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

1. Introduction
   - Motivation
   - Peptide identification

2. Mutation identification
   - Dymka
   - Enumeration Algorithm
   - Peptide Alteration Cracker
Introduction

- Motivation
- Peptide identification

Mutation identification

- Dymka
- Enumeration Algorithm
- Peptide Alteration Cracker
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

1. Introduction
   - Motivation
   - Peptide identification

2. Mutation identification
   - Dymka
   - Enumeration Algorithm
   - Peptide Alteration Cracker

ENBIK 2014
The task: Given MS² spectrum, determine the molecule, which produced it.¹

¹MS² spectrum shows fragment ion abundance and depends on fragmentation method.
Peptide database search

1. load protein sequences
2. create theoretical spectrum for candidate peptides
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

Spectrum for VGAHAGEYGAEALER/3

Mass-to-charge (m/z)
Relative intensity

Experimental spectrum
Theoretical spectrum
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

Spectrum for VGAHAGEYGAEEALER/3

Mass-to-charge (m/z)
Relative intensity

- Experimental spectrum
- Database spectrum

ENBIK 2014
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

1. Introduction
   - Motivation
   - Peptide identification

2. Mutation identification
   - Dymka
   - Enumeration Algorithm
   - Peptide Alteration Cracker
Dymka—reliable identification system

Motto: “Reliable identification of peptides from MS\textsuperscript{2} spectra.”

Integrated with:

- 5 peptide database search engines\textsuperscript{8}
- 2 spectral database search engines\textsuperscript{9}
- 3 de-novo systems\textsuperscript{10}

Other properties:

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

\textsuperscript{8}crux (Sequest), MASCOT, MyriMatch, OMSSA, X!Tandem
\textsuperscript{9}Pepitome, SpectraST
\textsuperscript{10}CompNovoCID, DirecTag, PepNovo
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

- consider a candidate peptide for a spectrum

<table>
<thead>
<tr>
<th>scan number</th>
<th>peptide</th>
<th>charge</th>
<th>MZ</th>
<th>RT</th>
</tr>
</thead>
<tbody>
<tr>
<td>12311</td>
<td>ALGFENATQALGR</td>
<td>2</td>
<td>674.8461</td>
<td>3192.8735</td>
</tr>
</tbody>
</table>

- its scores and associated q-values across search engines

<table>
<thead>
<tr>
<th>search engine</th>
<th>q-value</th>
<th>interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>crux, Pepitome, SpectraST</td>
<td>( \leq 0.01 )</td>
<td>confident match</td>
</tr>
<tr>
<td>MASCOT, MyriMatch</td>
<td>( &gt; 0.01 )</td>
<td>non-confident match</td>
</tr>
<tr>
<td>OMSSA, X!Tandem</td>
<td>NA</td>
<td>no report for match</td>
</tr>
</tbody>
</table>
Outline

1. Introduction
   - Motivation
   - Peptide identification

2. Mutation identification
   - Dymka
   - Enumeration Algorithm
   - Peptide Alteration Cracker
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
- we do not have to generate variously altered proteome, which becomes infeasible
- 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 $\geq 50$ alterations per protein.
Algorithm 1 Naïve enumeration algorithm

1: procedure NAÏVE-ENUMERATE($alts$, mRNA)
2: \hspace{1em} combs $\leftarrow$ COMBINATIONS($alts$)
3: \hspace{1em} for $c \in$ combs do
4: \hspace{2em} protein $\leftarrow$ TRANSLATE(UPDATE(mRNA, $c$))
5: \hspace{2em} peptides $\leftarrow$ DIGEST(protein)
6: \hspace{2em} APPEND-OUTPUT(peptides)
7: \hspace{1em} 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

ENBIK 2014
Mutation induced difference in pattern

Reference proteolytic pattern

DNA/RNA

Alteration induced proteolytic pattern
Mutation induced difference in pattern

Reference proteolytic pattern

DNA/RNA

Alteration induced proteolytic pattern
Mutation induced difference in pattern

Reference proteolytic pattern

DNA/RNA

Alteration induced proteolytic pattern
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

1. Introduction
   - Motivation
   - Peptide identification

2. Mutation identification
   - Dymka
   - Enumeration Algorithm
   - Peptide Alteration Cracker
System overview

- Sample
- Peptide DB engines
- De-novo engines
- Statistical evaluation
- Artefact analysis
- Output

Dymka

Recreated peptidome DB
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
  \[\Rightarrow\] 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:
- \(\text{mass(alt. AA)} \approx \text{mass(ref. AA)}\)\(^{15}\)
- \(\text{mass(alt. AA + variable PTM)} \approx \text{mass(ref. AA)}\)

\(^{15}\)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**

<table>
<thead>
<tr>
<th>Number of reads</th>
<th>$\geq$ 500</th>
<th>$\geq$ 1000</th>
<th>$\geq$ 2000</th>
<th>$\geq$ 4000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alterations</td>
<td>1239</td>
<td>580</td>
<td>245</td>
<td>153</td>
</tr>
<tr>
<td>Identified ($q \leq 0.1$)</td>
<td>100</td>
<td>93</td>
<td>91</td>
<td>58</td>
</tr>
<tr>
<td>Ratio</td>
<td>8.07 %</td>
<td>16 %</td>
<td>37.14 %</td>
<td>37.9 %</td>
</tr>
<tr>
<td>Identified ($q \leq 0.01$)</td>
<td>61</td>
<td>56</td>
<td>54</td>
<td>42</td>
</tr>
<tr>
<td>Ratio</td>
<td>4.92 %</td>
<td>9.65 %</td>
<td>22.04 %</td>
<td>27.45 %</td>
</tr>
</tbody>
</table>

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