Published online 27 May 2011 *Nucleic Acids Research, 2011, Vol. 39, Web Server issue* ***W505–W510***
doi:10.1093/nar/gkr403

ARISTO: ontological classification of small molecules by electron ionization-mass spectrometry

Manor Askenazi[1,2,*] and Michal Linial[1]

[1]Sudarsky Center for Computational Biology, Department of Biological Chemistry, Hebrew University of Jerusalem, Israel and [2]Blais Proteomics Center, Dana-Farber Cancer Institute, Boston, MA, USA

Received February 15, 2011; Revised May 2, 2011; Accepted May 5, 2011

ABSTRACT

Gas chromatography–mass spectrometry (GC–MS) acquisitions routinely yield hundreds to thousands of Electron Ionization (EI) mass spectra. The chemical identification of these spectra typically involves a search protocol that seeks an exact match to a reference spectrum. Reference spectra are found in comprehensive libraries of small molecule EI spectra curated by commercial and public entities. We developed ARISTO (Automatic Reduction of Ion Spectra To Ontology), a webtool, which provides information regarding the general chemical nature of the compound underlying an input EI mass spectrum. Importantly, ARISTO can provide such annotation without necessitating an exact match to a specific compound. ARISTO provides assignments to a subset of the ChEBI (Chemical Entities of Biological Interest) dictionary, an ontology, which aims to cover biologically relevant small molecules. Our system takes as input a mass spectrum represented as a series of mass and intensity pairs; the system returns a graphical representation of the supported ontology as well as a detailed table of suggested annotations along with their associated statistical evidence. ARISTO is accessible at this URL: http://www.ionspectra.org/aristo. The system is free, open to all and does not require registration of any sort.

INTRODUCTION

In a typical Gas chromatography–mass spectrometry (GC–MS) study, analytical chemists will usually submit their Electron Ionization (EI) mass spectra to a software system, which attempts to identify each spectrum by matching it exactly to a reference spectrum from a well curated library (1). These libraries are generated by systematically running pure compounds under standard conditions to generate high-quality reference spectra. One of the largest libraries to date is the NIST 08 EI library (2) which contains spectra for 191 436 compounds (in its main library).

While the coverage of the NIST library ensures that for many of the common small molecules an exact match can be provided, if an exact match is not found, the researcher is left with two choices: (i) inspection of the nearest, imperfect hits and (ii) *de novo* elucidation of the compound structure, typically with the help of specialized software. Several software systems attempt to provide such *de novo* assignments including AMDIS (3,4) and Mass Frontier™ (ThermoFisher Scientific, Waltham, MA, USA). These software solutions ultimately rely on a chemical substructure identification (5) as a key component of their solution strategy.

In contrast, ARISTO does not explicitly incorporate substructure matching. Instead, it attempts to match spectra directly to averaged spectra corresponding to a formal standardized set of annotations (ChEBI, described below), without seeking an explicit substructure match with the query spectrum. ARISTO leverages the development in the chemical informatics community of a formal ontology which aims to cover Chemical Entities of Biological Interest—namely the ChEBI (6) dictionary. ARISTO was derived by matching 3000 spectra from the NIST 08 EI library to their respective ChEBI entries.

The combination of these two resources is used to derive a directed acyclic graph (DAG) of ~400 canonical annotation spectra corresponding to broad (>10 member compounds) ChEBI concepts. For each ChEBI concept (e.g. CHEBI:46686 or azaalkane), the corresponding EI spectra are averaged to generate a canonical annotation spectrum. A user provided query spectrum can then be scored against each of these averaged annotation spectra. From this score a probability of correctness is derived by comparing each score to a previously stored, comprehensive leave-one-out study which allows the system to estimate the precision of each prediction by linear interpolation against an empirical precision-recall plot

*To whom correspondence should be addressed. Tel: +972 2 6585425; Fax: +972 2 6586448; Email: manoras@cs.huji.ac.il

© The Author(s) 2011. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

(accessible in the results table). The scores are represented both graphically and in tabular form. The table contains links allowing the user to inspect the query spectrum aligned with the associated annotation spectrum (in a so-called mirror plot) as well as precision-recall plots and receiver operating characteristic (ROC) curves for the training data. The table also contains the description of each ChEBI concept as well as a link to the associated entry on the ChEBI website. The ontology supported by ARISTO is expected to expand in parallel to growth in the ChEBI annotation scheme.

We propose ARISTO as a potential complementary approach in cases where exact spectral matches are unavailable and furthermore, in cases where approximate matches fail to support any obvious conserved substructure. A description of the implementation will be provided herein, along with example analyses and future directions.

RATIONALE AND IMPLEMENTATION

ARISTO is built on ChEBI, which is a relatively recent ontology of chemical entities with relevance to biological research. Though many other resources exist, which address the space of small-molecules, we focused on ChEBI exclusively due to its commitment to a comprehensive and fully hierarchical ontology. Broadly speaking ChEBI can be understood as a hierarchical tree of nodes (actually a DAG), each with a ChEBI ID (e.g. CHEBI:38728) and name (e.g. monocrotophos) as well as descriptions, synonyms, cross-references to other databases and detailed chemical structures where appropriate. The lower levels of the hierarchy, the 'leaves' of the DAG, typically correspond to specific compounds whereas internal nodes mostly correspond to increasingly general concepts in chemistry, culminating in 'chemical entity' at the root of the tree (e.g. CHEBI:24431). As is the case with the well-known gene ontology (GO) annotation scheme (7), most edges in ChEBI correspond to 'is a' edges, though chemistry specific relations such as 'is enantiomer of' and 'has parent hydride' are also supported.

While ChEBI has experienced phenomenal growth in recent years, it remains committed to a definitive and thorough annotation of all its entries. In support of this goal, ChEBI uses a 'rated' annotation scheme similar to the one in use by UniProt. Whereas UniProt distinguishes its manually curated (Swiss-Prot) entries, ChEBI highlights the fully curated subset of its database (so called, 3 star entries). ARISTO relies only on these, highest rated entries, and consequently, its growth is projected to be limited primarily by the rate of ChEBI's manual annotation efforts (in addition to limiting ARISTO's growth rate, ChEBI also limits ARISTO to essentially un-derivatized molecules since derivatized forms are not often explicitly represented in chemical ontologies). Furthermore, to ensure robustness, we have opted to focus exclusively on the subset of ChEBI made entirely of 'is a' edges and internal nodes for which we could find at least 10 compounds with EI-spectra.

All the spectra used as a foundation for the ARISTO system were extracted from the NIST 08 EI library, which contains 191 436 reference spectra in its main library. With each spectrum there is a structure provided as well as a compound name along with common synonyms and external accession numbers. While the compound nomenclature does not consistently correspond to any canonical nomenclature, the structure would be, in principle, sufficient to derive an InChiKey (8,9) identifier to match with the one systematically provided by the ChEBI ontology.

The matching of EI spectra to broad annotations and chemical identifiers is challenging and prone to severe inconsistency. To overcome such pitfalls, the following matching-guidelines were considered: (i) The chemical structures provided by the NIST library do not contain stereochemistry and therefore are only guaranteed to yield the first 14 characters of the InChiKey correctly for any given compound. Recall that the first 14 letters of any InChiKey e.g. WHWZLSFABNNENI-OAHLLOKOSA-N for S-epinastine (CHEBI:51036) versus WHWZLSFABNNENI-HNNXBMFYSA-N for R-epinastine (CHEBI:51035), are produced by hashing the connectivity layer in the original InChi (and are therefore equivalent for these two compounds). (ii) While enantiomers almost always produce identical mass-spectra, the remaining characters of the InChiKey do provide information on isotopic composition, which can impact the spectrum. Ultimately, in order to minimize the chance of an incorrect match, especially in a first version of the system, the mapping between NIST and ChEBI was achieved not by InChiKey, but rather by CAS number. CAS was used as a cross-referenced accession that unifies many of the NIST and ChEBI entries. The resulting basis-set of annotated spectra available to the ARISTO system was therefore reduced from a potential ~200 000 spectra in the NIST 08 EI library, through a maximum of ~26 000 ChEBI entries to a CAS-supported matching of 3000 ChEBI-annotated EI spectra. In future versions of the system, we will attempt a more extensive matching in order to support a broader ontology. This will be achieved by matching on the first 14 characters of the InChiKey ('connectivity hashing').

Having matched spectra to ChEBI concepts we applied an extremely simple 'learning' technique; we averaged the spectra per concept. This approach was selected for two main reasons: averaging replicate spectra constitutes a valid step in most library curation pipelines, so in a sense ARISTO is simply treating spectra from different compounds as replicates of the higher-level concept being represented. Second, spectral averaging simplifies the process of cross-validation since there is no need to 'retrain' when removing a spectrum from the training set. A simple weighted subtraction simulates the elimination of the spectrum from the input.

Given a query spectrum and the averaged training data, ARISTO can generate a score for each concept by applying a simple dot-product to the query and the representative average spectrum for the concept. In order to translate scores into probabilities, we conducted a complete leave-one-out cross-validation study from which we generated ROC curves and precision-recall curves per concept. ARISTO reports the area under the curve (AUC) of the ROC curves to help users filter out concepts, which cannot be effectively recognized by our

simple dot-product against-average-spectrum strategy. Based on numerous tests (data not shown), the users are encourage to ignore entries with AUC < 0.8. ARISTO also reports the precision-recall curve, which it uses to translate scores to probabilities on a per concept basis. While it is evident that more complex learning schemes are likely to yield better coverage in terms of the ChEBI concept-space, we were able to find 'learnable' concepts even with this very basic strategy (Figure 1). Indeed, these concepts correspond to classes of compounds which can be identified visually by experienced researchers that are trained in the elucidation of structure from mass-spectrometry data (such compound classes include, e.g. long-chain fatty acids, steroids, etc.).

TEST CASE ANALYSES

The system is extremely easy to use in that it requires only one element of input, namely, a nominal mass spectrum specified as a series of mass and intensity pairs, optionally separated by new lines. The user pastes the spectrum into the provided input text box and activates the 'Ontologize' button. Candidate spectra will most likely be generated by tools such as SpectConnect (10) which can indentify reproducibly detected spectra that, while resisting identification, seem to be instrumental in discriminating e.g. biological conditions of interest. In Figure 2A, the input field contains a spectrum from a lauric acid derivative (available as example 5 in the Examples tab). Note that the example spectra are all members of a randomly selected test set comprised of 32 spectra removed prior to the learning phase and made available in the Batch Mode tab of the ARISTO website (see description below).

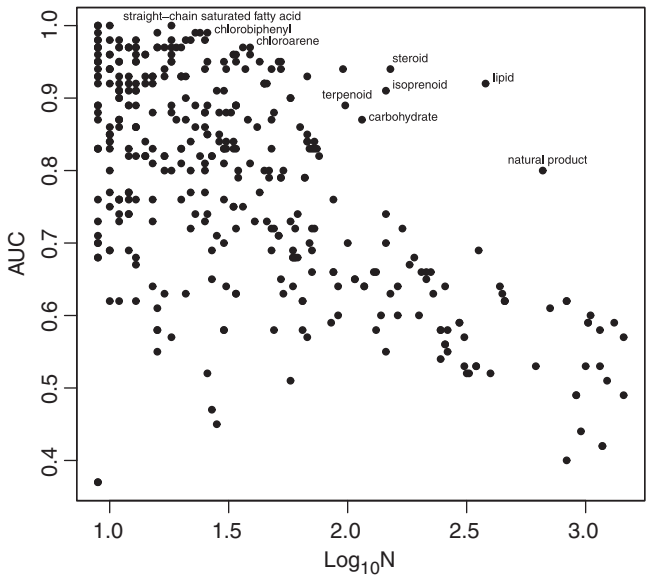

Figure 1. Learning higher level ChEBI concepts by a naive process of spectral averaging. The 'learnable' ChEBI concepts (AUC>0.8) include long-chain fatty acid and steroid which are considered to be 'easily recognizable' by mass spectrometry experts while other broad categories such as natural product are relatively unexpected. Note that with increasing compound coverage some of the broader categories (N>500) may decrease in their AUC while categories that currently lack adequate spectral support (N < 10) are likely to cross the threshold of learnability.

The system returns a complete assessment of whether the spectrum matches one of 388 chemical concepts. The results are provided in both graphical form (Figure 2B) and tabular form (Figure 2C). The directed acyclic graph represents the subset of the ChEBI ontology currently supported by ARISTO, where each node corresponds to a ChEBI concept color-coded by the probability that it is a valid annotation for the query spectrum (ranging from P = 0.0 in red to P = 1.0 in green). Edges in the DAG correspond to 'is-a' relationships. The size of each node is proportional to its AUC in the comprehensive leave-one-out analysis mentioned previously. Hovering over the node produces a tooltip specifying the node name and clicking on the node brings the user to the relevant ChEBI webpage. The tabular results contain sortable columns for the dot-product scores, the ChEBI node name, the number of compound spectra used to generate the representative average spectrum for the node, the AUC for the concept, as well as the estimated precision of the assignment. In addition, there is a link per annotation which produces a mirror plot (Figure 2D) allowing the visual comparison of the average annotation spectrum (top in blue) with the query spectrum (bottom in red) as well as links to plots showing the data collected during the leave-one-out analysis with an additional red data point corresponding to the score produced for the query spectrum. When a user submits a spectrum from the examples provided by the website itself, an additional column is appended to the table showing the correctness of the call (according to release 73 of the ChEBI ontology made in October 2010). By default, the table is filtered to include only annotations with an AUC > 0.8 and they are sorted according to their estimated precision ('Est. Precision' column), which is the preferred figure of merit by which to investigate ARISTO's output. This is because the significance of the raw dot-product score ('score' column) is dependent, among other factors, on the size of the category ('N' column). The AUC and precision-recall plots are particularly useful in this regard, since they allow the user to see the empirical basis for the translation from raw score to estimated precision. The user can see every element used in the learning phase and effectively compare the dot-score of the query spectrum to all the data points collected during the leave-one-out learning phase (the query data point is colored red).

In the example illustrated in Figure 2 (corresponding to example #5 in the examples page of the ARISTO website) we can see that straight-chain saturated fatty acid (CHEBI:39418) corresponds to the node with the strongest absolute score and is one of the four nodes showing maximal estimated precision (in the full, unfiltered table). However, two of these four predictions correspond to extremely broad and uninformative categories that also have relatively weak AUCs (they are in fact not visible in the default, filtered version of the results table). The entry for CHEBI:39418 is based on 19 compounds (N = 19) and an extremely high AUC. Consequently, CHEBI:39418 is selected as the more specific and well

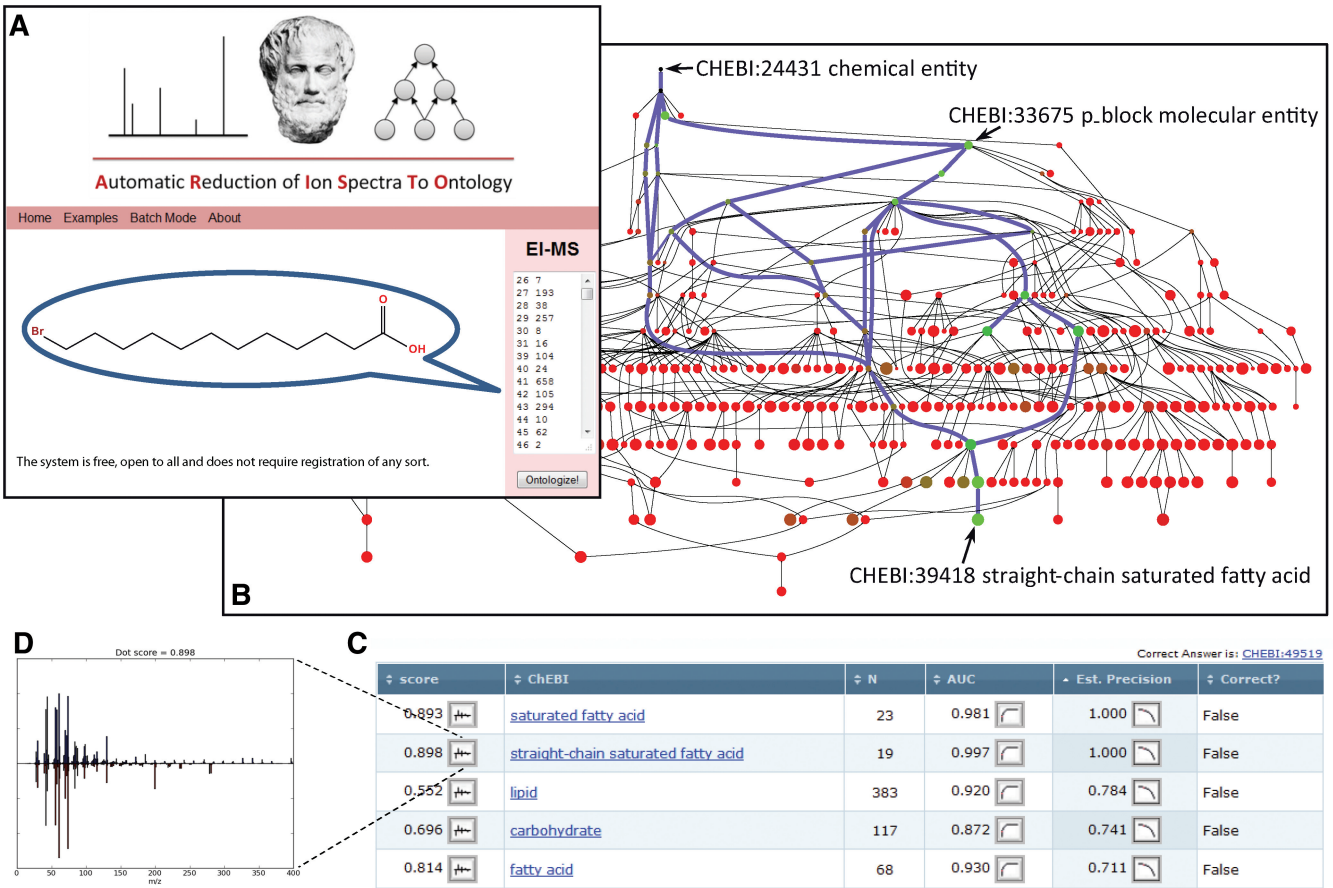

Figure 2. Screenshots from ARISTO's web interface. (**A**) The user loads a mass spectrum corresponding to an unknown compound—in this example the structure is known (CHEBI:49519) and corresponds to example spectrum #5 in the Examples section of the website. The resulting report produced by ChEBI shows: (**B**) a DAG representing the estimated precision and AUC for each concept represented as the color and size of each concept-node along with its location within the supported ontology (an induced subgraph is highlighted in blue which links all ChEBI annotations passing ARISTO's default filter back to the DAG's root node) and (**C**) a tabular report showing the estimated precision of each prediction, the correctness of the call (marked as True/False and available only for the example input provided by the website) along with image-links to more supporting information. The first image-link yields a mirror plot (**D**) showing the match between the average annotation-spectrum (top in blue) for a given annotation, in this case straight-chain saturated fatty acid (CHEBI:39418), and the query spectrum (bottom in red). See text for more details.

discriminated category. By clicking on the image links in the AUC column and estimated precision columns, the user can inspect the precision-recall and ROC curves for CHEBI:39418 and confirm that a score of 0.898 corresponds to a high-confidence identification. CHEBI:39418 is indeed a useful characterization of the input compound which would benefit an experimentalist interested in the chemical nature of the compound. It is noteworthy that this association is provided despite the fact that, strictly speaking, according to ChEBI, our input compound (CHEBI:49519) cannot be linked to the CHEBI:39418 category exclusively through a series of 'is-a' links (hence the False entry in the 'Correct?' column), rather it is necessary to go through a 'has-functional-parent' link to lauric acid (CHEBI:30805). While future versions of ARISTO may leverage these additional link-types explicitly, it is heartening that a version of ARISTO trained only on the 'is-a' subset of ChEBI can recover such relationships.

Example #6 (not shown in Figure 2 but available in the examples page) corresponds to a compound from the MassBank (11) site: 3alpha,6alpha-Dihydroxy-5alpha-cholan-24-oic acid Methyl ester which is not incorporated into the ChEBI ontology (and hence was never introduced into ARISTO) and is derived (12) from hyodeoxycholic acid (CHEBI:52023); when submitted to ARISTO this spectrum is correctly annotated as a 3-hydroxy steroid (CHEBI:36834). Note that the tabular report for this spectrum does not contain a 'Correct?' column since this compound has not yet been annotated by ChEBI. It is nevertheless characterizable by ARISTO.

Example #7 corresponds to Apoatropine (as in Example #6, this compound is not currently represented in ChEBI). The spectrum originates from the analysis of Datura inoxia seed extracts conducted by the Interdepartmental Equipment Unit of the Faculty of Agriculture, Food and Environment. This noisy GC–MS spectrum proved impossible to characterize using NIST's MS-Search

program and the NIST08 library: the correct structure was #36 on the list of suggested structures and only 1 of the top 10 identifications even contained a Nitrogen atom. In contrast, ARISTO correctly and confidently (estimated precision = 1.0) characterized the spectrum as containing an azabicycloalkane (CHEBI:38295).

A final test case is provided in the Batch Mode tab of the system where the user can upload a set of 32-spectra specifically excluded from the learning phase of the algorithm. The user can use this data set to get a sense of the system's coverage (for what percent of typical input spectra will the system issue a prediction) and whether or not the estimated precision is correct. Under the default filtering (which allows predictions with a minimal estimated precision of 0.5, i.e. with an expected failure rate of 50%) the system returns 33 predictions for 13 of the compounds (or 40% of the input data). Of these predictions 17 were correct corresponding to 51% of the predictions as expected under the default filtering. A stricter filtering at 0.75 yields 15 predictions of which 12 are correct, corresponding again to a well-calibrated 80% of the responses. However, the coverage has now dropped to only 6 compounds or ~19% of the input data (in fact, the user may recognize some of these compounds as examples 1–5 in the Examples tab). The coverage of the system is expected to increase as it grows beyond the initial basis-set of 3000 spectra. Despite the relatively low coverage, it is again heartening to see that the system makes correct predictions for every compound except for CHEBI:49519 where, as described previously, the predictions are indeed meaningful and the annotation in the 'Correct?' column is set to False essentially due to the fact that ARISTO currently considers only 'is-a' links in the ChEBI dictionary.

CONCLUSIONS AND FUTURE DIRECTION

ARISTO represents an existence proof for the idea of direct chemical characterization of mass spectra without explicit substructure matching. We expect ARISTO to be used primarily by experimentalists and to support the growing field of chemical informatics. ARISTO aims to close the gap between the enormous scale of small molecule spectral libraries and the nascent, relatively poor coverage of chemical ontologies associated with the small molecule concept space. The EI spectra are used as connectors between these spaces with the hope to circumvent the need for substructure identification. Recall that often the identity of a substructure is a poor indicator for the biological activity of the small molecule or its biological relevance.

Clearly, there is much room for improvements. As described earlier, the spectral basis set of the system can in principle be increased quite dramatically by shifting from a CAS number-based matching between the NIST EI library and the ChEBI ontology to a looser matching based on the connectivity layer of the InChiKey. Such a mapping is bound to generate false matches, however it is likely that, while the mismatched pairs will be chemically distinct (e.g. they may be enantiomers), with respect to their EI fragmentation they are likely to be legitimately considered as equivalent. Using this alternative mapping, we expect to cover most of the ChEBI ontology and more importantly, we can automatically track its growth.

In addition to improvements in coverage of chemical space, there is room for increased complexity in the algorithms used to classify spectra. The current version of ARISTO aims specifically to test the lower bound of algorithmic complexity, which is why it is based on a simple dot-product against each average class spectrum. More sophisticated machine learning algorithms are likely to increase the set of learnable concepts (e.g. algorithms that intelligently leverage the many TMS-derivatives present in the GC-MS centric NIST 08 EI library).

ACKNOWLEDGEMENTS

The authors thank members of the Linial Lab for helpful discussions; Dr Steve Stein, Dr Anzor Mikaia and Dr Edward White for their advice and for providing a text-formatted version of the NIST 08 EI library; Dr Julius Ben-Ari for invaluable feedback and the spectrum corresponding to Apoatropine; Zina Muzikansky for her support during the preparation of this manuscript.

FUNDING

Binational Science Foundation (BSF) (2007-219); Prospects consortium [EU FRVII]. Funding for open access charge: Prospects consortium (EU FRVII).

Conflict of interest statement. None declared.

REFERENCES

1. Ausloos,P., Clifton,C.L., Lias,S.G., Mikaya,A.I., Stein,S.E., Tchekhovskoi,D.V., Sparkman,O.D., Zaikin,V. and Zhu,D. (1999) The critical evaluation of a comprehensive mass spectral library. *J. Am. Soc. Mass Spectrom.*, **10**, 287–299.
2. Stein,S.E. and Scott,D.R. (1994) Optimization and testing of mass spectral library search algorithms for compound identification. *J. Am. Soc. Mass Spectrom.*, **5**, 859–866.
3. Stein,S.E. (1999) An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. *J. Am. Soc. Mass Spectrom.*, **10**, 770–781.
4. Stein,S.E. (1995) Chemical substructure identification by mass spectral library searching. *J. Am. Soc. Mass Spectrom.*, **6**, 644–655.
5. Gan,F., Yang,J.-h. and Liang,Y.-z. (2001) Library search of mass spectra with a new matching algorithm based on substructure similarity. *Anal. Sci.*, **17**, 635–638.
6. de Matos,P., Alcantara,R., Dekker,A., Ennis,M., Hastings,J., Haug,K., Spiteri,I., Turner,S. and Steinbeck,C. (2009) Chemical entities of biological interest: an update. *Nucleic Acids Res.*, **38**, D249–D254.
7. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, **25**, 25–29.
8. Heller,S. and McNaught,A. (2010) The status of the InChI project and the InChI trust. *J. Cheminf.*, **2**, P2.
9. Kidd,R. (2009) Changing the face of scientific publishing. *Integrative Biology*, **1**, 293.

10. Styczynski,M.P., Moxley,J.F., Tong,L.V., Walther,J.L., Jensen,K.L. and Stephanopoulos,G.N. (2007) systematic identification of conserved metabolites in GC/MS data for metabolomics and biomarker discovery. *Anal. Chem.*, **79**, 966–973.
11. Horai,H., Arita,M., Kanaya,S., Nihei,Y., Ikeda,T., Suwa,K., Ojima,Y., Tanaka,K., Tanaka,S., Aoshima,K. *et al.* (2010) MassBank: a public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.*, **45**, 703–714.
12. Iida,T., Tamaru,T., Chang,F.C., Niwa,T., Goto,J. and Nambara,T. (1993) Potential bile acid metabolites. 20. A new synthetic route to stereoisomeric 3,6-dihydroxy- and 6-hydroxy-5 alpha-cholanoic acids. *Steroids*, **58**, 362–369.