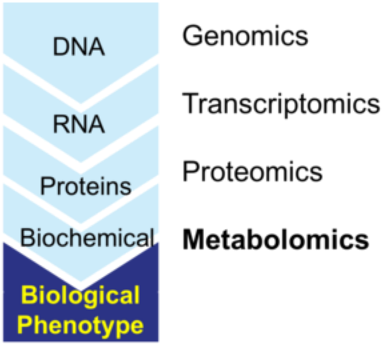
WIKIPEDIA

Metabolomics



The central dogma of biology showing the flow of information from DNA to the phenotype. Associated with each stage is the corresponding systems biology tool, from genomics to metabolomics.

Metabolomics is the scientific study of chemical processes involving metabolites, the small molecule substrates, intermediates and products of cell metabolism. Specifically, metabolomics is the "systematic study of the unique chemical fingerprints that specific cellular processes leave behind", the study of their small-molecule metabolite profiles.[1] The metabolome represents the complete set of metabolites in a biological cell, tissue, organ or organism, which are the end products of cellular processes.[2] Messenger RNA (mRNA), gene expression data and proteomic analyses reveal the set of gene products being produced in the cell, data that represents one aspect of cellular function. Conversely, metabolic profiling can give an instantaneous snapshot of the physiology of that cell,[3] and thus, metabolomics provides a direct "functional readout of the physiological state" of an organism.[4] One of the challenges of systems biology and functional genomics is to integrate genomics, transcriptomic, proteomic, and metabolomic information to provide a better understanding of cellular biology.

Contents

History

The concept that individuals might have a "metabolic profile" that could be reflected in the makeup of their biological fluids was introduced by Roger Williams in the late 1940s,[5] who used paper chromatography to suggest characteristic metabolic patterns in urine and saliva were associated with diseases such as schizophrenia. However, it was only through technological advancements in the 1960s and 1970s that it became feasible to quantitatively (as opposed to qualitatively) measure metabolic profiles. [6] The term "metabolic profile" was introduced by Horning, *et al.* in 1971 after they demonstrated that gas chromatography-mass spectrometry (GC-MS) could be used to measure compounds present in human urine and tissue extracts.[7][8] The Horning group, along with that of Linus Pauling and Arthur B. Robinson led the development of GC-MS methods to monitor the metabolites present in urine through the 1970s.[9]

Concurrently, NMR spectroscopy, which was discovered in the 1940s, was also undergoing rapid advances. In 1974, Seeley et al. demonstrated the utility of using NMR to detect metabolites in unmodified biological samples.[10] This first study on muscle highlighted the value of NMR in that it was determined that 90% of cellular ATP is complexed with magnesium. As sensitivity has improved with the evolution of higher magnetic field strengths and magic angle spinning, NMR continues to be a leading analytical tool to investigate metabolism.[7][11] Recent efforts to utilize NMR for metabolomics have been largely driven by the laboratory of Jeremy K. Nicholson at Birkbeck College, University of London and later at Imperial College London. In 1984, Nicholson showed ^{1}H NMR spectroscopy could potentially be used to diagnose diabetes mellitus, and later pioneered the application of pattern recognition methods to NMR spectroscopic data.[12][13]

In 1995 liquid chromatography mass spectrometry metabolomics experiments[14] were performed by Gary Siuzdak while working with Richard Lerner (then president of The Scripps Research Institute) and Benjamin Cravatt, to analyze the cerebral spinal fluid from sleep deprived animals. One molecule of particular interest, oleamide, was observed and later shown to have sleep inducing properties. This work is one of the earliest such experiments combining liquid chromatography and mass spectrometry in metabolomics.

In 2005, the first metabolomics tandem mass spectrometry database, METLIN,[15][16] for characterizing human metabolites was developed in the Siuzdak laboratory at The Scripps Research Institute. METLIN has since grown and as of July 1, 2019, METLIN contains over 450,000 metabolites and other chemical entities, each compound having experimental tandem mass spectrometry data generated from molecular standards at multiple collision energies and in positive and negative ionization modes. METLIN is the largest repository of tandem mass spectrometry data of its kind. 2005 was also the year in which the dedicated academic journal Metabolomics first appeared, founded by its current editor-in-chief Professor Roy Goodacre.

In 2005, the Siuzdak lab was engaged in identifying metabolites associated with sepsis and in an effort to address the issue of statistically identifying the most relevant dysregulated metabolites across hundreds of LC/MS datasets, the first algorithm was developed to allow for the nonlinear alignment of mass spectrometry metabolomics data. Called XCMS,[17] where the "X" constitutes any chromatographic technology, it has since (2012)[18] been developed as an online tool and as of 2019 (with METLIN) has over 30,000 registered users.

On 23 January 2007, the Human Metabolome Project, led by David Wishart of the University of Alberta, Canada, completed the first draft of the human metabolome, consisting of a database of approximately 2500 metabolites, 1200 drugs and 3500 food components.[19][20] Similar projects have been underway in several plant species, most notably *Medicago truncatula*[21] and *Arabidopsis thaliana*[22] for several years.

As late as mid-2010, metabolomics was still considered an "emerging field".[23] Further, it was noted that further progress in the field depended in large part, through addressing otherwise "irresolvable technical challenges", by technical evolution of mass spectrometry instrumentation.[23]

In 2015, real-time metabolome profiling was demonstrated for the first time.[24]

Metabolome

The metabolome refers to the complete set of small-molecule (<1.5 kDa)[19] metabolites (such as metabolic intermediates, hormones and other signaling molecules, and secondary metabolites) to be found within a biological sample, such as a single organism.[25][26] The word was coined in analogy with transcriptomics and proteomics; like the transcriptome and the proteome, the metabolome is dynamic, changing from second to second. Although the metabolome can be defined readily enough, it is not currently possible to analyse the entire range of metabolites by a single analytical method.

Human metabolome project

The first metabolite database (called METLIN) for searching fragmentation data from tandem mass spectrometry experiments was developed by the Siuzdak lab at The Scripps Research Institute in 2005.[15][16] METLIN contains over 450,000 metabolites and other chemical entities, each compound having experimental tandem mass spectrometry data. In 2006,[17] the Siuzdak lab also developed the first algorithm to allow for the nonlinear alignment of mass spectrometry metabolomics data. Called XCMS, where the "X" constitutes any chromatographic technology, it has since (2012)[18] been developed as an online tool and as of 2019 (with METLIN) has over 30,000 registered users.

In January 2007, scientists at the University of Alberta and the University of Calgary completed the first draft of the human metabolome. The Human Metabolome Database (HMDB) is perhaps the most extensive public metabolomic spectral database to date.[27] The HMDB stores more than 110,000 different metabolite entries. They catalogued approximately 1200 drugs and 3500 food components that can be found in the human body, as reported in the literature.[19] This information, available at the Human Metabolome Database (www.hmdb.ca) and based on analysis of information available in the current scientific literature, is far from complete.[28] In contrast, much more is known about the metabolomes of other organisms. For example, over 50,000 metabolites have been characterized from the plant kingdom, and many thousands of metabolites have been identified and/or characterized from single plants.[29][30]

Each type of cell and tissue has a unique metabolic ‘fingerprint’ that can elucidate organ or tissue-specific information. Bio-specimens used for metabolomics analysis include but not limit to plasma, serum, urine, saliva, feces, muscle, sweat, exhaled breath and gastrointestinal fluid.[31] The ease of collection facilitates high temporal resolution, and because they are always at dynamic equilibrium with the body, they can describe the host as a whole.[32] Genome can tell what could happen, transcriptome can tell what appears to be happening, proteome can tell what makes it happen and metabolome can tell what has happened and what is happening.[33]

Metabolites

Metabolites are the substrates, intermediates and products of metabolism. Within the context of metabolomics, a metabolite is usually defined as any molecule less than 1.5 kDa in size.[19] However, there are exceptions to this depending on the sample and detection method. For example, macromolecules such as lipoproteins and albumin are reliably detected in NMR-based metabolomics studies of blood plasma.[34] In plant-based metabolomics, it is common to refer to "primary" and "secondary" metabolites.[3] A primary metabolite is directly involved in the normal growth, development, and reproduction. A secondary metabolite is not directly involved in those processes, but usually has important ecological function. Examples include antibiotics and pigments.[35] By contrast, in human-based metabolomics, it is more common to describe metabolites as being either endogenous (produced by the host organism) or exogenous.[36] [37]Metabolites of foreign substances such as drugs are termed xenometabolites.[38]

The metabolome forms a large network of metabolic reactions, where outputs from one enzymatic chemical reaction are inputs to other chemical reactions. Such systems have been described as hypercycles.

Metabonomics

Metabonomics is defined as "the quantitative measurement of the dynamic multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification". The word origin is from the Greek *μεταβολή* meaning change and *nomos* meaning a rule set or set of laws.[39] This approach was pioneered by Jeremy Nicholson at Murdoch University and has been used in toxicology, disease diagnosis and a number of other fields. Historically, the metabonomics approach was one of the first methods to apply the scope of systems biology to studies of metabolism.[40][41][42]

There has been some disagreement over the exact differences between 'metabolomics' and 'metabonomics'. The difference between the two terms is not related to choice of analytical platform: although metabonomics is more associated with NMR spectroscopy and metabolomics with mass spectrometry-based techniques, this is simply because of usages amongst different groups that have popularized the different terms. While there is still no absolute agreement, there is a growing consensus that 'metabolomics' places a greater emphasis on metabolic profiling at a cellular or organ level and is primarily concerned with normal endogenous metabolism. 'Metabonomics' extends metabolic profiling to include information about perturbations of metabolism caused by environmental factors (including diet and toxins), disease processes, and the involvement of extragenomic influences, such as gut microflora. This is not a trivial difference; metabolomic studies should, by definition, exclude metabolic contributions from extragenomic sources, because these are external to the system being studied. However, in practice, within the field of human disease research there is still a large degree of overlap in the way both terms are used, and they are often in effect synonymous.[43]

Exometabolomics

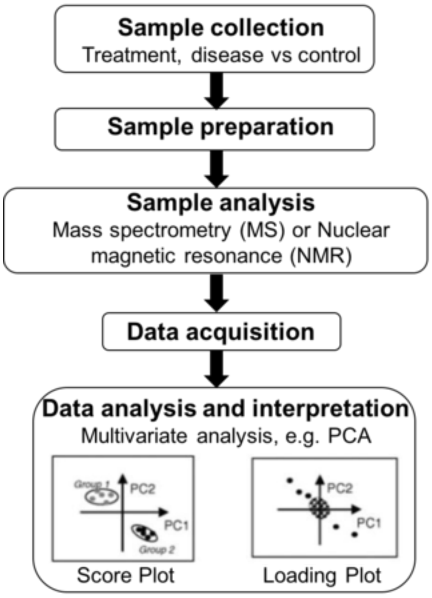
Exometabolomics, or "metabolic footprinting", is the study of extracellular metabolites. It uses many techniques from other subfields of metabolomics, and has applications in biofuel development, bioprocessing, determining drugs' mechanism of action, and studying intercellular interactions.[44]

Analytical technologies

The typical workflow of metabolomics studies is shown in the figure. First, samples are collected from tissue, plasma, urine, saliva, cells, etc. Next, metabolites extracted often with the addition of internal standards and derivatization.[45] During sample analysis, metabolites are quantified (liquid chromatography or gas chromatography coupled with MS and/or NMR spectroscopy). The raw output data can be used for metabolite feature extraction and further processed before statistical analysis (such as PCA). Many bioinformatic tools and software are available to identify associations with disease states and outcomes, determine significant correlations, and characterize metabolic signatures with existing biological knowledge.[46]

Separation methods

Initially, analytes in a metabolomic sample comprise a highly complex mixture. This complex mixture can be simplified prior to detection by separating some analytes from others. Separation achieves various goals: analytes which cannot be resolved by the detector may be separated in this step; in MS analysis ion suppression is reduced; the retention time of the analyte serves as information regarding its identity. This separation step is not mandatory and is often omitted in NMR and "shotgun" based approaches such as shotgun lipidomics.


Sample collection
Treatment, disease vs control
Sample preparation
Sample analysis
Mass spectrometry (MS) or Nuclear magnetic resonance (NMR)
Data acquisition
Data analysis and interpretation
Multivariate analysis, e.g. PCA
Group 1
PC2
PC1
Group 2
PC2
PC1
Score Plot
Loading Plot

Key stages of a metabolomics study

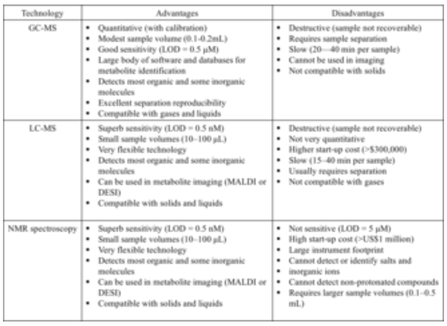
Gas chromatography (GC), especially when interfaced with mass spectrometry (GC-MS), is a widely used separation technique for metabolomic analysis.[47] GC offers very high chromatographic resolution, and can be used in conjunction with a flame ionization detector (GC/FID) or a mass spectrometer (GC-MS). The method is especially useful for identification and quantification of small and volatile molecules.[48] However, a practical limitation of GC is the requirement of chemical derivatization for many biomolecules as only volatile chemicals can be analysed without derivatization. In cases where greater resolving power is required, two-dimensional chromatography (GCxGC) can be applied.

High performance liquid chromatography (HPLC) has emerged as the most common separation technique for metabolomic analysis. With the advent of electrospray ionization, HPLC was coupled to MS. In contrast with GC, HPLC has lower chromatographic resolution, but requires no derivatization for polar molecules, and separates molecules in the liquid phase. Additionally HPLC has the advantage that a much wider range of analytes can be measured with a higher sensitivity than GC methods.[49]

Capillary electrophoresis (CE) has a higher theoretical separation efficiency than HPLC (although requiring much more time per separation), and is suitable for use with a wider range of metabolite classes than is GC. As for all electrophoretic techniques, it is most appropriate for charged analytes.[50]

Detection methods

Mass spectrometry (MS) is used to identify and quantify metabolites after optional separation by GC, HPLC, or CE. GC-MS was the first hyphenated technique to be developed. Identification leverages the distinct patterns in which analytes fragment which can be thought of as a mass spectral fingerprint; libraries exist that allow identification of a metabolite according to this fragmentation pattern. MS is both sensitive and can be very specific. There are also a number of techniques which use MS as a stand-alone technology: the sample is infused directly into the mass spectrometer with no prior separation, and the MS provides sufficient selectivity to both separate and to detect metabolites.



Technology	Advantages	Disadvantages
GC-MS	• Quantitative (with calibration) • Modest sample volume (0.1-0.2mL) • Good sensitivity (LOD = 0.5 μM) • Large body of software and databases for metabolite identification • Detects most organic and some inorganic molecules • Excellent separation reproducibility • Compatible with gases and liquids	• Destructive (sample not recoverable) • Requires sample separation • Slow (20—40 min per sample) • Cannot be used in imaging • Not compatible with solids
LC-MS	• Superb sensitivity (LOD = 0.5 nM) • Small sample volumes (10–100 μL) • Very flexible technology • Detects most organic and some inorganic molecules • Can be used in metabolite imaging (MALDI or DESI) • Compatible with solids and liquids	• Destructive (sample not recoverable) • Not very quantitative • Higher start-up cost (>$300,000) • Slow (15–40 min per sample) • Usually requires separation • Not compatible with gases
NMR spectroscopy	• Superb sensitivity (LOD = 0.5 nM) • Small sample volumes (10–100 μL) • Very flexible technology • Detects most organic and some inorganic molecules • Can be used in metabolite imaging (MALDI or DESI) • Compatible with solids and liquids	• Not sensitive (LOD = 5 μM) • High start-up cost (>US$1 million) • Large instrument footprint • Cannot detect or identify salts and • inorganic ions • Cannot detect non-protonated compounds • Requires larger sample volumes (0.1–0.5 mL)

Comparison of most common used metabolomics methods

For analysis by mass spectrometry the analytes must be imparted with a charge and transferred to the gas phase. Electron ionization (EI) is the most common ionization technique applies to GC separations as it is amenable to low pressures. EI also produces fragmentation of the analyte, both providing structural information while increasing the complexity of the data and possibly obscuring the molecular ion. Atmospheric-pressure chemical ionization (APCI) is an atmospheric pressure technique that can be applied to all the above separation techniques. APCI is a gas phase ionization method slightly more aggressive ionization than ESI which is suitable for less polar compounds. Electrospray ionization (ESI) is the most common ionization technique applied in LC/MS. This soft ionization is most successful for polar molecules with ionizable functional groups. Another commonly used soft ionization technique is secondary electrospray ionization (SESI).

Surface-based mass analysis has seen a resurgence in the past decade, with new MS technologies focused on increasing sensitivity, minimizing background, and reducing sample preparation. The ability to analyze metabolites directly from biofluids and tissues continues to challenge current MS technology, largely because of the limits imposed by the complexity of these samples, which contain thousands to tens of thousands of metabolites. Among the technologies being developed to address this challenge is Nanostructure-Initiator MS (NIMS),[51][52] a desorption/ ionization approach that does not require the application of matrix and thereby facilitates small-molecule (i.e., metabolite) identification. MALDI is also used however, the application of a MALDI matrix can add significant background at <1000 Da that complicates analysis of the low-mass range (i.e., metabolites). In addition, the size of the resulting matrix crystals limits the spatial resolution that can be achieved in tissue imaging. Because of these limitations, several other matrix-free desorption/ionization approaches have been applied to the analysis of biofluids and tissues.

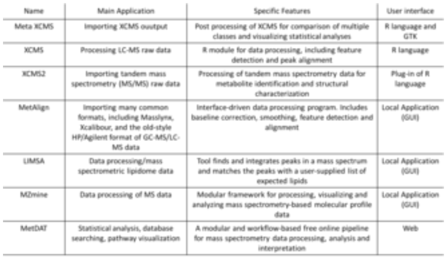
Secondary ion mass spectrometry (SIMS) was one of the first matrix-free desorption/ionization approaches used to analyze metabolites from biological samples. SIMS uses a high-energy primary ion beam to desorb and generate secondary ions from a surface. The primary advantage of SIMS is its high spatial resolution (as small as 50 nm), a powerful characteristic for tissue imaging with MS. However, SIMS has yet to be readily applied to the analysis of biofluids and tissues because of its limited sensitivity at >500 Da and analyte fragmentation generated by the high-energy primary ion beam. Desorption electrospray ionization (DESI) is a matrix-free technique for analyzing biological samples that uses a charged solvent spray to desorb ions from a surface. Advantages of DESI are that no special surface is required and the analysis is performed at ambient pressure with full access to the sample during acquisition. A limitation of DESI is spatial resolution because "focusing" the charged solvent spray is difficult. However, a recent development termed laser ablation ESI (LAESI) is a promising approach to circumvent this limitation. Most recently, ion trap techniques such as orbitrap mass spectrometry are also applied to metabolomics research.[53]

Nuclear magnetic resonance (NMR) spectroscopy is the only detection technique which does not rely on separation of the analytes, and the sample can thus be recovered for further analyses. All kinds of small molecule metabolites can be measured simultaneously - in this sense, NMR is close to being a universal detector. The main advantages of NMR are high analytical reproducibility and simplicity of sample preparation. Practically, however, it is relatively insensitive compared to mass spectrometry-based techniques.[54][55] Comparison of most common used metabolomics methods is shown in the table.

Although NMR and MS are the most widely used, modern day techniques other methods of detection that have been used. These include Fourier-transform ion cyclotron resonance,[56] ion-mobility spectrometry,[57] electrochemical detection (coupled to HPLC), Raman spectroscopy and radiolabel (when combined with thin-layer chromatography).

Statistical methods

The data generated in metabolomics usually consist of measurements performed on subjects under various conditions. These measurements may be digitized spectra, or a list of metabolite features. In its simplest form this generates a matrix with rows corresponding to subjects and columns corresponding with metabolite features (or vice versa).[7] Several statistical programs are currently available for analysis of both NMR and mass spectrometry data. A great number of free software are already available for the analysis of metabolomics data shown in the table. Some statistical tools listed in the table were designed for NMR data analyses were also useful for MS data.[58] For mass spectrometry data, software is available that identifies molecules that vary in subject groups on the basis of mass-over-charge value and sometimes retention time depending on the experimental design.[59]



Name	Main Application	Specific Features	User interface
Meta XCMS	Importing XCMS ouutput	Post processing of XCMS for comparison of multiple classes and visualizing statistical analyses	R language and GTK
XCMS	Processing LC-MS raw data	R module for data processing, including feature detection and peak alignment	R language
XCMS2	Importing tandem mass spectrometry (MS/MS) raw data	Processing of tandem mass spectrometry data for metabolite identification and structural characterization	Plug-in of R language
MetAlign	Importing many common formats, including Masslynx, Xcalibour, and the old-style HP/Agilent format of GC-MS/LC-MS data	Interface-driven data processing program. Includes baseline correction, smoothing, feature detection and alignment	Local Application (GUI)
LIMSA	Data processing/mass spectrometric lipidome data	Tool finds and integrates peaks in a mass spectrum and matches the peaks with a user-supplied list of expected lipids	Local Application (GUI)
MZmine	Data processing of MS data	Modular framework for processing, visualizing and analyzing mass spectrometry-based molecular profile data	Local Application (GUI)
MetDAT	Statistical analysis, database searching, pathway visualization	A modular and workflow-based free online pipeline for mass spectrometry data processing, analysis and interpretation	Web

Software List for Metabolomic Analysis

Once metabolite data matrix is determined, unsupervised data reduction techniques (e.g. PCA) can be used to elucidate patterns and connections. In many studies, including those evaluating drug-toxicity and some disease models, the metabolites of interest are not known *a priori*. This makes unsupervised methods, those with no prior assumptions of class membership, a popular first choice. The most common of these methods includes principal component analysis (PCA) which can efficiently reduce the dimensions of a dataset to a few which explain the greatest variation.[32] When analyzed in the lower-dimensional PCA space, clustering of samples with similar metabolic fingerprints can be detected. PCA algorithms aim to replace all correlated variables by a much smaller number of uncorrelated variables (referred to as principal components (PCs)) and retain most of the information in the original dataset.[60] This clustering can elucidate patterns and assist in the determination of disease biomarkers - metabolites that correlate most with class membership.

Linear models are commonly used for metabolomics data, but are affected by multicollinearity. On the other hand, multivariate statistics are thriving methods for high-dimensional correlated metabolomics data, of which the most popular one is Projection to Latent Structures (PLS) regression and its classification version PLS-DA. Other data mining methods, such as random forest, support-vector machines, etc. are received increasing attention for untargeted metabolomics data analysis.[61] In the case of univariate methods, variables are analyzed one by one using classical statistics tools (such as Student's t-test, ANOVA or mixed models) and only these with sufficient small p-values are considered relevant.[31] However, correction strategies should be used to reduce false discoveries when multiple comparisons are conducted. For multivariate analysis, models should always be validated to ensure the results can be generalized.

Machine learning and data mining

Machine learning is also a powerful tool that can be used in metabolomics analysis. Recently, the authors of a paper published in Analytical Chemistry, developed a retention time prediction software called Retip (https://www.retip.app/). This tool, developed in collaboration with NGALAB (https://www.atens.es/ngalab/), the West Coast Metabolomics Centre (https://metabolomics.ucdavis.edu/) and Riken empower all labs to apply artificial intelligence to the retention time prediction of small molecules in complex matrix, as human plasma, plants, food or microbials. Retention time prediction increases the identification rate in liquid chromatography and subsequently leads to an improved biological interpretation of metabolomics data.[62]

Key applications

Toxicity assessment/toxicology by metabolic profiling (especially of urine or blood plasma samples) detects the physiological changes caused by toxic insult of a chemical (or mixture of chemicals). In many cases, the observed changes can be related to specific syndromes, e.g. a specific lesion in liver or kidney. This is of particular relevance to pharmaceutical companies wanting to test the toxicity of potential drug candidates: if a compound can be eliminated before it reaches clinical trials on the grounds of adverse toxicity, it saves the enormous expense of the trials.[43]

For functional genomics, metabolomics can be an excellent tool for determining the phenotype caused by a genetic manipulation, such as gene deletion or insertion. Sometimes this can be a sufficient goal in itself—for instance, to detect any phenotypic changes in a genetically modified plant intended for human or animal consumption. More exciting is the prospect of predicting the function of unknown genes by comparison with the metabolic perturbations caused by deletion/insertion of known genes. Such advances are most likely to come from model organisms such as *Saccharomyces cerevisiae* and *Arabidopsis thaliana*. The Cravatt laboratory at The Scripps Research Institute has recently applied this technology to mammalian systems, identifying the *N*-acyltaurines as previously uncharacterized endogenous substrates for the enzyme fatty acid amide hydrolase (FAAH) and the monoalkylglycerol ethers (MAGEs) as endogenous substrates for the uncharacterized hydrolase KIAA1363.[63][64]

Metabologenomics is a novel approach to integrate metabolomics and genomics data by correlating microbial-exported metabolites with predicted biosynthetic genes.[65] This bioinformatics-based pairing method enables natural product discovery at a larger-scale by refining non-targeted metabolomic analyses to identify small molecules with related biosynthesis and to focus on those that may not have previously well known structures.

Fluxomics is a further development of metabolomics. The disadvantage of metabolomics is that it only provides the user with steady-state level information, while fluxomics determines the reaction rates of metabolic reactions and can trace metabolites in a biological system over time.

Nutrigenomics is a generalised term which links genomics, transcriptomics, proteomics and metabolomics to human nutrition. In general a metabolome in a given body fluid is influenced by endogenous factors such as age, sex, body composition and genetics as well as underlying pathologies. The large bowel microflora are also a very significant potential confounder of metabolic profiles and could be classified as either an endogenous or exogenous factor. The main exogenous factors are diet and drugs. Diet can then be broken down to nutrients and non-nutrients. Metabolomics is one means to determine a biological endpoint, or metabolic fingerprint, which reflects the balance of all these forces on an individual's metabolism.[66]

See also

- Genomics
- Epigenomics
- Transcriptomics
- Proteomics
- Molecular epidemiology
- Molecular medicine
- Molecular pathology
- Precision medicine
- Fluxomics
- Lipidomics

References

1. Daviss B (April 2005). "Growing pains for metabolomics" (http://www.the-scientist.com/article/display/15427/). *The Scientist*. **19** (8): 25–28.