**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**

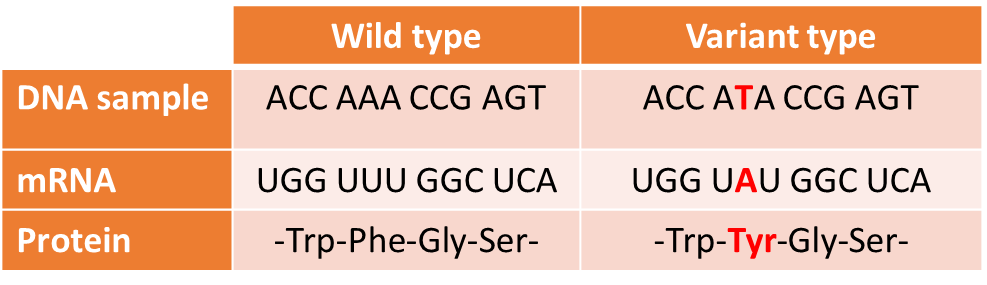
Белки играют важнейшую роль в нашей жизни. Ведь именно белки регулируют внутриклеточные . Если в белке произойдет какой-либо сбой, то он будет выполнять не ту функцию, которую должен, а это, в свою очередь, может привести к появлению тяжелых заболеваний у человека.

Причинами таких сбоев могут быть абсолютно разные факторы, но в данной работе рассматривался один из наиболее важных – однонуклеотидная замена в геноме.

Научившись находить места в белковой последовательности, где могла произойти замена, мы получим возможность выявлять, так называемые, вариантные белки, которые могут являться биомаркерами различных .

Сейчас существует два противоположных по смыслу подхода к анализу таких белков: качественный и количественный. В основе количественного метода лежит идея о том, что в здоровом организме должно сохраняться количественное отношение между разными белками. Если какое-нибудь отношение нарушится, то это значит, что в организме произошел сбой. Задачей количественного анализа является как раз выявление таких и на основе них делать определенные выводы.

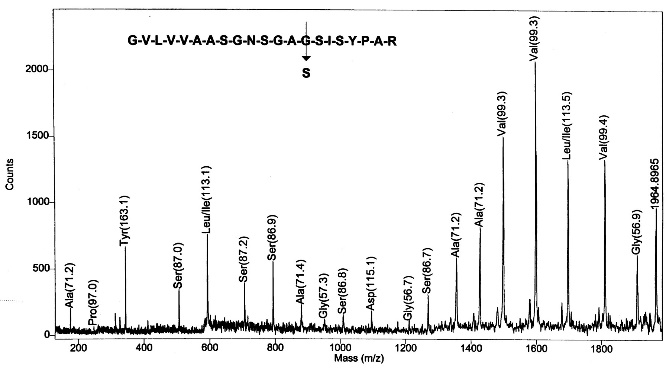
При качественном анализе интересно не количество, а именно качественный состав пептида (аминокислотная последовательность).

Так как аминокислоты кодируются тремя последовательными нуклеотидами в цепи ДНК (кодонами), то замена одного из нуклеотидов может привести к замене аминокислоты. А, как следствие, к изменению массы пептидов (последовательностей аминокислот), содержащих данную аминокислоту.

Example of SNP

Также к изменению массы пептида может привести посттрансляционная модификация (ПТМ). ПТМ – это модификация аминокислоты посредством изменения ее химического состава. Например, метионин (аминокислота) может окислиться, что приведет к изменению его массы приблизительно на 16 Да. В природе ПТМ встречаются чаще чем замены, поэтому если данному изменению массы может соответствовать, и замена, и модификация, то, вероятнее всего, произошла ПТМ.

Наша работа посвящена качественному анализу. Результатом такого анализа является масс-спектр, взятый с исследуемого пептида. Вначале растворенный пептид бомбардируется заряженными частицами, и в идеале каждая молекула этого пептида распадается на две (префикс и суффикс) и на каждой части находится по одному положительному заряду. Далее эти фрагменты прогоняют через масс-спектрометр (специальный прибор), и данный прибор выдает масс-спектр. Масс-спектр – это график, по оси абсцисс которого, в нашем случае, отложена масса ионизированного фрагмента молекулы, а по оси ординат – интенсивность ионного тока. (количество зарегистрированных частиц с таким значением массы). Соответственно, данный график представляет собой набор пиков, которые являются изначальной молекулы. По этому набору пиков удается восстановить некоторый фрагмент, исследуемой аминокислотной последовательности, а по полученному фрагменту можно восстановить (с помощью методов биологического выравнивания) весь пептид (наиболее похожий на данный), так как в природе существует конечное количество белков. Также по данному набору пиков можно восстановить массу, исследуемого пептида и соответственно разность масс модифицированного и существующего в природе пептида.



Example of mass spectrum

Анализировать полученные из эксперимента данные вручную – очень неудобно. Поэтому возникает идея, о том, чтобы разработать программное обеспечение для анализа таких данных.

Целью моей работы было разработать программное обеспечение для анализа данных, полученных в ходе работы с модифицированным пептидом.

Основной функцией данного программного интерфейса является выделение тех позиций в пептиде, где могла произойти ПТМ или замена.

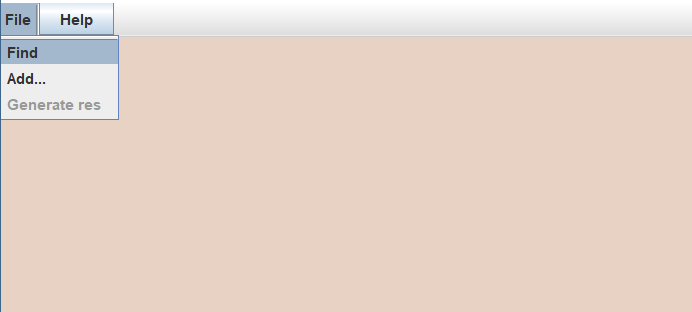
Данная работа разительно отличается от предыдущих, во-первых, тем, что в итоге был реализован *графический интерфейс,* а во-вторых, тем, что для анализа берутся данные высокого разрешения (масс-спектр), в отличие от предыдущих работ, посвященных данной .

**Methodology**

В ходе разработки данного интерфейса использовался язык программирования Java и библиотека Swing для создания графических интерфейсов.

В качестве входных данных итоговый интерфейс принимает файл с масс-спектром, полученный при анализе модифицированного пептида.

В реализованном интерфейсе имеется возможность добавления большого количества файлов с масс-спектрами и затем поиска пептидов в базе данных программы. Эти функции находятся во вкладке File (рис. 1)

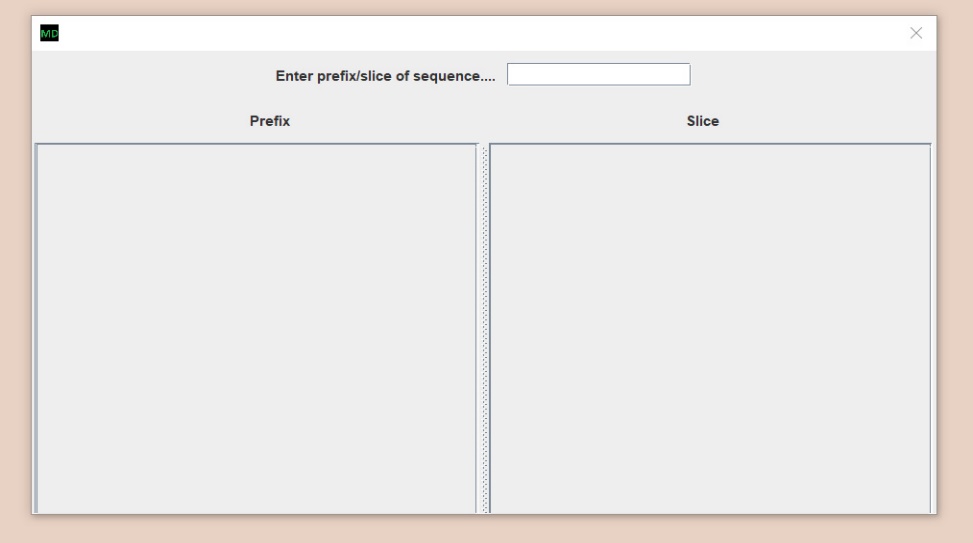


**Рис. 1 выпадающее меню**

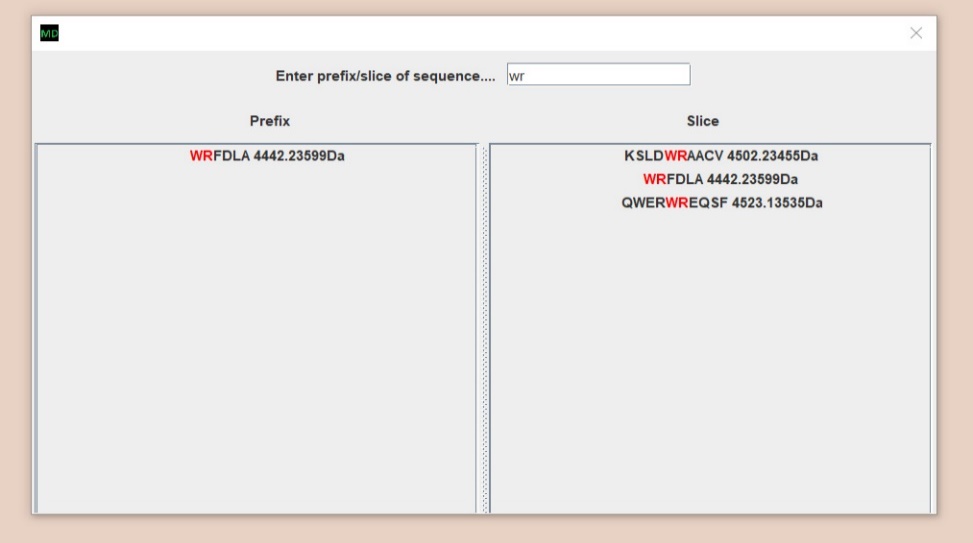
Также в этом меню присутствует пункт “Generate res”, о котором будет сказано чуть позже, он становится активным, когда выбран какой-нибудь пептид.

При нажатии пункта find появляется диалоговое окно

(рис. 3.1)

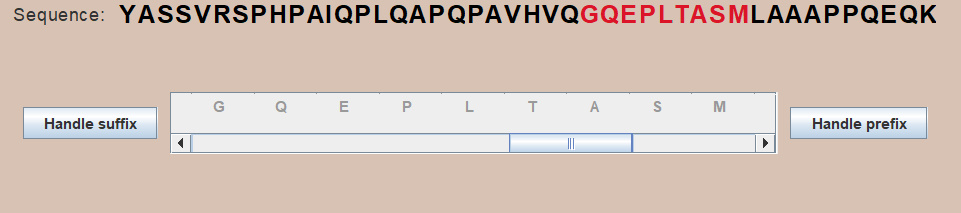
**Рис. 3.1 пустое диалоговое окно**

При вводе какой-нибудь последовательности аминокислот, в левой части данного окна появляются пептиды, где данная строка является префиксом пептида, а в правой части, где данная строка просто встречается в пептиде (рис. 3.2)



**Рис 3.2 введена строка wr**

При выборе пептида, открывается главное окно данного интерфейса, где в верхней части экрана появляется аминокислотная последовательность, чуть ниже прокручиваемое окно, где представлены те же аминокислоты, только в приближенном виде, и в верхней последовательности помечаются красным, те аминокислоты, которые видны в данном прокручиваемом окне. Также справа и слева от данного окна присутствуют кнопки handle suffix и handle prefix соответственно (рис. 4)



**Рис. 4 последовательность и прокручиваемое окно**

При нажатии handle prefix или handle suffix кнопки в прокручиваемом окне становятся активными, если нажать на какую-нибудь аминокислоту, то программа будет искать замену или ПТМ, подходящие под данное изменение по массе в суффиксе или в префиксе в зависимости от того, на что нажал пользователь. Программа ищет не во всей последовательности, потому что масс-спектр представляет собой набор пиков масс префиксов и суффиксов, также погрешность в случае префикса и в случае суффикса разные (это опять же следует из особенности методики качественного анализа):

Погрешность в префиксе:

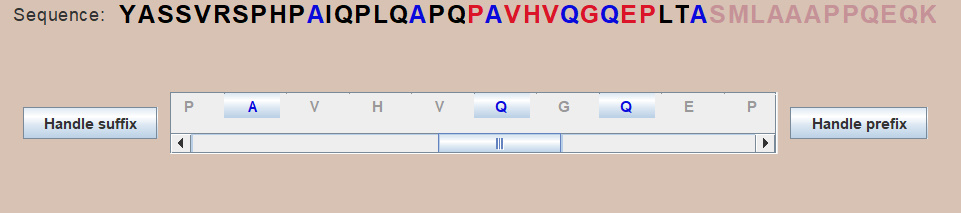
- масса префикса (сумма всех масс аминокислота в данном префиксе

ppm – parse par million (долей на миллион), погрешность прибора (масс – спектрометра),

Погрешность в суффиксе: , где – масса префикса, заканчивающегося началом суффикса.

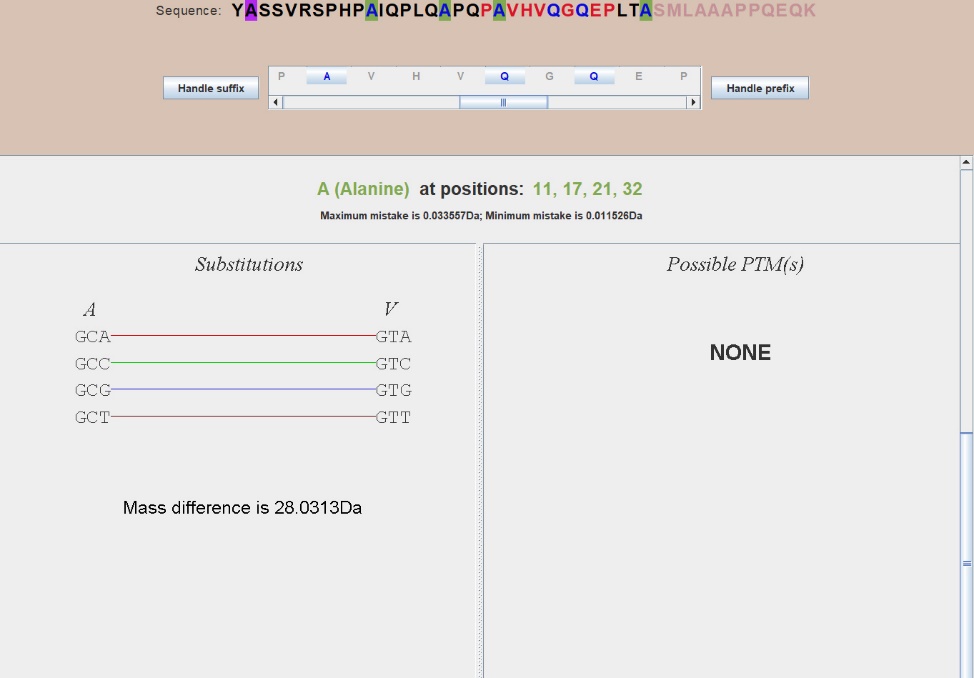
– масса всего пептида.

После того как программа для каждой позиции проверила есть ли замена или ПТМ, подходящая под данное изменение по массе. Позиции, где могла произойти замена подсвечиваются синим, где ПТМ – оранжевым (рис. 5) А бледным цветом подсвечиваются позиции, которые не входят в рассмотрение программы в данный момент (программа искала модификации в префиксе, заканчивается S)



**Рис. 5 подсвеченные позиции**

Далее при нажатии на активную кнопку, например на одну из A, то появляется информация о том, что могло произойти в данной позиции (рис. 6)

****

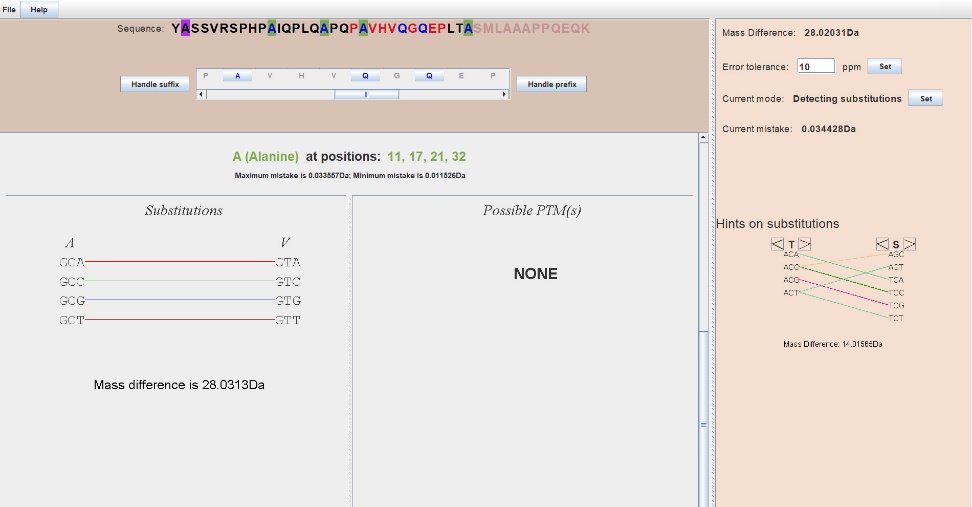
**Рис. 6 информация о замене**

После нажатия на А в верхней последовательности все предстали на разном фоне первая А на фиолетовом, а последующие на бледно-зеленом, потому первая А не подходит, так как погрешность недостаточно большая (масса префикса недостаточно большая).

Замена, которая могла произойти в данных позициях – это замена A >> V и некоторые кодоны, кодирующие данные аминокислоты, соединены разноцветными линиями. Они соединены, потому что могут перейти друг в друга посредством однонуклеотидной замены.

**Results**

В результате был реализован следующий интерфейс (рис. 7):



**Рис 7 общий вид приложения**

Помимо описанного функционала интерфейс имеет дополнительные возможности для упрощения работы в данном приложении.

* Горячие клавиши для некоторых действия, которые делаются по кнопкам (их полный список есть во вкладке Help в верхнем левом краю экрана)
* Подсказка по сравнению аминокислот. Она находится в правом нижнем углу экрана. Данная подсказка представляет собой два окна выбора аминокислоты (с помощью них пользователь может выбрать две интересующие его аминокислоты), и панель отображения кодонов данных аминокислот
* Упомянутая ранее вкладка Help, где пользователь может посмотреть справку об этом приложении

Чтобы удостовериться, что разработанная программа работает исправно, были проведены испытания.

Из базы данных Академического Университета были взяты два проанализированных пептида:

1. YASSVRSPHPAIQPLQAPQPAVHVQGQEPLTASMLAAAPPQEQK

В данном пептиде произошла замена A>>V (в префиксе).

1. EAATQEDPEQVPELAAHEVSASEAEERPVAEEEILL

В данном пептиде произошла замена A>>V (в суффиксе)

Проанализировав эти пептиды, моя программа выдала ответ, какой должен был быть.

**Conclusion**

Реализованный интерфейс отличается от приложений, которые делались ранее, тем что полученный продукт ориентирован на работу с данными высокого расширения, то есть на вход программа принимает файл, полученный сразу с масс-спектрометра, а не отдельно аминокислотную последовательность и изменения по массе.

В дальнейшем планируется расширение функционала для решения более специфических задач в данной области.

**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