17

High Content and Throughput Drug Discovery

Quin Wills
SimuGen, London and Kuala Lumpur
United Kingdom and Malaysia

1. Introduction

1.1 The marriage of 'high throughput' and 'high content'

While the pharmaceutical industry innovation crisis draws much debate (Kaitin & DiMasi, 2011; Macarron et al., 2011; Munos, 2009; Paul et al., 2010), there remains little consensus on how to cohesively deliver value throughout the drug development pipeline (Fig. 1). This chapter considers some of these issues in the context of a growing field for computational biology: drug discovery high throughput screening (HTS). HTS is the approach of rapidly studying physical, chemical, biological and genetic perturbations on the scale of of tens of thousands per day. Today we are faced with ultra-HTS daily screen rates of hundreds of thousands, in part thanks to the continued development of technologies such as micro-fluidics (Agresti et al., 2010). As a discovery tool, it traces it roots back over twenty years (An & Tolliday, 2010), however it is the more recent improvements in cell culture technique - with the potential for multivariate output such as gene expression - that brings it into the domain of high content computational biology. With this maturation of cell-based assays we also notice an increased focus on statistical rigour, analytical integration, and the apparent user-driven plateau in miniaturisation (Mayr & Bojanic, 2009). Rather than being faced with a continued improvement in simple assay throughput, these suggest a growing role for more data-rich high content HTS (hcHTS)[1].

Despite the implicit gains, there exists a notable and growing antipathy towards many 'big data' approaches as discovery tools. Much publication has refocused on data quality versus quantity, with some doubting the impact of high throughput science altogether (Douglas et al., 2010; Macarron et al., 2011; Mayr & Bojanic, 2009). There persists the very real hurdle of experimentalists and team leaders struggling with the interpretation, integration, and decision making based on such data. As a concern routinely witnessed in post-genome era science, it is doubtful that the blame rests primarily with problem-specific methodology. In this chapter, the need for screens to be more decision-centric and transparent across disciplines is proposed. The aim here is not just to provide the reader with specific tools that are likely to rapidy become dated, but introduce the scope and opportunities in drug screening science.

[1] In this chapter 'high content' is used interchangeably with 'high dimension', as applied to multiplexed technologies that produce many descriptors per sample, well or observation. A common example is gene expression microarrays. 'High content screening' is commonly used in the literature to indicate high content imaging. To avoid confusion here, the high content approach to HTS is referred to as hcHTS, and high content imaging is referred to as HCI.

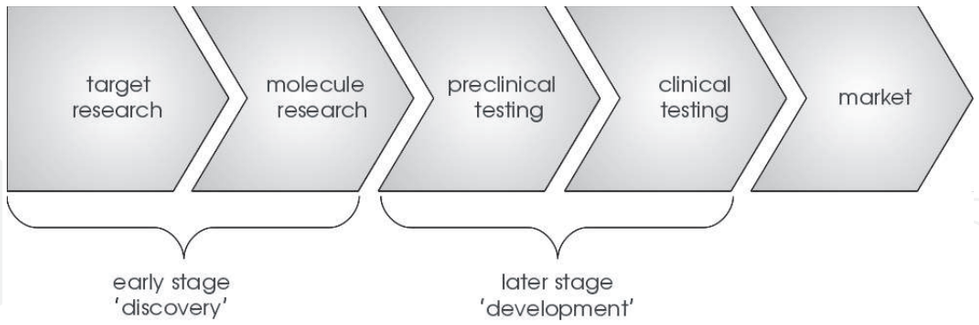


Fig. 1. No single approach is prototypical of drug development. Particularly as a growing number of therapeutic programmes focus on 'biologics' - such as proteins and RNA inhibitors - versus 'small molecule' chemical therapies. However, in general it remains an arduous, failure prone process of 10-15 years, costing hundreds of millions to billions of US$ (Adams & Brantner, 2006). The pipeline can be described as a task in managing attrition rates; a process with a very low success rate sometimes beginning with many hundreds of thousands of chemicals to launch a single successful therapy. The research and development life of a potential drug might be considered in five phases. The first being the identification, development and validation of a target for the drug; the most common targets being G-protein coupled receptors and kinases. The second phase involves the discovery of 'hit' chemicals affecting the target, and development of the hits into leads. Hit through to lead research often begins as high throughput assays, where large libraries of chemicals are screened for effect and sometimes side-effect. What remains are the development phases of animal (preclinical) and human (clinical) testing prior to market release and surveillance (pharmacovigilance).

1.2 The vital role of computational biology

hcHTS provides a unique challenge to the computational biologist more familiar with high dimension analysis. It increases the analytical demand from the 'few observations, many descriptors' paradigm of small sample multiplex genomics to that of 'many observations, many descriptors'. Drug discovery is also gradually devolving its chemo-centric dominance into an increasingly bio-centric approach. This positions computational biology as a crucial bridge between complex science and technology, and the challenging decisions that need be made from the data produced. The melting pot of *in vitro* (cell-based and biochemical) biology, cheminformatics, bioinformatics, systems biology, and 'big data' analysis requires broad inter-disciplinary scientific and computational strengths. It affords the computational biologist the opportunity to become part of a wide ranging science. A practice where hypotheses and data iteratively refine screens and studies, converging on greater scientific understanding and defined solutions.

This chapter is divided into two main parts. Section 2 contextualises some of the challenges and considerations to guide the choice of modelling strategy, whilst section 3 provides a simple predictive toxicology example that builds on these suggestions. Two traditionally medium throughput multiplex approaches - now increasingly being used in

higher throughput settings - will be discussed: gene expression and high content imaging (Bickle, 2010; Zanella et al., 2010). For other promising hcHTS technologies, such as flow cytometry (Edwards et al., 2009) and label-free methods for real time living cell assessment (Xie et al., 2009), the reader is referred to the provided citations. High content imaging (HCI) utilises high resolution multiplex fluorescence microscopy - typically immunofluoresence - to study cellular architecture and health (Karol Kozak, 2009; Zanella et al., 2010). Its strength as a tool is the single cell resolution of physiologically relevant endpoints. HCI together with transcriptomics might be thought of as high content cell and molecular phenotyping. While gene expression analysis is not typically considered part of phenotypic assays, in the context of hcHTS where perturbed pathways and their reporter genes are are studied as indicators of biological process and cell state, it should very much be seen as a proxy of the cell's molecular phenotype. An example of where the two approaches have become inextricably linked is RNA inhibition screens (Karol Kozak, 2009).

2. Modern high throughput drug discovery

2.1 'Big data' analysis paralysis

Not without its critics (Douglas et al., 2010), the ongoing drug discovery mantra has been one of managing attrition rates by 'failing early, failing often'. However, the biological and drugability knowledge around validated targets has remained poor. An often cited FDA white paper of the early post-genome years (FDA, 2004) drew widespread attention by calling for the greater use of biomarkers and computational approaches to improve this knowledge. With the strong political willpower to modernise drug discovery, HTS has continued to gain popularity as a brute force innovation tool, entering the public domain with resources such as ChemBank (http://chembank.broadinstitute.org), PubChem (http://pubchem.ncbi.nlm.nih.gov) and ChEMBL (https://www.ebi.ac.uk/chembldb). Progress has however faced persistent concerns, with common complaints being poor chemical library design (Gillet, 2008) and that of decision-makers drowning in data. While chemical library design is beyond the scope of this chapter, the data concern is one familiar to every high content computational biologist.

A contrasting argument to the suggested deluge of data as the core concern is that the principal challenge lies with modelling strategy; not the data per se. A case in point might be made of the much publicised Large Hadron Collider with its daily data quota exceeding 40 terabytes. This represents more data than that managed by a typical computational biology team and - while still requiring considerable computing resources - remains a manageable data flow. This is arguably due to well developed theoretical models around which physics expects the data to behave. In contrast, theoretical and systems biology still suffer from a paucity of rigorous quantitative models relevant to disease and chemical biology. The ship may be sinking not because the ocean is large, but because the water is bailed with teaspoons. What then are the most appropriate strategies? To answer this we first need to appreciate that the challenge with screening science is less that of providing narrowly focused yes/no answers. Rather it is more a task of iteratively triaging the optimal options, while managing the decision-making risk across heterogeneous studies spanning months to years. Though not a style of research unaccustomed to statisticians and decision analysts, this challenges computational biology culture with its often data-centric rather than decision-centric and translational mandates.

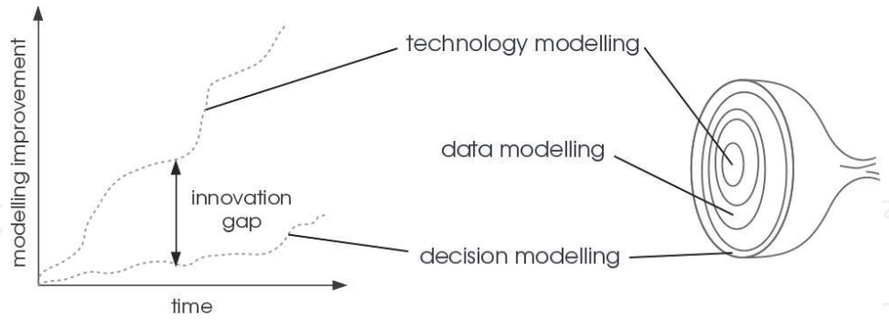


Fig. 2. A useful paradigm for HTS is that of the layers in a 'modelling onion', which emphasises the crucial role of the computational biologist, bridging technology and scientific decision making. Initial research is often driven by technology modelling: choosing the optimal biological models, experimental protocols and technologies to provide good data signal. Data modelling is the remit of computational biology, which might be divided up as a spectrum of low level data clean-up through to higher level theoretical and systems modelling. The screening computational biologist needs to balance the merits of providing detailed results versus fast results; the latter proving useful if they enable the research team to make real-time decisions and rapidly test hypotheses. In HTS this is often the balance of primary versus secondary screening strategies. The final, often neglected, layer is decision modelling. No matter how well the HTS technologists and informaticians consider their models to be performing, if these don't explicitly and transparently assist large discovery teams in making decisions, they are effectively of little use.

2.2 Improving your modelling IQ^2

Modern biology retains its distinctive knowledge-driven culture as a science; differentiated from more mature sciences as being heavily dependent on phenomenological 'stamp collecting'. Similar divides manifest in computational biology as low level data collection, clean-up and mining of bioinformatics versus computational biology modelling. In research with direct translational and economic goals - such as drug screening - it is helpful to remember that:

- Science exists to create explanatory and/or predictive models. Cohesive and comprehensive modelling practice along the entire drug development pipeline is the mandate of all researchers from *in vitro* to *in silico* to the patient.
- All models are wrong, some are useful. Particularly in HTS drug discovery, the development and use of models should be driven by their utility as transparent, triaging decision tools, not narrowly focused technological arguments.
- HTS combines three levels of modelling. Technology, data and decision models should be seen as essential layers in a 'modelling onion' (Fig. 2).

There is, of course, no silver bullet to address data rich problems in drug screening. Notwithstanding, there are general considerations before deciding on methods to optimise the screening model (Fig. 3). A few overlapping rules of thumb are suggested here as a measure of a screen's IQ^2. The test for IQ^2 summarises the need for better *integration*, assay *quantitative*

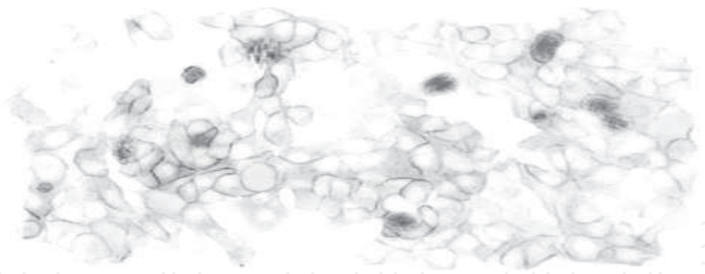


Fig. 3. The above immuno-stained cells - after timed exposure to a toxic chemical - provide a simple example of the screening model in three parts: technology, data and decision modelling. Here technology modelling involves the choice of an informative fluorescent biomarker in an appropriate cell model after an optimal perturbation duration. The data model might be to infer the lowest concentration at which 10% of cells are statistically significantly brighter than twice the average baseline fluorescence. This type of model is a 'lowest dose with an effect' model, where the combination of statistical and biological significance define the concentration output. The utility as a decision tool might be to provide a reproducible relative measure of cytotoxicity across collaborating laboratories interested in a simple ranking of cytotoxic effect, viz. a robust measure best suited as an ordinal triage of effect. The aim being hazard identification, with little explicit attention to risk management and translational or economic impact.

performance, and the decision-making *synergies* (the squared exponent) which present with the action-enabling results.

How well does your approach integrate? Integration entails more than just the use of all available data, but includes the effective integration along the entire flow of data through to knowledge and scientific wisdom (Fig. 4). This is the central tenet of translational bioinformatics, which aims to promote free flow of data between the lab and patient (Buchan et al., 2011). Still in the early stages, translational informatics projects such as Informatics for Integrating Biology and the Bedside (http://www.i2b2.org) hold much promise for feeding back into HTS.

How quantitative is your approach? The quality of the inference is limited by the quantitative performance of the screen. Too often it would seem that post hoc analysis attempts to stretch the assertions made by screening models not fit for purpose. In HTS the primary measures of interest are dose and time. If, for example, a screening programme is required to predict a new drug's safety concerns ('how toxic?'), these might be framed as one or more of many dose and time relevant questions. A few translational toxicity concerns are listed below:

- The concentration at which a percentage of the population begin to experience an effect.
- The concentration at which the risk of rare (unpredictable) adverse effects becomes too great.
- The extent of pathology after a set dose and time exposure.
- The optimal dosing schedule to minimise toxicity without significant loss of efficacy.
- Chronic affects - such as bioaccumulation - less easily extrapolated from acute and sub-acute testing.

Fig. 4. The flow of data into results and decisions reflects the well described flow of information into knowledge and wisdom. Bioinformatics began, in part, as a field to address data integration concerns (Searls, 2010). Today the integration of technologies, laboratories and heterogeneous databases is common practice, and remains vibrant with emerging resources such as cloud computing (Mak, 2011; Schadt et al., 2010). Less well practised is the routine and formal integration of results beyond simple score-based meta-analyses. Bayesian computation promises more formal approaches to update results and incorporate prior information, yet advanced statistical treatment remains underutilised in modern HCS (Malo et al., 2006). Again, the importance of assisting with decision making deserves greater attention. Modelling approaches need be transparent enough to allow a diverse community of scientists to easily communicate and understand the analytical assumptions and limitations.

A role of the screening computational biologist should be seen as providing reliable quantitative measures of concentration and time to hypotheses/questions; not just the provision of IC_{50} or EC_{50} values per variable. The concern, for example, is not the reliable measure of gene expression and the confidence around these measures per se. Rather, it's the transformation of these values into measures of concentration and time, and the confidence around these measures.
Three quantitative concerns often deserving better consideration are suggested:

- The first is the signal-to-noise ratio (Fig. 5A), commonly measured as $Z = 1 - \frac{3(\hat{\sigma}_p + \hat{\sigma}_n)}{|\hat{\mu}_p - \hat{\mu}_n|}$, for positive and negative controls p and n respectively. A Z-score greater than 0.5 is typically accepted to suggest a good assay. The Z-score is a narrowly focused measurement aimed at single-plex assays, which is not robust and assumes data normality. It also does not take into account the performance impact on decision making.
- A second concern is dynamic range. Not all cell models or technologies provide an adequate dynamic range in which drug effects over wide concentration ranges can reliably be measured by a broad spectrum of markers exhibiting near-linear correlation with the effects they proxy. A well known example of this is the poor dynamic performance of gene expression microarrays demonstrated by the Microarray Quality Control project (MAQC Consortium et al., 2006). The screening computational biologist needs to clearly demonstrate the adequate dynamic performance of their data prior to establishing any routine screening.
- The third and final consideration can be broadly defined as that of information resolution. Screens and their follow-up secondary screens/studies need to define clear goals of improving concentration and time sampling density to ensure accurate and sufficiently precise quantitative assertions. The results also need to be presented to the decision-making team at an optimal resolution to be informative without being overwhelming (Fig. 5B - 5D).

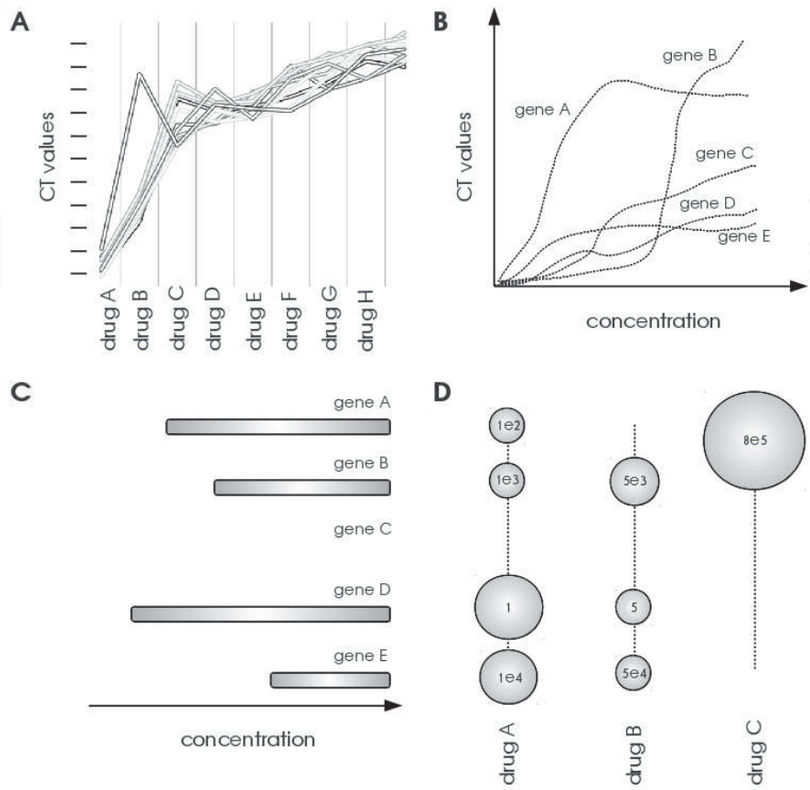


Fig. 5. **(A)** quantitative PCR measurement of gene expression remains the most common gold standard for assessing gene expression without the dynamic limitations of nucleic acid hybridisation technologies, such as gene expression microarrays. However, as can be seen in this well controlled example of highly replicated measurements for a single gene across several compounds, achieving repeated *in vitro* measurement within 1 CT unit remains a challenge. Assuming near optimal reaction efficiency, the CT scale approaches $\log_2$, indicating the cost and time challenge of adequate replication to confidently discern the doubling of a gene's expression under screening conditions. **(B-D)** The resolution at which it is optimal to present results affects the design and/or execution of the data modelling. Figure B provides a detailed trace of five genes perturbed by a compound in a secondary screen. While being detailed, it is ineffective at answering 'at what dose?' and quickly becomes intractable in terms of technical cost and analysis when comparing multiple compounds. Figure C partly resolves this by presenting the results as bars beginning at the lowest concentration at which the answer to the question becomes true. Figure D presents this information for the same genes, comparing tested compounds, numerically providing the concentrations and displaying the statistical confidence in the results as being proportional to the bubble size. Here we see that drug B behaves most similarly to drug A, but at a 5-fold higher concentration (lower potency). It would seem from the screen that we can be fairly confident that drug C behaves differently from drugs A and B.

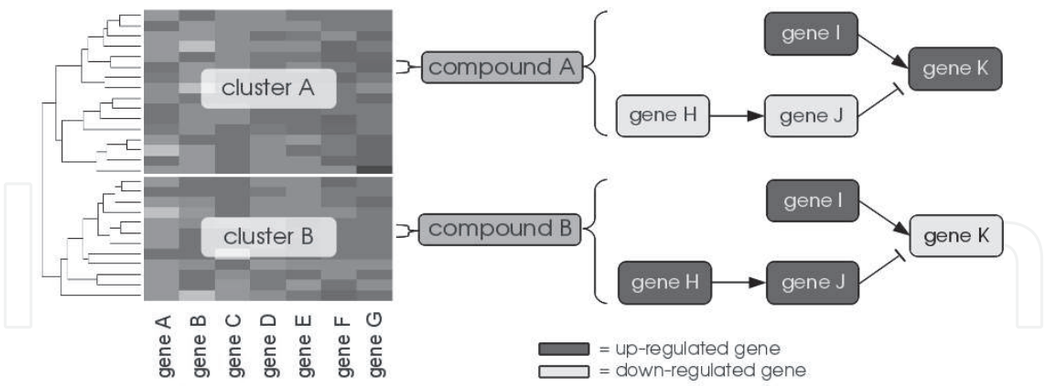


Fig. 6. Careful consideration need always be given to the actionability of screening methods used. If, for example, compound A clusters (or classifies) together with a prototypical compound in cluster A, while compound B clusters together with another prototype in cluster B, does this provide sufficient information to prioritise compound A with a defined dose and time dependent confidence? Similarly with pathway-driven approaches. If compound A does not up-regulate genes H and J at low concentrations, how does this translate into dose and time dependent effects for the purposes of screen prioritisation?

How actionable is your approach? The most important consideration is how effective the screening strategy is at enabling the team to make informed decisions that lead to clear actions where the utility, cost and risk attached to those actions are understood. These can again be considered at the technology, data and decision modelling levels.

Technologist bias will routinely be towards increasing technology complexity within time and cost restraints. However, increased complexity needs to translate into improved actionability. The debate on simple cell culture techniques replaced by the earlier use of lower throughput three-dimensional approaches (Fernandes et al., 2009) highlights this concern. Complex *in vitro* approaches run the risk of compromised data reproducibility. If reduced reproducibility and cost of technology complexity outweigh potential gains in insight, the technological improvements and necessary data modelling changes need be questioned.

The over-reliance on exploratory bioinformatics without clear quantitative questions, hypotheses and follow up is arguably core to current innovation failures (Fig. 6). Three pillars of result significance enable the rational implementation of screens:

- The first is statistical significance, which has traditionally played a minor role within single-plex HCS (Karol Kozak, 2009; Malo et al., 2006).
- Not to be confused with statistical significance is biological significance. A differentially expressed gene defined purely in terms of statistical confidence above baseline needn't represent its biological relevance as a useful marker to elucidate mechanism or enable clear actions based on screening questions.
- The economic argument forms the final pillar, where cost is considered together with utility (Swamidass et al., 2010).

Bayesian methods provide a useful framework to formally work with these different notions of importance, whilst also enabling the use of external data - such as cheminformatics and

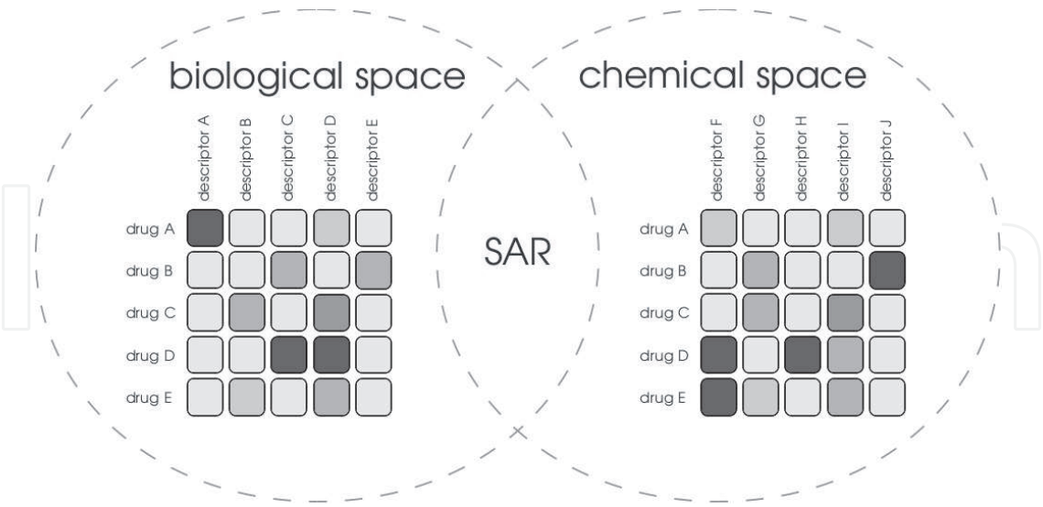


Fig. 7. Common cheminformatics practice is to define descriptors of structure and physiochemical properties in order to position a compound in chemical space. If the purpose of bioinformatics is not only to define molecular and cellular phenotype patterns and mechanisms, then the role of the screening computational biologist effectively becomes a collaborative counterpart to the cheminformatician, defining biological descriptors that are not merely useful as biologically predictive or mechanistic markers, but can be mapped to chemical descriptors in order to define structure-activity relationships (SAR) for rational drug design.

pharmacoeconomics - to establish meaningful priors. However Bayesian analysis appears notably absent in routine published practice (Klon, 2009; Nidhi et al., 2006). Computational limitations seem less likely than the poor understanding surrounding the use of these methods. The need for model understanding and transparency by the entire decision making team is paramount. So until formal frameworks can meet this need, we are left to rely on simpler approaches, some of which are discussed later in this chapter.

Finally, it might be argued that screening methods should ultimately strive not to provide a 'post-mortem' of results but actively assist the discovery team to design better therapies. If, for example, a screen has been developed to predict a spectrum of toxicities from tested chemicals, it should not only accurately identify the correct toxicities, but also their dose-time properties while guiding the chemists on how to alter the compound structures to improve their safety profiles. This re-emphasises the argument for generalist computational biologists in HTS who are able to collaborate with the chemical design and cheminformatics teams, identifying actionable structure-activity relationships (Fig. 7).

3. 'Next generation' drug screening

The title above has deliberately been borrowed from the same description applied to second generation nucleic acid sequencing technologies. It is used in part to stress the increasingly high content flavour of HCS, but also the need for a new screening paradigm focused on a cohesive, transparent and actionable modelling practice. Drawing from real data, this

section demonstrates how a screen for genotoxicity might be created using currently available software. The aim is to present how simple rules, weights and thresholds can be used as one approach to create screens not only with good performance characteristics but which can be easily understood and acted on by all team members.

3.1 Rules, weights and thresholds

Cheminformatics has routinely utilised machine-driven pattern recognition to distil large data sets into rule-based models as a form of 'human readable' modelling, or rules of thumb, to predict drug properties. A well known example applicable to ADME is Lipinski's Rule of Five, which assesses how likely it is that a chemical will be orally active. To pass Lipinski's rule, a chemical is limited to violating no more than one criterion:

- Less than or equal to five hydrogen bond donors.
- Less than or equal to ten hydrogen bond acceptors.
- Less than or equal to 500 daltons in molecular weight.
- An octanol-water partition coefficient log *P* less than or equal to 5

In a similar vein, standard bioinformatics methods can be distilled into combinations of rules, creating such models. These can be tested, refined and understood by non-specialists across the drug discovery team (Fig. 8), with biological and decision-relevant significance better ensured by applying transparent weights and thresholds (Fig. 9). As a consequence of increased computing power and data set size, it seems likely that rule-based approaches will grow in popularity as a tractable modelling strategy. Rule-based modelling has already proven popular in systems biology, where unmanageable lists of differential equations have yielded to agent-based rules of interaction used to drive simulations (Barnes, 2010; Krakauer et al., 2011; Yoav Shoham, 2009). As a methodological approach, rule-based modelling provides a natural bridge for team-driven hypothesis generation and the maturation of generalities for mechanistic and screening biology treated as information science. A particular benefit of rule-based models in screening is that it also allows for the seamless integration of multiple data types. A model might be a collection of rules from multiple cell culture models, multiple time scales, and multiple technologies such as quantitative PCR, HCI, and classical cytometry. Collectively, all of these benefits ensure a high IQ^2 for rule-based models.

3.2 Screening with HT-Stream™

In vitro drug safety screening currently falls within the domain of what are typically medium throughput models aimed at predicting drug absorption, distribution, metabolism and excretion (Ekins et al., 2005). These models are collectively referred to as predictive ADME. Combinatorial chemistry and the shift of ADME to early stage discovery have both significantly improved our ability to design efficacious pharmaceuticals. This has left drug toxicity as an important bottleneck contributing to the innovation crisis, and has prompted its shift to earlier 'off target' cell screens. Whilst born out of lower throughput toxicogenomics (Van Hummelen & Sasaki, 2010), this shift of high content application to high throughput screening requires new methodology and software.

Fig. 10 provides an example of two established DNA damage and stress markers tested on the hepaRG® human liver cell co-culture (Guguen-Guillouzo & Guillouzo, 2010) using quantitative PCR. Microfluidic 'lab-on-a-chip' improvements have enabled cost-effective, high throughput quantitative PCR (Stedtfeld et al., 2008); dramatically improving the scalability of this gold-standard technology as a stand-alone tool or in conjunction with

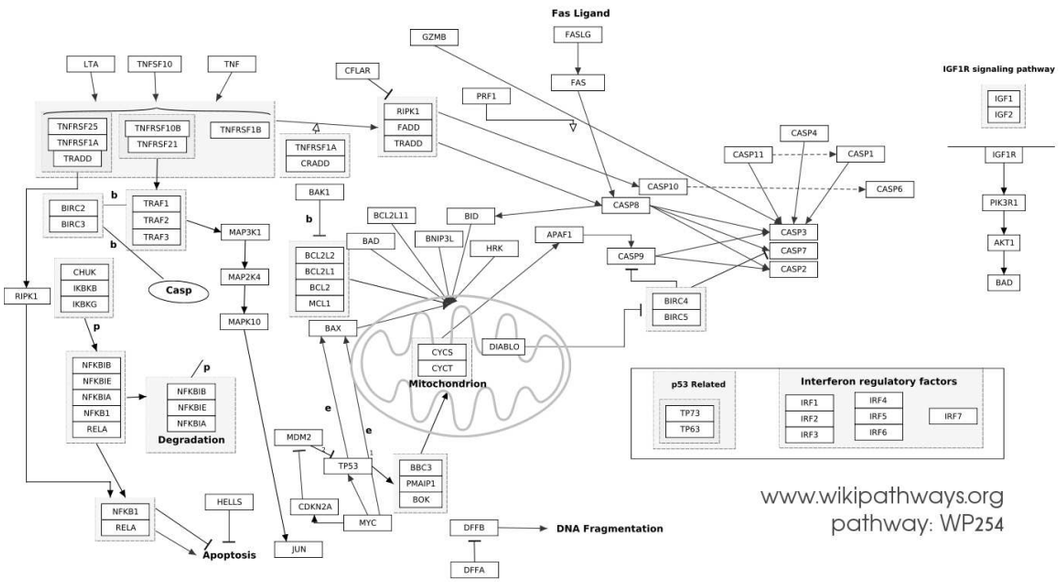


Fig. 8. The apoptosis pathway, as represented above, is commonly used to predict drug safety concerns. A typical bioinformatics approach in a high dimension setting might be to search for apoptosis gene enrichment. Gene set enrichment based on established pathways or ontologies is a rough exploratory tool that translates poorly into a setting where a precise dose-response relationship is required. Using unguided machine learning, literature mining and expert knowledge, complex pathways can be stripped down to collections of rules able to be refined over time and combined to form rule-based models. An example of a rule-based model might be 'clinically significant human apoptosis when at least one, but no greater than four of the following gene ratios hold true…'. In a dose-response setting, the lowest concentration at which the rule holds true is called, as a 'lowest dose with an effect' model.

other high content methods such as HCI. While it remains traditional to begin testing gene expression at cell cytotoxicity IC_{50} concentrations, these do not represent physiologically appropriate dosings. As suggested in the data correlations of Fig. 11, HCI allows for mechanistically relevant concentrations to refine classic viability assays in the absence of reliable human data; and discover transcriptomic biomarkers for use in rule-based models. Once the models have been developed, HT-Stream™ (www.ht-stream.com, www.simugen-global.com) proves useful as online collaborative software that presents the results in a decision-focused manner (Fig. 12). All submitted screening data undergoes automated quality control (Fig. 10A), prior to the inference of the lowest concentration at which each rule becomes true. HT-Stream™ uses the derived concentrations, together with weights and thresholds to help prioritise compounds, visualise results, and compare models (Fig. 13). Easy-to-use software such as this helps make it possible for teams to create 'ecosystems' of applications, rules and models; continuously refining them as collective interdisciplinary knowledge grows with transparent, decision-centric screens.

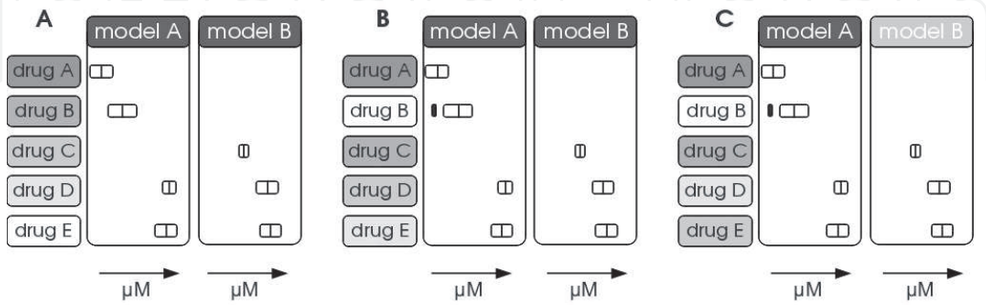


Fig. 9. The above plots represents the output for two 'lowest dose with an effect' models for five tested compounds. The aim here is to rank compounds from highest to lowest toxicity, viz. prioritise those presenting with toxicity at lower concentrations. The darker a compound's label, the higher it is prioritised. The x-axes represent increasing microMolar concentrations of the compounds, with bars representing 95% confidence intervals. **(A)** Here we observe the simplest prioritisation: drug A demonstrates high potency (from model A) and so is ranked first. The benefit of presenting data as concentration values, and the statistical confidence around those values, is evident. While drug A is prioritised over drug B, their confidence intervals overlap, suggesting insufficient statistical evidence to support the ranking. **(B)** The same results are plotted, but with a threshold added to drug B. In this example, drug B has prior information regarding its therapeutic efficacy. The discovery team have decided that any toxicity called above a certain threshold will be of little consequence, as it is unlikely to be reached at therapeutic concentrations. By including this threshold, drug B is de-prioritised. **(C)** The previous rankings assume an equal weighting of the two models. In reality this is rarely the case. If model A represents the drug's carcinogenic potential, whilst model B represents a low-grade safety concern, then model A requires a greater weighting in the global prioritisation. Here drugs D and E switch positions as model B is assigned a low weighting. Thresholds and weights ensure transparent assumptions of biological relevance. With the inclusion of prototypical compounds in the test rankings, transparent weights, thresholds, and statistical significance enable the team to collectively make informed, defensible decisions.

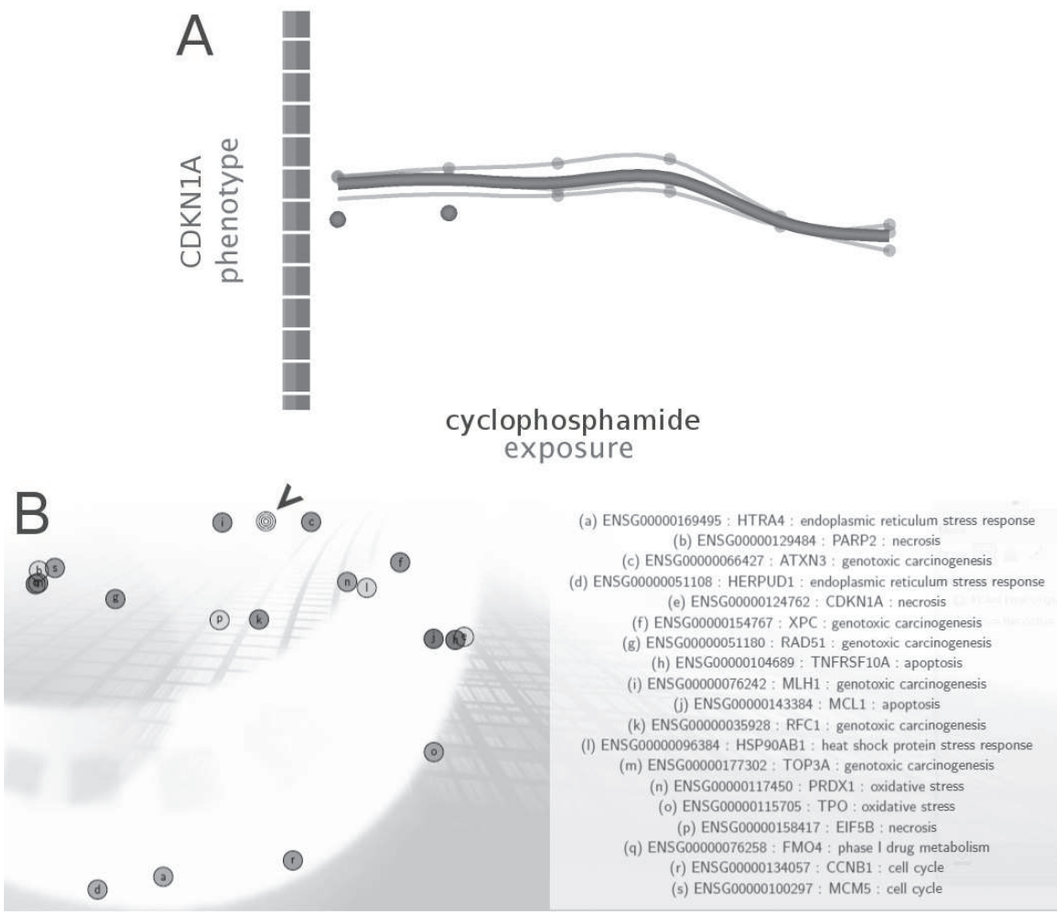


Fig. 10. Genotoxicity biomarkers in hepaRG® viewed using the online tools provided by SimuGen. **(A)** The doubling of CDKN1A's expression with high dose cyclophosphamide. The x-axis represents increasing compound concentration, while each tick in the y-axis represents a CT unit; a drop in one unit thus representing a doubling in gene expression. The analysis tools provide robust automated quality control, in this case identifying two measurements believed to be outliers in bold. **(B)** SimuGen's biomarker discovery tools provide a reference database for over 22,000 genes tested across multiple chemical perturbations in hepaRG®. The above result for GADD45A has identified its most strongly correlated toxic biomarkers, and plotted their first two principal components. GADD45A is highlighted with an arrow and can be seen to be closely clustered with, and enriching for, known genotoxic biomarkers.

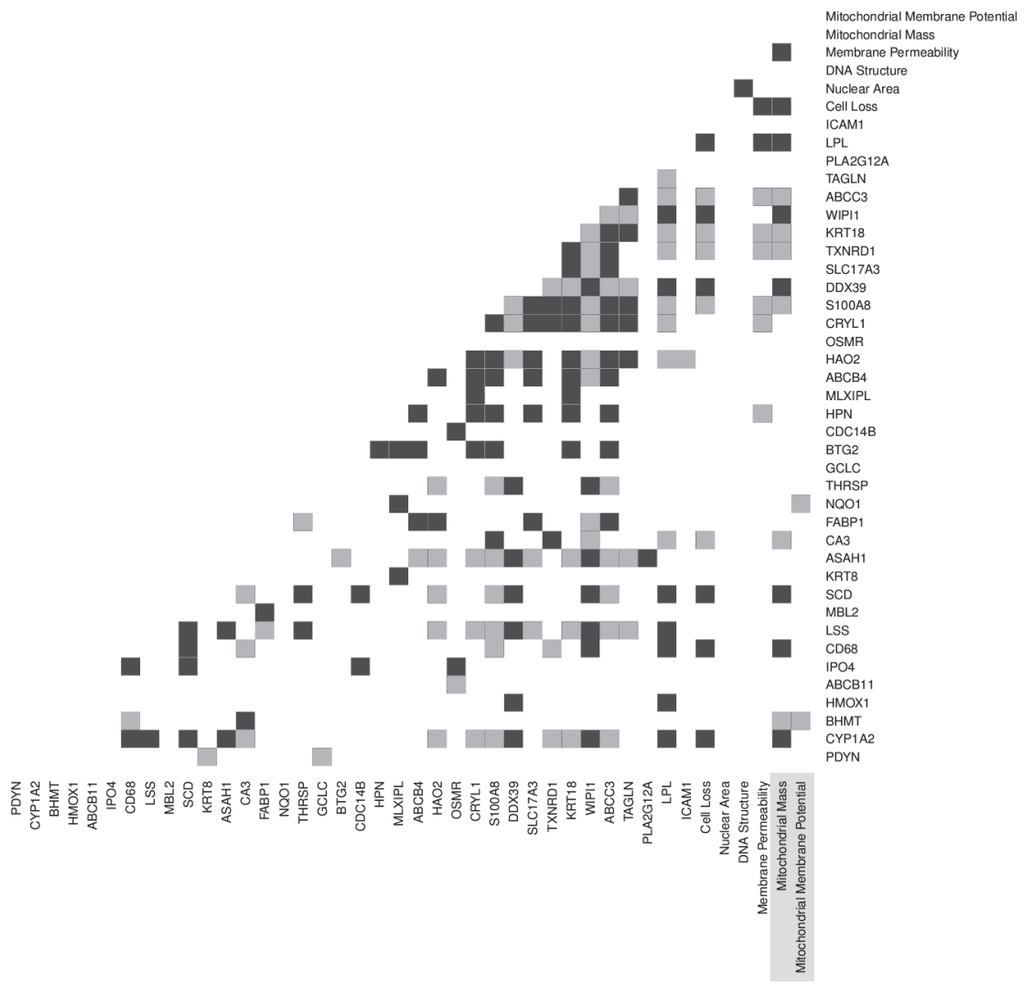


Fig. 11. HCI allows the standardisation of compound concentrations using mechanistic criteria. This correlation plot demonstrates strong correlation (dark squares: Pearson >0.8) and anti-correlation (light squares: Pearson<-08) between known hepatotoxic biomarkers and microscopic phenotypes. The gene expression profile for each compound is measured at the lowest concentration at which any HCI phenotype emerges. Considering drop in mitochondrial mass and potential as a joint phenotype (highlighted) shows strong association with with stress, metabolic and cirrhosis markers. Strong correlations such as these indicate the compatibility of the approaches and the ability to used joint HCI and gene expression data in rule-based models.

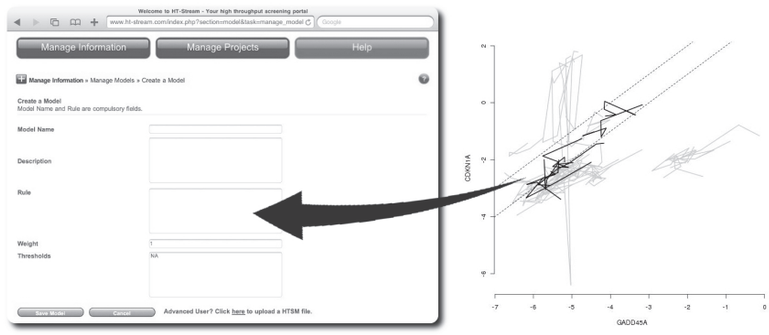


Fig. 12. The right-hand side plot traces the paths of GADD45A and CDKN1A over almost 50 chemicals as their concentrations increase. The black paths represent known genotoxic drugs. It can be seen that there is a 'golden ratio' for the two genes between the dotted lines. Most compounds fall below, whilst non-genotoxic compounds typically present above. HT-Stream^{TM} allows such rules to be entered.

Fig. 13. Using the weights and thresholds, all tested compounds are ranked in HT-Stream^{TM}. Any model with a positive result has its results plotted, and contrasted to similar behaving compounds, as described in Fig. 9. A principal components plot, using all models, is also provided to allow the computational biologist and chemist to identify overall patterns that might be related back to chemical structure.

4. References

Adams, C. P. & Brantner, V. V. (2006). Estimating the cost of new drug development: is it really 802 million dollars?, *Health affairs (Project Hope)* 25(2): 420–428.
URL: *http://dx.doi.org/10.1377/hlthaff.25.2.420*

Agresti, J. J., Antipov, E., Abate, A. R., Ahn, K., Rowat, A. C., Baret, J.-C. C., Marquez, M., Klibanov, A. M., Griffiths, A. D. & Weitz, D. A. (2010). Ultrahigh-throughput screening in drop-based microfluidics for directed evolution., *Proceedings of the National Academy of Sciences of the United States of America* 107(9): 4004–4009.
URL: *http://dx.doi.org/10.1073/pnas.0910781107*

An, W. F. & Tolliday, N. (2010). Cell-based assays for high-throughput screening., *Molecular biotechnology* 45(2): 180–186.
URL: *http://dx.doi.org/10.1007/s12033-010-9251-z*

Barnes, D. J. (2010). *Agent-based modeling*, 1 edn, Springer.
URL: *http://www.springer.com/computer/theoretical+computer+science/book/978-1-84996-325-1*

Bickle, M. (2010). The beautiful cell: high-content screening in drug discovery., *Analytical and bioanalytical chemistry* 398(1): 219–226.
URL: *http://dx.doi.org/10.1007/s00216-010-3788-3*

Buchan, N. S., Rajpal, D. K., Webster, Y., Alatorre, C., Gudivada, R. C., Zheng, C., Sanseau, P. & Koehler, J. (2011). The role of translational bioinformatics in drug discovery, *Drug Discovery Today* .
URL: *http://dx.doi.org/10.1016/j.drudis.2011.03.002*

Douglas, F. L., Narayanan, V. K., Mitchell, L. & Litan, R. E. (2010). The case for entrepreneurship in R&D in the pharmaceutical industry., *Nature reviews. Drug discovery* 9(9): 683–689.
URL: *http://dx.doi.org/10.1038/nrd3230*

Edwards, B. S., Young, S. M., Ivnitsky-Steele, I., Ye, R. D., Prossnitz, E. R. & Sklar, L. A. (2009). High-content screening: flow cytometry analysis., *Methods in molecular biology (Clifton, N.J.)* 486: 151–165.
URL: *http://dx.doi.org/10.1007/978-1-60327-545-3_11*

Ekins, S., Nikolsky, Y. & Nikolskaya, T. (2005). Techniques: application of systems biology to absorption, distribution, metabolism, excretion and toxicity., *Trends in pharmacological sciences* 26(4): 202–209.
URL: *http://dx.doi.org/10.1016/j.tips.2005.02.006*

FDA (2004). Innovation or Stagnation: Challenge and Opportunity on the Critical Path to New Medical Products.

Fernandes, T. G., Diogo, M. M., Clark, D. S., Dordick, J. S. & Cabral, J. M. S. (2009). High-throughput cellular microarray platforms: applications in drug discovery, toxicology and stem cell research, *Trends in Biotechnology* 27(6): 342–349.
URL: *http://dx.doi.org/10.1016/j.tibtech.2009.02.009*

Gillet, V. J. (2008). New directions in library design and analysis, *Current Opinion in Chemical Biology* 12(3): 372–378.
URL: *http://dx.doi.org/10.1016/j.cbpa.2008.02.015*

Guguen-Guillouzo, C. & Guillouzo, A. (2010). General review on in vitro hepatocyte models and their applications., *Methods in molecular biology (Clifton, N.J.)* 640: 1–40.
URL: *http://dx.doi.org/10.1007/978-1-60761-688-7_1*

Kaitin, K. I. & DiMasi, J. A. (2011). Pharmaceutical innovation in the 21st century: new drug approvals in the first decade, 2000-2009., *Clinical pharmacology and therapeutics*

89(2): 183–188.
URL: *http://dx.doi.org/10.1038/clpt.2010.286*

Karol Kozak, A. A. (2009). Data Mining Techniques in High Content Screening: A Survey, *J Comput Sci Syst Biol* 2: 219–239.
URL: *http://www.omicsonline.com/ArchiveJCSB/2009/August/04/JCSB2.219.xml*

Klon, A. E. (2009). Bayesian Modeling in Virtual High Throughput Screening, *Combinatorial Chemistry & High Throughput Screening* 12(5): 469–483.
URL: *http://dx.doi.org/10.2174/138620709788489046*

Krakauer, D. C., Collins, J. P., Erwin, D., Flack, J. C., Fontana, W., Laubichler, M. D., Prohaska, S. J., West, G. B. & Stadler, P. F. (2011). The challenges and scope of theoretical biology, *Journal of Theoretical Biology* .
URL: *http://dx.doi.org/10.1016/j.jtbi.2011.01.051*

Macarron, R., Banks, M. N., Bojanic, D., Burns, D. J., Cirovic, D. A., Garyantes, T., Green, D. V. S., Hertzberg, R. P., Janzen, W. P., Paslay, J. W., Schopfer, U. & Sittampalam, G. S. (2011). Impact of high-throughput screening in biomedical research, *Nature Reviews Drug Discovery* 10(3): 188–195.
URL: *http://dx.doi.org/10.1038/nrd3368*

Mak, H. C. (2011). Trends in computational biology[mdash]2010, *Nature Biotechnology* 29(1): 45.
URL: *http://dx.doi.org/10.1038/nbt.1747*

Malo, N., Hanley, J. A., Cerquozzi, S., Pelletier, J. & Nadon, R. (2006). Statistical practice in high-throughput screening data analysis., *Nature biotechnology* 24(2): 167–175.
URL: *http://dx.doi.org/10.1038/nbt1186*

MAQC Consortium, Shi, L., Reid, L. H., Jones, W. D., Shippy, R., Warrington, J. A., Baker, S. C., Collins, P. J., de Longueville, F., Kawasaki, E. S., Lee, K. Y., Luo, Y., Sun, Y. A. A., Willey, J. C., Setterquist, R. A., Fischer, G. M., Tong, W., Dragan, Y. P., Dix, D. J., Frueh, F. W., Goodsaid, F. M., Herman, D., Jensen, R. V., Johnson, C. D., Lobenhofer, E. K., Puri, R. K., Schrf, U., Thierry-Mieg, J., Wang, C., Wilson, M., Wolber, P. K., Zhang, L., Amur, S., Bao, W., Barbacioru, C. C., Lucas, A. B. B., Bertholet, V., Boysen, C., Bromley, B., Brown, D., Brunner, A., Canales, R., Cao, X. M. M., Cebula, T. A., Chen, J. J., Cheng, J., Chu, T.-M. M., Chudin, E., Corson, J., Corton, J. C., Croner, L. J., Davies, C., Davison, T. S., Delenstarr, G., Deng, X., Dorris, D., Eklund, A. C. & Fan, X.-h. H. (2006). The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements., *Nature biotechnology* 24(9): 1151–1161.
URL: *http://dx.doi.org/10.1038/nbt1239*

Mayr, L. M. & Bojanic, D. (2009). Novel trends in high-throughput screening, *Current Opinion in Pharmacology* 9(5): 580–588.
URL: *http://dx.doi.org/10.1016/j.coph.2009.08.004*

Munos, B. (2009). Lessons from 60 years of pharmaceutical innovation., *Nature reviews. Drug discovery* 8(12): 959–968.
URL: *http://dx.doi.org/10.1038/nrd2961*

Nidhi, Glick, M., Davies, J. W. & Jenkins, J. L. (2006). Prediction of Biological Targets for Compounds Using Multiple-Category Bayesian Models Trained on Chemogenomics Databases, *Journal of Chemical Information and Modeling* 46(3): 1124–1133.
URL: *http://dx.doi.org/10.1021/ci060003g*

Paul, S. M., Mytelka, D. S., Dunwiddie, C. T., Persinger, C. C., Munos, B. H., Lindborg, S. R. & Schacht, A. L. (2010). How to improve R&D productivity: the pharmaceutical industry's grand challenge, *Nature Reviews Drug Discovery* 9(3): 203–214.
URL: *http://dx.doi.org/10.1038/nrd3078*

Schadt, E. E., Linderman, M. D., Sorenson, J., Lee, L. & Nolan, G. P. (2010). Computational solutions to large-scale data management and analysis, *Nature Reviews Genetics* 11(9): 647–657.
URL: *http://dx.doi.org/10.1038/nrg2857*

Searls, D. B. (2010). The Roots of Bioinformatics, *PLoS Comput Biol* 6(6): e1000809+.
URL: *http://dx.doi.org/10.1371/journal.pcbi.1000809*

Stedtfeld, R. D., Baushke, S. W., Tourlousse, D. M., Miller, S. M., Stedtfeld, T. M., Gulari, E., Tiedje, J. M. & Hashsham, S. A. (2008). Development and experimental validation of a predictive threshold cycle equation for quantification of virulence and marker genes by high-throughput nanoliter-volume PCR on the OpenArray platform., *Applied and environmental microbiology* 74(12): 3831–3838.
URL: *http://dx.doi.org/10.1128/AEM.02743-07*

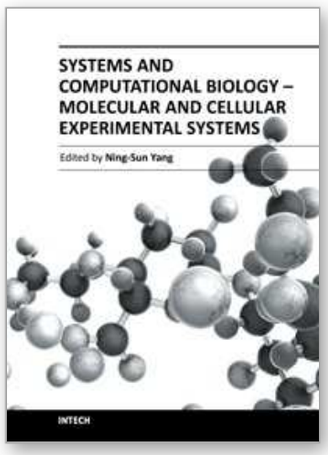
Swamidass, S. J., Bittker, J. A., Bodycombe, N. E., Ryder, S. P. & Clemons, P. A. (2010). An Economic Framework to Prioritize Confirmatory Tests after a High-Throughput Screen, *Journal of Biomolecular Screening* 15(6): 680–686.
URL: *http://dx.doi.org/10.1177/1087057110372803*

Van Hummelen, P. & Sasaki, J. (2010). State-of-the-art genomics approaches in toxicology, *Mutation Research/Reviews in Mutation Research* 705(3): 165–171.
URL: *http://dx.doi.org/10.1016/j.mrrev.2010.04.007*

Xie, J., Thapa, R., Reverdatto, S., Burz, D. S. & Shekhtman, A. (2009). Screening of Small Molecule Interactor Library by Using In-Cell NMR Spectroscopy (SMILI-NMR), *Journal of Medicinal Chemistry* 52(11): 3516–3522.
URL: *http://dx.doi.org/10.1021/jm9000743*

Yoav Shoham, K. L. (2009). *Multiagent Systems - Algorithmic, Game-Theoretic, and Logical Foundations*, Cambridge University Press.
URL: *http://www.cambridge.org/gb/knowledge/isbn/item1175725/?site_locale=en_GB*

Zanella, F., Lorens, J. B. & Link, W. (2010). High content screening: seeing is believing, *Trends in Biotechnology* 28(5): 237–245.
URL: *http://dx.doi.org/10.1016/j.tibtech.2010.02.005*



Systems and Computational Biology - Molecular and Cellular Experimental Systems

Edited by Prof. Ning-Sun Yang

ISBN 978-953-307-280-7
Hard cover, 332 pages
Publisher InTech
Published online 15, September, 2011
Published in print edition September, 2011

Whereas some â€œmicroarrayâ€ or â€œbioinformaticsâ€ scientists among us may have been criticized as doing â€œcataloging researchâ€, the majority of us believe that we are sincerely exploring new scientific and technological systems to benefit human health, human food and animal feed production, and environmental protections. Indeed, we are humbled by the complexity, extent and beauty of cross-talks in various biological systems; on the other hand, we are becoming more educated and are able to start addressing honestly and skillfully the various important issues concerning translational medicine, global agriculture, and the environment. The two volumes of this book presents a series of high-quality research or review articles in a timely fashion to this emerging research field of our scientific community.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Quin Wills (2011). High Content and Throughput Drug Discovery, Systems and Computational Biology - Molecular and Cellular Experimental Systems, Prof. Ning-Sun Yang (Ed.), ISBN: 978-953-307-280-7, InTech, Available from: http://www.intechopen.com/books/systems-and-computational-biology-molecular-and-cellular-experimental-systems/high-content-and-throughput-drug-discovery

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen