DRUG DISCOVERY TODAY BIOSILICO

Reengineering the pharmaceutical industry by crash-testing molecules

Peter W. Swaan and Sean Ekins

The recent decline in drug approvals and the increase in late-stage failures indicate that the ability to generate and screen large numbers of molecules has not improved the drug pipeline. Perhaps the pharmaceutical industry should follow the example of the automotive industry and agree upon a shared modeling language with vendors and academics to enable integration of predictive computational tools across the industry. This will then enable the virtual 'crash-testing' of drugs before synthesis, biological testing and, most importantly, clinical trials. This represents an ambitiously progressive approach using the models for simulating every stage of the drug discovery and development process. Combining the relevant computational algorithms into a grand unified model would enable prioritization of the best ideas before pursuing a discovery program, selecting a target or synthesizing a molecule. The successful application of these virtual crash-testing principles by any of its current proponents could revitalize the pharmaceutical industry so that failure is avoided.

> *'The best way to have a good idea is to have lots of ideas'*
> *– Linus Pauling (1901–1994)*

The impact of automation and high-throughput approaches on drug discovery has enabled the pharmaceutical industry to probe many ideas at the bench but surprisingly few of these eventually make it to the bedside. The resounding message over the past few years is that the discovery and subsequent development of a new therapeutic molecule involves over a decade of research and close to one billion in research dollars [1,2]. Unfortunately, even at the marketing stage there is no guarantee that the originator company will recoup its research investment. Recall scenarios owing to adverse events become evident only after the drug has been exposed to a heterogeneous patient population. Thus, the industry not only needs to increase productivity and development efficiency urgently but also implement strategies to avoid predictable failure [3].

Computer science has changed processes in every industry from product manufacturing to sales and marketing, by implementing predictions based on statistics, mathematics and risk assessment algorithms. In this context, it is of particular interest that the pharmaceutical industry has lagged far behind engineering-based businesses, such as the automobile, weapons and aircraft industries, who were the early adopters of computer-aided design (CAD). The arrival of supercomputers at pharmaceutical companies in the late 1980s heralded the era of rational drug design [3]. Over the past few decades, computational approaches have trickled down into other areas but have remained focused around medicinal chemistry, drug design and molecular biology. However, to date, computational methods have apparently had surprisingly limited impact on the design of

Peter W. Swaan
Department of
Pharmaceutical Sciences,
University of Maryland,
20 Penn Street,
Baltimore,
MD 21201, USA
email:
pswaan@rx.umaryland.edu
Sean Ekins
GeneGo,
500 Renaissance Drive,
Suite 106,
St Joseph,
MI 49085, USA
email: sean@genego.com

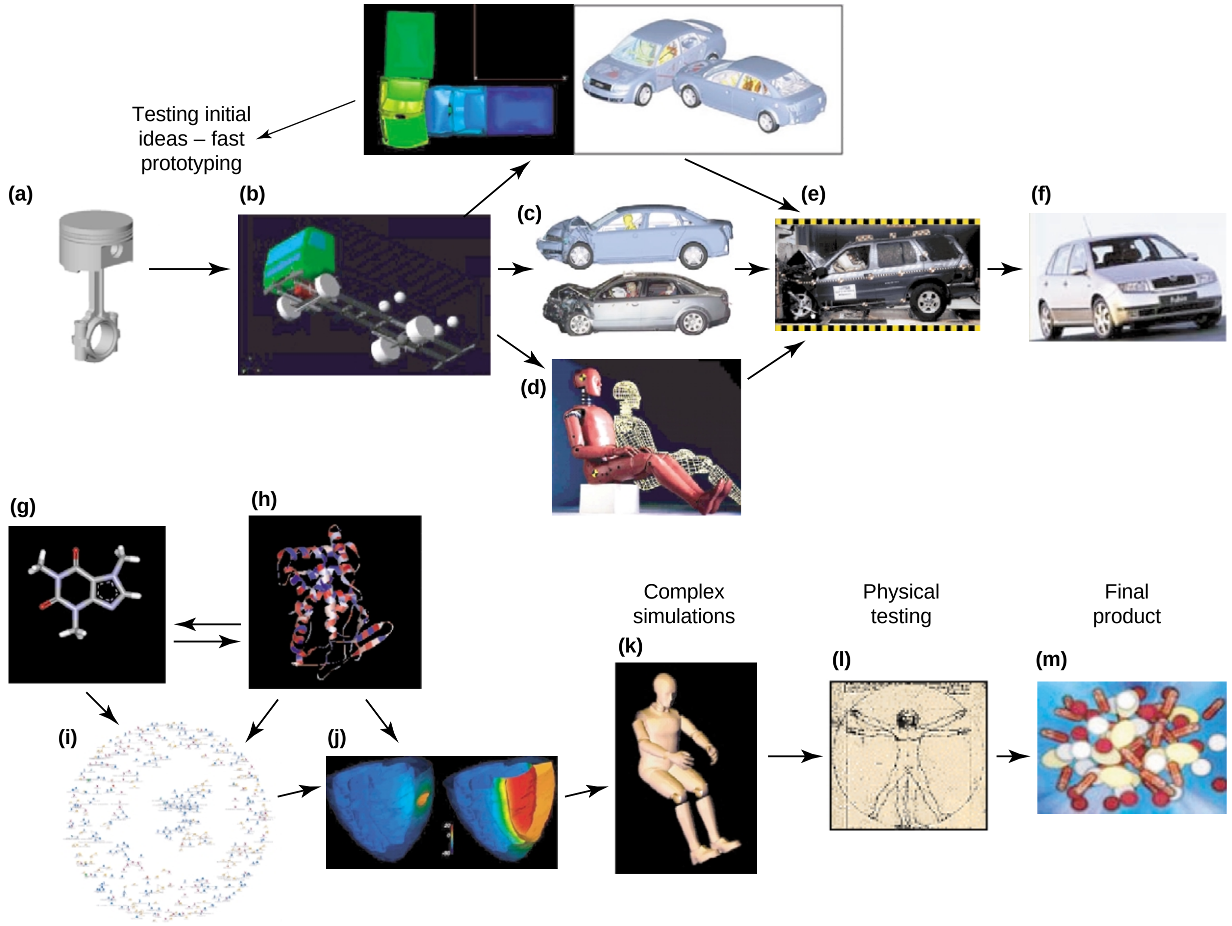
1359-6446/05/$ – see front matter ©2005 Elsevier Ltd. All rights reserved. PII: S1359-6446(05)03557-9

BOX 1

Crash testing the automotive design principle

The automobile industry has many parallels with the pharmaceutical industry in that the end products of both have to pass government-regulated safety criteria (Table 1). Vehicle safety engineering includes two areas: high-performance computing in early development of new prototypes and real-life testing at a later stage. Analogies can be drawn to *in silico* predictions for a molecule on its performance upon administration. Both crash and molecular simulations are now feasible through increased computing performance available over the past decade. The overall result for the automotive industry will be improved safety performance, fewer physical tests, reduced costs and shorter lead times when new cars are developed. The use of simulations has increased as government regulations on the industry have increased and the computational tools enable the evaluation of design ideas before analysis. Between 10 and 12 full-scale crashes can be accomplished per day, considerably shorter than a single real-life test, which can take several days.

Crash simulations use finite element analysis by dividing up the vehicle into a fine mesh of polygons, resembling a molecular van der Waals surface. Higher precision in calculating the deformation of a structure can be accomplished by a more closely-knit network. A real crash, whether with a stationary object or another vehicle, takes one-tenth of a second and is simulated in slow-motion (>100 000 stages), while calculating the deformation of each of the individual elements (e.g. chassis, engine components). Real crash tests cannot be dispensed with entirely because certification tests are required by legislation and companies need documentation in case of lawsuits. Simulation invariably describes the behavior of an idealized vehicle, whereas it is much more difficult to simulate reliably the passenger movements following a crash. Crash-test dummy movements and the injury values they record are best ascertained in actual crash tests, although simulations of crash dummies are now in use. The objective of computer-aided car crash test engineers is to find the ideal combination of the opportunities presented by simulations and real tests, to arrive at a high level of design maturity within a short development time involving a minimum of prototype tests. The pharmaceutical industry could learn much from the types of simulations that occur routinely in this industrial design process and align its processes similarly after first enabling the integration of models in different areas of the research and development pipeline.



Drug Discovery Today

FIGURE I

Simulations and crash testing to refine product design. Analogous roles of computer models to predict crash testing of cars or the effect of molecules on target proteins, organs and the complex human system. Ultimately, the goal of both research and development processes is a reliable, safe and commercially successful product. **(a)** Computer-aided component design – rapid prototyping. **(b)** Computer-aided design (CAD) of whole vehicle. **(c)** Computational crash testing of vehicle. **(d)** Simulating crash-test 'human' responses. **(e)** Physical simulation of vehicle. **(f)** Final launched vehicle. **(g)** Computed-aided molecule design. **(h)** Structure-aided design. **(i)** An interaction network for genes of interest as target validation. **(j)** Whole organ (heart) simulation. **(k)** More complex simulations, multi-organ response, disease response, PK–PD modeling, and so on. **(l)** Preclinical *in vitro*, *in vivo* and human clinical testing of the drug. **(m)** Successful launch of the drug.

TABLE 1

Analogous phases in the pharmaceutical and automotive development processes; these similarities can be exploited to improve drug discovery

Pharmaceutical industry	Automotive testing
Therapeutic area (CNS/cardiovascular/cancer)	Target market (midsize/sport/SUV/truck)
Computer-aided drug design – enzyme/target *de novo* design *in silico*	Computer-aided design (CAD) – components, engine, vehicle design and testing *in silico*
Medicinal chemistry, synthesis of molecule	Clay model, first physical model
SARs, lead optimization incorporating biological data to improve the initial design	Fast prototyping using sequential rounds of CAD
Testing for physical properties of molecules, such as permeability	Testing for physical behavior of the car, such as fluid dynamics/wind tunnel, aerodynamics
Testing the physical stability of the molecule	Testing the physical stability of the car under various conditions; steering, rollover
Undesirable interactions with other proteins, pathways evaluated during toxicity testing	Physical crash testing and continuous use of vehicle and parts to simulate product lifetime to evaluate durability
Phase I clinical trials	Test track evaluation
Phase III trial and FDA	Vehicle registration and approval
A successful molecule is a blockbuster	A well-received car is a commercial success
Post marketing problems, blackbox warning, recall, drug–drug interactions; could result in litigation	Post sales problems (e.g. rollover) might result in recall for repairs, vehicle replacement or money back; could result in litigation

marketed compounds. In stark contrast, the automobile industry has used simulations for a similar period to prospectively design and crash-test vehicles from the ground up to decrease cycle time and improve profitability (Box 1, Table 1) [4]. We could extend this analogy further to the discovery and development of a drug entirely by computer, currently a Utopian vision. Here, the complexity of the human body and inter-individual variations in physiology, diet and environment could ultimately hamper the development of a comprehensive virtual human test-bed. Thus, pharmacogenomics and pharmacogenetics add a wrinkle to the engineering-based analogies in which our genes, as well as other factors, can complicate prediction accuracy significantly. On a more positive note, the currently available and growing numbers of computational models could improve the quality of molecules developed, or at least help in reducing late stage failure owing to poor ADME–Tox properties [5–7]. Though these parameters represent merely a part of the drug discovery and development pipeline, it can be envisioned that unified algorithms combining these and other computational strategies might improve efficiency. Generally, the pharmaceutical industry has used computers and algorithms for many different facets of drug discovery and development, thereby creating a vast database of knowledge throughout the integrated steps of the research and development cycle [8]. To date, however, there has not been a concerted effort to agree upon a common modeling language for predictive computational tools or a unified output model format. These steps would facilitate a modular approach to constructing a complete computational test-bed for developing pharmaceuticals.

Selecting targets

It has been suggested that there might be 600–1500 tractable targets for small molecules in the entire human genome [9]. Identifying the targets with a justifiable likelihood for pharmaceutical success would be an obvious place to start. Prior to initiating drug discovery, a potential target should be fully evaluated computationally in terms of therapeutic area and one or a range of disease modalities. Intrinsically, target selection is not only disease driven but impacted by many other factors, such as the cost of goods, prior art, and the number and size of clinical trials required. Additionally, target selection analysis would integrate prospective financial returns (value) from present market conditions and forecast influences of changing demographics (e.g. the aging population, potential spread of resistance to antibiotics). The importance of accurate financial projections – incorporating sales potential, cost, time, risk and options [10] – before embarking on empirical pharmaceutical research cannot be stressed too highly.

Once a target has been selected, further information can be obtained using common bioinformatics tools (e.g. http://www.expasy.org). These techniques serve as a virtual HTS approach to mine and select pharmacological targets [11] with regard to gene regulation, function, structure and folding. In turn, detailed topological knowledge about the target, such as binding-site location and volume, can be translated into desirable ligand properties that demarcate molecular dimensions [12]. Furthermore, identification of closely related proteins enables specification of ligand selectivity to decrease potential toxicity owing to overlapping affinities with closely related proteins. In general, knowledge of the molecular structure of the target protein provides an ideal starting point for structure-based drug design. In the absence of high-resolution structural data, however, indirect molecular modeling approaches, such as pharmacophore development or 3D quantitative SAR (QSAR) techniques, are invaluable for initial hit discovery and screening for affinity overlap [13–15]. A recent study has applied a virtual set of molecules and uses their docking scores to gauge similarity of binding sites across multiple proteins [16]. However, the example method does not appear as effective with open solvent-accessible binding-sites, thereby potentially limiting the number of proteins to which it can be applied. There is currently no consensus on approaches for target validation, therefore, the application of such computational methods will remain an integral component alongside molecular biology, *in vitro*, *in vivo* and other methods [17,18].

In a rational approach, the ability to prioritize targets based on suitability for computer-based drug design will determine whether a project can be taken forward. Once virtual lead molecules have been identified, a selection process for synthesis based on desirable chemical, biological or other properties is necessary. This is ultimately validated by the synthesis of active molecules that can be fed back into the design process. Up until this point, it has been of most value to select only those molecules with both desirable biological activity and satisfactory ADME–Tox properties [5,19–21]. Available computational approaches to predicting these and other binding-related properties include ligand docking [22], 3D QSAR, similarity searching [23], pharmacophores or a holistic integration of all available tools [24,25]. The iterative refinement of multiple computational models with experimental data introduces smarter screening approaches that have the potential to accelerate drug discovery [26,27]. As a long-term strategy, exhaustive computational simulations for all aspects of a project would help generate a viable backup strategy that can balance the risk of using virtual insight alone or in combination with experimental data [28].

Moving targets

As mentioned earlier, structural information is not readily available for many target proteins, especially when these are membrane-embedded. Although algorithms are available to predict various protein features, detection of functional binding sites is not a straightforward process, particularly when protein–protein interactions occur (interestingly, Thornton and colleagues [29] developed a neural network algorithm to enable just these types of predictions). In addition, dynamic variations in the binding-site dimensions might allow for flexibility, thereby allowing different sized ligands and their conformational variations to bind [30]. This, in turn, severely hampers the design of rigid molecules. Newer, protein-based computational methods incorporate protein flexibility that allow for overlap between ligand and protein, conformational sampling of the side chains, or ligand docking to multiple conformations of the protein [31].

The complexities of protein-based ligand design have not discouraged the development of ligand-specificity prediction algorithms. For example, Brinkworth and co-workers [32] developed a set of rules to predict the binding of highly flexible heptapeptide substrates to kinases, enabling the prediction of an optimal substrate from merely a protein sequence. Another approach, aptly named 'inverse docking', screens specific ligands for binding to multiple proteins that are not considered primary targets but that could potentially lead to toxic side-effects [33,34]. This approach successfully identified molecules demonstrating known toxicity but also provided a high false-positive rate [34].

In summary, once a target has been selected, it alone will not guarantee that *in silico* design will succeed. However, numerous computational approaches so far show considerable promise in aiding in target validation and handling the intrinsic complexities of protein–ligand interactions.

Discovery: finding computational synergy

Traditionally, and perhaps surprisingly, protein-based molecular modeling has remained independent from QSAR type modeling. For example, computational tools for scoring of molecular docking in a protein structure [35,36] have been developed separately to those for predicting physicochemical properties of small molecules [37]. The integration of computational technologies requires concurrent optimization [21,38] of virtual properties for both ADME–Tox, target affinity and possibly other properties, to improve the quality of molecules selected for synthesis [39]. To this end, we have witnessed a trend over the past 5 years in the synergistic merger of technologies that simultaneously incorporate ligand and target flexibility [40–42].

A comparison of structure-based virtual screening methods suggests that pre-filtering (based on calculated molecular properties) before docking is desirable [22]. Lyne and colleagues have used a tiered approach for virtual screening, including property filters, a simple pharmacophore, followed by docking and scoring to discover Chk-1 kinase inhibitors [43]. Having filtered a database from ~560 000 molecules to 103 for testing, 36 were found to have activities from 110 nM to 68 μM. Others have had some success combining pharmacophore and *de novo* design to find hits from ~3000 compounds [35]. By reducing the number of virtual hits to a few hundred or less, this process is more appealing for lead discovery, especially if the majority of hits can be identified with minimal experimental effort [24]. Not only is it important to identify potential leads, it is also desirable to select only those molecules that can be synthesized readily. The chemical search space can be narrowed initially by retrosynthesis of available molecules [44] that are known to be ligands for the target. In turn, focusing on available molecules or fragments from commercial vendors via either similarity searches, scaffold-based classification [45] or bioisosteric replacement searches [46,47] creates a more diverse pool of fragments for *de novo* design that are likely to be synthesizable. The use of algorithms to predict ligand–protein binding energies reliably, based on databases of ligands with known binding energies [48], could also be useful. The real utility of these methods, such as those using geometrical descriptors and machine learning approaches like Kernel-partial least squares, will come once these have been validated suitably with experimental data and additional crystal structures. To facilitate this, the pharmaceutical industry has moved towards high-throughput X-ray crystallography of protein–ligand complexes [49]. In turn, this requires automated approaches for the determination of protein structures, including molecular replacement searching, automated ligand fitting, water placement and structure refinement of the solvated proteins [50,51]; all these approaches require the development of novel algorithms. Additionally, progress made in developing fold recognition tools enables the reliable prediction of protein structure, such that the structure determination and structure–function insights can be made earlier in the drug discovery process. In some cases, this might lead to structural information of a protein before its physiological function can be ascribed, thereby bringing this emerging field to the forefront of target discovery. This exemplifies the current breakdown of the traditional strictly linear drug discovery process and moves towards the use of concurrent multiple discovery processes. This paradigm shift, however, does require complex informatics tracking and optimal scheduling of all the processes involved to be effective.

As well as the work with scoring algorithms for docking, substructure searching studies have shown that employing multiple methods is essential because each selects different lists of molecules. This suggests that various search algorithms are required for a single project, which, ultimately, can be combined and filtered [23]. Comparing the physicochemical properties of the designed molecule to known drugs either on the market or in development might also influence the decision to ultimately synthesize a molecule or structural series [52,53]. In this context, many studies have defined the appropriate physicochemical properties of successfully marketed drugs and potential lead compounds or have derived rules to predict physicochemical properties [52,54–59]. Whether rules derived from predicted physicochemical parameters for known molecules alone can modify the drug discovery process remains to be seen but they are more likely to be a valuable addition to virtual library screening cascades.

Preclinical *in silico* toxicology

As the FDA recently recognized, toxicity testing has changed little over the decades [60] and animal studies still represent the final hurdle before human clinical studies. In addition, the breadth of the toxicology modeling problem has been underestimated. Evidently, throughput of toxicity models is limited and pharmacodynamic inter-species differences can provide an additional challenge in developing predictive models. The FDA has already indicated the emerging role of computational models in toxicology. Although many ADME models are available [5,20], as well as commercially available software to design such models [61], there are relatively few options for *in silico* toxicology. The current status of toxicology modeling includes rule-based modeling [62] and simple descriptor analysis [42,63–67]. Several ecotoxicology, acute toxicity and reproductive toxicology models have been derived generally from small datasets based around a congeneric, mostly non-drug-like series of molecules [68,69]. Drugs have been used to generate a QSAR for maximum recommended therapeutic dose, which can be used as a mechanism to predict the toxic dose threshold in humans [70]. Furthermore, toxicology data are rarely integrated with other physicochemical properties, let alone stored in a central repository. One exception is a recent study by Duart and co-workers who used molecular topology descriptors with lipophilicity (logP), pharmacokinetic (T_{max}) and toxicity data (LD_{50}) to develop a model for a series of antihistamines [71]. Key initiatives such as DSSTox (http://www.epa.gov/nheerl/dsstox/) [72] and the Chemical Effects in Biological Systems (CEBS) knowledgebase [73] represent significant government funded efforts to centralize *in silico* toxicity models and high content data, respectively, which might enhance future computational modeling efforts.

Immunotoxicology is another relatively underdeveloped area of high significance. The complexity of effectively modeling therapeutic proteins and antisense molecules is extremely challenging. Therefore, it might be considered advantageous to look to the area of vaccine development for insights, where bioinformatics and computational chemistry have been applied to epitope and protein structure prediction, respectively [74]. Protein surface features are important determinants for protein–protein interactions. For example, the scorpion toxins BeKm-1, BmTx3 and CnErg1 are inhibitors of the hERG potassium channel [75–77]. The protein exterior of these hERG inhibitors possess key pharmacophoric features dominated by hydrophobic areas, previously identified in small molecule drugs [76,78,79]. In contrast to small-molecule inhibitors these toxins are likely to bind in a different location outside of the channel pore. Therefore, we can envision that basic research aiming to record protein surface features computationally, in correlation with protein–protein interactions, will lead to a database that can be used to predict various toxicities of new protein-based drugs.

Beyond computational drug discovery

At present we are faced with simultaneously growing computational fields, such as computational molecular biology [80], the modeling of genetic and biochemical networks [81], which covers aspects from alignment of sequences, modeling activity of genes, gene expression, cell-cycle regulation and proteomics, amongst others. One important result achieved in this area is the realization that biological networks of different origins (e.g. metabolic, regulatory, protein interactions, networks for different organisms) share the same global architecture [82,83]. An early attempt to illustrate the many levels of relationships between genetics and physiology was made by Palsson [84], who captured and linked process databases from genes to proteins, to whole cells. Although biological systems contain many non-linear processes that are continually interacting (Box 1), a reductionist viewpoint is to treat parallel systems as an engineering process [85]. Several companies, such as Gene Network Sciences [86], Entelos [87] and BioSeek [88], have emerged in recent years and focus on simulating cellular pathways, organs or whole cells. Large, curated interaction databases, combined with powerful analytical and network building tools, are available from companies like GeneGo (MetaCore™), Ariadne (Pathway Assist™), Ingenuity (Pathways Analysis™) and Jubilant (PathArt™), which cover human metabolism, regulation and signaling. These tools can enable visualization of global cellular mechanisms driving differences in gene expression to discover relationships in such complex data. To date, these approaches have been applied to: modeling nuclear hormone interactions [89], the generation of compound related gene network signatures [90], understanding the pathways affected by the tumor suppressor DBC2 [91] and combining networks with metabolite prediction tools [92]. These systems biology methods would have clear value in drug discovery when combined with the other computational and empirical approaches described previously to identify biomarkers and to understand inter-individual variability in response to drugs [21,93].

Form(ulation) and function

The development of a drug is strongly influenced by its formulation. Capturing the complexities of this process for individual molecules could enable the introduction of algorithms that predict ‘overall developability’ from structure alone. The field of quantitative structure–property relationships (QSPR) has successfully modeled physical properties using multiple different algorithms and descriptors [94]. Although QSPR studies can be applied conceivably to predict many molecular properties, there are ominous gaps in such applications. For example, there have been relatively few attempts to model the melting point of a molecule even though large databases have been collated with this information [95–97].

An example of formulation-based prediction is the optimization of controlled-release dosage forms using a neural network, which also simulated *in vivo* plasma concentrations [98]. Amongst the input variables for 22 different formulations were tableting factors such as moisture, particle size and hardness. Correlations of input to output parameters, such as *in vitro* dissolution time profiles, were used to simulate *in vivo* plasma concentration profiles. Assuming a direct relationship between *in vitro* dissolution and bioavailability, this method can therefore be used to predict the *in vivo* characteristics of a particular formulation based on its physicochemical parameters. This assumption might prove valid for molecules absorbed via passive diffusion but it is likely to be problematic for molecules that cross the intestinal tract via protein-mediated interactions.

Other parameters that are of particular importance to pharmaceutical technology and preformulation that have been simulated successfully are: water uptake profile, glass transition temperature [99,100], viscosity [101,102], crystalline polymorphic forms [103,104] and plasticization efficiency of pharmaceutical coating formulations [105]. Additional predictive computational models are essential for an accurate forecast of the physical aging of tablet coatings, as well as emulsion stability or charge.

A leap forward: virtual forecasting of clinical studies

The ultimate cost-saving in drug development lies in the accurate forecasting of human pharmacokinetic data. Early computational models for human clinical data used simple molecular descriptors to predict oral bioavailability and volume of distribution [106,107]. Neural networks have been applied to predict pharmacokinetic properties from simple physicochemical properties, such as lipophilicity, pK_a and the fraction of free drug in plasma [108]. As a result of the small sample size inherent to clinical studies, a general issue with models generated from human data is the lack of, or small size of, control groups (test sets). Comparisons of algorithms for pharmacokinetic predictions have shown the equivalence of the limited sampling model and the Bayesian approach to modeling AUC and C_{max} [109].

The recent ability to simulate populations of patients with a disease, and their controls, using integrated differential equations enables the prediction of response to a particular type of therapy based on relatively few input parameters [110–112]. One specific example created a mathematical model called Archimedes, which incorporated >50 independent variables, such as patient anatomy, pathophysiology, tests, treatments and outcomes related to diabetes [113,114]. Such an approach could be useful for designing and simulating actual clinical studies and has yet to be embraced broadly by the pharmaceutical industry. However, as part of a total computational strategy for drug discovery and development it would represent a key component before investing in costly clinical trials [2,115]. During the planning phase of clinical trials, the statistical power of a study is often estimated beforehand and computational algorithms are under development that can provide accurate predictions [116]. An area that is currently also in need of automated decision-making routines is post-marketing clinical trials. Here, data are generated on safety, cost effectiveness and efficacy, which are ultimately used to improve labeling and suggest new indications. Although post-marketing results are analyzed by standard protocols, prospective forecasting analysis could aid in the eventual decision-making process [117]. One example of such an advanced approach has optimized patient gains, cost and future cost for Phase II oncology trials. This enabled inferior treatments to be dropped early in the clinical trial process [118].

The potential for avoiding drug–drug interactions with patients on multi-drug therapy is also possible using computer-based management systems to check orders in the pharmacy. Studies have shown that the reliability reaches 100%, which is of high importance for drugs with a narrow therapeutic range [119]. Currently, the FDA stores and maintains a large drug–drug interaction database; mining this database with correlations to chemical structure could be used to derive new robust algorithms that could relate structure to adverse reactions more readily. To date this has not been widely attempted.

Crash testing molecules and potential roadblocks

Ultimately, algorithms can be used to adaptively design lead molecules faster by testing them *in silico*. Numerous companies use computational technologies as a small part of either the drug discovery or drug development process. Few have integrated a complete portfolio of these technologies. The reasons for this can be explained, in part, by the resistance to adopt computational strategies at each individual level of the process. Therefore, at present true innovation in applying virtual screens for drugs might be expected in smaller, more specialized biotechnology companies that are less risk averse and already have higher success rates [2,115]. Surprisingly, at present no pharmaceutical company appears to be crash-testing molecules from discovery through development entirely using computational approaches in the same way that engineering-based companies are developing products. This would include using not only technologies for optimizing small molecule but also those algorithms that model complete biological systems and other processes described earlier. Although drug discovery tools are generally limited to the optimization of a single property, the development of more holistic systems-based models using multiple integrated software tools has also not occurred to date. Perhaps a reason for this is that all modeling technologies used in the pharmaceutical industry have a unique ‘language’ and model or result output format, depending on the vendor and software type. This ultimately prevents sharing of models between groups after publication or even their integration into meta-models within companies. However, a trend can be detected in the recent combination of computational approaches from distantly related areas, for example, small molecule QSAR and protein–ligand docking. In the current competitive environment, smaller drug companies will need to rely more heavily on computational approaches, which will minimize HTS and combinatorial chemistry, enabling researchers to crash-test their ideas and focus on exercising the best ones. The pharmaceutical industry can learn further from other defense and automotive engineering-based industries, which worked on open standards for software and models to enable integration. This was facilitated by ESCHER (http://www.escherinstitute.org/), a consortium of software vendors and industries funded by the US government, resulting in an open source software repository. A recent Defense Advanced Research Projects Agency (DARPA) sponsored workshop (Tool and Software Infrastructure in Systems Biology Workshop, Arlington VA USA, 17–18 February 2005) proposed the development of a similar independent organization to set up open standards for systems biology. Taking this a few steps further, perhaps all predictive computational methods used by the pharmaceutical industry could be incorporated simultaneously into this future initiative to foster the overall integration of the methods and models. At the same time, a more modular approach to model development would ensure that each drug company would not repeat the same software failures as its peers but instead truly advance the industry. The involvement of DARPA could indicate that computational models for the prediction of human health have strategic importance.

Conclusion

It is crucial that the pharmaceutical industry reengineers its process of discovering and evaluating molecules from target discovery through clinical trials and beyond as has been suggested here. In the post-genomic age, providing integrated computational biology models of whole cells, organs and disease states has aroused a great deal of interest, bringing us closer to simulating diverse and complex stages of drug discovery and development simultaneously. The combination of computational models for desired chemical, biological, physiological and economical properties earlier can all be used to filter virtual molecules. This could represent a limited view but in this case will enable the simulation of many steps in the drug discovery pipeline to ultimately improve research and development success. By integrating many of the computational approaches described over the past decade, we foresee the beginning of the total *in silico* design and testing of molecules within the next decade. This will be greatly facilitated by a common software language and the sharing of models, which will need to be sponsored by government agencies. It might be indicative of the direction the industry is taking today that numerous patents on computational models and technologies have already been filed [120]. However, this could be counter to what is required for rapid development and deployment of these technologies. To direct the drug discovery process away from a trial-and-error, brute force synthesis and HTS paradigm towards a more rationally driven design approach, today's chemists will need to take more direction from computational scientists. However, there appears to be a natural resistance and skepticism to *in silico* technologies. This is despite the mounting evidence that, for example, the computational chemistry approaches have a throughput greater than *in vitro* methods and, therefore, already have considerable value. Drug discovery requires the active collaboration of those skilled in computational sciences if we are to focus not only our *in vitro* and *in vivo* resources but also those at later discovery and development stages. The tools for reengineering the pharmaceutical industry are likely to be complex but they are also within our reach; however, are we ready to embrace them? Ultimately, computational approaches will enable us to ensure the best ideas result in bringing safer and better designed drugs to the waiting patient more rapidly.

Acknowledgements

S.E. gratefully acknowledges Maggie A.Z. Hupcey for valuable discussions. P.W.S. is funded by grants from the National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases (DK56631 and DK061425) Images reproduced with kind permission and assistance from Olivier Morisot (ESI North America), Moustafa El-Gindy (Pennsylvania State University), Ronald Barr (University of Texas) and Raimond Winslow (Johns Hopkins University School of Medicine).

References

1 Bain, W. (2004) Failure rates in drug discovery and development: will we ever get any better? *Drug Discov. World* Fall, 9-18

2 Kola, I. and Landis, J. (2004) Can the pharmaceutical industry reduce attrition rates. *Nat. Rev. Drug Discov.* 3, 711–715

3 Pramik, M.J. (1989) Molecular modeling ushers in the age of rational biotech drug design. *Genet. Eng News* June 11, 38

4 Kaufmann, W.J.I. and Smarr, L.L. (1993) *Supercomputing and the Transformation of Science*, Scientific American Library

5 van de Waterbeemd, H. and Gifford, E. (2003) ADMET *in silico* modelling towards prediction paradise? *Nat. Rev. Drug Discov.* 2, 192–204

6 Beresford, A.P. *et al.* (2002) The emerging importance of predictive ADME simulation in drug discovery. *Drug Discov. Today* 7, 109–116

7 Butina, D. *et al.* (2002) Predicting ADME properties *in silico*: methods and models. *Drug Discov. Today* 7 (Suppl.), S83–S88

8 Weintraub, H.J.R. (2003) The race to integrate. *Curr. Drug Discov.* Feb., 23–27

9 Hopkins, A.L. and Groom, C.R. (2002) The druggable genome. *Nat. Rev. Drug Discov.* 1, 727–730

10 Greuel, J.M. (2002) The R&D value conundrum. *Curr. Drug Discov.* July, 37–42

11 Terstappen, G.C. and Reggiani, A. (2001) *In silico* research in drug discovery. *Trends Pharmacol. Sci.* 22, 23–26

12 Schneider, M. (2004) A rational approach to maximize success rate in target discovery. *Arch. Pharm. (Weinheim)* 337, 625–633

13 Ekins, S. (2004) Predicting undesirable drug interactions with promiscuous proteins *in silico*. *Drug Discov. Today* 9, 276–285

14 Ekins, S. and Swaan, P.W. (2004) Computational models for enzymes, transporters, channels and receptors relevant to ADME/TOX. *Rev. Comp. Chem.* 20, 333–415

15 Ekins, S. *et al.* (2004) Applying computational and *in vitro* approaches to lead selection. In *Pharmaceutical Profiling in Drug Discovery for Lead Selection* (Borchardt, R.T. *et al.*, eds), pp. 361–389, AAPS Press

16 Yoon, S. *et al.* (2005) Computational identification of proteins for selectivity assays. *Proteins* 59, 434–443

17 Whittaker, P.A. (2003) What is the relevance of bioinformatics to pharmacology? *Trends Pharmacol. Sci.* 24, 434–439

18 Kopec, K.K. *et al.* (2005) Target identification and validation in drug discovery: the role of proteomics. *Biochem. Pharmacol.* 69, 1133–1139

19 Ekins, S. *et al.* (2000) Predicting drug-drug interactions *in silico* using pharmacophores: a paradigm for the next millennium. In *Pharmacophore Perception, Development and Use in Drug Design* (Guner, O.F., ed.), pp. 269–299, IUL

20 Ekins, S. *et al.* (2000) Progress in predicting human ADME parameters *in silico*. *J. Pharmacol. Toxicol. Methods* 44, 251–272

21 Ekins, S. *et al.* (2002) Towards a new age of virtual ADME/TOX and multidimensional drug discovery. *J. Comput. Aided Mol. Des.* 16, 381–401

22 Lyne, P.D. (2002) Structure-based virtual screening: an overview. *Drug Discov. Today* 7, 1047–1055

23 Sheridan, R.P. and Kearsley, S.K. (2002) Why do we need so many chemical simialrity search methods? *Drug Discov. Today* 7, 903–911

24 Bajorath, J. (2002) Integration of virtula and high-throughput screening. *Nat. Rev. Drug Discov.* 1, 882–894

25 Klebe, G. (2002) Innovative lead discovery: from geometry to function and ligand design. *Curr. Drug Discov.* July, 27–30

26 Engels, M.F.M. and Venkatarangan, P. (2001) Smart screening: approaches to efficient HTS. *Curr. Opin. Drug Discov. Dev.* 4, 275–283

27 Manly, C.J. *et al.* (2001) The impact of informatics and computational chemistry on synthesis and screening. *Drug Discov. Today* 6, 1101–1110

28 Kennedy, T. (1997) Managing the drug discovery / development interface. *Drug Discov. Today* 2, 436–444

29 Thornton, J.M. *et al.* (2000) From structure to function: approaches and limitations. *Nat. Struct. Biol.* 7 (Suppl.), 991–994

30 Ma, B. *et al.* (2002) Multiple diverse ligands binding at a single protein site: A matter of pre-existing populations. *Protein Sci.* 11, 184–197

31 Carlson, H.A. and McCammon, J.A. (2000) Accomodating protein flexibility in computational drug design. *Mol. Pharmacol.* 57, 213–218

32 Brinkworth, R.I. *et al.* (2003) Structural basis and prediction of substrate specificity in protein serine/threonine kinases. *Proc. Natl. Acad. Sci. U. S. A.* 100, 74–79

33 Rockey, W.M. and Elcock, A.H. (2002) Progress toward virtual screening for drug side effects. *Proteins* 48, 664–671

34 Chen, Y.Z. and Ung, C.Y. (2001) Prediction of potential toxicity and side effect protein targets of a small molecule by a ligand-protein inverse docking approach. *J. Mol. Graph. Model.* 20, 199–218

35 Schneider, G. and Bohm, H-J. (2002) Virtual screening and fast automated docking methods. *Drug Discov. Today* 7, 64–70

36 Kubinyi, H. (1998) Combinatorial and computational approaches in structure-based drug design. *Drug Discov. Dev.* 1, 16–27

37 Shimada, J. *et al.* (2002) Integrating computer-based *de novo* drug design and multidimensional filtering for desirable drugs. *Targets* 1, 196–205

38 Egan, W.J. *et al.* (2000) Prediction of drug absorption using multivariate statistics. *J. Med. Chem.* 43, 3867–3877

39 Cheng, A. *et al.* (2002) Computation of the physio-chemical properties and data mining of large molecular libraries. *J. Comput. Chem.* 23, 172–183

40 Matter, H. *et al.* (2002) Design and quantitative structure-activity relationship of 3-amidinobenzyl-1H-indole-2-carboxamides as potent, nonchiral, and selective inhibitors of blood coagulation factor Xa. *J. Med. Chem.* 45, 2749–2769

41 Tikhonova, I.G. *et al.* (2003) CoMFA and homology-based models of the glycine binding site of N-methyl-d-aspartate receptor. *J. Med. Chem.* 46, 1609–1616

42 Sharma, V. and Duffel, M.W. (2002) Comparative molecular field analysis of substrates for an aryl sulfotransferase based on catalytic mechanism and protein homology modeling. *J. Med. Chem.* 45, 5514–5522

43 Lyne, P.D. *et al.* (2004) Identification of compounds with nanomolar binding affinity for checkpoint kinase-1 using knowledge-based virtual screening. *J. Med. Chem.* 47, 1962–1968

44 Lewell, X.Q. *et al.* (1998) RECAP-retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying priveleged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comp. Sci.* 38, 511–522

45 Xu, J. (2002) A new approach to finding natural chemical structure classes. *J. Med. Chem.* 45, 5311–5320

46 Patani, G.A. and LaVoie, E.J. (1996) Bioisosterism: A rational approach in drug design. *Chem. Rev.* 96, 3147–3176

47 Ertl, P. (2003) Cheminformatics analysis of organic substituents: identification of the most common substituents, calculation of substituent

properties, and automatic identification of drug-like properties. *J. Chem. Inf. Comp. Sci.* 43, 374–380

48 Andrews, P.R. *et al.* (1984) Functional group contributions to drug-receptor interactions. *J. Med. Chem.* 27, 1648–1657

49 Blundell, T.L. *et al.* (2002) High-throughput crystallography for lead discovery in drug design. *Nat. Rev. Drug Discov.* 1, 45–54

50 Mochalkin, I. *et al.* (2003) *Structure-Based Drug Discovery in Informatics Environments,* Accelrys

51 Adams, P.D. *et al.* (2004) Recent developments in the PHENIX software for automated crystallographic structure determination. *J. Synchrotron Radiat.* 11, 53–55

52 Lipinski, C.A. *et al.* (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* 23, 3–25

53 Navia, M.A. and Chaturvedi, P.R. (1996) Design principles for orally bioavailable drugs. *Drug Discov. Today* 1, 179–189

54 Veber, D.F. *et al.* (2002) Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* 45, 2615–2623

55 Blake, J. (2003) Examination of the computed molecular properties of compounds selected for clinical development. *Biotechniques* (June Suppl.) 16–20

56 Oprea, T.I. *et al.* (2001) Is there a difference between leads and drugs? A historical perspective. *J. Chem. Inf. Comput. Sci.* 41, 1308–1315

57 Walters, W.P. and Murcko, M.A. (2002) Prediction of 'drug-likeness'. *Adv. Drug Deliv. Rev.* 54, 255–271

58 Wenlock, M.C. *et al.* (2003) A comparison of physicochemical property profiles of development and marketed oral drugs. *J. Med. Chem.* 46, 1250–1256

59 Vieth, M. *et al.* (2004) Characteristic physical properties and structural fragments of marketed oral drugs. *J. Med. Chem.* 47, 224–232

60 US *Food and Drug Administration* (2004) Innovation stagnation: challenge and opportunity on the critical path to new medicinal products (http://www.fda.gov/oc/initiatives/criticalpath/whitepaper.html).

61 Boobis, A. *et al.* (2002) *In silico* prediction of ADME and pharmacokinetics: Report of an expert meeting organised by COST B15. *Eur. J. Pharm. Sci.* 17, 183–193

62 Greene, N. (2002) Computer systems for the prediction of toxicity: an update. *Adv. Drug Deliv. Rev.* 54, 417–431

63 King, L.A. (1985) Ferguson's principle and the prediction of fatal drug levels in blood. *Hum. Toxicol.* 4, 273–278

64 Tichy, M. (1991) QSAR approach to estimation of the distribution of xenobiotics and the target organ in the body. *Drug Metabol. Drug Interact.* 9, 191–200

65 Mekenyan, O. *et al.* (1993) Bronchospasmolytic activity and toxicity modeling of theophylline derivatives by a microcomputer based method. *Arzneim. Forsch./Drug. Res.* 43, 1341–1350

66 Blower, P. *et al.* (2002) On combining recursive partitioning and simulated annealing to detect groups of biologically active compounds. *J. Chem. Inf. Comput. Sci.* 42, 393–404

67 Young, S.S. *et al.* (2002) Mixture deconvolution and analysis of Ames mutagenicity data. *Chemomet. Intell. Lab. Syst.* 60, 5–11

68 Espinosa, G. *et al.* (2002) An integrated SOM-fuzzy ARTMAP neural system for the evaluation of toxicity. *J. Chem. Inf. Comput. Sci.* 42, 343–359

69 Giampaolo, C. *et al.* (1991) Predicting chemically induced duodenal ulcer and adrenal necrosis with classification trees. *Proc. Natl. Acad. Sci. U. S. A.* 88, 6298–6302

70 Contrera, J.F. *et al.* (2003) Predicting the carcinogenic potential of pharmaceuticals in rodents using molecular structural similarity and E-state indices. *Regul. Toxicol. Pharmacol.* 38, 243–259

71 Duart, M.J. *et al.* (2002) Use of molecular topology for the prediction of physico-chemical, pharmacokinetic and toxicological properties of a group of antihistaminic drugs. *Int. J. Pharmaceut.* 246, 111–119

72 Richard, A.M. and Williams, C.R. (2002) Distributed structure-searchable toxicity (DSSTox) public database network: a proposal. *Mutat. Res.* 499, 27–52

73 Waters, M. *et al.* (2003) Systems toxicology and the Chemical Effects in Biological Systems (CEBS) knowledge base. *EHP Toxicogenomics* 111(*1T*), 15–28

74 Flower, D. (2003) Towards *in silico* prediction of immunogenic epitopes. *Trends Immunol.* 24, 667–674

75 Korolkova, Y.V. *et al.* (2002) New binding site on common molecular scaffold provides HERG channel specificity of scorpion toxin BeKm-1. *J. Biol. Chem.* 277, 43104–43109

76 Torres, A. *et al.* (2003) Solution structure of CnErg1 (Ergtoxin), a HERG specific scorpion toxin. *FEBS Lett.* 539, 138–142

77 Huys, I. *et al.* (2004) BmTx3, a scorpion toxin with two putative functional faces separately active on A-type K+ and HERG currents. *Biochem. J.* 378, 745–752

78 Ekins, S. *et al.* (2002) Three-dimensional quantitative structure activity relationship for the inhibition of the hERG (human ether-a-gogo related gene) potassium channel. *J. Pharmacol. Exp. Therapeut.* 301, 427–434

79 Cavalli, A. *et al.* (2002) Toward a pharmacophore for drugs inducing the long QT syndrome: insights from a CoMFA study of HERG K+ channel blockers. *J. Med. Chem.* 45, 3844–3853

80 Pevzner, P.A. (2000) *Computational Molecular Biology*, MIT Press

81 Bower, J.M. and Bolouri, H., eds (2001) *Computational Modeling of Genetic and Biochemical Networks*, MIT Press

82 Barabasi, A-L. and Oltvai, Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5, 101–113

83 Giot, L. *et al.* (2003) A protein interaction map of *Drosophila melanogaster. Science* 302, 1727–1736

84 Palsson, B.O. (1997) What lies beyond bioinformatics? *Nat. Biotechnol.* 15, 3–4

85 Kitano, H. (2002) Computational systems biology. *Nature* 420, 206–210

86 Christopher, R. *et al.* (2004) Data-driven computer simulation of human cancer cell. *Ann. N. Y. Acad. Sci.* 1020, 132–153

87 Defranoux, N.A. *et al.* (2005) *In silico* modeling and simulation of bone biology: a proposal. *J. Bone Miner. Res.* 20, 1079–1084

88 Plavec, I. *et al.* (2004) Method for analyzing signaling networks in complex cellular systems. *Proc. Natl. Acad. Sci. U. S. A.* 101, 1223–1228

89 Ekins, S. *et al.* (2005) A Novel Method for Visualizing Nuclear Hormone Receptor Networks Relevant to Drug Metabolism. *Drug Metab. Dispos.* 33, 474–481

90 Nikolsky, Y. *et al.* (2005) A novel method for generation of signature networks as biomarkers from complex high throughput data. *Toxicol. Lett.* 158, 20–29

91 Siripurapu, V. *et al.* (2005) DBC2 significantly influences cell-cycle, apoptosis, cytoskeleton and membrane-trafficking pathways. *J. Mol. Biol.* 346, 83–89

92 Ekins, S. *et al.* (2005) Techniques: Application of Systems Biology to Absorption, Distribution, Metabolism, Excretion and Toxicity. *Trends Pharmacol. Sci.* 26, 202–209

93 Ekins, S. *et al.* (2005) Systems biology: applications in drug discovery. In *Drug Discovery Handbook* (Gad, S., ed.), pp. 123–183, Wiley

94 Wanchana, S. *et al.* (2002) Quantitative structure/property relationship analysis on aqueous solubility using genetic algorithm-combined partial least squares method. *Pharmazie* 57, 127–129

95 Young, S.S. *et al.* (2002) So many targets, so many compounds, but so few resources. *Curr. Drug Disc.* December, 17–22

96 Dearden, J.C. (2003) Quantitative structure-property relationships for prediction of boiling point, vapor pressure, and melting point. *Environ. Toxicol. Chem.* 22, 1696–1709

97 Dearden, J.C. (1991) The QSAR prediction of melting point, a property of environmental relevance. *Sci. Total Environ.* 109-110, 59–68

98 Chen, Y. *et al.* (1999) the application of an artificial neural network and pharmacokinetic simulations in the design of controlled-release dosage forms. *J. Control. Release* 59, 33–41

99 Mattioni, B.E. and Jurs, P.C. (2002) Prediction of glass transition temperatures from monomer and repeat unit structure using computational neural networks. *J. Chem. Inf. Comput. Sci.* 42, 232–240

100 Bourquin, J. *et al.* (1998) Advantages of artificial neural networks (ANNs) as alternative modelling technique for data sets showing non-linear relationships using data from a galenical study on a solid dosage form. *Eur. J. Pharm. Sci.* 7, 5–16

101 Ebube, N.K. *et al.* (2000) Preformulation studies and characterization of the physicochemical properties of amorphous polymers using artificial neural networks. *Int. J. Pharmaceut.* 196, 27–35

102 Van den Mooter, G. *et al.* (1999) Stability prediction of amorphous benzodiazepines by calculation of the mean relaxation time constant using the Williams-Watts decay function. *Eur. J. Pharm. Biopharm.* 48, 43–48

103 Almarsson, O. and Gardner, C.R. (2003) Novel approaches to issues of developability. *Curr. Drug Discov.* Jan, 21–26

104 Morissette, S.L. *et al.* (2003) Elucidation of crystal form diversity of the HIV protease inhibitor ritonavir by high-throughput crystallization. *Proc. Natl. Acad. Sci. U. S. A.* 100, 2180–2184

105 Tarvainen, M. *et al.* (2001) Predicting the plasticization efficiency from the three-dimensional molecular structure of a polymer plasticizer. *Pharm. Res.* 18, 1760–1766

106 Hirono, S. *et al.* (1994) Non-congeneric structure-pharmacokinetic property correlation studies using fuzzy adaptive least-squares: oral bioavailability. *Biol. Pharm. Bull.* 17, 306–309

107 Hirono, S. *et al.* (1994) Non-congeneric structure-pharmacokinetic property correlation studies using fuzzy adaptive least squares:

volume of distribution. *Biol. Pharm. Bull.* 17, 686–690

108 Gobburu, J.V.S. and Shelver, W.H. (1995) Quantitative structure-pharmacokinetic relationships (QSPR) of beta blockers derived using neural networks. *J. Pharm. Sci.* 84, 862–865

109 Mahmoud, I. and Miller, R. (1999) Comparison of the Bayesian approach and a limited sampling model for the estimation of AUC and Cmax: a computer simulation analysis. *Int. J. Clin. Pharmacol. Therapeut.* 37, 439–445

110 Michelson, S. (2003) Assessing the impact of predictive biosimulation on drug discovery and development. *J. Bioinform. Comput. Biol.* 1, 169–177

111 Kansal, A.R. (2004) Modeling approaches to type 2 diabetes. *Diabetes Technol. Ther.* 6, 39–47

112 Musante, C.J. *et al.* (2002) Small- and large-scale biosimulation applied to drug discovery and development. *Drug Discov. Today* 7 (Suppl. 20), S192–S196

113 Eddy, D.M. and Schlessinger, L. (2003) Validation of the archimedes diabetes model. *Diabetes Care* 26, 3102–3110

114 Eddy, D.M. and Schlessinger, L. (2003) Archimedes: a trial-validated model of diabetes. *Diabetes Care* 26, 3093–3101

115 Bain, W. (2004) Failure rates in drug discovery and development: will we ever get any better? *Drug Discov. World,* 9–18

116 Natarajan, R. *et al.* (1996) A computer program for sample size and power calculations in the design of multi-arm and factorial clinical trials with survival time endpoints. *Comput. Meth. Prog. Biomed.* 49, 137–147

117 Toscani, M.R. and Resnick, G. (1992) Postmarketing studies: methods for implemetation and potential use of data. *Drug Inf. J.* 26, 261–265

118 Stallard, N. and Thall, P.F. (2001) Decision-theoretic designs for pre-Phase II screening trials in oncology. *Biometrics* 57, 1089–1095

119 Kuhlmann, J. and Muck, W. (2001) Clinical-Pharmacological strategies to assess drug interaction potential during drug development. *Drug Saf.* 24, 715–725

120 Pieraccioli, D. (2002) Patenting the pharmacophore. *Curr. Drug Discov.* Oct, 40–43