Chapter 16
Metabolomics and Metabonomics

Metabolism is the ensemble of chemical transformations carried out in living tissue (§10.2); operationally it is embodied in the matter and energy fluxes through organisms. Metabolomics is defined as the measurement of the amounts (concentrations) and locations of the all the metabolites in a cell, the metabolites being the small molecules ($M_r \lesssim 1000$; e.g., glucose, cAMP,[1] GMP,[2] glutamate, etc.) transformed in the process of metabolism (i.e., mostly the substrates and products of enzymes).[3] The quantification of the amounts of expressed enzymes is, as we have seen, proteomics; metabolomics is essentially an extension of proteomics to the activities of the expressed enzymes, and it is of major interest to examine correlations between expression data and metabolite data.[4]

Metabonomics is a subset of metabolomics and is defined as the quantitative measurement of the multiparametric metabolic responses of living systems to pathophysiological stimuli or genetic modification, with particular emphasis on the elucidation of differences in population groups due to genetic modification, disease, and environmental (including nutritional) stress. In the numerous cases of diseases not obviously linked to genetic alteration (mutation), metabolites are the most revealing markers of disease or chronic exposure to toxins from the environment and of the effect of drugs. As far as drugs are concerned, metabonomics is effectively a subset of the investigation of the absorption, distribution, metabolism, and excretion (ADME) of drugs.

[1] Cyclic adenosine monophosphate.

[2] Guanosine monophosphate.

[3] The official classification of enzyme function is that of the Enzyme Commission (EC), which recognizes six main classes: 1, oxidoreductases; 2, transferases; 3, hydrolases; 4, lyases; 5, isomerases; and 6, ligases. The main class number is followed by three further numbers (separated by points), whose significance depends on the main class. For class 1, the second number denotes the substrate and the third number denotes the acceptor; whereas for class 3, the second number denotes the type of bond cleaved and the third number denotes the molecule in which that bond is embedded. For all classes, the fourth number signifies some specific feature such as a particular cofactor.

[4] These correlations are crucial for understanding the links between genome and epigenetics.

J.J. Ramsden, *Bioinformatics,* Computational Biology 10,
DOI 10.1007/978-1-84800-257-9_16, © Springer-Verlag London Limited 2009

Metabonomics usually includes not only intracellular molecules but also the components of extracellular biofluids. Of course, many such molecules have been analysed in clinical practice for centuries; the novelty of metabonomics lies above all in the vast increase of the scale of analysis; high-throughput techniques allow large numbers (hundreds) of metabolites to be analysed simultaneously and repeat measurements can be carried out in rapid succession, enabling the temporal evolution of physiological states to be monitored. The concentrations of a fairly small number of metabolites has been shown in many cases to be so well correlated with a pathological state of the organism that these metabolite concentrations could well serve as the essential variables of the organism, whose physiology is, as we may recall, primarily directed toward maintaining the essential variables within viable limits.

Metabonomics is being integrated with genomics and proteomics in order to create a new systems biology, fully cognizant of the intense interrelationships of genome, proteome, and metabolome; for example, ingestion of a toxin may trigger expression of a certain gene, which is enzymatically involved in a metabolic pathway, thereby changing it, and those changes may, in turn, influence other proteins, and hence (if some of those proteins are transcription factors or cofactors) gene expression.

16.1 Data Collection

The basic principle is the same as in genomics and proteomics: separation of the components followed by their identification. Unlike genomics and transcriptomics, metabonomics has to deal with a diverse set of metabolites even more varied than proteins (which are at least all polypeptides). Typical approaches are to use chromatography to separate the components one is interested in and mass spectrometry to identify them. Alternatively, high-resolution nuclear magnetic resonance spectroscopy can be applied directly to many biofluids and even organ or tissue samples.

Metabolic microarrays operate on the same principle as other kinds of microarrays (§14.1) in which large numbers of small molecules are synthesized, typically using combinatorial or other chemistry for generating high diversity. The array is then exposed to the target, whose components of interest are usually labelled (although their chemical diversity makes this more difficult than in the case of nucleic acids, for example; moreover, the small size of metabolites makes it more likely that the label chemically perturbs them). This technique can be used to answer questions such as "to which metabolite(s) does macromolecule X bind?"

Much ingenuity is currently being applied to determine spatial variations in selected metabolites. An example of a method developed for that purpose is PEBBLES (Probes Encapsulated By Biologically Localized Embedding): fluorescent dyes, entrapped inside larger cage molecules, and which respond (i.e., change their fluorescence) to certain ions or molecules. Their spatial location in the cell can be mapped using fluorescence microscopy. Another example is the development of

high-resolution scanning secondary ion mass spectrometry (“nanoSIMS”), whereby a focused ion beam (usually Cs^+ or O^-) is scanned across a (somewhat conducting) sample and the secondary ions released from the sample are detected mass spectrometrically with a spatial resolution of some tens of nanometres. This method is very favourable for certain metal ions, which can be detected at mole fractions of as little as 10^{-6}. If biomolecules are to be detected, it is advantageous to label the molecule or molecules of interest with non-natural isotopes (e.g., ^{15}N); the enriched molecule can then easily be distinguished via the masses of its fragments in the mass spectrometer.

As far as whole bodies are concerned, the blood is an extremely valuable organ to analyse, since its composition sensitively depends on the state of the organism, to the extent that the blood is sometimes called the “sentinel of the body.”

16.2 Data Analysis

The first task in metabonomics is typically to correlate the presence of metabolites with gene expression. One is therefore trying to correlate two datasets, each containing hundreds of points, with each other. This in essence is a problem of pattern recognition. There are two categories of algorithms used for this task: unsupervised and supervised.

The unsupervised techniques determine whether there is any intrinsic clustering within the dataset. Initial information is given as object descriptions, but the classes to which the objects belong is not known beforehand. A widely used unsupervised technique is principal component analysis (PCA, see §8.3.2). Essentially, the original dataset is projected onto a space of lower dimension; for example, a set of metabonomic data consisting of a snapshot of the concentrations of 100 metabolites is a point in a space of 100 dimensions. One rotates the original axes to find a new axis along which there is the highest variation in the data. This axis becomes the first principal component. The second one is orthogonal to the first and has the highest residual variation (i.e., that remaining after the variation along the first axis has been taken out), the third axis is again orthogonal and has the next highest residual variation, and so on. Very often, the first two or three axes are sufficient to account for an overwhelming proportion of the variation in the original data. Since they are orthogonal, the principle components are uncorrelated (have zero covariance).

In supervised methods, the initial information is given as learning descriptions (i.e., sequences of parameter values (features) characterizing the object whose class is known beforehand).[5] The classes are nonoverlapping. During the first stage, decision functions are elaborated, enabling new objects from a dataset to be recognized, and during the second stage, those objects are recognized. Neural networks are often used as supervised methods.

[5] See, e.g., Tkemaladze.

16.3 Metabolic Regulation

Once all of the data have been gathered and analysed, one attempts to interpret the regularities (patterns). *Simple regulation* describes the direct chemical relationship between regulatory effector molecules, together with their immediate effects, such as feedback inhibition of enzyme activity or the repression of enzyme biosynthesis. *Complex regulation* deals with specific metabolic symbols and their domains. These "symbols" are intracellular effector molecules that accumulate whenever the cell is exposed to a particular environment (cf. Table 16.1). Their domains are the metabolic processes controlled by them; for example, hormones encode a certain metabolic state; they are synthesized and secreted, circulate in the blood and, finally, are decoded into primary intracellular symbols (§16.3.2).

16.3.1 Metabolic Control Analysis

Metabolic control analysis (MCA) is the application of systems theory (§7.1) or synergetics (§7.3) to metabolism. Let $\mathbf{X} = \{x_1, x_2, \ldots, x_m\}$, where x_i is the concentration of the ith metabolite in the cell; that is, the set $\mathbf{X}$ constitutes the metabolome. These concentrations vary in both time and space. Let $\mathbf{v} = \{v_1, v_2, \ldots, v_r\}$, where v_j is the rate of the jth process. To a first approximation, each process corresponds to an enzyme. Then

$$\frac{\mathrm{d}\mathbf{X}}{\mathrm{d}t} = \mathbf{N}\mathbf{v} \,, \tag{16.1}$$

where the "stoichiometry matrix" $\mathbf{N}$ specifies how each process depends on the metabolites. Metabolic control theory (MCT) is concerned with solutions to equation (16.1) and their properties. The dynamical system is generally too complicated for explicit solutions to be attempted, and numerical solutions are of little use unless one knows more of the parameters (enzyme rate coefficients) and can measure more of the variables than are generally available at present. Hence, much current discussion about metabolism centres on qualitative features. Some are especially noteworthy: It is well known, from numerous documented examples, that large changes in enzyme concentration may cause negligible changes in flux through pathways of which they are a part. Metabolic networks are truly many-component systems, as discussed in Chapter 7, and, hence, the concept of feedback, so valuable in dealing with systems of just two components, is of little value in understanding metabolic networks.

Problem. Write $\mathbf{X}$ and $\mathbf{v}$ in equation (16.1) as column matrices and $\mathbf{N}$ as an $m \times r$ matrix. Construct, solve, and discuss an explicit example with only two or three metabolites and processes.

Table 16.1 Some examples of metabolic coding

condition	symbol	domain
glucose deficiency	cAMP	starvation response
N-deficiency	ppGpp	stringent response
redox level	NADH	DNA transcription

16.3.2 The Metabolic Code

It is apparent that certain molecules mediating intracellular function (e.g., cAMP) are ubiquitous in the cell (see Table 16.1). Tomkins has pointed out that these molecules are essentially symbols encoding environmental conditions. The domain of these symbols is defined as the metabolic responses controlled by them. Note that the symbols are metabolically labile and are not chemically related to molecules promoting their accumulation. The concept applies to both within and without cells. Cells affected by a symbol may secrete a hormone, which circulates (e.g., via the blood) to another cell, where the hormone-signal is decoded—often back into the same symbol.

16.4 Metabolic Networks

Metabolism can be represented as a network in which the nodes are the enzymes and the edges connecting them are the substrates and products of the enzymes. There are two main lines of investigation in this area, which have hitherto been pursued fairly independently from one another.

The first line is centred on metabolic pathways, defined as series of consecutive enzyme-catalysed reactions producing specific products; "intermediates" in the series are defined as substances with a sole reaction producing them and a sole reaction consuming them. The complexity of the ensemble of metabolic pathways in a cell is typified by Gerhard Michal's famous chart found on the walls of biochemistry laboratories throughout the world. Current work focuses on ways of rendering this ensemble tractable; for example, a set of transformations can be decomposed into elementary flux modes. An *elementary flux mode* is a minimal set of enzymes able to operate at steady state for a selected group of transformations ("minimal" implies that inhibition of any one enzyme in the set would block the flux). A related approach is to construct linearly independent basis vectors in flux space, combinations of which express observed flux distributions. The extent to which the requirement of a steady state is realistic for living cells remains an open question. In analogy to electrical circuits, use has also been made of Kirchhoff's laws to analyse metabolic networks, especially his first law stating that the sum of all (metabolite) currents at a node is zero.

The second line is to disregard the dynamic aspects and focus on the distribution of the density of connexions between the nodes. The number of nodes of degree k appears to follow a power law distribution (i.e., the probability that a node has

k edges $\sim k^{-\gamma}$).[6] Moreover, there is evidence that metabolic networks thus defined have small world properties (cf. §7.2).

Just as in the abstract networks (automata) discussed previously (Chapter 7), a major challenge in metabolomics is to understand the relationship between the physical structure (the nodes and their connecting edges) and the state structure. As the elementary demonstrations showed (cf. the discussion around Fig. 7.1), physical and state structures are only tenuously related. Much work is still needed to integrate the two approaches to metabolic networks and to further integrate metabolic networks into expression networks. Life is represented by essentially one network, in which the nodes are characterized by both their amounts and their activities, and the edges likewise.

[6] See Wagner & Fell or Raine & Norris.