The search for validated biomarkers in the face of biosystems complexity

"It is always a mistake to apply notions to things external to the human experience – for these are dangerous analogies" *Albert Einstein*

As the pharmaceutical industry is all too well aware, the genomics and post-genomic sciences have delivered an excessive number of drug targets – few of which have been well validated. Indeed, apart from a few rare exceptions, genomics and many millions of dollars in expenditure have yet to greatly impact drug development. This situation has been further exacerbated by the 'hype' and over-selling of the biotechnology industry, much to the chagrin of the world's stock markets and investors. Nonetheless, the genomic sciences have much to offer. For the first time in the history of the pharmaceutical industry, we possess a full parts-list for all the protein targets in the human body. This information is invaluable and should allow us to move forward in a better-informed manner, particularly with respect to target selectivity analysis and prediction. Here, we will examine the need to work better in the biomedical sciences to deliver biomarkers of high worth.

By Professor Ian Humphery-Smith and Will Dracup

The search for biomarkers is laudable in that they allow for:

- Discovery of new targets.
- Early diagnosis of life-threatening disease.
- Monitoring of adverse responses to one or more drugs.
- Monitoring new treatments in clinical and pre-clinical trials.
- Predicting disease and treatment outcomes.
- Improved economic and practical efficiencies in healthcare delivery.
- Better targeting of high cost treatments to those patients most likely to benefit.

For these reasons, much effort in genome mutation analysis, transcriptomics, proteomics, metabolomics and increasingly systems biology is being devoted to the search for biomarkers. However, there are a number of features inherent in biomedical data that make this task highly non-trivial. We must realise from the outset that proteomics or transcriptomics experiments performed in isolation will always offer up prominent features

Figure 1
The ever increasing queue of clinical and academic researchers waiting outside the door to the untold riches of the pharmaceutical industry armed with 'biomarkers' to their favourite disease! Pharma has unfortunately heard this story before and awaits the day when these biomarkers have first been well validated.
Image of nurses queuing courtesy of University of Connecticut Libraries Archives and Special Collections, showing nurses outside the Wilcox College of Nursing in 1918

indicative of ' biomarkers'. The non-intuitive message is that it is literally impossible not to find biomarker candidates when one is examining many hundreds of gene transcripts or proteins in a parallelised analytical environment, eg cDNA or oligo biochips, mass spectrometry and two-dimensional gels. Unfortunately, most practitioners working in the genome sciences have yet to comprehend this mathematical reality, or the need to work better in high dimensional space. To better justify this statement, we must first examine in more detail the nature of the high dimensional space in which we are operating. When confronted with the biological complexity associated with living organisms, and in fact with life itself, the parameter space is almost always effectively boundless. Data sets are never complete, while the quality and statistical confidence of data will always be far from homogeneous. Furthermore, individual data sets will consistently give results below statistical cut-off limits due to the practical inability to do the large numbers (100s) of replicates required before observed frequencies surpass those expected by chance alone.

The biomedical sciences must increasingly learn to be mathematically driven. That is to say that clinical practitioners and biologists must learn to ignore their well-informed 'gut feelings' and a priori intuition and turn increasingly to the symbolic logic of mathematics. As biologists, we have been slow to realise this central tenet so essential to other scientific disciplines and introduced by Newton and Leibniz some three centuries ago. We must realise that, when faced with highly-complex systems, our intuitive reasoning, no matter how well-informed, will not be capable of clarity. In the biological kingdom, environmental factors and genetic influences are highly multi-factorial and characterised by both positive and negative feedback loops and molecular processes working within highly-diverse spatial and temporal constraints. In the face of such complexity, investigators' thinking must be question-driven in a more holistic, systems-modelled environment. Of course, this approach must continue to be followed by the more traditional hypothesis-driven science, but only once the whole has first been subjected to mathematical scrutiny. This mathematical scrutiny must in turn be an iterative process of repeated cycles of prediction, wet lab validation and refinement. The genomic sciences have permitted the acquisition of data relating to entire organisms, but much progress is needed in the manner we analyse such systems.

In the pharmaceutical industry today, there is a strong tendency within IT departments towards the integration of highly-disparate data sets ranging from clinical observations through to SNP frequencies and multi-drug effects. Regrettably, this alone tends only to make the situation worse. Well-informed decision making during drug development is only rendered more difficult in that pharma executives must first read a larger encyclopaedia of information before proceeding to take a novel compound into formal drug development. What is needed are mathematical guidelines to assist our decision making, rather than therapeutic candidates championed as part of the routine competitive path to career advancement and a myriad of both rational and less rational influences. As was seen only too recently (the withdrawal of Vioxx™ by Merck & Co on September 30, 2004), unforeseen side-effects caused a multimillion dollar corporation to lose almost half its value in just a few days. Reducing the likelihood of these unforeseen consequences will defy intuitive reasoning and must increasingly be accompanied by mathematically-established guidelines based upon a more holistic appreciation of the system being analysed.

Such mathematical tools are not the realm of science fiction. Disparate data can be accommodated through Boolean, Bayesian, Heuristic, Differential and other means to provide an improved synthesis. The good news is that, given access to so much information, surprisingly today a superior understanding can be achieved. Pharma executives must start demanding such support and contextualisation of their industry-wide information gathering.

Returning to the quest for validated biomarkers, most experimental biologists have yet to realise just how daunting is the task at hand. Examples will be provided to further justify statements made above.

Binding site diversity in the human proteome

Unpublished estimations (Ian Humphery-Smith and David Gestel), believed to be accurate to within one order of magnitude, suggest that some 20 billion 5-mer drug binding-sites are likely to exist within the human body. This astounding figure is the reason all drugs possess Adverse Drug Effects (ADE). The challenge for the pharmaceutical industry is to minimise such effects and take to market those iterations, isoforms and/or formulations most likely to be linked with the least number of ADEs as indicated by off-target binding. These estimates of binding-site diversity were based upon 37 antigen/antibody crystal structures and were employed to establish a reliable estimate of a mean interaction patch radius as determined by Van de Waals and electrostatistic forces. A spherical minimalist model for surface-exposed residues was then corrected with real data derived from 680 SCOP domain structures. Each interaction patch corresponded to 2,408 potential and distinct 5-mer interactions, once disc overlap was subtracted. This mean interaction patch was then walked across the surface of all proteins encoded by every Open Reading Frame in the human genome. The estimated 20 billion binding-sites takes into account reliable estimates of the number of protein isoforms engendered by post-translational modifications and splice variants. A contingency table was then produced to calculate the binding potential associated with distinct protein-protein interactions in the human body, with the result being some 600 trillion. This figure was obtained by multiplying the number of potential surface-exposed binding sites by the number of Open Reading Frames in the Human genome, minus respectively 500,000 unique binding sites for a single mean molecular mass protein of 30k Da and one gene.

Given the potential for so many distinct bimolecular interactions within the human body, useful molecular work and ordered biological processes only become possible within living cells as a consequence of temporal and spatial separation. These processes are too frequently simplified in the minds of experimental biologists into linear biochemical pathways and the inaccurate notion that this protein does or does not bind with another in the world of intra-cellular molecular thermodynamics. All biomolecules interact with one another to a greater or lesser extent. The resultant interaction between any two proteins is never zero. Each interaction exists along a continuum up to 100% and can be expressed in terms of target recognition; degree of cross-reactivity and low to high affinity. In turn,

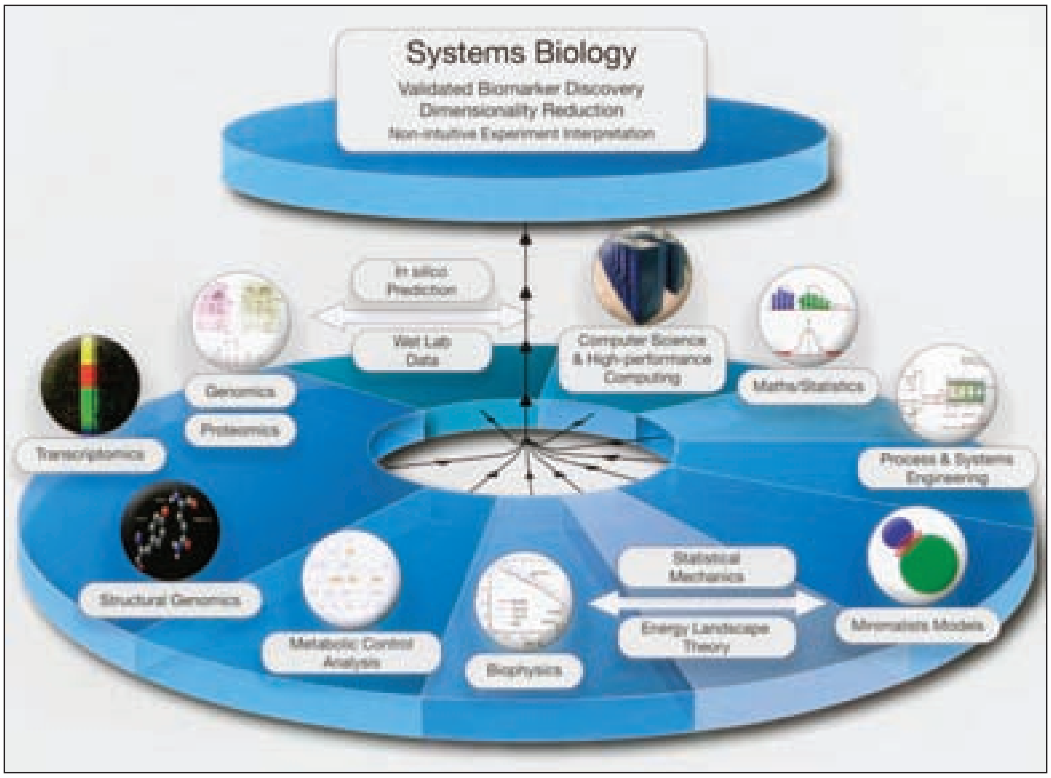

Figure 2
A schematic overview of data integration via a systems approach

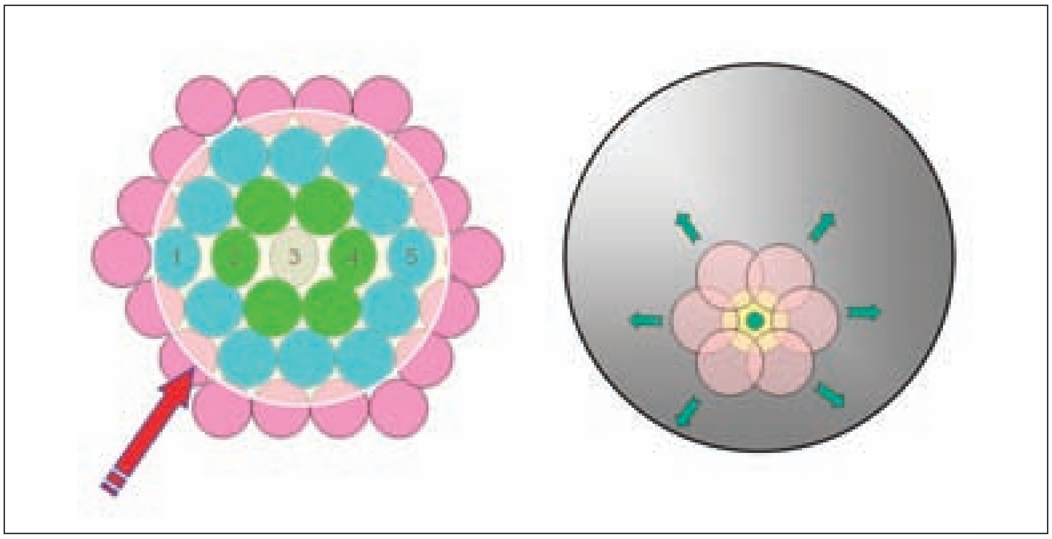

Figure 3
A mean protein-protein interaction patch (touching zone) as determined by 37 same-class x-ray co-crystallisation events was walked across the surface of all molecules in the human proteome. Molecular surface area estimations and exposure of amino acid residues were based upon a spherical minimalist model and corrections derived from exact measurements on 680 SCOP domain structures

many factors then combine to influence a physiological outcome. These include the concentration of target and ligand; time; pressure (eg partial pressure of oxygen); both physical (intracellular organelles and cell types within tissues) and physiological (eg inactive precursors) compartmentalisation and, crucially important, populations of molecules.

Here too, our current conceptual framework needs to be challenged. Much of our thinking in cellular biochemistry is heavily influenced by the manner in which we have been taught to think about biochemical pathways. The basis for these pathway reconstructions has been entirely dependent upon *in vitro* experimentation whereby reaction kinetics are measured in a test-tube in a one-on-one environment. It is noteworthy, however, that a much more complex scenario is the rule in biology. Our biochemical experimentation has rarely been conducted with a biomolecule of interest being exposed to all the protein isoforms encoded by the human genome. Under the latter conditions, specific binding to a single target molecule is totally impossible. Thus, our simplistic view of the biological universe must be further refined in the light of a detailed knowledge of the potential target space within the human body. For eons, life processes have had to deal with the constraints of doing molecular work in a complex environment. The latter is achieved through the collaboration of both high and low affinity binders; and appropriate degree of reaction reversibility for the biological function in question; the end-point of one or more pathways; and populations of molecules combining to surpass a physiological threshold, for example, to commit to cells either to apoptosis or cellular division.

It is impossible not to find biomarkers!

Statistically-validated molecular observations (biomarkers) may occur as individual events or combinations of events. These tell-tale events allow us to seek-out prominent processes within biology and understand them or diagnose them correctly. As mentioned earlier, it is somewhat non-intuitive to consider that biomarkers will almost always be present if sufficient experimental parameters are gathered. The dilemma is the fact that most of the biomarkers we detect during high throughput screening of nucleic acids, proteins and/or clinical observations will inevitably be false positives. The task at hand is very much to seek-out validated biomarkers within this sea of false positives.

To better understand the scale of the problem at hand the following example is cited. If one considers all the unique combinations of single events, pairs of observations, 3s, 4s, 5s, 47s and 167s all the way out to the potential co-occurrence of 1,000 parameters, the correct number of possible biomarkers in this restricted space equals 5.0 x 10^{130}.

The acquisition of 1,000 different measurements from a single patient in a single assay has nowadays almost become routine. In fact, parallel analysis of gene transcript activity is possible for every gene in the human body. Nonetheless, even without going to this extreme, 5.0 x 10^{130} distinct combinations is already an extremely large number. This number is greater than the number of seconds that have elapsed in the history of the universe, ie 13.7 billion years! Thus, we can conclude that finding reliable biomarkers is highly likely to be computationally intensive, no less so than modelling the universe on nuclear physics research today. In the biological universe, finding reliable biomarkers will be equivalent to winning consecutively the National Lottery every week for a whole year. This is the reason why genomics and post-genomic sciences are delivering poor quality, non-validated biomarkers.

Many people were alarmed to learn that human beings possess a mere 30,000 genes in their genome, and that these are almost identical to those of mice. In a human-centric view of the universe, we may feel superior, but the facts suggest otherwise. The almost limitless diversity in facial features, the differences between feet and brains, embryos and old people lies in the combinations of biomolecules produced by the genetic code and interactions with the environment, and not entirely within the DNA itself. In a living cell, such molecular interactions number in the trillions. When searching for the 'cause of common cancers', one can be confident it will not be a single answer, but combinations of many subtle effects. Thus, computational efforts need to be better funded as part of our efforts to cure human disease and improve human well-being and healthcare delivery.

The added hindrance of biological and experimental variance in our data sets

The reality of the figures cited above should instill some degree of apprehension among researchers and executives alike within the pharmaceutical and biotechnological industries, but the litany of woes does not end here. At every turn, biological and experimental variance in our datasets makes valid conclusions increasingly difficult. For example, we must also acknowledge the variance in gene and gene product expression that occurs before and after coffee, breakfast or exercise, while patients in a test group will never possess identical genetic backgrounds; exposure to environmental risk factors; diets or disease time-lines. These factors all combine to generate variance in our data. In summary, the factors outlined above clearly show why the fruits of genomics and the high-throughput screening of biomedical data have yet to impact

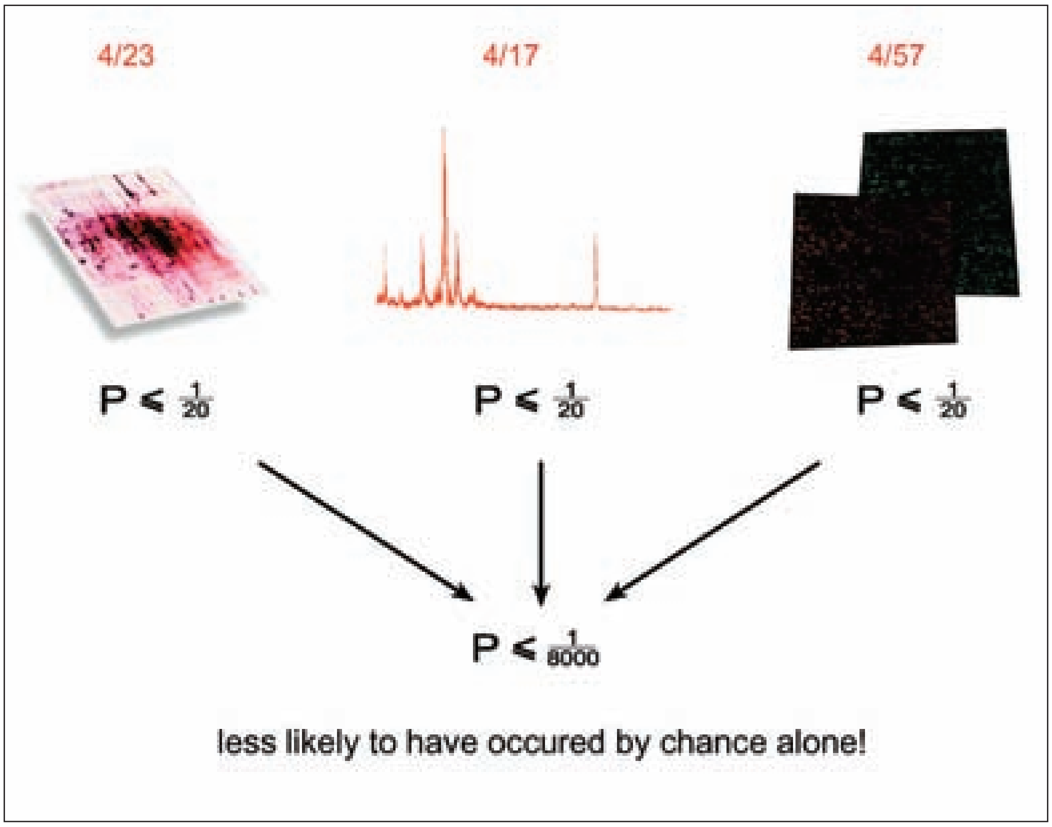

Figure 4
Co-occurrence of gene-product across three noisy, low replicate, but orthogonal experimental approaches. The numbers in red indicate respectively the number of common significant findings within the total number of significant findings obtained. Here, the pharmaceutical industry is well placed because it has at its disposal the prior existence of this resource base, namely, data acquired via orthogonal means on an identical sample set. Unfortunately, it is most often underexploited. Data integrated in a query-specific environment can almost always lead to an enhanced level of understanding, particularly if observations are contextualisation through a systems approach

drug development beyond the more traditional path of trial and error and fortuitous serendipity. Most importantly, this genomic revolution has yet to be exploited in designing better therapeutics with a focus on improved target selectivity when confronted with an almost boundless universe of drug binding options.

Solutions – either continue to hide our heads in the sand or confront this daunting reality

One option for the discipline of genomic and post-genomic technology is to continue with the status quo, a setting not far removed from the scenario presented by Hans Christian Anderson in *The Emperor's New Clothes*.

Much of the current biomarker rhetoric would appear little short of this. For example, approaches such as 'quantiles exclusion' and 'shaving', as currently employed to deal with the high variance characteristic of array data sets, are too blunt an instrument, whereby overlapping variance between test and control or drug and placebo is simply excluded to render data sets seemingly distinct. Populations with overlapping variance and too few replicates add up to poor science.

Alternatively, as thinking scientists, we could face-up to this reality and work better in the face of high dimensional space. The challenge is to constrain high dimensional space by employing:

- Multiple analytical techniques.
  For example, applying microarrays for transcriptomics, 2D gels and mass spectrometry in proteomics – all on the same experiment.
  This use of orthogonal techniques can give greatly improved reliability.
- Multiple analysis algorithms.
  Individual techniques (support vector machines for example) have particular weaknesses. Multiple techniques again increase reliability.
- Integration of observations from the clinic and literature (both patent and scientific).
  Concordance builds confidence.
- Appropriate statistics.
  Sufficient numbers of replicates are crucial to robust results.
- Good experimental design.
  Calibration and quality control are also vital.

Importantly, both multiple analytical techniques and multiple algorithms (collectively we would describe this as Cross Technique Analysis, or CTA) should be employed in the same experimental setting, hopefully to be then underwritten by previous observations from the clinic and scientific literature.

We mentioned above the issue of varying data quality and an inability to ensure homogeneous quality across large datasets. However, that does not mean that we should not strive continuously for data quality assurance. Encouragingly, valid biomarkers will be able to be extracted from the sea of false positives, particularly if more attention is paid to:

- More replicates.
- Data QA.
- Calibration.
- Noise Elimination.
- Signal Extraction.

For example, the mathematics of noise elimination and signal extraction is well evolved in seismology, neurology, cosmology and the semi-conductor industry, and certainly needs to be applied to gene expression datasets. Similarly, in cosmology, dimensionality reduction has been the aim of physicists for many decades in dealing with complex datasets. We can learn from their work – in fact, most often, the biomedical sciences do not have to develop new mathematics, but rather apply existing tools to their data sets in an appropriate manner.

Computational tools can assist in:

- Experimental design.
- Sample collection/preparation/annotation.
- Quality Assurance of incoming and outgoing data at multiple levels.
- Improved dataflows and data warehousing.
- Data acquisition and management.
  Storage, Back-up, Retrieval, Interrogation.
- Noise elimination/signal extraction.
- Linking of disparate data sets and experimental techniques.
- Statistical analysis.
- Use of alternate algorithms to increase confidence in conclusions.
- Summative visualisation/comparison.
- Based increasingly on intelligent knowledge-preening software.

The pharmaceutical industry has realised over the last few years that the idealistic schema presented in **Figure 4** can add value to experimental findings, whereby simplistic Boolean operators can enhance statistical confidence. As in most of biology, things are not going to be as straightforward as one would have wished. In fact, discussions with colleagues currently combining both proteomics and transcriptomics analyses suggest that the experimental commonalities are most frequently encountered at the level of target pathways. Here too, improved mathematics can come to our collective rescue. Dynamic pathway modelling designed to take into account both positive and negative feedback loops have certainly shown their worth in bacterial systems and will increasingly do so in a eukaryotic context. These approaches have much to offer by way of improvements, albeit in a non-intuitive manner, with respect to the more traditional view afforded by static biochemical pathway reconstruction.

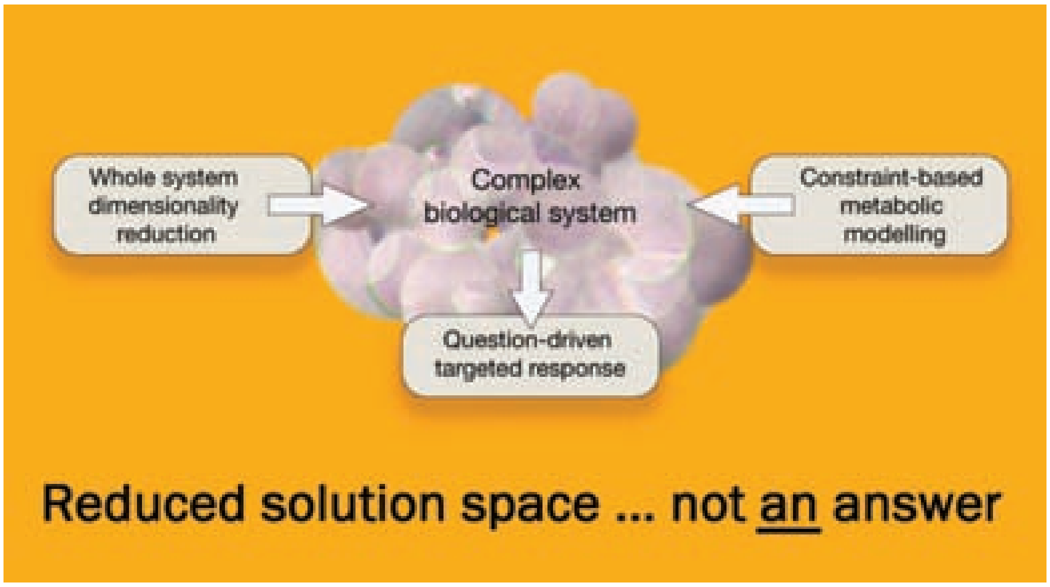
Whole system dimensionality reduction
Complex biological system
Constraint-based metabolic modelling
Question-driven targeted response
Reduced solution space ... not an answer

Figure 5
Biological complexity examined from a question-specific viewpoint. Whole systems cannot yet be accurately modelled, but value-added comprehension can be afforded by such an approach

Figure 6
Pandora considering whether or not to recognise the existence of the multiplicity of drug induced effects (mostly 'off-target effects') and attempting to better contextualise these associated biomarkers in preparation for Voluntary Genome Data Submissions (VGDS). A final version of VGDS was expected from the FDA by March 18, 2005 but was unfortunately postponed. Such a decision on the part of Pandora might allow the art of bioinformatics to impact for the first time the highest value end of drug development pipeline, namely, during final registration

Conclusion

For the first time, an accurate parts-list of all the drug targets (proteins) in the human body is available. As a consequence, we can now better estimate the real potential for adverse drug effects from off-target binding in among some 600 trillion potential binding sites. The FDA understands this and as of spring 2005 we are awaiting the final guidelines for Voluntary Genome Data Submissions (VGDS) for use during final drug registration. The aim of VGDS is a better understanding and contextualisation of the multitude of drug-induced changes in the human body.

The future

The post-genomic technologies of transcriptomics, proteomics, systems biology and target-ligand interaction monitoring and prediction can add value to data contextualisation, especially when both clinical observations and literature are also integrated, but the 'omics' industries have produced too many targets with too little inherent value. Correct interpretation of post-genomic data demands a higher level of mathematics than previously employed by either academia or the pharmaceutical industry. Biomedical scientists and clinicians must increasingly be driven by the symbolic logic of mathematics, thereby avoiding those intuitive 'notions' alluded to by Einstein.

An accurate comprehension of high dimensional space is beyond the realms of intuitive reasoning. Therefore, as in computer science and cosmology, biomedical sciences must set about reducing the dimensionality of their experimental systems. Factors involved in reducing dimensionality of biomolecular interaction space include the concentration of both targets and ligand, on-and off-rates, spatial and temporal segregation, molecular mass of binders, numbers of binders in a biological complex, combinatorial effects and physiological thresholds. These top down approaches can then be combined with 'bottom-up' and dynamic metabolic modelling tools to reduce the solution space.

A single answer is almost always not to be expected in the biological universe, but a reduced solution space can dramatically lessen the quantity and cost of pharmaceutical screening required. Furthermore, an assessment of potential binding-site diversity *in silico* and in the wet lab in pharma can facilitate more informed selection of iterations, isoforms and/or formulations with the least likelihood of adverse drug effects. For these high worth applications and innumerable other uses, well-validated biomarkers must be the goal of the biomedical sciences in the post-genomic era. DDW

__Professor Ian Humphery-Smith__ joined the Biosystems Informatics Institute as the Chief Executive in 2004. He is trained in infectious diseases and molecular biology and is recognised internationally as a pioneer in the field of proteomics. He founded the Human Proteome Organisation and has significant experience in both academia and industry. He continues to provide advice in the sector to both established biotech firms and government funding agencies internationally. For many years, he has been a keen proponent of the need to transform medicine into a more mathematically-driven discipline and was honoured for these efforts in 2001 by TIME *magazine as a member of their Top Digital 25.*

__Will Dracup__ founded Nonlinear Dynamics in 1989. An economics graduate and skillful software engineer, he became interested in the analysis of protein separations after working for a life science imaging company. Deciding that the available technology was inadequate, he went on to develop software that has since become the industry standard in its field. Since then, Will has led Nonlinear to become a multi-million pound company. Will is also Chairman of the newly established Biosystems Informatics Institute (Bii) in the UK, undertaking groundbreaking research into bioinformatics applications in the Life Sciences.