Contact Information

Postal address:

Department of Computer Science
P.O.Box 68 (Gustaf Hällströminkatu 2b)
FIN-00014 University of Helsinki
Finland

URL: http://www.cs.helsinki.fi

Telephone: +358 9 1911
Telefax: +358 9 191 51120

Series of Publications C, Report C-2010-40
Helsinki 2010
Preface

Mass spectrometry (MS) became a standard tool for identifying metabolites in biological tissues, and metabolomics is slowly acknowledged as a legitimate research discipline for characterizing biological conditions. The computational analyses of metabolomics, however, lag behind compared with the rapid advances in analytical aspects for two reasons. First is the lack of standardized data repository for mass spectra: each research institution is flooded with gigabytes of mass-spectral data from its own analytical groups and cannot host a world-class repository for mass spectra. The second reason is the lack of informatics experts that are fully experienced with spectral analyses. The two barriers must be overcome to establish a publicly free data server for MS analysis in metabolomics as does GenBank in genomics and UniProt in proteomics.

The workshop brought together bioinformaticians working on mass spectral analyses in Finland and Japan with the goal to establish a consortium to freely exchange and publicize mass spectra of metabolites measured on various platforms computational tools to analyze spectra spectral knowledge that are computationally predicted from standardized data.

This book contains the abstracts of the presentations given in the workshop. The programme of the workshop consisted of oral presentations from Japan and Finland, invited lectures from Steffen Neumann (Leibniz Institute of Plant Biochemistry), Matej Orešič (VTT), Merja Penttilä (VTT) and Nicola Zamboni (ETH Zürich) as well as free form discussion among the participants.

The event was funded by Academy of Finland (grants 139203 and 118653), Japan Society for the Promotion of Science (JSPS Japan-Finland Bilateral Seminar Program 2010) and Department of Computer Science University of Helsinki.

We would like to thank all the people contributing to the technical programme and the sponsors for making the workshop possible.
Organizers
Masanori Arita (University of Tokyo, Japan)
Juho Rousu (University of Helsinki, Finland)
Markus Heinonen (University of Helsinki, Finland)

Sponsors
Academy of Finland, grants 139203 and 118653 (Algodan)
Japan Society for the Promotion of Science
Department of Computer Science, University of Helsinki

Contributors
Masanori Arita
Markus Heinonen
Kensuke Nakamura
Steffen Neumann
Takaaki Nishioka
Matej Orešič
Merja Penttilä
Juho Rousu
Masahiro Sugimoto
Ken Tanaka
Nicola Zamboni
## Workshop Programme

**Thursday October 28, 2010**

*Venue: Exactum lecture room D122, Kumpula Campus, University of Helsinki*  
*Gustaf Hällströmin katu 2b, Helsinki, Finland*

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00-9:30</td>
<td>Registration and coffee</td>
</tr>
<tr>
<td>9:30-9:45</td>
<td>Opening of the workshop</td>
</tr>
<tr>
<td>9:45-10:15</td>
<td>Session 1, Chair: Juho Rousu</td>
</tr>
</tbody>
</table>
| 9:45-10:15  | “MassBank: A public repository of mass spectral data”  
**Takaaki Nishioka** |
| 10:15-10:45 | “Energy-based identification of molecular fragments in tandem mass spectrometry”  
**Markus Heinonen** |
| 10:45-11:15 | Break & discussion                                                        |
| 11:15-11:45 | Session 2, Chair: Masanori Arita                                          |
| 11:15-11:45 | “Circulating information and knowledge through a cyber-infrastructure”   
**Masanori Arita** |
| 11:45-12:30 | Invited Talk: “Opportunities and bottlenecks of high-throughput metabolomics”  
**Nicola Zamboni** |
| 12:30-14:30 | Lunch break                                                               |
| 14:30-15:00 | “Algorithms and Machine Learning for Small Molecules”  
**Juho Rousu** |
| 15:00-15:30 | “Species-metabolite relation database KNApSAcK and its multifaceted retrieval system KNApSAcK Family”  
**Kensuke Nakamura** |
| 15:30-16:00 | Coffee break & discussion                                                 |
| 16:00-16:45 | Invited Talk: “Processing of mass spectrometry based molecular profile data”  
**Matej Orešič** |
| 16:45-17:00 | Open discussion                                                           |
## Friday October 29, 2010

Venue: Exactum lecture room D122, Kumpula Campus, University of Helsinki

Gustaf Hällströmin katu 2b, Helsinki, Finland

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00</td>
<td>Registration and coffee</td>
</tr>
<tr>
<td>9:30-10:15</td>
<td>Invited Talk: “Spectral libraries and computational mass spectrometry – A symbiosis?” Steffen Neumann</td>
</tr>
<tr>
<td>10:15-10:45</td>
<td>“Analysis of biosynthetic fluctuations of cultured Taxus seedlings using LC-IT-TOF MS and a metabolomic approach” Ken Tanaka</td>
</tr>
<tr>
<td>10:45-11:15</td>
<td>Break &amp; discussion</td>
</tr>
<tr>
<td>11:15-11:45</td>
<td>“Non-targeted analysis for CE-MS data” Masahiro Sugimoto</td>
</tr>
<tr>
<td>11:45-12:30</td>
<td>Invited Talk: “Cell factory research at VTT: Applications of metabolomics and -omics technologies” Merja Penttilä:</td>
</tr>
<tr>
<td>12:30–13:00</td>
<td>Open discussion</td>
</tr>
<tr>
<td>13:00</td>
<td>End of Scientific Programme</td>
</tr>
</tbody>
</table>
Table of Contents

MassBank: A public repository of mass spectral data ..................... 1
   Takaaki Nishioka

Energy-based identification of molecular fragments in tandem mass
spectrometry ......................................................... 3
   Markus Heinonen

Circulating information and knowledge through a cyber-infrastructure ... 5
   Masanori Arita

Opportunities and bottlenecks of high-throughput metabolomics .......... 7
   Nicola Zamboni

Algorithms and Machine Learning for Small Molecules ................... 9
   Juho Rousu

Species-metabolite relation database KNAPSAcK, and its multifaceted
retrieval system KNAPSAcK Family ................................... 11
   Kensuke Nakamura

Processing of mass spectrometry based molecular profile data .......... 13
   Matej Orešič

Spectral libraries and computational mass spectrometry – A symbiosis? .. 15
   Steffen Neumann

Analysis of biosynthetic fluctuations of cultured Taxus seedlings using
LC-IT-TOF MS and a metabolomic approach .......................... 17
   Ken Tanaka

Non-targeted analysis for CE-MS data .................................. 19
   Masahiro Sugimoto

Cell factory research at VTT: Applications of metabolomics and -omics
technologies ......................................................... 21
   Merja Penttilä
MassBank: A public repository of mass spectral data

Takaaki Nishioka

Institute for Advanced Biosciences
Keio University, Japan
and MassBank Project, JST-BIRD

Statistics of the Data: MassBank (http://www.massbank.jp/) is the first public repository of mass spectral data of biological, and chemically synthetic small molecules. Contributors prepare their data in a common record format, “MassBank record format”. Ideally, each contributor has a local data server for publication of the formatted data. As of August 2010, a total of 17 research groups (14 in Japan, 2 in the USA, and 1 in Germany) contribute their data to MassBank from 8 data servers distributed on the internet. Nine small research groups, which currently do not have their own database servers, contribute their data from one of the 8 data servers. Contributed MassBank data are 12,362 mass spectra (11,620 ESI-MS\textsuperscript{n}, 605 EI-MS, 136 FAB-MS, 1 MALDI-MS) of 2,811 biological compounds, 3,504 ESI-MS\textsuperscript{2} data of 749 synthetic drugs, and 12,380 EI-, CE- and other MS data of 10,287 volatile synthetic compounds. The query interface of MassBank functions as an access point to the distributed data servers.

Merged ESI-MS\textsuperscript{2} data as reference data: No standard experimental protocol is available for ESI-MS\textsuperscript{n}, because researchers have individually optimized their experimental methods of ESI-MS\textsuperscript{n} to their target compounds. The ESI-MS\textsuperscript{n} data contributed to MassBank were analyzed by 14 different, well-controlled analytical methods. However, slight differences in their experimental methods gave non-identical (i.e., similar) mass spectra to identical chemical compounds. This raises concerns about the suitability of the contributed ESI-MS\textsuperscript{2} data as the reference for the identification of chemical compounds detected by mass spectrometry. MassBank project found a solution to the problem. Research groups in IAB, Keio University and PSC, RIKEN contributed the ESI-MS\textsuperscript{2} data that each chemical compound was analyzed at five different collision energies, 5-60 V. When the ESI-MS\textsuperscript{2} data for each chemical compound were overlaid into an artificially merged single MS\textsuperscript{2} spectrum (or a ramp-mode MS\textsuperscript{2} spectrum), such merged MS\textsuperscript{2} spectra yielded representative fragmentation patterns. A total of 1,449 merged MS\textsuperscript{2} spectra are available as the reference for chemical identification. Data merging has significantly improved the precision in the identification of chemical compounds through the MassBank service “Spectral Similarity Search”.
Tools for contributors: Previously preparation and management of MassBank records on data servers involved tedious manual works. To reduce such contributors’ works, MassBank project has developed two tools, “Record Editor” and “Administration Tool”. Additionally in collaboration with the Mass++ project (http://groups.google.com/group/massplusplus), peak data and analytical parameters from the binary raw data from different types of mass spectrometers can be now extracted and transformed into the MassBank record format. These tools considerably reduced the contributors’ workload.

Application program interface (SOAP-API): MassBank project provides an interface, SOAP-API, for users’ programs to search MassBank data. For example, the Mass++ program can extract peak data from binary raw data and submit the peak data as a query of Spectral Similarity Search to get the search results by using MassBank SOAP-API.

Fragmentation library: In the MassBank project, molecular formulae were assigned to about 4,000 observed major product ions in merged ESI-QqTOF-MS² data of 453 chemical compounds (positive and negative modes). From the molecular formulae, we successfully predicted the covalent bonds the collision cleaved. This information can be accessed from the MassBank service, “Peak Search Advanced”, in which users can retrieve the MS² spectra that gave the product ions or neutral losses composed of the specified molecular formulae. It is useful for analyzing the chemical relations between product ions and chemical substructures. The accumulation of such relations is essential to elucidate chemical structure of unidentified chemical compounds from mass spectral data.

Energy-based identification of molecular fragments in tandem mass spectrometry

Markus Heinonen

Department of Computer Science
University of Helsinki, Finland

Mass Spectrometry is one of the key enabling measurement technologies for systems biology, due to its ability to quantify molecules in small concentrations. Tandem mass spectrometry tackles the main shortcoming of mass spectrometry: the fact that molecules with equal mass-to-charge ratios are not separated. Its fragmentation patterns create a need for methods for identifying the generated fragments.

In contrast to manual annotations to fragment peaks in the mass spectrum, computational methods allow for processing and analyzing thousands of spectra, looking for patterns to provide reasonable annotations. Here, the open databases of mass spectra are crucial for methodological progress.

Fragment identification methods have long relied on rule-based methods, where domain knowledge has been used to infer fragmentation rules mapping structural patterns into fragments. We present an energy-based combinatorial approach for predicting the fragment structures. We enumerate the fragment set and rank them according to the bond dissociation energies of the cleaved bonds. Unlike many existing methods, our method does not rely on hand-coded fragmentation rules. Extensions to the method allow simultaneous prediction of fragmentation reaction chains and prediction of rearrangement reactions, where in addition to cleavages bond formations also occur. Our method is able to predict the correct fragmentation of medium-sized molecules with high accuracy.
Circulating information and knowledge through a cyber-infrastructure

Masanori Arita

Department of Biophysics and Biochemistry
Graduate School of Science
The University of Tokyo and RIKEN Plant Science Center, Japan

A MediaWiki-based system (http://metabolomics.jp/) is introduced for organizing metabolomic knowledge across multiple research groups. The system not only achieves efficiency by searching and summarizing structured data entries but also flexibility by accepting free formatted description for ambiguous data. In addition, a user will be inevitably exposed to multidisciplinary data from different academic disciplines while navigating through the wiki pages. As an example, MS² spectra for 453 compounds in MassBank have been annotated on our wiki platform (http://metabolomics.jp/wiki/Index:MassBank). These pages only list the relationship of fragments’ dissociation in terms of molecular formulas, and all other factors such as page layouts, display styles, and score rankings are automatically computed using the Lua programming language embedded inside pages. The benefits of this style are the familiarity of wiki inputs and the flexibility of style changes. Annotation activities are often ad-hoc. Acceptance of free-formatted inputs is indeed necessary. On the other hand, constraints by using the text-based wiki obstruct the design of highly formalized, large-scale numeric database. Inefficiency is its fatal disadvantage. The system is now used by several academic groups in Japan. One closely collaborating project is the LipidBank of Japanese Conference on the Biochemistry of Lipids (http://lipidbank.jp/).
Opportunities and bottlenecks of high-throughput metabolomics

Nicola Zamboni
Institute of Molecular Systems Biology
ETH Zürich, Switzerland

In all questions of biology that aim at - or rely on - gathering a holistic picture of cellular life, there is a pressing demand for large scale studies with thousands of strains or conditions. This applies for example to functional genomics, causal elucidation of regulatory networks, mapping of gene-environment interactions in quantitative traits, inhibitor screens, etc.

To meet current and upcoming requirements of biology, biotechnology, and clinics, we set out to develop high-throughput metabolomics. We successfully established a pipeline capable of handling 1000 samples/day and yet provide a broad coverage of metabolism. The effort included scale-down to microscale cultivation, parallelization of sample preparation, development of high-throughput metabolomics by flow injection—time-of-flight mass spectrometry, and novel data processing workflow. The analytical platform was thoroughly optimized for sensitivity, linearity, robustness, accuracy to ultimately attain reproducible detection of typically 2000-4000 ions in a single sample in less than a minute. Our ad-hoc processing software exploits the size of the dataset to filter background signal, remove noise, convolute multimers, adducts and isotopomers, and merge thousands of samples without sacrificing rare markers. The routine throughput is of 1400 samples/day and thus far beyond any other existing -omics workflow. To date the platform delivers accurate monitoring of ca. 1000-2000 biological compounds with good coverage of primary metabolism, but only a minor fraction can be successfully annotated. With this platform, we completed a genome-wide analysis of the metabolome in ca. 8500 single-knockout mutants of Escherichia coli (Keio collection), comprising more than 35000 analyses in 6 weeks on a single instrument with excellent reproducibility.

More recently, we started developing a workflow for comprehensively targeted metabolomics, which efficiently employs the hybrid nature of the AB SCIEX QTRAP 5500 which is anticipated to enable precise quantification of all detectable metabolites in any matrix and hundreds of samples per day. Exemplary cases for these developments and the information content of both setups and the unprecedented potential of this platform will be presented, together with the current limitations inherent to ion annotation which largely call for structural elucidation by MSⁿ.
Algorithms and Machine Learning for Small Molecules

Juho Rousu
Department of Computer Science, University of Helsinki, Finland

This presentation will give an overview of the CSBB group in the realm of metabolism. The first part of the talk will focus on the computational methods developed in the context 13C metabolic flux analysis, including tools for 13C mass spectra deconvolution, metabolic network construction and in particular flux estimation and experimental planning methods.

The second part describes our recent activity in developing structural machine learning methods for various prediction problems involving small molecules such as enzyme function prediction and drug bioactivity prediction. I will also outline how this family of methods with MassBank resources could be used to build tools for spectral annotation.
Species-metabolite relation database  
KNApSAcK, and its multifaceted retrieval  
system KNApSAcK Family

Kensuke Nakamura
Comparative Genomics Laboratory, Graduate School of Information Science  
Nara Institute of Science and Technology, Japan

In the metabolomics research, assignment of measured mass spectra to specific  
metabolites is one of the most fundamental processes. The assignment is usually  
made by taking match of molecular weight with all known compounds. This means  
the search goes through whole natural products reported in literatures, which may be  
extremely tedious and daunting process. In order to make this process more  
feasible, we have developed a metabolite database that contains species-metabolite  
relationship called KNApSAcK, which currently contains 102,009 species-metabolite  
relationships involving 50,054 metabolites. In the present study, we review the  
current status of the KNApSAcK database and its application to metabolomics  
research. Also, we would like to introduce the multifaceted retrieval system,  
KNApSAcK Family, which consists of seven parts for the purpose of retrieving  
metabolites from several different aspects. “Pocket” includes the search system  
for species closely related to people’s life, such as edible plant in Japan (“Lunch  
Box”), herb tea (“Tea Pot”, in progress), traditional Japanese herbal medicine (“KAMPO”),  
poisonous plant (“Poison”, in progress), and the source of bio-fuel (“Fuel”, in  
progress). “KNApSAcK from around the world” includes medicinal and edible plants  
utilized in each country. So far, 7,356 pairwise relations between medicinal/edible  
plants and 119 nations worldwide have been accumulated from scientific literatures.
Processing of mass spectrometry based molecular profile data

Matej Orešič
Quantitative Biology and Bioinformatics Group
VTT Technical Research Centre of Finland

Modern analytical technologies afford comprehensive and quantitative investigation of a multitude of different metabolites. Typical metabolomic experiments can therefore produce large amounts of data. Handling such complex datasets is an important step that has big impact on extent and quality at which the metabolite identification and quantification can be made, and thus on the ultimate biological interpretation of results. Increasing interest in metabolomics thus led to resurgence of interest in related data processing. A wide variety of methods and software tools have been developed for metabolomics during recent years, and this trend is likely going to continue. The key steps of metabolomic data processing will be overviewed, with the focus on methods for handling data from liquid chromatography mass spectrometry (LC-MS) experiments.
Spectral libraries and computational mass spectrometry – A symbiosis?

Steffen Neumann
Leibniz Institute of Plant Biochemistry, Germany

Computational Mass Spectrometry is starting to deliver powerful tools for the analysis of mass spectra. For training and validation, large gold standards are a must: “There is no data like more data” (Mercer at Arden. House, 1985). That’s why we need Open Data.

With modern MS instruments, we can produce large spectral libraries, but reliable annotations (and quality control !) remain a tedious process. That’s why we need Computational Mass Spectrometry.
Analysis of biosynthetic fluctuations of cultured Taxus seedlings using LC-IT-TOF MS and a metabolomic approach

Ken Tanaka

Department of Medicinal Resources
Institute of Natural Medicine
University of Toyama, Japan

Fluctuations in the biosynthesis of taxoids in 1 to 5 year old cultured seedlings of “T. chinensis” var. “mairei” were investigated using LC-IT-TOF MS and a metabolomic approach. In the TIC of the extracts, 16 prominent peaks were observed. Ten compounds were identified by comparison of the retention times and MS/MS spectra with those of reference compounds. An additional 6 taxoids were annotated following a detailed analysis of the MS/MS spectral data. It was clarified that the relative concentrations of basic taxoids with 4(20) double bonds are high in the early stages of cultivation. On the other hand, relatively higher amounts of 5-acetoxy taxoids oxidated at positions 4- and 10- and taxoids having 5(20)-oxetane rings were found in the later stages of cultivation. The approach provides practical information on the biosynthetic flow of taxoids in cultured yew seedlings.
Non-targeted analysis for CE-MS data

Masahiro Sugimoto
Institute for Advanced Biosciences
Keio University, Japan

Capillary electrophoresis-mass spectrometry (CE-MS), a powerful platform to quantify numerous charged metabolites simultaneously, has made a large contribution in advancing metabolomics research. Since charged metabolites are the most common intermediates in central metabolic pathways, comprehensive CE-MS profiling has a great potential to facilitate the understanding of cellular metabolism at the systems level. However, large run-to-run variability in migration times and the variety of peak shapes observed during CE-MS data hinder data processing and quantification. Especially, metabolite identification of the detected peaks without matching compound in standard library is still a difficult procedure in analyzing CE-MS data. Here, we introduce a data processing algorithm optimized for non-targeted analysis of CE-time-of-flight (TOF)-MS data, including feature detection, alignment of multiple samples, noise-filtering and interpretation of detected peaks, and assignment of possible metabolite names to the metabolite-derived peaks.
Cell factory research at VTT: Applications of metabolomics and -omics technologies

Merja Penttilä

Finnish CoE in White Biotechnology - Green Chemistry
VTT Technical Research Centre of Finland