*W258–W263* *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
*doi:10.1093/nar/gkl245*

# MOD$^i$ : a powerful and convenient web server for identifying multiple post-translational peptide modifications from tandem mass spectra

**Sangtae Kim, Seungjin Na[1], Ji Woong Sim[2], Heejin Park[3], Jaeho Jeong[4], Hokeun Kim[5], Younghwan Seo[6], Jawon Seo[4], Kong-Joo Lee[4] and Eunok Paek[1,*]**

Research Center for Computer Technology, Hanyang University, Seoul, 133-791, Korea, [1]Department of Mechanical and Information Engineering, University of Seoul, Seoul 130-743, Korea, [2]School of Computer Science and Engineering, Seoul National University, Seoul 151-742, Korea, [3]College of Information and Communications, Hanyang University, Seoul 133-791, Korea, [4]Center for Cell Signaling Research, Division of Molecular Life Sciences and College of Pharmacy, Ewha Womans University, Seoul 120-750, Korea, [5]Functional Proteomics Center, Korea Institute of Science and Technology, Seoul 136-791, Korea and [6]Lab Frontier Co. Ltd, Seoul 120-750, Korea

Received February 14, 2006; Revised March 14, 2006; Accepted March 29, 2006

## ABSTRACT

**MOD$^i$ (http://modi.uos.ac.kr/modi/) is a powerful and convenient web service that facilitates the interpretation of tandem mass spectra for identifying post-translational modifications (PTMs) in a peptide. It is powerful in that it can interpret a tandem mass spectrum even when hundreds of modification types are considered and the number of potential PTMs in a peptide is large, in contrast to most of the methods currently available for spectra interpretation that limit the number of PTM sites and types being used for PTM analysis. For example, using MOD$^i$, one can consider for analysis both the entire PTM list published on the unimod webpage (http://www.unimod.org) and user-defined PTMs simultaneously, and one can also identify multiple PTM sites in a spectrum. MOD$^i$ is convenient in that it can take various input file formats such as .mzXML, .dta, .pkl and .mgf files, and it is equipped with a graphical tool called MassPective developed to display MOD$^i$'s output in a user-friendly manner and helps users understand MOD$^i$'s output quickly. In addition, one can perform manual *de novo* sequencing using MassPective.**

## INTRODUCTION

Identification of post-translational modifications (PTMs) is important to understand cellular functions of proteins (1). Sensitive methodologies based on conventional biochemical methods are lacking for the identification of PTMs *in vivo*, but recent advances in proteomic technology including mass spectrometry provide an approach to identify PTMs (2–5). Sequencing by tandem mass spectrometry (MS/MS) which has aided protein identification offers a tremendous potential for detecting PTMs (1).

Three approaches have been used to automatically interpret tandem mass spectra for peptide sequencing (6), namely, database searching (7–9), *de novo* peptide sequencing (10–12) and sequence tag approach (13,14), and there have also been efforts to combine these methods (15,16). Interpretation of experimental spectra is harder if a peptide sequence contains PTMs. One might consider extending one of these approaches in a straightforward manner to sequence a peptide with any number, any kind and any combination of PTMs, i.e. developing a virtual peptide database by incorporating peptides with all possible combinations of PTMs, or extend the set of amino acids by introducing new virtual amino acids such that the mass of each virtual amino acid corresponds to the mass of a post-transitionally modified amino acid. However, such extensions yield an exponential time algorithm or a polynomial time algorithm, the degree of which is very high. Thus these algorithms are not appropriate to interpret tandem mass spectra with multiple PTMs in a reasonable amount of time. Recently there have been efforts to formally define this problem and suggest a way to reduce the time requirement of PTM identification (17,18).

We have developed a convenient method called MOD$^i$ for rapidly interpreting tandem mass spectra of peptides with multiple PTMs. This method adopts a hybrid approach that combines *de novo* sequencing with database searching. It performs well even when a large number and types (>100 modification

*To whom correspondence should be addressed. Tel: +82 2 2210 2680; Fax: +82 2 2248 5110; Email: paek@uos.ac.kr

types) of potential PTMs are considered. In addition, we developed a graphical tool called MassPective that shows MOD$^i$'s output in a user-friendly manner. MassPective enables a user not only to quickly view and understand MOD$^i$'s output, but also to perform additional manual *de novo* sequencing so that a MOD$^i$'s interpretation can be manually inspected.

By incorporating MOD$^i$ and MassPective with web-based interface, we have produced MOD$^i$ web service. MOD$^i$ web will provide the biological community a fast and convenient vehicle for identifying PTMs from tandem mass spectra.

## METHODS

MOD$^i$ consists of five stages: peak selection, tag discovery, database search, tag chain generation and PTM identification. It assumes that the number of candidate proteins has already been reduced to 20 or less by protein identification, before the spectra set is analyzed for PTM identification.

(i) Peak selection: We select peaks with relatively high intensities (both globally and locally). The number of peaks selected is proportional to a parent ion mass.

(ii) Tag discovery: We perform partial *de novo* sequencing on the selected peaks to identify all the tags (partial amino acid sequences that do not contain PTMs) of length up to 3.

(iii) Database search: Using the tags identified, we search the peptide database for candidate peptides that contain any of the identified tags (called forward tags) or the reverse sequences of the identified tags (called reverse tags) of length at least 3. It should be noted that the peptide database we use does not contain any PTM information. Thus the scalability requirement is satisfied.

(iv) Tag chain generation: For each candidate peptide, we build a tag chain. A tag chain for a candidate peptide consists of non-overlapping forward or reverse tags of length at least 2, occurring in the candidate peptide and in-between gaps, where each gap is a maximal consecutive amino acid subsequence of the candidate peptide that is not covered by any tags. The difference between the mass of a gap and the size of its aligned segment of the spectrum is called mass offset for the gap. (Figure 1)

(v) PTM identification: For each gap of a tag chain, we find a set of PTMs that best interprets the gap. We first enumerate candidate sets of PTMs that correspond to the mass offset of each gap, and then select the best candidate set by comparing the partial theoretical spectra generated by each candidate set with the partial experimental spectrum of the gap.

## INPUT, OUTPUT AND PARAMETERS

### Input

MOD$^i$ requires users to input spectra, protein database and PTM database.

(i) *Spectra*. MOD$^i$ can take several different formats of spectra: ISB mzXML format (*.xml), Thermo Finnigan dta format (possibly compressed to a *.zip format),

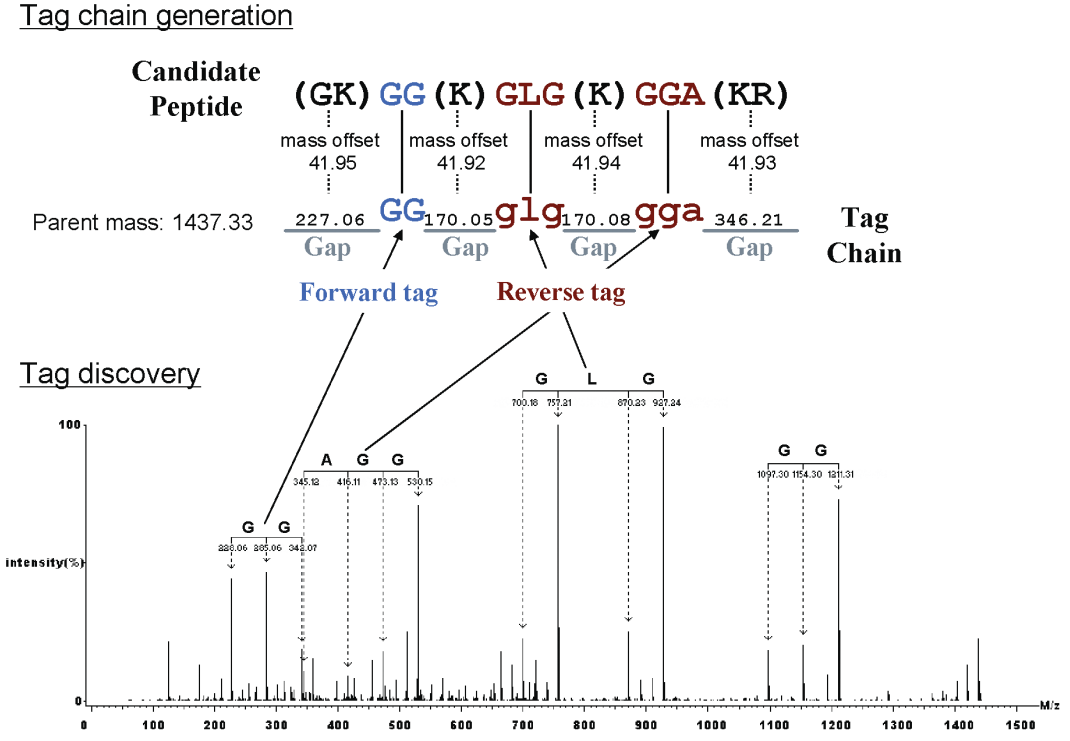


**Figure 1.** A tag chain '__GG_glg_gga__' for a spectrum from histone. A sequence of capital letters represents a forward tag, a sequence of small letters a reverse tag and each underline a gap. This tag chain consists of one forward tag 'GG', two reverse tags 'glg' and 'gga' and four gaps in between. A mass offset for a gap can be calculated by subtracting the mass of a gap from the size of its aligned segment of the spectrum. For example, the mass offset of the leftmost gap can be calculated as 41.947 Da, based on the size of aligned segment of 227.063(228.063 − 1) and the mass of the sequence 'GK' corresponding to the gap of 185.116 (Glycine: 57.021, Lysine: 128.095).

Micromass pkl format (*.pkl) and Mascot mgf format (*.mgf). For reliable interpretation of PTMs, we recommend users to input spectra with mass error tolerance <1 Da.

(ii) *Protein database*. Protein database should be in fasta format (*.fasta) and contain 20 or less protein sequences because a large protein database may produce bulky false positive results.

(iii) *PTM database*. MOD$^i$ can consider the entire unimod PTMs published in http://www.unimod.org. In addition, users can configure their tailored PTM list composed of selected unimod PTMs and user-defined PTMs. It is possible to save and load user-selected PTM lists.

## Parameters

Users can adjust MOD$^i$ parameters appropriately according to one's experimental conditions. The parameters include maximum number of missed cleavages, mass tolerance, precursor mass tolerance and enzymes used. In addition, users can fine-tune MOD$^i$ by fitting advanced parameters such as offset minimum/maximum value per gap, tag chain discard rate, minimum normalized intensity to consider, peak selection window size and minimum/maximum peaks in a window. Users can save parameters on a local host to re-use them for later analysis. A more detailed description of each parameter can be found in help pages of the MOD$^i$ website.

## Output

MOD$^i$ outputs a unidta file (*.unidta) and a unidrawing file (*.unidrawing). To guarantee random access, each file has a tailing offset list which indicates file offsets of input spectra. The unidta file is a set of input spectra and the unidrawing file is an XML formatted file which contains interpretations of input spectra.

## MassPective

MassPective has been developed to help users understand the interpretation from unidta and unidrawing files. MassPective is a graphical tool that displays MOD$^i$'s output in a user-friendly manner and helps users understand MOD$^i$'s output quickly. It shows each spectrum with ion-type annotation, candidate peptides for the spectrum, tag-chains for the candidate peptides and possible PTM interpretations for the tag-chains using graphical user interface. In addition, it enables users to perform additional manual *de novo* sequencing so that a MOD$^i$'s partial interpretation can be manually annotated with additional sequencing information.

Given a spectrum, MassPective displays MOD$^i$'s output in three tiled windows (candidate peptide list, detailed information, and spectrum windows) and a pop-up gap list window (Figure 2). Candidate peptide list window shows candidate peptides for the current spectrum and tag-chains of each candidate peptide. By selecting a candidate peptide or a tag chain, a user can see useful information such as score, offset list or gap list in the detailed information window. If a user clicks a tag chain, a pop-up window showing gap list appears. Gap list window shows combinations of PTMs that may occur in each gap and their scores and a user can select an interpretation of each gap so that the corresponding annotation in the spectrum window can be displayed. Spectrum window displays the selected spectrum together with various annotations a user selects. By clicking 

buttons in the toolbar, a user can see spectral alignment of y ion tags, b ion tags, theoretical y-ion peaks and theoretical b-ion peaks respectively.

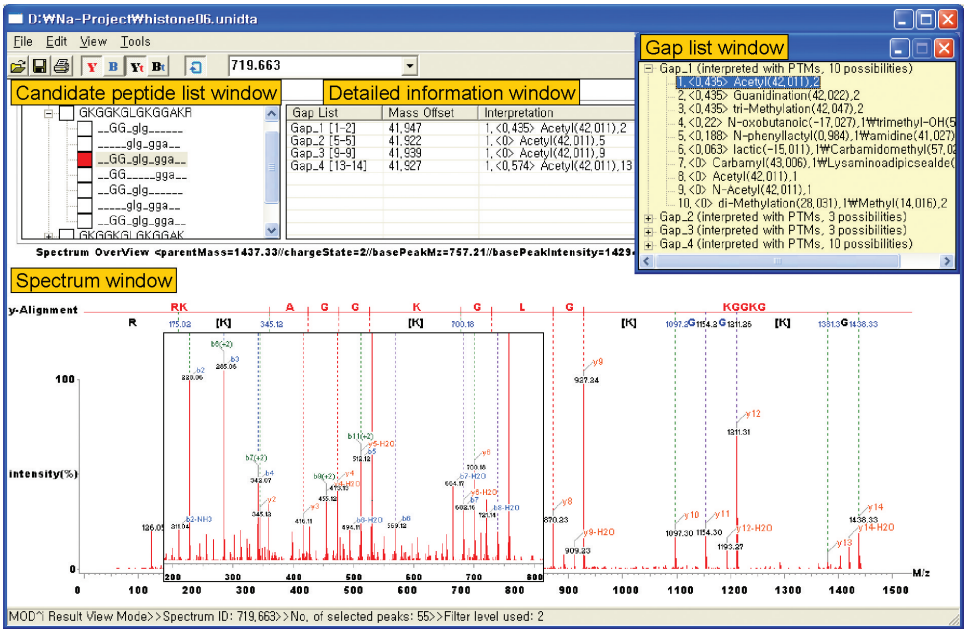


**Figure 2.** MassPective screen shot. This is the interpretation of a spectrum corresponding to the peptide 'GKGGKGLGKGGAKR' of histone. It shows spectral alignments of y ion tags, theoretical y-ion peaks, and their ion annotation in red and those of their b counterparts in blue.

| | A | C | D | E | F | G | H | I | J | K | L | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 | accession NO | Seq.Cov(%) | sub Mass | site | peptide sequence | score | spectrum NO | PTM | | | | |
| 4 | gi\|122098\|sp\|P02 | 13.59 | 1395.339/ 3 | 5-18 | GKGGKGLGKGGAKR | 22.44 | 466.113 | Acetyl(1) Acetyl(9) Methyl(13) di-Methylation(14) | | | | |
| 5 | | 13.59 | 1395.358/ 2 | 5-18 | GKGGKGLGKGGAKR | 19.66 | 698.679 | Acetyl(5) Acetyl(9) Acetyl(13) | | | | |
| 6 | | 13.59 | 1437.283/ 3 | 5-18 | GKGGKGLGKGGAKR | 20.65 | 480.094 | di-Methylation(1) Methyl(2) Acetyl(5) Acetyl(9) Methyl(13) di-Methylation(14) | | | | |
| 7 | | 13.59 | 1437.326/ 2 | 5-18 | GKGGKGLGKGGAKR | 22.44 | 719.663 | Acetyl(2) Acetyl(5) Acetyl(9) Acetyl(13) | | | | |
| 8 | | 11.65 | 1210.179/ 2 | 7-18 | GGKGLGKGGAKR | 19.79 | 606.09 | di-Methylation(1) Methyl(3) Acetyl(7) Acetyl(11) | | | | |
| 9 | | 15.53 | 1848.491/ 4 | 21-36 | KVLRDNIQGITKPAIR | 22.78 | 463.123 | di-Methylation(1) | | | | |
| 10 | | 11.65 | 1324.279/ 2 | 25-36 | DNIQGITKPAIR | 17.62 | 663.139 | | | | | |
| 11 | | 11.65 | 1324.272/ 3 | 25-36 | DNIQGITKPAIR | 10.98 | 442.424 | | | | | |
| 12 | | 12.62 | 1480.376/ 3 | 25-37 | DNIQGITKPAIRR | 14.84 | 494.459 | | | | | |
| 13 | | 10.68 | 1335.270/ 2 | 46-56 | RISGLIYEETR | 18.01 | 668.635 | | | | | |
| 14 | | 9.71 | 1179.184/ 2 | 47-56 | ISGLIYEETR | 16.75 | 590.592 | | | | | |
| 15 | | 13.59 | 1576.316/ 3 | 47-60 | ISGLIYEETRGVLK | 16.32 | 526.439 | | | | | |
| 16 | | 13.59 | 1576.346/ 2 | 47-60 | ISGLIYEETRGVLK | 8.6 | 789.173 | | | | | |
| 17 | | 13.59 | 1558.332/ 2 | 47-60 | ISGLIYEETRGVLK | 9.19 | 780.166 | Phospho+PL(9) | | | | |
| 18 | | 7.77 | 988.248/ 2 | 61-68 | VFLENVIR | 13.63 | 495.124 | | | | | |
| 19 | | 17.48 | 2103.396/ 3 | 61-78 | VFLENVIRDAVTYTEHAK | 20.15 | 702.132 | | | | | |
| 20 | | 10.68 | 1289.198/ 2 | 69-79 | DAVTYTEHAKR | 14.41 | 645.599 | | | | | |
| 21 | | 13.59 | 1609.342/ 3 | 80-93 | KTVTAMDVVYALKR | 28.62 | 537.447 | Oxidation(6) | | | | |
| 22 | | 13.59 | 1637.302/ 3 | 80-93 | KTVTAMDVVYALKR | 19.12 | 546.768 | di-Methylation(1) Oxidation(6) | | | | |
| 23 | | 11.65 | 1325.302/ 2 | 81-92 | TVTAMDVVYALK | 21.02 | 663.651 | Oxidation(5) | | | | |
| 24 | | 11.65 | 1309.221/ 2 | 81-92 | TVTAMDVVYALK | 17.34 | 655.61 | | | | | |
| 25 | | 12.62 | 1481.304/ 2 | 81-93 | TVTAMDVVYALKR | 21.59 | 741.652 | Oxidation(5) | | | | |
| 26 | | 12.62 | 1465.303/ 2 | 81-93 | TVTAMDVVYALKR | 21.19 | 733.652 | | | | | |
| 27 | | 12.62 | 1465.285/ 3 | 81-93 | TVTAMDVVYALKR | 14.34 | 489.428 | | | | | |
| 28 | | 12.62 | 1481.281/ 3 | 81-93 | TVTAMDVVYALKR | 19.58 | 494.76 | Oxidation(5) | | | | |
| 29 | | | | | | | | | | | | |
| 30 | | Protein Seq.Cov(%)=76.70 Seq.Len=103 Match.Len=79 | | | | | | | | | | |

**Figure 3.** A protein summary report generated by MassPective. It is a summary of the output of MOD$^i$ for each identified protein, containing protein information and sequence coverage for proteins which correspond to input spectra. Also, for each interpreted spectrum, it gives a submitted mass, matched peptide sites in protein sequence, match score, a spectrum identifier and identified PTMs.

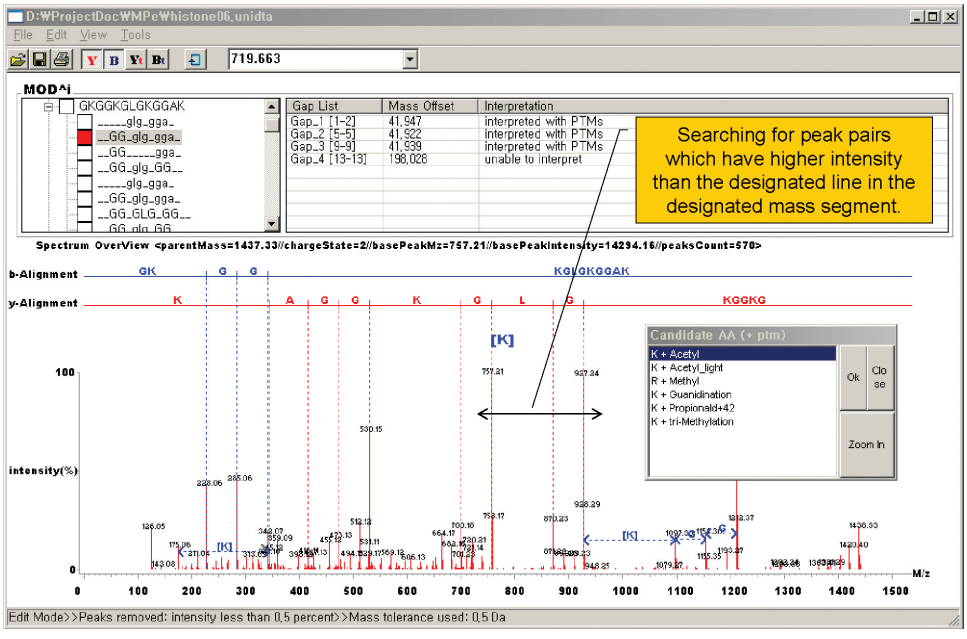


**Figure 4.** Manual sequencing by MassPective. If a user designates an area in the spectrum, a window pops up showing all possible combinations of one amino acid mass possibly with one PTM whose mass agrees with the difference of two peaks in the area with higher intensity than the intensity value of the designated line.

Another function of MassPective is to report protein summary in CSV (Comma-Seperated Values) file format (Figure 3) or save spectral image in *.bmp or *.jpeg format and print out the currently displayed spectrum.

MassPective also supports manual *de novo* sequencing with PTMs by identifying every amino acid (possibly with a PTM) of which mass corresponds to the mass difference of any two peaks in a designated spectral segment. (Figure 4)

## RESULTS

MOD$^i$ web server has been tested by four groups: a group that uses a nano-LC/MS–MS system consisting of an Ultimate HPLC system (LC Packings) for nano-LC and a Q-TOF Ultima Global mass spectrometer (Micromass) equipped with a nano-ESI source, a group that uses a Finnigan LTQ equipped with 'in-house' nano-ESI source, a group that uses Finnigan LCQ equipped with 'in-house' nano-LC system and a group that uses Applied Biosystems 4700 Proteomics Analyzer MALDI-TOF/TOF equipped with Agilent 1100 series capillary HPLC system.

Figure 5 shows how MOD$^i$ interprets a spectrum with multiple modifications successfully. MOD$^i$ interprets the spectrum as a peptide 'GKGGKGLGKGGAKR' with four acetylation sites at every Lysine in the peptide. As shown in Figure 5, MOD$^i$ first finds a tag chain __GG _glg_gga__, computes the mass offsets of the gaps, which are 41.95, 41.92, 41.94 and

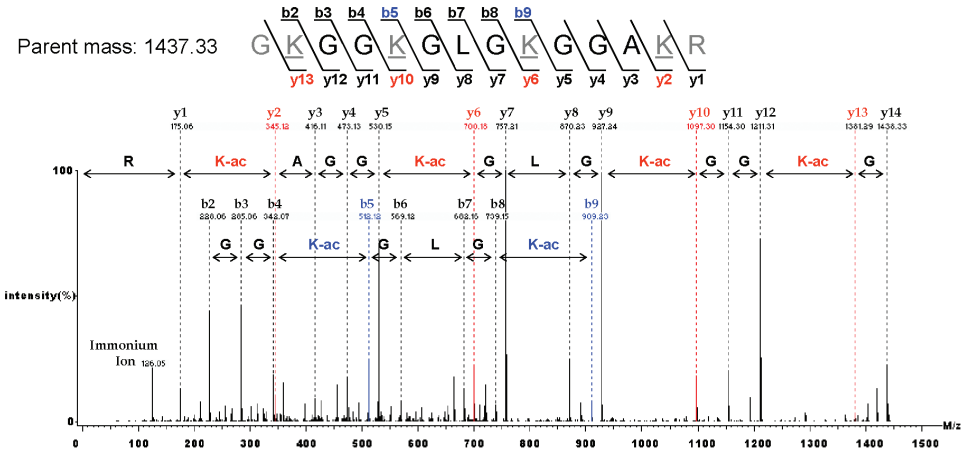


**Figure 5.** Interpretation for a spectrum with precursor ion 719.663 (2+) corresponding to a peptide 'GKGGKGLGKGGAKR' of histone. MOD$^i$ first finds a tag chain __GG _glg_gga__, computes mass offsets of the gaps, which are 41.95, 41.92, 41.94 and 41.93 Da, respectively (shown in Figure 1). Every Lysine is interpreted with an acetylation and the estimation is evaluated by a scoring scheme.

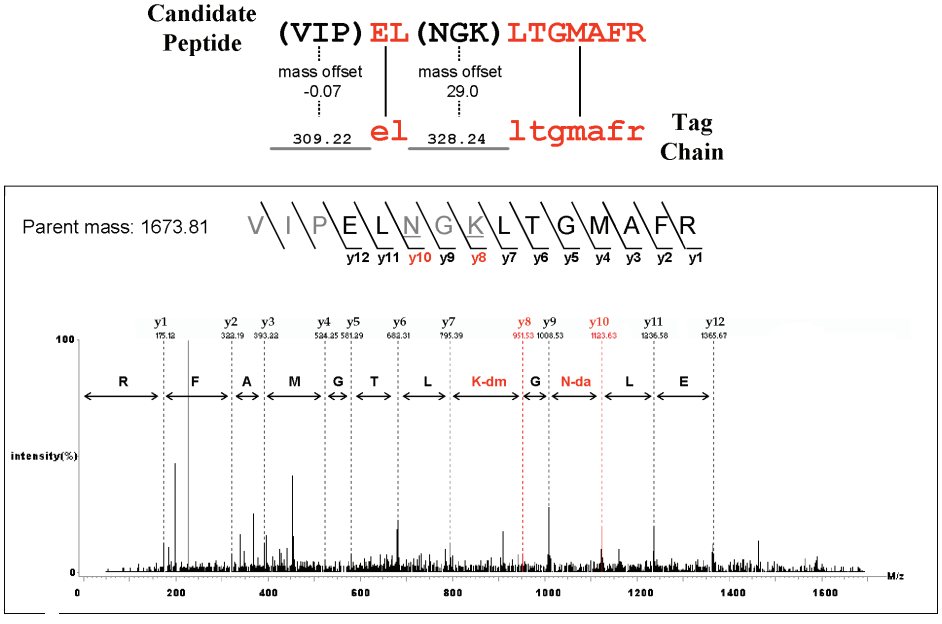


**Figure 6.** Shows that MOD$^i$ can find multiple PTMs in a single gap. It identifies a deamidation on 'N' and a di-methylation on 'K' in a gap 'NGK'.

41.93 Da, respectively, estimates four acetylation sites from the mass offsets and evaluates the estimation by a simple scoring scheme.

MOD$^i$ can find multiple PTMs in a single gap. In Figure 6, MOD$^i$ identifies a deamidation on 'N' and a di-methylation on 'K' in a gap 'NGK'. In general, considering various combinations of possible modifications requires prohibitively huge computational time as the numbers of possible modification sites and types increase. This example demonstrates how MOD$^i$ successfully manages the time complexity of modification analysis problem so that it can spend time on identifying multiple closely located PTMs in a peptide. This is possible because MOD$^i$ has already identified regions in a peptide that do not contain modified residues and are anchored on those sites before starting modification analysis for each gap. Thus, it narrowed down the search space for multiple PTMs to each gap region of a peptide, which is generally a lot shorter than the entire peptide length.

## DISCUSSION

Identifying multiple PTMs in a single gap mandates trying all possible combinations of feasible PTMs. This implies that as the number and the type of PTMs being considered grow, the number of combinations grows exponentially. This is the reason why most of the previous methods limit the number of PTM sites and PTM types being considered. However, we manage the computational complexity of PTM identification

innovatively and therefore MOD$^i$ can interpret a tandem mass spectrum when hundreds of modification types are considered and the number of potential PTMs in a peptide is large. MOD$^i$ is different from existing methods based on sequence tags in that it aligns multiple sequence tags to a candidate peptide and isolates regions that might include post-translationally modified amino acids, while most of the previous methods align a single tag to a candidate peptide and try to identify modifications over the entire peptide. By first fixing unmodified regions in a peptide and then interpreting potentially modified amino acids in relatively small regions in between these unmodified tags, MOD$^i$ can greatly reduce the search space for PTMs. In our experiments, MOD$^i$ demonstrates its power by identifying uncommon modifications and artefacts such as di-methylation, acrylamide adduct (propionamide), cysteine oxidation to cysteic acid and tryptophan oxidation to formylkynurenin. Such results are in accordance with the recent publication by Pevzner group where a variety of PTMs are reported (18).

Owing to the efficiency gain obtained by our method, MOD$^i$ runs in a reasonable amount of time. We tested MOD$^i$ on a web server of Intel Xeon 3.06 GHz dual processor with 2 GB memory, running Windows Server 2003. For a dataset of 5684 spectra, with 211 different PTM types downloaded from http://www.unimod.org and nine protein sequences, obtained from Mascot's protein ID results, in the Protein DB, it took ~360 s. When the same dataset was run with a database of 20 proteins (additional 11 random proteins from IPI human database), it took ~1070 s.

The current version of MOD$^i$ assumes that the number of candidate proteins is limited to 20 or less, and focuses on finding PTMs in tandem mass spectra. Candidate proteins can be easily identified using tandem mass spectra that do not contain PTMs using any of existing methods such as SEQUEST, Mascot or X!Tandem (7,8,19). In order to conduct protein identification on the same web server, instead of using separate software tools, we are planning to generalize the current version of MOD$^i$ so that it also includes the protein identification step.

## ACKNOWLEDGEMENTS

The authors thank Dr J. B. Kwon (Ewha Womans University) for providing purified histone from Hela cells. This work was supported by 21C Frontier Functional Proteomics Project from Korean Ministry of Science and Technology (FPR05A2-340 and FPR05A2-480) and through the Center for Cell Signaling Research (CCSR) at Ewha Womans University. Funding to pay the Open Access publication charges for this article was provided by 21C Frontier Functional Proteomics Project from Korean Ministry of Science and Technology (FPR05A2-340).

*Conflict of interest statement*. None declared.

## REFERENCES

1. Baenziger,J.U. (2003) A major step on the road to understanding a unique posttranslational modification and its role in a genetic disease. *Cell*, **113**, 421–422.
2. Wilkins,M.R. *et al*. (1999) High-throughput mass spectrometric discovery of protein post-translational modifications. *J. Mol. Biol*., **289**, 645–657.
3. Mann,M. and Jensen,O.N. (2003) Proteomic analysis of post-translational modifications. *Nat. Biotechnol*., **21**, 255–261.
4. Jensen,O.N. (2004) Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. *Curr. Opin. Chem. Biol*., **8**, 33–41.
5. Seo,J. and Lee,K.J. (2004) Post-translational modifications and their biological functions: proteomic analysis and systematic approaches. *J. Biochem. Mol. Biol*., **37**, 35–44.
6. Aebersold,R. and Mann,M. (2003) Mass spectrometry-based proteomics. *Nature*, **422**, 198–207.
7. Eng,J.K., McCormack,A.L. and Yates,J.R. (1994) An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database. *J. Am. Soc. Mass Spectrom*., **5**, 976–989.
8. Perkins,D.N., Pappin,D.J.C., Creasy,D.M. and Cottrell,J.S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551–3567.
9. Bafna,V. and Edwards,N. (2001) SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics*, **17** (Suppl. 1), 13–21.
10. Chen,T., Kao,M., Tepel,M., Rush,J. and Church,G.M. (2001) Dynamic programming approach to *de novo* peptide sequencing via tandem mass spectrometry. *J. Comput. Biol*., **8**, 325–337.
11. Frank,A. and Pevzner,P. (2005) PepNovo: *de novo* peptide sequencing via probabilistic network modeling. *Anal. Chem*., **77**, 964–973.
12. Ma,B., Zhang,K., Hendrie,C., Liang,C., Li,M., Doherty-Kirby,A. and Lajoie,G. (2003) PEAKS: powerful software for peptide *de novo* sequencing by tandem mass spectrometry. *Rapid Comm. Mass Spec*., **17**, 2337–2342.
13. Mann,M. and Wilm,M. (1994) Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem*., **66**, 4390–4399.
14. Tabb,D.L., Saraf,A. and Yates,J.R. (2003) GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. *Anal. Chem*., **75**, 6415–6421.
15. Tanner,S., Hongjun,S., Frank,A., Wang,L.C., Zandi,E., Mumby,M., Pevzner,P. and Bafna,V. (2005) InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal. Chem*., **77**, 4626–4639.
16. Matthiesen,R., Bunkenborg,J., Stensballe,A., Jensen,O., Welinder,K. and Bauw,G. (2004) Database-independent, database-dependent, and extended interpretation of peptide mass spectra in VEMS V2.0. *Proteomics*, **4**, 2583–2593.
17. Matthiesen,R., Trelle,M., Hojrup,P., Bunkenborg,J. and Jensen,O. (2005) VEMS 3.0: algorithms and computational tools for tandem mass spectrometry based identification of post-translational modifications in proteins. *J. Proteome Res*., **4**, 2338–2347.
18. Tsur,D., Tanner,S., Zandi,E., Bafna,V. and Pevzner,P. (2005) Identification of post-translational modifications by blind search of mass spectra. *Nat. Biotechnol*., **23**, 1562–1566.
19. Craig,R. and Beavis,R. (2003) A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Comm. Mass Spec*., **17**, 2310–2316.