Putting the Predictive Toxicology Challenge into perspective: reflections on the results

Romualdo Benigni* and Alessandro Giuliani

Laboratory of Comparative Toxicology and Ecotoxicology, ‘Istituto Superiore di Sanita’, Viale Regina Elena 299, 00161 Rome, Italy

Received on November 21, 2001; revised on May 21, 2002; accepted on June 3, 2002

ABSTRACT
Motivation: Chemical carcinogenicity is of primary interest, because it drives much of the current regulatory actions regarding new and existing chemicals, and its experimental determination involves time-consuming and expensive animal testing. Both academia and private companies are actively trying to develop SAR and QSAR models. This paper reviews the new Predictive Toxicology Challenge (PTC) results, by putting them into the context of previous attempts.

Results: A marked dependency of the prediction ability of the different algorithms on the training sets was observed, pointing to a still insufficient coverage of the chemical carcinogens ‘universe’. A theoretical treatment of the possible developments of the Artificial Intelligence approaches is sketched.

Contact: rbenigni@iss.it

INTRODUCTION
The environment—in its broadest sense, as opposed to the genetic component—is the major determinant of cancer, and is responsible for at least 50–80% of cancers (Tomatis and Huff, 2001; Tomatis et al., 1997). Among the other toxicological endpoints, chemical carcinogenicity is of primary interest, because it drives much of the current regulatory actions, and its experimental determination involves time-consuming and expensive animal testing. Both academia and private companies are actively involved in the development of structure-activity relationships (SAR) and quantitative structure-activity relationships (QSAR) models for the rodent carcinogenicity among others. For approaches that relied solely on chemical structure information (mainly presence or absence of structural alerts (SA) for carcinogenicity), the overall accuracy in terms of positive or negative predictions was in the range 50–65%, whereas the biologically based approaches attained 75% accuracy. The latter combined the SA’s information with information on experimental results from mutagenicity, acute toxicity, etc… Most of the prediction systems were concordant in the identification of the most powerful carcinogens. The chemicals with mixed carcinogenicity profiles were mostly predicted to be positive. The greatest difficulties were found with the non-carcinogens, since many of them were predicted to be positive by different systems. The results also revealed that the various (Q)SAR approaches essentially acted as gross ‘class-identifiers’, i.e. they pointed to the presence or absence of the alerting chemical functionalities (SA), but were not able to make gradations within each potentially harmful class (Benigni, 1997).

THE TWO NTP PREDICTION EXERCISES
The first comparative exercise regarded 44 chemicals of various classes (Anonymous, 1993). Participating in the exercise were models based on expert systems, mechanistic hypotheses, statistically based approaches, and biological data filtered through human expert judgement, among others. For approaches that relied solely on chemical structure information (mainly presence or absence of structural alerts (SA) for carcinogenicity), the overall accuracy in terms of positive or negative predictions was in the range 50–65%, whereas the biologically based approaches attained 75% accuracy. The latter combined the SA’s information with information on experimental results from mutagenicity, acute toxicity, etc… Most of the prediction systems were concordant in the identification of the most powerful carcinogens. The chemicals with mixed carcinogenicity profiles were mostly predicted to be positive. The greatest difficulties were found with the non-carcinogens, since many of them were predicted to be positive by different systems. The results also revealed that the various (Q)SAR approaches essentially acted as gross ‘class-identifiers’, i.e. they pointed to the presence or absence of the alerting chemical functionalities (SA), but were not able to make gradations within each potentially harmful class (Benigni, 1997).

The second NTP comparative exercise regarded further 30 chemicals (Bristol et al., 1996). In a preliminary analysis of 26 results (out of 30) it was possible to identify three main patterns of predictions. Cluster 1 consisted mainly of human experts: the predictions were mainly elaborated...
Reflections on the Predictive Toxicology Challenge

by scientists who considered, in an informal way, several types of information. (Huff et al., 1996) considered the presence of structural alerts and analogy with known carcinogens and non-carcinogens, previous bioassays when available, subchronic toxicity and relative toxicity, and reported or predicted metabolites. (Benigni et al., 1996) relied mostly on chemical analogy reasoning. (Tennant and Spalding, 1996; Ashby, 1996) relied largely on biological evidence, as interpreted according to their own theories and models. Oncologic is a computerized expert system, but since it does not have rules for all chemical classes, many predictions were formulated by the Oncologic team at the US Environmental Protection Agency (EPA) (Woo et al., 1997). Application of the FALS method involved eight separate QSAR models, the universe of chemicals being divided into eight chemical classes (Moriguchi et al., 1996). Cluster 2 was composed of an experimental system (transformation assay in Syrian hamster embryo [SHE] cells) (Kerckaert et al., 1996), and of two computerised models (decision trees) based on organ-specific toxicity (Lee et al., 1996). Cluster 3 was composed of very diverse methods, among which were some Artificial Intelligence (AI) approaches: MULTICASE (Zhang et al., 1996) and Progol (King and Srinivasan, 1996), which identify the structural alerts in an unbiased way, and the rule-based expert system DEREK (Marchant, 1996).

No prediction approach reached 100% accuracy. However, the trends were much different. Cluster 1 (including many human experts) tended to increase the percentage of chemicals predicted as positive (carcinogens) going from the noncarcinogens to the carcinogens, which is good. However, the rate of false positives was quite high (0.54). Cluster 2 showed a discontinuous trend, together with a similar high percentage of false positives (0.53). Both Clusters 1 and 2 had a good performance in the identification of the carcinogens. The Cluster 3 methods had a puzzling trend, with more positive predictions for the noncarninogens than for the clear evidence carcinogens, and they failed to identify some of the most powerful carcinogens.

A common weakness of the prediction systems was that often the noncarninogens were predicted as carcinogens. Thus, it appears that the prediction approaches were almost invariably unable to make gradations between potential and actual carcinogenicity. The same evidence was shown by the first comparative exercise (Benigni, 2000).

Figure 1 summarizes the accuracy of the individual predictions in the second NTP exercise. Some of the human expert-based predictions (especially those that relied mainly on chemical analogy) performed best (together with SHE). The maximum accuracy attained was around 65–70%. The AI systems generally had quite bad performance, mainly because of their failure in identifying many of the powerful carcinogens.

When judging the above accuracy figures, it should be remembered that the selection process for chemicals tested in the rodent bioassay is biased toward chemicals suspected of potential carcinogenicity (because of obvious practical limitations). For example, out of ~400 chemicals tested by National Cancer Institute/NTP, two-thirds were selected as suspect carcinogens, and one-third on production/exposure considerations. In the first class, 68% were demonstrated to be rodent carcinogens, whereas only 20% of the second class were positive (only 7% were positive in two animal species) (Fung et al., 1995). On a general ground, this demonstrates that the rodent bioassay is not oversensitive and does not appear to ‘produce’ false positives. More specifically, the chemicals considered in

<table>
<thead>
<tr>
<th>Code</th>
<th>System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Huff</td>
<td>(Huff et al., 1996)</td>
</tr>
<tr>
<td>Ben</td>
<td>(Benigni et al., 1996)</td>
</tr>
<tr>
<td>Tenn</td>
<td>(Tennant and Spalding, 1996)</td>
</tr>
<tr>
<td>Ashby</td>
<td>(Ashby, 1996)</td>
</tr>
<tr>
<td>Onc</td>
<td>Oncologic (Woo et al., 1997)</td>
</tr>
<tr>
<td>SHE</td>
<td>SHE transformation assay (Kerckaert et al., 1996)</td>
</tr>
<tr>
<td>R1, R2</td>
<td>decision-trees (Lee et al., 1996)</td>
</tr>
<tr>
<td>CASE</td>
<td>MULTICASE (Zhang et al., 1996)</td>
</tr>
<tr>
<td>Pro</td>
<td>Progol (King and Srinivasan, 1996)</td>
</tr>
<tr>
<td>DEREK</td>
<td>DEREK (Marchant, 1996)</td>
</tr>
</tbody>
</table>

For the definition of a ROC curve, see the site: http://www.informatik.uni-freiburg.de/~ml/ptc/index.html.
the comparative prediction exercises were not a random sample from the universe of chemicals, but were a particularly ‘difficult’ test set, including a majority of suspect structures. Thus, the real goal of the exercise was not the separation of carcinogens from non-carcinogens, but the harder task of separating potential from actual carcinogens. On a more balanced sample of chemicals, the performance of the systems would probably be higher.

THE CHALLENGE OF AN EVOLVING CHEMICAL WORLD

The composition of the database has a crucial importance. For example, the chemicals of the two NTP exercises, about 5 years apart from each other, were quite dissimilar. The first NTP set was largely populated by carcinogens which were also mutagenic. In the second set most of the carcinogens were nonmutagenic. This means that the types of chemicals, and the underlying mechanisms of carcinogenicity were different. In the two exercises, the different performance of the biologically based approaches (quite good in the first one, less good in the second one), that relied also on the information from mutagenicity tests, can be explained by the above difference in composition.

The different composition of the databases is a consequence of the evolving pattern of chemicals in use in the world. The chemicals with a practical role change during the years; new dyes are produced to follow the evolution of fashion and new pharmaceutical drugs are designed to improve health. Many SAs are now well known, and drug designers do not include them in the new molecules. This has the consequence that the training sets of chemicals are always ‘old’ in respect to the test sets of prospective exercises like the NTP ones. It is likely that the temporal gap between training and test sets has a more negative impact on the automatic prediction systems than on the human experts that can exploit a more free access to various sources of information.

ABOUT MODELING THE IN VIVO DATA

A criticism sometimes held is that much noise is carried by the measurement of the biological activity (the rodent bioassay). This criticism maintains that the rodent bioassay has little reproducibility and for this reason is fundamentally out of the reach of modeling. However, various evidence shows that when the rodent data are selected carefully, and when a meaningful analysis is designed, the in vivo data from rodent carcinogenicity experiments can be modeled as efficiently as other biological end-points. A notable example is represented by a paper from the Hanisch’s laboratory relative to the skin carcinogenicity of the aromatic hydrocarbons (Zhang et al., 1992). A second example is from our laboratory, and regards the aromatic amines (Franke et al., 2001; Benigni et al., 2000). We were able to generate separate models for the potency of the carcinogenic aromatic amines, and for the separation of carcinogens from noncarcinogens: in both cases, we attained a satisfactory accuracy (80–90%). This rules out the criticism that the rodent bioassay data cannot be modeled adequately because of their intrinsic limitations.

Since the aromatic amines are the class most represented in the carcinogenicity database, our work is the ideal case for QSAR applications (one congeneric class of chemicals acting through the same mechanism, reliable for the powerful Hansch approach): thus a 80–90% accuracy can be considered as an estimate of the top accuracy reachable. In this perspective, the 65–70% accuracy of the approaches to noncongeneric (and poorly balanced) sets of chemicals used in the comparative exercises is understandable.

THE PTC RESULTS IN PERSPECTIVE

Let us now consider the PTC (http://www.informatik.uni-freiburg.de/~ml/ptc/index.html).

First, the ROC curves relative to the four experimental animal systems (see the above site) show that most of the various AI approaches produced more incorrect predictions for the carcinogens than for the noncarcinogens (left bottom quadrant of the ROC plane). This is the opposite of what happened with the NTP (Fig. 1).

Second, most of the predictions lie along and around the diagonal of the ROC plane, which corresponds to random results. As a matter of fact, the elegant statistical analysis by (Toivonen et al., 2003) demonstrated that, out of 111 sets of predictions, only five models performed better than random guessing at a significance level of 0.05.

Third, the largely random nature of the carcinogenicity predictions is confirmed by the fact that, in a comparative inspection of the four different ROC planes (for the four experimental systems: rat male and female, mouse male and female), the location of the systems varies in the different planes. Among the statistically significant models (Toivonen et al., 2003), there was no approach consistently effective in the four experimental rodent groups.

Overall, this evidence points to a limited performance of the AI approaches in the PTC. On the contrary, some of the approaches challenged in the NTP were remarkably displaced from the diagonal line, heading towards the left top corner of the ROC plane (predominance of correct predictions) (Fig. 1). These were the human experts approaches and the SHE experimental system. Even if a direct comparison of the evidence from the two exercises cannot be done, the suggestion of a better performance of the human experts in respect to more automatized approaches is quite strong. This also comes out from the
evidence from within the second NTP exercise itself, with the poor performance of the participating AI approaches. What can be the origin of the difficulties experienced by AI prediction approaches? The problems can be twofold: first, the representation of the problem; second, the representativity of the learning database. Regarding the first issue, the successful examples of QSAR modeling of rodent carcinogenicity rule out that this is due to an ‘intrinsic’ fallacy of the experimental carcinogenicity data. The representation of the chemical structures may be more problematic, and this point should be investigated.

In our opinion, a key point is the representativity of the learning set in respect to the test set. The chemicals in use change with the time. In the PTC, the difference between the training set and the test set should be remarked. The database used as training set was composed of chemicals tested by the NTP, mostly small molecules, whereas the test set was composed of pharmaceutical drugs from the Food and Drug Administration compilation (hence different chemical classes were differentially represented). The test set was predominantly composed by noncarcinogens, whereas the proportion of carcinogens in the NTP training set was much higher; moreover, many carcinogenic pharmaceuticals are non genotoxic and act through epigenetic mechanisms, whereas the majority of the NTP carcinogens have structural alerts typical of the genotoxic carcinogenicity mechanisms. In our opinion, this difference had a strong impact on the PTC results. Despite the ‘surface’ resemblance of the training (NTP molecules) and test sets as assessed by (Toivonen et al., 2003), we think that the crucial difference between the two datasets should be assessed at the level of the Structural Alerts. The presence or absence in the molecules of functional groups and substructures known to be related to the carcinogenicity mechanisms, has the effect of shifting the individual molecules towards either toxifying or detoxifying cellular pathways. This is the basic reason for the dependence of successful QSARs on the careful selection of molecules belonging to specific, well defined chemical series.

Both the biochemical and QSAR evidence on the carcinogens indicate that each chemical class acts with a different mechanism, and that different descriptors and different models are necessary for each class (Debnath et al., 1994; Benigni and Giuliani, 1996; Cronin and Dardeen, 1995; Hansch, 1991). A suitable learning set should contain a balanced representation of each class of carcinogens, with their noncarcinogenic counterparts. However, the carcinogenicity experimentation for both practical and ethical reasons has generated until now a quite limited database. For example, we have selected from the literature a database of about 800 chemicals tested for rodent carcinogenicity. The composition in the most represented classes was: (a) aromatic amines: 197; (b) nitroaromatics: 32; (c) halogenated alkanes: 27; (d) halogenated alkenes: 9; (e) halogenated alcohols: 8. The consequence is: insufficient information for classifying chemicals according to mechanisms and insufficient representation in chemical classes. The final result is an insufficient learning set: no matter how smart are the AI approaches, they cannot cope with problems without enough theory to formalize rules, or for which adequate learning sets do not exist. We think that the limited representation of the ‘universe’ of carcinogens in the learning set is the main problem faced by the approaches aimed at providing ‘general’ models of chemical carcinogenicity, and that this problem obscures other problems. It is interesting that the human experts, which can exploit simultaneously, in a non-formalized way, different types of information at different hierarchical levels (chemical analogy, presence of SAs, reported or predicted metabolism, knowledge of action mechanisms, etc…) appeared to reach a better performance.

A MORE THEORETICAL INSIGHT

The problem of: (1) if; (2) how; and (3) when automatic prediction methods will reach reliable performance in carcinogenicity estimation allows for a theoretical treatment. Let us start with the ‘if’. The basic condition to be fulfilled for an AI approach to be successful is that the event to be predicted fully depends upon the examined features. That is to say that the tumor burden FULLY depends upon the administered molecule. While this condition is obviously false in general (due to individual susceptibility, intrinsic stochastic character of the event, unobserved boundary conditions…), it can be assumed to be true in relative terms (at least asymptotically) given the adoption of controlled experimental conditions and the statistical character of the observable. What is modeled is not the probability of developing cancer for a single animal, but a statistically significant increase in the number of cancers in a drug-administered group. The good results obtained by QSAR methodologies in the case of organic series, consisting of chemicals supposedly acting through the same mechanism, demonstrates the validity of this hypothesis.

If we can assume that the increase in tumors FULLY depends upon the molecule, a specific representation of the molecule allowing us to predict its carcinogenic potential MUST exist. Carcinogenicity IS one of the properties of the molecule and thus is derivable from the knowledge of the structural and chemico-physical properties.

Given a positive answer to the ‘if’ issue, let us discuss the ‘how’. There is a potentially infinite number of properties with which the molecules are endowed. Some of these are futile (the number of letters of their IUPAC name, the color of the container in which they are stored, etc…). Some of them may be of value for their biological
action (the energy of the highest occupied molecular orbital, the partition coefficient between water and lipid phase, etc...). Without a substantial information loss, we can focus only on the features directly computable from the structural formula: we know from quantum mechanics that in the structural formula are embedded all the relevant information on the behavior of the molecule.

Thus the problem of generating a prediction system is equivalent to finding an algorithm applicable to the structural formula for deriving the ‘carcinogenic’ property. We cannot find this algorithm *ab initio*, because we cannot map carcinogenesis at the quantum chemistry level. Thus we must approach the problem in a statistical-empirical way.

The empirical approach implies the generation of local models for subsets of molecules, and the reconstruction of experimental carcinogenicity by vectors of structure-derived features. The amount of *a priori* ‘chemico-biological’ knowledge embedded in the model qualifies the different approaches: they range from classical QSARs with maximum *a priori* knowledge (Hansch and Leo, 1995) to chemically ‘blind’ models (like most of the AI predictors) based on the automatic derivation of descriptions. The answer to the ‘how’ question is then: ‘Do perform statistical experiments by generating empirical prediction algorithms’.

The existence of a large spectrum of empirical models introduces the ‘when’ issue. The chemico-physical intensive models can work even with relatively small data sets, provided that the starting premises are sound. In contrast, ‘blind’ models need an extensive sampling of the chemical space. Insufficient sampling only reproduces the singularities of the training set, and provides scarce hope for generalization. We have no idea of the extension and density of the space of the molecules necessary for carcinogenicity prediction, so we cannot sketch credible estimates of the numerosity and composition of the training set for an optimal blind approach.

Here, we have two possible avenues for future research: a ‘chemico-physical’ and an ‘information technology’ one. The first implies the replication of classical QSAR models for the individual chemical classes, to cover all the organic chemicals space. This corresponds to replicate the aromatic amines and aromatic hydrocarbons examples for more and more chemical classes. As more QSARs are collected, we can imagine to compare the different models and to derive higher-level, general information (Hansch *et al.*, 1996).

The information technology approach consists in collecting the experimental data on the universe of chemicals, until a ‘blind’ AI efficient prediction model comes out. We have no idea which of the two approaches can reach in a lesser time the final goal of predicting the carcinogenicity from the chemical formula. What we know is that the first approach will IMMEDIATELY start to give us ‘pay-offs’. When a chemical class is modeled via the classical QSAR approaches, the results are immediately applicable to other chemicals of the same class. However this reasoning is based on an ideal situation, with an almost unlimited effort in bioassaying to provide the ‘construction material’ for the predictive models: this does not correspond to reality. The reality is the sporadic experimentation on molecules of diverse origin, that accumulate relatively slowly and with no particular order. In this situation, progressively collecting ALL the bioassay results and generating self-correcting AI systems can be preferable to the classical QSAR strategy. When the ‘critical mass’ (in terms of representativity of the unknown chemical space) of the training set is reached, we expect that an AI ‘blind’ system will start to produce reliable predictions. This AI system, in a first phase, will produce reliable predictions ‘at class level’, by identifying the classes of risk of the chemicals with results similar to those attained by the human experts. This procedure is already able to significantly ‘enrich’ the frequency of positive compounds with respect to an unbiased search (Fung *et al.*, 1995). In a subsequent phase the AI performance will improve, but very slowly with respect to the accumulation of new data. Once the ‘class allocation’ step has been solved, any further improvement depends upon the ability to model the modulation of carcinogenicity INSIDE the individual classes. The determinants of carcinogenicity modulation are not the same in the different classes, thus a general, blind strategy is sub-optimal in this phase. At the (probably very far) ‘thermodynamical limit’ even the most ‘dull’ AI system will reach the goal, because of the dependency of the biological activity from the structural formula.

In the same scenario, the classical QSAR approach will initially remain not applicable until the first well formed ‘class’ appears; then it will rapidly reach the ‘blind’ AI performance, and will overcome the AI systems during their ‘latency’ period after the completion of the ‘gross class identification’ phase. This depends on the above sketched property of QSAR methods to give pay-offs at the end of each ‘completed’ class (see Fig. 2).

**CONCLUSION**

No model can replace the reality. When dealing with chemicals with a large impact on the human population, only experiments can give reliable estimates of the human risk (Haseman *et al.*, 2001; Fung *et al.*, 1995; Huff, 1993). At the same time the structure-activity knowledge has already large room for use. Estimates for individual chemicals cannot be taken at face value: a prediction for an individual chemical can be useful if it complements other available information and if the model provides background information on the reasons of the prediction.
Fig. 2. Expected trends in efficiency increase for the QSAR and AI approaches, at the increase of the coverage of the ‘universe’ of chemical carcinogens. The steps in the QSAR curve correspond to the coverage of individual classes of chemical carcinogens (see details in the text). a: reach of the ‘class allocation’ level of performance; b: completion of the first congenic class of carcinogens. Solid line: QSAR. Dashed line: AI.

The situation is completely different at the statistical level, with large numbers of chemicals. For example the application of SAR concepts in a non formalized way for setting priorities in the NTP experimentation has been quite successful, and the human experts attained a good, non random accuracy in the NTP prediction exercises. This demonstrates that a chemical knowledge basis for developing models, including the AI ones, already exists. More specifically, for the optimal use of the chemical knowledge it is of paramount importance that the AI techniques are not confined only to the formalization of the chemical structures in a purely abstract and neutral way (for example, as mathematical graphs). In contrast, the AI approaches should be challenged also in the formalization of the chemical reasoning, and of the scientific principles developed in the one and a half centuries of progress in organic physical chemistry.

REFERENCES


