**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 one – 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. The output is the sequence fragments, which might incorporate an appropriate amino acid substitution or a PTM, appear highlighted.

**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).

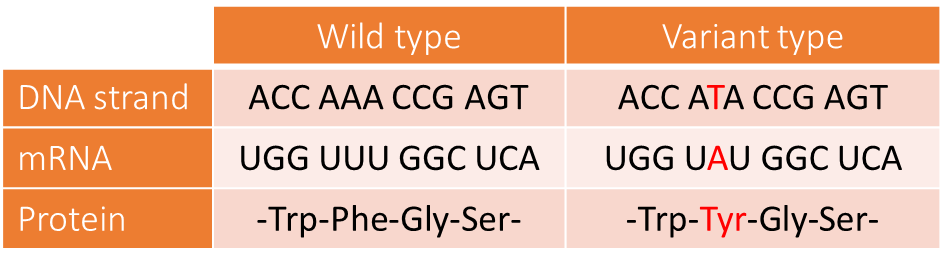
Having learned to find positions in protein sequence where the substitution might occur, we would get a possibility to identify the so-called variant proteins, which can be biomarkers for a variety of hard .

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 and make 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. (Fig.1)

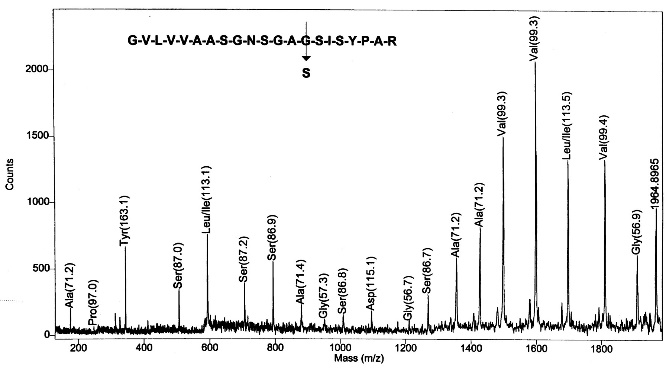
**Fig. 1 Example of SNP**



Post-translational modifications (PTMs) of the amino acids can also change the protein’s 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 more often than substitutions in nature, therefore if some PTM and some substitution correspond to present difference in mass, most probably the PTM occurred.

Our work concentrates on the qualitative 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 a mass spectrometer (special device) and this device gives us the mass spectrum (Fig. 2). Mass spectrum is a histogram, where X – axis is the 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 a certain fragment of investigated amino acid sequence, and based on that fragment, we can establish (with the help of methods of biological alignment) the whole peptide, which is most similar to the present one, in other words we can establish the peptide which contains 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 a protein which exists in nature (wild-type peptide).



**Fig. 2 Example of mass spectrum**

Analyzing this data manually is inconvenient. Therefore, the idea to develop a software tool for analyzing such data arises.

The goal of our 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 to handle the positions in the peptide, where a PTM or a substitution might occur.

This work differs from previous works concentrated on this theme. Firstly, the result of this the work is a *graphical interface.* Secondly, this interface takes input high-resolution data

**Methodology**

**The algorithm**

One of the main aims of this work was to find the positions in the peptide, where a PTM or a substitution could occur. For this purpose, we developed the algorithm. Since the mass spectrum is a set of masses of prefixes and suffixes (ideally) of source peptide, we can’t search through the whole peptide, hence we have to search through a prefix or a suffix.

Algorithm:

Input data: Amino acid sequence, difference in mass (∆M), error tolerance, the position which is the end of prefix or the start of suffix (k).

1. For each position (i) calculate the error
   1. If it is a prefix:
   2. If it is a suffix:

– mass of the prefix ends with i-position. – mass of the whole peptide. – error tolerance of a mass spectrometer. These formulas are results of the view of source data (set of peaks of masses of prefixes and suffixes)

1. Find a substitution or a PTM which provide a mass in difference ∆m: ∆m (∆M – ε; ∆M + ε)
2. Give a result

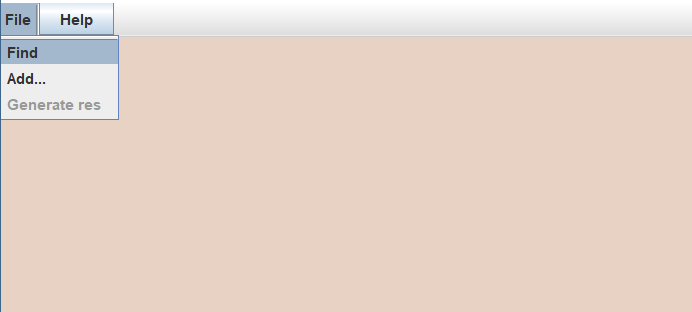
**Graphical interface**

To develop this application, we used Java and Swing (GUI widget toolkit)

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

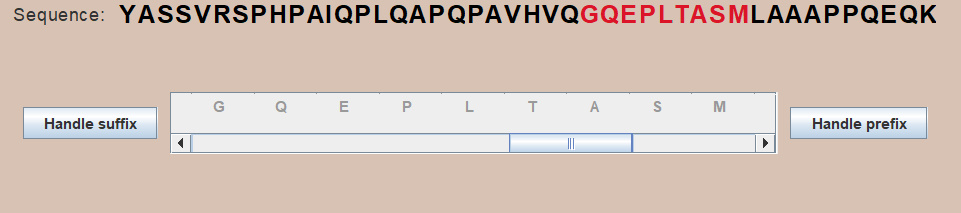
There is a possibility of adding of a large number of files of mass-spectrums and searching for peptides through the program’s database. These functions are in the “File” tab (Fig. 3)

**Fig. 3 pull-down menu**



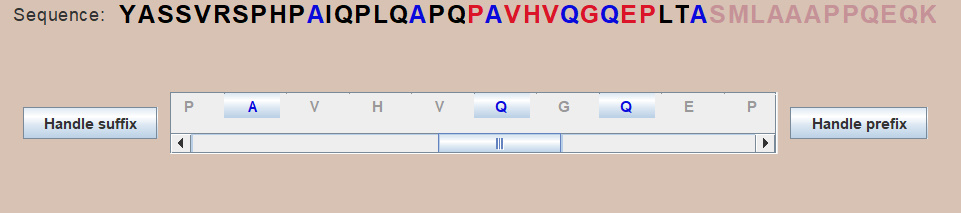
Also, there is an item named “Generate res”, it allows a user to generate a file loaded with information about chosen peptide (the positions where a PTM or a substitution could occur).

When a user chooses a peptide the main frame of present interfaces appears. At the top of this frame there is an amino acid sequence, just below there is a scrollable panel, where there are the amino acids, from sequence, but a user can interact with them. Some of the amino acids in the top panel are red, because a user can see them now in the scrollable panel. To the right and to the left from the scrollable panel there are buttons “handle suffix” and “handle prefix” (Fig. 4)



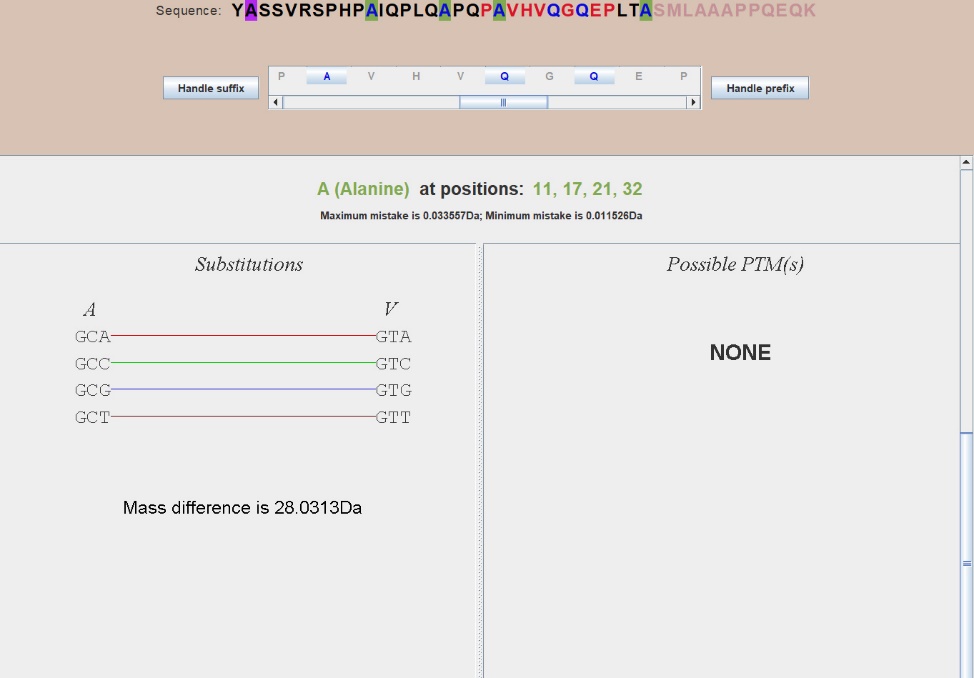
**Fig. 4 sequence and scrollable panel**

When a user clicks on one of the buttons “handle prefix” or “handle suffix” the algorithm described earlier begins. According to results of this algorithm, the position where a substitution might occur appears highlighted in blue color, where a PTM might occur appears highlighted in orange color. The positions, which are not in the suffix or the prefix under observation are highlighted in pale color (the observed prefix ended with “S”) (Fig. 5)



**Fig. 5 highlighted positions**

When a user clicks on one of the “A”s in the top sequence, first of these “A”s appears on the purple background while the remaining ones appear on pale green background, because the first “A” does not fit the conditions since the error 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 encodes A, GTA encodes V, the difference between them is that in the second position there are different nucleotides. If a SNP (C>>T) occurs, a substitution A>>V occurs.



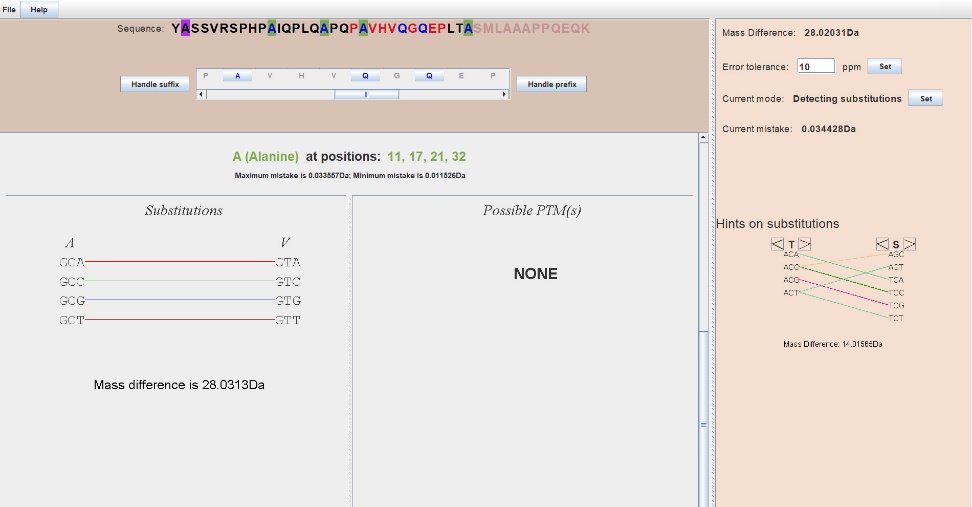
**Fig. 6 information about the substitution**

**Results**

As a result, the following interface has been developed (Fig. 7):

There are some additional features that make work with this application more convenient besides the functionality described above.

**Fig. 7 general view of the application**



* Tab “Help”, where user can learn how to work with this application.
* Hot keys for some actions (there is a list of them in “Help”).
* Hints on substitutions. It is located in the bottom right corner. A 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 carried out 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).

1. EAATQEDPEQVPELAAHEVSASEAEERPVAEEEILL

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

The developed program gave the right result in both cases.

**Conclusion**

In 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*. BINOM 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