Establishing metabolome analysis technology for seeking out unknown useful natural products

SAKURAI, Nozomu Project Associate Professor Bioinformation and DDBJ Center

Graduated from the Faculty of Agriculture, Tohoku University (under the tutelage of Prof. Tomoyuki Yamaya). Finished a doctoral course at the Graduate School of the same university (Doctor of Agriculture). After serving as a researcher at the Kazusa DNA Research Institute and as a team leader of the metabolomics team at the same institution, assumed the position of Project Associate Professor at the National Institute of Genetics in June 2018. Coawarded a Technology Award by the Japanese Society for Plant Cell and Molecular Biology in 2012.

Project Associate Professor Nozomu Sakurai assumed his current position in June 2018. Before that, he had been engaged in the technological development of metabolome analysis at the Kazusa DNA Research Institute for approximately 17 years. During his career, he has developed mass spectrometry tools and software that do not restrict chemical components of analysis objects. These tools and software are made publicly available on his self-administered portal site. With the ability to identify chemical characteristics and lists of candidate substances without the need to determine exact substance names of unknown components, he believes that it becomes possible to make full use of previously untapped data.

Starting metabolome analysis at the Kazusa DNA Research Institute

Initially a student of agriculture at the Tohoku University, I was studying the internal circulation of nitrogen with the aim of increasing the production of rice. In 2001, I was appointed at the Kazusa DNA Research Institute, and the following year, I participated in the newly initiated metabolome analysis project commissioned by the New Energy and Industrial Technology Development Organization (NEDO). The Kazusa DNA Research Institute was founded as an institution specializing in DNA research in 1994, then proceeding with studies such as genome sequencing, DNA analysis technology development, and gene prediction program development. With the belief that “comprehensive information on biologically metabolized substances is necessary to better understand living organisms,” we also started metabolome analysis. At the time, genome sequencing of various organisms was rapidly advancing, and it was kind of “trendy” to perform comprehensive (ome) analysis on genetic expression and other subjects.

In the NEDO project, we performed metabolome analysis primarily of plants with the aim of increasing the production of and searching for substances applicable to industrial materials. Besides our team, the Keio University and Riken independently launched their own metabolome projects.

Amid these circumstances, I joined the field of bioinformatics and have since been engaged in metabolome analysis and development of computer analytical technology to this date.

Metabolome analysis at the initial stages was not very comprehensive

However, metabolome analysis back then was not what you would describe as “comprehensive.” Although research was underway to examine specific metabolites using mass spectroscopy, no computer

software that was capable of capturing and comprehensively analyzing raw data as a whole was available. The ability to depict the entirety of tens of thousands of metabolites contained in samples was beyond anyone's wildest dreams back then. Currently, the technologies have improved to the point where this dream can be, not fully, but partially realized.

In the past, researchers used to purchase a commercially available "purified single substance (specimen)" for mass spectrometry (MS) and used obtained signal data as indices. For example, by comparing the index of Substance A with the "signals yielded from MS of a sample," it was possible to determine the absence or presence of Substance A and its quantity. However, specimens were available only for a limited number of substances, with only about 1000 types purchasable at the most. Most analytical software applications were capable of visualizing individual pieces of component information. Thus, it was impossible to comprehensively capture and analyze component signals contained in datasets.

In the NEDO project, we developed a software that could capture and digitalize all signals output by mass spectroscopy (PowerGet). Currently, it is still under improvement to achieve better performance. That said, the use of this software makes it possible to list the mass and characteristics of all signals. Using such lists, researchers can compare specimen to specimen and analyze a single specific substance in detail if that piques their interest. In other words, it has become possible to perform analysis of unspecified (nontargeted) substances rather than targeting specific components based on specimens.

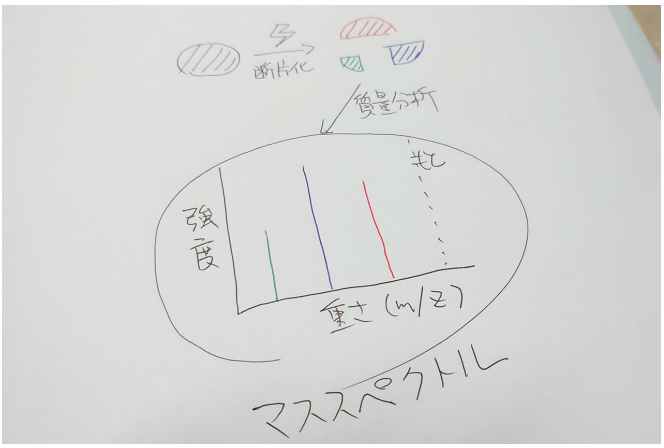
I can say that by performing metabolome analysis, we have achieved the strategy of transcriptome analysis to some extent: "focusing on specific genes after overviewing the entire picture of gene expression using DNA chips."

◀ link to PowerGet

A tool for nontargeted prediction of 7000 flavonoids

After the completion of the NEDO project, we continued taking one of the hardest challenges in metabolome analysis, namely the "development of tools for estimating components without specimens," at the Kazusa DNA Research Institute, motivated by the desire to uncover all the information that is supposedly recorded by mass spectrometers.

For instance, we have established a technology for comprehensively detecting flavonoids (FlavonoidSearch). Flavonoids are a general term for a group of plant-derived metabolites, including pigments. There are more than 7,000 known flavonoids. Because many of them exhibit antioxidant and hormone-like activities, flavonoids have been attracting attention as ingredients for functional food products. Although flavonoids such as anthocyanin, tea catechin, and soy isoflavone are well known, there are still many varieties whose functions are unknown. Meanwhile, there are only around 100 commercially available flavonoid specimens that can be used as indices for component identification. In other words, even if an analytical sample contains flavonoids with useful physiological properties, they remain undetectable.

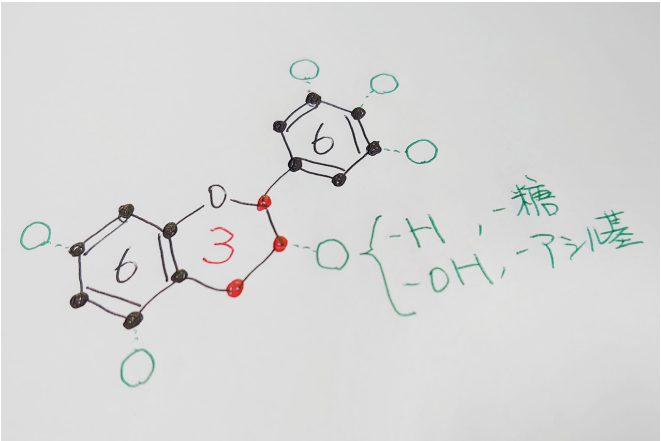


Against this backdrop, I thought about developing a tool for detecting all substances that are "possibly flavonoids" using MS-derived data (mass spectra) based on liquid chromatography-MS (LC-MS). Mass spectra are a sort of "fingerprint reflecting the chemical structures of components." By making good use of this information, we can distinguish and estimate chemical structures to a certain extent.

◀ link to FlavonoidSearch

A flavonoid has a C6-C3-C6 core (diphenyl propane structure), to which chemical groups such as sugars and acyl groups are attached. First, we performed precise MS using commercially available to obtain accurate MS data. Next, referring to the past literature information, we constructed a set of rules to predict what type of flavonoid chemical structure will yield what type of spectrum. This attempt was successful thanks to the "craftsman-like feats" of a former colleague of mine, Dr. Nayumi Akimoto, who was a postdoctoral researcher specializing in chemistry back then (currently working for the Kazusa DNA Research Institute). Dr. Akimoto was able to discern the correlation between subtle chemical structural differences and their mass spectra at levels higher than the comprehension of most of the staff, including me. She then constructed prediction rules. Over a period of 3 years, we converted structures of 7000 flavonoids into mass spectrum data. Using this data, I wrote a program that lets you compare them with actual LC-MS measured spectrum signals, eventually completing the FlavonoidSearch software capable of comprehensively

detecting flavonoid candidates in diverse species of organisms.



In fact, when we comprehensively analyzed flavonoids contained in parsley, we found that 36 of approximately 1000 components whose mass spectrum data was obtained were flavonoid candidates. Of these candidates, some had previously unknown chemical groups attached to the core, which were, namely, novel flavonoids. Given that only about six flavonoids were previously reported to be contained in parsley, I am convinced that FlavonoidSearch will contribute to seeking out novel functional flavonoids.

In 2016, the year FlavonoidSearch was completed, an artificial intelligence (AI) called AlphaGo caused a sensation by defeating a professional go player. Perhaps because of this, people tended to think “What if we used AI for developing tools for component estimation?”However, it is impracticable for AI to set parameters that can properly process various chemical structures. Therefore, it may still take a while before such applications are put to practical use.

Construction and publication of the Food Metabolome Repository

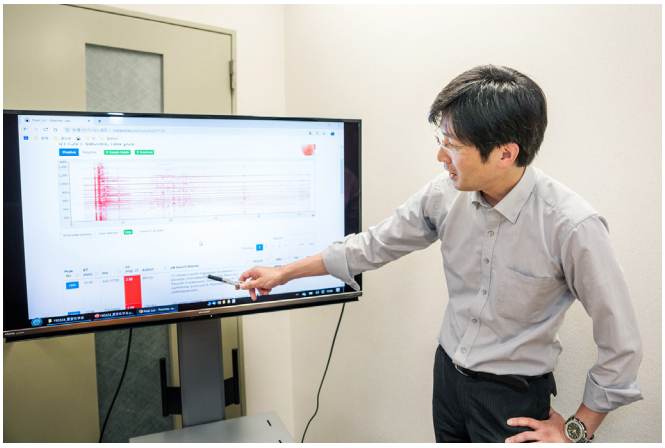
In 2017, we constructed an information resource called the “Food Metabolome Repository,” which is simply abbreviated to “Shoku-Repo” (Food Repo). We focused on food because foodstuffs are created from processed living organisms, including plants, livestock, and fish. Therefore, they are likely to contain diverse metabolites and thus prove very useful in metabolome analysis as a whole.

◀link to the Food Metabolome Repository

We analyzed 222 ordinary foodstuffs available at supermarkets and other retailers using LC-MS. We then collected data on the LC elution time (shorter for highly water-soluble components and longer for poorly water-soluble components) and accurate mass value (mass-to-charge ratio) of each component. As a result, approximately 4000 substances were extracted per food on average. The Food Repo provides the “results of flavonoid prediction by FlavonoidSearch” for components whose spectra are available. Additionally, it also provides two-dimensional mass chromatograms, with horizontal and vertical axes representing the elution time and mass-to-charge ratio, respectively. This allows the user to visually differentiate components from foodstuff to foodstuff.

The standard tables of food composition in Japan list quantitative values of approximately 140 components of main foodstuffs, including carbohydrates, proteins, lipids, fatty acids, amino acids, organic acids, minerals, and vitamins. However, actual foodstuffs contain thousands of substances not listed in these tables. Some may be beneficial for living organisms even at trace amounts, and others may contain remarkably different components depending on the manufacturing process even when the raw material was the same. It would be ideal if we could uncover these unknown components and put them to good use. However, the new challenge is “how to decide which unknown component to pay attention to.” Even when the name of the substance cannot be identified, information on which foodstuffs the same component is detected from will be helpful in prioritizing substances that should be analyzed. The Food Repo makes it easy to gain such relevant information.

Let me elaborate on one possible use. Imagine that you were studying yogurt and found only products using a specific strain of lactic acid bacteria had properties to “lower blood pressure.” You performed a nontargeted analysis using a software such as PowerGet and identified 20 candidate components that possibly correlated with the effects to lower blood pressure. It is virtually impossible to analyze all these components in detail, so you need to narrow down the targets. In such cases, the Food Repo allows the user to search for and list foodstuffs containing the same unknown components by referencing the mass-to-charge value the user obtained from LC-MS. Of the 20 candidates, if Substance A was contained particularly in large quantities in lactic acid bacteria-derived products, such as yogurt, and was not detected in raw milk or cheese, this Substance A is highly likely to be a useful component produced by lactic acid bacteria; therefore, it should be highly prioritized. On the other hand, if Substance B was found to be commonly contained in other foodstuffs such as vegetables and fish, its priority should be lowered.



Let me talk about an instance where the Food Repo actually proved useful in my research. Nontargeted analysis of soil samples collected from areas around crop roots in several regions detected a substance only extracted from soil samples from a particular region. When I searched for this substance on the Food Repo, no foodstuffs contained the identical substance. Taking a hint from the fact that “no foodstuffs” contained the same substance, I inferred that this substance might be produced by soil bacteria rather than by plants. Lo and behold, it was in fact derived from a certain soil bacterium. Proceeding with studies based on this approach may lead to the inception of novel antibiotics

However, please note that the analytical conditions adopted in LC-MS for the Food Repo make it impossible to measure proteins, highly

hydrophobic substances, and substances that are hard to ionize.
I have introduced three tools so far. They are all made publicly available on a portal site called KOMICS (Kazusa Metabolomics Database). Anyone can access them for research purposes.

◀link to KOMICS (Kazusa Metabolomics Database)

Calling for academic–industrial collaboration for developing a metabolome analytical data repository

Based on my work at the Kazusa DNA Research Institute, my mission at the National Institute of Genetics is to construct a public metabolome analytical repository that is not affected by the type of MS used. In Japan or globally, there is no truly practical metabolome analytical repository. Although databases such as METABOLOMICSWORKBENCH (US) and MetaboLights (Europe) are available, they are merely information archives and do not function as comparative analytical tools.

◀link to METABOLOMICSWORKBENCH [US]

◀link to MetaboLights [Europe]

The biggest factor of the above situation is that simple comparison of data is unfeasible when the elution time of the sample, characteristics of ionization, and detection sensitivity differ depending on the analytical equipment and conditions. Making good use of my experience in developing the Food Repo, I hope to continue developing easy-to-use “searchable and comparable databases” irrespective of the MS equipment and conditions.

Metabolome analysis has the potential to be applied to disease prevention, drug creation, and novel remedy development. For instance, MS of blood yields more than 1000 signals, most of which are probably derived from foodstuffs, intestinal bacteria, and liver metabolites. However, it is impossible to identify the exact substances. If we can identify substances that fluctuate greatly depending on one’s health state (biomarkers) or substances that reflect a prior stage of onset (nondisease state as referred to in oriental medicine), they would prove very useful. For this reason, promoting academic–industrial collaboration is another one of my major missions. Already, the National Institute of Genetics, Kazusa DNA Research Institute, and Tsumura have joined hands and initiated comprehensive analysis and structural estimation of medicinal herb components. In Japan, oriental medicine has developed in its own way and has become an integral part of modern medicine, with the usefulness of herbal medicines being well established for treating diseases. Despite this, their active ingredients and action mechanisms remain largely unknown. By sharing the tools and data with Tsumura, we hope to proceed with analyses of oriental herbal medicines, thereby accumulating useful findings.

Certainly, there are inherent complications in academic–industrial collaboration in terms of patents and intellectual property rights; however, good collaboration is essential for putting the results into practice. For both academic and industrial sectors to reap profits, I believe the environment and legislation need to be optimized.

My dream is to develop an inexpensive mass spectroscope and novel sensing technology

A strength of metabolome analysis is the ability to discover useful unknown substances by making full use of untapped data. At the same time, there are many drawbacks that need to be resolved. One is that MS equipment is prohibitively expensive, with cheaper devices costing around 10 million yen and high-end models requiring whapping 200 million yen. Although this is still a mere pipe dream, I wish to develop cheaper mass spectrometers myself. Moreover, I also want to develop more affordable sensing technology that can replace MS. To that end, I have already taken the first step.

Hopefully, my study results will encourage more researchers to take up metabolome analysis, who would otherwise have stayed away from this field because of its apparent complexity. It is true that metabolome analysis is convoluted, with an infinite number of analytical targets. However, I would like to continue developing easy-to-use technologies so that genomics researchers would be encouraged to dabble in.

Interviewer:
Naoko Nishimura, Science writer

Photographer:
Mitsuhiko Kurusu, Office for Research Development

July 2019