precise phenotypic anchoring for drug target identification, validation and biomarker discovery using an advanced SYSTEMS BIOLOGY approach

The pharmaceutical industry has not seen the hoped-for productivity gains from the various omics datastreams over the last decade. This article discusses how systems biology can exploit the natural interlinkages between these datastreams and put in place a powerful system for modern therapeutic development.

By Dr Peter C. Johnson and Dr Alan J. Higgins

The pharmaceutical industry is facing a major productivity challenge that has been emphasised by the recent NIH Roadmap[1] and FDA Critical Path initiatives[2]. Research and development costs for new drugs are skyrocketing and yet the number of new drug approvals and drug pipeline numbers are declining[3,4]. A decade of mergers has not had the desired effect. The sequencing of the human genome, while a remarkable and necessary achievement, has also not resolved the problem of productivity. Nor have the isolated additions of data pools emanating from proteomic, metabolomic, cellomic or tissue analysis technologies. In fact, some complain that the sheer volume of data requiring interpretation is outstripping human capability and enhancing confusion. It has become clear that past practices of drug discovery and development, while effective in the approach to some diseases, need to be revolutionised for the complex diseases that demand treatment. Despite these problems, the path forward appears to be brightening.

The emergence of Systems Biology

The past 10 years have seen the development or refinement of multiple unique tools whose basic function is to convert analog biological data to digital data. These include:

- Rapid DNA Sequencers
- Gene Expression Microarrays
- Multiprotein Detection Systems (Protein Chips, Mass Spectrometry)
- Multimetabolite Detection Systems (NMR, Mass Spectrometry)
- Cell Function Imaging Systems Using Fluorescent Multiprobe Analysis
- Tissue Analysis Using Machine Vision
- Advanced Medical Record Systems and Research LIMS

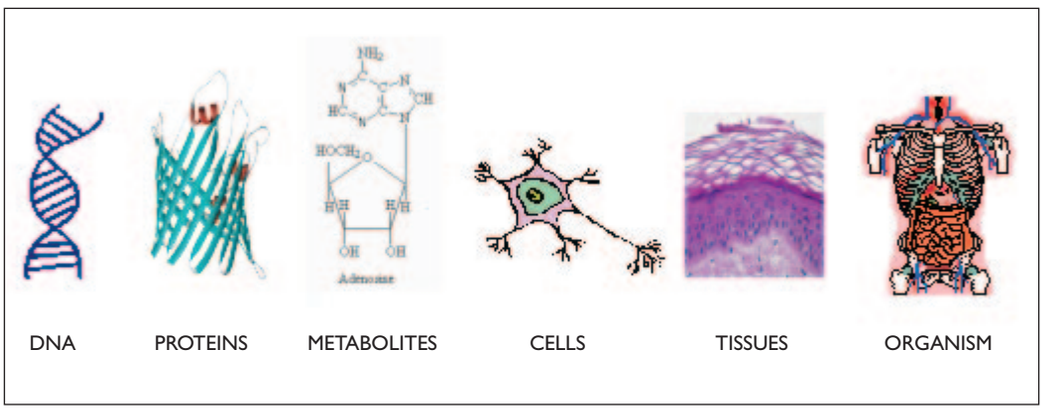

Figure 1
The 'omics' data streams: An organism is characterised by events occurring at multiple scales, from gene to cell to system. Systems biology recognises the need to measure these interrelated streams to effect precise understanding of cause and effect in the organism

● Informatics Systems Enabling Analysis of Coherent Data

At this time, these are not yet being applied optimally in unison. A natural tendency in industry and even in academia has been to organise working units around single 'omics' platforms, each of which depends upon one or more of the listed technologies for its data stream. Seldom have informatics systems been developed that encompass the whole to enable ready analysis of the interdependence of biological activities across the scale from molecule to organism. That time has arrived.

In the past, tools were simply not available to measure relevant biological activities. Today, when almost all biological activity can be measured in a high throughput fashion, we are limited primarily by the inability to leverage the power of the data. The innate interdependence of measurable events in a biological system, if readily analysed, would enable us to more deeply understand specific mechanisms of biology and also the consequence of exposure to drugs in the context of that system.

As such, we are fundamentally facing an informatics problem having two components. The first is the need to create a common denominator for the data being generated by different measurement tools that access biological data on different scales. This is known as the creation of 'coherent' data. The second problem is the need to provide an information management system that delivers the following:

● Storage and access to coherent, multiscale biological data.
● Data visualisation (sequence visualisation, metabolic pathway maps, etc).
● Access to available online sources of biological information (BLAST, etc).
● Statistical analysis software.
● Intuitive user interface.

Fortunately, such functional information systems are now in development. Once applied, their benefits will include the following:

● Rapid data storage, access and sharing.
● Continuous accumulation of information value through the provision of historical context.
● Correlation of biological data across all levels of organism scale.

The latter benefit is the most compelling because it enables an investigator to engage in an analysis of cause and effect from molecule to organism, an approach known as Systems Biology[5]. Systems Biology can be defined as the science that combines multiple biological data streams (from gene to organism) to enable a profound understanding of interdependent biological processes. Not just another 'omic' data set, Systems Biology is the final integrator of all omics datasets into a cohesive network that can be probed deeply to understand the mechanisms of disease and/or drug effects (**Figure 1**).

Consider the impact that Systems Biology brings to biological understanding. Previous functional genomics approaches have attempted to compare gene expression profiles between normal and diseased tissues in a search for drug targets. However, many functional genomics approaches have lacked genuine function; they merely establish some degree of probability that certain genes may be linked to a disease without defining causal mechanisms. Another issue is that many genes that are altered by an intervention (eg drugs or disease) cannot even be identified, so how does the researcher prioritise hits for target selection? The answer may be to filter the highly granular gene expression data using additional data streams.

Almost always, the utility of this process has been limited by the noise due to interindividual variability even among groups having the same tis-

sue diagnosis. Systems Biology turns such analyses in a new direction, taking advantage of the tight intraindividual linkage between biological cause and effect to see how gene expression patterns result in specific protein expression, pathway activation and resulting cellular and tissue changes that may be unique to an individual or shared within groups. The strength of the underlying information system enables each individual organism's responses to be placed within the context of any related group, allowing the investigator to determine those responses that lend themselves to personalised therapies versus therapies that can be generalised to populations.

Systems Biology has been likened to a Global Positioning System (GPS) to the extent that the interlinkage of multiple streams of related data sharpens the ability to localise clinically meaningful cause and effect beyond that which analysis of any one stream can provide. The larger the number of related data sets one can apply, the stronger the analytic result. However, it has been shown that the use of only two such sets (gene expression and automated tissue feature analysis) already enables the identification of genes related to toxicity-associated tissue change that lie well within the two-fold noise cut-off normally required in expression analysis (**Figure 2**)[6,7]. This approach increases the probability that the leads so identified are linked to the desired endpoint. This has powerful implications for drug discovery in pharmaceutical and biotechnology companies because in addition to the identified targets, the Systems Biology approach enables concurrent validation through analysis of cause and effect throughout the organism.

If we use pathway mapping as our frame of reference, our approach becomes more mechanistic so instead of deriving a series of individual gene or protein targets, we identify mechanisms and pathways of disease that broaden our choice of druggable targets. For instance, genes and proteins that are significantly altered may be parts of critical disease mechanisms but may not be the best choice of targets within those pathways (eg not control points or rate-limiting steps, or lack specificity).

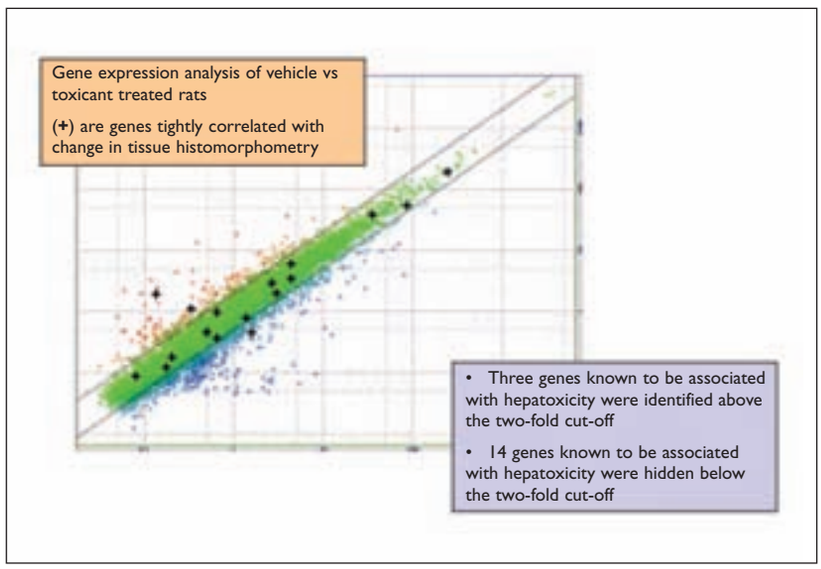

Figure 2
The power of correlation between system data sets: Rats treated with Carbon Tetrachloride were studied using both gene expression analysis and hyperquantitative tissue analysis. Correlative change at the gene and tissue feature levels allowed genetic influences upon tissues to be detected at a greater than usual level of precision, as illustrated by the + signs within the two-fold noise variation that is usually accepted (14 otherwise inaccessible target candidates found)

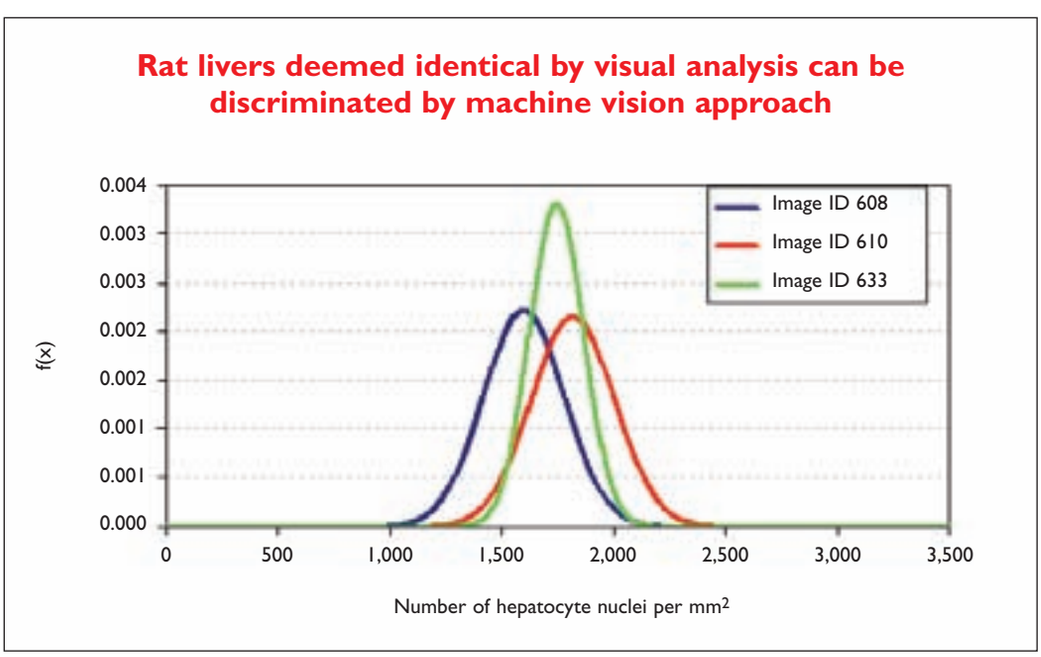

Figure 3
Machine vision detection of tissue change. Three sets of livers, each considered by pathologists to be "identical and normal" were assayed using HTA. Each liver had a distinct gene expression profile using microarray analysis. HTA was able to show that the tissues were distinct and potentially correlable with their expressed genetic changes

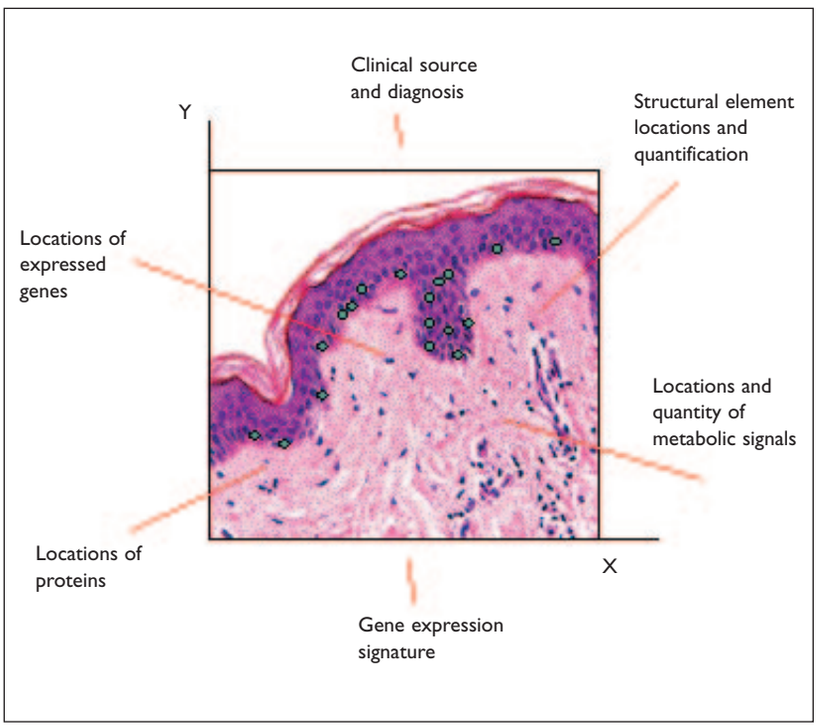

Figure 4
Tissue information. HTA is designed to quantitate anything that can be made visible within tissue. This figure illustrates the Cartesian relationship of the components of tissue, each of which is controlled by gene expression at some stage in development

Advanced Systems Biology and the problem of diagnosis

Systems Biology creates a cause and effect map across levels of biological information, from the molecular to the organism. In this way, the story of an organism's life unfolds in a linked fashion, enabling paths of information to lead back to the cause of disease (eg target(s)). Not all Systems Biology approaches are equal, however. For example, a Systems Biology approach that incorporates the analysis of gene expression, proteomic, metabolomic and cellular data streams alone, while powerful, will always be limited in its precision by the diagnosis assigned to the tissue of origin. The assignment of a pathological diagnosis (Gr. dia + gnosis, literally to 'know through') at the tissue level has been a time-honoured way of classifying an organism's phenotypic response to normal development, toxicity or disease. In classical functional genomics studies, differential gene expression is assessed through comparison of two sets of samples, one typically diagnosed as normal and the other having the diagnosis of a disease of interest. This approach unfortunately incorporates an extraordinary amount of noise, contributed in part by the following:

- Tissue diagnosis is a 'binning' mechanism that reflects generally agreed upon distributions of tissue features within often large acceptable bounds of variability.
- Pathologists disagree in the assignment of diagnoses for the same tissue a significant percentage of the time[8].

Disease generally manifests itself along a continuum, whereas a diagnosis is most often a static designator. It has been shown that gene expression patterns among samples grouped by diagnosis can vary widely, even when pathologists consider the samples to be identical in appearance (**Figure 3**). An Advanced Systems Biology approach recognises this problem and rectifies it through the incorporation of one additional 'omic' dataset: Hyperquantitative Tissue Analysis (HTA; **Figure 4**)[7]. HTA is a recently developed methodology that links robotic microscopy (to automatically capture the full image in a histological section as a digital image file) with tissue-specific machine vision image analysis software. Machine vision is an image processing methodology that segments an image into its components on the basis of their colour, shape, texture and contextual association. When machine vision software is created to analyse tissue, pathologists comprise a vital part of the design team, since they identify the appropriate cell types, matrices and structures to be analysed. They are also responsible for validating the output of the software during its development. The software is consequently able to detect the relative locations of all of the tissue's components at once, enabling precise quantitation of tissue components as well as the metric interrelationships of any number of these components. As shown in **Table 1**, this reduces information derived from the tissue image to a form consistent with the discrete digital data obtained at other levels of the biological continuum.

HTA creates two streams of value. First, the capacity to precisely quantitate components within a tissue simplifies the detection of differences between tissue samples. By measuring the proper set of tissue components, one can automate the differentiation of normal from abnormal tissues in preliminary drug safety screening, for example. It has also been shown that tissues identified as normal and identical by pathologists' visual examination can be shown to be significantly different when measured in this more precise fashion (**Figure 3**). Second, since tissue components and their interrelationships do not arise randomly, they are likely to be under substantial genetic control. The reduction of the histological image to discrete digital component data in an Advanced Systems Biology context renders possible the identification of covarying events (such as gene expression patterns associated with tissue change) at antecedent stages in the biological continuum, thereby using the granularity

afforded by HTA to exploit observed changes in tissue features (histopathology) in a mechanistic context. This has been shown to be the case in CCl4 hepatic toxicity studies[6] (**Figure 2**) and in hepatic changes associated with Type II diabetes.

The introduction of HTA into an Advanced Systems Biology context creates a new opportunity to group samples not by diagnosis alone but by actual tissue component patterns known to be associated with a disease. Through correlation of these patterns with actual graded clinical responses, a new and more accurate form of phenotype assessment is emerging. This concept, known as 'Phenotype Anchoring' will lead to the identification of subgroups that are more likely to respond to specific therapeutics[6].

Another opportunity is fostered in the following way. As Phenotype Anchoring becomes more precise, analysis of uniquely associated gene expression, proteomic and metabolite changes immediately provides a suite of biomarkers that can be used in the identification of candidate subject groups for therapeutics and can reduce the inter-subject noise in clinical trials. Indeed, biomarkers and mechanisms derived from invasive sampling can also be correlated with plasma, serum or urine metabolite changes to identify 'bridging biomarkers' of both efficacy and toxicity (on-target and off-target effects) that can further ease the management of clinical trials.

This approach conforms well to the recently published FDA Critical Path Initiative that requests enhanced tools and biomarkers for the improvement of clinical trial design and throughput[2].

Practical applications of Advanced Systems Biology

Assume that you are engaged in the identification of novel drug targets and therapeutics for a disease such as alcohol-induced cirrhosis of the liver[9]. You have identified an appropriate animal model and have clinical partners who can ethically access

BIOLOGICAL DATA SUBSET	OMIC DESIGNATION	APPLICABLE TECHNOLOGY	MINIMUM DATUM	DIGITAL DISCRETION
DNA	Genomics	Nucleotide Sequencing	Nucleotide	High
RNA	Functional Genomics	Gene Expression Profiling	Expressed Sequence	High
Protein	Proteomics	2D Gels; Mass Spectrometry; Protein Chips	Amino Acid Sequences; Whole Proteins	High
Metabolites	Metabolomics/onomics	Mass Spectrometry NMR	Metabolite Molecule	High
Cells	Cellomics	Fluorescence Probe Digital Imaging; Immunohistochemistry; Cytology	Photons	High
Tissue*	Pathology	Microscope	Tissue image and components	Low
*Tissue***	*Hyperquantitative Tissue Analysis*	*Robotic Microscopy and Machine Vision Tissue Analysis Software*	*Pixel*	*High*
Whole Organism	Clinical Informatics	Relational Databasing; Support Technologies	Byte	High

Table 1: Omics data streams, detection technologies and level of discretion of the resulting data. Note that standard pathology analysis (with reporting of diagnosis as text) does not enable tissue information to be utilised at as high a level of discretion as is possible with other -omic datasets.

* Before the development of machine vision tissue analysis software and robotic microscopy
** After the development and application of these technologies

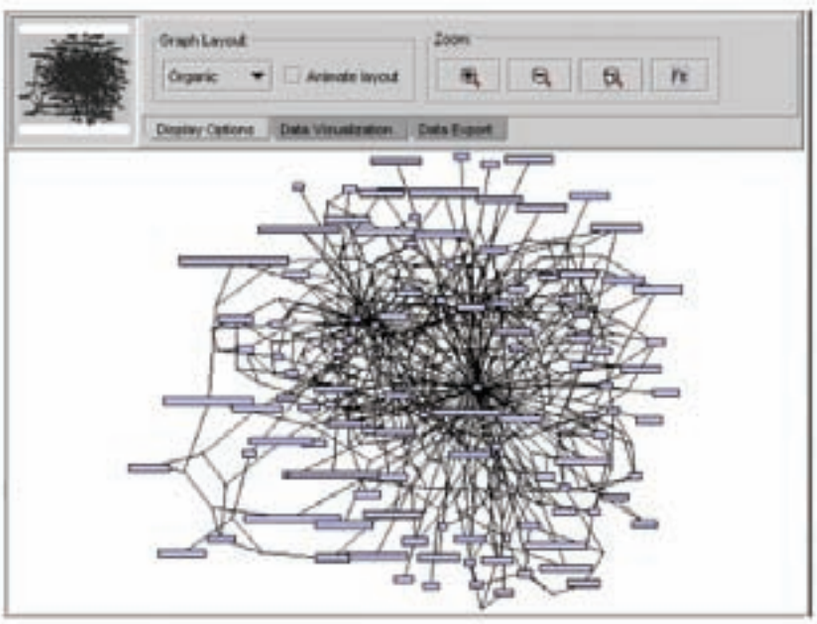

Figure 5
Pathway viewer: Metabolic pathways are carefully curated using the literature, then mapped within a system that enables identification of perturbed metabolic elements. Knowledge of the essential proteins within the perturbed area assists with target validation and hypothesis testing

diseased human liver samples. These samples are used to identify mechanisms and biomarkers that are perturbed in the target population and that can be correlated with the animal model to validate it or to guide development of an alternative. The animal trial is designed with multiple time endpoints so as to observe the time course of the disease once alcohol exposure is initiated. A normal control group is similarly studied. At appropriate time points, clinical measurements such as weight and blood pressure are measured, urine and serum are obtained and liver tissue is removed and prepared for gene expression, metabolomic and HTA studies. In addition to measurement of standard liver-associated serum tests, metabolite profiling is also carried out in urine and serum using mass spectrometry to identify and track peripheral biomarkers of the target mechanism. HTA is performed and up to 100 separate liver tissue components are concurrently measured. The n is five per group for each time point. All data are entered into a data coherence informatics platform that enables the interrelationships between all sets of data to be identified.

Each individual animal varies slightly in its responses at all levels from the other animals in each group, but taken as a whole, there is only modest variability at each time point. There is, however, sufficient variability to enable quantitative gene expression levels to be ranked by relative magnitude within the group and the same ranking step is carried out using the metabolomic and HTA data. Co-variance between the data sets is sought and genes whose quantitative expression changes in a similar fashion to metabolite patterns and tissue component patterns are singled out for further analysis.

Using a pathway-viewing tool (**Figure 5**) that enables metabolic perturbances (**Figure 6**) to be mapped against pathways driven by known proteins, the candidate genes can be assessed for their potential role in the metabolic processes observed. The pathway viewer is also capable of providing some indication of whether the suspect gene product is located at an appropriate signalling point to potentially serve as a 'druggable target'. Human tissue samples are then probed to determine if the protein of interest is present in human tissue. RNAi or knockout studies can then be done in a repeat of the animal model to determine the necessity of the protein in the disease process.

As all of this is being done, all of the information so derived is being captured in the data coherence system to enable sharing of the data, generation of data context and to enhance the opportunity for serendipitous discovery by additional observers. Cellular or protein assays are then created and chemical libraries are tested against the candidate target. Hits are refined to leads and leads are tested in animal models using standard approaches. HTA is applied to detect change occurring in tissue at levels normally not detected visually. Any toxicities induced by the lead(s) can be evaluated using information previously derived from the disease state as a background, in comparison to the full set of measurements being made in the drug safety study.

As the lead becomes a candidate for clinical trial, prior measurements of the disease state are reassessed to identify constant or near constant gene expression, metabolite and tissue component changes that reflect the ongoing development of the disease. This is known as a biomarker suite and, if validated, can serve as a surrogate marker in a clinical trial. If there are responder and non-responder animal populations, variance in these markers by group can provide valuable responder biomarkers to enable recruitment of optimal sets of subjects into the clinical trial. In addition, such biomarkers represent valuable intellectual property since they may be convertible into clinical diagnostics that determine who will be treated with the drug, once it is released for marketing.

As this is done, a correlative analysis of metabolite patterns in serum and urine is carried out with respect to the previously mentioned biomarkers. If there is tight correlation, then it may be possible to use blood or urine tests as opposed to tissue biopsy in order to guide the clinical trial or identify responder populations. These are known as

'Bridging Biomarkers', since they enable events occurring in an accessible fluid compartment (such as blood or urine) to accurately reflect what is occurring in relatively inaccessible compartments such as tissue. These bridging biomarkers can then also be converted into diagnostics to guide disease management.

Should problems develop in the clinical trial or after marketing has begun, the cache of interrelated data in the data coherence system can make investigation of toxic changes more transparent. The integrated system works to hasten and enhance decision-making, investigation and context-based discoveries in ways that have not been available to investigators before.

Summary

The past 10 years of isolated development of multiple omics datastreams have not produced the productivity gains hoped for by the pharmaceutical industry. However, the natural interlinkages between these datastreams are now being exploited. Systems Biology approaches allow previously inaccessible targets to be 'seen', target validation to be based logically on pathway perturbation analysis, the previous coarse classification by diagnosis to be refined and disease management biomarkers to be developed. New informatics platforms that enable disparate data to be analysed in a coherent fashion are changing the ways in which we will pursue drug discovery and development and investigative toxicology in the near future. The past 10 years of technology development, it is hoped, will eventually merely be seen as essential development time needed to put such a powerful system in place for modern therapeutic development. DDW

***Peter Johnson, MD**, serves as Executive Vice-President, Life Sciences, Chief Medical Officer and Head, Corporate Development for Paradigm Genetics. He was a Co-Founder of Tissue Informatics Inc and has served as Chairman and Chief Executive Officer since 1999. In 1994, he founded and became the first Executive Director and in 1996 the first President of the Pittsburgh Tissue Engineering Initiative.*

***Alan J. Higgins, PhD**, serves as Senior Director, Investigational Medicine, for Paradigm Genetics. Prior to Paradigm, he held various senior R&D management positions with both major Pharma (Pfizer and Hoffmann-LaRoche) and biotechnology companies (Nobex, Allelix and Oncogene Science).*

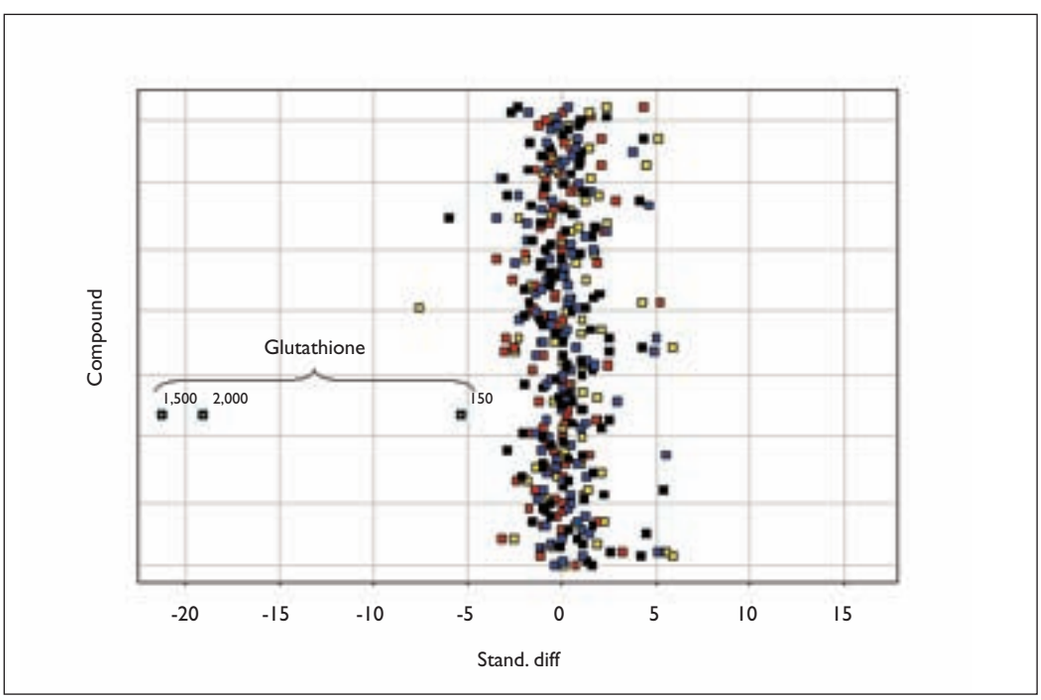

Figure 6
Metabolomic profiling: In a metabolomic analysis of up to thousands of peaks in a mass spectrum, output analysis enables detection of departures from normal, as in this instance of change in glutathione levels that are significantly greater than in the normal population. This data is fed into a pathway analysis system in order to identify toxic or disease-associated pathways and their related protein targets

References

1 NIH Roadmap: Accelerating Medical Discovery To Improve Health. http://nihroadmap.nih.gov/overview.asp

2 Innovation Stagnation: Challenge and Opportunity on the Critical Path to New Medical Products; US Food and Drug Administration, March 2004. www.fda.gov/oc/initiatives/criticalpath/whitepaper.html

3 DiMasi, JA, Hansen, RW, Grabowski, HG, The price of innovation: New estimates of drug development costs, J. Health Economics 2003 Mar: 22(2) 151-85.

4 Burrill, GS. Biotech 2003, Burrill and Co.

5 Weston, AD, Hood, L. Systems biology, proteomics and the future of healthcare: toward predictive, preventative, and personalised medicine. J. Proteome Res. 2004 Mar-Apr;3(2):179-96.

6 Kriete, A, Anderson, MK, Love, B, Freund, J, Caffrey, JJ, Young, MB, Sendera, TJ, Magnuson, SR, Braughler, JM. Combined histomorphometric and gene-expression profiling applied to toxicology, Genome Biology 2003 4(5):R32.

7 Johnson, PC, Braughler, JM. Automated pathology software in toxicology and drug safety. Current Drug Discovery, December 2003, 23 –28.

8 Terry, MB, Neugut, AI, Bostick, RM, Potter, JD, Haile, RW, Fenoglio-Preiser, CM. Reliability in the classification of advanced colorectal adenomas. Cancer Epidemiol Biomarkers Prev. 2002 Jul;11(7):660-3.

9 French, SW, Miyamoto, K, Tsukamoto, H. Ethanol-induced hepatic fibrosis in the rat: Role of the amount of dietary fat, Alcoholism: Clinical and Experimental Research 10(6):13S-19S, 1986.