WILEY
Interdisciplinary Reviews
COMPUTATIONAL MOLECULAR SCIENCE

Article Type: Advanced Review

In silico toxicology: computational methods for prediction of chemical toxicity

Arwa Bin Raies: Computational Bioscience Research Centre (CBRC), Computer, Electrical and Mathematical Sciences and Engineering Division (CEMCE), King Abdullah University of Science and Technology (KAUST)

Vladimir B. Bajic: Computational Bioscience Research Centre (CBRC), Computer, Electrical and Mathematical Sciences and Engineering Division (CEMCE), King Abdullah University of Science and Technology (KAUST). Email: vladimir.bajic@kaust.edu.sa

Supplementary Information

Brief descriptions of items in Supplementary Information

Table S1. Different types of PK and PD models.

Table S2. Different method to determine uncertainty factors.

Table S3. Different Types of models and molecular descriptors in the QSAR family.

All references are included in the main manuscript

Table S1. Different types of PK and PD models (summarized from [74]).

Model Category	Model Type	Model
Simple direct Effects: Assume direct relationship between concentraion and effect. However, these models assume rapid equilibrium conditions between plasma and biophase concentrations, and consequently assume maximum response occurs at maximum concentration. However, responses maybe delayed.	**Linear or log-linear models:** These models show a linear or log-linear correlation between response and concentration or log concentration respectively. They are easily calculated, but they have limited applicability for extrapolating responses.	$E = E_0 \pm S \times C_p$ $E = E_0 \pm m \times \log C_p$ E: effect or response E_0: baseline effect S: slope of the linear model m: slope of log-linear model C_p: plasma concentration
	Hill models: These models overcome the limitations of linear or log-linear models. This model is based on drug-receptor rate of change, and assumes rapid equilibrium conditions between plasma and biophase concentrations. Also it assumes drug effect is directly proportional to the fraction of occupied receptors density, and agonism is the relevant action of the drug or chemical.	$E = \frac{E_{max} \times C_p}{EC_{50} + C_p}$ E: effect or response E_{max}: maximum effect EC_{50}: concentration producing 50% of E_{max} C_p: plasma concentration
Biophase distribution model: These models are used when chemical experience response delays. It was proposed using a hypothetical effect-compartment to link between time course plasma concentrations with drug effect. These models relate change of biophase concentration with plasma concentration.		$\frac{dC_e}{dt} = k_{1e} \times C_p - k_{e0} \times C_e$ C_e: Biophase concentration C_p: plasma concentration k_{1e} and k_{e0}: first order distribution rate constants t: time
Slow receptor binding model: These models eliminate equilibrium conditions and rapid and reversible response assumptions. Instead, these models take into account slow		$\frac{dE}{dt} = k_{on} \times (E_{max} - E) \times C_p - k$ E: Response k_{on}: second-order association rate constant

rates of association and disassociation. The response is directly related to the concentration.		k_{off}: first-order dissociation rate constant C_p: plasma concentration E_{max}: maximum response t: time
Irreversible Effects: Many chemicals (such as those used in chemotherapy) exhibit responses using irreversible interactions with cells or proteins.	**Cell Proliferation Model with Irreversible Inactivation:** Models for cellular proliferation and phase-nonspecific cell killing for chemotherapeutic drugs.	$\frac{dR}{dt} = g(R) - f(C) \times R$ $g(R)$: density of the viable cells $g(R) = k_g \times R$ Or $g(R) = k_g \times R \times (1 - \frac{R}{R_{ss}})$ $f(C)$: plasma or effect-compartment drug or chemical concentrations that are involved in irreversible interaction $f(C) = k \times C$ Or $f(C) = \frac{K_{max} \times C}{KC_{50} + C}$ R: Response or effect C: either plasma or biophase concentration k_g: first-order cell growth rate constant R_{ss}: upper limit of cell number k and K_{max}: second-order cell-kill rate constants KC_{50}: chemical or drug concentration that produces 50% of K_{max} t: time
	Cell Proliferation Model with Cycle-Specific Inactivation: A two-compartment model to characterize chemotherapeutic drugs that exhibit responses during certain phases of the cell cycle. The model splits the cells	$\frac{dR_s}{dt} = g(R_s) - f(C) \times R_s - k_{sr} \times$ $\frac{dR_r}{dt} = k_{sr} \times R_s - k_{rs} \times R_r$ R_s: response of proliferating cells R_r: response of quiescent cells

	into two groups: proliferating (R_s) and quiescent (R_r) cells.	k_{sr} and k_{rs}: first-order transformation rate constants between the two groups t: time $g(R_s)$ and $f(C)$: as described above
	Turnover Model: Models interaction between some drugs and endogenous enzymes.	$\frac{dR}{dt} = k_{in} - k_{out} \times R - f(C) \times R$ R: response or effect k_{in}: zero-order production rate of the response k_{out}: first-order loss rate constant t: time $f(C)$: as described above
Tolerance Models: Some drugs or chemical exhibit reduction in response after repeated or prolonged exposure.	**Counter-regulation models:** These models use and opposing response that reduce the target response.	$\frac{dM}{dt} = k_1 \times R - k_2 \times M$ R: response or effect M: opposing response k_1 and k_2: first-order constants of production and loss respectively.
	Desensitization Models: These models show a decrease in response on continuous exposure. One modeling technique is to allow the response to be temporarily lost.	$\frac{dR_i}{dt} = k_d \times (R - R_i)$ R: response or effect R_i: inactive response k_d: first order process t: time

Table S2. Different method to determine uncertainty factors.

Uncertainty Factor	Description	Advantages	Disadvantages
Default Uncertainty Factors and sub-factors [88]	$100 = 10_A \times 10_H = (10^{0.6} \times 10^{0.4}) \times (10^{0.5}{}_{PK} \times 10^{0.5}{}_{PD}) = (4 \times 2.5) \times (3.162 \times 3.162)$ 10_A: factor for inter-species differences 10_H: factor for intra-species differences $10^{0.6}$: Toxicokinetic factor for inter-species $10^{0.4}$: Toxicodynamics factor for inter-species $10^{0.5}{}_{PK}$: sub-factor for intra-species Toxicokinetics $10^{0.5}{}_{PD}$: sub-factor for intraspcies Toxicodynamics	It's based on two factors of 10. Each factor accounts for animal-to-animal and human-to-human differences. The sub-factors provide smaller uncertainty factors that account for Toxicokinetics and Toxicodynamics of inter-species and intra-species differences. [88]	The default value is arbitrary, do not represent a worst case scenario. Rather these values represent 'adequate' scenario. Also, it has been shown that the default values are not overly conservative. [88] However, in other cases these values are criticized for overestimating toxicity levels [89]. And the default value of 100 does not cover all inter-species and inter-individual differences. [87]
Body surface area correction [89]	$\left(\frac{\text{human body weight}}{\text{animal body weight}}\right)^{1/3}$	It accounts for differences between human and animal sizes.	
allometric scaling [89]	$\left(\frac{\text{human body weight}}{\text{animal body weight}}\right)^{1/4}$	A more appropriate adjustments based on allometric scaling according to caloric demand or metabolic size. Very useful for cancer risk assessment.	Provides uncertainty factors that are much smaller than 10 for non-cancer endpoints
Probabilistic-based species-dependent default values	It assumes log-normal distribution of safety levels, and uses statistical analysis such as geometric mean and geometric standard deviation to derive the uncertainty factors. Therefore, instead of generating a single uncertainty factor, these methods generate a probability distribution of these values, and certain percentile of the distribution is used [87]. The values of the factors vary depending on the statistical approach and species. Comparison of different values from	It would be more reliable for UFs to be estimated on the basis of actual experimental data and be developed using statistical elements such as the median or geometric mean rather than use conventional default UFs. This	The values vary depending on the statistical approach. Also these methods depend on NOAEL values which can vary based on experimental design. Therefore, it is suggested to use critical effect doses to derive the probability

	different approaches is available in [89]	method accounts for inter- and intra-species differences and PK and PD variations.	distribution. Probability approaches may eliminate over-conservativeness and worst-case assumptions [87].
chemical-specific assessment factors [87]	Use chemical-specific factors from toxicokinetics or toxicodynamics to replace sub-factors.	Estimation of sub-factors is based on scientific data.	It requires determining if acting chemical is the parent or a metabolite, or whether to consider external dose or tissue internal concentration. Such scientific data may not be available or may require in vivo studies on humans.
Pathway-related assessment factors [87]	The factors that account for metabolism and excretion are developed for either inter or intra-species differences	Toxicokinetics data can be found in literature.	

Table S3. Different Types of models and molecular descriptors in the QSAR family.

Models	Types of Descriptors	Advantages	Disadvantages
1D-QSAR	1D descriptors: represent the structure of the chemicals such as atoms Functional groups [19].	Easy to implement and interpret. Can be represented as binary features (0 or 1 for absence or presence of structure) or frequency: the number of chemicals that contain a certain atom [19]. Can be used to identify SAs [6]	Requires feature selection. Also, can be used to predict toxicity of chemicals that have the same structure as the chemicals used to generate the model. When using the frequency encoding, it requires having high frequency to obtain statistical significance. But multivariate regression can be used to avoid this problem. [19]
2D-QSAR	2D descriptors: represent the physico-chemical, physico-biological properties [119] and topological indices [19]	Many physico-chemical and physico-biological properties are useful for understanding ADME properties [19]: • logP (1-octanol/water partition coefficient) : to explain hydrophobicity [19, 5] • Lipophilicity, integrity and stability [10]: Define the pharmaceutical potential of drugs • Permeability and solubility [10]: Useful for gastrointestinal absorption for oral exposure • Other properties include LogD (1-octanol/water distribution coefficient), topological surface area polar surface area (TPSA), oral bioavailability, intestinal absorption, plasma-protein binding, volume distribution, blood-brain barrier and excretion [119]. Also topological indices [19] are useful for conformational independency and simplistic and less time-consuming computation to yield good results especially with small molecules since small changes in structural features can lead to large changes in toxicity [5]. Additionally, Extended topo-chemical atom (ETA) indices [19]	There are a large number of 2D descriptors. It is necessary to ensure that the chosen descriptors are uncorrelated and meaningful. Many calculated 2D descriptors are meaningless although they have a large discriminant power. PCA can be used to find independent descriptors. [19] 2D descriptors provide a flat representation of chemicals, and ignore stereochemistry, and can miss key elements that depend on chirality [18].

		have been shown to include sufficient chemical information and encode chemical structure parameters [5].	
3D-QSAR	3D descriptors: represent filed properties in 3D such as energy fields: steric, electrostatic and hydrophobic. [19]	Allows to understand receptor-ligand interactions	May require alignment of molecular, which is time consuming and requires experience. Therefore, these descriptors are useful for a small number of molecules [18].
CoMFA (Comparative Molecular Field Analysis) [19]	3D descriptors field descriptors: steric (van der Waals) and electrostatic (Coulombic)	Same as 3D-QSAR	Same as 3D-QSAR
CoMSIA (Comparative Molecular Similarity Indices Analysis) [19]	3D descriptors field descriptors: steric, electrostatic, hydrophobic, hydrogen bond donor, hydrogen bond acceptor.	Use of more fields allows for a deeper analysis to emphasize on space regions where the contributions of various fields are necessary for the biological activity (i.e., Toxicity)	Same as 3D-QSAR
COREPA (Common Reactivity Pattern) [19]	Conformational distribution of chemical across local and global reactivity parameters that are linked to the biological activity.	Does not require alignment of molecules.	
Pseudo-3D [19]	Eigen values derived from IR and Raman range molecular vibrational frequencies. Weighted holistic invariant molecular (WHIM) descriptors.	Does not require alignment.	
QSIIR (Quantitative structure in-vitro in-vivo relationship) [8]	Chemical and biological descriptors (from high throughput screening in vitro data)	Uses in vitro data to predict in vivo toxicity	