**NEW HORIZONS IN TOXICITY PREDICTION. LHASA LIMITED SYMPOSIUM EVENT IN COLLABORATION WITH THE UNIVERSITY OF CAMBRIDGE**

*A report by Wendy A. Warr, wendy@warr.com, http://www.warr.com*

## Introduction

David Hawkins, Lhasa Limited; Robert Glen, University of Cambridge

Toxicology is a multidisciplinary science that examines the adverse effects of chemicals on organisms. It is a rapidly developing area with many new scientists entering the field. There is a shift from primarily *in vivo* animal studies to *in vitro* assays, *in vivo* assays with lower organisms, and computational modeling for toxicity assessments.[1] The conference explored various current approaches to toxicity prediction, covering and comparing the tools and methods available today, uses by regulators, industry and academia, and a look at emerging areas and technologies.

## Current Approaches for Toxicity Prediction

### Pharmaceutical perspective

Edwin Matthews, FDA

The goal of the Food and Drug Administration (FDA) Center for Drug Evaluation and Research (CDER) ComTox program is to be able to predict accurately chemical toxicities with *in silico* software for all toxicological and clinical effect endpoints of interest to the U.S. FDA. A benefit would be substantially to reduce, replace and refine the need for animal toxicological testing in establishing the safety of chemical substances. The Informatics and Computational Safety Analysis Staff (ICSAS), part of CDER's Office of Pharmaceutical Science, is facilitating an orderly transition to a new *in silico* testing paradigm. This is being articulated in parallel to the current Organization for Economic Cooperation and Development (OECD) and European Union QSAR efforts, but there are substantial strategic differences in these approaches, e.g., ICSAS employs commercial software products (which are excluded from the EU effort); freeware is only used for special applications. There is a commitment to global QSAR and expert systems. Commitment to global QSAR means that new models must be added all the time.

Three software platforms are already validated and two additional platforms are being validated. The programs are Derek for Windows and Meteor; Leadscope FDA Model Applier and Predictive Data Miner; MCASE/MC4PC and META; Prous BioEpisteme and Integrity; and QSARIS (formerly MDL-QSAR, now from Scimatics).[2] Insilicofirst (founding members Lhasa, Leadscope, MCASE and Molecular Networks)[3] is a collaborative endeavor working to develop a computational prediction system to support the environmental safety assessment of chemicals.

The FDA has several reasons for using more than one QSAR software program. None of the programs has all the necessary functionalities, and none has 100% coverage, sensitivity, and specificity. All of the programs are complementary and can be used for consensus prediction strategies. Moreover, FDA cannot endorse a single (Q)SAR program. Components of the multiple platform strategy are predicted or experimental value; bioavailability; structural analogues; coverage (i.e., domain of applicability); metabolites; weight of evidence predictions (combining predictions from multiple QSAR programs); and mechanism of action.

Matthews listed some unmet needs in QSARs and expert systems. These include integrated fragment and descriptor paradigms and 3D descriptors; QSARs based upon pure active ingredient (PAI) and metabolites; QSARs for drug-drug interaction, for animal organ toxicities, and for regulatory dose concentration endpoints (e.g., lowest observed effect level (LOEL) and no observed effect level (NOEL)); and expert system rules for toxicities of substances such as biologicals which cannot be predicted by QSAR. Other unmet needs are databases of pharmaceutical off-target activities, of pharmaceutical Investigational New Drug (IND), of confidential business information and of regulatory dose concentration endpoints; integration of FDA and Environmental Protection Agency (EPA) archival data; and advanced linguistic software to extract data.

### ***In silico*** **tools and guidance developed by the Joint Research Centre**

Andrew Worth, European Commission Joint Research Centre (JRC)

Under the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) regulation information on intrinsic properties of substances may be generated by means other than tests, provided that certain conditions are met, so animal testing can be reduced or avoided by replacing traditional test data with predictions or equivalent data.

Integrated testing strategies (ITS), including *in vitro* assays, QSARs, and "read-across", can be used in a combined "non-testing" strategy, i.e., as an alternative to the use of animals. In read across, known information on the property of a substance is used to make a prediction of the same property for another substance that is considered similar. This avoids the need to test every substance for every endpoint, but there are conditions. QSARs are allowed under REACH if the method is scientifically valid, the domain is applicable, the endpoint is relevant, and adequate documentation is provided.

Step 1 in the tiered ITS approach is information collection: the European chemical Substances Information System (ESIS)[4] has been developed, together with some more specific databases. Step 2 is the preliminary assessment of reactivity and fate. Commercial software and databases are available but JRC has chosen to develop some freely available and open-source software:[5] CRAFT (Chemical Reactivity & Fate Tool), START (Structural Alerts in Toxtree[6]) and the OECD Toolbox.[7] CRAFT and START are being developed in collaboration with Molecular Networks of Germany. A bewildering array of SAR and expert system tools could be used, but again JRC has concentrated on freely available and open-source software such as Toxtree[6] and the OECD Toolbox. Toxtree is an application which is able to classify chemicals into modes of action and estimate toxic hazard by applying decision tree approaches. It is being developed in collaboration with Ideaconsult of Bulgaria. DART (Decision Analysis by Ranking Techniques) is a flexible, user-friendly, open source application, which is able to rank and group chemicals according to properties of concern. This is developed in conjunction with Talete, Italy. Toxmatch[8] is a chemical similarity tool which supports chemical grouping and read across. In the interests of international collaboration and harmonization, the JRC is also contributing to the development of the OECD QSAR Toolbox.

Finally, the JRC QSAR Model Database is an inventory of information on (Q)SAR models (also developed in collaboration with Ideaconsult). This can be searched in various ways including substructure and similarity search. Further guidance is needed on how to assess the adequacy of non-testing data by weight-of-evidence approaches.

### **Modeling and informatics support for safety and metabolism studies in early drug discovery**

Scott Boyer, AstraZeneca

Drug candidates may fail because of target pharmacology, off-target pharmacology, or chemically related toxicity. As a generalization, on-target pharmacology (efficacy) is easy; the other two areas (safety) are hard. A pharmacologist's view of Cyclooxygenase 2 (COX-2) is simple; a toxicologist's view is complicated. The scientist must insure that the "obvious" compound liabilities (cardiac arrythmias, genetic toxicity, hepatotoxicity) are addressed, and must use hypothesis generation when things go wrong.

Human Ether-a-go-go Related Gene (hERG) encodes an ion channel, abnormalities in which may lead to either long or short QT syndrome, both of them potentially fatal cardiac arrhythmias. In *in silico* prediction of hERG activity in drug discovery, the models get more sophisticated as the pipeline is traversed. As a general strategy, most models are tuned to enhance the negative prediction rate, since false negatives in safety are expensive, and positives are tested if they are real compounds, and reprioritized if they are virtual ones. Because the interactions in hERG mechanisms are diverse, chemical descriptors must be diverse: a docking score for size/shape complementarity, pharmacophore features (correct spatial orientation of features), and traditional descriptors such as physicochemical properties. AstraZeneca gets consistently better results from a consensus prediction using all three.

Local QSAR models are validated to make sure that they can predict the future but models lose their accuracy over time: as the chemical space expands the quality of prediction degrades. At AstraZeneca machine learning is automated and QSAR models are used by chemists in library design. It is very important that the system is user-friendly or the model will not be used. The system could have predicted that the antihistamine Allegra (fexofenadine) would be “safe” and Seldane (terfenadine) “unsafe”. (Seldane is thought to have been involved in more than 10 hERG related deaths.) Results such as these are recorded in the AstraZeneca system with a link to the full text of the original publication to give the chemists evidence they can believe. In 2003 more than a quarter of compounds in the company’s compound collection were predicted to be hERG blockers. This trend has been reversed since 2004 when multiple computational and experimental hERG methods were introduced.

AstraZeneca’s Genetox database has non-validated data from the Chemical Carcinogenesis Research Information System (CCRIS), FDA-approved data from MCASE, the quality of which is roughly known, and data of known quality generated in-house. The Ames risk assessment system runs automatically and by “inverse QSAR” shows the chemist which substructure is most significant for a negative or positive prediction.

There are more than 10 different pathologies for hepatotoxicity. Reactive metabolites should be avoided if possible. AstraZeneca uses essentially the same procedure for structural warnings as it uses for hERG. Glen, Boyer and colleagues have shown how predictive metabolism methods in drug discovery projects can be used to enhance the understanding of structure-metabolism relationships.[9] In the SPORCalc system the Symyx Metabolite database was mined to exploit biotransformation data. Reaction center fingerprints were derived from a comparison of reactants and products to give two fingerprint databases: all atoms in all reactants and all reacting centers. The metabolic reaction data are then mined by submitting a new molecule and searching for fingerprint matches to every atom in the new molecule in both databases. A normalized occurrence ratio derived from the fingerprint matches enables the search results to be rank-ordered as a measure of the relative frequency of a reaction occurring at a specific site within the submitted molecule. Boyer has also worked with Mestres’s team on biological fingerprinting. using SHED molecular descriptors.[10] Hypothesis generation is critical for rapid problem solving.

Boyer’s final comments concerned physicochemical properties. In work as yet unpublished, he and Tudor Oprea have used the maximum recommended therapeutic dose (MRTD) data and “classes” from Matthews *et al.* (2004) now available in DSSTox[11]. The MRTD classes were defined as low (active), medium (marginal) and high (inactive). The classes were compared in terms of log*P* and volume of distribution. Low MRTD was indicative of toleration problems. Low MRTD drugs are more lipophilic, interact with more targets, and are more widely distributed. Optimization of ligand efficiency is important in lead selection.

In summary, QSARs should be accurate, to the point the data will allow, should reflect a testable endpoint, and should be supported by interpretations and past experience. Data mining should reflect summary data in terms of structure, and help develop focused hypotheses and experiments. Control of physicochemical properties is critical. In the discussion session after his talk, Boyer remarked that log*P* estimation is pretty good, but $pK_a$ estimation is pretty poor. Unfortunately, log*D,* which is what matters, depends on $pK_a$.

**Recent developments in toxico-cheminformatics: supporting a new paradigm for predictive toxicology**
Ann Richard, EPA

"*A major focus for the future of computational toxicology will be integration and analysis of large data sets. The current state of toxicity databases is something of a mess. There are a number of databases, each with differing content, architecture, and searchability, that makes the task of integration extremely difficult.*" Lawrence Marnett, editorial in *Chemical Research in Toxicology*.

The Distributed Structure Searchable Toxicity (DSSTox) public database and website[11] provide a public forum for publishing downloadable, structure-searchable, standardized chemical structure files associated with toxicity data. The data are put into a model where they are easier to manipulate. Data are deposited in PubChem: 11 DSSTox "bioassays" are already in PubChem. Structure search is possible in DSSTox and there are links out to other resources: ChemSpider, PubChem, the EPA Aggregated Computational Toxicology Resource (ACToR), Lazar in silico tox,[12] the National Toxicology Program (NTP), the National Center for Biotechnology Information (NCBI), and the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). EPA is linking people to information, with chemical structure as the key, and is working toward a public toxico-chemogenomics capability by chemical indexing of EMBL-EBI and links to NCBI Gene Expression Omnibus (GEO). Structure and similarity searches can be used to produce a meta data set for a given chemical.

A National Academy of Sciences (NAS) panel has called for a major shift in how EPA assesses the toxicity of chemicals.[1] In 2007, EPA launched the ToxCast program[13] in order to develop a cost-effective approach for prioritizing the toxicity testing of large numbers of chemicals in a short period of time. Using data from state-of-the-art high throughput screening (HTS) bioassays developed in the pharmaceutical industry, ToxCast is building computational models to forecast the potential human toxicity of chemicals. The goal is to derive "signatures" from *in vitro* and *in silico* assays to predict *in vivo* endpoints.

In its first phase, ToxCast is profiling over 300 well-characterized chemicals (primarily pesticides) in over 400 HTS endpoints. Various chemical classes and diverse mechanisms of action are included. *In vivo* data have been extracted from PDF, TIF files etc., and put into a relational database, ToxRefDB. ToxCast will have millions of dollars worth of *in vivo* chronic and cancer bioassay effects and endpoints. ToxRefDB has been used in profiling of liver effects for pesticides. Liver non-neoplastic histopathology and increased organ weight are often associated with tumors and cancer. The activity profile of a compound is the refined "endpoint" for SAR modeling.[14] Nine EPA contracts provide chemical procurement; hundreds of biochemical, cellular, tissue and genomic assays; model organisms; and the capacity to screen up to 10,000 chemicals.

SAR concepts are being incorporated into ToxCast. The system holds chemical structures, HTS data ("fast biology") and bioassay (*in vivo*) data ("slow biology"). We have to use "fast biology" to begin to address the backlog of untested chemicals. One SAR approach to toxicity prediction is global modeling and another is chemical class-based modeling. A third approach, using a bioactivity profile of a structure class is richer information. Chemical structure classes are identified by clustering according to activity and mechanism. Differences in activity profiles can discriminate within a structure class; a bioactivity profile class can be projected onto multiple chemical classes. This gives potentially broader coverage of chemical space and implies mechanistic similarity. HTS assay data, positive or negative, is incorporated as biological "descriptors". *In vivo* activity clusters can also be used. It could be that biology predicts chemical similarity better than chemistry predicts biological similarity.

The 320 pesticides in ToxCast have been deposited in PubChem. As you move away from the 320, there are fewer and fewer *in vivo* data. Phase I of ToxCast is proof of concept. Later phases will produce an affordable science-based system for categorizing chemicals. There will be increasing confidence as the database grows. ToxCast will identify potential mechanisms of action, and refine and reduce animal use for hazard identification and risk assessment.

## Strengths and Limitations of Current Toxicity Prediction Systems

### Understanding toxicity from predictive data mining

Chihae Yang, ORISE fellow, US FDA Center for Food Safety and Applied Nutrition (CFSAN)

US FDA CFSAN has initiated a project to develop a knowledge base for food additives and ways to implement computational risk assessment methods within the workflow of reviewers. One of the goals of this project is to develop the center's knowledge base, which will be disseminated through structural categories and predictive models. Predictive data mining is the process that is being used at US FDA CFSAN to build these components of the knowledge base.

The currently available SAR paradigm is associated with a couple of inherent issues. First, when linking chemistry and biology, there is an inadequate description of biology per chemical feature: complex biology is compressed into a highly summarized outcome, while the chemistry domain is extremely diverse and sparse. Chemical diversity is of the order of $10^{59}$ while biology diversity is of the order of $10^{9}$ and is very highly summarized (0,1 or an $LC_{50}$ value). Most of the time, the training sets of the models suffer from this issue and the data mining process has been a black box. Second, most of the QSAR models are "global" and data-driven while the original SAR paradigm is defined only within a mechanistic domain. Hence the issue of "global" *versus* "mechanistic" models requires the clear definition of the applicability domain, where valid ranges of independent variables must be established. We must make both the data mining process of the training data set and the limitations of the knowledge base transparent and sustainable.

The predictive data mining process begins with data preparation, which is followed by data mining and analysis, then knowledge base development of structural rules and prediction models, and then applying and disseminating knowledge. There are two types of learning goal: (1) identifying relationships between structural classes and various toxicity endpoints so that an intelligent testing decision can be made ("what do I make next?"); and (2) developing structural alerts and QSAR models to assist safety decisions.

Data sources included the Leadscope-FDA genetic and carcinogenic toxicity databases (constructed according to the criteria of the ToxML standard)[15] and biological assay data from the first National Toxicology Program High Throughput Screening (NTPHTS) campaign. The toxicity endpoints considered in this study are genetic toxicity (*salmonella*, mouse lymphoma, *in vitro* chromosome aberration and *in vivo* micronucleus) and rodent carcinogenicity (mouse and rat). The sources of the Leadscope genetic toxicity and carcinogenicity databases include NTP, CCRIS, Tokyo-Eiken (Tokyo Metropolitan Institute of Public Health Epidemiological Information Office), US FDA, and primary publications.

The NTPHTS campaign data are from an HTS project which NTP has initiated to explore new approaches to evaluating chemicals across a spectrum of high-throughput biological assays. Assays are being selected based on their potential to be informative of animal bioassay results and relevant to human health risk assessments. As an initial phase of this project, NTP has provided a set of 1,408 chemicals from NTP inventories for HTS in bioassays relevant to toxicology, to the NIH Chemical Genomics Center (NCGC), part of the NIH Molecular Libraries and Imaging Roadmap (MLR) initiative. Assays are described and assay results reported in PubChem for this NTPHTS chemical data set in the same manner as for compounds from the Molecular Libraries Small Molecule Repository. The DSSTox project[11] is collaborating with the NTPHTS project to provide structure annotation and cheminformatics support for this effort. Drawing largely from the contents of the existing NTP Bioassay Online Indicator (BSI) Structure-Index Locator File, the DSSTox NTPHTS Structure-Index File provides the full complement of DSSTox standard chemical fields for the NTPHTS chemical set.

Once the data set is prepared, the data mining and analysis steps follow. The compound level profile, a data matrix of compounds and activity (positive or negative) for the six endpoints is sparse. For example, out of a total of 3,548 structures included in this study, only 45 have all the data for four genetic toxicity endpoints. To profile the association between chemistry and biological endpoint better, chemical structures are decomposed into features. A feature level profile is a data matrix of structural classes and average endpoint results for each class, which has very few empty cells. The structural classes can be any chemically meaningful fragments and Yang used Leadscope features as an example. Multivariate analysis of structure classes allows one to detect, for example, *salmonella* negatives which are mouse lymphoma positive. Non-concordant chemical classes can give insights, e.g., the pyrrolidine, 2-oxo class is *salmonella* positive, and mouse lymphoma negative; aromatic amines and alkyl halides are *in vitro* chromosome aberration positive but micronucleus negative.[15]

Probabilistic analysis is possible when the database is sufficiently large. For example, a probability of a compound or a structural class to be mutagenic or carcinogenic can be estimated from a large database. Probabilities can be marginal, conditional, or joint. If 2000 out of 8000 compounds are *salmonella* positive, the marginal probability of a *salmonella* positive result is 0.25. Conditional probability is defined under a given condition; for example, if 75 out of 150 compounds are *salmonella* positive given that micronucleus is positive, then the conditional probability is 0.5. Joint probability is the probability of both salmonella and micronucleus being positive. If the marginal probability of the micronucleus is 0.25, then the joint probability of a compound being positive for both *salmonella* and micronucleus positive is 0.25x0.25 (0.0625).

This probability analysis can be extended to the structure-class level. If a class is a structural alert for *salmonella*, then the probability of the class for *salmonella* should be high. The mean values from each structural class can be used in the probabilistic analysis and further for predictive likelihood. Since a chemical is made of these structural classes, a joint probability can be calculated to estimate the likelihood of a chemical to be *salmonella* mutagenic, for example. These structural classes provide a chemical features dimension and act as a link between the structure toxicity matrix for compounds tested *in vivo* and the structure assay matrix for compounds tested *in vitro*. Significant features describing compounds can be related to the probabilities, and probabilistic feature analysis can be carried out. This probabilistic analysis provides validation for using these classes as structural alerts and molecular descriptors in QSAR models.

After training sets and endpoints have been prepared and descriptors selected and validated, a weight of evidence approach can be applied to statistical QSAR modeling. For example, Yang presented a rat carcinogenicity model based on molecular descriptors including the structural classes, physicochemical calculated properties, and NTPHTS screening assays. To model *salmonella* negative but rat carcinogenic, NTPHTS screening assays reflecting mostly apoptosis cycle were used. Four submodels based on structural classes were put together by optimizing the weights for individual structural class models to result in one final model. A combination of descriptors selected based on our knowledge of chemistry and biology leads to a much simpler interpretation of the domain of applicability and weight of evidence optimization improves reliability.

In the FDA CFSAN critical path project, this predictive data mining method will be transparently documented for reproducibility. The plan is eventually to disseminate knowledge in a decision tree type of algorithm for making the computational knowledge available on demand to the reviewers. This plan also includes making some of this knowledge base publicly available.

### Consensus QSAR models

Mark Cronin, Liverpool John Moores University

Integrated testing strategies (ITS) can involve compilation of *in silico* predictions from the same or similar techniques, such as regression based models; compilation of *in silico* predictions from different techniques, weighting or averaging predictions; and compilation of *in silico* predictions with *in vitro* and *in chemico* data. Consensus modeling of regression-based QSARs for large, heterogeneous data sets requires large groups of physicochemical descriptors and/or properties and a method to select them. A pool of models is created (usually regression models but neural networks can also be used) and the best (statistically) QSARs and/or most diverse QSARs are determined. Predictions are weighted or averaged. The method often performs better than a single QSAR.[16] One study by Cronin's own team shows that the use of consensus models does not seem warranted given the minimal improvement in model statistics for the data sets in question.[17]

Consensus QSAR has also been applied to models developed from different techniques and different data sets. Matthews *et al.* used MC4PC, MDL-QSAR, BioEpisteme, Leadscope PDM, and Derek for Windows, with the same data sets, to predict carcinogenicity.[2] The QSAR models were based upon a weight-of-evidence paradigm, which has a bigger "cost" than weighting or averaging. The individual models made complementary predictions of carcinogenesis and had equivalent predictive performance. Consensus predictions for two programs achieved better performance, better confidence predictions, and better sensitivity. Four QSAR programs predicted carcinogenicity with high specificity (85%). Consensus positive predictions identified clusters of carcinogens with reasonable mechanisms of action. Consensus models from three different expert systems have also been used with some success in prediction of mutagenicity using the commercial system KnowItAll.[18] The system discussed by Boyer[9] earlier in the meeting is a good example of how consensus models can be tailored to a risk assessment scenario.

Consensus can improve models, it confirms predictions in expert systems, it provides greater confidence in predictions, it accounts for outliers and it has regulatory significance and proven use. There are, however, disadvantages. Consensus models hide outliers, incorrect data and interesting parts of the data set. They lack portability, transparency and mechanistic interpretation. It is not clear how to characterize and develop a QSAR Model Reporting Format (QMRF) for compliance with the REACH regulations, for example. Defining applicability domain, statistical concerns, how to carry out validation, cost, and difficulty in use are other concerns.

The initial stage in an ITS is *in silico* assessment and this may include consensus QSAR. If there is insufficient confidence in the *in silico* assessment, *in chemico* assessment, bringing in reactivity data, can be used. From stage to stage in an ITS, more and more information is gathered. *In vitro* assessment follows *in chemico* (although it is not clear whether this will be acceptable under the REACH regulations) and only after all the other methods have been used, is *in vivo* assessment necessary. A special supplement[19] to *Alternatives to Laboratory Animals* deals with the development of ITS for REACH. Integrated testing strategies will reduce animal usage by providing frameworks to use non-test data but the European Chemicals Agency supplies only guidance;[20] it does not deal with all the requirements (weighting factors, costs, probabilistic techniques for decision making, tools and case studies) for making an ITS functional. Expertise is required for success.

Ensemble models are controversial. Ann Richard pointed out that you cannot just use eight models some of which are awful: you must apply some judgment. Douglas Hawkins added that you will not achieve much from consensus of good models; consensus adds improvement if you have several weak methods. You should not mix good and bad. Bobby Glen is suspicious of mixing models. Mark Cronin says that if your base model is poor it may tell you something about your data set. Bobby thought that it might be better to model the *process*; he referred to phenomenological models. Mark Cronin is currently looking at reactivity, specifically at modeling glutathione activity.

### QSAR approaches, models and statistics relating to toxicity prediction

Douglas M. Hawkins, University of Minnesota

(Jessica J. Kraker, University of Wisconsin, was a co-author.) Consider a set of dependent measures, *Y*, and predictors, *X.* The dependent measures may be binary (e.g., toxic or non-toxic) or numeric. The predictors are almost unlimited topological descriptors, atom pairs, Burden numbers etc. QSAR modeling relates *Y* to *X.* Models can be broadly categorized as global, where a single *X:Y* relationship is used for the whole data set, or local, where different models are used in different parts of predictor space. Note that some people may use a slightly different definition and equate local with congeneric.

Global models may arise from linear methods (ordinary regression, ridge regression, least angle regression, lasso, elastic net, partial least squares, principal component regression, logistic regression) or nonlinear methods (primarily neural nets). Local models are derived using *k* nearest neighbors, kernel methods, Support Vector Machines (SVM) or tree models. Global methods are good when true, but potentially disastrous if not true. Local methods are arguably conservative and safe, but they are of lower statistical efficiency, squeezing less information from the data.

In global, additive models a predictor set *X* is written as $x_1, x_2, \ldots x_p$, there are *n* compounds for the fitting, and the model determines suitable functions $h_j$ and predicts *Y* on the basis of

$$\sum_{j=1}^{p} h_j\left(x_j\right)$$

Neural nets take this form. Usually functions *h* are monotonic, so they require that "more *x* is better" or "more *x* is worse". Additive models assume no interaction between predictors. Linear methods further specialize the additive model to the form

$$X^T\beta = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$$

In feature selection, some coefficients are zero. Prediction is not necessarily $X^T\beta$; a link function $g(X^T\beta)$ could be needed if a curve is produced instead of a straight line.

The linear regression family is intended for numeric dependents, but it also works for a binary classification (by regression formulation of linear discriminant analysis (LDA)). The traditional method is ordinary least squares (OLS), but this requires *n* to be much greater than *p*, so it is not useful in many QSAR applications. If *n* is less than *p*, variants to get round under-determination are ridge regression, least angle regression, lasso, and elastic net. The last three methods can also do feature selection. OLS often works better than it has a right to.

Two other linear methods are Partial Least Squares (PLS) and Principal Component Regression (PCR). PLS is computationally fast and empirically performs well, though formal statistical proofs of good properties are sparse. PCR relies on the assumption that a few latent dimensions drive both *X* and *Y*. Its performance is spotty and it is probably safe to ignore it.

At first glance, kernel Support Vector Machines are linear regressions applied to transforms of *x*. In practice though, SVM is a local method. It rests on the choice of a kernel function and its effectiveness rests on how well this is selected. A kernel regression method predicts the *Y* at some future *X* as a weighted average of all $Y_i$ in the calibration data, weighted by the distance between *X* and $X_i$. The quality of the results depends on the weighting function. Any prediction in principle requires the full calibration data set, so this method does not scale well.

Nearest neighbor is a cousin of the kernel regression methods. To predict at *X*, it finds the *k* calibration cases closest to *X* and uses these cases' *Y* values to predict *Y* by the average if *Y* is numeric, or the modal class if *Y* is a classification. *k*NN has some drawbacks. Prediction requires the full data set. A distance metric is needed. Conventionally Euclidean distance is used, ignoring the impact of correlation among predictors. Scaling is also a concern.

Recursive partitioning (RP) produces a tree model. This has minimal statistical assumptions and making predictions is easy but one drawback is that RP needs big samples. Random forests improve on single trees, squeezing more information out of data. The goal of feature selection is to pick the predictors that matter, and eliminate redundant ones. It is vital in drug discovery but may be less so in toxicology. Some linear regression models can do feature selection, and RP relies on it. Feature selection is harder with the other methods.

Models must be validated. If many compounds are available, a learning set and a validation set can be split out. All model building is done on the learning set and testing is done on the validation set. If only a moderate number of compounds is available then it is advisable to use cross validation instead.

Hawkins presented two examples from his own work on two data sets using about 300 mainly topological descriptors. The first data set was a mutagenicity one: 508 compounds with binary data from the *CRC Handbook of Identified Carcinogens and Non-Carcinogens*. The second was the Crebelli data set of 55 halocarbons assessed for D37 toxicity (with a numeric objective). The models were evaluated by cross validation. For the *CRC Handbook* data set there were appreciable differences in method capability. Random Forest was best, in line with notion that "ensemble" methods work well. PLSLDA is second best. SVM was worse than random. For the halocarbons, elastic net was best, with RP methods a little behind.

Hawkins also summarized some results reported by Young and Hughes-Oliver at the 2008 Spring National ACS Meeting (to be published in *Cheminformatics*). For 57,821 compounds tested in cathepsin L, these authors found that Random Forest and atom pairs were a good choice. These are only three examples. Sometimes global methods win; sometimes local ones. It depends on the descriptors and the dependent measures. The lesson is perhaps not to be wedded to a single QSAR methodology. From the audience, Stephen Pickett commented that in other QSAR areas SVMs have been shown to perform comparably to the other methods employed here.[21] SVM should not be used without feature selection.

**Knowledge-based approaches for toxicity prediction**
Nigel Greene, Pfizer

Early hazard identification in the pharmaceutical industry is very important because development costs increase exponentially over time and stage. Adverse safety effects may be due to primary pharmacology (e.g., phosphodiesterase-4, PDE-4, inhibitors are linked to emesis and vasculitis), secondary pharmacology (e.g., D-1 activity is linked to tremors), chemical structure (e.g., clozapine causes agranulocytosis and forms reactive metabolites), or physicochemical properties (e.g., lipophilic basic compounds have a risk of causing phospholipidosis, hepatotoxicity, and QT interval prolongation].

Approaches to toxicity prediction based on machine learning are fast to find relationships and can deal with complex data and relationships, but they are dependent on high quality data and can be difficult to interpret. Knowledge-based approaches can cope with "fuzzy" data sets and the results are easy to interpret, but they are slow to develop and they may not identify complex relationships. It is always important to consider exposure. Greene illustrated this point with a data set where area under the curve (AUC) and maximum concentration (Cmax) vary over seven orders of magnitude for compounds that are administered at the same dose. [In bioequivalence studies, Cmax is shown to reflect not only the rate but also the extent of absorption. Cmax is highly correlated with AUC contrasting blood concentration with time. Therefore, use of the Cmax/AUC ratio is recommended for assessing the equivalence of absorption rates.]

In the hit identification to lead optimization phases of drug discovery, knowledge-based approaches can quickly identify a toxicophore and a potential mechanism. In addition any new information can be added quickly to the knowledge base. In later phases, knowledge-based approaches can help with risk assessment for synthetic intermediates, low-level impurities, metabolites, degredants and excipients and the information may form part of a regulatory submission.

Legacy and public data are fairly readily available for *in vivo* outcomes that can be used for building *in silico* models. It is easy to correlate the predictions of these models to clinical outcomes but often it is not clear if there is an *in vitro* assay to use for confirming the prediction. It is difficult and costly to validate these systems and often model performance may be limited by the complexity of the biological systems. Model development is also hindered by a lack of exposure information in preclinical species.

*In silico* models to predict the results of *in vitro* assays can be used to prioritize screening and thus reduce assay capacity requirements, and give a clear next step for exploring a potential safety issue. It is relatively easy to validate these models, but they require a training set of compounds, and often the *in vitro* assay correlates poorly with *in vivo* outcomes. Cell based systems reduce complexity of the system to some extent because fewer mechanisms are involved.

*In vitro* assays have higher throughput and are cheaper to run than *in vivo* ones. They reduce the time taken to make a decision and enable SAR and comparison of series. They implement the “three R’s”: reducing, refining and replacing animals. The relatively simple readouts are easy to understand. Thus many common *in vitro* assays (e.g., hERG patch clamp assay, Ames test) are in use, although they frequently have poor correlation to *in vivo* toxicity and may not be broadly applicable. A high false positive rate would eliminate too many compounds that might be useful, but the assays may help resolve mechanisms of toxicity were an *in vivo* issue identified.

Derek for Windows is an example of a knowledge-based approach. It is continually improving, users can store their own knowledge using the editor, and it is not an isolated system since it can interact with other software, such as physicochemical property predictors, through an adapter. Greene gave an example of the development of an alert. From 156 compounds in a database, more than 71 with a 2-aminopyrimidine substructure were positive in an *in vitro* micronucleus assay. SAR revealed several distinct subclasses. Mechanisms of action were studied. Greene showed some colored matrices of ranges of values for inhibition, showing that one structural class consisted of non-selective kinase inhibitors while another contained selective inhibitors of an unrevealed enzyme.

Greene has also developed hepatotoxicity SARs. About 50 structural classes known to cause liver injury in humans were identified and implemented in Derek for Windows. Performance was evaluated against about 600 compounds compiled from internal and external sources. The validation set contained both idiosyncratic and dose-dependent hepatotoxicants. Derek for Windows predictivity for positives was 45% and for negatives was 76.3%. The results were rather better if weak and animal-only results were ignored. Sensitivity was reduced due to animal hepatotoxicants not identified correctly and specificity suffered due to compounds that have fewer than 10 case reports of liver injury. The development is addressing these issues.

Predictions based on physicochemical properties are also being developed. The Ploeman model[22] involving CLogP and $pK_a$ has been implemented in Derek using the adapter to computational models. A statistically significant correlation between CLogP and Topological Polar Surface Area (TPSA) and increased incidence of findings in *in vivo* toxicology studies is also being considered.

There has been much research looking at using batteries of *in vitro* assays to predict an *in vivo* outcome, for example in the EPA’s ToxCast, CEREP Bioprint profiling, and work by Roche and Pfizer on kinase selectivity as a surrogate for *in vitro* micronucleus. Some success stories have been reported but the patterns may not be broadly applicable.

## Models and databases for genetic/carcinogenic toxicity

Romualdo Benigni, Istituto Superiore di Sanitá (ISS), Rome, Italy

In the framework of a collaboration between the ISS and the European Chemicals Bureau (ECB), a series of non-commercial (Q)SARs for mutagenicity and carcinogenicity have been evaluated.[23] These include structure alerts, and QSARs for congeneric classes of chemicals. Structure alerts are a coarse-grained approach to SAR, whereas QSARs are fine-tuned.

Knowledge about the action mechanisms as exemplified by structure alerts is routinely used in SAR assessment in a regulatory context. In addition, alerts are at the basis of popular commercial systems such as Derek for Windows. Benigni and co-workers identified four structural alert models as particularly promising.[24-27] The four did not differ to a large extent in their performance. In the general databases of chemicals the alerts appear to agree around 65% with rodent carcinogenicity data, and 75% with *salmonella* mutagenicity data.

The alert-based models do not seem to work equally efficiently in discriminating between active and inactive chemicals within individual chemical classes. Thus, their main role is that of preliminary, or large-scale screenings. They are excellent tools for coarse-grain characterization of chemicals, for example description of sets of chemicals, preliminary hazard characterization, category formation and priority setting (enrichment). A priority for future research is the expansion of structural alerts to include alerts for nongenotoxic carcinogens.

Based on the experience gathered from the above survey on the structure alerts, a rule base for mutagens and carcinogens has been designed and implemented in Toxtree 1.50.[6] It uses a structure-based approach consisting of a new compilation of structure alerts, for both genotoxicity and nongenotoxicity. It also offers three mechanistically based QSARs for congeneric classes (aromatic amines and aldehydes).

In the same survey, local QSARs for congeneric classes were short listed based on the following criteria: interpretability from a scientific (mechanistic) point of view, good internal statistics, and domain applicability. A crucial point is that of “validation”. Whereas it is generally accepted that the gold standard is to test the model on a set of chemicals not used for the derivation of the model, in practice many investigators use different statistical procedures to generate artificial test sets, for example, splitting the chemicals into training and test sets. On the contrary, in this survey the short listed QSARs were challenged to predict the activity of external sets of chemicals, never considered by the authors.

Benigni presented tables summarizing the external prediction outcomes for regression based models (i.e., QSAR models for potency), and the outcomes for discriminating models (i.e., QSAR models for activity). The two tables reported also parameters for goodness of fit and different internal validations of the training set. In summary, all the short listed local QSARs are scientifically interpretable and have good internal statistics, but they vary in their external predictivity. In QSARs for potency the predictions are 30–70% correct and in QSARs for activity the predictions are 70–100 % correct. Estimating intervals is more reliable than estimating points. In addition, it appears that internal validation measures do not correlate with external predictivity.[28]

Mechanistically-based models should be preferred, since this gives a common ground for modelers, toxicologists and regulators, and provides an additional tool for minimizing chance correlations, and intelligible information for synthesizing safer chemicals. Unfortunately, existing local, mechanistic QSARs are limited in number and the mechanistic understanding of many human health effects is not possible at this time. In many instances there is no alternative to models for noncongeneric chemicals aimed at modeling simultaneously “all” chemical classes. There are many commercial systems of this type. Often they use non-mechanistically based descriptors and offer no mechanistic interpretation. They are mostly validated through internal statistics alone. Independent external validation studies of these models have pointed to a great variability of their predictivity in the different regions of the chemical space.

The recent progress in the technology and availability of chemical relational databases provides new opportunities to QSAR modeling.[29] New fine-tuned QSARs can be created by intelligent interrogation of databases. For example, a published QSAR model for the mutagenicity of αβ-unsaturated aldehydes has been proposed by Benigni to the European Food Safety Agency's FLAVIS group for their priority setting of αβ-unsaturated carbonyls.[30] Since ketones were not considered in the paper, databases were interrogated and data on their mutagenicity were retrieved. This permitted the generation of a new mechanistically-based QSAR model for the mutagenicity of the αβ-unsaturated ketones (Benigni, unpublished).

## Emerging Areas and Technologies in Toxicity Prediction

### The cardiovascular PhysioLab platform and its applications in toxicity prediction

Héctor de Léon, Entelos

Entelos has developed a set of biosimulation and gene expression profiling tools aimed at the early identification of effective drug candidates with low toxicity. Its main strength is in a dynamic representation of whole-body lipoprotein synthesis, distribution, processing and uptake. The company has assembled the cardiovascular PhysioLab platform, a large-scale mathematical model of human lipid metabolism and cardiovascular pathology, to evaluate the potential efficacy of alternative therapeutic approaches. The model uses differential equations to represent interactions of cells and biomolecules linked to key cardiovascular clinical outcomes such as myocardial infarction. Finite-element modeling is used to simulate the temporal changes in the structure of atherosclerotic plaques that lead to rupture. The structural stability of the plaque can be linked to an estimated risk of a cardiovascular event. Virtual patients and patient populations are used to represent different pathophysiological hypotheses and to analyze the impact of phenotypic variability in response to therapies and drug-induced toxicity. The PhysioLab platform has been validated against data from a number of clinical trials.

De Léon presented a case study of identification of novel candidate biomarkers for patient stratification. Raising high-density lipoprotein cholesterol (HDL-C) is a promising strategy in prevention of cardiovascular disease, and cholesteryl ester transfer protein (CETP) inhibitors have been developed to reduce atherosclerosis. Entelos used a biologically diverse cohort of 60 virtual patients and simulated their response, after two years, to treatment with either a statin or a statin plus a CETP inhibitor. Only a third of patients responded to CETP inhibition, determined by percent atheroma volume. Baseline lipids did not correlate with response to CETP inhibitor treatment. A novel candidate multivariate biomarker was identified for exclusion of CETP adverse responders prior to treatment.

Entelos can identify novel, optimal collections of measurements (candidate multivariate biomarkers) predictive of efficacy and/or safety; determine different classes of biomarkers; provide assessments of biomarker robustness (using sensitivity and specificity, and $R^2$) and optimality; and provide recommendations for means to validate candidate biomarkers.

The company has also developed DrugMatrix, a toxicogenomic database of microarray expression data linked to classic preclinical and clinical toxicology measurements, to identify predictive gene expression profiles. Hypotheses generated from these profiles can be simulated in the cardiovascular PhysioLab platform to identify biomarkers predictive of adverse events. De Léon presented a case study identifying putative mechanisms to differentiate efficacy and safety of two compounds.

Peroxisome proliferator-activated receptor γ (PPARγ) agonists, such as Actos (pioglitazone) and Avandia (rosiglitazone), activate nuclear hormone receptors which improves insulin sensitivity. De Léon hypothesized that reported differences in lipoprotein particle distributions in patients treated with Actos *versus* Avandia correspond with associated differences in hepatic gene expression. He queried the DrugMatrix database and compared hepatic gene expression profiles from animals treated with Actos or Avandia, and established differences between Actos- and Avandia-induced changes in gene expression. When administered in high doses sub-chronically, Actos and Avandia evoke differential patterns in hepatic gene expression. Entelos is still analyzing the significance of these results. De Léon also hypothesized that the reported differences in lipoprotein particle distributions in patients treated with Actos *versus* Avandia yield differential effects on plaque growth and stability, and he presented evidence for his hypothesis.

In the discussion session, the chairman was concerned about incorporation of exposure and dose rate; an attendee from GSK commented that this was not really the forum for making unsubstantiated (or non-validated) claims in public about a marketed drug [Avandia]; and another attendee asked whether we need to understand differential equations (asking about the finite element modeling system used by the PhysioLab platform to simulate plaque rupture).

**Emerging areas in toxicity prediction: an NIHS perspective**
Akihiko Hirose, National Institute of Health Sciences (NIHS), Japan

We urgently need to develop a high throughput evaluation system for the risk to humans of environmental chemicals. Since no individual QSAR system is powerful enough, NIHS has started to develop a workflow to assess genotoxicity using a combination of three *in silico* systems: Derek for Windows (a rule-based system), MCASE (a database and substructure-based system) and ADMEWorks (an unsupervised regression classification system from Fujitsu Kyushu System Engineering).[31]

Each system was customized for mutagenicity prediction, using bacterial gene mutation and *in vitro* chromosomal aberration assays. In a combination approach, the concordance between *in vitro* and *in silico* assays on bacterial gene mutation reached around 94%, although applicability decreased to 55%. Next, NIHS tried a similar approach for developing a chromosomal aberration prediction system. The performance in this case was even lower than that of bacterial mutagenicity prediction and further development is required.

In addition to these genotoxicity studies, repeated dose rat toxicity studies are commonly used to evaluate the risk of industrial chemicals, but no suitable *in silico* general toxicity evaluation system is available at present. NIHS analyzed the toxicity profiles of hundreds of 28-day repeated dose studies and focused on developing a prediction system for hepatotoxicity and/or renal toxicity endpoints, by searching new substructural alerts for Derek for Windows and using Leadscope Predictive Data Miner, a discriminant-based QSAR model builder. Rapid alerts are being developed for Derek for Windows in order to improve the sensitivity, although this may cause an increase in the number of false positives. With Leadscope Predictive Data Miner high concordance models could be obtained by using a consensus approach or by restricting the probability thresholds, although the applicability was decreased to about 40-50%. With the ADMEWorks model builder a high concordance model to predict liver weight changes, using an SVM method, was obtained as a single prediction model but other models had relatively low concordance. In order to improve predictability, a combination approach of Derek and the statistical data mining models would be required. In addition, more accurate structural alerts and endpoint-specific prediction models could be constructed by using a more precise learning data set.

NIHS has also joined a multi-institutional Japanese project developing a repeated-dose toxicity knowledge base system, which could assist toxicological expert judgment, or support preliminary governmental decisions. The system consists of three parts: a detailed subchronic toxicity studies database, a toxicity mechanisms database, and a metabolite prediction system. The project is led by the National Institute of Technology and Evaluation (NITE), Tohoku University, Kwansei Gakuin University, Fujitsu Co. Ltd., and NIHS. Parts of this *in silico* knowledge based system will be integrated in the OECD (Q)SAR Application Tool Box,[7] and will support a categorical approach to evaluation of high production volume chemicals. A repeated-dose toxicity (Q)SAR system will also be developed in future.

**Application of *in silico* modeling in guiding alternatives research in skin allergy**
Cameron MacKay, Unilever

Assuring the safety of consumer products without the need to conduct animal tests is a considerable challenge. The mouse local lymph node assay (LLNA) is now used widely to generate data for assessing the risk of chemical-induced skin sensitization, but changes in EU legislation (in the seventh amendment to the EU Cosmetics Directive) have made developing non-animal approaches to provide the data for skin sensitization risk assessment a key business need.

Skin sensitization is a complicated multistage process and a single *in vitro* assay system is not feasible at present. It is difficult to know where to target assays and how to interpret a battery of disjoint assays. Unilever decided to try applying mathematical models, or "systems biology". The purpose was to explain and elucidate mechanisms, not to predict sensitization *a priori*; this is not a QSAR model. In collaboration with Entelos, Unilever has developed a large-scale *in silico* model of skin sensitization induction (comprising nonlinear ordinary differential equations) to characterize and quantify the contribution of implicated pathways to the overall biological response. Such knowledge is crucial in guiding the development of *in vitro* assay development for use in consumer safety risk assessment.

The model describes the developing immune response in mice over a 7-day period following exposure to dinitrochlorobenzene (a well known contact allergen) and includes both epidermal and lymph node cellular processes implicated in skin sensitization such as cytokine responses, cell surface marker regulation, cellular migration and proliferation. In order to populate the model, *in vivo* and *in vitro* data from the published literature were used. Some of the data, such as epidermal cytokine release in response to chemical insult, were used to build focused submodels of the biology. Cellular data from mouse LLNA were used in order to ensure that, acting together, these submodels could model the full system response effectively.

The modeling uncovered a previously underappreciated pathway in skin sensitization and showed it to be key to the sensitization response. Additionally, the modeling revealed a number of gaps in both the current mechanistic knowledge and the available data. Unilever is using the model to focus and guide its future research in the area of skin sensitization. The session chairman pointed out the importance of using a model to develop an *in vitro* assay.

**Challenges in predicting metabolism and toxicity with known and abstract targets**
Fred Guengerich, Vanderbilt University School of Medicine

The costs of drug development and environmental risk assessment continue to increase, and the availability of better *in silico* and *in vitro* methods has the potential to yield better and more economical predictions. Metabolism issues are key to some toxicities, but an accurate assessment of the fraction of all drug and other toxicities due to metabolism is unavailable and, even when metabolism is agreed to be central, the ensuing biological events are not well understood.

A need exists for better biomarkers and assays to predict not only bioactivation but also off-target toxicity, immune-related toxicity, and idiosyncratic reactions.[32] It is hard to predict toxicity because of lack of understanding of mechanisms. Few protein target structures are available and there is limited information about linear pathways and networks, so SAR has to be done with gross endpoints. The issue of relevance of the parameters used in comparisons is critical in judging the usefulness of the analyzer. Although much has been learned about the enzymes involved in the metabolism of many drugs and other xenobiotics, predictions are not trivial. We do now have crystal structures for the main five cytochrome P450s. Actual protein structures are much preferable to homology models, but even when these are available they often do not predict products accurately.

Boyer and co-workers have studied biotransformation in early drug discovery. Their SPORCalc system[9] is described above. Ligand-based methods for in-house screening can be used when complete P450-specific data are unavailable; SPORCalc compares favorably with docking methods at picking the top three sites of oxidation. Unfortunately, it is very hard to predict rates *a priori*, i.e., to study how fast the compound will be metabolized.

MacDonald and co-workers have published a strategy for identifying off-target effects and hidden phenotypes of drugs by directly probing biochemical pathways that underlie therapeutic or toxic mechanisms.[33] Iconix Biosciences (now part of Entelos)[34] has an *in vivo* predictive toxicogenomics paradigm. In an idealized system, principle component analysis (PCA) could be applied to data from transcriptonomics, metabolomics and other disciplines.

Guengerich closed by summarizing three more big problems. Is the animal model relevant to humans for the toxicity issue? Is the *in vitro* system relevant to the *in vivo* system? Many cell lines lack critical features, e.g., bioactivation. Finally, what about dose? This is a problem in *in vitro* work. What is the human exposure?

## Panel Discussion

*Question*: Where are the critical gaps and needs?
*Hirose*: How to use.
*Cronin*: Guidance, case studies, and workflows especially with respect to REACH. What will the European Chemicals Agency (ECHA) accept? So far, it will accept a valid prediction. In future it will need more data, especially repeated dose data.
*Matthews*: Most molecules have multiple off-target activity. We do not look at these activities enough. We need to look again at QSAR: a compound and its metabolites are a constellation.
*Boyer*: There is a gap between some modelers and some experimentalists. Knowledge about mechanism and basic facts would help in constructing and judging models. Another problem is that models built on 50 compounds are applied to 1000. Reviews are transferred into experimental systems.
*Richard*: There is a problem with the question. What *area* of toxicology are we predicting? What exactly are we modeling? Our models are limited by regulatory requirements and they may or may not be relevant to humans. What endpoints are we modeling? Each endpoint may need a different approach.
*Question*: But what about application in the pharmaceutical industry?
*Richard*: Target to the most relevant endpoint for the drug.
*Guengerich*: More mechanistic information must be built in, with relevance to human toxicology.
*Boyer*: One of the biggest gaps is cultural: the data and the data generators are separate from the modelers. They sometimes do not understand the term “multivariate data”. If they did they might change the data gathered. Iteration in relationships would help here. Clinicians are less receptive. These guys have their own problems and we modelers are giving them too complicated a message.
*Benigni*: They appreciate simple tools such as Toxtree.
*Greene*: Going back to the first question of the key gaps, we have a very limited understanding of the biological processes behind toxicity. If we understood them better we could model them better. The second problem is access to data.
*Richard*: One solution is to engage toxicologists to inject more biology into the model at the level of a structured database and data mining. Chihae Yang’s work matters here.
*Greene*: We could look at how we describe things: people use different terms for the same thing.
*Richard*: ToxML tries to answer this. The International Life Sciences Institute (ILSI) group will make the database available.

*From the floor*: We are good at prediction for rats and mice but we should be looking at human-relevant toxicology. It is convenient to class things as drugs, chemicals etc., but they are all xenobiotic. As Paracelsus said “Everything is toxic”.

*Richard:* We do not have the human data so we work on what we have, e.g., rodent *in vivo* data.

Matthews: There are two problems: the vocabulary (we use Medical Dictionary for Regulatory Activities, MedDRA)[35] and the denominator for exposure. In the case of a rare occurrence in just two people it is hard to establish a mechanism. You need to look at the majority of the population, therefore most pharmaceutical companies use the whole population as denominator to get the most significant results.

*Question*: [Inaudible]

*Matthews*: Gather a large database of adverse drug reactions (ADR) and add the whole population exposure. This is a straightforward computer problem but it has not been done.

*Guengerich*: You can save blood samples in our hospital and track ADRs for any drug, and then tie a hypothesis back to the DNA.

*Matthews*: Tens of thousands of clinical trials are available at the Center for Drugs. There is computer power there but not the other resources.

*Question*: Has the time for *in silico* come? AstraZeneca and Pfizer have put in significant effort.

*Matthews*: Many tools are designed to be used as tool *boxes*, for example the OECD toolbox, and these tools are applied naively. If you developed your own database it would be better. The whole area has taken off. It will explode.

*Question*: But in big pharma resources are challenged.

*Greene*: There has been a massive expansion in the computational area recently.

*Boyer*: This is true for AstraZeneca too but you have to make a reasoned case for an activity.

*Glen*: Large systems must be broken into smaller components so much research is needed. We need to break QSAR down into understandable bits. How do we calculate solubility? Why use octanol/water?

*Richard*: We need a fundamental change in attitude. I know toxicologists who have worked on one group of chemicals for 20 years. We need to pull all the data together. It can cost $250,000 to do one Multigenerational Developmental Toxicity study and $4 million to do a rodent bioassay study. We can do lots with that sort of money. Even in companies you can standardize data even if it is not shared. You need standardization so you can look across data. Compare genomics. You can look at common patterns of effects if you have standards, and at low cost too.

*From the floor*: It is not easy to detect small differences, e.g., small perturbances that may cause a tumor in 30 years time. It is impossible to separate these from the background.

*Boyer*: We should be humble when we look at the magnitude of the task. My DNA is 99% the same as that of a chimpanzee but the 1% difference results in a large phenotypic change. We must recognize how large the problem is and at the same time try to model the small details. We may not be able to model ourselves.

*Benigni*: Microarrays show that gene expression is not very specific: you can reach the same point from different paths. We have big maps of pathways on the wall but we need to know the kinetics. It is not simple to model all this. In an interconnected network of biochemical pathways, the presence of only one feedback loop makes modeling impossible. We have a long way to go.

*Guengerich*: People look for the things that change most, and the smaller changes might be more important. We will understand the central points eventually.

*From the floor*: There will be unknowns but you can still get useful predictions.

*Question from the floor*: Will it be easier with biologicals?

*Boyer*: No. The problems are different but there is the same number of problems.

*Greene*: Agreed.

*Question from the floor*: Will you spend more time on biologicals now?

*Boyer*: Biologicals cannot be used in all areas. We must run studies in species closer to humans. The ambiguity of exposure is another issue.

*Matthews*: This is a significant gap: 40% of our information is on biologicals. We need expert system rules for these, for example peptides with immunologic toxicity.

*Question submitted in writing*: Is rodent bioassay for carcinogens useful?

*Benigni*: Yes. All human carcinogens, when adequately tested, were correctly identified in rodents. You cannot test potential carcinogens on humans; anyway it would take 30 years.

*Richard*: Multispecies hits are of the greatest significance, so we must test in multiple species, and multiple cancer sites, to get significant results.
*Benigni*: In addition, there is a correlation between rat and mouse, and between animals and humans in terms of carcinogenic potency. This is a strong support for extrapolation from rodent bioassay to humans. On the contrary, the results in terms of target organs are quite idiosyncratic.
*Matthews*: With new pesticides there is a data gap in our understanding of pancreas, thyroid, etc. variation among animals. We need to get to grips with this.
*Question from the floor*: Concerning carcinogenicity across species and sexes: how many carcinogens have not shown up in shorter studies in animals? They should be detectable in shorter studies.
*Matthews*: Expert systems were very successful; 90-day and six month studies can be useful.
*Benigni*: There is no relationship between short and long term. Long term rodent positives and negatives are reflected in humans.
*Glen*: In identifying flu epidemics Google searches have been monitored. Could we do this with adverse drug reactions? How about social networks in toxicology?
*Boyer*: We could also look at people's electronic patient records.
*Richard*: Going back to an earlier point, the huge variation in individuals, look at Jeremy Nicholson's work on the metabolome. We can look at general trends and gross generalizations. It is interesting how much you can explain by what people eat. Nicholson can see changes before there is any histopathology.
*Guengerich*: In a study of two sets of rats it turned out that an environmental factor (i.e., which room the rats were in) could be involved. Think also about transcriptonomics.
*Matthews*: There are so many chemicals for which we have no toxicology data. Of 160,000-180,000 common chemicals we have toxicology on only 5%. Let us find the really bad ones: it could be a huge success in a short time.
*Alan Wilson*: Pick out the really bad molecules in pharma.
*From the floor*: We need data curation and ontologies to put the data in the right format. In metabolomics this is being done. Apply the tools to safety as well as to efficacy.
*Boyer*: The experimentalists will change the way they do experiments when they see how successful models are.
*Question from the floor*: Should we be suspicious of 70% specificity?
*Matthews*: We can model some toxicities (e.g., endocrine, heart and kidney) but the liver is harder.
*Question from the floor*: Will there be a balance in future of *in silico versus* experimental?
*Matthews*: The National Cancer Institute (NCI) uses a battery of cell lines and they are extraordinarily successful. QSAR could not do that. But at the next stage we will want to know adverse effects of the hits etc. so you will need *both* methods in future.
*Hirose:* Look at the overall success of Ames tests. *In silico* gives some information sometimes. We can use it on a case by case basis.
*Cronin*: This is not a case of one solution fitting all situations. There are roles for both *in silico* and *in vitro*. Hence the importance of ITS.
*Richard*: You will always need *in silico* where you do not have the chemicals (for example, virtual libraries), but we want to use *both* methods.
*Guengerich*: You might ask whether organic chemists at Cambridge should spend all their time in the laboratory or all their time in the library.
*Greene*: Experimental data from a well validated *biological* assay should always be considered to be more reliable than an *in silico* prediction.
*Richard*: I disagree: it depends.
*Matthews*: in picking the first dose for clinical trials you set up a QSAR and you will get very near, but the result from complex animal analysis is *worse*. Animals have different bioavailability and different metabolism.
*Benigni*: The Ames test is one of the best *in vitro* assays, but results vary from laboratory to laboratory.
*From the Floor*: Carcinogenicity prediction works within an application domain.
*Greene*: Even with one receptor the chemistry "space" being synthesized changes over time and so an *in silico* model fails to predict for new compounds coming through.
*Question from the Floor*: What about metabolism based toxicology?

*Richard*: Some of the ToxCast *in vitro* assays will have metabolic capability.
*Guengerich*: For the FDA you have to do the experiments to find the metabolites, to get round expensive retesting.
*Matthews*: QSAR tools can be used as a method of prioritizing. They can cope with the most likely off-target activities for metabolites as well as for the original chemicals. You confirm your *in silico* observations with wet work.
*Boyer*: You start with a molecule of molecular weight 300 or 400. You scale back the functionality to make the molecule specific but when it is metabolized, more functional groups appear, and the metabolite is therefore less specific.
*From the floor*: Pharma is making molecules live longer in the body. Now we need to predict the effect of the drug on the environment.
*David Hawkins*: Are there examples?
*Benigni*: There is no experience in Europe.
*Matthews*: In pre-manufacture notice in the United States for food you have to use *in silico* methods.
*Richard*: Pre-manufacture notification in the Toxics program requires EPA to make a toxicity determination without data; hence, SAR and *in silico* methods have been necessary and essential to this program. In the pesticide program, in contrast, EPA has had legal authority to require lots of test data; hence, SAR historically has not been used. Congress has recently mandated, however, that pesticides programs evaluate tolerances for impurities in pesticides, without giving them the authority to request new data; hence, the pesticides program now has to look to SAR and *in silico* methods.
*Matthews*: Tier testing can detect the highly toxic compounds so you can make early decisions on some pesticides.

## References

1. Collins, F. S.; Gray, G. M.; Bucher, J. R. Transforming Environmental Health Protection. *Science (Washington, DC, U. S.)* **2008**, *319*, 906-907.

2. Matthews; E. J.; Kruhlak, N. L.; Benz, R. D; Contrera; J. F.; Marchant, C. A.; Yang, C. Combined use of MC4PC, MDL-QSAR, BioEpisteme, Leadscope Predictive Data Miner, and Derek for Windows software to achieve high performance, high confidence, mode of action-based predictions of chemical carcinogenesis in rodents. *Toxicol. Mechan. Methods* **2008**, *18*, 189-206.

3. Insilicofirst. http://www.insilicofirst.com/index.html (accessed January 25, 2009).

4. ESIS http://ecb.jrc.ec.europa.eu/esis/ (accessed January 25, 2009).

5. JRC QSAR tools http://ecb.jrc.ec.europa.eu/qsar/qsar-tools/ (accessed January 25, 2009).

6. *Toxtree*. http://ecb.jrc.ec.europa.eu/qsar/qsar-tools/index.php?c=TOXTREE%20h and http://ecb.jrc.ec.europa.eu/documents/QSAR/EUR_23241_EN.pdf (accessed January 25, 2009).

7. OECD QSAR Toolbox. http://www.oecd.org/document/23/0,3343,en_2649_34377_33957015_1_1_1_37465,00.html (accessed February 12, 2009).

8. Patlewicz, G.; Jeliazkova, N.; Gallegos Saliner, A.; Worth A. P. Toxmatch – A new software tool to aid in the development and evaluation of chemically similar groups. *SAR QSAR Environ. Res.* **2008**, *19*, 397-412.

9. Boyer, S.; Arnby, C.H.; Carlsson, L.; Smith, J.; Stein, V.; Glen, R. C. Reaction Site Mapping of Xenobiotic Biotransformations. *J. Chem. Inf. Model.* **2007**, *47*l, 583–590.

10. Mestres, J.; Martín-Couce, L.; Gregori-Puigjané, E.; Cases, M.; Boyer, S. Ligand-Based Approach to In Silico Pharmacology: Nuclear Receptor Profiling. *J. Chem. Inf. Model.* **2006**, *46*, 2725–2736.

11. DSSTox http://www.epa.gov/ncct/dsstox/ (accessed January 25, 2009).

12. Lazar Toxicity Predictions. http://lazar.in-silico.de/ (accessed January 25, 2009).

13. TheToxcast program. http://www.epa.gov/ncct/toxcast (accessed January 25, 2009).

14. Judson, R.; Elloumi, F.; Setzer, R. W. Li, Z.; Shah, I. A comparison of machine learning algorithms for chemical toxicity classification using a simulated multi-scale data model. *BMC Bioinformatics* **2008**, *9*, 241.

15. Yang, C., Hasselgren, C.H., Boyer, S., Arvidson, K., Aveston, S., Dierkes, P., Benigni, R., Benz, R.D., Contrera, J., Kruhlak, N.L., Matthews, E.J., Han, X., Jaworska, J., Kemper, R.A., Rathman, J.F., Richard, A.M. Understanding genetic toxicity through data mining: the process of building knowledge by integrating multiple genetic toxicity databases. *Toxicol. Mech. Methods* **2008**, *18*, 277-295.

16. Gramatica, P.; Pilutti, P.; Papa. E. Validated QSAR Prediction of OH Tropospheric Degradation of VOCs: Splitting into Training−Test Sets and Consensus Modeling. *J. Chem. Inf. Comput. Sci.*, **2004**, *44*, 1794–1802.

17. Hewitt, M.; Cronin, M. T. D.; Madden, J. C.; Rowe, P. H.; Johnson, C.; Obi, A.; Enoch, S. J. Consensus QSAR Models. Do the benefits outweigh the complexity? *J. Chem. Inf. Model.* **2007***,* *47*, 1460-1468.

18. Abshear, T.; Banik, G. M.; D'Souza, M. L.; Nedwed, K.; Peng, C. A model validation and consensus building environment. *SAR QSAR Environ. Res.* **2006**, *17*(3): 311–321.

19. Trigwell, S. (Ed.) The FRAME/Liverpool John Moores University Defra REACH Project. *Alternatives to Laboratory Animals,* **2008**, *36*, Supplement 1. (Copies available from Mark Cronin.)

20. European Chemicals Agency Guidance, including ITS, can be obtained from http://guidance.echa.europa.eu/ (accessed January 25, 2009).

21. Bruce, C. L.; Melville, J. L.; Pickett, S. D.; Hirst, J. D. Contemporary QSAR Classifiers Compared. *J. Chem. Inf. Model.* **2007**, *47*, 219–227.

22. Ploemen, J. P.; Kelder, J.; Hafmans, T.; van de Sandt, H.; van Burgsteden, J. A.; Saleminki, P. J.; van Esch, E. Use of Physicochemical Calculation of pKa and CLogP to Predict Phospholipidosis-Inducing Potential: A Case Study with Structurally Related Piperazines. *Exp. Toxicol. Pathol*. **2004**, *55*, 347−55.

23. Benigni, R.; Bossa, C.; Netzeva, T.; Worth, A. Collection and Evaluation of (Q)SAR Models for Mutagenicity and Carcinogenicity. 2007. http://ecb.jrc.ec.europa.eu/documents/QSAR/EUR_22772_EN.pdf (accessed January 25, 2009).

24. Ashby J. Fundamental Structural Alerts to Potential Carcinogenicity or Noncarcinogenicity. *Environ. Mutagen* **1985**, *7*, 919-921.

25. Bailey AB, Chanderbhan N, Collazo-Braier N, Cheeseman MA, Twaroski ML. The use of structure-activity relationship analysis in the food contact notification program. *Regulat Pharmacol Toxicol* (2005) 42:225-235

26. Kazius, J.; McGuire, R.; Bursi, R. Derivation and Validation of Toxicophores for Mutagenicity Prediction. *J. Med. Chem.* **2005**, *48*, 312–320.

27. Kazius, J.; Nijssen, S.; Kok, J.; Bäck, T.; Ijzerman, A. P. Substructure Mining Using Elaborate Chemical Representation. *J. Chem. Inf. Model.* **2006**, *46*, 597–605.

28. Benigni, R.; Bossa, C. Predictivity of QSAR. *J. Chem. Inf. Model.* **2008**, *48*, 971-980.

29. Benigni, R.; Bossa, C., Richard, A. M.; Yang, C. A novel approach: chemical relational databases, and the role of the ISSCAN database on assessing chemical carcinogenicity. *Ann. Ist. Super. Sanità* **2008**, *44*, 48-56.

30. Benigni, R.; Andreoli, C.; Conti, L.; Tafani, P.; Cotta-Ramusino, M.; Carere, A.; Crebelli, R. Quantitave structure-activity relationship models correctly predict the toxic and aneuploidizing properties of six halogenated methanes in Aspergillus nidulans. *Mutagenesis* 1993, 8, 301-305.

31. Hayashi, M.; Kamata, E.; Hirose, A.; Takahashi, M; Morita, T.; Ema, M. *In silico* assessment of chemical mutagenesis in comparison with results of salmobella microsome assay on 909 chemicals, *Mut. Res,* **2005**, 588, 129-135.

32. Liebler, D. C.; Guengerich, F. P. Elucidating mechanisms of drug-induced toxicity. *Nat. Rev. Drug Discovery* **2005**, *4*, 410-420.

33. MacDonald, M.L.; Lamerdin, J.; Owens, S.; Keon, B. H.; Bilter, G.K.; Shang, Z.; Huang, Z.; Yu, H.; Dias, J.; Minami, T.; Michnick, S. W.; Westwick, J. K. Identifying off-target effects and hidden phenotypes of drugs in human cells. *Nat. Chem. Biol.* **2006**, *2*, 329 - 337

34. Iconix Biosciences (now owned by Entelos) http://www.iconixbiosciences.com/ (accessed January 25, 2009).

35. MedDRA. http://www.meddramsso.com/MSSOWeb/index.htm (accessed February 12, 2009).