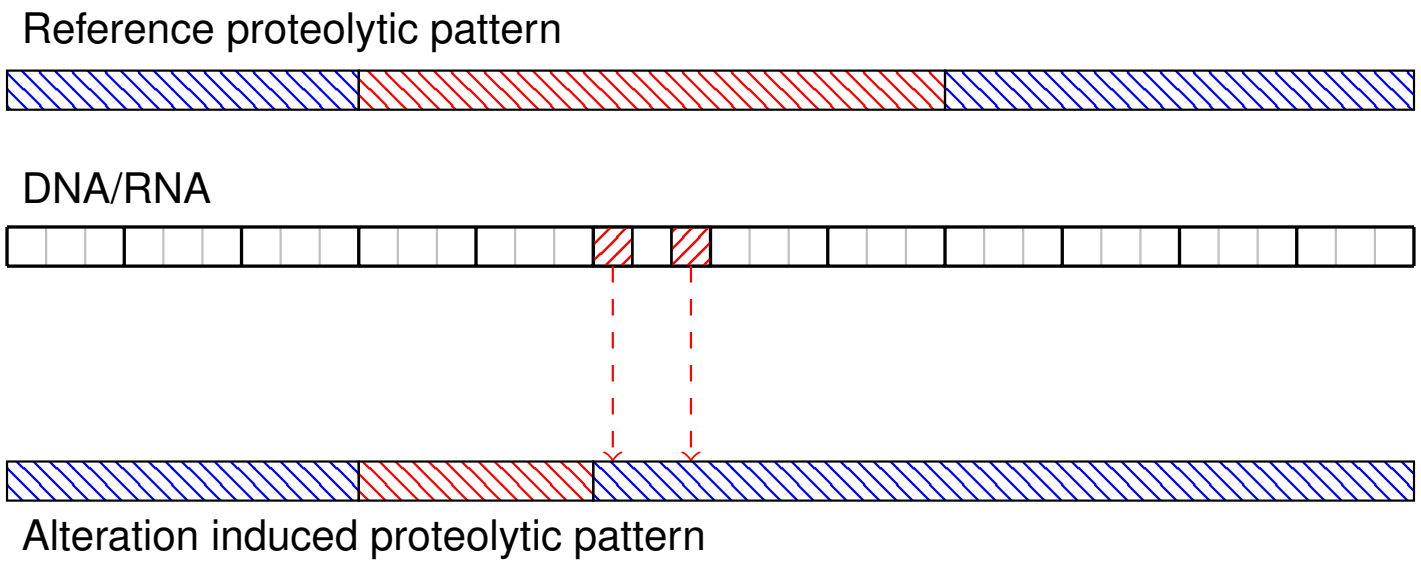
Reference proteolytic pattern

DNA/RNA

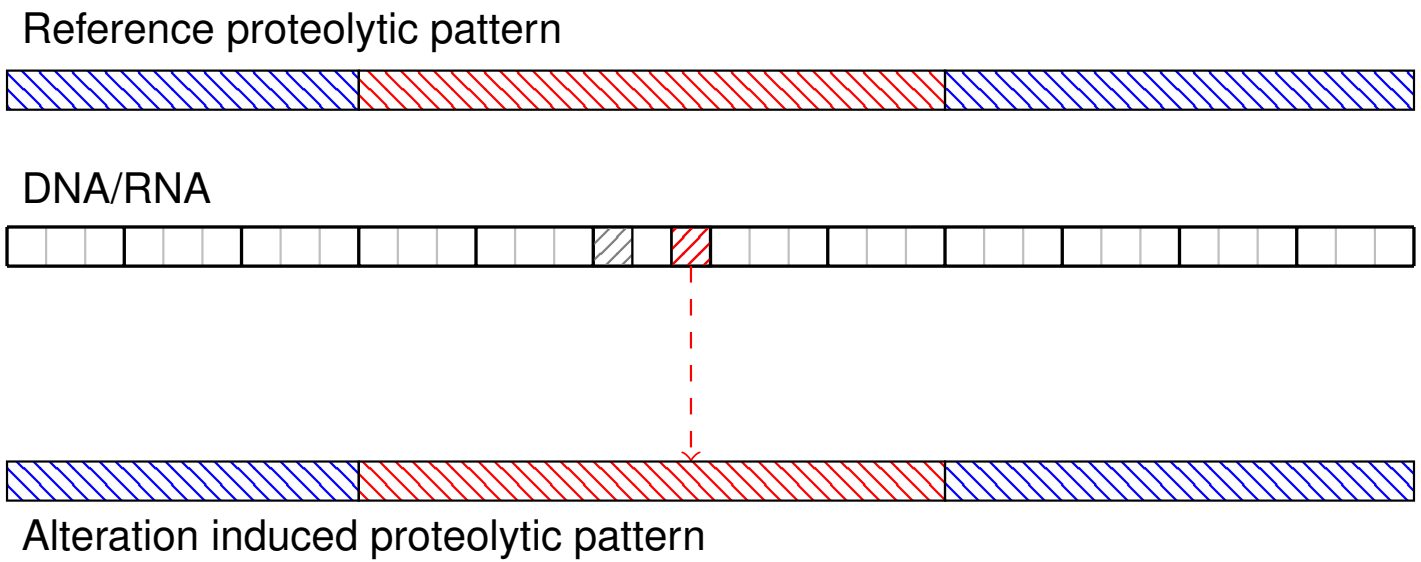
Mutation induced difference in pattern



Mutation induced difference in pattern



Mutation induced difference in pattern



Enumeration algorithm

Definition

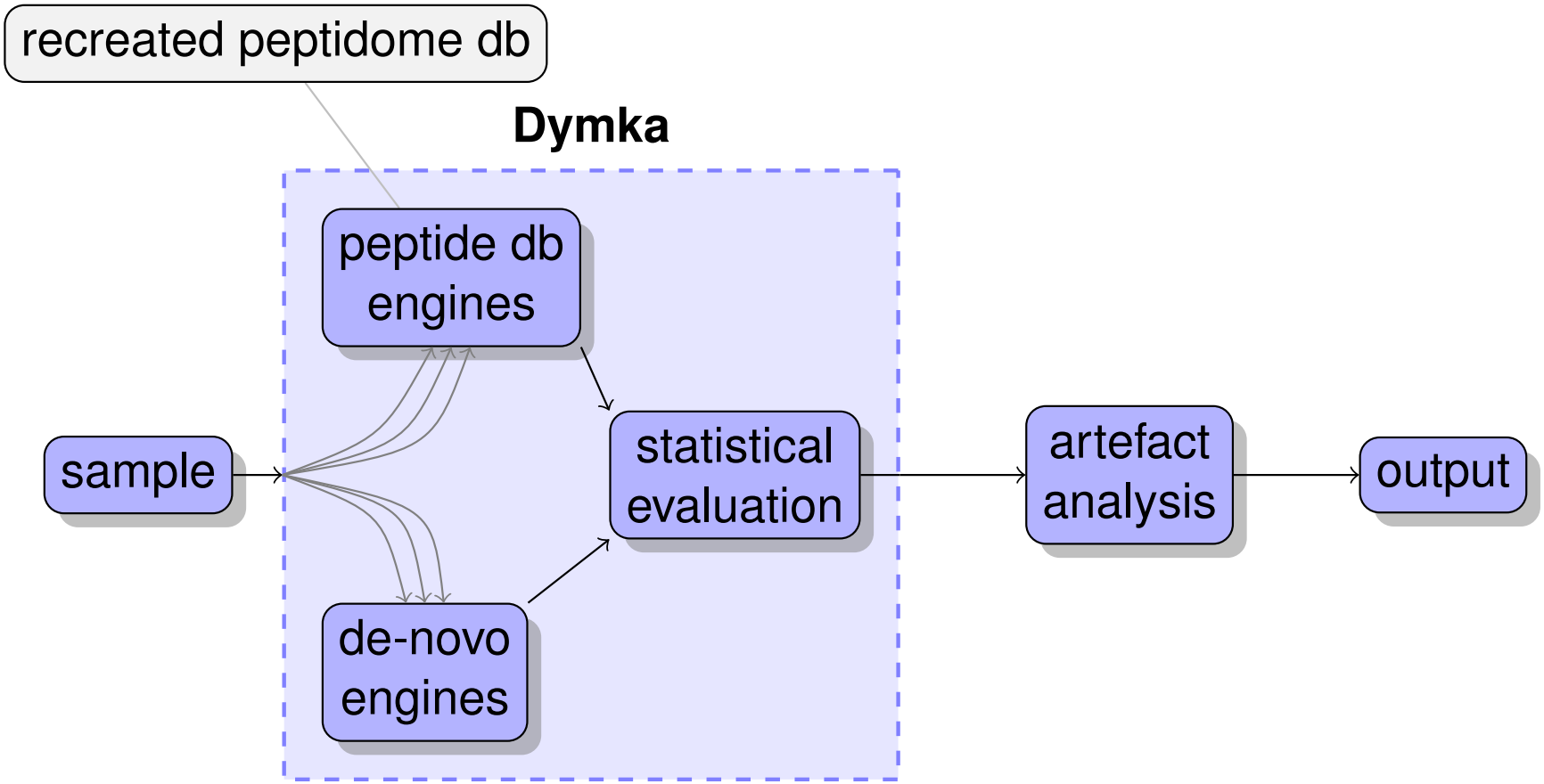
Any sequence of alterations which applied to given mRNA changes proteolytic digest pattern when translated is called Proteolytic-Digest Difference Introducer, *shortened as* PDDI.

Algorithm's main steps:

1. identification of relevant PDDIs—these change digest
2. for each combination of non-overlapping PDDIs: digestion of protein into peptides
3. then just combinations over alterations in scope of peptide—because digest pattern remains the same

Outline

System overview



Peptide Alteration Cracker

- identification of altered peptides using:
  - peptide database search and generated peptidome
  - de-novo approach and peptide reconstruction
- generation of peptidome based on user-provided genomic alterations
  - support for multiple formats—vcf, COSMIC, ICGC, raw csv
  - automatic detection of coordinate system, strand information inference[13]
  - support for different protein models[14]
- encapsulated in web interface

[13] Done by searching for maximum correlation of reference nucleotides (from alterations source) with genome.

[14] Currently, only ENSEMBL protein models are available.

Artefact analysis

Incomplete fragmentation artefacts:

- fragmentation prior to MS^2 is often incomplete process $\Longrightarrow$ subchain of peptide can have no support from fragment ions
- however, the altered part of peptide should be supported by fragment ions to establish presence of alteration

Other artefacts:

- mass(alt. AA) $\approx$ mass(ref. AA)[15]
- mass(alt. AA + variable PTM[16]) $\approx$ mass(ref. AA)

[15] Leucine/isoleucine as an example.
[16] Post-translational modification.

Transcriptomics–proteomics experiment

Experiment:

- transcriptome sequencing and mass spectrometry of proteome performed at IMTM
- cancer cell-line HCT116

Expectations:

- mass spectrometry is less sensitive than NGS—thus we would expect to identify higher ratio of more abundant alterations

Results from experiment

- the table sums up the behavior with different thresholds of number of reads of alterations

Number of reads	≥ 500	≥ 1000	≥ 2000	≥ 4000
Alterations	1239	580	245	153
Identified ($q \leq 0.1$)	100	93	91	58
Ratio	8.07 %	16 %	37.14 %	37.9 %
Identified ($q \leq 0.01$)	61	56	54	42
Ratio	4.92 %	9.65 %	22.04 %	27.45 %

- we can identify about 20–30% of high-abundant alterations sequenced on the transcriptomics level

Conclusion

- mass spectrometry can be, in limited way, used for screening of high-abundant alterations
- mass spectrometers are continuously improving, so it is expected that their sensitivity will be higher as time progresses
- one advantage over genomic/transcriptomic sequencing is the ability to observe post-translational modifications

Thank you for your attention.