BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 2 2013, pages 290–291
doi:10.1093/bioinformatics/bts660

Databases and ontologies

Advance Access publication November 18, 2012

An application of a relational database system for high-throughput prediction of elemental compositions from accurate mass values

Nozomu Sakurai[1,*], Takeshi Ara[1], Shigehiko Kanaya[2], Yukiko Nakamura[2], Yoko Iijima[1,†], Mitsuo Enomoto[1], Takeshi Motegi[1], Koh Aoki[1,‡], Hideyuki Suzuki[1] and Daisuke Shibata[1]

[1]Department of Biotechnology Research, Kazusa DNA Research Institute, Kisarazu, Chiba 292-0818, Japan and [2]Department of Information Science, Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, Nara 630-0101, Japan

Associate Editor: Jonathan Wren

ABSTRACT

Summary: High-accuracy mass values detected by high-resolution mass spectrometry analysis enable prediction of elemental compositions, and thus are used for metabolite annotations in metabolomic studies. Here, we report an application of a relational database to significantly improve the rate of elemental composition predictions. By searching a database of pre-calculated elemental compositions with fixed kinds and numbers of atoms, the approach eliminates redundant evaluations of the same formula that occur in repeated calculations with other tools. When our approach is compared with HR2, which is one of the fastest tools available, our database search times were at least 109 times shorter than those of HR2. When a solid-state drive (SSD) was applied, the search time was 488 times shorter at 5 ppm mass tolerance and 1833 times at 0.1 ppm. Even if the search by HR2 was performed with 8 threads in a high-spec Windows 7 PC, the database search times were at least 26 and 115 times shorter without and with the SSD. These improvements were enhanced in a low spec Windows XP PC. We constructed a web service 'MFSearcher' to query the database in a RESTful manner.

Availability and implementation: Available for free at http://webs2.kazusa.or.jp/mfsearcher. The web service is implemented in Java, MySQL, Apache and Tomcat, with all major browsers supported.

Contact: sakurai@kazusa.or.jp

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on September 22, 2012; revised on November 3, 2012; accepted on November 5, 2012

1 INTRODUCTION

Identification or annotation of metabolite peaks detected by mass spectrometers (MSs) is one of the crucial steps in MS-based metabolomic studies (Saito and Matsuda, 2010). Accurate mass values measured by high-resolution MS instruments, such as Orbitrap MS and Time-of-Flight MS, are used in calculations of elemental compositions for the peaks and/or the product ions generated in multistage MS (MS^n) analyses (Feng and Siegel, 2007; Rojas-Chertó *et al.*, 2011), by which comprehensive identifications/annotations of metabolites have been reported (Beale and Sussman, 2011; Böttcher *et al.*, 2008; Iijima *et al.*, 2008).

Several software tools are available for calculation of elemental compositions: HR2 (Kind and Fiehn, 2007), AutoMFCalculator (Nakamura *et al.*, 2008), MassToFormula function of the Chemistry Development Kit (Steinbeck *et al.*, 2006) and Molecular Weight Calculator (http://www.alchemistmatt.com/). These tools generate all possible formulae under the condition of user-defined kinds and numbers of atoms (hereafter described as 'atom condition') and return elemental compositions that match the given mass value with the mass tolerance. An advantage of these tools is that users can change the atom conditions for each calculation. A disadvantage is that generations and evaluations of the same formulae cannot be avoided when similar mass values are repeatedly examined. This latent redundancy causes a limitation of the calculation rate when the tools are applied for the large number of mass data detected in metabolomic studies. To raise the throughput of metabolite annotations and then metabolomic data productions, acceleration of the calculation rate of elemental compositions is required.

We report here a database approach that remarkably improves the prediction of elemental compositions. Because the atom condition can be fixed for high-throughput surveys, all possible elemental compositions under the atom conditions are calculated and stored in a relational database system. The elemental compositions that match the given mass values with the mass tolerances are searched from the database. Therefore, the latent redundancy occurring in other tools is eliminated.

2 METHODS

We chose the following atoms and their maximum numbers in the calculations of all possible elemental compositions—C: 100, H: 200, O: 50, N: 10, P: 10, S: 10. The elemental compositions that fulfil the Senior and the Lewis valence rules were selected (Nakamura *et al.*, 2008). The formula weights of ~400 billion elemental compositions were calculated and stored in a MySQL database (Oracle Corporation) named 'ExactMassDB'. To evaluate the formula-searching performance of our database strategy, the HR2 tool, which is one of the fastest programs for

*To whom correspondence should be addressed.
†Present address: Department of Nutrition and Life Science, Kanagawa Institute of Technology, Atsugi, Kanagawa 243-0292, Japan.
‡Present address: Graduate School of Life and Environmental Sciences, Osaka Prefecture University, Sakai, Osaka 599-8531, Japan.

formula calculations that we tested (Supplementary Fig. S1), was selected for comparison. As HR2 filters some elemental compositions according to the Seven Golden Rules (Kind and Fiehn, 2007), another database 'ExactMassDB-HR2' (EX-HR2), which returns the same results as does HR2, was prepared. The search time was measured by an in-house Java program on Windows PCs. To evaluate the search time, we prepared lists of mass values that consist of 1000 mass peaks and 1000 MS^2 peaks detected by high-resolution MSs in practical analyses (practical mass lists). To help users query their mass values in ExactMassDB, we constructed a RESTful web service named 'MFSearcher' with Apache, Tomcat and MySQL. The web service also provides querying functions for other databases, e.g. EX-HR2 database, KEGG, PubChem, KNApSAcK, FlavonoidViewer and a database of possible linear polypeptides. The search time for the MFSearcher web service was measured by querying from several PCs in different countries. Details are described in the Supplementary Material.

3 RESULTS AND DISCUSSION

The search times of the pre-calculated MySQL database EX-HR2 were significantly shorter than those of HR2 by 109–123 times (Table 1). The creation of an index for the formula weight column of the MySQL table contributed to the reduction of the search time; however, it was still at the same level as HR2 at 5 ppm mass tolerance (Supplementary Table S1). The search time was largely dependent on the record numbers found, and hence on the mass tolerance (Supplementary Figs. S1–S4). A drastic improvement of the search rate was observed when the records of EX-HR2 were sorted by formula weight (Supplementary Table S1 and Fig. S4). It was speculated that the minimization of time for searching record data on the hard disk drive contributed to the reduction of the search time. This is proven by a further remarkable speed-up (488–1833 fold) when a solid-state drive (SSD) was applied (Table 1). These improvements were enhanced in a lower-spec Windows XP PC (Machine B) (Supplementary Tables S1–S3). The search time of HR2 was largely dependent on the evaluated formula number during the calculation (Supplementary Fig. S5). A parallel execution of HR2 by threads reduces the total search time for a mass list, although it requires higher CPU occupations (Supplementary Fig. S7). EX-HR2 was still 26–29 times faster than HR2 executed with threads in the Windows 7 PC (8 threads), and 55–61 times faster in the Windows XP PC (2 threads) (Table 1, Supplementary Table S2). In a practical usage of EX-HR2, further speed-up was expected by the cache of MySQL and the OS (Supplementary Table S3 and Figure S8). Searching with the MFSearcher web service needs a longer time than those on the PCs, but was still faster than HR2 as far as we tested, except for the search at 5 ppm mass tolerance from Peru (Supplementary Table S4).

A high-throughput prediction of elemental compositions by our database approach will improve comprehensiveness and throughput of annotations of metabolite peaks, and so will accelerate metabolomic data productions. For example, our approach boosts the analysis of elemental compositions of MS^n fragments, which is one of the perspective strategies towards semi-automatic *de novo* identification of metabolites (Rojas-Chertó *et al*., 2011). Other tools such as HR2 are nevertheless powerful for thorough investigation of a small number of metabolites by searching various atom settings. Our database approach can be complementary with all these other tools.

Table 1. Comparison of search times of EX-HR2 database with those of HR2 for practical mass lists consisting of 1000 metabolite peaks and 1000 MS^2 fragment peaks

ppm[a]	Search time (s)[b]				Fold			
	EX-HR2[c]	EX-HR2 (SSD)[d]	HR2	HR2 (thread)[e]	HR2/EX-HR2	HR2/EX-HR2 (SSD)	HR2 (thread)/EX-HR2	HR2 (thread)/EX-HR2 (SSD)
0.1	15.48	1.04	1900	449	123	1833	29	433
0.5	15.84	1.38	1911	449	121	1387	28	326
1	16.13	1.69	1901	449	118	1128	28	267
5	17.59	3.92	1912	449	109	488	26	115

Windows 7 PC (Machine A) was used (See Supplementary Material). [a]Mass tolerance given for elemental formula searching. [b]Ten sets of practical mass lists were searched, and average values of processing time are shown. [c]EX-HR2 database was indexed and sorted by the formula weight column. [d]Database files of EX-HR2 were placed on the solid-state drive (SSD). [e]Maximum of 8 threads of HR2 were executed in parallel.

ACKNOWLEDGEMENTS

We are grateful to all the people who kindly helped measure the search times of MFSearcher (Supplementary Material).

Funding: This work was partly supported by the New Energy and Industrial Technology Development Organization (NEDO, Japan) as part of a project named 'Development of Fundamental Technologies for Controlling the Material Production Process of Plants' [P02001], and by National Bioscience Database Center (NBDC) of Japan Science and Technology Agency (JST).

Conflict of Interest: none declared.

REFERENCES

Böttcher,C. *et al.* (2008) Metabolome analysis of biosynthetic mutants reveals a diversity of metabolic changes and allows identification of a large number of new compounds in Arabidopsis. *Plant Physiol.*, **147**, 2107–2120.

Beale,M.H. and Sussman,M.R. (2011) Metabolomics of Arabidopsis thaliana. In Hall,R.D. (ed.) *Annual Plant Reviews*. Vol. 43, Wiley-Blackwell, Chichester, UK, pp. 157–180.

Feng,X. and Siegel,M.M. (2007) FTICR-MS applications for the structure determination of natural products. *Anal. Bioanal. Chem.*, **389**, 1341–1363.

Iijima,Y. *et al.* (2008) Metabolite annotations based on the integration of mass spectral information. *Plant J.*, **54**, 949–962.

Kind,T. and Fiehn,O. (2007) Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics*, **8**, 105.

Nakamura,Y. *et al.* (2008) A tool for high-throughput prediction of molecular formulas and identification of isotopic peaks from large-scale mass spectrometry data. *Plant Biotechnol.*, **25**, 377–380.

Rojas-Chertó,M. *et al.* (2011) Elemental composition determination based on MS^n. *Bioinformatics*, **27**, 2376–2383.

Saito,K. and Matsuda,F. (2010) Metabolomics for functional genomics, systems biology, and biotechnology. *Annu. Rev. Plant. Biol.*, **61**, 463–489.

Steinbeck,C. *et al.* (2006) Recent developments of the chemistry development kit (CDK)—an open-source java library for chemo- and bioinformatics, *Curr. Pharm. Des.*, **12**, 2111–2120.