Identification of mutations in protein-coding DNA sequences

by mass-spectrometry of proteome

Miroslav Hruška

INSTITUTE OF MOLECULAR AND TRANSLATIONAL MEDICINE

Outline

Outline

Study of proteins

What is proteomics?

Large-scale study of proteins.

proteins	large macromolecules[1] performing variety of biological functions
peptides	macromolecules of the same kind as proteins, but significantly shorter and usually without specific biological function

[1] Specifically, a long chains of amino-acid residues.

Bottom-up proteomics

1. proteins are separated using biochemical methods from sample of interest
2. proteins are digested using protease to peptides[2]
3. mass spectra of peptides are measured
4. **peptide identification is performed**
5. proteins are assembled from identified peptides

[2] This is because of identification using mass-spectrometry. Proteins are in general too large for mass-spectrometry.

Mass-spectrometry

- destructive analytical chemistry technique for identification of analytes
- molecules are ionized—based on acquired charge, they have specific mass-to-charge ratio (mz)[3]

Fundamental ability of mass-spectrometer

- isolation of charged ion with specific mz from a pool of molecules

[3] For charge 1, this essentialy means mass, however with mass of the charge-giving particle, i.e. a proton.

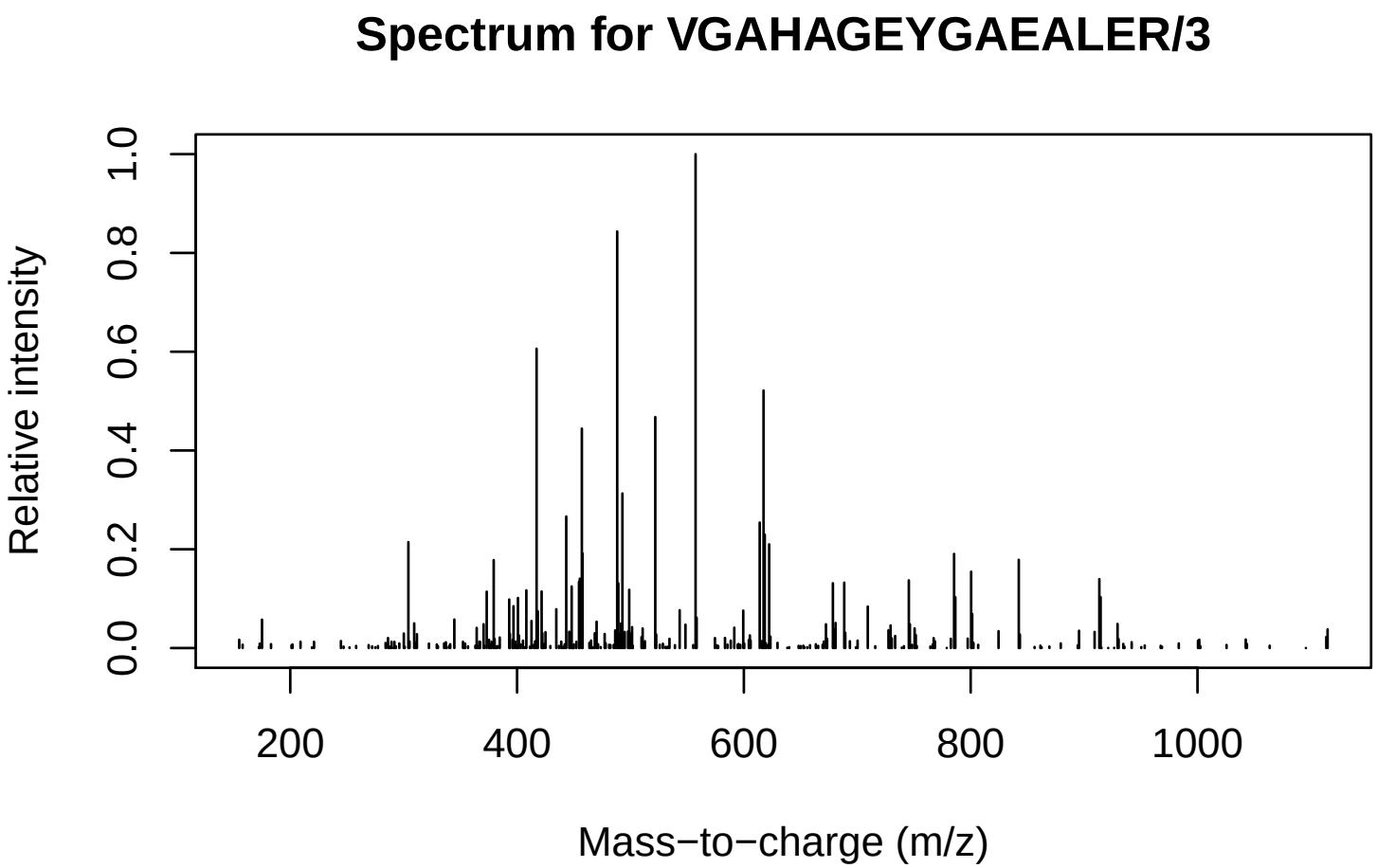
Tandem mass spectrometry for peptide identification

One step in mass-spectrometer cycle:

1. molecules are entering the mass-spectrometer for some short amount of time[4] and are ionized
2. mass spectrum of these molecules is created[5]
3. candidate mz's representing peptides are selected[6]
4. selected molecules then undergo fragmentation
5. the mass spectrum after fragmentation is created[7]

[4] Usually at most hundreds of milliseconds.
[5] So-called MS^1 spectrum.
[6] By means of isotopic envelope.
[7] So-called MS^2 spectrum.

Example: MS^2 spectrum



Outline

Outline

Peptide identification task

Notation

$\mathbb{S}$	set of all spectra; $\mathbb{S} \equiv \langle 0, 1 \rangle^{\mathbb{R}^+}$
$\mathbb{AA}^C$	set of coded amino-acids; $\mathbb{AA}^C = \{A, C, D, E, \dots\}$
$\mathbb{P}_e$	set of peptides; $\langle a_1, \dots, a_n \rangle \in \mathbb{P}_e, n \geqslant 1, a_i \in \mathbb{AA}^C, i \in \{1, \dots, n\}$
$m(p)$	mass of peptide p $\in \mathbb{P}_e$

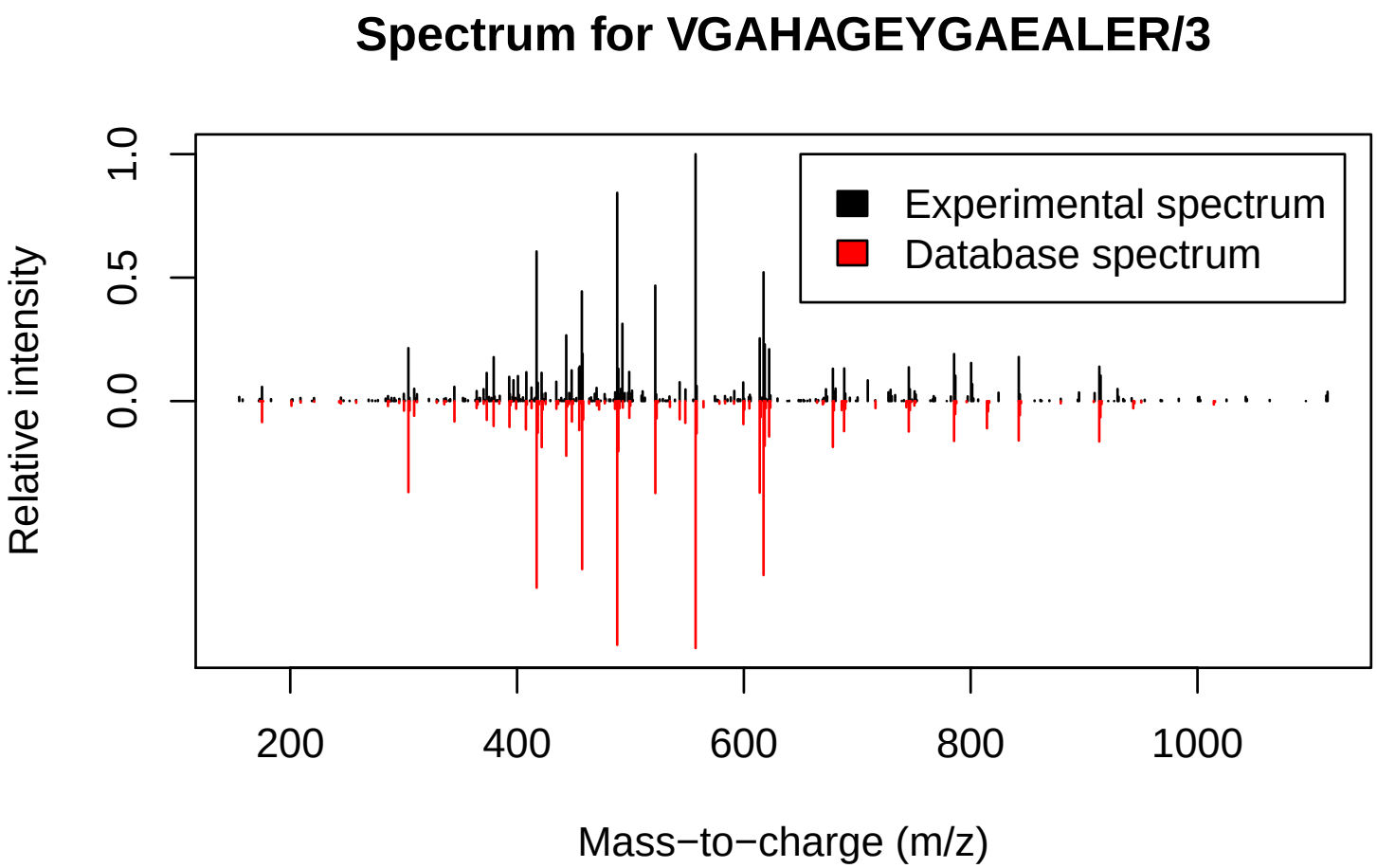
- suppose $\psi : \mathbb{P}_e \to \mathbb{S}$ is a function representing the fragmentation of peptide and construction of MS^2 spectrum from molecular fragments
- the identification task is the reversed process—given MS^2 spectrum $s \in \mathbb{S}$ obtain $p \in \mathbb{P}_e$ which produced it

Identification methods

- database methods
  - spectral database search
  - peptide database search[8]
- de novo methods
  - peptide tagging[9]
  - de novo peptide reconstruction

[8] Theoretical spectrum database search.
[9] Partial identification.

Spectral database search



General fragmentation model Ψ

Let peptide $p = \langle a_1, \ldots, a_n \rangle \in \mathbb{P}_e$, consider $\Psi^+ : \mathbb{P}_e \to \mathbb{S}$, where

$$\Psi^+(p)\Big(\psi_j\big(\sum_{i=1}^{k} m(a_i)\big)\Big) = 1$$

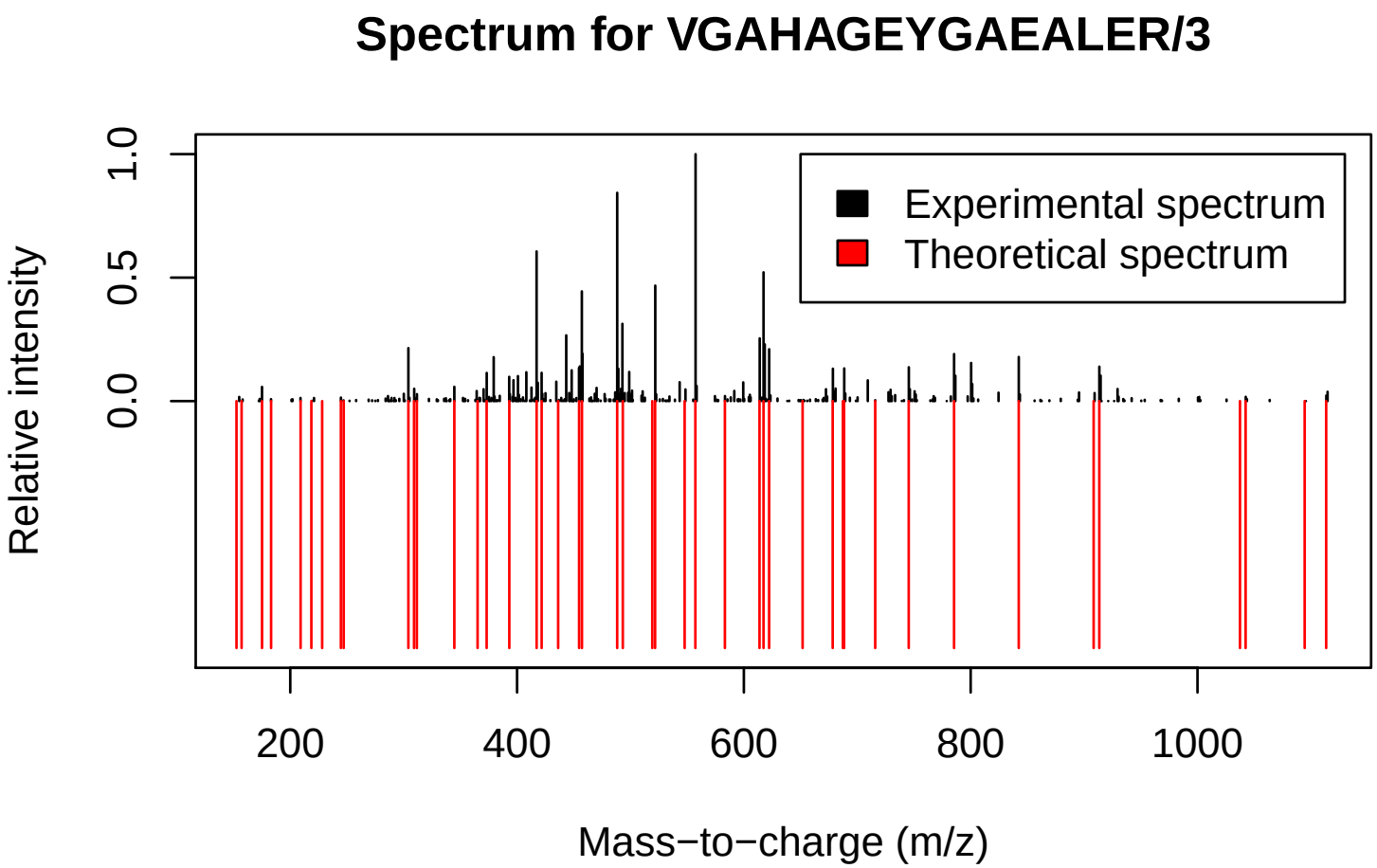
$$k \in \{1, \ldots, n\}, j \in \{1, \ldots, m\}$$

and 0 everywhere else.
Each $\psi_j : \mathbb{R}^+ \to \mathbb{R}^+, j \in \{1, \ldots, m\}$ is of following form:

$$\psi_j(r) = \frac{r + q + zp}{z}, q \in \mathbb{R}, z \in \mathbb{N}, p \approx 1.007276 \text{ (Da)}$$

ψ_j are functions mapping the mass of peptide into mass-to-charge ratio of specific type of peptide fragment (q) at given charge (z).

Peptide database search



De-novo peptide reconstruction

Have $s \in \mathbb{S}$ with peaks $q = \{a \in \mathbb{R}^+ \mid s(a) > 0\}$. Using fragmentation model Ψ with set ψ_j, $j \in \{1, \ldots, l\}$ of mass–to–fragment-mz functions obtain set of candidate masses q_m:

$$q_m = \bigcup_{j=1}^{l} \psi_j^{-1}(q)$$

Construct directed graph $G = \langle q_m, E \rangle$, where

$$E = \left\{ \langle a, b \rangle \in q_m \times q_m \mid (\exists x \in \mathbb{AA}^C)\ m(b) - m(a) \approx_\epsilon m(x) \right\}$$

Then return highest scoring path starting at zero and ending at mass of the molecule prior to fragmentation.

Outline

Fragmentation considerations

- molecular fragmentation ψ is not well understood
- we are not able to tell ψ is injective
- in practice the situation is more complicated because of noise peaks when performing the identification task

Definition

$\Phi : \mathbb{S} \times \mathbb{S} \to \langle 0, 1 \rangle$ is a spectra similarity measure, if following holds:

$$\Phi(x, y) = \Phi(y, x)$$
$$\Phi(x, y) = 1 \iff x = y$$

Peptide homeometricity

Definition

Let $p, q \in \mathbb{P}_e$, ϕ a spectra similarity measure and ψ a fragmentation function. Then call p, q ϕ_t-homeometric if

$$\phi(\psi(p), \psi(q)) \geqslant t, t \in \langle 0, 1 \rangle.$$

Definition

Let $p, q \in \mathbb{P}_e$ and $\Psi : \mathbb{P}_e \to \mathbb{S}$, ϕ a spectra similarity measure. Then call p, q Ψ-model-ϕ_t-homeometric if

$$\phi(\Psi(p), \Psi(q)) \geqslant t, t \in \langle 0, 1 \rangle.$$

We would like to have approximately this behaviour:

- p, q ϕ_t-homeometric $\iff$ p, q Ψ-model-ϕ_t-homeometric

Ψ_{b_1} fragmentation model

- Ψ_{b_1} considers only so-called b fragments and only charge 1
- the only mass-to-fragment-mz function is $\psi_{b_1}(r) = r - b + p$
- the inverse fragment-mz-to-mass function is $\theta_{b_1}(s) = s + b - p$

$$\Psi_{b_1}(\langle a_1, \ldots, a_n \rangle)\left(\psi_{b_1}\left(\sum_{i=1}^{k} m(a_i)\right)\right) = 1$$

$$k \in \{1, \ldots, n\}$$

Properties of Ψ_{b_1}

- non-injectivity of Ψ_{b_1} follows from existence of $a, b \in \mathbb{AA}^C, a \neq b$ with $m(a) = m(b)$[10]
- thus there are $p, q \in \mathbb{P}_e$ which are Ψ_{b_1}-model-ϕ_1-homeometric for any ϕ

Equivalence relation on $\mathbb{P}_e$

Let $\theta \subseteq \mathbb{P}_e \times \mathbb{P}_e$, $\langle p, q \rangle \in \theta \iff \Psi_{b_1}(p) = \Psi_{b_1}(q)$. Directly by definition, θ is an equivalence relation.

Thus it is meaningful to consider the identification task as a function: $\Psi_{b_1}(\mathbb{P}_e) \to \mathbb{P}_e/\theta$

[10] Leucine and Isoleucine are coded amino-acids, that are molecular isomers; having the same chemical formula, but different structure.

Ψ_{b_1}-model-homeometric peptides enumeration

So having a similarity measure ϕ, the set of Ψ_{b_1}-model-ϕ_t-homeometric peptides for p whose mass differ[11] at most ϵ is

$$H_\epsilon(p) = \{q \in \mathbb{P}_e \mid \phi(\Psi_{b_1}(p), \Psi_{b_1}(q)) \geqslant t \text{ and } m(p) \approx_\epsilon m(q)\}$$

We will approach the enumeration of $H_\epsilon(p)$.

[11] The mass difference condition is because the mz of molecule is measured before fragmentation and mass can be deduced.

Properties of peptides with given mass

Define $f_\epsilon(x)$ as a function which returns all peptides with mass equal to x (up to ϵ).

$$f_\epsilon(x) = \left\{ \langle a_1, \ldots, a_n \rangle \in \mathbb{P}_e \mid x \approx_\epsilon m(\langle a_1, \ldots, a_n \rangle) \right\}$$

Note that for any $y = f_\epsilon(x)$ if $a \in y$ then any permutation b of a is in y, so $b \in y$; which follows from commutativity of addition.

$$H_\epsilon(p) \subseteq f_\epsilon(m(p))$$

Related problem

The homeometricity peptides enumeration problem is related to the following problem:

Inputs

Let $x \in \mathbb{R}^+$ be desired value, $\epsilon \geqslant 0$ a tolerance, finite set of atoms $\mathbb{A} = \left\{a_i \in \mathbb{R}^+ \mid i \in \{1, \ldots, k\}\right\}$, $k \geqslant 1$ and finite set of "checkpoints" $\mathbb{C} = \left\{c_i \in \mathbb{R}^+ \mid i \in \{1, \ldots, l\}\right\}$, $l \geqslant 1$.

What we are interested is following:

Output

Obtain sequences of atoms that sum up to desired value x (up to ϵ), and for each checkpoint c, there is some prefix subsequence which sums up to c (up to ϵ).

Problem decomposition [1/2]

Formally:

$$g_\epsilon^{\mathbb{C}}(x) = \Big\{ \langle a_1, \ldots, a_n \rangle \in \mathbb{A}^n \mid \sum_{i=1}^{n} a_i \approx_\epsilon x \text{ and } (\forall c \in \mathbb{C})\ (\exists j \in \{1, \ldots, n\})\ m(\langle a_1, \ldots, a_j \rangle) \approx_\epsilon c \Big\}$$

We can decompose the problem and consider summing up to each "checkpoint" separately. This is obvious for $\epsilon = 0$, however there is a subtle change when $\epsilon > 0$.
Consider having ordered elements of $\mathbb{C}$ as $c_1 \leqslant c_2 \leqslant \cdots \leqslant c_l$ and define:

$$x_1 = c_2 - c_1$$

$$\vdots$$

$$x_{l-1} = c_l - c_{l-1}$$

$$x_l = x - c_l$$

Problem decomposition [2/2]

$$\begin{aligned}
\mathbb{X}_1 &= f_\epsilon(x_1) \\
\mathbb{X}_2 &= f_{2\epsilon}(x_2) \\
\mathbb{X}_3 &= f_{2\epsilon}(x_3) \\
&\vdots \\
\mathbb{X}_l &= f_{2\epsilon}(x_l)
\end{aligned}$$

Then the union of solutions obtained by concatenating candidate subsolutions contains $g_\epsilon^{\mathbb{C}}(x)$.

$$\bigcup \left\{ a_1 \oplus \cdots \oplus a_n \mid (\forall i \in \{1, \ldots, l\})\ a_i \in \mathbb{X}_i \right\} \supseteq g_\epsilon^{\mathbb{C}}(x)$$

Application for homeometric peptides

Have $\langle a_1, \ldots, a_n \rangle = p \in \mathbb{P}_e$. Observe that p has n peaks in Ψ_{b_1}.

$$q = \{a \in \mathbb{R}^+ \mid \Psi_{b_1}(a) > 0\}$$

$$x = m(p)$$
$$\mathbb{A} = m(\mathbb{A}\mathbb{A}^C)$$
$$\mathbb{C}_q = \theta_{b_1}(q)$$

Pick number $m \leqslant n$ as the desired least amount of intersecting peaks. Let $\mathbb{C} \subseteq \mathbb{C}_q, |\mathbb{C}| = m$. Then each solution to previously addressed problem (for a given $x, \mathbb{A}, \mathbb{C}$) will contain at least m intersecting peaks (up to ϵ).

Outline

Motivation

- typical cancer cell carry mutations in up to hundreds of genes
- early diagnostics of potential disease-relevant information
- knowledge of mutation profile helps in selection of therapy[12]

[12] There are well-known cases where mutations are the reason why patients do not respond to drug treatment.

Outline

Database construction

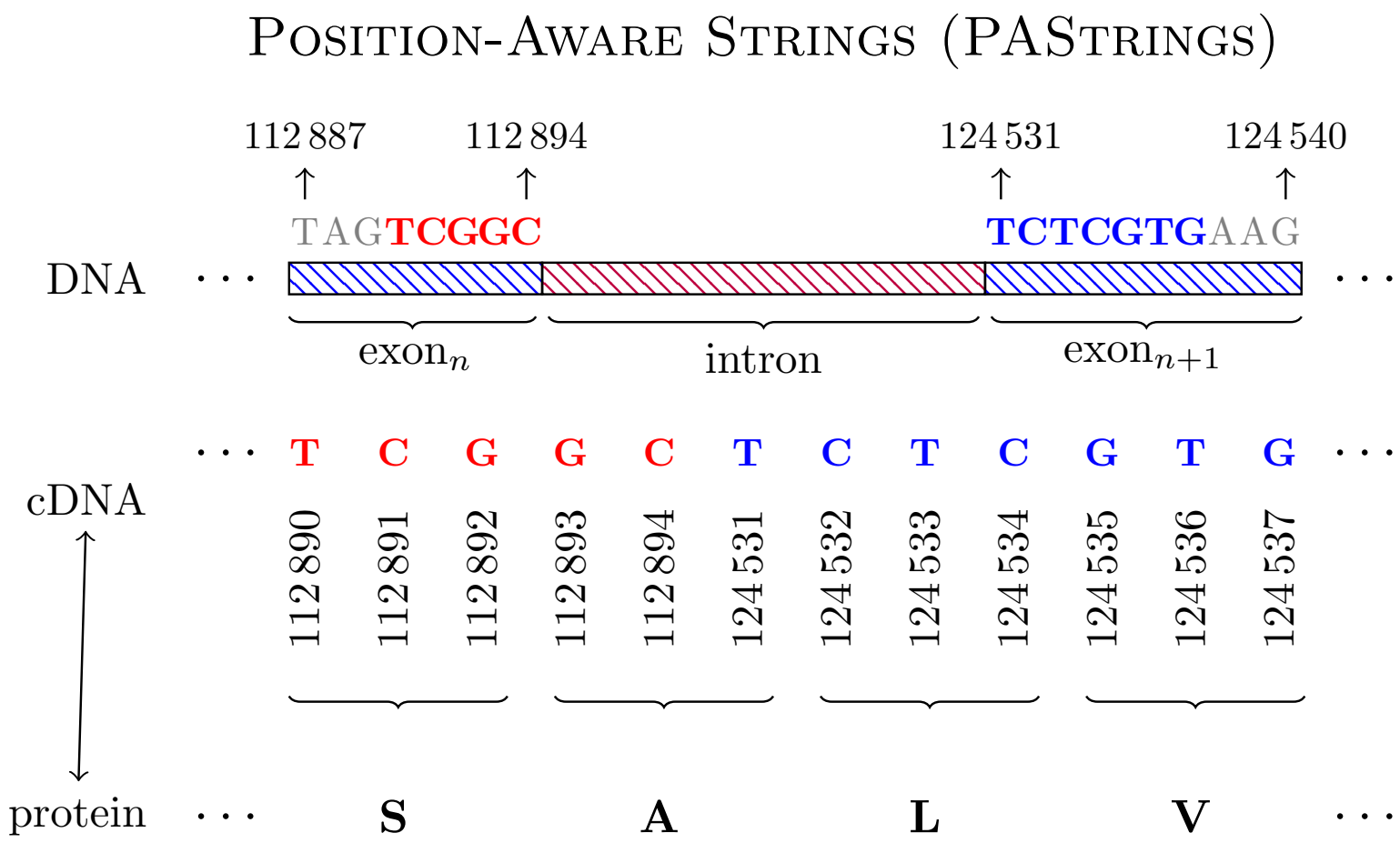
1. obtain DNA sequences
2. transcribe to RNA and obtain protein-coding sequences by cutting out non-coding subsequences $\Longrightarrow$ obtain mRNA
3. update sequences by DNA/RNA alterations from known, disease-relevant sources
4. translate altered mRNA to proteins
5. digest to peptides

mRNA translation

- biological process which synthesizes proteins from mRNA
- the translation machinery maps triplets of RNA bases to one amino-acid, the mapping is dictated by so-called genetic code[13]
- the genetic code could be thought of as function $\{A, C, G, U\}^3 \rightarrow \mathbb{AA}^C$ and it is non-injective, surjective mapping
- denote Ω function that maps sequences of RNA bases to peptides (the mapping is induced by genetic code)

[13] Genetic code is highly similar between organisms.

Position-Aware strings



Simplified example of database record

Peptide:	AAIEQSMK
Protein:	ENSP00000377197
Protein position:	2098
Protein reference:	V (Valine)
Protein altered:	M (Methionine)
Peptide position:	6
Chromosome:	16
Chromosome position:	70,989,298
Chromosome reference:	G (Guanine)
Chromosome altered:	A (Adenine)

Enumeration of peptides

Algorithm 1 Naïve enumeration pseudo-algorithm

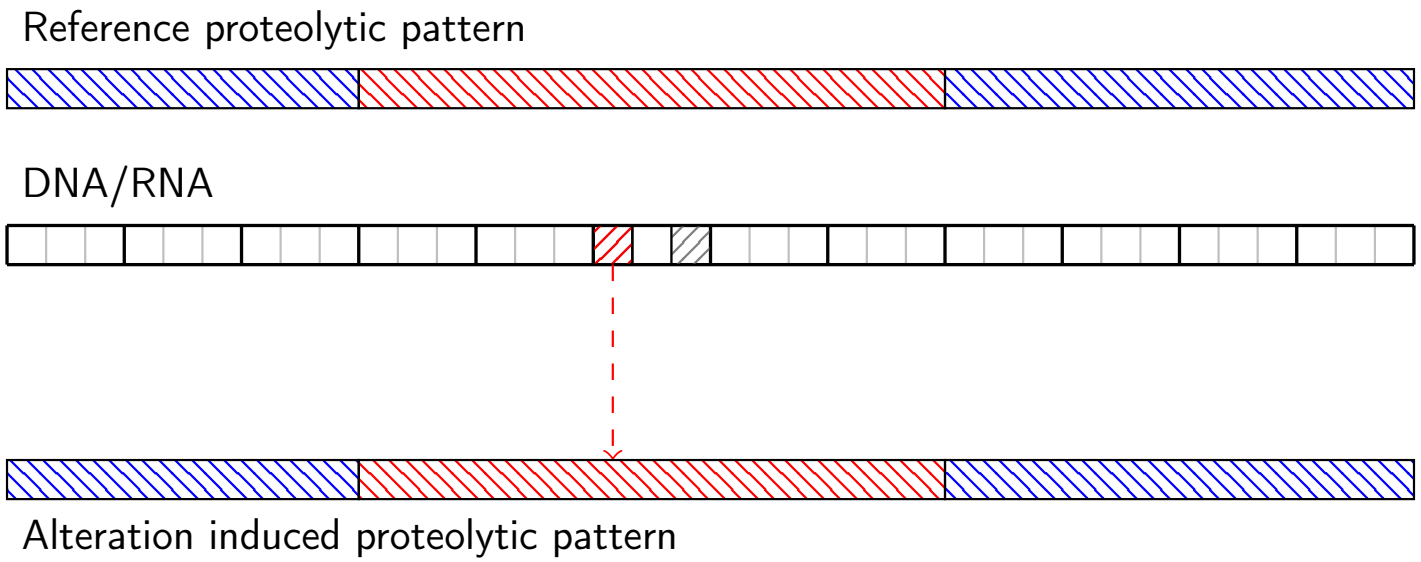
```
1: procedure naïve-enumerate(alts, mRNA)
2:     combs ← Combinations(alts)
3:     for c ∈ combs do
4:         protein ← Translate(Update(mRNA, c))
5:         peptides ← Digest(protein)
6:         Append-Output(peptides)
7:     end for
8: end procedure
```

Mutation induced difference in pattern

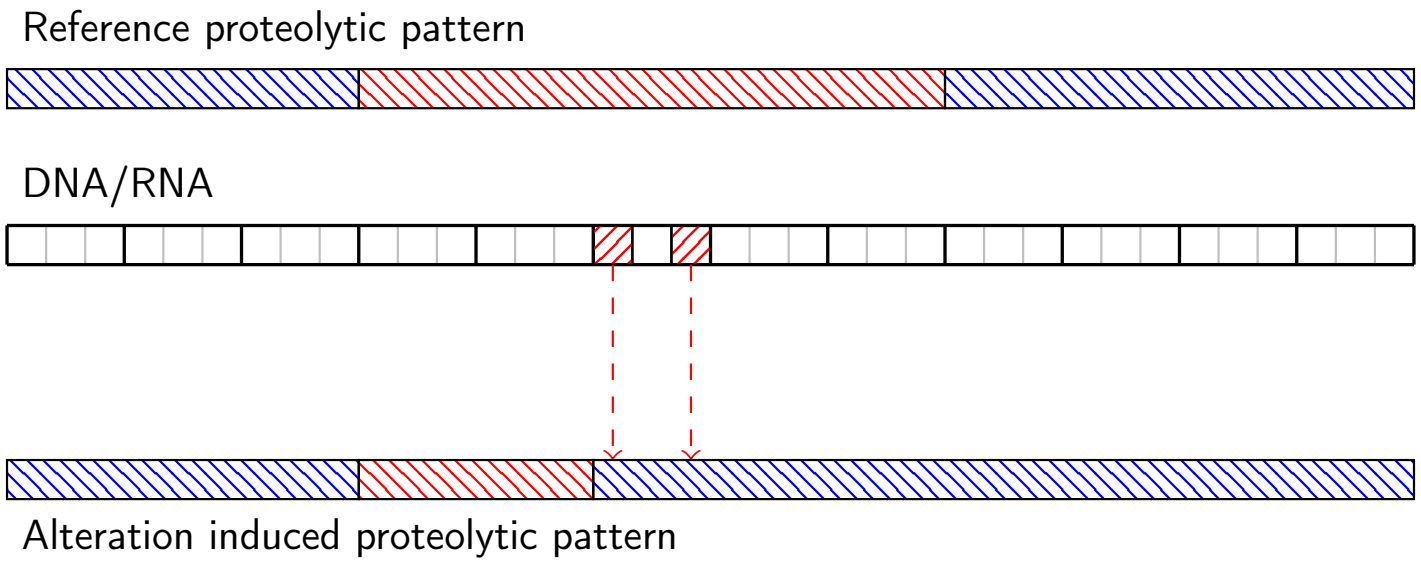
Reference proteolytic pattern

DNA/RNA

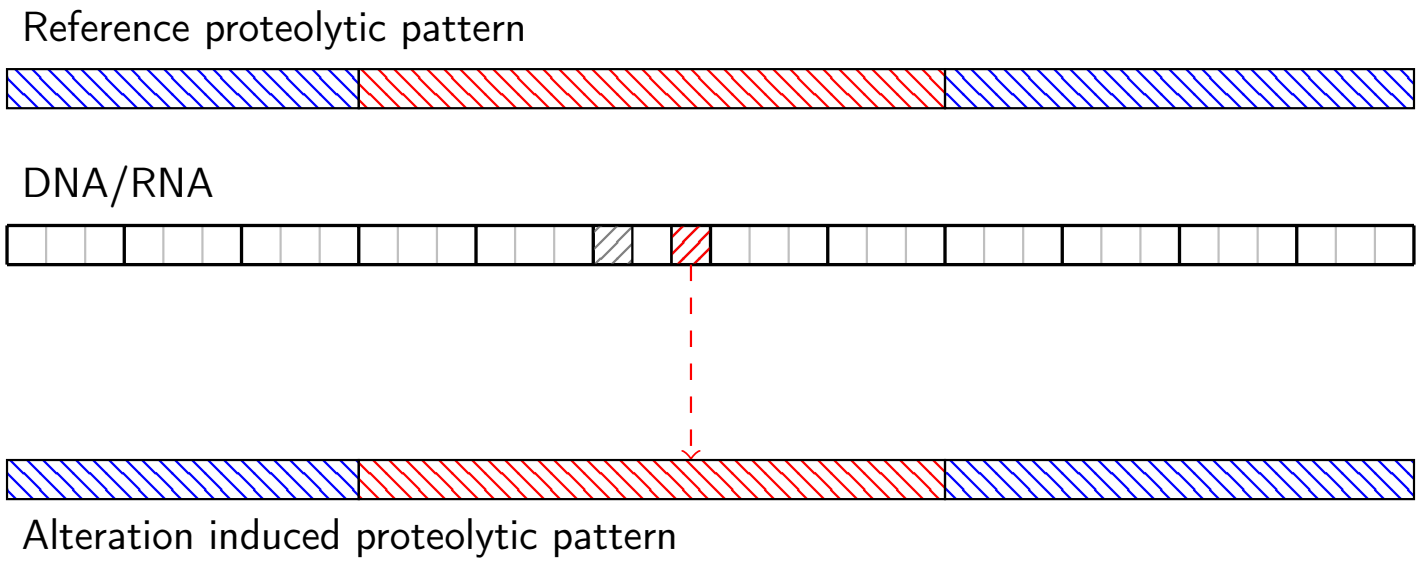
Mutation induced difference in pattern



Mutation induced difference in pattern



Mutation induced difference in pattern



PDDI algorithm

Definition

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

Outline

Dymka—identification system

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

Integrated with:

- 5 peptide database search engines[14]
- 2 spectral database search engines[15]
- 3 de-novo systems[16]

Other properties:

- cluster-powered, deployed at IMTM (250+ cores)
- statistical evaluation based on target-decoy approach[17]

[14] crux (Sequest), MASCOT, MyriMatch, OMSSA, X!Tandem
[15] Pepitome, SpectraST
[16] CompNovoCID, DirecTag, PepNovo
[17] This is not true anymore.

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 there
- assumption—incorrect target match is equally likely as match to decoy database
- then each match to decoy part is incorrect
- each score, say s, is associated with q-value
  - the proportion of decoy matches with score $\geqslant s$

Example of conflicting information

- 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	$\leqslant 0.01$	confident match
MASCOT, MyriMatch	> 0.01	non-confident match
OMSSA, X!Tandem	NA	no report for match

Outline

Observations regarding the identification

- search engines do not address the homeometricity problem
  $\Longrightarrow$ even high-scoring matches are incorrect
- this problem does not show up so often for reference peptides[18] because in majority of cases their presence is more likely than their non-reference homeometric cognates
- this is in direct contrast with mutant peptides which often have homeometric peptide among reference peptides[19]

Outcome

Only mutant peptides with unlikely interpretation by homeometric peptides are selected.

[18] This is also probably the reason why it was not studied in detail.

[19] Mainly post-translationally modified reference peptides.

Observations regarding the peptide origin

- suppose $p \in \mathbb{P}_e$ is correctly identified reference peptide and there is only one reference mRNA sequence r, such that $\Omega(r) = p$

warning this doesn't neccessarily mean that p originated from r

- it could happen that p originated from other "reference" mRNA, which was adequately mutated
- we use Occam's razor principle

Outcome

Especially, in identification of non-reference peptides we're interested in those that originated from unique (non-reference) mRNA.

Results

- the system was recently validated on cancer cell line HCT116
- both RNA and proteins were separated from the sample
  - RNA underwent sequencing
  - peptide spectra were measured using mass-spectrometry
- the system was used to deduce DNA/RNA alterations and these were compared to alterations obtained by RNA sequencing[20]
- without artefact analysis—enormous amount of false positives[21]
- 73 alterations were identified, of which 13 were cancer-related

[20] The comparison is not as trivial as it may seem.

[21] Alteration which was not found using RNA sequencing.

Conclusions & Future work

Conclusions

- system capable of reliable identification of small mutations using mass-spectrometry was developed
- treatment of homeometricity was shown to be important to remove one class of artefacts

Future work

- construction of spectras for mutant peptides from spectral databases
- extension of system to work reliably with large mutations and splice site alterations

Thank you for your attention!