MutationDetector – a software tool for detecting single amino acids polymorphisms

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).

Since three consequent nucleotides, together forming a *codon*, encode an amino acid, an SNP can lead to a single amino acid substitution, also termed as *single amino acid polymorphism* (SAP), thereby potentially affecting the functionality of the protein, and implying a change in its mass.

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. Reliably distinguishing SAPs from PTMs represents an important yet challenging task.

In this work, we present MutationDetector – a software tool for detecting and localizing SAPs in peptides. It accepts as input a wild-type peptide 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 possibly incorporate a SAP or PTM appear highlighted.

Brilliantov Kirill

MutationDetector – a software tool for detecting single amino acids polymorphisms

Research Report

Introduction

In any living organism, sub-cellar processes are regulated by proteins[1] – and thus, alternation of the structure of proteins, along with their functionality, may drastically impact its condition.

From this point of view, of particular importance are single nucleotide polymorphisms (SNPs), which may result in single amino acid substitutions, or *single amino acid polymorphisms* (SAPs), in proteins and peptides. An example of such situation is provided in Fig. 1.

	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-

Figure 1. An example of an SNP causing a SAP.

Consequently, it is important to learn detecting potential SAPs in proteins, or, stated otherwise, identifying so-called *variant proteins*, which can be biomarkers for a variety of severe diseases[2]. In particular, it is essential to be able to reliably distinguish SAPs from posttranslational modifications (PTMs), the role of which is very different. However, this may be challenging, particularly due to the fact that a SAP and PTM may lead to the same change in the protein mass.

Since PTMs are encountered much more frequently than SAPs, a strong evidence of the latter is required for reliably classifying the alternation of the protein structure as a SAP.

In this work, we present a method and a software tool MutationDetector for detecting and localizing SAPs in peptides,

based on high-resolution tandem mass spectra acquired from those.

Experimental data

A melanoma cancer cell line sample was analyzed by high-performance liquid chromatography coupled online with tandem mass spectrometry (HPLC-MS/MS). High-resolution full MS and MS/MS spectra were acquired on Thermo Orbitrap Q Exactive Plus at a resolution of 140000 and 17500 for MS and MS/MS, respectively.

According to this technology, peptide molecules are ionized, then separated by the *m*/*z* value (where *m* and *z* denote the ion mass and charge, respectively), and finally, fragmented. A tandem (MS/MS) spectrum contains the information on the ion current for each registered value of *m*/*z*. An example of such spectrum is given in Fig. 2.

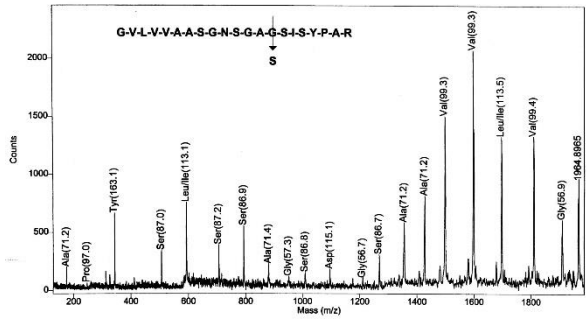

Figure 2. An MS/MS spectrum acquired from the peptide GVLVVAASGNSGAGSISYPAR.

Data preprocessing

The obtained high-resolution MS/MS spectra were deisotoped and charge state deconvoluted with MS-Deconv[3]. As a result, we obtained mass spectra for neutral masses rather than the *m*/*z* ratio.

Subsequently, those mass spectra were processed with an appropriately modified version of the Twister algorithm[4]. In this way, for several mass spectra, we derived highly accurate

fragments of the underlying protein sequences. Based on those fragments, the respective mass spectra were matched against the SwissProt (v. 02/2015) human proteome database[5]. Further, we restricted our attention to the matches suggesting that the mass spectrum could have been acquired from a peptide with the sequence differing from that of a certain wild-type peptide by a SAP, which occurred either before or after the retrieved fragment.

Algorithm scheme

As input, we are given an MS/MS spectrum *S*, along with a peptide sequence *P*, a fragment *F* of which was extracted from *S*. In addition, it is known than either the prefix of *P* preceding *F* or the suffix of *S* succeeding *F* matches the spectrum as well, while the remaining part of the sequence should be modified in order to match *S*. The mass ΔM of the required modification can be calculated as the difference between the precursor mass of *S* and the mass of *P*.

Thus, our goal is to determine whether ΔM could occur due to a SAP, or a PTM, or both. In case both options seem valid, we attribute ΔM to a PTM, since PTMs are observed substantially more often than SAPs. Below we explain the validation procedure in more detail.

To decide whether ΔM can be explained by a SAP in the corresponding prefix or suffix of the sequence, we check whether any of the amino acid contained in the respective sequence fragment could bear a modification resulting in a similar mass difference. Thereby, we need to allow for a potential error in the mass measurements. Since we deal with high-resolution data, the error threshold should be specified relatively to the absolute masses measured in the experiment, and expressed in *parts per million*, or *ppm*. For example, given an error threshold of 10 ppm, and an absolute mass of 10 kDa, we allow for an error in the measured mass up to $10{,}000 * 10 / 10^6 = 0{,}1$ kDa.

In case a putative SAP is located inside the prefix, the allowable error is calculated relative to prefix mass. Otherwise, it is calculated with respect to the precursor mass of the spectrum *S*.

The software tool: MutationDetector

The proposed algorithm was implemented in a software tool MutationDetector, in the Java programming language using the Swing library.

MutationDetector takes as input a file containing a mass spectrum acquired from a putative variant peptide, and the respective wild-type sequence.

As soon as a user specifies a peptide, the main frame appears. At its top the amino acid sequence is displayed; below it, there is a scrollable panel that allows the user to navigate along the sequence; the fragment currently visible at the navigation panel is highlighted in red at the top. To the right and left of the scrollable panel, there are buttons “Handle suffix” and “Handle prefix” (Fig. 3).

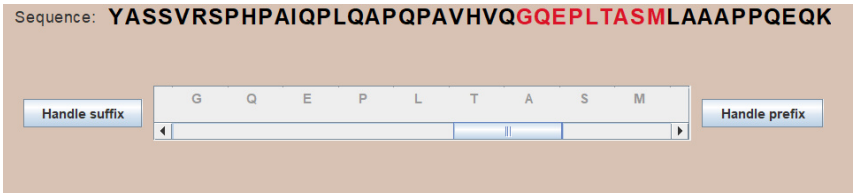

Figure 3. The graphical interface.

When the user clicks on one of the buttons “Handle prefix” or “Handle suffix”, the algorithm described above is launched. Upon its execution, the positions, at which SAPs and PTMs might have occurred, get highlighted in blue and orange, respectively. The amino acids not contained in the suffix or prefix under consideration are displayed in a pale color (Fig. 4).

Figure 4. The output sequence with highlighted amino acids (the prefix under consideration ends at 'S').

However, at this stage, the error tolerance is not yet taken into account, and in order to see, at which places a SAP might have occurred indeed, the user needs to select a particular amino acid. In the example being discussed, if the user clicks on any "A" from the sequence at the top, the background of the first of those becomes purple, while that of the other ones becomes pale green (Fig. 5): this means that the first "A" is not eligible for a SAP, taken into account the tolerance specified. The substitution, which could have occurred at the other positions, is A -> V. The three-codon representations of those amino acids witnessing for that are displayed below the sequence, and the corresponding pairs of those are connected with colored line segments. For example, the codon GCA and GTA encode A and V, respectively, and differ only at the second nucleotide. If a SNP (C->T) occurs, it causes an amino acid substitution A->V.

Figure 5. Putative SNPs and corresponding SAPs.

Results

To summarize, we have developed a software tool MutationDetector, the graphical user interface of which is shown on Fig. 6.

Apart from the main functionality described above, the interface provides a few more features, and namely:

- The tab “Help” instructing user how to work with this application.
- Hot keys for certain actions (a list of those is provided in the section “Help”).
- Hints on substitutions located in the bottom-right corner. A user can select any two amino acids and see whether any SNP can lead to a substitution of one of those with the other.

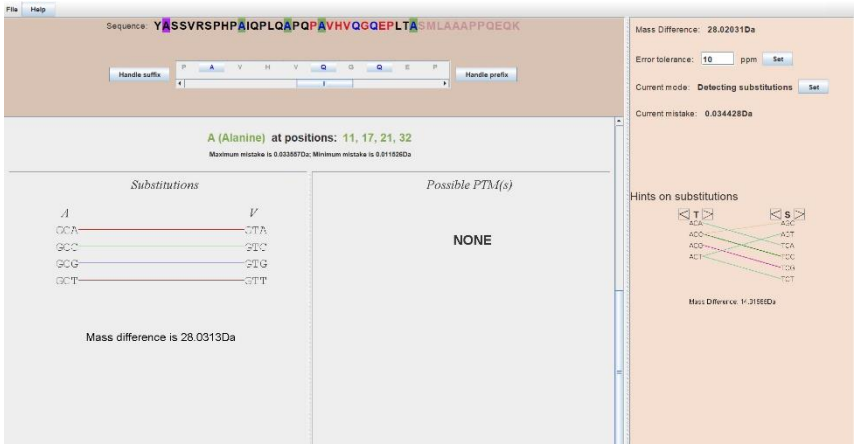

Figure 6. The interface of the software tool MutationDetector.

We tested MutationDetector on the mass spectrometry data acquired from the melanoma cell line (see above). Below we provide two examples illustrating its behavior.

1. YASSVRSPHPAIQPLQAPQPAVHVQGQEPLTASMLA AAPPQEQK

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

2. EAATQEDPEQVPELAAHEVSASEAEERPVAEEEILL

In this peptide, a substitution A->V occurred (in the suffix)

In either case, MutationDetector provides as ouput a correct result.

Conclusion

We have developed a software tool MutationDetector for identifying variant peptides. It was tested on a melanoma cancer cell line, and demonstrated correct behavior

In the future, we plan to extend the functionality of MutationDetector, thereby adjusting it for solving a number of particular problems.

References

1. B. Lewin. *Cells*. BINOM Russia, 2011. 951 c.
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. X. Liu, Y. Inbar, P. C. Dorrestein, C. Wynne, N. Edwards, P. Souda, J. P. Whitelegge, V. Bafna, P. A. Pevzner. *Deconvolution and database search of complex tandem mass spectra of intact proteins: a combinatorial approach*.
   Molecular and Cellular Proteomics, 9:2772-2782, 2010.
4. 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.
5. K.V. Vyatkina, A.A. Lobas, L.I. Levitsky, M.V. Ivanov, E.M. Solovyeva, S.A. Moshkovskii, M.V. Gorshkov. *Automated detection and validation of variant peptides in cancer cell lines via de novo sequencing assisted database search*. Proc. 5th International Conference "POSTGENOME'2018" "In Search of Models for Personalized Medicine", Kazan, Russia, October 29-November 2, 2018, pp. 41-42.