# Systems Biology and Metabolomics: Experimental and Computational Challenges

M. CAO, A. KOULMAN, D. PACHECO, G.A. LANE AND S. RASMUSSEN

AgResearch Grasslands, Private Bag 11008, Palmerston North, New Zealand

## ABSTRACT

High-throughput technologies in the era of post-genomics provide new opportunities to investigate complex biological systems. Data can be collected for a system of interest at the level of gene expression, protein expression and metabolite concentration changes and then used for the assessment of interactions of molecules. However, the vast amount of information also poses challenges for data management, analysis, and interpretation. This paper aims to outline the technical complexity and the sources of variation in the technologies involved in molecular systems biology and computational strategies for data integration.

**Keywords:** metabolomics; systems biology; computational biology.

## INTRODUCTION

Most economically important traits, such as milk composition and yield, meat and fibre yield are complex traits. These traits are controlled to varying degrees by complex genetic systems and affected by environmental factors. Understanding the mechanisms underlying these complex phenotypes is of fundamental importance in biology and will provide new clues and strategies for trait improvement.

However, identifying relevant genes has been difficult because the genetic effect of each causal gene is often modest and only a small number of genes can be studied at a time by traditional genetics. The interactions of genes are often neglected and difficult to detect (Carlborg & Haley, 2004). The completion of the genome sequence projects in a number of species now makes it possible to investigate genome-wide gene interactions for complex traits (Hirschhorn & Daly, 2005). Techniques developed in the era of post-genomics (transcriptomics, proteomics and metabolomics) make it practical to gather information on complex systems from different perspectives. For investigation of a particular complex system, microarrays provide a technical platform in transcriptomics to capture snapshots of gene expression in the genomic scale; while mass spectrometry and other spectroscopic tools adopted in proteomics and metabolomics can evaluate protein expression and monitor the variations in the accumulation of small metabolites on a large scale.

The advent of these high-throughput techniques calls for good practice in data management and novel approaches for making sense of data. Systems biology is devoted to providing new research paradigms for integrating data from different levels of analysis to understand the complex biological systems at the molecular level. The aim of this paper is to outline the technical complexity and the sources of variation in the technologies involved and computational strategies for data integration.

## METABOLOMICS AND OTHER COMPONENTS IN SYSTEMS BIOLOGY

The use of systematic genomic, proteomic and metabolomic technologies to construct models of complex biological systems is becoming a common practice, which is collectively known as systems biology (Ideker *et al*., 2001; Kitano, 2002a, 2002b). Although the notation of "systems" science has existed for some time in traditional biology (e.g. ecology), the goal, research scope and methodology of systems biology in the context of genomics is still under current discussion (Kirschner, 2005, and other commentaries in the same issue of *Cell*). Systems approaches have recently become far more powerful in molecular biology because of the availability of new high-throughput and quantitative experimental technologies (see commentary: Ideker, 2004).

The genome sequencing projects had a very clear objective, the determination of the complete genome sequences of the species in question. The objective was ambitious but proved achievable through remarkable technical advances and huge international collaborative efforts (Collins *et al*., 2003). However, the initial goal of proteomics and metabolomics, which is the rapid determination of the "complete" set of proteins or metabolites expressed in a cell or tissue has yet to be achieved for any species (Patterson & Aebersold, 2003; Summer *et al*., 2003).

Genome-wide determination of functions of genes and their products is a more challenging

task than deciphering static genetic codes. It is difficult to catalogue the properties and functions of all the proteins in the proteome and all the metabolites in the metabolome, not only because the current analytical methodologies are limited in their capacity to detect and characterize all the molecules in the cell or tissue but also the interpretation of the functions is impossible without knowledge of the biological context. Compared with structural genomics, data collected from high-throughput techniques in functional genomics are complex and more prone to errors. Details of both pre-processing and biological context must be recorded for effective data sharing and reproducible research. The complexity of the data makes it difficult to deposit and query from a relational database as the metadata (data about data) must be accompanied with numerical measurements. Data standards (data models) have been established to facilitate information exchange in functional genomics including MIAME (Brazma *et al*., 2001) for microarray experiments, PEDRO (Taylor *et al*., 2003) and PSI (Orchard *et al*., 2004) for proteomics and ArMet (Jenkins *et al*., 2004) for metabolomics. EXtensible Markup Language (XML) has become the technique of choice for data exchange because of its capability and flexibility for describing data. Computational improvements in the storage and query of data in XML formats has been an important factor in its adoption as a mainstream data exchange technology.

There is no single technology platform that can satisfy all of the desired measurements of expression of genes, proteins and metabolites on the genomic scale. Microarrays have been one of the main tools for investigating gene functions by providing a global view of transcription. Customized cDNA arrays and Affymetrix oligo arrays are the two main platforms. Tiling arrays were originally designed for unbiased detection of transcription and unexpected transcription active sites were often found (Bertone *et al*., 2004). SAGE (serial analysis of gene expression) provides an alternative to chip-based techniques. Low throughput RT-PCR is often considered as a gold standard to verify microarray results.

In contrast to Watson-Crick base-paired nucleic acids, proteins and metabolites have diverse structural and functional properties which make the development of technology platforms for comprehensive proteomic and metabolomic profiling intrinsically much more difficult. Many efforts are being undertaken for their improvement. Any current practical analysis provides only a limited view of the proteome and metabolome. Two-dimensional gel electrophoresis (2DE), MALDI-TOF mass spectrometry, LC-MS/MS and chip-based methods are being used in proteomics for protein measurements (Patterson & Aebersold, 2003). The two-hybrid system enables the construction of a map of interactions among proteins (Tucker *et al*., 2001). Depending on the properties of metabolites of interest, a range of chromatography-coupled analytical techniques (GC-MS, LC-MS and LC-MS/MS), FT-MS and NMR could be selected to provide broad-spectrum metabolite profiles (Weckwerth, 2003; Kell, 2004). Techniques such as stable isotope labelling GC-IRMS or NMR can provide additional information on metabolic fluxes. An understanding of the machinery used for the high-throughput data generation is a prerequisite to coping with technical sources of variation.

The primary experimental goal of "omics" technologies is the non-targeted identification and global profiling of mRNA, proteins and metabolites present in specific biological samples. It is common to see these technologies reveal unexpected properties of biological systems. The global data sets are rich in information but difficult to analyze using traditional knowledge-based interpretation (Patterson & Aebersold, 2003). New methods must be employed or developed to extract biological insights or to formulate hypotheses.

## COMPUTATIONAL CHALLENGES

Identification and measurement of biological variance of interest among all other biological and non-biological variances (environmental factors, instrumentation etc.) is a key requirement of quantitative biological research. However, the challenges are magnified in the domain we are investigating as the sources of variation are manifold.

### Sources of variations

The understanding and proper assessment of these variations are prerequisites for effective data analysis and integration. General good practice from sample preparation to instrument manipulation is essential. High-throughput assays in functional genomics typically produce noisy data. The process of sample preparation is more complicated at least in microarray experiments. Conditions of DNA preparation and hybridizations need to be optimised beforehand.

Modern instruments also come with sophisticated software, experience is required to handle the machines and optimise parameters for data acquisition. Sources of variation in microarray experiments have been well documented (Kerr & Churchill, 2001). Careful experimental designs, such as dye-swap, have been

advocated to reduce the variation due to dye-bias. It may be relatively easy to identify sources of variation, but it is often tricky to choose appropriate approaches to eliminate variations in high-throughput data generation. Technical limitations and sources of variation such as batch effects need to be considered during statistical consultations on experimental design and data processing to make a cost-effective choice. For example, relying on experimental design to remove batch effect may be more costly than choosing a proper data pre-processing procedure to reduce the effect.

As said, commercial software is often provided by manufacturers for data capture and preliminary data processing. It is important to understand these low-level data processing steps (e.g. missing value treatment) to ensure high-quality data generation. Although algorithms for signal quantification, peak alignment, deconvolution and structural identification are currently implemented in the commercial software packages, these algorithms are still the subjects of active research in the area of chemometrics (Johnson *et al*., 2003) and proteomics (Listgarten & Emili, 2005). The best use of some complicated algorithms, peak alignment for example, requires human intervention. Therefore, it may be necessary to start from raw data rather than relying on the default data pre-processing provided by instrument manufacturers as novel algorithms continue to be invented. Fortunately, it is now feasible to access those raw data in proprietary formats on a range of analytical instrumentations thanks again to XML technology (Pedrioli *et al*., 2004).

**Make Sense of Data -- The Art of Modelling**

Once the data collected for studying a complex biological system are properly cleaned up and pre-processed, exploratory data analysis is usually the first step to search for structures and patterns in the data. Unsupervised learning algorithms including hierarchical clustering, k-means and self-organizing map (SOM) etc. can be used. Investigation of clusters within the data may give new insight into the systems often with the aid of other prior information about the individuals in each cluster. All clustering algorithms rely on distance metrics to measure the similarity between individual observations or groups. The pattern of clustering found depends on the choice of distance metric, such as Euclidean, Manhattan, Pearson correlation coefficient, just to name a few. Co-clustering (co-occurrence) analysis of genes, proteins and metabolites data may reveal coherent relationships among these variables, which may be unexpected and provide the basis of new hypothesis formation. Supervised learning algorithms take advantage of current knowledge to guide learning process to achieve better classification and prediction. Brown *et al*. (2000) used support vector machine (SVM) learning algorithms to re-classify genes from the same data set (Eisen *et al*., 1998) by incorporating functional categories calculated from MIPS yeast genome database (MYGD) and showed that SVM outperformed many unsupervised algorithms. Information on pathways of genes, proteins and metabolites, their cellular location and functional categories can be incorporated into supervised learning processes to obtain models with better predictive power.

Genes, proteins and metabolites interact with themselves and each other to manifest cellular functions. The reconstruction of cellular networks and exploration of universal laws that govern structure and dynamics of molecular networks is developing as the emerging field of network biology (Barabasi & Oltvai, 2004). Network models can be inferred from transcriptomics, proteomics and metabolomics data respectively (Samoilov *et al*., 2001, Segal *et al*., 2003) or joint analysis of the data. Combined analysis of microarray, proteomics and/or metabolomics data has been used to identify novel gene function and regulatory networks. The approaches to derive network structure include multiple regressions (Gardner *et al*., 2003), correlation (Stuart *et al*., 2003) and Bayesian network models (Friedman, 2004). Evaluation (model selection) is usually done by ranking models with score functions, such as Bayesian information criteria. The best model is selected with the highest score (Segal *et al*., 2003) or a combined model is pursued (Hastie *et al*., 2001).

There are pitfalls to be avoided in network modelling. Searching network structure from *de novo* is computationally expensive or prohibitive, so integration of prior knowledge is an efficient way to set constraints in searching for optimal network models. Prior knowledge here includes cellular location, DNA/protein sequence, pathway information and exhaustively-searched literatures. Defining the right granularity for modelling is another important issue. Granularity here refers to the level of detail or abstraction at which a particular problem is analyzed. Specifying the right level of details, which often requires sophisticated knowledge in the domain being modelled, is a key to finding effective solutions for a complex problem. The same mathematical theory can be used for totally different domains, but the interpretation of the models will suffer if

the granularity is not carefully defined. For example, a network may be modelled as a graph G (V, E) where V denotes node and E edges in the graph (G). In terms of a metabolic network, metabolites correspond to nodes and reactions correspond to edge(s) between nodes. For gene interactions, it may be important to select a subset of genes rather than all genes in the whole genome to model their interactions. Granularity can be addressed by asking questions like: what kinds of relationships do we need to model and what details about the relations will give us sufficient power of prediction and interpretation? The answer to these questions depends on both prior knowledge and the availability of data. The relationships may be modelled as Boolean (one gene turns on/off another) or other logic relationships (fuzzy logic); correlation (one gene is highly correlated to another); probabilistic (one gene affects another with probability of 0.8, for example), or kinetic (reaction rate) etc.

Each "omics" provides one side of information towards the understanding of a complex system. This review is far from complete as there are several missing links which need to be connected for functional characterization of cellular molecules and their interactions. For example, molecular fluxes through metabolic networks have their role in the determination of cellular phenotypes. High-throughput flux analysis has been reviewed elsewhere (Sauer, 2004). Each "omics" experiment alone has a limited utility. The integration of data from different sources helps reinforce *bona fide* observations and reduce false negatives (Hwang *et al*., 2005). However, data integration in systems biology from different observations and prior knowledge is not a trivial task. The lack of suitable software tools and resources currently limits essentially all areas of proteomics (and metabolomics) data analysis from searching MS/MS spectra of known metabolites to deriving value from data that goes beyond an initial scan for "interesting observations" (Patterson & Aebersold, 2003).

## DISCUSSION AND CONCLUSIONS

It is imperative to develop new paradigms to understand complex traits and provide detailed mapping from genotypes to phenotypes (Botstein & Risch, 2003). The mechanisms behind the adaptability and versatility of complex biological system need to be elucidated to establish more reliable biomarkers for the prediction of complex traits. The integration of the "omics" information will help toward understanding cellular interactions of molecules and bridging gaps between genotypes and phenotypes.

However, complex traits, such as lactation, growth rate and stress resistance, in multi-cellular organisms operate across a range of scales from intracellular, to intercellular to inter-tissue/organ, and this requires integrating understanding across these different scales. For lactation to occur, communication networks among liver, muscle, adipose and mammary tissues are needed to orchestrate the appropriate metabolic responses (Lucy, 2004; Ribaux *et al*., 2002). For example, the liver senses nutrient supply to the body and elicits an endocrine message to communicate to the rest of the body how nutrients are partitioned away from muscle and adipose tissue to the lactating mammary gland (Lobley *et al*., 2000). However, this control slides back towards muscle and adipose tissue as lactation advances. Understanding the molecular networks underpinning these "switches" is challenging in modelling but the key to establish hypotheses on how to manipulate specific networks to modify lactation and nutrient partitioning in a predictable manner.

Techniques (e.g. from nanotechnology to real-time imaging) will continue to be developed to collect quality data toward understanding complex biological systems. While we agree that data will need to be collected at fine resolution in space and time to reveal subtle interactions and dynamics, it is also important to take full advantage of current capabilities of data collection. The sophistication of data produced in transcriptomics, proteomics and metabolomics is enough to justify increased effort to develop improved computational tools to aid knowledge discovery. However, because of the complexity and diversity of biological systems, it will not make sense to try to create a one-size-fits-all software solution. We need to do better in developing effective and devoted collaborations among biologists, chemists, mathematicians and computer scientists.

## ACKNOWLEDGEMENTS

The authors acknowledge the support of the Foundation for Research Science and Technology through the programme "New opportunities from Forage Plant Genomics".

## REFERENCES

Barabasi, A.L.; Oltvai, Z.N. 2004: Network biology: understanding the cell's functional organization. *Nature reviews genetics 5*: 101-113

Bertone, P.; Stolc, V.; Royce, T.E.; Rozowsky, J.S.; Urban, A.E.; Zhu, X.; Rinn, J.L.; Tongprasit,

W.; Samanta, M.; Weissman, S.; Gerstein, M.; Snyder, M. 2004: Global identification of human transcribed sequences with genome tiling arrays. *Science 306*: 2242-2246

Botstein, D.; Risch, N. 2003: Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature genetics 33* (Suppl.) 228-237

Brazma, A.; Hingamp, P.; Quackenbush, J.; Sherlock, G.; Spellman, P.; Stoeckert, C.; Aach, J.; Ansorge, W.; Ball, C.A.; Causton H.C.; Gaasterland, T.; Glenisson, P.; Holstege, F.C.P.; Kim, I.F.; Markowitz, V.; Matese, J.C.; Parkinson, H.; Robinson, A.; Sarkans, U.; Schulze-Kremer, S.; Stewart, J.; Taylor, R.; Vilo, J.; Vingron, M. 2001: Minimum information about a microarray experiment (MIAME) - toward standards for microarray data. *Nature genetics*. *29*: 365-371

Brown, M,P.; Grundy, W.N.; Lin, D.; Cristianini, N.; Sugnet, C.W.; Furey, T.S.; Ares, M. Jr.; Haussler, D. 2000: Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences, USA 97*: 262-267

Carlborg, Ö.; Haley, C.S. 2004: Epistasis: too often neglected in complex traits studies. *Nature review genetics 5*: 618-625

Collins, F.S.; Morgan, M.; Patrinos, A. 2003: The human genome project: lessons from large-scale biology. *Science 300*: 286-290

Friedman, N. 2004: Inferring cellular networks using probabilistic graphical models. *Science 303*: 799-805

Eisen, M.B.; Spellman, P.T.; Brown, P.O.; Bostein, D. 1998: Cluster analysis and display of genome-wide expression patters. *Proceedings of the National Academy of Sciences, USA 95*: 14863-14868

Gardner, T.S.; di Bernardo, D.; Lorenz, D.; Collins, J.J. 2003: Inferring genetic networks and identifying compound mode of action via expression profiling. *Science 301*: 102-105

Hastie, T.; Tibshirani, R.; Friedman, J. 2001: The elements of statistical learning. Data mining, inference, and prediction. Springer series in statistics

Hirschhorn, J.N.; Daly, M.J. 2005: genome-wide association studies for common diseases and complex traits. *Nature reviews genetics 6*: 95-108

Hwang, D.; Rust, A.G.; Ramsey, S.; Smith, J.J.; Leslie, D.M.; Weston, A.D.; de Atauri, P.; Aitchison, J.D.; Hood, L.; Siegel, A.F.; Bolouri, H. 2005: A data integration methodology for systems biology. *Proceedings of the National Academy of Sciences, USA 102(48)*: 17296-17301

Ideker, T.; Galitski, T.; Hood, L. 2001: A new approach to decoding life: Systems Biology. *Annu. rev. genomics hum. genetics 2*: 343-372

Ideker, T. 2004: Systems biology 101 – what you need to know. *Nature Biotechnology 22(4)*: 473-475

Jenkins, H. Hardy N.; Beckmann, M.; Draper, J.; Smith, A.R.; Taylor, J.; Fiehn, O.; Goodacre, R.; Bino, R.J.; Hall, R.; Kopka, J.; Lane, G.A.; Lange, B.M.; Liu, J.R.; Mendes, P.; Nikolau, B.J.; Oliver, S.G.; Paton, N.W.; Roessner-Tunali, U.; Saito, K.; Smedsgaard, J.; Sumner, L.W.; Wang, T.; Walsh, S.; Wurtele, E.S.; Kell, D.B. 2004: A proposed framework for the description of plant metabolomics experiments and their results. *Nature biotechnology 22(12)*: 1601-1606

Johnson, K.J.; Wright, B.W.; Jarman K.H.; Synovec R.E. 2003: High-speed peak matching algorithm for retention time alignment of gas chromatographic data for chemometric analysis. *Journal of chromatography A, 996*: 141-155.

Kell, D.B. 2004: Metabolomics and systems biology: making sense of the soup. *Current opinion in microbiology 7*:296-307

Kerr, M.K.; Churchill, G.A. 2001: Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments. *Proceedings of the National Academy of Sciences, USA 98(16):* 8961-8965

Kirschner, M.W. 2005: The meaning of systems biology. *Cell 121(4)*: 503-504

Kitano, H. 2002a: Looking beyond the details: a rise in system-oriented approaches in genetics and molecular biology. *Current genetics 41*: 1-10

Kitano, H. 2002b: Computational systems biology. *Nature 420*: 206-210

Listgarten, J.; Emili, A. 2005: Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Molecular & Cellular Proteomics 4.4*: 419-434

Lobley, G.E.; Milano, G.D.; Van der Walt, J.G. 2000: The Liver: Integrator of Nitrogen Metabolism. In *Ruminant Physiology: Digestion, Metabolism, Growth, and Reproduction* (Ed. Cronjé, P. B.) 149-168. Wallingford, UK, New York: CABI Pub.

Lucy, M.C. 2004: Mechanisms linking the somatotropic axis with insulin: Lessons from the postpartum dairy cow. *Proceedings of the New Zealand Society of Animal Production 64*:19-23

Orchard, S., Hermjakob, H.; Apweiler, R. 2003: The proteomics standards initiative. *Proteomics 3*: 1374-1376

Patterson, S.; Aebersold, R. H. 2003: Proteomics: the first decade and beyond. *Nature genetics 33* (Suppl.) 311-323

Pedrioli, P.G.A.; Eng, J.K.; Hubley, R.; Vogelzang, M.; Deutsch, E.W.; Raught, B.; Pratt, B.; Nilsson, E.; Angeletti, R.; Apweiler, R.; Cheung, K.; Costello, C.E.; Hermjakob, H.; Huang, S.; Julian, R.K. Jr.; Kapp, E.; McComb, M.E.; Oliver, S.G.; Omenn, G.; Paton, N.W.; Simpson, R.; Smith, R.; Taylor, C.F.; Zhu, W.; Aebersold, R. 2004: A common open

representation of mass spectrometry data and its application in a proteomics research environment. *Nature biotechnology 22(11)*: 1459-1466

Ribaux, P.; Gjinovci, A.; Sadowski, H.B.; Iynedjian, P.B. 2002: Discrimination between signaling pathways in regulation of specific gene expression by insulin and growth hormone in hepatocytes. *Endocrinology 143(10):* 3766-3772

Samoilov, M.; Arkin, A.; Ross, J. 2001: On the deduction of chemical reaction pathways from measurements of time series of concentrations. *Chaos 11(1):*108-114

Sauer, U. 2004: High-throughput phenomics: experimental methods for mapping fluxomes. *Current opinion in biotechnology 15*: 58-63

Segal, E.; Shapira, M.; Regev, A.; Pe'er, D.; Bostein, D.; Koller, D.; Friedman, N. 2003: Module networks: identify regulatory modules and their condition-specific regulators from gene expression data. *Nature genetics 34(2)*: 166-176

Stuart, J.M.; Segal, E.; Koller, D.; Kim, S.K. 2003: A gene-coexpression netwrok for global discovery of conserved genetic modules. *Science 302(10):* 249-255

Summer, L.W.; Mendes, P.; Dixon, R.A. 2003: Plant metabolomics: large-scale phytochemistry on the function genomics era. *Phytochemistry* 62: 817-836

Taylor, C.F.; Paton, N.W.; Garwood, K.L.; Kirby P.D.; Stead, D.A.; Yin, Z.; Deutsch, E.W.; Selway, L.; Walker, J.; Riba-Garcia, I.; Mohammed, S.; Deery, M.J.; Howard, J.A.; Dunkley, T.; Aebersold, R.; Kell, D.B.; Lilley, K.S.; Roepstorff, P.; Yates III, J.R.; Brass, A.; Brown, A.J.P.; Cash, P.; Gaskell, S.J.; Hubbard S. J.; Oliver S.G. 2003: A systematic approach to modeling, capturing, and disseminating proteomics experimental data. *Nature biotechnology 21*: 247-254

Tucker, C.L.; Gera, J.F.; Uetz, P. 2001: Towards an understanding of complex protein networks. *Trends in cell biology 11*: 102-106

Weckwerth, W. 2003: Metabolomics in systems biology. *Annu. rev. plant biol. 54*: 669-689