

The challenges of modeling mammalian biocomplexity

Jeremy K Nicholson¹, Elaine Holmes¹, John C Lindon¹ & Ian D Wilson²

Understanding the relationships between human genetic factors, the risks of developing major diseases and the molecular basis of drug efficacy and toxicity is a fundamental problem in modern biology. Predicting biological outcomes on the basis of genomic data is a major challenge because of the interactions of specific genetic profiles with numerous environmental factors that may conditionally influence disease risks in a nonlinear fashion. 'Global' systems biology attempts to integrate multivariate biological information to better understand the interaction of genes with the environment. The measurement and modeling of such diverse information sets is difficult at the analytical and bioinformatic modeling levels. Highly complex animals such as humans can be considered 'superorganisms' with an internal ecosystem of diverse symbiotic microbiota and parasites that have interactive metabolic processes. We now need novel approaches to measure and model metabolic compartments in interacting cell types and genomes that are connected by cometabolic processes in symbiotic mammalian systems.

Human populations face many diverse and aggressive biological challenges, including new infectious agents, antibiotic resistance, the increased incidence of cancer and age-related neurodegenerative conditions and the rapid and insidious rise in insulin resistance. All these problems involve interactions of multiple gene loci, environmental factors and, in many cases, interacting nonhuman genomes. In the quest to improve our understanding of disease processes, researchers have applied advanced analytical platforms to generate new physiological information to complement data supplied by modern genomics^{1,2}. The hope is that judicious use of genomic knowledge within a framework of physiology and metabolism will yield improvements in the health of whole populations and in the health of individuals by personalized healthcare solutions¹⁻⁴.

The growth of a wide range of 'omics' sciences enables the measurement of multiple features of complex systems at various levels of biomolecular organization from the cell to the whole organism^{3,4}. However, these technologies generate massive amounts of data and it is a major task to model these robustly in a way that allows predictive disease modeling. This is a particular challenge because of the level of complexity of the mammalian system in its entirety, with its many spatially heterogeneous arrays of disparate cell types. Thus, the question is what needs to be measured and modeled to describe the integrated function of the system in a way that can be used to predict modes of failure accurately.

In this review, we consider some aspects of mammalian biocomplexity that are currently poorly understood, but may be of great

importance in understanding certain aspects of human disease development and drug action or drug toxicity. We first examine temporal and spatial variation in data, then describe the hierarchy of different systems that can be modeled, including multiple genome interactions, and then conclude by discussing trends in the modeling of systems of increasing levels of complexity.

Timescales of 'omics' events

To measure a system, even at the single-cell level, one must first understand the time-displacement that exists between gene, protein, metabolic and physiological events and their end points³. This is one of the confounding issues to be gauged when relating gene expression data with, for example, proteomic data using classic correlation methods or multivariate statistics. Attempts to achieve correlative 'omics' have often proved to be unsatisfactory even for simple systems such as yeasts⁵. In addition, the timescales of various biological control functions are either very different or simply unknown. **Figure 1** illustrates a theoretical view of the problems that can occur when trying to cross-correlate protein levels with rises in gene products. It should also be remembered that the levels of mRNA transcripts (which have highly variable half-lives in any case) are indirect measures of genome activity, and they in turn relate to the switching on or off of genes, which operate on other undetermined timescales. At the single-cell level, gene events can be considered to be 'quantized,' that is, either on or off at any given time.

Understanding the true quantitative relationship between the variation in activity of every one of the thousands of hypothetical gene-protein couples in a cellular system is complicated by the time displacement of the genetic and protein synthetic and post-translational events, their different timescales and their half-lives; the frequency and times of measurement of the transcripts and the proteins can markedly alter the modeled statistical relationships between these variables and, therefore, the conclusions to be drawn. Given that the

¹Biological Chemistry, Biomedical Sciences Division, Imperial College London, Sir Alexander Fleming Building, South Kensington, London SW7 2AZ, UK.

²Dept. of Drug Metabolism and Pharmacokinetics, AstraZeneca Pharmaceuticals, Mereside, Alderley Park, Macclesfield, Cheshire SK10 4TG, UK. Correspondence should be addressed to J.K.N. (j.nicholson@imperial.ac.uk).

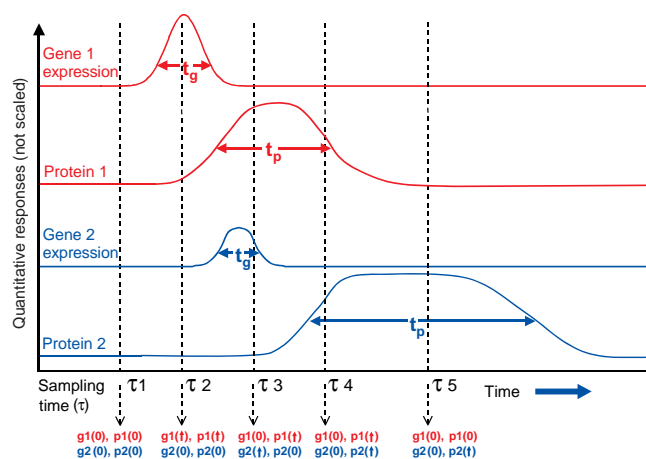


Figure 1 Time courses for two hypothetical gene–protein couples (relating transcription activity to protein level) that are up-regulated following a system stimulus, such as a drug intervention at time zero. It can be seen that if the action of the stimulus is at the genetic level in the first instance it will take a finite amount of time in a cell for the associated protein synthesis (or post-translational modification) to occur and that the duration of the gene events (t_g) and protein events (t_p) may be very different. The practical consequence of this is that the observed co-variance of the gene and protein events is highly dependent on sampling time point and frequency and in some instances a single sampling point, say τ_3 might lead to the incorrect assumption that gene 2 co-varied with protein 1. The differential displacement of the times of maximal activity or expression cannot be assumed to be constant and the variation in possible values for t_p and related turnover times is known to be large. Key: $g_1(0)$, $p_1(0)$ etc describe the relative condition with respect to up regulation of each gene and protein at a given time-point (e.g. τ_1 , τ_2 , etc.), where (0) indicates baseline level and (\uparrow) indicates upregulation. Thus the post intervention observation of relative state can be seen to be dependent on sampling time-point.

timescale of a gene switching event itself is short and difficult to measure and the half-life of a cytosolic protein can vary from a few to hundreds of hours^{6,7}, it is easy to understand why transcriptomic and proteomic data from the same system do not always agree. Another weakness when attempting to correlate proteomic with, say, metabolomic data, is the fact that currently proteomic data sets contain no information on the activity of specific proteins, which is dependent on their location in the cell and the presence of cofactors or inhibitors.

An example of this problem is provided by work on the effects of orotic acid, an endogenous metabolite that when administered to mammals causes profound fatty changes in the liver⁸. Both transcripts and metabolites have been measured and cross-correlated in animals treated with orotic acid. Statistically, there was little agreement between the level of gene expression and the level of many metabolites in the liver, urine or plasma. However, in some lipid metabolism pathways in the system, connections between changes in gene regulation and metabolite levels could be rationalized in biochemical terms⁸. In simpler systems, it has been shown that metabolic data can be effectively used to generate functional genomic information and to uncover the phenotype of silent mutations, thus truly integrating the omics sciences involved⁹. This is highly appropriate for relatively static systems, such as cells in culture (assuming a stationary growth phase), but hypercomplex systems, such as mammals, which function with many interacting and spatially dispersed cell types and show constant time-related variations, require more sophisticated approaches where detailed time-responses must be measured.

For metabolic studies, it is usually crucial to measure the timed responses of the system to obtain a complete and evolving picture of metabolic injury after toxic insult or during a disease process^{10–16}. For practical reasons, this may not be economical using DNA microarray technology or proteomics, although this may change in the future. Because of differences in the magnitude of effects in, for example, comparative species studies in toxicology^{14,15}, data comparisons and the relative scaling of time-related metabolic phenomena also present problems (but these are not confined to metabolic studies). In the field of metabonomics^{3,4} we are observing complex or multicellular system responses to stressors, and the measurement of integrated time-related changes is intrinsic to this approach.

Overall, omics approaches should lead to the formulation of new hypotheses about pathway control and dysfunction, but they do not actually prove that a given data model provides a complete or even a correct biological explanation of the observed condition. Omics-generated hypotheses should of course be tested to make sure that the

biological understanding is complete; however, this is not frequently done. When omic-generated hypotheses are tested properly, the results can be rewarding; for example, it was possible to deduce the mechanism of toxicity of a drug that failed regulatory approval using a metabonomic approach that identified by timed post-dose biofluid analysis key points in mitochondrial metabolism that were disrupted *in vivo*¹⁷. This was then tested *in vitro* proving that the hypothesis of the toxic mechanism generated by the exploratory metabolic studies was correct¹⁷.

In the future, as both transcriptomic and proteomic technologies become faster and cheaper, it should be possible to measure time-related fluctuations in much more detail leading to new levels of understanding. Indeed, one can already consider what such data might look like and how they could be analyzed. Thus, at another level of time-related complexity in cellular responses to a stressor, the geometric shape of the time response (rather than just its magnitude) could be considered as a descriptor and the mean time to recovery of gene expression in a reversibly perturbed system, for example, after a low dose of a drug, would also be new parameters that could be modeled (Fig. 2). Thus, the time course/activity graph for a given species (mRNA, protein or metabolite) after the initial perturbation can take on many forms and Figure 2 illustrates some hypothetical time responses of mRNA levels for a series of genes after administration of a drug causing a reversible pharmacological or toxicological effect.

This is an interesting concept as one could consider a new set of time-related parameters for the DNA microarray set that would describe the effects of the overall perturbation on the basis not only of the level of expression but also of how long levels of expression take to reach equilibrium after the intervention; we term this the ‘genetic (or proteomic) relaxation time’ (which could have individual values for each gene or protein, but also global values for recovery of the system). The time-related relaxation patterns (with values in seconds or hours) would then provide an alternative signature of response to a specific intervention that did not rely on measurement of magnitudes of change.

Such differential geometric time responses can already be observed in real metabolic data showing responses to drug or toxin treatments. Because in many cases there must be time-displaced relationships among metabolite level changes and transcript and protein levels, we may surmise that similar responses would be observed in transcriptomic and proteomic data. These vary from time responses of simple mathematical form, where a particular metabolite closely

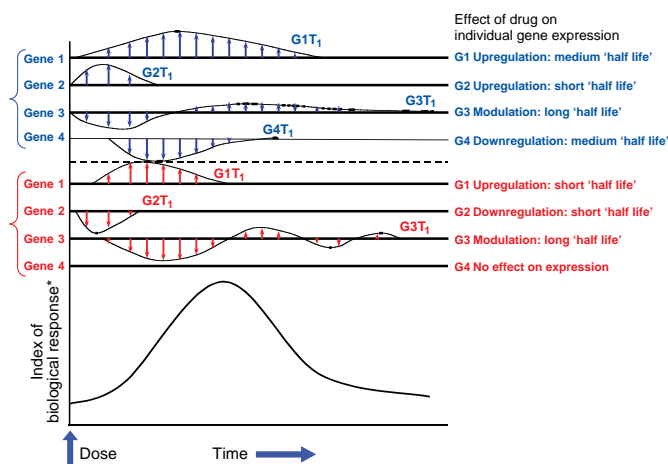


Figure 2 The 'genetic relaxation time' hypothesis. Different genes will have different up- or down-regulation responses to a given stimulus, in this case a drug intervention, but the detailed time course of such responses will also carry information about the nature of the biological interaction that is not specifically dependent on the degree of expression but on the relationships between individual gene activities throughout the time-course. Similar arguments can be applied to protein and metabolite levels and time-courses. The time to return to equilibrium (in a reversible process) for a particular gene activity (or transcript level) can be termed the relaxation time, T_1 , the relationships between the T_1 s of many genes in the system will be characteristic of the particular intervention. $G1T_1$, $G2T_1$ represent the specific times of each 'gene' to recover equilibrium following a stressor or intervention. An independent indicator of biological response might be a plasma enzyme level that marks the level of parenchymal damage following a toxic insult.

follows the onset and recovery from the lesion, to complex forms such as damped oscillations as multiple organ systems interact to establish post-traumatic homeostatic control (Fig. 3).

Single cells, multicellular organisms and 'super-organisms'

It is important to measure system responses of stimuli (drug or disease) through time as this can show both the development of a lesion with multiple organ effects occurring at different times and also the recovery process. This is particularly important where processes involve multiple cell or tissue systems that vary in metabolic composition and biosynthetic activity. When a noxious stimulus occurs, there will be not only site- and cell-specific effects but also downstream consequences in terms of changed interactions with other tissue and organ systems.

In a single-cell system, such as a bacterial or plant cell culture, or in a primary tissue cell line (Fig. 4a), the systems biology challenge is to understand the upstream and downstream relationships between the transcripts, proteins and intracellular and extracellular metabolites. The metabolite concentrations are influenced by substrate availability in the extracellular milieu, the enzyme activities and presence of cofactors within the cell and the activities of membrane transporters.

In multicellular organisms with many interacting cell types (Fig. 4b), the system is regulated at the cellular level and higher levels that are also under neurohormonal control. Here we can consider metabolic linkages of disparate nature that are dispersed through the medium of the extracellular fluids such as blood plasma and lymph, but are also removed through multiple discrete secretory and excretory drains on the metabolite pool such as urine, bile and sweat. Sampling these fluids also gives many clues as to what is happening in the integrated system, but can be more complex to map against

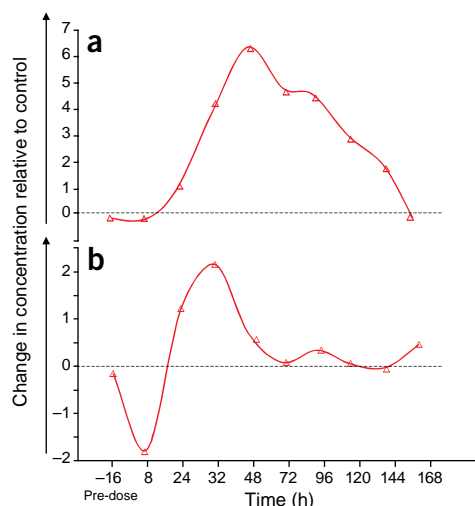


Figure 3 Time-related concentrations of two metabolites, a and b, that are perturbed by dosing an animal with a drug (in this case, these are from real data sets generated from liquid chromatography-mass spectrometric analysis of sequential urine samples taken from an animal dosed with a model liver toxin). One marker (a) has a simple response that corresponds exactly with the level of maximal liver damage. The other (b) gives a complex damped oscillation characteristic of system overshoot due to multiple tissue interactions and the re-establishment of homeostasis over many hours.

specific pathway activity because analytical data represent the weighted average of the whole system. However, such measurements provide a means of studying the whole system's responses collectively¹⁵.

Mammalian-microbe-environment interactions. It is widely accepted that most major disease classes have significant environmental and genetic components and that the incidence of disease in a population or an individual is a complex product of the conditional probabilities of certain gene combinations interacting with a diverse range of environmental triggers.

Diet clearly has a major influence on many diseases and modulates the complex internal community of gut microorganisms^{18,19} (the particular microbial community in an individual mammalian host is referred to as the microbiome^{20,21}). These microorganisms, weighing up to 1 kg in a normal adult human, may total ~100 trillion cells²². This means that the 1,000-plus known species of symbionts probably contain more than 100 times as many genes as exist in the host²³. Together these interacting genomes can be considered to operate as a super-organism, with extensive coordination of metabolic and physiological responses, particularly at the gut-liver and the gut-immune system levels. Not all of these interactions are necessarily obligate but the degree of true physiological association is difficult to study as all mammals possess a microbiome. Indeed, because of the level of comanagement of many biological processes by alien symbiotic genomes, Xu *et al.*²⁴ recently stated that "sequencing the components of the microbiome can be viewed as a logical albeit ambitious expansion of the human genome project".

We have recently discussed the microbiome-host relationship with respect to drug metabolism and toxicity and proposed a new type of probabilistic model based on conditional mammalian genome-microfloral interactions that may account for some aspects of individual variation in drug responses as well as the idiosyncratic toxicity of certain compounds⁴. The gut microbiome undoubtedly influences cytochrome P450 levels in the host, has intrinsic drug metabolizing capabilities²⁵⁻²⁸ and can influence the immune status as well as

such factors as peroxisome proliferator activated receptor γ nuclear cytoplasmic shuttling in the host, once thought to be purely under mammalian genome control²⁹. The fact that modulation of the gut microbial populations results in significant changes at the macroscopic level of metabolism has been demonstrated by observing fluctuations of cometabolized substrates, such as hippuric acid and hydroxyphenylpropionic acids and their conjugates present in urine at millimolar concentrations^{30–32}.

Parasites increase biocomplexity. The biocomplexity of mammalian systems is further increased by multidirectional interactions between the mammalian host, parasites and microbes. Each of these organisms metabolize and modify substrates interactively. For example cytochrome P450 activities have been demonstrated in some parasites³³, but additionally, several species of parasite increase cytochrome P450 activity and alter the activity of phase I drug-metabolizing enzymes in the liver of the host, which may have an impact on the liver's capacity to activate or detoxify both endogenous and xenobiotic compounds^{34–36}.

Recently, we have proposed a three-way interaction in a laboratory mouse model for *S. mansoni* whereby the introduction of the parasite into the host caused an indirect urinary depletion of gut microbial products such as *n*-butyrate, propionate and hippurate followed by increased excretion of 4-cresol together with its ether glucuronide and sulfate³⁷. Cresol metabolites are known to be antimicrobial products of *Clostridium difficile*, a mildly deleterious obligate anaerobe, which possesses 4-hydroxyphenylacetate decarboxylase activity³⁸. Thus, the parasitic infection resulted in a perturbation of the intestinal microbiome, allowing colonization of *C. difficile* and resulting in a change in the metabolite signature of the host. This interaction is mediated by an unknown mechanism but probably involves selective bacteriocidal secretions of the parasite implying another level of biological selection and interaction exerted on the gut microbiome. The intestinal environment must therefore be visualized as an entire ecosystem where chemical interactions occur at multiple organizational levels with cross-talk between the mammalian, parasitic and microbial systems (Fig. 5).

Implications for therapy and drug design. There is arguably a great deal of information exchange between symbionts, parasites and their hosts through cometabolism of substrates, some of which are pharmacologically active⁴. Because of the conditional nature of these interactions, the exact nature of the mammalian host-microbe interaction could influence certain aspects of drug metabolism and drug toxicity. The possible interactions that could affect drug toxicity and efficacy are summarized in Figure 6. From this, we can conclude that pharmacogenomic approaches² relying only on mammalian genetic information (polymorphisms or transcripts) are unlikely to provide a general solution to the problem of predicting drug activity in individuals (which is needed for personalized healthcare) as these measure only a

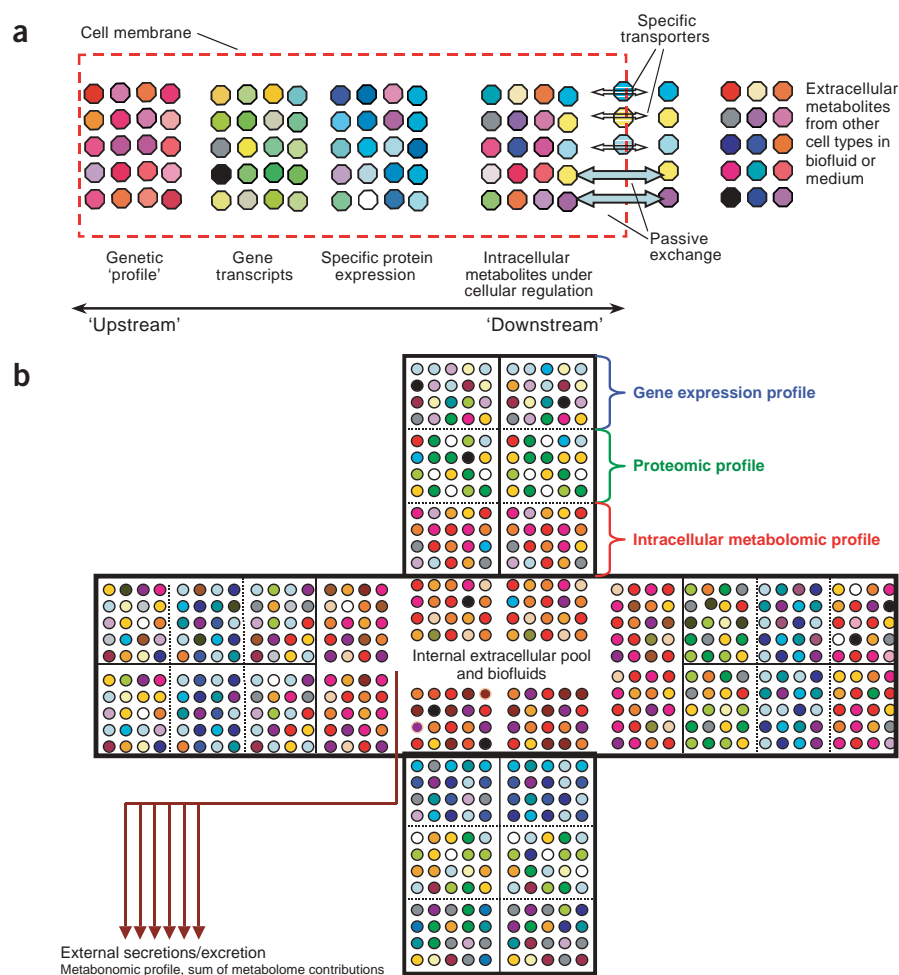


Figure 4 The global systems approach. (a) Representation of the omics-measurable variables from gene to metabolites. (b) Representation of multiple (in this case, 8) cell lines contributing to a common extracellular pool, such as blood plasma, which may be sampled to detect the effects of interventions, such as drug treatment, in the whole system.

small part of the relevant biological interactions. Such a conclusion may profoundly change the way that basic drug research is conducted in the future.

Integrated multiple 'omics' approaches

Most studies in the world of 'omics' tend to be in a single discipline (e.g., genomics/transcriptomics, proteomics). This does not necessarily imply a logical approach to the problem under investigation, but often merely reflects the particular expertise and specialization of the groups involved. In any integrated systems approach, it would seem beneficial to combine studies in all of the 'omics' to provide an overview of what is going on. Clearly, the downside of such a strategy is the requirement for a wide range of technical expertise, and experimental approaches to combining data from the different 'omics' platforms into a single coherent stream.

An indication of what is now possible using integrated multiple 'omics' approaches relevant to the pharmaceutical industry is given in two toxicological examples. Genomics and proteomics were combined, together with conventional techniques such as histopathology, to study the effects of paracetamol (acetaminophen), at a range of doses from 0 to 500 mg/kg, on liver in the mouse³⁹. Liver was sampled

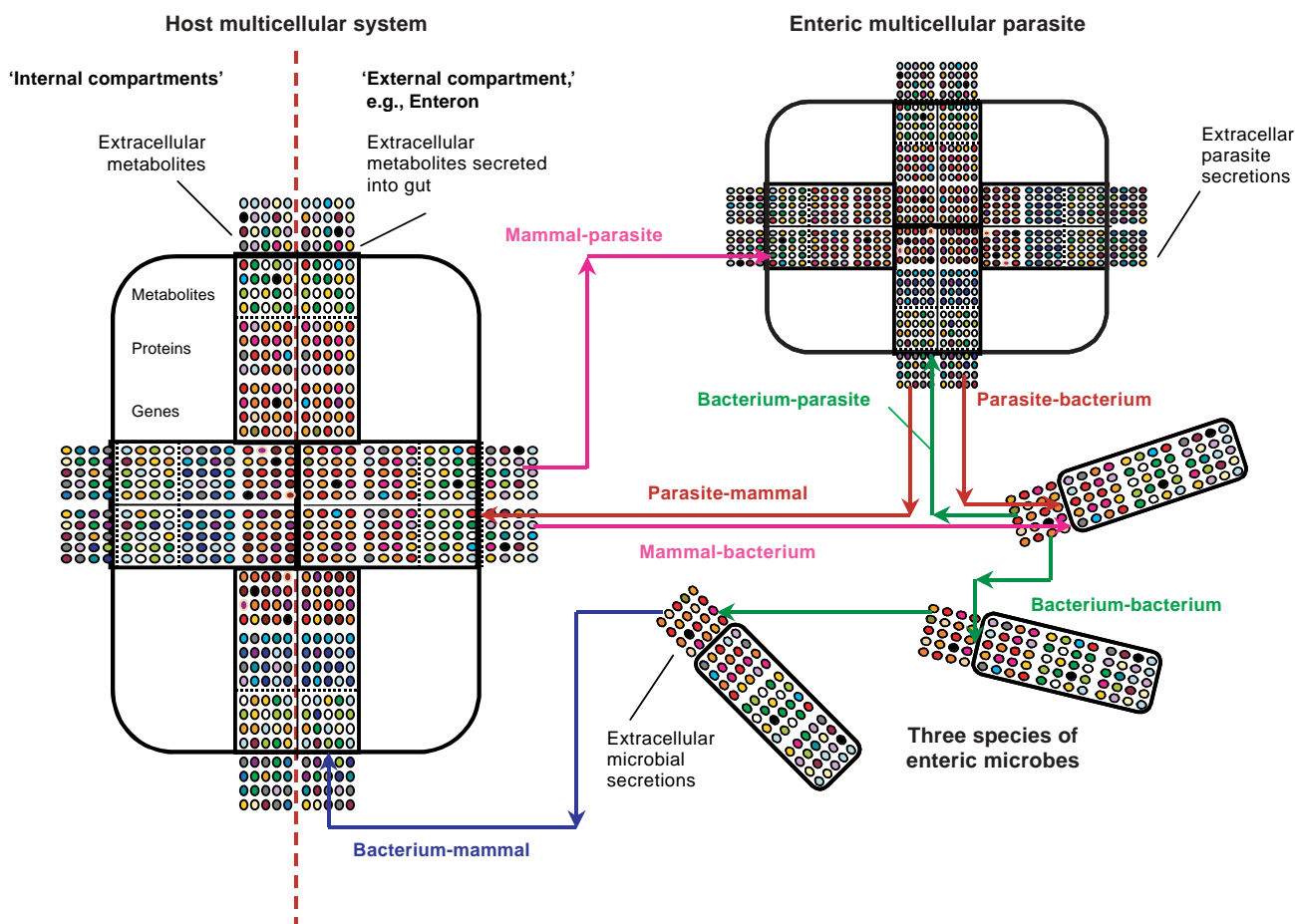


Figure 5 Depiction of multiple genome interactions between mammalian host, macroparasites and gut microbiome in terms of exchange and cometabolism of substrates. Arrows indicate the direction of metabolite transfer.

from 15 min to 4 h after dosing, with changes detected as early as 15 min after administration in mitochondrial proteins (apparently preceding most of the gene transcript changes). The bulk of the effects noted by both transcript profiling and proteomics were associated with loss of energy production (a decrease in ATP synthase subunits and β -oxidation pathway proteins). As the authors noted “transcript profiling and proteomics did not usually detect expression level changes of one mRNA and the corresponding protein, but genes, proteins and pathways identified by transcript profiling and proteomics told a similar story.”

A subsequent experiment from the same laboratories, using essentially the same study design, examined the relationship between the transcript profiles and metabolic profiles⁴⁰. The results of the metabonomic investigation also showed changes consistent with an alteration in energy metabolism (dramatic reductions in hepatic glycogen and glucose and concomitant increases in saturated lipids and reductions in unsaturated lipids and phospholipids). Here, the gene changes both preceded and were concurrent with the observed perturbations in metabolism. Taken together, the genomic-proteomic and genomic-metabonomic studies reveal a consistent qualitative picture of cells attempting to respond to the consequences of a global energy failure.

Orotic acid hepatotoxicity in the rat has also been studied in an integrated reverse functional genomic and metabolic study using transcriptional and metabonomic analysis combined with multivariate analysis and statistical bootstrapping⁸. Perturbed metabolic pathways

included those involved with fatty acid, triglyceride and phospholipid biosynthesis, β -oxidation, carbohydrate metabolism and altered nucleotide and methyl donor metabolism and stress responses, whereas transcriptome analysis showed effects on stearyl-CoA desaturase and other lipid-related transcripts. The relative success of these initial studies suggests that there may indeed be useful synergy to be gained from performing integrated ‘omics’ studies. However, these approaches still do not address fully the problems of modeling the complexity of the mammalian superorganism.

Current models of metabolic interactions

For simple systems, the way in which intracellular biochemistry can be modeled is now becoming mature, especially for single-celled organisms such as *Saccharomyces cerevisiae*, and a clear approach that combines conventional metabolic pathway analysis at the metabolite level with pathway prediction from the genome confirmed by proteomic studies⁴¹ is being followed. Even here though, and with complete knowledge of the genome of a simple organism, such as a pathogen, interpretation of genomic data has not led as quickly as expected to new drug targets, but rather it has been proposed that comprehensive modeling of metabolic networks will be more likely to lead to new avenues of opportunity for pharmaceutical development. This is particularly true for simple organisms where the known metabolic pathways are considerably simpler than those for typical mammalian cells with all their secondary interactions. In this case, genes can be mapped

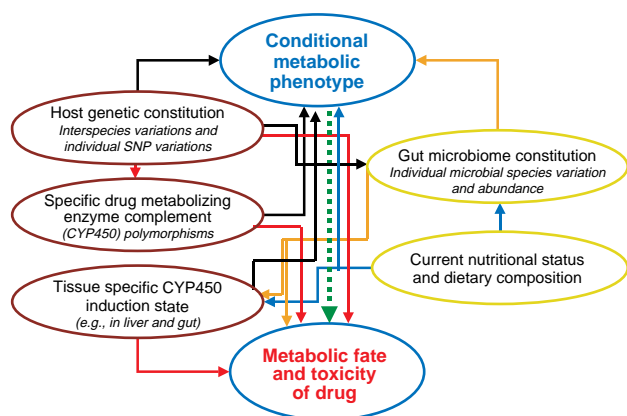


Figure 6 Mammalian–microbe interactions that can impact on drug metabolism and toxicity.

to specific proteins/enzymes and hence to metabolic pathways at several levels, that is, for the whole genome for a given organism, or for a specific part of the genome to compare across organisms (e.g., *Trypanosoma brucei* and *Trypanosoma cruzi*, which cause sleeping sickness and Chagas disease, respectively)⁴².

Operating at the metabolic level is particularly advantageous when making comparisons across species. However, many metabolites have not been identified (e.g., bile acid metabolites involving mammalian-microbial cometabolism⁴), and thus many gaps remain in conventional pathways models, such as KEGG⁴³ (<http://www.genome.ad.jp/kegg>) or ExPASy²¹ (<http://www.expasy.ch/cgi-bin/search-biochem-index>). Also, it has to be remembered that any mathematical model will always be simpler than the real system that it attempts to emulate. Thus, different mathematical models could, in principle, provide equally good predictions. Nevertheless, integrated modeling of biological systems at the transcriptomic, proteomic and metabolic levels is feasible using bioinformatics techniques, including rule-based methods such as are now used for predicting pathways for xenobiotic metabolism⁴⁴.

However, even at this low level of complexity, many of the feedback and regulatory mechanisms evident in metabolic pathways cannot be addressed. Such metabolic models can be broken down into studies of stoichiometry where metabolic fluxes are calculated to give a view of the system under a particular set of conditions using mathematical optimization techniques⁴⁵ often incorporating stable isotope labeling experiments⁴⁶. This approach has been used to describe all the known reactions in *Escherichia coli*⁴⁷. A second approach attempts to model the reaction kinetics of a system and to use the derived kinetic parameters to predict time profiles⁴⁸.

Conclusions and future trends

While significant progress has been made in modeling simple systems, a major problem arises at the next level of complexity—multicellular organisms and superorganisms—because of the need to consider intercellular transport and other kinetic parameters. Full mathematical treatment of the physicochemical phenomena is necessary. Given the differences in gene and protein expression in different cells, the only coherent and integrated approach is to model the biochemistry at the metabolic level.

Giersch⁴⁹ has reviewed the techniques used for metabolic modeling, including metabolic control analysis in cells, the effects of oscillations and chaos, the use of flux analysis and has provided pointers to soft-

ware sources. It is possible to conceive of each cell as a node with a set of metabolic pathways within the node as above, but with each node connected to other nodes and then the problem reduces to the modeling of the internodal connections. This could be achieved using conventional mass transfer methods used in process engineering or by taking advantage of the methods used to monitor transport of drugs across cell membranes or for more generalized pharmacokinetics. The topology and regulatory feedback aspects of such internodal modeling would be complex and would be organism dependent⁵⁰. It is here that the use of nonparametric multivariate statistics will provide a better understanding of the complexities of the situation in that connections and correlations between metabolic pathways from different cell types and different organs separated in space and with effects displaced in time according to locations and event are capable of being unraveled. One unexpected outcome of such an approach would be the discovery of the natural substrates for a number of promiscuous enzymes involved in xenobiotic metabolism, such as cytochrome P450 families⁵¹. Additionally, the multivariate statistical approach allows the analysis of combined gene array and metabolic data so that events identified at the transcriptome level can be related to those seen at the metabolic level in a reverse functional genomics approach as described above.

At the level of interactions between genomes, such as between host and pathogen or symbiotic bacterium, modeling approaches using probabilistic methods are expected to be the only way to model interactions between organisms. Bayesian approaches and the concept of maximum entropy, allow the construction of a metabolic model at the most sparse and simplest degree of complexity. At its simplest, Bayes's theory states:

$$\Pr(h|D) = \Pr(h) + \Pr(D|h)/\Pr(D)$$

that is, it gives the probability of a hypothesis h given a set of data D , and this requires an estimate of the prior probability of the hypothesis $\Pr(h)$, the likelihood of the data-given hypothesis $\Pr(D|h)$, and the probability of the data or the evidence, $\Pr(D)$. One recent example is the determination of the biochemical sources of glucose found in blood plasma using automated Bayesian analysis of the nuclear magnetic resonance-detected deuterium incorporation patterns after ingestion of D_2O ⁵². Thus, highly accurate and precise data, such as metabolite concentrations, are very valuable when attempting to model a system, given the recognized errors on quantification at the transcriptomic and proteomic levels.

It is possible to envisage an approach whereby both probabilistic and multivariate statistical models can be constructed at the microscale (i.e., using individual genomes) and united using probabilistic criteria to produce models at the macroscale (genome-genome). Thus, the probability of a given pathway or the level of a given substance will be assignable based on a combination of the data as measured and the information available. For instance, human cytochrome P450-mediated drug metabolism has been modeled probabilistically using a machine-learning approach⁵³ and a heuristic approach to prediction of more general xenobiotic metabolism has been described based on a comprehensive library of biotransformations and other reactions and on system-specific transformation probabilities⁵⁴. We conclude that only by probabilistic modeling approaches, and at the metabolic level using the host metabolome or sum of the interacting metabolomes as the blueprint, will overall system biology models be forthcoming because the genomes of many of the interacting organisms will not be known and indeed ultimately may be unknowable.

ACKNOWLEDGMENTS

We thank the Biotechnology and Biological Science Research Council, Engineering and Physical Sciences Research Council, The Wellcome Trust and the National Institutes of Health for funding this and related work. We also thank Paul Elliot and James Scott, Yueung Utzinger and Burt Singer, Jeremy Everett and Felicity Nicholson for their helpful comments and discussion on this work and related subjects.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at <http://www.nature.com/naturebiotechnology/>

- Hood, L. & Galas, D. The digital code of DNA. *Nature* **421**, 444–448 (2003).
- Smith, L.L. Key challenges for toxicologists in the 21st Century. *Trend. Pharm. Sci.* **22**, 281–285 (2001).
- Nicholson, J.K., Connelly, J., Lindon, J.C. & Holmes, E. Metabonomics a platform for studying drug toxicity and gene function. *Nat. Reviews Drug Disc.* **1**, 153–161 (2002).
- Nicholson, J.K. & Wilson, I.D. Understanding 'global' systems biology: metabonomics and the continuum of metabolism. *Nat. Reviews Drug Disc.* **2**, 668–676 (2003).
- Gygi, S.P., Rochon, Y., Franza, B.R. & Aebersold, R. Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.* **19**, 1720–1730 (1999).
- Schoenheimer, R. *The Dynamic State of Body Constituents* (Harvard University Press, Boston, 1942).
- Mayer, R.J. The meteoric rise of regulated intracellular proteolysis. *Nat. Reviews Cell Biol.* **1**, 145–148 (2000).
- Griffin, J.L. *et al.* An integrated reverse functional genomic and metabolic approach to understanding orotic acid-induced fatty liver. *Physiol. Genomics* **17**, 140–149 (2004).
- Raamsdonk, L.M. *et al.* A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nat. Biotechnol.* **19**, 45–50 (2001).
- Bollard, M.E. *et al.* Investigations into biochemical changes due to diurnal variation and estrus cycle in female rats using high resolution ¹H NMR spectroscopy and pattern recognition. *Anal. Biochem.* **295**, 194–202 (2001).
- Holmes, E. *et al.* NMR spectroscopy and pattern recognition analysis of the biochemical processes associated with the progression and recovery from nephrotoxic lesions in the rat induced by mercury II chloride and 2-bromoethanamine. *Molecular Pharmacology* **42**, 922–930 (1992).
- Waters, N.J. *et al.* NMR and pattern recognition studies on the time-related metabolic effects of α -naphthylisothiocyanate on liver, urine, and plasma in the rat: an integrative metabonomic approach. *Chem. Res. Toxicol.* **14**, 1401–1412 (2001).
- Azmi, J. *et al.* Metabolic trajectory characterisation of xenobiotic-induced hepatotoxic lesions using statistical batch processing of NMR data. *Analyst* **127**, 271–276 (2002).
- Keun, H.C. *et al.* Geometric trajectory analysis of metabolic responses to toxicity can define treatment-specific profiles. *Chem. Res. Toxicol.* **17**, 579–587 (2004).
- Lindon, J.C., Nicholson, J.K., Holmes, E. & Everett, J.R. Metabonomics: Metabolic processes studied by NMR spectroscopy of biofluids. *Concepts Magn. Reson.* **12**, 289–320 (2000).
- Robertson, D.G. *et al.* Metabonomics: evaluation of nuclear magnetic resonance (NMR) and pattern recognition technology for rapid *in vivo* screening of liver and kidney toxicants. *Toxicol. Sci.* **57**, 326–337 (2000).
- Mortishire-Smith, R.J. *et al.* Use of metabonomics to identify impaired fatty acid metabolism as the mechanism of drug-induced toxicity. *Chem. Res. Toxicol.* **17**, 165–173 (2004).
- Guarner, F. & Malagelada, J.R. Gut flora in health and disease. *Lancet* **361**, 512–519 (2003).
- Tannock, G.W. *Normal Microflora* (Chapman and Hall, London, 1995).
- Xu, J. *et al.* A genomic view of the human-*Bacteroides thetaiotaomicron* symbiosis. *Science* **299**, 2074–2076 (2003).
- Gilmore, M.S. & Ferretti, J.J. The thin line between gut commensal and pathogen. *Science* **299**, 1999–2002 (2003).
- Berg, R.D. The indigenous gastrointestinal microflora. *Trends Microbiol.* **4**, 430–435 (1996).
- Relman, D.A. & Falkow, S. The meaning and impact of the human genome sequence for microbiology. *Trends Microbiol.* **9**, 206–208 (2001).
- Xu, J., Chaing, H.C., Bjursell, M.K. & Gordon, J.I. Message from a human gut symbiont: sensitivity is a prerequisite for sharing. *Trends Microbiol.* **12**, 21–28 (2004).
- Gingell, R., Bridges, J.W. & Williams, R.T. The role of the intestinal flora in the metabolism of prontosil and neoprontosil in the rat. *Xenobiotica* **1**, 143–156 (1971).
- Rawls, J.F., Samuel, B.S. & Gordon, J.I. Gnotobiotic zebrafish reveal evolutionarily conserved responses to the gut microflora. *Proc. Natl. Acad. Sci. USA* **101**, 4596–4601 (2004).
- Peppercorn, M.A. & Goldman, P. Caffeic acid metabolism by bacteria of the human gastrointestinal tract. *Proc. Natl. Acad. Sci. USA* **69**, 1413–1415 (1972).
- Ogawa, H. *et al.* Sodium butyrate inhibits human intestinal microvascular endothelial cells through COX-2 inhibition. *FEBS Lett.* **554**, 88–94 (2003).
- Kelly, D., Campbell, J.I., King, T.P., Grant, G., Jansson, E.A., Coutts, A.G.P., Pettersson, S. & Conway, S. Commensal anaerobic gut bacteria attenuate inflammation by regulating nuclear-cytoplasmic shuttling of PPAR γ and RelA. *Nat. Immunol.* **5**, 104–112 (2003).
- Phipps, A.N., Stewart, J., Wright, B. & Wilson, I.D. Effect of diet on the urinary excretion of hippuric acid and other dietary derived aromatics in the rat. A complex interaction between the diet, intestinal microflora and substrate specificity. *Xenobiotica* **28**, 527–537 (1998).
- Williams, R.E., Eyton-Jones, H.W., Farnworth, M.J., Gallagher, R. & Provan, W.M. Effect of intestinal microflora on the urinary metabolite profile of rats: a ¹H nuclear magnetic resonance spectroscopy study. *Xenobiotica* **32**, 783–794 (2002).
- Nicholls, A., Mortishire-Smith, R. & Nicholson, J.K. NMR spectroscopic-based metabonomic studies of urinary metabolite variation in acclimatizing germ free rats. *Chem. Res. Toxicol.* **16**, 1395–1404 (2003).
- Saeed, H.M., Mostafa, M.H., O'Connor, P.J., Rafferty, J.A. & Doenhoff, M.J. Evidence for the presence of active cytochrome P450 systems in *Schistosoma mansoni* and *Schistosoma haematobium* adult worms. *FEBS Lett.* **519**, 205–209 (2002).
- Sheweita, S.A., Mangoura, S.A. & El Shemi, A.G. Different levels of *Schistosoma mansoni* infection induce changes in drug-metabolizing enzymes. *J. Helminthol.* **72**, 71–77 (1998).
- Sheweita, S.A. *et al.* Changes in the expression of cytochrome P450 isoenzymes and related carcinogen metabolizing enzyme activities in *Schistosoma mansoni*-infected mice. *J. Helminthol.* **76**, 71–78 (2002).
- Satarug, S. & Haswell-Elkins, M.R. Induction of cytochrome P450 2A6 expression in humans by the carcinogenic parasite infection, opisthorchiasis viverrini. *Cancer Epidemiol Biomarkers Prev.* **5**, 795–800 (1996).
- Wang, Y. *et al.* Metabonomic investigations in mice infected with *Schistosoma mansoni*: an approach for biomarker identification. *Proc. Natl. Acad. Sci. USA* **101**, 12676–12681 (2004).
- Selmer, T. & Andrei, P.I. p-hydroxyphenylacetate decarboxylase from *Clostridium difficile*: a novel glycol radical enzyme catalysing the formation of p-cresol. *Eur. J. Biochem.* **268**, 1363–1372 (2001).
- Ruepp, R., Tonge, R.P., Shaw, J., Wallis, N. & Pognan, F. Genomics and proteomics analysis of acetaminophen toxicity in mouse liver. *Toxicol. Sci.* **65**, 135–150 (2002).
- Coen, M. *et al.* Integrated application of transcriptomics and metabonomics yields new insight into the toxicity due to paracetamol in the mouse. *J. Pharm. Biomed. Anal.* **35**, 93–105 (2004).
- Ideker, T. *et al.* Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292**, 929–934 (2001).
- Fairlamb, A.H. Metabolic pathway analysis in trypanosomes and malaria parasites. *Phil. Trans. R. Soc. Lond. B* **357**, 101–107 (2003).
- KEGG: Kyoto Encyclopedia of Genes and Genomes. Release 26. April 2003.
- Greene, N. *et al.* Knowledge-based expert systems for toxicity and metabolism prediction: DEREK SAR and METEOR SAR. *QSAR Environ. Res.* **10**, 299–314 (1999).
- Varma, A. & Pálsson, B.O. Metabolic flux balancing: basic concepts, scientific and practical use. *BioTechnology* **12**, 994–998 (1994).
- Szyperski, T. ¹³C NMR, MS and metabolic flux balancing in biotechnology research. *Q. Rev. Biophys.* **31**, 41–106 (1998).
- Schilling, C.H., Edwards, J.S. & Pálsson, B.O. Towards metabolic phenomics: analysis of genomic data using flux balances. *Biotechnol. Prog.* **15**, 288–295 (1999).
- Vaseghi, S., Baumeister, A., Rizzi, M. & Reuss, M. *In vivo* dynamics of the pentose phosphate pathway in *Saccharomyces cerevisiae*. *Metab. Eng.* **1**, 128–140 (1999).
- Giersch, C. Mathematical modelling of metabolism. *Curr. Opin. Plant Biol.* **3**, 249–253 (2000).
- Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N. & Barabási, A.-L. The large-scale organization of metabolic networks. *Nature* **407**, 651–654 (2000).
- Mohan, R. & Heyman, R.A. Orphan nuclear receptor modulators. *Curr. Top. Med. Chem.* **3**, 1637–1647 (2003).
- Merritt, M., Bretthorst, G.L., Burgess, S.C., Sherry, A.D. & Malloy, C.R. Sources of plasma glucose by automated Bayesian analysis of H-2 NMR spectra. *Magn. Reson. Med.* **50**, 659–663 (2003).
- Korolev, D. *et al.* Modelling of human cytochrome P450-mediated drug metabolism using unsupervised machine learning approach. *J. Med. Chem.* **46**, 3631–3643 (2003).
- Mekenyan, O.G., Dimitrov, S.D., Pavlov, T.S. & Veith, G.D. A systematic approach to simulating metabolism in computational toxicology. I. The TIMES heuristic modelling framework. *Curr. Pharm. Des.* **10**, 1273–1293 (2004).