MutationDetector – software tool for detecting single amino acids substitutions

Brilliantov Kirill

Abstract

Proteins play an essential role in our lives, because they provide structure to cells. If any disruption occurs, a protein will cease acting properly and may cause severe diseases.

Many factors can result in such disruptions. We consider the most important of them – a single nucleotide polymorphism (SNP).

The goal of this work is to develop a software tool for detecting and localizing single amino acid substitutions.

Since three consequent nucleotides, together forming a *codon*, encode an amino acid, SNP can lead to an amino acid substitution, thereby implying a change of the mass of the protein.

Post-translational modifications (PTMs) of the amino acids can also change the protein mass. For example, the mass of methionine increases by approximately 16Da upon oxidation.

The software tool named MutationDetector accepts as input: a wild-type sequence, the difference between its mass and that of a putative variant peptide and an error tolerance threshold. In the output the sequence fragments which might incorporate an appropriate amino acid substitution or a PTM, appear highlighted. Also, this tool has some useful features such as drawn lines between codons, one of the covered amino acid and another, symbolising the SNPs, which can cause the substitution.

In the future, we intend to extend functionality of MutationDetector in various ways thereby adapting it to solving special problems.

Word Count: *218*

Brilliantov Kirill

MutationDetector – software tool for detecting single amino acids substitutions

Research Report

Introduction

Proteins play an essential role in our lives, because they are regulating sub-cellar processes[1]. If any disruption occurs, a protein will cease acting properly and may cause severe diseases.

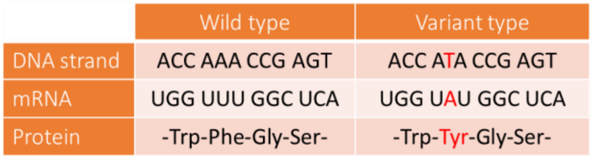
Many factors can result in such disruptions. We consider the most important of those – a single nucleotide polymorphism (SNP).

Having learned to find positions in protein sequence where the substitution might occur, we would get a possibility to identify the so-called variant proteins, which can be biomarkers for a variety of hard diseases[2].

Nowadays there are two opposite approaches to analyzing such proteins: quantitative and qualitative. The idea that the quantitative ratio between various proteins must be constant in healthy organism is foundation of the quantitative analysis. If any disruption in organism occurred, some ratio would be violated. The main aim of quantitative analysis is to identify such violations[2] and making conclusions based on them.

The primary interest of qualitive analysis is not a quantity but a qualitive precursor of protein (amino acid sequence).

Since three consequent nucleotides in a DNA strand, together forming a codon, encode an amino acid, SNP can lead to an amino acid substitution, thereby implying a change of the mass of the protein.

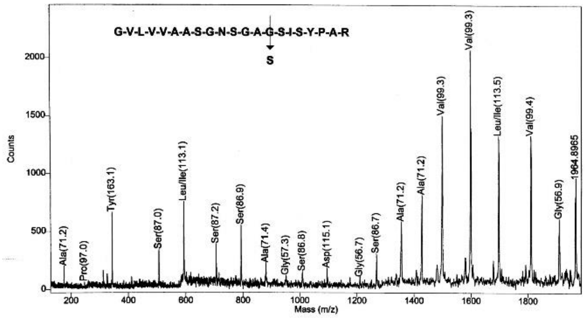


	Wild type	Variant type
DNA strand	ACC AAA CCG AGT	ACC ATA CCG AGT
mRNA	UGG UUU GGC UCA	UGG UAU GGC UCA
Protein	-Trp-Phe-Gly-Ser-	-Trp-Tyr-Gly-Ser-

Example of SNP

Post-translational modifications (PTMs) of the amino acids can also change the protein mass. PTM – is a change of amino acid's chemical composition through adherence of some chemical radicals. For example, the mass of methionine increases by approximately 16Da upon oxidation.

PTMs occur often than substitutions in nature, therefore if some PTM and some substitution correspond to present difference in mass, most probably the PTM occurred.



Example of mass spectrum

Our work concentrated on the qualitive analysis. The result of such analysis is a mass spectrum, taken from an investigated peptide. In the beginning, the dissolved peptide is being bombarded with charged particles, so ideally each molecule divides in two parts (a prefix and a suffix) and each of these parts is charged with one positive particle. Further, these fragments go through mass spectrometer (special device) and this device gives us the mass spectrum. Mass spectrum is a histogram, where X – axis is a mass of an ionized fragment, Y-axis is the ion current intensity (the quantity of registered particles with such mass). Accordingly, this graph is a set of peaks. Based on that set of peaks, we can establish certain fragment of investigated amino acid sequence, and based on that fragment, we can establish (with help of methods of biological alignment) the whole peptide, which is most similar to present, in another words we can establish the peptide which contain the established fragment and has the smallest difference in mass. It is possible, because there is a finite number of proteins in nature. In addition to this, we can establish the mass of the investigated protein and difference in mass between the variant peptide and the protein which existing in nature (wild-type peptide).

Analyzing this data manually is uncomfortable. Therefore, the idea about developing a software tool for analyzing such data arises.

The goal of my work was to develop a software tool for analyzes of the data, which was obtained from exploration of the modified peptide.

The main function of this programming interface is a handling the positions in the peptide, where a PTM or a substitution might occur.

This work differes from previous works concentrated on this theme, first, the result of this the work is the *graphical interface*, secondly, this interface takes as input a high-resolution data (data, which was taken from mass – spectrum).[4]

Methodology

For developing this application, I used Java and Swing, a library for developing graphical interfaces.

The input of this interface is a file with mass-spectrum, obtained from an investigated modified peptide.

There is a possibility of adding a big quantity of files with mass-spectrums and of searching for peptides through program's database. These functions are in "File" tab (fig. 1)

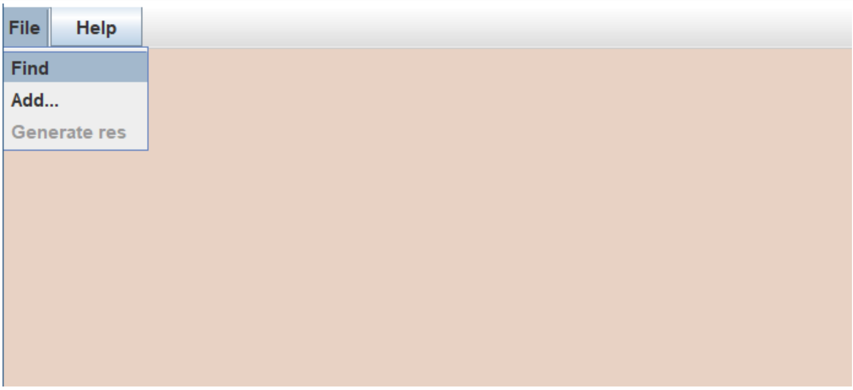


Fig. 1 pull-down menu

Also, there is an item named “Generate res”, it will subsequently be enabled (now it isn’t enabled). The dialog window appears by clicking the item “Find” (fig. 3.1)

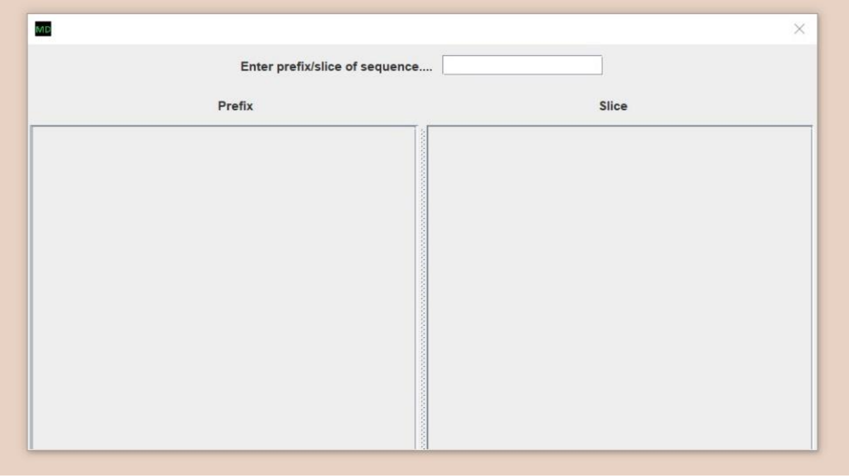
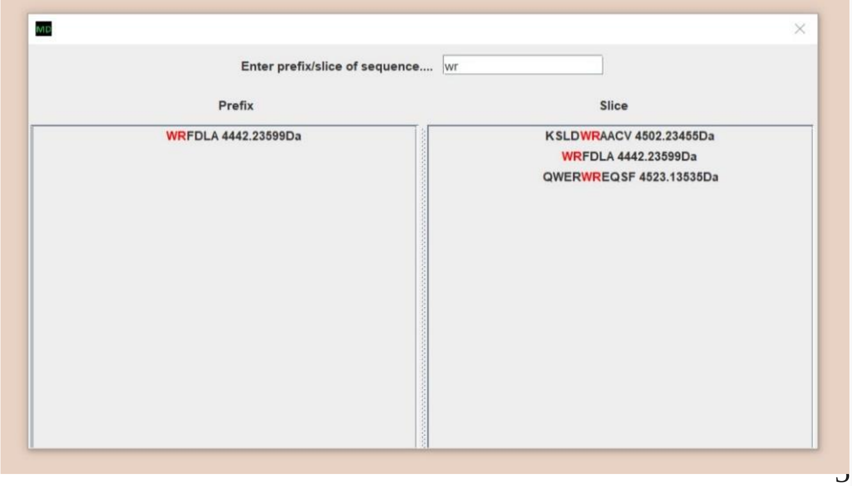


Fig 3.1 the empty dialog window

When a user inputs some sequence of amino acids into the input field, the peptides where this fragment is a prefix are appearing in the left part of this window, and the peptides which contain this fragment appeares in the right part of this window. (fig. 3.2).



Enter prefix/slice of sequence.... wr

```
WRFDLA 4442.23599Da
```

```
KSLDWRAACV 4502.23455Da
WRFDLA 4442.23599Da
QWERWREQSF 4523.13535Da
```

Fig 3.2 user has inputted “WR”

The main frame of present interfaces appears by selecting one of these peptides. In the top of this frame there is an amino acid sequence, just below there is a scrollable panel, where are the amino acids, from sequence, appears zoomed, and these amino acids which are visible on this panel now are red in the top sequence. To the right and to the left from the scrollable panel there are buttons “handle suffix” and “handle prefix” (fig. 4)

The buttons in the scrollable panel are get enabled when a user clicks on one of the buttons “handle prefix” or “handle suffix”. Then if a user clicks on some of amino acids (buttons in the scrollable panel), the program starts searching for positions in the suffix, which starts with pressed amino acid, or in the prefix, which ends with pressed amino acid, (it depends on which of prefix or suffix button user clicked before) where the PTM or the substitution fitting the present mass difference might occur. The program does not search through whole sequence, because mass-spectrum (source of data) is a set of prefixes and suffixes masses. When a user clicks on one of the amino acids, the following algorithm begins: first, program calculates the error for each position, because a mass-spectrometer is not ideal, so it has some mistake. Here are the formulas:

The mistake when the program searches through the prefix:

$$\varepsilon = \frac{M_i * ppm}{10^6}$$

M_i – mass of the prefix (the sum of all amino acids masses in this prefix) ends with i-position, i – is the number of the position.

ppm – parse par million, error tolerance.

The mistake when the program searches through suffix:

$\varepsilon = M_0 + \frac{M_i * ppm}{10^6}$, where M_i – mass of the prefix ends with i-position.

M_0 – mass of the whole peptide.

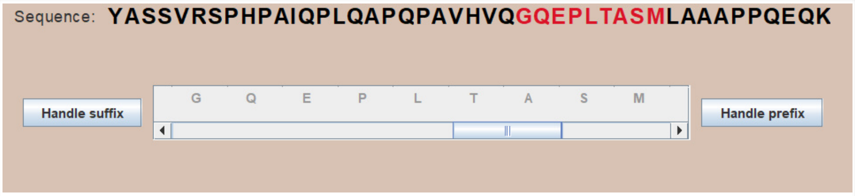

Sequence: YASSVRSPHPAIQPLQAPQPAVHVQGQEPLTASMLAAAPPQEQK

Fig. 4 sequence and scrollable panel

Then the program checks if there is some substitution or PTM with difference in mass which get into this interval: $(\Delta M - \varepsilon;\ \Delta M + \varepsilon)$ After the program checked for all accessible position whether there is a PTM or a substitution fitting the mass difference, the positions where the substitutions might occur are highlighted with blue color, where the PTM highlighted with orange color (fig. 5). The positions, which are not in the suffix or prefix under observation highlighted with pale color (the observed prefix ended with “S”)

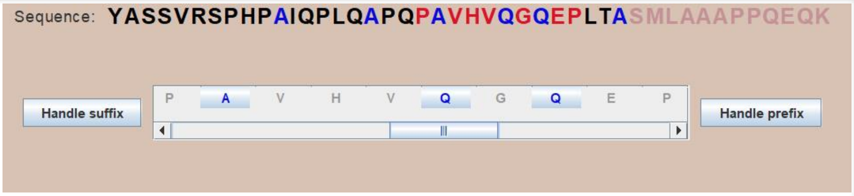
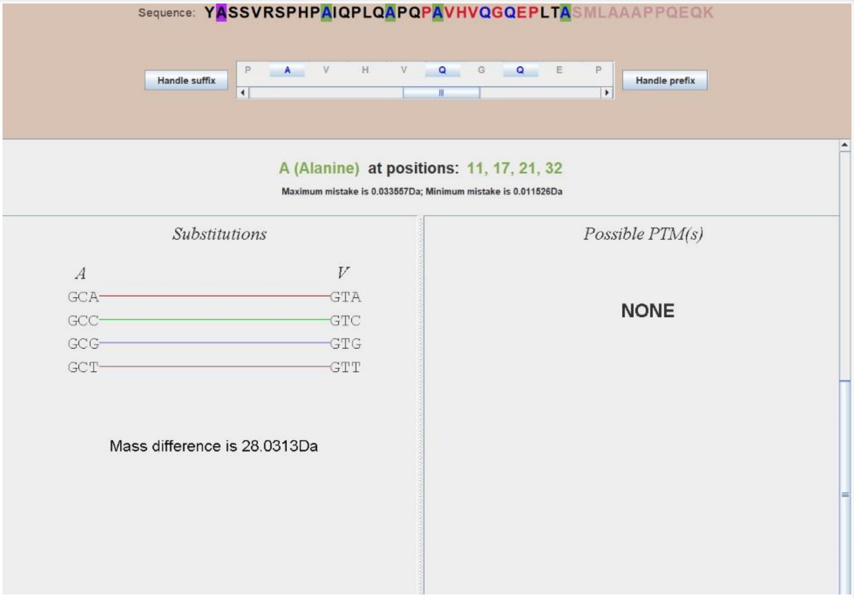

Sequence: YASSVRSPHPAIQPLQAPQPAVHVQGQEPLTASMLAAAPPQEQK

Fig. 5 highlighted positions

At this moment “Generate res” item gets enabled, it allows to generate a file loaded with strings like: ”at position 23 A>>V” which means that in some peptide at position 23 the amino acid V could substitute the amino acid A. Further, when user clicks on some of the enabled buttons (for example on of “A”s) the information about what could happen at this position appears (fig. 6).

When a user clicks on one of the “A”s in the top sequence, first of these “A”s appears on the purple background since the remaining appear on pale green background, because first “A”

does not fit the conditions since the mistake is not enough big (mass of prefix is not big enough).
The substitution which could occur in those positions is A >> V some of the codons (encoding these amino acids) are connected with colorful lines. They are connected because the only difference between them is one nucleotide. For example, codon GCA encode A, GTA encode V, the difference between them is that on the second position there are different nucleotides. If SNP in which nucleotide C substitutes with nucleotide T occurs, a substitution A>>V occurs.



Sequence: YASSVRSPHPAIQPLQAPQPAVHVQGQEPLTASMLAAAPPQEQK

A (Alanine) at positions: 11, 17, 21, 32

Maximum mistake is 0.033557Da; Minimum mistake is 0.011526Da

Substitutions

A	*V*
GCA	GTA
GCC	GTC
GCG	GTG
GCT	GTT

Mass difference is 28.0313Da

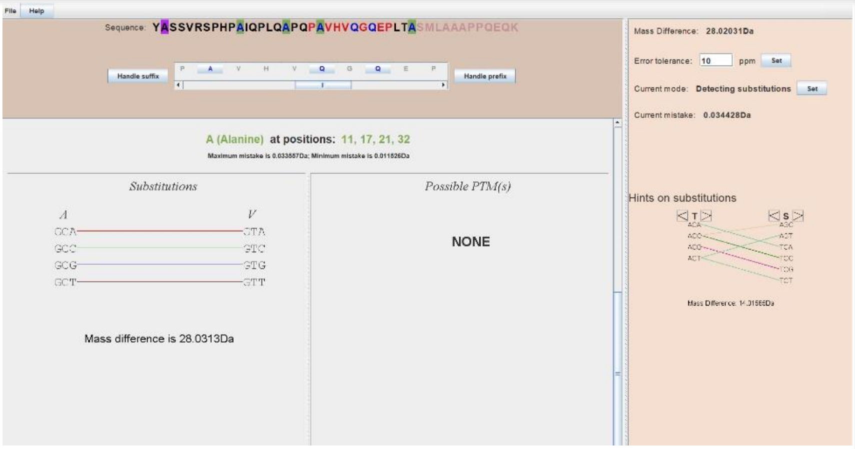
Possible PTM(s)

NONE

Fig. 6 information about the substitution

Results

As a result such interface had been developed (fig. 7):



Sequence: YASSVRSPHPAIQPLQAPQPAVHVQGQEPLTASMLAAAPPQEQK

A (Alanine) at positions: 11, 17, 21, 32

Maximum mistake is 0.033557Da; Minimum mistake is 0.011526Da

Possible PTM(s): NONE

Mass difference is 28.0313Da

Mass Difference: 28.02031Da

Error tolerance: 10 ppm

Current mode: Detecting substitutions

Current mistake: 0.034428Da

Mass Difference: 14.01565Da

Fig. 7 general view of the application

There are some additional possibilities for making work with this application more convenient besides the functionality described before.

- Tab “Help”, where user can see how to work with this application.
- Hot keys for some actions (there is a list of them in “Help”).
- Hints on substitutions. It located in the bottom right corner. User can choose two amino acids and see is there are any SNP which can lead to the substitution between these two amino acids.

To make sure that the developed software tool works well we did some tests

We took peptides which were investigated before (we took them from the database of the University where I was working)

1. YASSVRSPHPAIQPLQAPQPAVHVQGQEPLTASMLAAAPPQEQK

In this peptide A>>V and Q>>R occurred (in the prefix).

2. EAATQEDPEQVPELAAHEVSASEAEERPVAEEEILL

In this peptide A>>V occurred (in the suffix)

The developed program gave the right result in both cases.

Conclusion

During this work, the software tool for analyzing data obtained from modified peptide has been developed. It has been tested, it works correctly.

In the future, we intend to extend the functionality of MutationDetector in various ways thereby adapting it to solving special problems.

References

1. B. Lewin. *Cells*. БИНОМ Russia, 2011. 951 с.
2. S. Nie, H. Yin, Z. Tan, M. A. Anderson, M. T. Ruffin, D. M. Simeone, D. M. Lubman. *Quantitative Analysis of Single Amino Acid Variant Peptides Associated with Pancreatic Cancer in Serum by an Isobaric Labeling Quantitative Method*. J Proteome Res. 2014, 13(12):6058–6066.
3. K. Vyatkina, S. Wu, L. J. M. Dekker, M. M. VanDuijn, X. Liu, N. Tolic, M. Dvorkin, S. Alexandrova, T. M. Luider, L. Pasa-Tolic, P. A. Pevzner. *De Novo Sequencing of Peptides from Top-Down Tandem Mass Spectra*. J Proteome Res. 2015, 14(11):4450-4462.
4. Qisheng Peng, Zijian Wang, Donglin Wu, Xiaoou Li, Xiaofeng Liu, Wanchun Sun, Ning Liu. *Identification of single amino acid substitutions (SAAS) in neuraminidase from influenza a virus (H1N1) via mass spectrometry analysis coupled with de novo peptide sequencing.* Rapid Commun. Mass Spectrom. 2016