**MutationDetector – software tool for detecting single amino acids substitutions**

*Brilliantov Kirill*

**Abstract**

Proteins play an essential role in our lives, because they are providing 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 those – 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 all viewing 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: *221*

**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 . 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).

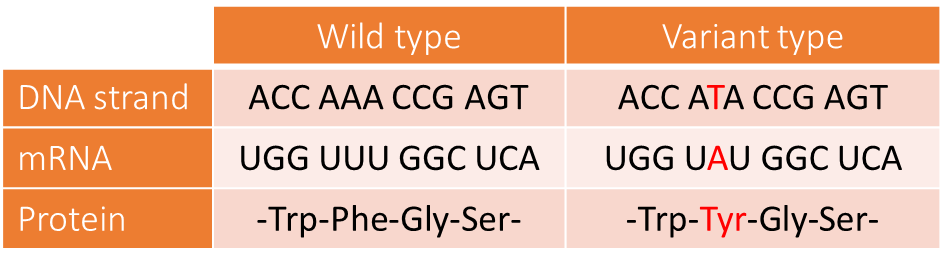
Having learned to find positions in protein sequence where the substitution might occur, we would get a possibility to identify so-called variant proteins, which can be biomarkers for a variety of hard .

Nowadays there are two opposite approaches to analyzing such proteins: quantitative and qualitative. Idea that the quantitative ratio between various proteins must be constant in healthy organism underlines the quantitative analysis. If any disruption in organism occurred, some ratio would violate. The main aim of quantitative analysis is identifying such and based on them making conclusions.

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

Since three consequent nucleotides in 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.

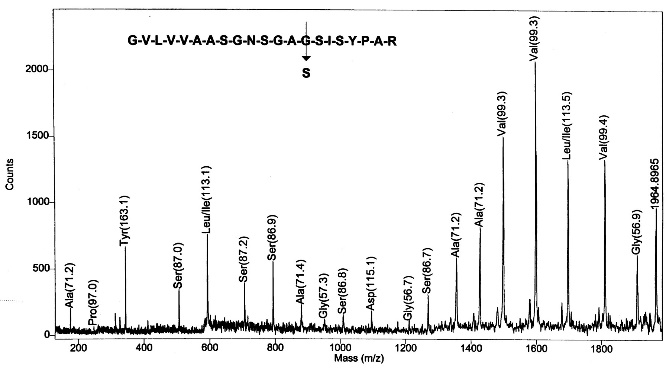
Example of SNP



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 are found more than substitutions in nature, therefore if some PTM and some substitution correspond to present difference in mass, most probably the PTM occurred.

Our work concerned with qualitive analysis. Result of such analysis is a mass spectrum, taken from investigated peptide. In the beginning the dissolved peptide is being bombarded with charged particles, so ideally each molecule divides on two parts (prefix and suffix) and the each of these parts is charged with one positive particle. Further these fragments go through mass spectra (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 ion current intensity (quantity of registered particles with such mass). Respectively, this graph is a set of peaks. On that set of peaks, we can establish certain fragment of investigated amino acid sequence, and on that fragment, we can establish (with help of methods of biological alignment) the whole peptide most similar to present. It is possible, because there are 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 variant peptide and protein which existing in nature.



Example of mass spectrum

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

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

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

Present work is really differing from previous works concerned with this theme, firstly, that during the work the *graphical interface* had been developed, secondly, that the input data is high-resolution data

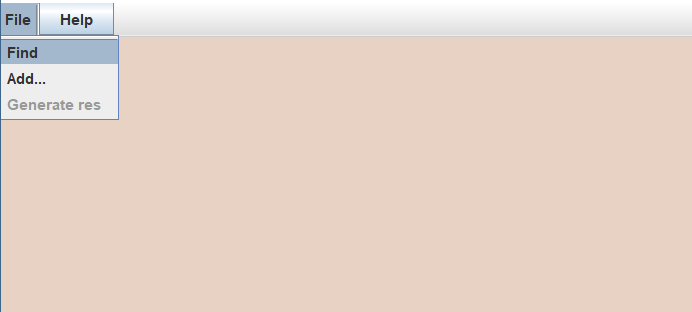
**Methodology**

The programming language which was used is Java and the library Swing for creating graphics interfaces.

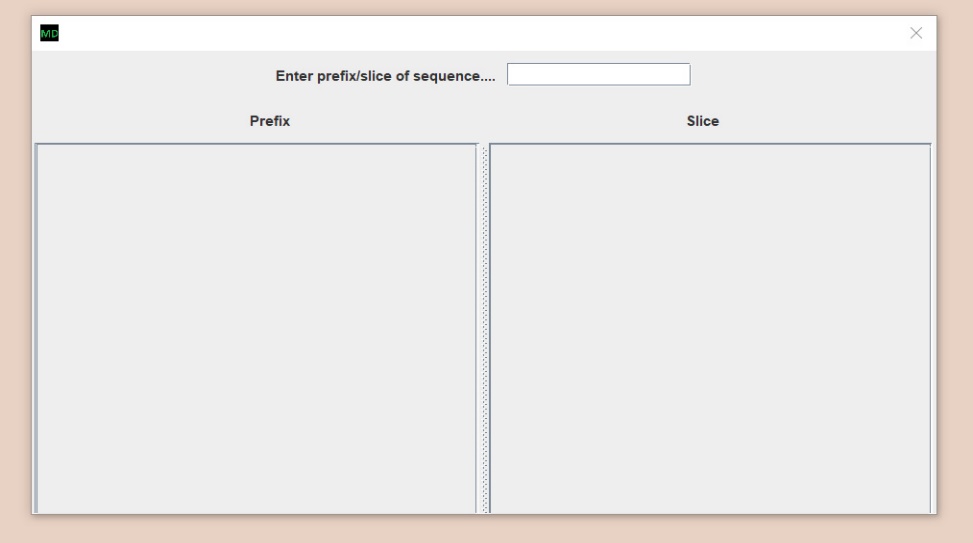
The input of this interface is a file with mass-spectrum, obtained from investigated modified peptide.

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

**Fig. 1 pull-down menu**

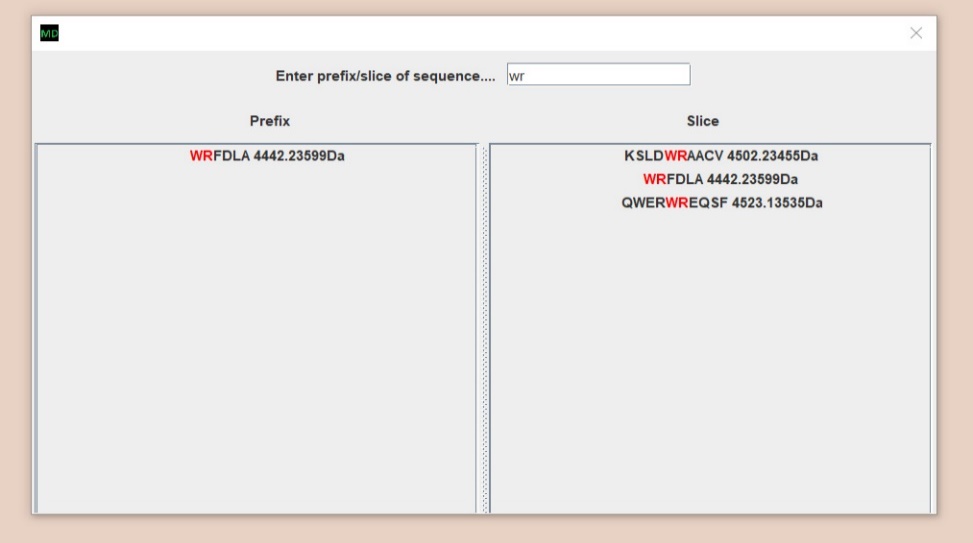


Also, there is an item named “Generate res”, it will subsequently highlighted. The dialog window appears by clicking item “Find” (fig. 3.1)



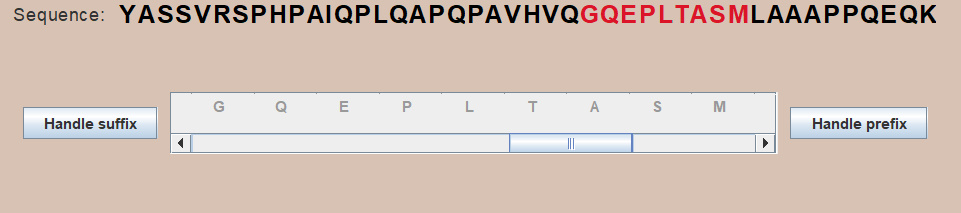
**Fig 3.1 the empty dialog window**

When user inputting some sequence of amino acids to 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 are appearing in the right part of this window. (fig. 3.2)



**Fig 3.2 wr was inputted**

The main frame of present interfaces appears by selecting one these peptides. In the top there is an amino acid sequence just below there is a scrollable panel, where are the same amino acids but 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 scrollable panel there are buttons “handle suffix” and “handle prefix” (fig. 4)



**Fig. 4 sequence and scrollable panel**

The buttons in the scrollable panel are getting enabled by clicking buttons “handle prefix” or “handle suffix”. Then if user click on some of amino acids (buttons in scrollable panel), the program starts searching positions in suffix, which starts with pressed amino acid, or prefix, which ends with pressed amino acid, (it depends on which of prefix or suffix 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’ peaks. Also, the mistake in determining does some PTM or substitution fit present mass difference is different when the program search through prefix and suffix (it is cause of peculiarity of methodology):

The mistake when the program searches through the prefix:

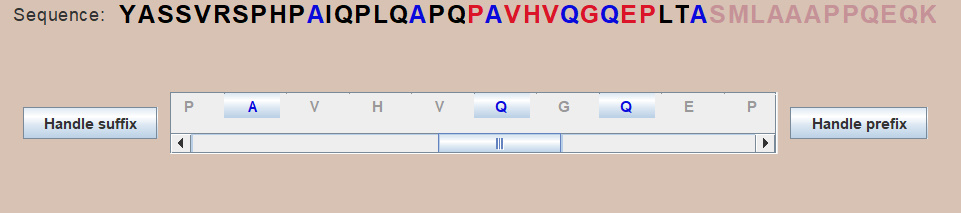
– mass of the prefix ends with i-position (the sum of all amino acids masses in this prefix)

ppm – parse par million, error tolerance.

The mistake when the program searches through suffix: , где – mass of the prefix ends with i-position.

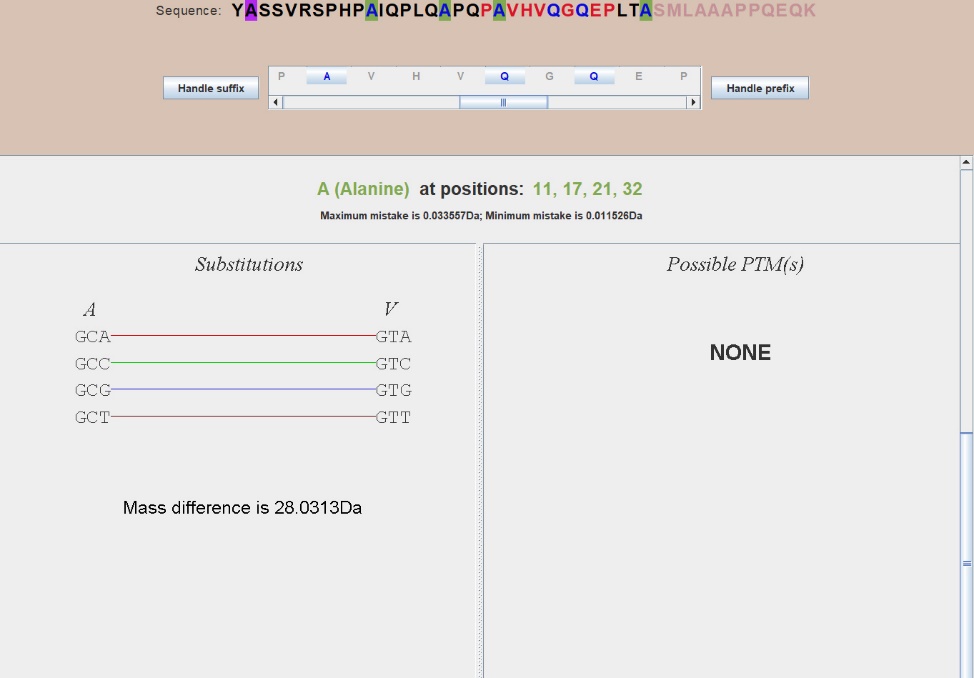
– mass of the whole peptide.

After the program checked for all allowable position whether there is a PTM or substitution fitting the mass difference, the positions where the substitutions might occur are highlighting with blue color, where the PTM – with orange color (fig. 5) The positions, which are not in the suffix or prefix under observation highlight with pale color (observed prefix ended with “S”)



**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 amino acid V could substitute 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).

****

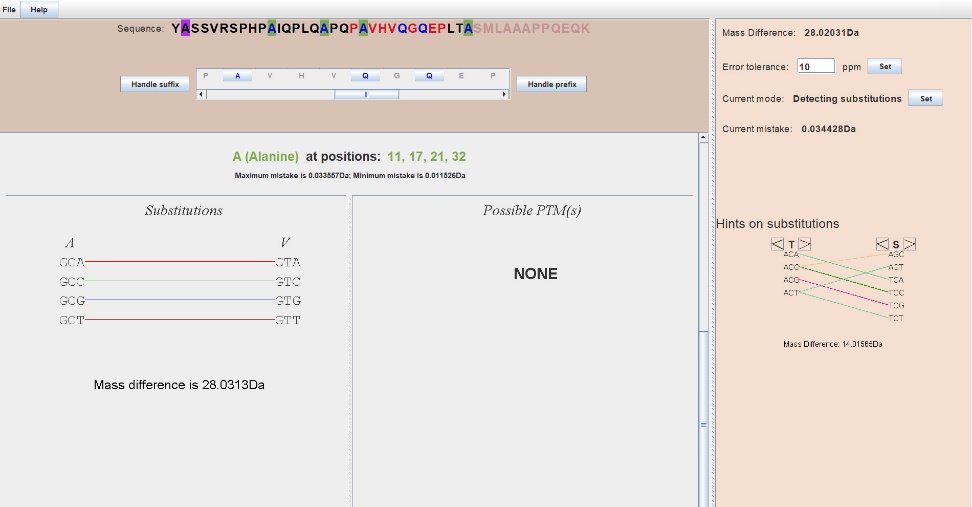
**Fig. 6 information about the substitution**

After clicking on “A” in the top sequence first “A” is on the purple background since the remaining are on pale green background, because first “A” does not fit the conditions since the mistake is not enough big (mass of prefix is not enough big).

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 they are differing at one nucleotide this means that if such SNP occurs such amino acid substitution occurs.

**Results**

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

**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 is 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 University where I was working)

1. YASSVRSPHPAIQPLQAPQPAVHVQGQEPLTASMLAAAPPQEQK

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

1. EAATQEDPEQVPELAAHEVSASEAEERPVAEEEILL

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

Developed program gave a right result in both cases.

**Conclusion**

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

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

**References**

1. Б. Льюин. *Клетки*. БИНОМ Россия, 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