Traceable Analysis of Multiple-Stage Mass Spectra through Precursor-Product Annotations

Hisayuki Horai[1,*], Masanori Arita[1,2,3,*], Yuya Ojima[1], Yoshito Nihei[1],
Shigehiko Kanaya[3,4] and Takaaki Nishioka[1]

[1]Institute for Advanced Biosciences, Keio University, Tsuruoka 997-0035, Japan
[2]Department of Computational Biology, Graduate School of Frontier Sciences,
The University of Tokyo, Kashiwa 277-8561, Japan
[3]RIKEN Plant Science Center, Yokohama 230-0045, Japan
[4]Laboratory of Comparative Genomics, Graduate School of Information Science,
Nara Institute of Science and Technology, Ikoma 630-0192, Japan
[*]Both authors contributed equally

Abstract: We present a wiki-based interface for multiple-stage mass spectra with molecular structures and their physicochemical properties. Spectra for 453 metabolites were measured on QqTOF-MS^n and their ion peaks were annotated with consideration of fragmentation patterns, especially bond cleavages. The resulting information was classified on wiki pages, where related molecular formulas and their relationships were likewise accumulated. Each page is rendered with search operation(s) using formulas as keys, and related information is automatically updated as database contents increase. Our data management model allows internet beginners to collaboratively input and organize information in a multi-user environment. The system, with links to our MassBank database (`http://massbank.jp/`), is available at `http://metabolomics.jp/wiki/Index:MassBank`.

1 Introduction

Metabolomics has become a standard technology in analyzing natural products [VBNS+07], and metabolite identification from mass spectra (MS) is a much investigated research topic. Identification from electrospray-ionization (ESI) spectra has been hampered by practical problems, however, because metabolites share similar molecular structures and physicochemical properties. First, there is no comprehensive database for ESI-MS. Fragmentation pattern of ESI has been considered machine-type dependent, and few research groups have attempted to accumulate and provide freely downloadable spectral information [Nat08, SRL+08, WKG+09]. Second, fragmentation rules have not been well understood compared to what has been accomplished in electron-ionization (EI) MS [McL59]. To overcome these difficulties, we have designed and implemented a distributed database called MassBank (`http://massbank.jp/`) for ESI spectra. Over 10 institutions joined our consortium and share spectra as well as data management systems. In this short article, we introduce a recent activity on our wiki-based interface to MassBank. Hereafter, MS stands for a deconvoluted set of ion peaks p whose m/z (mass to charge ratio) and scaled intensity (from 0 to 1000) can be accessed by functions $\mathrm{mass}(p)$ and $\mathrm{intensity}(p)$, respectively. We use these functions rather informally for elements other than ion peaks when the context is unambiguous. We also use the word 'mass' to refer to m/z hereafter.

2 Data Acquisition

2.1 Statistics of Mass Spectral Data

As of June 2009, twelve laboratories contribute their mass spectra to MassBank (see `http://massbank.jp/en/published.html`). The total number of ESI spectra is $> 10,000$ for over $1,500$ molecules with overlap. All records are accessible for free, and software programs are also available under the GNU General Public License. Because the overview of the database including supported search methods will be presented elsewhere, we focus on the analysis of precursor-product ion relationships here.

2.2 Peak Annotation

The current study used spectra of 453 metabolite standards measured on QqTOF-MS^{n} (Applied Biosystems Japan, Tokyo) at Keio University. Peaks were annotated with consideration of fragmentation patterns, especially bond cleavage. Let us assume that, for a standard compound M, a set of spectral peaks

$$\{P_M | \forall p \in P_M \ \mathrm{intensity}(p) > 5\}$$

is obtained. Our annotation process consisted of the following steps.

1. For each ion peak $p \in P_M$, find a molecular formula f_p that is a subset of molecular composition of M and whose mass is within 50 ppm from the $\mathrm{mass}(p)$. If not, remove p from P_M.
2. For each remaining $p \in P_M$, assign a connected molecular substructure m_p of M that corresponds to f_p. If such a structure is not found, then remove p from P_M.
3. For each remaining $p \in P_M$, find all peaks $q \in P_M$ such that its structure m_q can be obtained by a single fragmentation step (i.e. cleaving up to 2 bonds) from m_p. Output all pairs (p, q). This step tries to list precursor-product pairs only, not an arbitrary pair of fragments.

Assignments were manually checked using two commercial software programs: Mass Frontier (Thermo Fisher Scientific Inc., Waltham MA, USA) and ACD/MS Fragmenter (Fujitsu Inc., Kawasaki, Japan). Through this process, 1,483 different molecular formulas

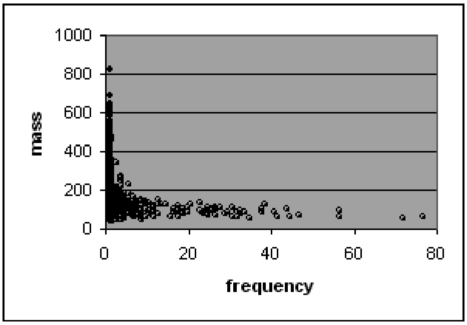

Figure 1: Statistics of Precursor-Product Ions

were identified, among which 5,557 precursor-product ion pairs were assigned. Note that the vast majority of peaks were left unannotated: we could annotate only 3,925 out of 130,246 peaks (3%, 9 peaks per spectra on average). Among the assigned molecular formulas, as much as 985 appeared only once. Most frequently appearing ions (and their frequency in parentheses) were C_6H_5 (100), C_5H_5 (80), C_3H_6N (71), and C_4H_3 (70), respectively. Not surprisingly, many ions were unsaturated fragments of C and H only and the mass of most frequent molecular formulas were under 200. Details of annotation results will be presented elsewhere (Ojima *et al.* in preparation).

3 Results and Discussion

3.1 Statistics of ESI Fragmentation

The notable character of annotated ions is the absence of clear correlation between frequency and mass (Figure 1). Usually MS contain more peaks of smaller mass, and such peaks are not informative for metabolite identification. However, except for some highly frequent masses (lower right dots in Figure 1), annotated ions showed low frequency and were distributed almost evenly up to mass 600. This indicates that annotated ions of smaller mass are equally as informative in structure prediction as those of larger mass. It must be noted that our annotation process is easier than the identification of metabolites [RSGB09]. Since we know the molecular structure in advance, we only need to traverse its possible, connected substructures (we did not consider a coupling of isolated fragments) in steps in Section 2.2. Although the enumeration problem of such substructures is NP-hard [HRM$^+$08], it is feasible for small metabolites under our strict condition.

3.2 Similarity Measure using Fragment Ions

The main purpose of annotation is to use it for metabolite identification from spectra in a future. For a fragment ion p to be used for identification, its frequency $\mathrm{freq}(p)$ should not be high. To identify informative ions, Shannon's information content $\mathrm{H}(p) = -\log(\mathrm{freq}(p))$ was used. Then, the similarity of two molecular structures M_1 and M_2 was defined as $\sum_{p \in P_{M_1} \cap P_{M_2}} \mathrm{H}(p)$. Note that all ions were equally weighted regardless of their mass by considering the result of Section 3.1.

3.3 Wiki-based Interface for Spectral Information

The information source of the analyses is essentially molecular formulas of product-precursor pairs, and all analyses are straightforward. Our novelty is not the analysis contents but the accessibility of processes and results on MediaWiki pages, i.e. traceability of research [AS08]. In other words, all operations are performed at the user-level on the wiki-based system and any user can reproduce, verify, and edit details just like editing a Wikipedia article.

Fragmentations observed in ESI-MS^{n} are a quite different type of chemical reactions from those observed in EI-MS; all the ions produced in EI-MS have odd-number electrons ("radical ions"), whereas those in ESI-MS have even-number electrons. Empirical rules that have been accumulated for the fragmentations in EI-MS are never applicable to the degradation reactions in ESI-MS^{n}. Only a few empirical rules are known in ESI-MS^{n} [Nak02], and we need to accumulate more rules on its degradation reactions. To provide a web-based forum of more chemical discussions among the mass spectral research communities, we provide a wiki-based platform linked with MassBank. Since the wiki part is used for annotation and discussion, not for actual spectra, default page contents should be as succinct as possible. For this reason, our wiki pages contain minimum possible information for drawing fragmentation scheme.

Let us explain an example at

`http://metabolomics.jp/wiki/MassBank:KOX00284`[1].

The page source is the simplest: identified molecular formulas and their precursor-product relationships only. At the time of page access, the minimum information is processed into a precursor-product table as its display, and the search for related pages is performed on demand. This molecule is Glycolate (MassBank ID: KOX00284), and its structural neighbor, Taurocolate (KOX00601), is automatically detected through the similarity of fragment ions. All results are always up-to-date even if other users add or remove data pages asynchronously. By checking such information, users can add comments and questions on precursor-product relationships as ordinary texts on wikis. Its advantage is obvious for a

[1] Currently, related pages are password-protected. To access, please login using the name "MassBank" and password "GCB2009".

collaborative project like our MassBank because page edit is open to all registered contributors.

The difference from conventional approaches such as Semantic Wiki is its simplicity [Ari09, HBB+09]. Data are plainly organized in a tabular form, and are exempted from site-specific predicates or manual addition of page links. Users' task is drastically alleviated since formatting and linking can be delegated to an embedded Lua programming language [Ier06], whose programs are also written inside wiki pages. Its computational power is restricted by running time and by closed I/O libraries to avoid web vandalism. Using Lua functionality, pages can be designed to minimize redundancy of data.

4 Conclusion

We implemented spectral annotation of precursor-product relationships on a MediaWiki based platform. All pages and Lua programs can be managed at the user-level, and this consequently guarantees traceability of research. Wiki users are also encouraged to leave references and traces of thinking in the annotation so that fragmentation rules can be later summarized from input information. A login account can be obtained on request to `massbank@iab.keio.ac.jp`.

Acknowledgments

This work is supported by JST-BIRD and Grant-in-Aid for Scientific Research on Priority Areas "Systems Genomics" from the Ministry of Education, Culture, Sports, Science and Technology, Japan. Authors also thank anonymous referees for their valuable comments.

References

[Ari09] Masanori Arita. A pitfall of wiki solution for biological databases. *Briefings in Bioinformatics*, 10(3):295–296, 2009.

[AS08] Masanori Arita and Kazuhiro Suwa. Search extension transforms Wiki into a relational system: A case for flavonoid metabolite database. *BioData Mining*, 1(1):7, 2008.

[HBB+09] Robert Hoehndorf, Joshua Bacher, Michael Backhaus, Sergio Gregorio, Frank Loebe, Kay Prufer, Alexandr Uciteli, Johann Visagie, Heinrich Herre, and Janet Kelso. BOWiki: an ontology-based wiki for annotation of data and integration of knowledge in biology. *BMC Bioinformatics*, 10(Suppl 5):S5, 2009.

[HRM+08] Markus Heinonen, Ari Rantanen, Taneli Mielikäinen, Juha Kokkonen, Jari Kiuru, Raimo A. Ketola, and Juho Rousu. FiD: a software for ab initio structural identification of product ions from tandem mass spectrometric data. *Rapid Communications in Mass Spectrometry*, 22(19):3043–3052, 2008.

[Ier06] Roberto Ierusalimschy. *Programming in Lua*. Lua.org, 2 edition, 2006.

[McL59] F. W. McLafferty. Mass Spectrometric Analysis. Molecular Rearrangements. *Analytical Chemistry*, 31(1):82–87, 1959.

[Nak02] H. Nakata. A Rule to account for mass shifts in fragmentations of even-electron organic ions in mass spectrometry. *Journal of the Mass Spectrometry Society of Japan*, 50(4):173–188, 2002.

[Nat08] National Institute of Standards and Technology. NIST Chemistry WebBook, 2008.

[RSGB09] Simon Rogers, Richard A. Scheltema, Mark Girolami, and Rainer Breitling. Probabilistic assignment of formulas to mass peaks in metabolomics experiments. *Bioinformatics*, 25(4):512–518, 2009.

[SRL$^+$08] T. R. Sana, J. C. Roark, X. Li, K. Waddell, and S. M. Fischer. Molecular formula and METLIN Personal Metabolite Database matching applied to the identification of compounds generated by LC/TOF-MS. *Journal of Biomolecular Techniques*, 19(4):258–266, 2008.

[VBNS$^+$07] Silas G. Villas-Boas, Jens Nielsen, Jorn Smedsgaard, Michael A. E. Hansen, and Ute Roessner-Tunali. *Metabolome Analysis: An Introduction*. Wiley - Interscience Series on Mass Spectrometry. Wiley-Interscience, 1 edition, 2007.

[WKG$^+$09] David S. Wishart, Craig Knox, An Chi Guo, Roman Eisner, Nelson Young, Bijaya Gautam, David D. Hau, Nick Psychogios, Edison Dong, Souhaila Bouatra, Rupasri Mandal, Igor Sinelnikov, Jianguo Xia, Leslie Jia, Joseph A. Cruz, Emilia Lim, Constance A. Sobsey, Savita Shrivastava, Paul Huang, Philip Liu, Lydia Fang, Jun Peng, Ryan Fradette, Dean Cheng, Dan Tzur, Melisa Clements, Avalyn Lewis, Andrea De Souza, Azaret Zuniga, Margot Dawe, Yeping Xiong, Derrick Clive, Russ Greiner, Alsu Nazyrova, Rustem Shaykhutdinov, Liang Li, Hans J. Vogel, and Ian Forsythe. HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Research*, 37(suppl_1):D603–610, 2009.