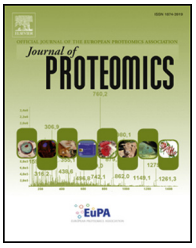
JOURNAL OF PROTEOMICS XX (2015) XXX–XXX

Available online at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/jprot



Methionine to isothreonine conversion as a source of false discovery identifications of genetically encoded variants in proteogenomics

Alexey L. Chernobrovkin[a,b], Arthur T. Kopylov[a], Victor G. Zgoda[a], Alexander A. Moysa[a], Mikhail A. Pyatnitskiy[a], Ksenia G. Kuznetsova[a], Irina Y. Ilina[a], Maria A. Karpova[a], Dmitry S. Karpov[a,c], Alexander V. Veselovsky[a], Mark V. Ivanov[d,e], Mikhail V. Gorshkov[d,e], Alexander I. Archakov[a], Sergei A. Moshkovskii[a,f,*]

[a]Institute of Biomedical Chemistry, Moscow 119121, Russia
[b]Karolinska Institutet, Stockholm SE-171 77, Sweden
[c]Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow 119991, Russia
[d]Institute for Energy Problems of Chemical Physics, Russian Academy of Sciences, Moscow 119334, Russia
[e]Moscow Institute of Physics and Technology (State University), Moscow region, Dolgoprudny 141700, Russia
[f]Pirogov Russian National Research Medical University, Moscow 117997, Russia

ARTICLE INFO

Article history:
Received 21 December 2014
Accepted 7 March 2015

Keywords:
Proteogenomics
Tandem mass spectrometry
Cancer cell line
Methionine
Isothreonine
Iodoacetamide

ABSTRACT

Searching deep proteome data for 9 NCI-60 cancer cell lines obtained earlier by Moghaddas Gholami et al. (Cell Reports, 2013) against a database from cancer genomes returned a variant tryptic peptide fragment 57-72 of molecular chaperone HSC70, in which methionine residue at 61 position is replaced by threonine, or isothreonine (homoserine), residue. However, no traces of the corresponding genetic alteration were found in the cell line genomes reported by Abaan et al. (Cancer Research, 2013). Studying on the background of this modification led us to conclude that a conversion of methionine into isothreonine resulted from iodoacetamide treatment of the probe during a sample preparation step. We found that up to 10% of methionine containing peptides experienced the above conversion for the datasets under study. The artifact was confirmed by model experiment with bovine albumin, where three of four methionine residues were partly converted to isothreonine by conventional iodoacetamide treatment. This experimental side reaction has to be taken into account when searching for genetically encoded peptide variants in the proteogenomics studies.

Biological significance
A lot of effort is currently put into proteogenomics of cancer. Studies detect non-synonymous cancer mutations at protein level by search of high-throughput LC–MS/MS data against customized genomic databases. In such studies, much attention is paid to

Abbreviations: HSC70, heat shock cognate 71 kDa protein; isoT, isothreonine; HCD, high energy collision dissociation; FDR, false discovery rate; IAA, iodoacetamide; PSM, peptide-spectrum match.

* *Corresponding author at:* Institute of Biomedical Chemistry, 10 Pogodinskaya Str., Moscow 119121, Russia. Fax: +7 499 245 0857.
E-mail address: smosh@mail.ru (S.A. Moshkovskii).

http://dx.doi.org/10.1016/j.jprot.2015.03.003
1874-3919/

potential false positive identifications. Here we describe one possible cause of such false identifications, an artifact of sample preparation which mimics methionine to threonine nucleic acid-encoded variant. The methionine to isothreonine conversion should be taken into consideration for correct interpretation of proteogenomic data.

1. Introduction

Recent introduction of mass-spectrometers with high mass accuracy and high resolving power, such as the Orbitrap [1], to proteomics has dramatically improved the capabilities of HPLC-MS/MS-based bottom-up proteome analysis. These advances enabled a dozen of thousands of tryptic peptides and several thousands of protein groups being identified with high confidence in a single experimental run [2–4]. In these experiments the LC–MS/MS data is analyzed by a database search using an applicable search engine [5] or a variety of different search engines applied to the same datasets [6]. However, database search generates "the streetlight effect", i.e. only those peptides may be found, which are predicted from the custom genomic database. Thus, a plethora of mass spectra used to be abandoned and not assigned to any peptide sequence. A number of alternative approaches to peptide identifications based on de novo sequencing have also been developed [7–11]. A method for accounting of the missing information in the protein database using a coding genome polymorphism data has been explored recently [12]. In general, a use of one or more custom DNA and/or mRNA sequence databases for LC–MS/MS data search is becoming a current trend in identifying the encoded variants of amino acid sequence that originated from single amino acid polymorphism or alternative splicing [13]. These areas of research, as well as the studies on genome re-annotations using proteomics data, are often referred to as proteogenomics [14]. Cancer proteogenomics is especially an important case of using the customized sequence databases because of a large amount of biologically relevant non-synonymous somatic mutations across the tumor genomes [15–17].

One of the challenges associated with the large size of the database that combined several genomes is the growing level of false positive identifications, which may constitute a significant portion of proteogenomic results [18]. Moreover, a number of chemical and post-translational modifications mimic single amino acid variants. One of the most obvious examples is spontaneous asparagine and glutamine deamidations that resulted in aspartic and glutamic acid residues, respectively [19]. These modifications occurring both in vivo and in vitro are indistinguishable from genetically encoded point mutations of Asn to Asp or Gln to Glu in the proteins.

In this work, we describe yet the other chemical modification which may produce false positive identification of peptides in shotgun proteogenomics study by imitation of genetically encoded mutations. This is methionine to isothreonine (homoserine) in vitro conversion which mimics methionine to threonine mutation. Analyzing publicly available data of NCI-60 cancer cell line proteomes [20], we found that the scale of this artifact can be significant to affect the interpretation of the results of proteogenomics studies. In order to explore the above conversion and distinguish MS/MS spectra of peptides containing isothreonine or threonine residues at the same locations in the sequences, synthetic peptides were obtained and their mass spectra were analyzed. Bovine serum albumin was used to confirm methionine to isothreonine conversion during conventional sample preparation for shotgun proteome analysis.

2. Materials and methods

2.1. Modified database for search of shotgun proteome data

NCI-60 cell line data were downloaded from the web at http://wzw.tum.de/proteomics/nci60 [20]. The concatenated database comprised the UniProt complete human proteome database (release from January 2013; 87,638 records) and the colon cancer database generated from available genomic data [21] (127,486 records) were used for peptide identification. The search of the data against the modified database was performed using X!Tandem, Mascot and Andromeda search engines using so-called separate FDR, or one-by-one approach as described in [16,22].

2.2. Peptide and protein identification and quantification

In order to determine the relative abundances of peptides and proteins, intensity-based label-free quantification was used with a minimum of two unique peptides. All raw files were reprocessed using MaxQuant package, version 1.5.2.8 [23], which has Andromeda as a database search engine. The MaxQuant analysis included an initial search with precursor mass tolerance of 20 ppm, the results of which were used for mass recalibration. In the main Andromeda search, precursor and the fragment mass tolerances were set to 6 ppm and 20 ppm, respectively. The search included variable modifications, such as methionine oxidation, deamidation of asparagine and glutamine and N-terminal acetylation, as well as carbamidomethyl cysteine as a fixed modification. The minimal peptide length was set to seven amino acids and a maximum of two missed cleavages was allowed. The false discovery rate (FDR) was set to 1% for both peptide and protein identifications.

Global analysis of M/(iso)T conversion in NCI-60 cell line deep proteomes was performed using X!Tandem search engine [24], version 2012.10.01.1, and MPscore post-search validation tools [25]. The precursor and the fragment mass tolerances were 15 ppm and 0.03 Da for the X!Tandem search, respectively. Carbamidomethylation of C (+57.021464) was applied as fixed modification and N-terminal acetylation (+42.010565) was used as potential modification. Enzyme specificity was set to "trypsin" with maximum of 2 missed cleavages. The UniProt complete human proteome database was used for all X!Tandem searches. The total number of

Please cite this article as: Chernobrovkin AL., et al, Methionine to isothreonine conversion as a source of false discovery identifications of genetically encoded variants in proteogeno..., J Prot (2015), http://dx.doi.org/10.1016/j.jprot.2015.03.003