BTO 2014.009 | April 2014

BTO rapport

Innovative testing strategies and their relevance for evaluating chemical drinking water quality

KWR Watercycle Research Institute

BTO

Innovative testing strategies and their relevance for evaluating chemical drinking water quality

BTO 2014.009 | Maart 2014

Opdrachtnummer
400554/019 (B222016)

Projectmanager
Stefan Kools

Opdrachtgever
BTO- Thematisch onderzoek - Nieuwe meetmethoden en sensoring

Kwaliteitsborger(s)
Kees van Leeuwen

Auteur(s)
Kirsten Baken en Stefan Kools

Verzonden aan
Verzonden naar BTO-participanten en is na een jaar openbaar.

Jaar van publicatie
2014

Meer informatie

T 030 - 60 69 703
E kirsten.baken@kwrwater.nl

PO Box 1072
3430 BB Nieuwegein
The Netherlands

T +31 (0)30 60 69 511
F +31 (0)30 60 61 165
E info@kwrwater.nl
I www.kwrwater.nl

KWR Watercycle Research Institute

Summary

Approaches to evaluate human health hazard of chemical exposure are modernized in order to reduce experimental animal use and investments in time and money, and to increase the predictability of test systems. This report describes the state-of-the-art of *in vitro* and non-testing hazard evaluation strategies. Besides an overview of available tools, the predictive value of the tools, advantages and opportunities that the testing strategies offer as well as their limitations, and required and foreseen developments in this field are addressed.

In vitro tests focusing at toxic events or endpoints in primary cells or cell lines, stem cells, cell fractions, co-cultures, or engineered tissues can predict adverse human health effects of chemicals or chemical mixtures *in vivo*. A large array of *in vitro* assays have been developed. Receptor binding, (reporter) gene expression, protein production and flow cytometry are for instance used as read-outs in these systems. Many of the available *in vitro* toxicity tests are still at the research and development stage. Only a small number of *in vitro* tests addressing specific endpoints have at present been formally validated and gained regulatory acceptance. A limitation of the use of *in vitro* tools for hazard assessment, is that *in vitro* models may poorly resemble their *in vivo* equivalents. One reason for this is the complex interplay between different cell types and the involvement of mediators released by other tissues that occurs *in vivo*. Besides, the toxic outcome of exposure may differ per organ or dose, duration, and timing of exposure. A single *in vitro* test is therefore often a too much simplified picture of reality, and ideally, a test battery of complementary *in vitro* tests should be applied. New developments that may improve the predictive value of *in vitro* assays are the inclusion of metabolic capacity, realizing more physiologic culture conditions, producing organotypic 3D (co)-cultures, using cells from human origin and stem cell-derived systems, better understanding of mechanism of action thereby refining and expanding endpoints measured, and appropriate statistics and prediction models. The increasing knowledge of biological systems and advances in molecular and computational tools, coupled with the concomitant development of high-throughput and high-content screening assays (such as transcriptomics), are providing mechanistic insight into the mode of action of chemicals. This enables establishment and application of 'Pathways of Toxicity' and 'Adverse Outcome Pathways', which may aid in the prediction of human health hazard. There are currently however no reliable ways to quantitatively extrapolate dose–response data from concentrations of the test substance in *in vitro* experimental systems to toxic exposure levels in the whole body.

Non-testing methods are based on the principle that the activity of a chemical can be predicted from its molecular structure and substructure, and inferred from the physicochemical properties and biological effects of similar substances. (Quantitative) Structure Activity Relationships or (Q)SARs are theoretical models that predict the physicochemical, biological (e.g. toxicological) and environmental fate properties of molecules from the knowledge of the chemical structure. (Q)SARs are more prevalent for endpoints for which large databases exist, such as ecotoxicity, mutagenicity and carcinogenicity, skin sensitisation, and endocrine disruption. Chemical categorizing is an approach in which only prototypic compounds out of a group of similar ones are tested. Chemicals whose physicochemical and toxicological properties or environmental fate properties are likely to be similar or follow a regular pattern as a result of structural similarity are in this case grouped. Read-across techniques fill data gaps on a specific

chemical by interpolating or extrapolating existing data of related chemicals within a category of substances. Physiologically based toxicokinetic (PBTK) modeling predicts the toxicokinetics of chemicals based on physiological and chemical parameters. Recent developments in computing power, the ability to create extensive databases and the use of the internet to compile, organise and distribute information, have increased the capability to investigate relationships between chemical structure and biological activity. For all the so called *in silico* approaches, the size and quality of the underlying databases and availability of physicochemical parameters are critical for reliable predictions. For complex endpoints, the models are still not overly realistic because they may ignore essential processes. Just like *in vitro* tools, increased understanding of mechanisms of toxicity will improve the quality of the predictions made by *in silico* models. The Organisation for Economic Co-operation and Development (OECD) has developed criteria for the validation of *in silico* methods.

The information produced by the different innovative hazard assessment approaches can be combined to enable further development of the tools. Moreover, the tools can be of added value when multiple information sources are systematically combined in integrated testing strategies (ITS). ITS approaches are useful when not all possible outcomes of interest, classes of test substances, or severity classes of effect are covered in a single test. ITS are also valuable when the human predictivity of a single test is not satisfactory. In addition, a tiered ITS approach provides the opportunity to combine existing data with new data, and to filter out certain substances before costly tests or animal tests are performed. By using a 'Weight of Evidence' approach, different pieces of evidence and test data be weighed and combined. Several recommendations are derived from the information gathered:

1. A selection of complementary tools that assess a series of health outcomes relevant for exposure to drinking water contaminants should be combined.
2. Multiple non-testing approaches should be used for quick hazard assessment, prioritization of compounds for further testing and to provide mechanistic information.
3. Immediate access to and familiarity with non-testing tools should be established. Currently, the OECD QSAR Toolbox and EPISUITE are non-commercial standard applications that are freely available and with high use in regulatory settings, but many other commercial and non-commercial tools have been developed. Expertise is needed to perform and evaluate predictions derived from *in silico* tools.
4. Besides, *in vitro* assays should be used to confirm non-testing information and to demonstrate potential biological effects of unidentified substances; in both cases water sources that need further examination can be prioritized.
5. The potential to evaluate a large range of substances at low dose levels and the predictive value for human health hazard should be taken into account when selecting *in vitro* assays
6. Development of guidelines for ITS should be closely watched.

The combination of rapid assessment by carefully selected up-to-date tools and expert judgement will provide a more reliable and relevant estimation of the potential hazard of currently present and emerging compounds in the (aquatic) environment.

Contents

Summary 2

1 Introduction 5

2 *In vitro* toxicity testing 6
2.1 Repeated dose toxicity 7
2.2 Carcinogenicity 12
2.3 Developmental toxicity 20
2.4 Toxicokinetics 24

3 Non-testing approaches 30
3.1 Structure activity relationships 30
3.2 Chemical categorizing and data gap filling 33
3.3 PBTK modelling 37
3.4 Expert systems 39

4 Opportunities and pitfalls of innovative testing strategies 47
4.1 *In vitro* toxicity testing 47
4.2 Non-testing approaches 53
4.3 Integration of testing strategies 57

5 Recommendations for assessment of water quality 62

References 65

1 Introduction

Chronic exposure to environmental contaminants may result in adverse health effects in humans. Evaluation of human health hazard of chemical exposure is a prerequisite for risk assessment. In the area of toxicological risk assessment, classic hazard evaluation is performed using experimental animal studies, which identify adverse health outcomes and safe dose levels of substances. In order to reduce experimental animal use and investments in time and money, and increase the predictability of test systems, alternative approaches are sought to evaluate toxic effects of chemicals. This shift in hazard assessment practices is advanced by the large number of chemicals that needs to be evaluated as a result of the Regulation on Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) and by the ban on animal testing for cosmetics according to the EU Cosmetics Directive.

There is international consensus that modern tools for toxicity testing should be based on mechanisms of toxicity and mode of action. *In vitro* tests focusing at specific toxic events or endpoints in cultured cells, tissues, or microorganisms can predict adverse human health effects of chemicals or chemical mixtures *in vivo*. Physiologically based toxicokinetic (PBTK) models describe the expected toxicokinetic properties of compounds. Chemical categorizing, read-across techniques, and (Quantitative) Structure Activity Relationships or (Q)SARs allow identification of biological reactivity of chemicals without the use of laboratory experiments to reveal biological effects ('non-testing'), by using physicochemical properties and known effects of similar substances.

In vitro assays and non-testing strategies are at present applied to a limited extent as hazard assessment tools in the evaluation of chemical drinking water quality. The state-of the-art and expected developments of these approaches are outlined in this report, as well as the potential applications for identification of potential human health effects of chemical drinking water contaminants when limited toxicological information is available. The aim of this report is to provide an overview and selection of *in vitro* and non-testing strategies for the joint research program of the drinking water companies in the Netherlands (BTO). The focus lies on drinking water exposure and the report therefore predominantly includes chronic toxicity, mutagenicity, carcinogenicity, and reproductive toxicology in humans.

The following two chapters summarize available *in vitro* and non-testing strategies, respectively. This is followed by a discussion of the predictive value of the tools, the advantages and opportunities that the testing strategies offer as well as their limitations, and the required and foreseen developments in this field. The information in these sections was obtained from recent reviews and governmental documents on relevant topics, and inquiry of practices at diverse institutions. The report is concluded with recommendations for the application of innovative testing strategies within the drinking water production sector.

2 *In vitro* toxicity testing

In vitro assays study effects of chemical exposure on the molecular (binding of or damage by a chemical) or cellular (altered protein production or signalling) level, thereby predicting adverse health outcomes *in vivo*. The majority of the *in vitro* models are based on dispersed cell cultures, either as primary cell cultures (derived directly from living tissue) or as continuous cell lines (tumour or artificially immortalized cells). In addition, *in vitro* tests applying stem cells (undifferentiated cells) are currently evaluated. Cells applied in *in vitro* systems may originate from many different species, including humans. The use of extracellular support systems such as scaffolds and extracellular support matrix improves survival of cells in culture. Other approaches include the use of microporous supports (mono-cultures or co-cultures) in combination with perfusion systems to provide an organotypic environment that can improve differentiation and lifespan of cells in culture. The various types of *in vitro* models available are listed below (Balls et al. 2012).

Varieties of *in vitro* models

Cell fractions (including postmitochondrial supernatant [S9], cytosolic [S100] and microsomal fractions for biotransformation studies)
Primary cell monolayer or suspension cultures
Continuous cell lines
Immortalized cell lines
Stem cells
Genetically-engineered cells
Co-cultures
Organotypic cultures
Precision-cut slices
Perfused cultures
Reconstituted tissue equivalents
Engineered tissues
Dynamic bioreactors
Multi-organ systems
Cell-/organ-/human-on-a-chip

In the context of the ban on animal testing for cosmetics and cosmetic ingredients by the 7th amendment to the EU Cosmetics Directive, an EU expert panel produced a review on the status and prospects of alternative methods for toxicity testing (Adler et al. 2011). The European Centre for the Validation of Alternative Methods (ECVAM), hosted by the Institute for Health and Consumer Protection of the European Commission's Joint Research Centre (JRC), coordinated this activity. The evaluations of *in vitro* toxicity assays assessing repeated dose toxicity, carcinogenicity, developmental toxicity (all relevant for chronic exposure to low doses), and toxicokinetics are adopted from this review. These endpoints are most relevant to chronic exposure to low levels of contaminants in drinking water. Unless indicated otherwise, all information in this section is subtracted from the ECVAM review.

2.1 Repeated dose toxicity

The term repeated dose toxicity comprises the general toxicological effects (excluding genotoxic, carcinogenic, and developmental effects) occurring as a result of repeated daily exposure to a substance for a part (sub-chronic exposure) or the major part (chronic exposure) of the lifespan of an organism. The onset and progression of the toxicity is influenced by the interplay between different cell types, tissues, and organs, including the concomitant contribution of toxicokinetics (see section 2.4), hormonal effects, autonomic nervous system, immune system and other complex systems. In some cases, toxic effects may be due to a build-up of toxic substance(s) in one or more sensitive areas of the body. In other cases, defence mechanisms may be exhausted, the tissue may be altered by regulations and counter-regulations, or immunological reactions involving the specific or non-specific immune system may be triggered (Basketter et al. 2012).

In current Organisation for Economic Co-operation and Development (OECD) guidelines for repeated dose toxicity testing in animals, a comprehensive series of about 30 tissues and organs is examined histopathologically, in addition to gross examination, clinical signs, clinical chemistry, and haematology. However, some tissues are much more frequently targets of toxicity than others. The targets of most concern are the heart, liver, kidney, lung, nervous system, and reproductive system. Other targets of potential concern include the endocrine system, which covers many different organs and systems and will overlap with the reproductive system, the immune system, haematological system, including bone marrow, the musculoskeletal system, and the gastrointestinal system. To date, *in vitro* methods for repeated dose toxicity have been developed mainly with the aim of producing stand-alone methods predicting effects in specific target organs. No formally validated methods are available at present.

2.1.1 Hepatotoxicity

A wide variety of *in vitro* systems is available for toxicity testing with regard to the liver, ranging from **subcellular hepatocyte fractions** to **whole isolated perfused livers**. Endpoints that are often assessed include measurement of apoptosis (programmed cell death), necrosis (premature cell death), cholestasis (accumulation of bile), steatosis (deposition of fat), phospholipidosis (accumulation of phospholipids), and fibrosis (formation of excess fibrous connective tissue). A prerequisite for repeated dose toxicity testing is that the method accurately and consistently predicts long-term effects. However, many of the hepatic cellular models undergo progressive changes in the functional and morphological phenotype, rendering them applicable for only short-term purposes. In the last decade, several innovative strategies have been introduced to counteract this dedifferentiation process, including **genetic** and **epigenetic modulations**. A notable exception in this respect is the human hepatoma-derived HepaRG **cell line**, which persistently displays high functionality.

> ***Epigenetic modulations*** *Changes in gene activity that are not caused by changes in the DNA sequence*

Traditional high-order systems, such as **precision-cut liver slices** and isolated perfused livers, are still widely used for toxicity testing. Meanwhile, there has been considerable focus in recent years on the development of **miniaturized bio-artificial liver devices** and **perfused bioreactors**: primary hepatocytes or liver-based cell lines cultivated on microelectronic sensors, micropatterned, or microfluidic systems that mimic the cellular microenvironment (also referred to as 'lab on a chip') (**Figure 1**). A parallel track that has been followed in the last few years concerns the *in vitro* differentiation of hepatocyte-like cells from **stem cell** sources from different species.

> ***Stem cell*** *An unspecialized cell characterized by the ability to self-renew by cell division while in undifferentiated state, and the capacity to give rise to various differentiated cell types*

This research area is still in its infancy, but given the exponentially growing interest, it can be anticipated that this field will be fully exploited in the upcoming years. Significant progress has also been made lately with respect to the refinement of read-outs for toxicity testing in the available liver-based *in vitro* models, especially by combining them with **gene and protein expression** analysis.

> ***Gene expression*** *The conversion of the information from the gene into messenger RNA via transcription and then to protein via translation resulting in the phenotypic manifestation of the gene.*

The potential use of all of these approaches for repeated dose toxicity testing is unclear because the interaction between different cells of the liver is hardly addressed, which might be of particular importance for assessing the capacity of the organ to regenerate after initial damage or for assessing adaptation processes.

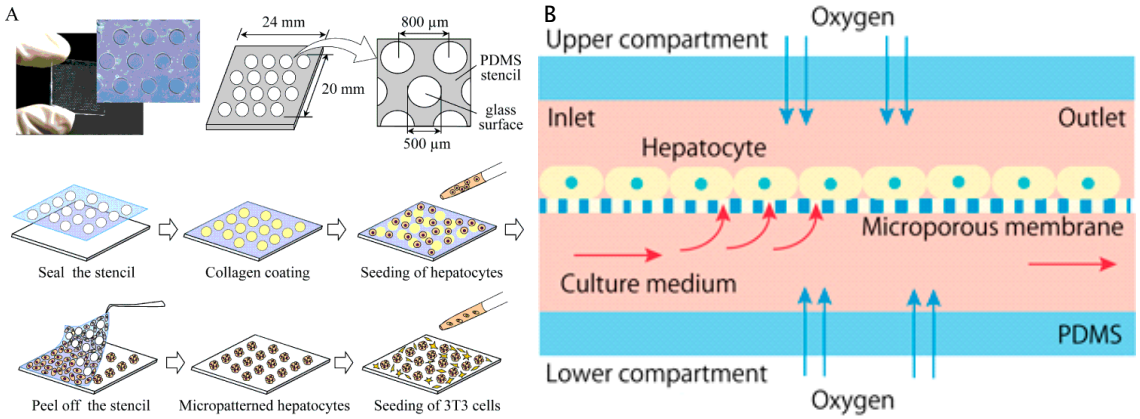

Figure 1. Reconstitution of tissue-like conditions to culture cells in physiologically relevant conditions, in order to maintain hepatic functions. (A) Preparation of a micropatterned hepatocyte co-culture, (Nakazawa et al. 2011) (B) Schematic representation of a microfluidic device in which hepatocytes cultures on a membrane are perfused for active supplement of nutrients and removal of waste material (www.microfluidics.iis.u-tokyo.ac.jp).

2.1.2 Nephrotoxicity

The kidney is a highly complex organ, composed of many different cell types, and it has a complex functional anatomy (**Figure 2**), which renders the assessment of kidney function difficult to explore in *in vitro* studies. The most frequently used techniques are **renal slices**, **perfused nephron segments**, **isolated tubules** as well as **isolated tubular cells in suspension**, longer-lasting **primary cell cultures of isolated tubular cells**, and **cell lines**. Primary proximal tubular cells from rodents and humans are widely used as a renal cell model and many improvements in culture techniques have been reported. In addition, a number of cell lines have been developed and used, most of which are derived from proximal tubular epithelium. They do not usually exhibit biotransformation of chemicals, but they retain at least some capacity for the active transport of chemicals. Renal epithelial cell phenotypes with extensive glycolytic metabolism and morphology very close to the *in vivo* parent cell type are available, e.g. RPTEC/TERT1 cells, and can be maintained in culture up to 6 weeks. Critical **molecular pathways** associated with toxicity have been identified, such as the Nrf2 antioxidant and detoxification pathway in human HK-2 cells. *In vitro* models for

other segments of the nephron are limited. No model is available for assessment of the potential for toxicity to the kidney medulla.

Molecular pathway *A series of actions among molecules in a cell that leads to a certain end point or cell function*

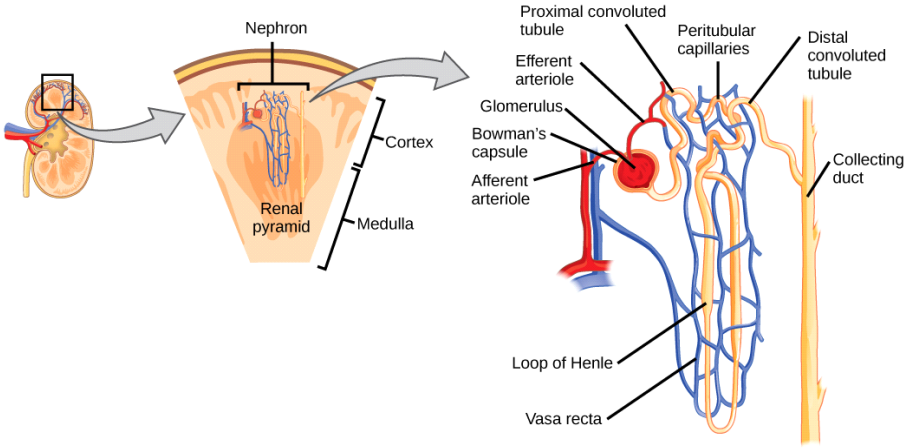

Figure 2. Schematic representation of kidney anatomy (OpenStax College "Biology" Connexions 2013, http://cnx.org/content/col11448/1.9).

2.1.3 Cardiovascular toxicity

Primary cardiomyocytes (cardiac muscle cells) have been isolated from a variety of animal species. An advantage of using these cells is their expression of all the ion channels underlying the cardiac action potential. Adult cells, however, have a low proliferative capacity. Foetal and neonatal cardiomyocyte isolations may contain additional cell types such as fibroblasts, which will overgrow the cardiomyocytes after a few days in culture. Techniques for the long-term culture of primary mouse cardiomyocytes are therefore refined. Cardiomyocytes have also been derived from **embryonic stem (ES) cells**. To assess cardiac-specific toxicity, the effects of compounds in ES cell-derived cardiomyocytes can be compared to those in non-cardiac cells such as fibroblasts. Markers of cardiac damage released by ES cell-derived cardiomyocytes can be identified, and electrophysiological behaviour can be assessed. In addition, efforts are ongoing to develop **three-dimensional tissue engineering models** of cardiac tissue. Engineered heart tissue models derived from neonatal rat cardiac myocytes and human ES cells are commercially available. Many of the *in vitro* assays for cardiovascular toxicity focus on detection of the ability to block the hERG potassium channel, including **receptor binding assays**, **ion efflux assays**, and **patch clamp studies**. **Pressurized human arteries** have been used for the investigation of microvascular dysfunction including vascular permeability and flow-mediated dilatation. Overall, *in vitro* tests for hazard identification in the field of arrhythmia (irregular heartbeat) are in a relatively advanced stage, whereas fewer methods are currently available to evaluate the potential for compounds to cause contractility toxicity, ischemic effects (reduced blood supply), secondary cardiotoxicity, and valve toxicity. Methods sufficient for quantitative risk assessment of repeated dose cardiac toxicity are not yet available.

2.1.4 Neurotoxicity

Detection of neurotoxicity induced by chemicals represents a major challenge due to the physiological and morphological complexity of the central (CNS) and peripheral (PNS) nervous system. The CNS (brain and spinal cord) is comprised of various cell types (**Figure 3**), complex cell–cell interactions, and unique protein interactions where functional coupling via synapses (junctions between neurons), gap junctions (intercellular connections), signalling

molecules, and growth factors has to be preserved. At present, validated *in vitro* methods for neurotoxicity that can provide quantitative predictions for use in risk assessment are not available, although there are methods which are used for screening purposes. Several *in vitro* systems from single cell types to systems that preserve some aspects of tissue structure and function are currently available for toxicity testing. Dissociated cell cultures are more accessible and easier to obtain and maintain than more complex systems. The **primary cell culture monolayer** is the most widely used *in vitro* system for neurotoxicity evaluation. Neuronal cells and glial cells (such as astrocytes, that provide nutrients and structural support, and oligodendrocytes, that form myelin: an insulating layer around nerves) are often mixed. The cell cultures allow visualization of individual living cells (neuronal and glial) and monitoring of both morphological and electrophysiological features. Additional purification methods can be used to enrich a particular cell type. However, *in vivo*–like structures cannot be achieved by this technique. **Reaggregate cultures** (or explant/slice cultures) offer a more structured, three-dimensional histotypic organization that more closely resembles *in vivo* conditions for cell growth and development. Processes such as formation of synapses, myelination, neurotransmission (transmission of nerve impulses across synapses by a chemical substance), and vesicular release of neurotransmitters (**Figure 4**) are the most classical endpoints studied in **3D models**. *In vitro* models of the blood–brain barrier (BBB) are also available, as it is necessary to define whether a compound crosses the BBB and whether it induces a direct toxic effect on the BBB.

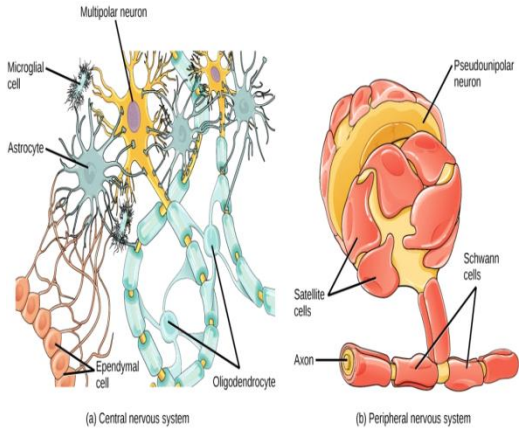

Figure 3. Schematic representation of the cell types of the central nervous system (OpenStax College, Connexions. Biology, 2013, http://cnx.org/content/col11448/1.9).

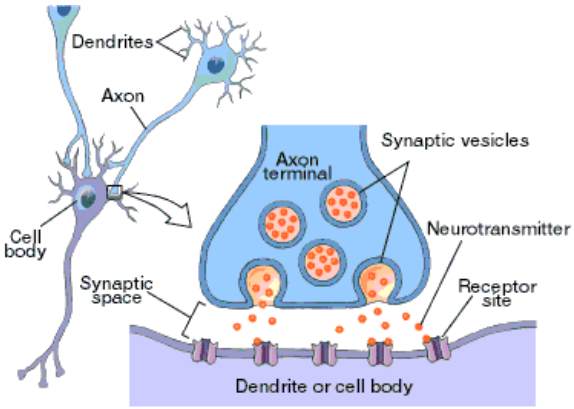

Figure 4. A simplified representation of the process of neurotransmission (Prentice Hall. Psychology: An Introduction, 11/E. Morris, C.G. and Maisto, A.A. http://cwx.prenhall.com/bookbind/pubbooks/morris5).

Continuous **cell lines** of tumoural origin (e.g. SH-SY5Y or PC12) provide homogeneous cell populations (neuronal or glial) in large quantities in a very reproducible manner. However, neuronal–glial cell interaction is lost. Many of these cell lines display properties of their normal cell counterpart when a differentiated state is induced by nerve growth factor (NGF), retinoic acid, etc. Recently, different types of **stem cells** (either adult, embryonic, or those derived from cord blood) are used as a source of neural progenitor cells that can be differentiated into functional neuronal and glial cells with possible applications for neurotoxicity testing. The advantage of this approach is that it is a human model and can be maintained in culture at different developmental stages: as non-differentiated stem cells, committed progenitors, and lineage directed into neuronal, astrocytic, and oligodendroglial cells. **Protein expression** specific for each cell type can be used to evaluate the cell population.

With regard to the PNS (the ganglia and the peripheral nerves), also various cell culture models exist including primary culture and cell lines. However, only limited aspects of functioning of Schwann cells (that surround neural cells and cause myelination of the PNS, similar to oligodendrocytes in the CNS) can be examined.

These *in vitro* approaches allow the assessment of cell viability, general but critical cell functions such as energy metabolism, oxidative stress and calcium homeostasis, and neuronal specific functions. The latter include neurite outgrowth and axonal transport, synaptogenesis, myelination, neurotransmission (assessed by for example up or down regulation of neurotransmitter receptors and activity of acetylcholine esterase) and vesicular release, signalling between neurons and glia, receptor pharmacology, ion channel activation, electrical activity (assessed by micro-electrode arrays, MEA), etc. Based on such models, a complex *in vitro* testing strategy needs to be developed with a battery of complementary endpoints using high-throughput and/or high content screening platforms to be able to test large number of substances with different mechanisms of toxicity.

> ***High-throughput screening*** *Rapid testing of large numbers of compounds for selected activities or interactions with specific proteins, receptors or other cell components*
>
> ***High-content screening*** *Valuation of multiple biochemical and morphological parameters in cells*

However, before these tests can be used for routine screening, the sensitivity, specificity, and reliability of the endpoints and models, and their capacity to predict human neurotoxic effects should be established. *In vitro* methods are not relevant for behavioural studies, which can at present only be performed using non-mammalian models such as zebra fish, medaka or *C. elegans*.

2.1.5 Immunotoxicity and myelotoxicity

Immunotoxicity is defined as the toxicological effect of substances on the functioning of the immune system and can be induced in either direct (caused by the effects of chemicals on the immune system) or indirect (caused by specific immune responses to the compounds themselves or to self-antigens altered by these compounds) ways. The different cells of the immune system are depicted in **Figure 5**. The general view is that effects on the immune system are very difficult to reproduce *in vitro*, because of the requirement of complex cellular interactions. However, some isolated processes may be studied *in vitro* such as proliferation of T lymphocytes and release of cytokines (signalling molecules secreted by lymphocytes). *In vitro* models could be used for pre-screening of immunotoxic potential (i.e. hazard identification), as part of a strategy. The first tier will consist of measuring myelotoxicity (bone marrow suppression). At present, a scientifically validated human and murine *in vitro* colony-forming unit granulocyte/macrophage (CFU-GM) assay is available for evaluating the potential myelotoxicity of chemicals. Toxic effects on **proliferation and differentiation** of progenitor cells of different blood cell lineages can also be measured *in vitro*, as well as long-term repopulating capacity of more primitive haemopoietic stem cells and the stromal microenvironment. If the compound is not myelotoxic, it can be tested for lymphotoxicity (toxicity towards white blood cells involved in adaptive immune responses) in the second tier. Several *in vitro* assays for lymphotoxicity exist, each comprising specific functions of the immune system, B- and T-cell proliferation, cytokine production, cytotoxic T-cell activity, natural killer cell activity, antibody (or immunoglobulins, involved in recognition of foreign molecules) production, and dendritic (antigen presenting) cell maturation.

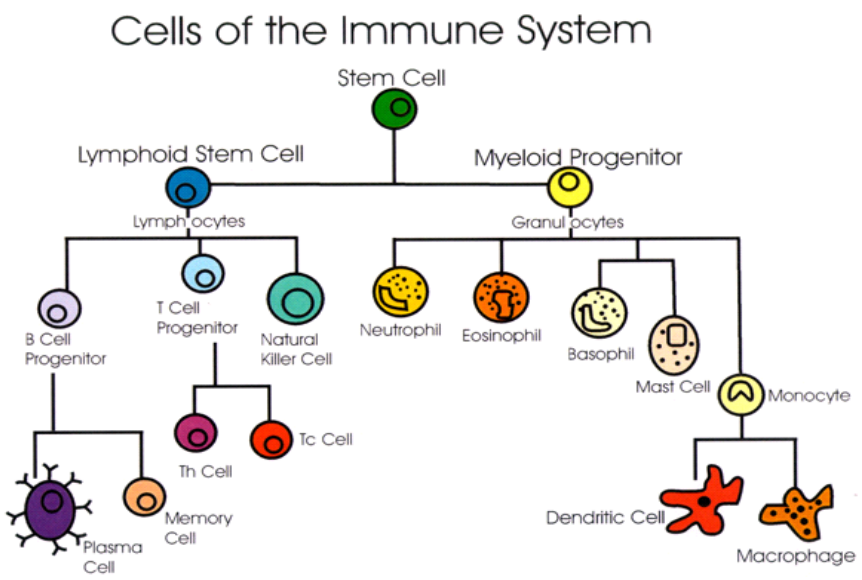

Figure 5. Cell types involved in the immune response (Todra's Online Textbook of Bacteriology, www.textbookofbacteriology.net).

2.2 Carcinogenicity

Substances are defined as carcinogenic if they induce (malignant) tumours, increase their incidence or malignancy, or shorten the time of tumour occurrence. Carcinogens have conventionally been divided into two categories according to their presumed mode of action: genotoxic carcinogens that affect the integrity of the genome by interacting with DNA and/or DNA replication processes and thereby altering structure, information content, or segregation of DNA; and non-genotoxic carcinogens that exert their carcinogenic effects through other mechanisms. Carcinogenesis is a long-term, highly complex process that is characterized by a sequence of stages, complex biological interactions, and many different modes of action. Besides, it is strongly influenced by factors such as genetics, age, diet, environment, hormonal balance, etc. It is recognized that even for one chemical substance, the mode of action can be different in different target organs, and/or in different species. Such complex adverse effects are to date neither fully understood nor can they be completely mimicked by the use of non-animal tests. Some *in vitro* test are available, however, that investigate key events in the carcinogenic process. Available *in vitro* tests focus on hazard evaluation only and cannot currently be used to support a full safety assessment with adequate dose-response information. It is expected that an integrated approach involving multiple *in vitro* models will be needed, but a better understanding of the carcinogenic process is required first. However, for some chemical classes, the available non-animal methods might be sufficient to rule out carcinogenic potential in a Weight of Evidence approach (see section 4.3).

2.2.1 Genotoxic carcinogens

The rationale behind using genotoxicity tests for identifying potential carcinogens is that gene mutations (**Figure 6a**), chromosomal aberrations or breakage (clastogenicity) (**Figure 6b**), and chromosome loss or gain (aneuploidy) are strongly associated with the process of carcinogenesis. A mutation is defined as a permanent change in the amount or structure of the genetic material in a cell. *In vitro* genotoxicity tests which measure a mutation endpoint (gene or chromosomal) are the gene mutation test in bacteria (OECD Test Guideline 471), the gene mutation test in mammalian cells (OECD Test Guideline 476), the chromosome aberration test (OECD 473), and the *in vitro* micronucleus test (OECD Test Guideline 487). Of

these tests, only the *in vitro* micronucleus test is formally validated by ECVAM, but these are nonetheless all established and scientifically accepted tests used for screening and hazard identification.

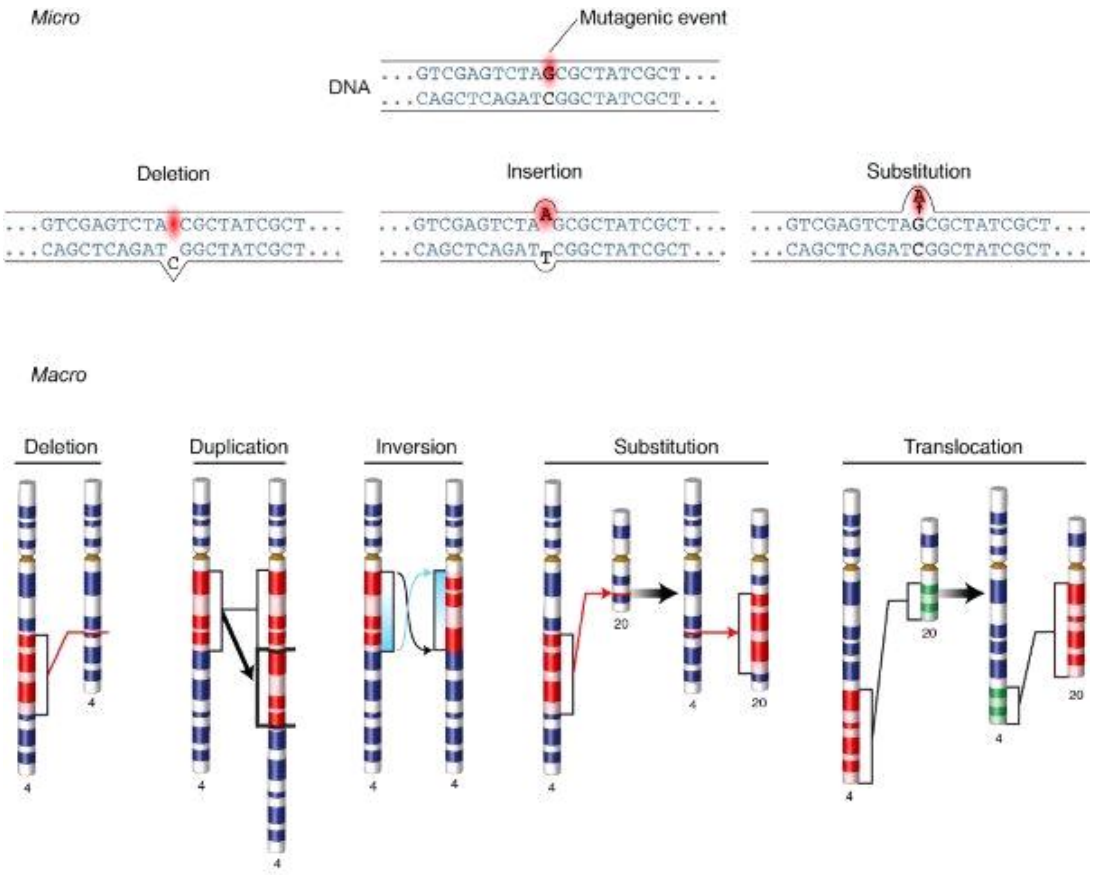

Figure 6. Schematic representation of (a) gene mutations and (b) chromosomal mutations (www.genome.gov).

The **bacterial reverse mutation assay** in *Salmonella typhimurium* and *Escherichia coli* allows detection of base pair substitutions and frameshift mutations. This test detects gene mutations which revert mutations present in the test strains and restore the functional capability of the bacteria to synthesize an essential amino acid. Besides the classic variant of the test in which bacterial colonies are grown and counted on a culture plate, a modern variant in which the bacteria grow in suspension and that includes a **pH indicator** to assess colony formation (the Ames fluctuation test) has been developed (Reifferscheid et al. 2012). An alternative test using *Salmonella typhimurium* is the **umuC assay**, which measures a wider range of genotoxic effects in a single bacterial strain by assessing the SOS repair response. This response is induced as a consequence to DNA damage or disruption of DNA replication and measured by expression of a **reporter gene** (**Figure 7**) (Reifferscheid and Heil 1996).

> ***Reporter gene*** *An endogenous or introduced gene that produces a measurable product in response to the toxic insult*

A comparable assay is the rapid and more sensitive **Vitotox** test (van der Lelie et al. 1997). Another bacterial gene mutation test, also making use of a reporter gene, is the **MutaGen**

test. This test rapidly detects reversions of inactivating base-pair substitutions and frameshifts in a TEM-1 class A beta-lactamase (ampicillinase) gene (Reifferscheid et al. 2005).

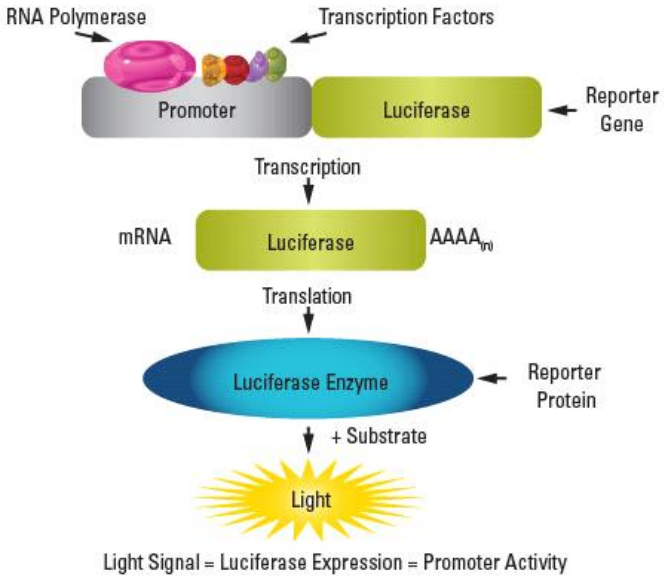

Figure 7. Example of a reporter gene. The reporter gene is attached to a regulatory sequence of the gene of interest. When the regulatory sequence (e.g. promotor) is activated, both genes are transcribed into mRNA and the accompanying protein is formed. The reporter gene product can easily be detected, for instance by emission of light (www.piercenet.com).

The *in vitro* **gene mutation test in mammalian cells** can include different endpoints to detect different spectra of genetic events. The most commonly-used endpoints assess mutations at thymidine kinase (TK) and hypoxanthine-guanine phosphoribosyl transferase (HPRT) gene, and a transgene of xanthineguanine phosphoribosyl transferase (XPRT). Structural and numerical chromosome damage is identified when mouse lymphoma L5178Y cells are used. The endpoint in this **mouse lymphoma assay** (MLA) is the acquired **resistance** to trifluorothymidine as a result of a forward mutation at the TK locus. The HPRT enzyme is important for DNA synthesis. The **HPRT forward mutation assay** involves exposure to the nucleoside analogue 6-thioguanine (6-TG), which is toxic to cells with intact HPRT, while mutant cells survive and form colonies.

The *in vitro* **chromosome aberration assay** in mammalian cells is used to measure structural and numerical chromosomal damage in mammalian cells (i.e. clastogenicity and polyploidy). The *in vitro* **micronucleus assay** in mammalian cells detects structural and numerical chromosome damage as well. Breakage of chromatids or chromosomes can result in micronucleus (a small nucleus that forms when a chromosome (fragment) is not incorporated into one of the daughter nuclei during cell division) formation if an acentric fragment is produced; therefore, assays that detect either chromosomal aberrations or micronuclei are considered appropriate for detecting clastogens. The micronucleus assay however offers advantages over the chromosome aberration assay, since it is less expensive and time-consuming, requires less investment in training, allows a greater statistical power, and has the potential to enhance the basic package of *in vitro* tests to detect aneugens (micronuclei can also result from lagging of one or more whole chromosome(s) at anaphase). Recently, micronuclei counting has been automated by high-content image analysis and flow cytometry platforms, such as the **MicroFlow®** flow cytometric MNT (www.litronlabs.com; Westerink et al. 2011). An alternative test for measuring clastogenicity is the **RadarScreen**

assay, which is a RAD54 promoter-linked beta-galactosidase reporter assay in yeast. The expression of beta-galactosidase is quantified by use of the substrate d-luciferin-o-beta-galactopyranoside, which is converted into galactose and **luciferin** that can be measured luminometrically (Westerink et al. 2009).

> ***Luciferin*** *Organic substance that produces practically heatless light when undergoing oxidation, e.g. such as promoted by the luciferase enzyme*

Other *in vitro* genotoxicity assays for which OECD guidelines have been established are the *in vitro* **sister chromatid exchange** test (OECD Test Guideline 479) and the **DNA damage and repair, unscheduled DNA synthesis** *in vitro* (OECD Test Guideline 482) in mammalian cells, and gene mutation assay (OECD Test Guideline 480) and mitotic recombination assay (OECD Test Guideline 481) in *Saccharomyces cerevisiae* (yeast). These assays are currently much less used since other assays offer several advantages, and their OECD guidelines will be removed in 2014.

Although a number of well-established and regulatory accepted *in vitro* genotoxicity tests are in place, a caveat to the use of these tests is the relatively low specificity and high rate of misleading positive results (i.e. not indicative of an increased cancer risk associated with DNA reactivity, as generally assumed from these tests). The combination of *in vitro* genotoxicity tests yields a high sensitivity of the test battery (to up to 90%), but the specificity (ability to identify non-carcinogens) has been shown to be very low (down to below 25%). The development of new *in vitro* tests with better predictivity is in focus. Examples are the already mentioned Vitotox and RadarScreen assays, which have been shown to yield a low number of false positive results (Westerink et al. 2009). Improved predictivity may also be achieved by choice of **cell or tissue models with higher relevance** (e.g. DNA repair capacity and human origin), use of **cell proliferation** as a measure of cytotoxicity, and applying lower maximum concentrations of test substances.

The *in vitro* **Comet assay** is used for hazard identification and clarification of positive responses in other assays. This assay is based on **single cell gel electrophoresis**, which detects DNA strand breaks on the level of the individual cell, and can be applied in mammalian cells. It is currently in the process of validation by ECVAM. Preliminary review of experimental data has shown it to be more specific than other *in vitro* genotoxicity tests and less prone to false positives (Burlinson 2012). The *in vitro* Comet assay has also been automated to provide a high-throughput screening method (Bajpayee et al. 2013).

Another promising system as a follow-up for confirmation of positive *in vitro* results is the **hens' egg test** for micronucleus induction (HET-MN). The HET-MN combines the use of the commonly accepted genetic endpoint ''formation of micronuclei'' (related to clastogenicity and aneuploidy) with the well characterized and complex model of the incubated hen's egg, which enables metabolic activation, elimination and excretion of xenobiotics, including those that are mutagens or promutagens. At present, only few laboratories have established this test for screening purposes. Up to now, the transferability and intra-laboratory reproducibility is not provided. An improvement may be the inclusion of **flow cytometric** analysis where higher cell numbers can be evaluated in a shorter time and which could improve the sensitivity of the assay as the sample size can be dramatically increased.

The **GreenScreen HC** is a commercially available assay for genotoxicity testing, using human lymphoblastoid TK6 cells transfected with the GADD45a gene (related to growth arrest) linked to a **green fluorescent protein** (GFP) reporter. This assay is based on the p53 tumour

suppressor gene dependent upregulation of GADD45a-GFP transcription and the subsequent increase in fluorescence. The p53 gene is upregulated in response to genome damage and genotoxic stress and controls the cell cycle, DNA repair, and apoptosis. Standard protocols have been developed, and the transferability, within-laboratory reproducibility, and between-laboratory reproducibility have been evaluated. In the absence of metabolic activation, detection takes place using a microplate spectrophotometer, which enables automation, whereas in the protocol with metabolic activation by S9-liver fraction a flow cytometer is used in order to avoid interference with the light absorbing and fluorescent properties of S9-particulates. A variant of the S9-protocol has been developed, which was adapted to microplate readers by the use of a fluorescent cell stain and fluorescence (instead of absorbance) measurement to estimate cell number. This approach is however less sensitive than the use of flow cytometry (Birrell et al. 2010; Hastwell et al. 2006). The **BlueScreen HC** is a more recently developed assay that uses the same GADD45a reporter gene as the GreenScreen HC assay but linked to Gaussia luciferase gene, which leads to a greater signal-to-noise ratio than with GFP. Moreover, it has full compatibility with S9-liver fraction use and thus with high-throughput screening capability (Hughes et al. 2012).

Another assay that studies p53 activity in mammalian cells is the **p53 CALUX** (Chemical Activated Luciferase gene eXpression) reporter gene (www.biodetectionsystems.com). Several other assays that examine the transcriptional response to genotoxic stress (e.g. p21, Nrf2, and AhR gene activity) are being developed as well (Hendriks et al. 2012; Mizota et al. 2011; Westerink et al. 2010). Sensitivity and specificity of such assays have been reported to be higher than those of the Ames test.

2.2.2 Non-genotoxic carcinogens

Non-genotoxic carcinogens act through disturbance of normal physiological control of cellular proliferation, differentiation, and survival by a variety of mechanisms, which are listed below (Adler et al. 2011; Basketter et al. 2012).

Modes of action of non-genotoxic carcinogens

Block of gap junctional intercellular communication
CYP450 enzyme induction
Chronic cell injury
Endocrine disruption
Epigenetic silencing; DNA methylation
Estrogenic activity
Immunosuppression
Inflammation
Inhibition of apoptosis
Reduced immune surveillance
Increased secretion of trophic hormones
Oxidative stress; reactive oxygen and nitrogen species
Peroxisome proliferation
Regenerative cell proliferation
Receptor activation
Signal transduction
Stimulation of angiogenesis

For non-genotoxic chemicals, the standard approach for risk assessment has been to assume a safety threshold based on the results of repeated dose toxicity studies. Despite the fact

that some of the major mechanisms behind non-genotoxic carcinogenicity are known, multiple unknown mechanisms of action and insufficient knowledge of the cellular and molecular events have not yet allowed for the implementation of a battery of *in vitro* tests that could predict and/or explain their carcinogenic potential to man. Moreover, the mechanisms by which non-genotoxic carcinogens cause tumours are in most cases tissue- and species-specific.

There is a need for the development of alternative methods for the detection and risk assessment of non-genotoxic carcinogens. Ideally, these alternatives should include individual endpoints that are typically targeted by non-genotoxic carcinogens, such as the induction of **oxidative stress** and the inhibition of gap junction **intercellular communication**. Currently, however, these tests are research methods primarily used for evaluating mechanism, and cannot yet be used to predict carcinogenicity and potency of substances. Non-animal testing methods for quantitative assessment of non-genotoxic carcinogenic risks are thus limited to non-testing tools.

2.2.3 Genotoxic and non-genotoxic carcinogens

Cell transformation has been suggested as a key event in carcinogenesis reflecting mutagenicity and some initial effects on cell replication, reduced apoptosis, DNA repair, oncogene activation, suppressor gene inactivation, and epigenetic effects (Basketter et al. 2012). Mammalian **cell transformation assays** (CTA) may be used to assess the ability of chemicals to induce changes in the **morphological** and **growth** properties of cultured mammalian cells that are presumed to be similar to phenotypic changes that accompany the development of neoplastic or pre-neoplastic lesions *in vivo*. Widely used cells include SHE, C3H10T1/2, Balb/3T3 and Bhas 42 cells. The SHE and Balb/c 3T3 CTA are now validated by ECVAM and should be developed into OECD Test Guidelines. The SHE and BALB/c 3T3 assays had a strong ability to detect rodent carcinogens, with a good positive and negative predictive capacity and sensitivities and specificities in the 80% range (Basketter et al. 2012). The CTA tests assess changes in cell colony morphology and monolayer formation after exposure of cultured cells (**Figure 8**). Less widely used systems exist which detect other **physiological** or morphological changes in cells following exposure to carcinogenic chemicals. **Cytotoxicity** is determined by measuring the effect of the test material on colony-forming abilities (cloning efficiency) or growth rates of the cultures. When primary cells are used that possess intrinsic metabolic activity, additional metabolic activation is not needed. Some other cell types and substances may require an appropriate external metabolic activation system. The scoring of transformed colonies and foci may require some training and experience. Certain improvements for investigating the transformed phenotype have been proposed, such as discrimination of the transformation phenotype by using ATR-FTIR (attenuated total reflectance fourier transform infrared spectroscopy) **spectroscopy**, **automated image analysis**, addition of **metabolic competence**, and the inclusion of molecular biomarkers.

> ***Molecular biomarker*** *A biological indicator that signals a changed physiological state, stress, or injury due to disease or the environment*

Gene and **protein expression** analysis are included for mechanistic investigation of cellular transformation and also the throughput has been increased by using soft agar colony screening and Bhas 42 96-well plate method. The cell transformation assays are currently used for clarification of *in vitro* positive results from genotoxicity assays. Cell transformation assays are to date the only *in vitro* tests for detection of both genotoxic and non-genotoxic carcinogens that have reached a level of standardisation. Importantly, these assays are currently useful only as one among a larger set of tools in the hazard identification of

carcinogens; there are no methods yet to use data from these tests to support a human risk assessment.

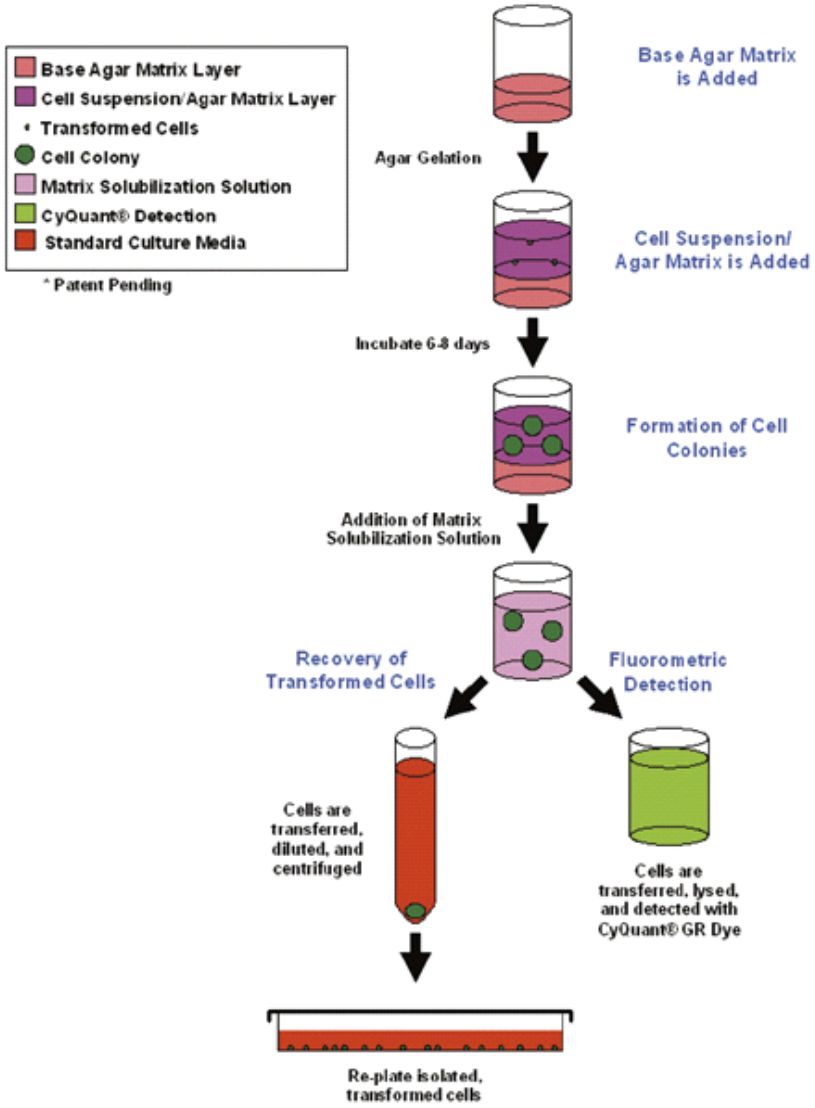

Figure 8. Cell transformation assay principle (www.cellbiolabs.com).

In vitro gene expression approaches provide a wealth of information on the molecular processes that are affected by chemical exposure. Test systems based transcriptomics (**Figure 9**) for genotoxicity and carcinogenicity in general, or for specific mechanisms therein, can be foreseen in the near future.

> ***Transcriptomics*** *The study of the complete set of RNA transcripts of the genome ('transcriptome'), produced by a cell across a variety of biological conditions*

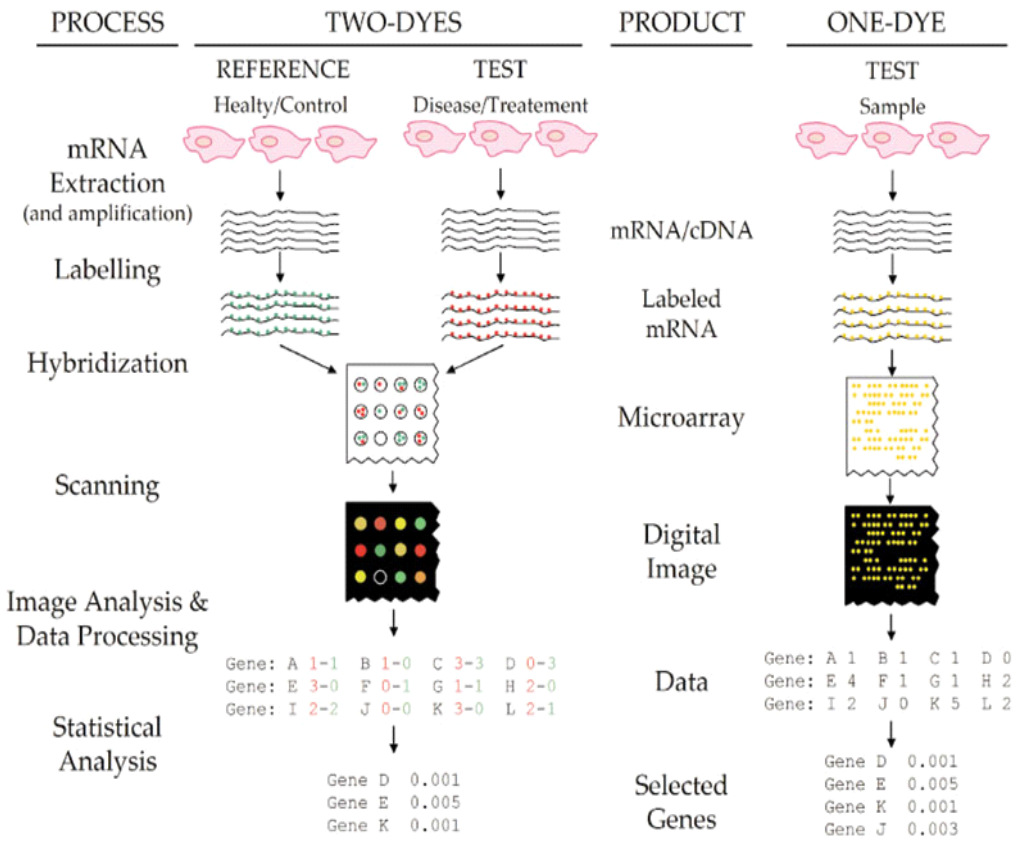

Figure 9. Schematic representation of the transcriptomics technique. The level of gene activation is represented by expression of mRNA molecules, which are labelled in different colours in the sample exposed to the test chemical and a relevant reference sample. During incubation of the samples on a test slide onto which gene fragment have been attached (a 'microarray'), the mRNA fragments will bind to their corresponding gene fragments. Each spot represents a different gene. Imaging techniques show which genes were expressed in which of the samples (http://www.landesbioscience.com/curie/chapter/4407).

Transcriptomics analyses in the area of genotoxicity and carcinogenicity mostly make use of TK6, HepG2, and primary liver cells (possessing intrinsic metabolic activity). As part of the EU-funded project ''carcinoGENOMICS'' several *in vitro* models for liver, lung and kidney are compared. Results so far have indicated that the most uniformly expressed pathway following genotoxicant exposure is the p53 pathway and its subsequently induced networks. The HepaRG **cell line** was the best performing human liver-based *in vitro* model showing few misclassifications, and human **embryonic stem cell**-derived hepatocyte-like cells were shown to have the potential to serve as an *in vitro* model for hazard assessment of chemical carcinogenesis. Furthermore, the novel human RPTEC/TERT1 cell line was found to maintain excellent characteristics of the proximal renal tubule and showed normal chromosomes and nuclear stability, and a successful prevalidation study for identification of (non-)genotoxic carcinogens using this cell line was performed (www.carcinogenomics.eu). Transcriptomics biomarkers will be complex, consisting of profiles for multiple genes. Gene expression analysis by **quantitative RT-PCR** of a relevant gene set derived from transcriptomics data, has been shown to be capable of distinguishing compounds that are DNA reactive genotoxins from those that are non-DNA reactive genotoxins. RT-PCR provides a cheaper and faster test for gene expression profiling, when relatively small gene sets are involved.

2.3 Developmental toxicity

Available assays for screening of developmental toxicity have been reviewed in a recent report (Baken 2013). The information in this paragraph was adopted from this report. The information has been derived from the ECVAM review by Adler et al. (2012), unless indicated otherwise.

Developmental toxicity refers to adverse effects on the developing organism that are induced prior to conception, during pregnancy, or postnatally up to the time of sexual maturity (EPA 1991). The developing human is often more sensitive to toxic insults than adults because of the rapid (cell) growth, development of organs and tissues, increased permeability of skin and gastrointestinal tract, high consumption of air, water and food per unit body weight, and metabolic differences (WHO/EEA 2002). Effects of exposure to agents during critical exposure windows can be reversible or irreversible and may appear at any point in the life span of the organism. The major manifestations of developmental toxicity include mortality (such as implantation loss, resorption, abortion or neonatal death), dysmorphogenesis (structural anomalies like malformations of the skeleton), alterations to growth (generally retardation), and functional impairment (any alteration of normal physiologic or biochemical function) (EPA 1991; FDA 2011). *In vitro* developmental toxicity tests are unable to cover all aspects of the reproductive cycle (e.g. implantation, embryogenesis, fetogenesis, postnatal development, and sexual maturation) and typically study prenatal effects of substances in pre-implantation, peri-implantation and post-implantation stages of development. The biological complexity of these tests ranges from primary and immortalized cell cultures, organ and tissue cultures, and whole embryo cultures using intact, viable conceptuses explanted from a variety of species (Harris 2012a).

2.3.1 Embryonic toxicity

The rodent post-implantation **whole embryo culture** (WEC) is a standardised and validated ex vivo test that uses intact embryos from gestation day 10-12 and covers the critical developmental phase of organogenesis. This allows the study of malformations as they may occur in real life. A series of well-defined **morphological** endpoints are used, which result in a Total Morphological Score (TMS). Besides, malformations and **size** measurements (yolk sac diameter, head length, and crown-rump length) are noted. This assay is currently extended with **gene expression** analysis to improve predictability (Harris 2012b).

The **zebrafish embryo teratogenicity** (ZFET) assay in *Danio rerio* investigates effects on the developing vertebrate organism, **lethality**, **malformations** and **growth retardation** being the primary endpoints. In this test, fertilised fish eggs are exposed to a test substance in the presence or absence of a metabolic activation system. The development of the zebrafish embryo is very similar to embryogenesis in humans, and the test has been reported to demonstrate correct classifications of teratogens and non-teratogens. A draft OEDC guideline is currently under review.

The **frog embryo teratogenesis assay xenopus** (FETAX) is a whole embryo screening assay is based on the South African clawed frog *Xenopus laevis*. Fertilised eggs in the mid- to late-blastula stage are incubated in the presence of the test substance for 96 hours. Lethality, growth retardation and malformations are evaluated at different time points. FETAX requires further development to improve predictability.

In the **chicken embryotoxicity screening test** (CHEST), compounds are administered to windowed eggs, and effects on the developing embryo such as **mortality**, malformations, **embryo development**, **blood vessel development** and **blood vessel coloration** are investigated. The chick embryo possesses its own basic metabolic capacity providing the

possibility to screen for effects of metabolites. This test has been criticised for not being able to distinguish general toxicity from specific developmental effects, a high rate of false positives especially among irritant and corrosive substances, and the absence of mammalian maternal–foetal relations.

The **micromass test** (MM) is a validated test that comprises cell cultures of limb bud and/or neuronal cells isolated from mid-organogenesis embryos. These cells undergo differentiation into chondrocytes and neurons *in vitro* without additional stimulation. **Cell staining** after exposure to test substances shows cell viability and differentiation.

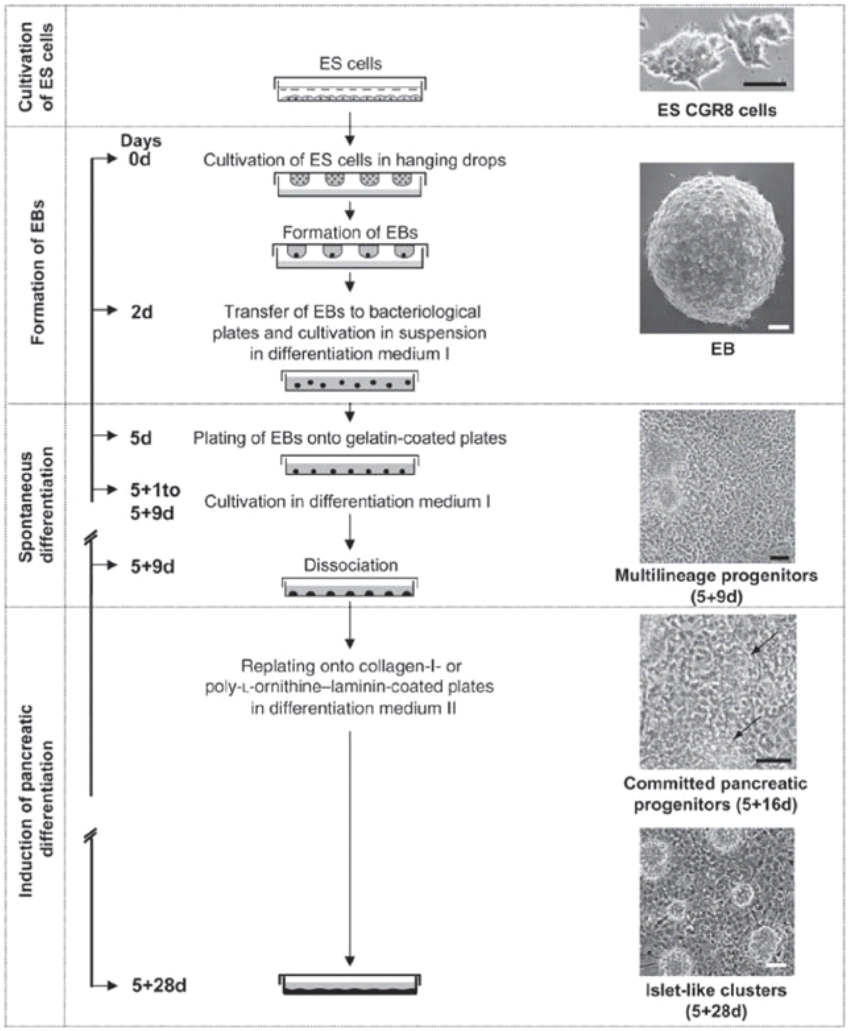

Figure 10. Schematic representation of the differentiation of embryonic stem cells. Undifferentiated embryonic stem cells are cultured under controlled conditions to allow differentiation into specific cell types, in this example pancreatic cells (Schroeder, S. et al. 2006 Differentiation of mouse embryonic stem cells to insulin-producing cells. Nature Protocols 1, 495 - 507).

Embryonic stem cells are pluripotent, which means that they may differentiate into all cell types of the mammalian organism. In the **embryonic stem cell test** (EST), adverse effects on differentiating murine embryonic stem cells, cultured in a hanging drop method to form embryoid bodies (**Figure 10**), are measured as cytotoxicity and differentiation inhibition. The test was validated by ECVAM, but has nevertheless been shown to have insufficient predictive capacity. The EST is however used to screen chemicals for effects on

differentiating cardiomyocytes, neural cells, and skeletal cells. The quantity of differentiated target cells can be evaluated by **flow cytometry**, **reverse transcriptase PCR** or **transcriptomics**. **Genetically engineered** embryonic stem cells, in which for instance **reporter genes** are incorporated that indicate disruption of key toxicity pathways, are also developed. An example is the **ReProGlo** assay that detects alterations in the canonical Wnt/β-catenin signaling pathway, which is involved in the regulation of early embryonic development, in a high-throughput 96-well microtiter format (Uibel et al. 2010). Besides applications for screening of substances for hazardous effects, the EST in combination with gene expression information is also useful for gaining mechanistic insight and identification of predictive biomarkers. Establishment of stable differentiation protocols is a challenge that still needs attention (Schulpen and Piersma 2013; Theunissen and Piersma 2012). Protocol standardization and adaptation, such as the use of a 96 well plate culture format and use of alternative endpoints, are in progress (Theunissen and Piersma 2012).

Ongoing elaboration of alternative testing strategies for developmental toxicity takes place within the ChemScreen and ReProTect collaborative research projects funded by the European Union (Piersma et al. 2013; Schenk et al. 2010) and U.S. EPA's ToxCast program (Sipes et al. 2011). Efforts are made to apply transcriptomics approaches for developmental toxicity testing, and potential biomarkers of developmental effects (such as changes in **hormone concentrations**, **oxidative stress** variables, changes in **gene expression levels**, and **epigenetic changes**) are sought (Louisse et al. 2012; Robinson and Piersma 2013; Schoeters et al. 2011). A key development would be to translate the EST to **human stem cells** (Basketter et al. 2012).

2.3.2 Placental toxicity and transport

Ex vivo **placental perfusion assays** can be performed to assess the potential of compounds to be transported from maternal blood to the embryo or foetus. After birth, the placenta can be used in this test for several hours. The application of this assay is however restricted due to placenta to placenta variations, the limited relevance of the placenta at term for the period of embryonic development, and the complexity of the assay which renders it hard to apply in routine testing of high numbers of samples. The **trophoblast cell assay** which uses the BeWo cell line which represents an immortalised trophoblastic line of human origin, is also useful in transport studies on the rate-limiting barrier to maternal–foetal exchange.

2.3.3 Preimplantation toxicity

Male fertility can be assessed in several *in vitro* tests. The **computer-assisted Male fertility Computer-assisted sperm analysis** (CASA) monitors effects of chemicals on spermatozoa. **Viability**, **motility**, **velocity**, **motion**, and **morphology** of mammalian semen are analysed in real time, which allows the detection of reversible and irreversible damage to the mature sperm as well as repeated dose effects. The test has been evaluated by several independent laboratories. The lower sensitivity of mature sperm in comparison with earlier stages of spermatogenesis may limit the relevance of this test. When Leydig cells are harmed, development of spermatozoa may decrease. In an assay including the **Leydig cell systems**, **cytotoxicity** and **testosterone** production are evaluated. Cytotoxicity towards primary rat **Sertoli cells** and **cells line SerW3**, and secretion of **inhibin B** are indications of blood-testes barrier passage and testicular toxicity. The **ReProComet** assay (Repair Proficient Comet assay) was developed to detect chemically induced DNA strand breaks in bull sperm cells. After incubation with the test chemicals for two hours, sperm cell viability and **DNA damage** are assessed.

Female fertility can be assessed using the **female fertility follicle culture bioassay** (FBA), which still needs further standardization. Mouse ovarian pre-antral follicles are grown *in*

vitro for 12 days until the preovulatory stage, followed by *in vitro* ovulation induction and mature oocyte retrieval. Effects of chemicals on the different biological processes of folliculogenesis, steroidogenesis and oogenesis are analysed with **morphological**, **biochemical** and **functional** parameters. The **bovine oocyte maturation assay** (bIVM) screens for potential adverse effects on the process of oocyte maturation. The endpoint is the successful achievement of the maturation stage. This test has been evaluated by several independent laboratories. The **bovine fertilization test** (bIVF), that focuses on the use of bovine oocytes and sperms for toxicity testing during the process of *in vitro* fertilization, is still in a very early phase of development.

The **mouse peri-implantation assay** (MEPA) allows studying the effect of compounds on the development of the pre-implantation embryo and its capacity to survive. Mouse zygotes are cultured in groups of 10 for 7 days with daily observation and scoring of embryo **development**. The bioassay has high intra-laboratory reproducibility but requires further standardization.

2.3.4 Endocrine disruption

Although validated test methods do not exist for all endocrine disrupting effects, many scientific tools and laboratory methods are available. Assessment of the competitive binding of substances to hormonal receptors (isolated from tissues or recombinant proteins) is one of the tools used to detect endocrine disruption potency. The regulatory acceptance of oestrogen and androgen **receptor binding tests**, which, among others, predict disruption of development of secondary sexual characteristics, are in preparation. Progesterone receptor binding assays have been developed in order to assess effects that might have an influence on the menstrual cycle, the pregnancy and/or embryogenesis. Tests monitoring the binding of the thyroid hormone triiodothyronine (T3) to its receptor, which is highly relevant for the development of the central nervous system, are in the development phase. The same is true for tests studying the binding of hormones produced by the hypothalamus (gonadotropin releasing hormone) or pituitary gland (follicle-stimulating hormone, luteinizing hormone), which are involved in the feedback loop controlling the reproductive system.

In vitro **proliferation** of cell lines containing oestrogen receptors ER-α and ER-β after chemical exposure is used as a measure of estrogenic activity. An example is the **MCF-7 cell proliferation assay**. The **Ishikawa cell test** aims to identify chemicals with estrogenic activity in human endometrial adenocarcinoma Ishikawa cells by assessing the **expression** of the embryo-implantation-associated progesterone receptor (PR) gene. This test is still in the phase of development.

OECD Test Guideline 455 comprises the **stably transfected human oestrogen receptor-α transcriptional activation assay** for detection of estrogenic agonist-activity of chemicals. This assay provides mechanistic information and can be used for screening and prioritization purposes (OECD 2006). Activation of **luciferase** gene expression is used as a readout of hER-α activation in a genetically engineered human cell line. Such so-called **transcriptional activation** assays are able to distinguish between agonist and antagonistic effects of substances, in contrast to the receptor binding tests. These assays make use of cells that express hormone receptors as well as hormone responsive genes and **reporter genes** and exist for various hormones. The **BG1Luc oestrogen receptor (ER) transactivation (TA) agonist** and **antagonist assays** are validated for identification of substances that induce or inhibit human ER activity *in vitro* (OECD Test Guideline 457). Other advanced ER transcriptional assays are the **MELN** and **ER-α-CALUX** assay which both make use of human cell lines, and yeast oestrogen and androgen screens (**YES** and **YAS**, respectively). The CALUX assay generally seems more sensitive than the yeast assays. Anti-estrogenic activities can

also be mediated through the activation of the aryl hydrocarbon receptor. A Japanese **stably transfected transcriptional activation** (STTA) assay for the detection of androgenic and anti-androgenic activity of chemicals is under consideration by OECD. Other transcriptional assays that assess for example progesterone transcriptional activity or interaction with the thyroid receptor are in their early phase of development.

Finally, *in vitro* cell-based assays aiming to detect substances that affect the **synthesis of the sex steroid hormones** are being developed, which is challenging due to involvement of a complex target enzyme and multiple receptors. An assay using H295R cells, designed to measure effects on estradiol and testosterone production, has been validated and is described in the OECD Test Guideline 456. Other steroidogenesis assays based on various **cell lines** derived from rat, mouse and human **tissues**, as well as **genetically engineered cells** and **primary Leydig cells** are developed.

2.4 Toxicokinetics

Toxicokinetics refers to the penetration into and fate within the body of a toxic substance, including the possible emergence of metabolites, and their absorption, distribution, metabolism, and excretion (ADME) (**Figure 11**).

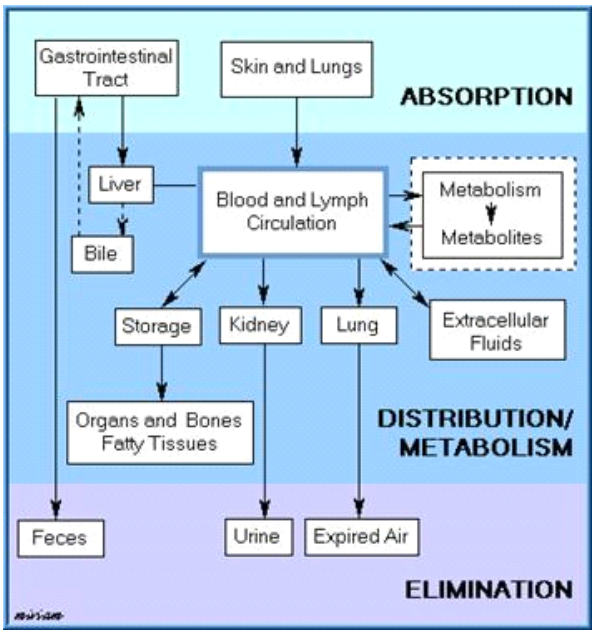

Figure 11. Toxicokinetic processes (www.biology.iupui.edu/biocourses).

As a key starting point for any toxicological testing, it is essential to know whether a compound and/or its metabolites will be bioavailable by one of the relevant uptake routes. Only in cases where a substance is bioavailable following dermal, oral, or inhalation exposure, further tests on systemic toxicity will be necessary.

> ***Bioavailability*** *The degree to which a substance becomes available to the target tissue after administration*

When developing *in vitro* studies, toxicokinetics becomes crucial for translating the observations *in vitro* to the human in vivo situation. ADME processes are lacking in vitro, which may lead to misinterpretation of test result data. In addition, the relationship between the effect of the parent compound and/or the metabolites on the *in vitro* test system and the health effect of interest must be defined in order to derive a relevant *in vitro* (no) effect

concentration. Moreover, the dose level causing no adverse effects *in vitro* needs to be converted to a predicted *in vivo* dose. Applied concentrations in *in vitro* culture media may greatly differ from the actual intracellular concentration due to altered bioavailability (interactions with the medium, the plate, the cell itself) or to physiological cellular processes (mechanism of transport across the membranes, biotransformation, bioaccumulation). This information is required upfront and can be obtained from toxicokinetic alternative methods that identify the main metabolites and the clearance rates of the parent compound and/or its metabolites.

Toxicokinetic studies can make use of the OECD Test Guideline 417 for toxicokinetics testing, which also comprises *in vitro* (e.g. use of **liver cell fractions** or **cell lines** to address metabolism) and *in silico* (computational modelling for the prediction of systemic exposure and internal tissue dose) methods. A whole array of *in vitro*/*in silico* methods at various levels of development is available for most of the steps and mechanisms which govern toxicokinetics. Toxicokinetic modelling is currently seen as the most adequate approach to simulate the fate of compounds in the human body. *In silico* tools that can be used for toxicokinetic modelling are dealt with in the next chapter. *In vitro* approached for studying ADME processes are outlined below.

2.4.1 Absorption

Absorption depends on the compound-specific properties and physiology (and pathology) of the epithelial tissue. Some of the properties of importance are physicochemical properties of a compound and the availability of specific influx and efflux transporters in the tissue. Three processes (partially linear, partially in parallel) can be distinguished that determine bioavailability: (1) release of the compound from its matrix (bioaccessibility), (2) absorption of the released fraction and (3) metabolism before reaching the systemic circulation. **Organotypic culture models** applicable to the intestinal and pulmonary barriers are being actively developed. However, these are laborious experimental systems, which are not easy to handle, and are currently restricted to mechanistic investigations or specific questions facilitating *in vitro*–*in vivo* comparisons. Absorption after oral exposure can be predicted by *in vitro* intestinal models such as the Caco-2 **cell line**. This model predicts absorption over a single barrier and is rather standard. However, its predictive capacity is limited for highly lipophilic compounds or substances in the low-to-moderate absorption area as well as when transporter-mediated routes and/or first pass metabolism (metabolic conversion by the liver before the substance reaches the systemic circulation) are involved (Coecke et al. 2013). It could be incorporated in a medium-throughput test strategy. Dermal absorption can be assessed by the *in vitro* method described in OECD Test Guideline 428, which is the only *in vitro* toxicokinetics test covered by an OECD test guideline. The absorption of several concentrations of a test substance in typical formulations through skin preparations from human or animal sources is assessed during 24 hours. This test method has been used as a stand-alone test method for several regulatory requirements. However, to adequately measure the flux across the dermal barrier and extrapolate results to different doses and exposure times for the purpose of integration into modelling approaches, some additional work needs to be carried out. Information on the biotransformation capacity of the model is for instance limited. *In vitro* methods for absorption in the lung are lacking.

2.4.2 Distribution

The distribution of a compound and its metabolites inside the body is again governed by three main factors: (1) binding of the substance to plasma proteins, (2) the partition of the substance between blood and specific tissues, and (3) the ability of the substance to cross specialized membranes, so-called barriers, such as the blood–brain barrier, blood–placental barrier, and blood–testis barrier.

Only the free (unbound) fraction of a compound is available for diffusion and transport across cell membranes. Therefore, it is essential to determine the binding of a compound to plasma or serum proteins. The easy availability of human plasma makes it possible to determine the unbound fraction of compounds by performing *in vitro* incubations directly in human plasma. There are three methods generally used for determination of **plasma protein binding**: **equilibrium dialysis** (ED), **ultrafiltration** (UF) and **ultracentrifugation**. All methods can be automated for high-throughput, are easy to perform and have good precision and reproducibility. The use of combined **LC/MS/MS** allows high selectivity and sensitivity. Equilibrium dialysis, in which a membrane is used that is only permeable to the unbound chemical and the unbound fraction is determined by measuring the amount of chemical that crosses the membrane, is regarded as the ''gold standard'' approach.

The fate of a compound in the body is determined by partitioning into the human tissues. The measurement of tissue storage and a molecular understanding of tissue affinity have, historically, not been studied to the same extent as plasma protein binding. However, the knowledge of these partitioning coefficients is essential for the development of toxicokinetic models. Quite a large number of approaches have been developed over the last years. The tissue–blood partition coefficient can be determined using **tissue–buffer homogenates**, in which tissues can be mixed to obtain average values for, for example, richly perfused tissue groups. Olive oil or octanol are often used instead of adipose tissue. The unbound concentration is typically assessed by one of the following techniques: **equilibrium dialysis**, **ultracentrifugation**, **headspace analysis** (for volatiles) or **solid-phase (micro-)extraction** followed by a classical analysis such as **HPLC UV** or **MS**. The purpose of this technique is the prediction of the *in vivo* tissue blood partitioning and the prediction of an *in vivo* volume of distribution. However, various transporters (influx and/or efflux proteins) in cells of excretory organs (intestine, liver, kidney) and on various blood/organ barriers (such as brain, testis, placenta) regulate the passage of substances across membranes. Knowledge on regulation of expression and function of such transporters is lacking (Coecke et al. 2013).

The blood–brain barrier (BBB, **Figure 12**) is a regulatory interface that separates the central nervous system from the systemic blood circulation and may limit or impair the delivery of certain compounds, which makes the brain different from other tissues. There are several passive and active mechanisms of transport through the BBB. Several *in vitro* BBB models are under development which integrate various **cells** of vascular and neural origin. Also single **cell lines** containing transfected transporters have been proposed as models to study BBB permeability. However, all available models are in early stages of development. Recently, some molecular models have been developed to consider the BBB **transporters**.

The blood–placenta barrier (BPB) serves to transport nutrients and waste, and other compounds such as hormones. However, the placenta does not provide a true barrier to protect the foetus from exposure to compounds present in the mother's systemic circulation, although it might reduce the transport of certain molecules. The transfer across the placenta can occur by several active or passive processes. There are both primary and permanent trophoblast-derived **cell models** available. An ex vivo model, human **perfused placenta** cotyledon, offers information about transplacental transfer, placental metabolism, storage, acute toxicity and the role of transporters, as well as an estimation of foetal exposure.

The blood–testis barrier (BTB) is a physical and physiological barrier which assures functions in hormonal regulation and spermatogenesis. Many **organ cultures**, **co-cultures** or **single cell cultures** have been evaluated, but none has developed into a testing system for toxicokinetic processes.

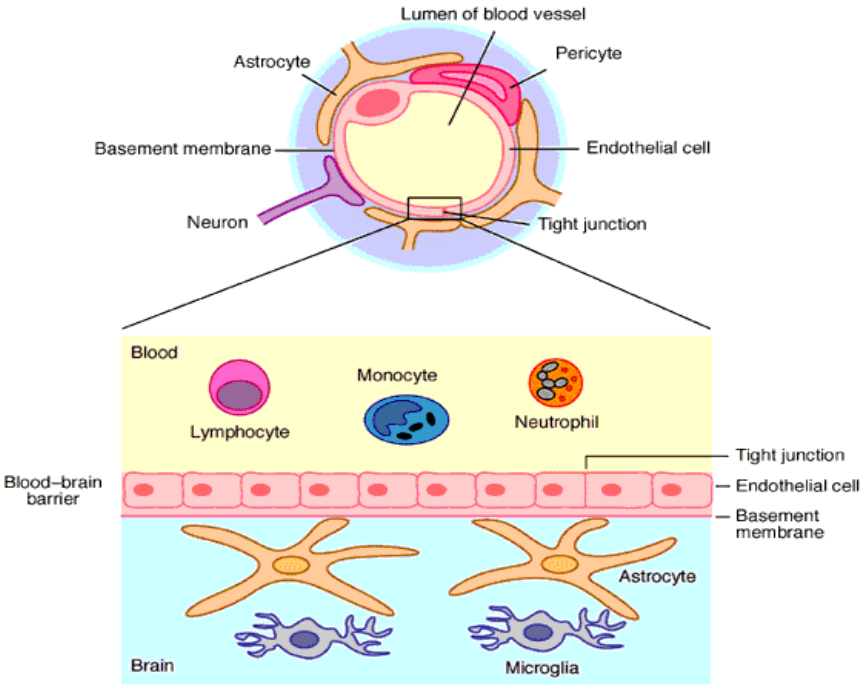

Figure 12. Schematic representation of the blood-brain barrier. The tight junctions between endothelial cells lining blood vessels in the brain limit transport of certain cells and molecules from blood to brain (www.expertreviews.org).

2.4.3 Metabolism

Metabolism or biotransformation is the principal elimination route of organic chemicals. A multitude of xenobiotic-metabolizing enzymes may act on a chemical in different metabolic pathways (**Figure 13**). Because liver is the principal site of metabolism, the enzyme component in *in vitro* systems should preferably be liver-derived and of human origin, to avoid species differences. There is a generally accepted consensus that metabolically competent human hepatocytes or hepatocyte-like cell lines are the best enzyme source to perform the primary screening of metabolism. However, the limited availability of human cells and tissues, and ethical concerns which are often raised, should be taken into account, although the use of human **recombinant enzymes**, **transgenic cells** *in vitro*, and the possibility to cryopreserve human **hepatocytes** are of great help. Attention is still needed in selecting the more adequate test system (**cDNA expressed enzymes**, **microsomes** (subcellular components containing many metabolic enzymes), **homogenates**, hepatocytes) for studies with different purposes (Coecke et al. 2013).

The two most important endpoints measured are (1) intrinsic clearance which can be extrapolated into hepatic metabolic clearance and (2) the identification of metabolites (stable, inactive, active or reactive metabolites of concern). The **metabolic stability test** is a relatively simple, fast-to-perform, but specialized **mass spectrometry** based study, to find out whether a compound is metabolically stable or labile. It measures the disappearance of the parent compound over time when incubated with a metabolically competent tissue preparation (e.g. a human liver preparation, preferably human hepatocytes). The rate of parent compound disappearance gives a measure of its metabolic stability and allows for the calculation of intrinsic clearance and extrapolation to hepatic (metabolic) clearance. The use of liver-based experimental systems should give a fairly reliable view of hepatic intrinsic clearance. However, to be able to predict *in vivo* clearance, a number of assumptions concerning the substance under study must be made, so an extrapolation model is needed.

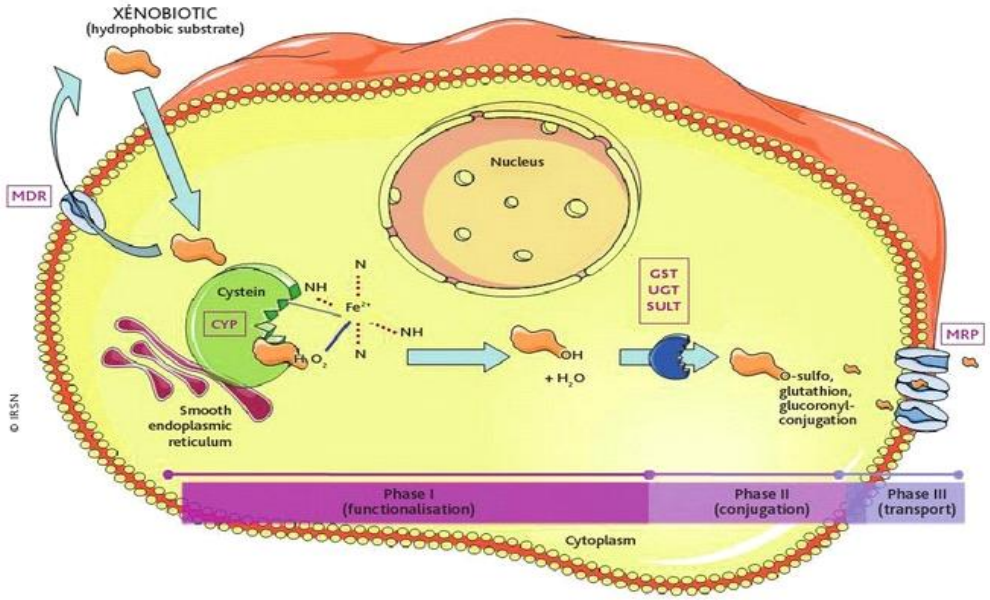

Figure 13. Cellular xenobiotic metabolism. Phase I enzymes (essentially the cytochromes P450, CYP) essentially catalyse oxido-reduction and hydrolysis reactions. Phase II enzymes (glutathione-S-transferases or GST, UDP glucuronosyl transferases or UGT, sulfotransferases or SULT) catalyse conjugation reactions. Multidrug resistance proteins (MRP) transport xenobiotics and their, usually more hydrophilic, metabolites through the membranes in order to eliminate them from the cell (www.irsn.fr).

Although no formal validation studies are known, the screening test for metabolic clearance should be relatively ready for validation after the availability of standard procedures. To cover extrahepatic biotransformation as well, the method for metabolic stability can be combined with the use of other tissues. With the advent of modern MS techniques, it is possible and feasible to study both the detailed qualitative and quantitative metabolic profiles of a compound. The use of recombinant expressed enzymes (principally cytochrome P450 (CYP) and UDP-glucuronosyltransferase (UGT) enzymes but increasingly also other xenobiotic-metabolizing enzymes), transfected cells and metabolically competent human (liver-derived) cell lines or subcellular fractions is being very actively employed in pharmaceutical industry and academia. In this way, it is possible to gain an indication of the enzymes participating in the metabolism which allows a number of predictions about physiological, pathological, and environmental factors affecting the kinetics of a compound of interest.

There are direct and indirect methods to test the potential formation of reactive metabolites. Most direct assays use **trapping agents** (i.e. glutathione or its derivatives, semicarbazide, methoxylamine or potassium cyanide) that are able to trap both soft and hard electrophiles; conjugates are then analytically measured. The Ames test is an example of an indirect method for the bioactivation assay making use of metabolically competent enzyme system (which could be human-derived, if needed) and properly engineered bacteria to detect reactive, DNA-bound metabolites. Since enzyme induction (i.e. increased expression after chemical exposure) has a complex underlying mechanism, it is a good indicator for high-quality metabolic competent systems that can be used for long-term purposes. That is why developments are ongoing to assess CYP induction in bioreactor-based systems. A large number of test systems ranging from **nuclear receptor binding** assays to **induction-competent cell lines** and cryopreserved human **hepatocytes** are currently available.

The reliability of two hepatic metabolically competent test systems (cryopreserved human hepatocytes and the human HepaRG cell line) is currently assessed by ECVAM. The test is standardized using coded test items. The coded test items are selected on the basis evidence from *in vivo* human data on their CYP induction potential and activation of the three main nuclear receptors (CAR, AhR, PXR) involved in CYP induction. Induction of four CYP isoforms that have been selected as the main CYP enzymes involved in the metabolisms of drugs and xenobiotics (CYP1A2, CYP2B6, CYP2C9 and CYP3A4), is measured following treatment with test compounds and two reference compounds (β-naphthoflavone and rifampicin), using a cocktail of prototypical substrates for different CYP isoforms incubated directly with the two test systems. These test systems are widely used in pharmaceutical industry to help early drug development and are designed to detect induction of CYP enzymes relevant for the pharmaceutical area. This can represent a potential limitation since for other substances, other CYP forms might play an additional or more prominent role. Thus, further progress is needed to cover this potential gap.

Due to the broad substrate specificity of metabolizing enzymes, there is always a possibility that compounds would interfere with each other's biotransformation. Inhibition of biotransformation leads to higher concentrations and delayed clearance and may cause adverse effects. At the site of entry, inhibition of the first-pass metabolism would increase the blood concentration of the parent compound.

2.4.4 Excretion

Predicting major excretion pathways of compounds is important in relation to their kinetic behaviour and the relationship to toxicological effects. The kidneys and the hepato-biliary system have the capacity to excrete substances either as the parent compound or as metabolites and are important routes for elimination of chemicals and their metabolites. Unfortunately, excretory processes seem to be the least developed area in the context of *in vitro* toxicokinetic methods, probably because renal and biliary excretion, the major excretory routes, are complex processes with a number of passive components and active processes involved. Excretion by the kidney encompasses different mechanisms and they all include the interplay of both passive movement of drugs and the participation of a number of active transporters. Even if there are examples how the involved transporters can be identified, it is difficult to use the findings to feed into a physiological model of renal excretion which includes tubular secretion and tubular reabsorption. In humans, biliary excretion does not seem to play an important role for most of the substances. However, in cases where it matters, the process is rather complex, first preceded with the entry of the substance to the hepatocyte and its possible metabolism by the hepatic metabolic machinery. Metabolizing enzymes produce conjugates, which are then transported across the canalicular membrane to be excreted into the bile.

There have been few attempts for developing computational approaches to predict renal excretion from some basic molecular and physicochemical properties. Efforts have been undertaken to use **collagen-sandwich cultures** of hepatocytes as an *in vitro* test system for testing biliary excretion. Due to the fact that progress in the field is very recent, no systematic efforts have been undertaken to standardize approaches and no formal validation studies are known. An understanding of mechanisms that determine these processes is required for the prediction of renal and biliary excretion. Whereas for biliary excretion some advances have been made with *in vitro* models, no reports could be identified in the literature on *in vitro* models of renal excretion nor were reports available on *in silico* methods.

3 Non-testing approaches

Several strategies exist that do not involve the use of any biological system, but are built utilizing chemical and toxicological data, which meets the requirement to reduce animal testing and to speed the process of identifying potentially harmful compounds. These non-testing methods are based on the principle that the properties of a chemical can be predicted from its molecular structure and substructure, and inferred from the properties of similar compounds of which activities are known. The term substructure refers to an atom, or group of adjacently connected atoms, in a molecule. A substructure associated with the presence of a biological activity is also called a structural alert (SA). SAs allow identification of structural characteristics that raise concerns or rule out possible hazards through the application of **structure activity relationships** (**SARs**) or **quantitative structure activity relationships** (**QSARs**) (Basketter et al. 2012). **Chemical categorizing** is an approach in which only prototypic compounds out of a group of similar ones are tested. **Read-across** techniques fill data gaps on a specific chemical by interpolating or extrapolating existing data of related chemicals within a category of substances. **Physiologically based toxicokinetic** (**PBTK**) **modelling** predicts the toxicokinetics of chemicals based on physiological and chemical parameters. These approaches, also referred to as *in silico* models, as compared to *in vitro* and *in vivo* methods, and their applications are described in more detail below.

> ***In silico models*** *Approaches for the assessment of chemicals based on the use computer-based estimations or simulations.*

3.1 Structure activity relationships

Structure activity relationships are mathematical based models relating one or more quantitative parameters, which are derived from the chemical (sub)structure, to a qualitative (in case of classification models) or quantitative (in case of statistical regression models) measure of a property or activity. SARs and QSARs are collectively referred to as (Q)SARs.

> ***(Q)SARs*** *Theoretical models that can be used to predict the physicochemical, biological (e.g. toxicological) and environmental fate properties of molecules from the knowledge of chemical structure*

Physicochemical properties include for instance boiling point, melting point, vapour pressure, water solubility, Henry's Law Constant over a temperature range, logarithmic octanol-water partition coefficient (log Kow), octanol-air partition coefficient, soil adsorption coefficient (Koc), bioconcentration factor (BCF), bioaccumulation factor (BAF), biotransformation rate (kM), aqueous hydrolysis rate constants, acid- and base-catalysed rate constants, hydrolysis half-lives at selected pH values. atmospheric half-lives, and probability of rapid aerobic and anaerobic biodegradation of organic compounds.

For the development of (Q)SAR models, a step-wise approach is used in which each step needs specific consideration. After all, the confidence in the outcome is highly dependent on data availability and understanding of the limitations of the system. The justification of the approach followed is important, and for this purpose a training set and validation set of chemicals needs to be included in the analysis. The main steps involved in developing a non-testing approach are as follows:

1. Selection of a training set of chemicals, which consists of molecules that vary in structure and have well-characterized toxicities and mechanisms of action;
2. Entry of structural information using codes (such as SMILES) or a chemical drawing package (like Isis/Draw, ChemDraw or CORINA);
3. Analysis of molecular descriptors such as those enumerated below:

Molecular descriptors relevant for developing non-testing strategies

molecular volume
molecular shape
molecular connectivity
the overall number of atoms in the molecule
dipole moment
molar refractivity
electronegativity
steric effects
quantum chemistry or electronic configuration states (intrinsic reactivity)
acidic dissociation constant (pKa)
octanol/water partition coefficient (logP; lipophilicity or hydrophobicity: a measure of bioavailability)

4. Processing of physicochemical and toxicological information to identify sub-structural features associated with toxicity ('toxicophores'); toxicophores determine whether chemicals can specifically interact with receptors, enzymes, or macromolecules such as proteins and DNA, thereby perturbing their functions;
5. Model development through analysis of molecular descriptors and their relative distributions in active and inactive molecules in a training set and in a query chemical in a test set using similarity searching, cluster analysis, and mathematical and statistical methods;
6. Testing for goodness of fit;
7. Validation using a validation set of chemicals;
8. Defining an applicability domain (AD);
9. Application for toxicological assessment of specific endpoints (Combes 2012).

Applicability domain *The physicochemical, structural, or biological space, knowledge or information on which the training set of the model has been developed and for which it is applicable to make predictions for new compounds*

This stepwise approach for (Q)SAR development is outlined in **Figure 14**. In addition to the (Q)SARs that have been reported in the scientific literature (more than 20,000 models), a number of expert systems have been developed, generally as commercial products. Computerized expert systems, which are available tools in software packages, use rules developed from pre-existing information to predict biological activity. These rules can either be designed by humans (knowledge-based systems) or by algorithms (automated rule induction systems). Two examples of KBS for hazard prediction are **HazardExpert** and **DEREK** for Windows (Deductive Estimation of Risk from Existing Knowledge). KBS for biotransformation prediction also exist and include **MetabolExpert** and **Meteor**. Examples of ARIs are **CASE** (Computer Automated Structure Evaluation), **MULTI-CASE**, and **TopKat** (Toxicity Prediction by Komputer Assisted Technology), while **TIMES** and **METACASE** can be

used in predicting biotransformation (Combes 2012).

Knowledge-Based Systems (KBS) *Use rules developed by human experts, which are programmed into software that then predicts the likely activity based on the presence or absence of toxicophores using reasoning engines*

Automated Rule Induction systems (ARIs) *Develop rules using algorithms and are useful for analysing complex data.*

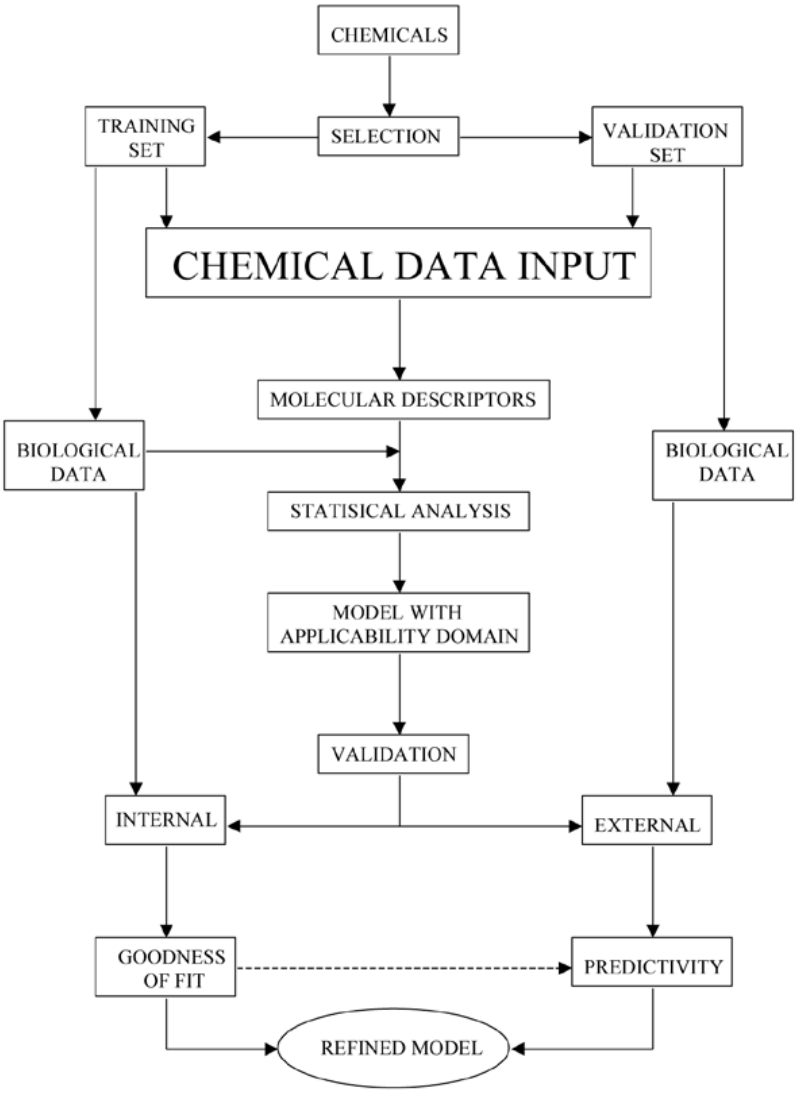

Figure 14. Overall scheme for developing and validating in silico methods for toxicity prediction (Combes 2012).

In toxicological hazard assessment, (Q)SARs are used to predict the impacts of chemicals on human health, wildlife, and the environment. Within the REACH procedures, persistence (recalcitrance, or resistance to bio-degradation), bioaccumulation and evidence of toxicity (PBT) have been identified as important properties of chemicals, and are used to prioritize chemicals for further testing. **PBT Profiler** (which incorporates the US EPA's **BCFWIN** and

ECOSAR software for determining environmental persistence, bio-concentration potential, and aquatic toxicity) and a set of QSAR models developed by using the Danish Environmental Protection Agency's (EPA's) extensive database on chemicals, are examples of models that can predict PBT properties. Expert systems customised for specific biological endpoints are available too. As the development of *in silico* models for toxicity is dependent on the availability of data and defined mechanisms of action, it is not surprising that (Q)SARs are more prevalent for endpoints where there are large databases such as ecotoxicity, mutagenicity and carcinogenicity, skin sensitisation, and endocrine disruption via ER receptor binding and ER reporter gene activation (see paragraph 3.4.3). The *in silico* prediction of some other endpoints, like eye and skin irritation and corrosion, is being improved as more data are being accumulated, but few models for general mammalian toxicity, particularly long-term effects, are available due to lack of knowledge about many of the mechanisms involved. Now that more data become available, attempts are also being made to develop QSAR models for predicting the inhalation toxicity of nanoparticles by using descriptors such as the ratio of size to surface area, surface charge and surface derivatization (Combes 2012). Those (Q)SARs of which the descriptors can be interpreted and rationalised in terms of the plausible mode of action (MOA) can also provide useful mechanistic information to help substantiate a read-across (next paragraph). In addition, (Q)SARs may provide the anchor or the context of similarity to justify why a given pair or set of substances are expected to behave similarly in terms of their biological activity (Combes 2012). In other disciplines beyond the scope of this report, (Q)SARs are used to predict the behaviour of compounds, for example in analytical chemistry (e.g. Schulten and Leinweber, 1996) and drinking water treatment (e.g. Luilo and Cabaniss, 2011).

Various decision-tree schemes have been devised for categorizing chemicals according to their predicted hazards. One of the earliest was developed by Cramer and Ford for mammalian acute toxicity, which classifies organic chemicals according to structure and biochemical and physiological chemistry into low priority substances (low human exposure and toxicity), high priority substances (high human exposure and toxicity), and intermediate priority substances (possible human exposure and toxicity). This scheme is now one of four methods originally forming the basis of a software package released by the European Chemicals Agency (ECHA), which coordinates the REACH processes, called **Toxtree**. The three other decision-tree schemes initially involved in Toxtree are for aquatic toxicity, skin and eye irritation and corrosion, and mutagenicity and carcinogenicity. Toxtree now has additional modules, for biodegradation and persistence, as well as chromosomal damage (the *in vivo* micronucleus assay). A decision-tree scheme for oestrogenicity has also been developed (Combes 2012).

3.2 Chemical categorizing and data gap filling

Information may be generated by addressing compounds which share a similar structure as a group of closely related chemicals, instead of single chemicals. This is referred to as a 'analogue approach' or 'category approach' (Combes 2012). An analogue approach is often used when the grouping is based on a very limited number of chemicals, whereas a chemical category describes a larger group of chemicals (van Leeuwen et al., 2009; ECETOC 2012).

> ***Category approach*** *Technique for grouping chemicals whose physicochemical and human health and/or environmental toxicological properties and/or environmental fate properties are likely to be similar or follow a regular pattern as a result of structural similarity*

Figure 15 shows a stepwise approach to construct a category as proposed by the OECD (2007).

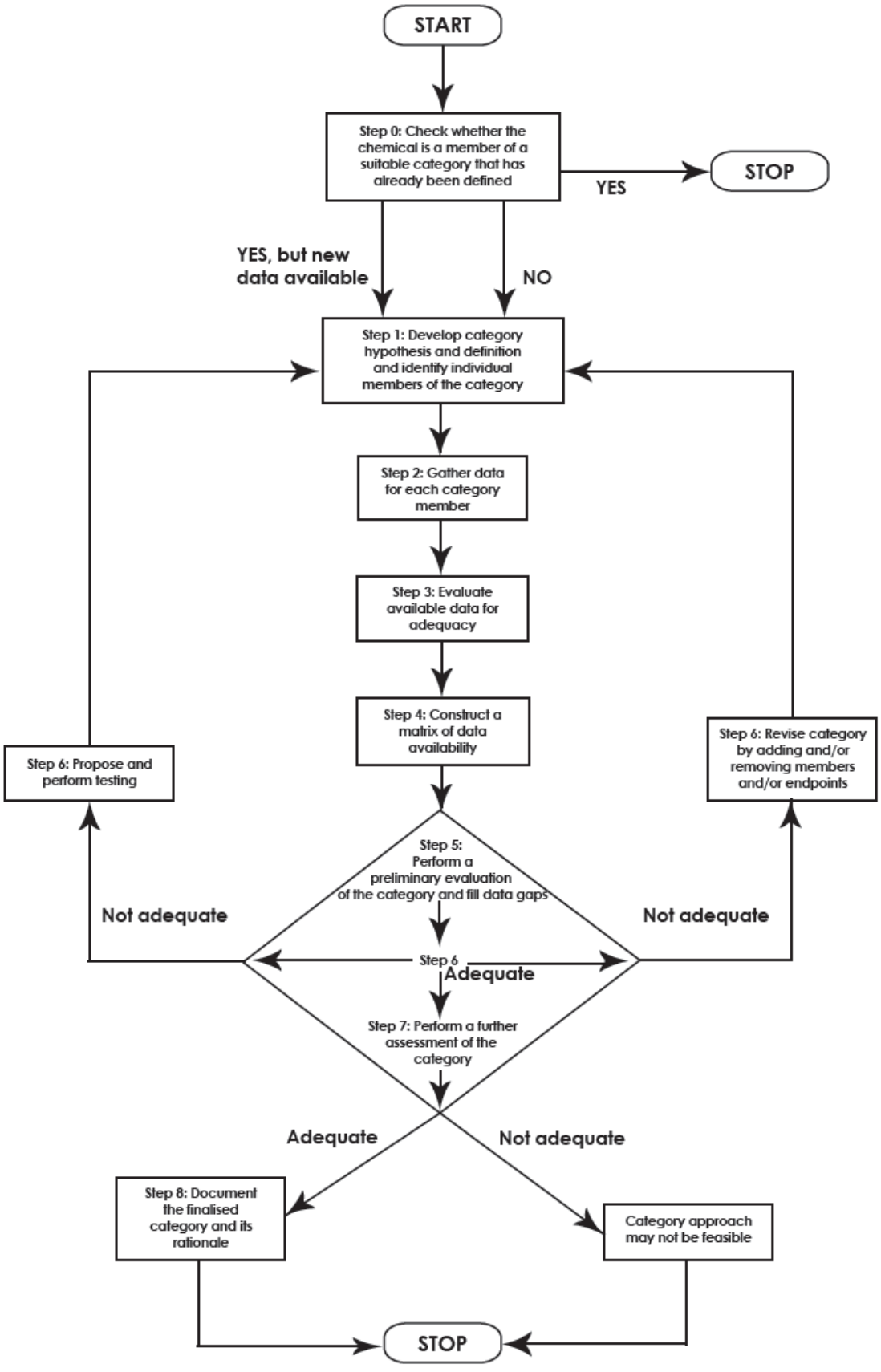

Figure 15. Stepwise approach to construct a chemical category (OECD 2007).

The similarities between chemicals within a group may be based on the following (OECD 2007):

- common functional group(s) (e.g. aldehyde, epoxide, ester, specific metal ion);
- common constituents or chemical classes, similar carbon range numbers, mainly in case of complex substances often known as "substances of Unknown or Variable composition, Complex reaction products or Biological material" (UVCB substances);
- an incremental and constant change across the category (e.g. a chain-length category), often observed in physicochemical properties, e.g. boiling point range;
- the likelihood of common precursors and/or breakdown products, via physical or biological processes, which result in structurally similar chemicals (e.g. the "metabolic pathway approach" of examining related chemicals such as acid/ester/salt).

Tools that can be used to identify analogues have been reviewed by ECETOC and are listed below (ECETOC 2012). In its report, ECETOC also addresses tools that can facilitate the grouping of chemicals; examples are **Toxmatch** and the **OECD QSAR Toolbox** (see paragraph 3.4).

Tools for the identification of analogues

Chemispider	http://www.chemspider.com
Chemifinder	http://www.chemfinder.com
ChemIDplus	http://chem.sis.nlm.nih.gov/chemidplus
Pubchem	http://pubchem.ncbi.nlm.nih.gov/search/search.cgi
ACToR	http://actor.epa.gov/actor/faces/ACToRHome.jsp
OECD (Q)SAR Toolbox	http://www.oecd.org/env/ehs/risk-assessment/theoecdqsartoolbox.htm
Leadscope	http://www.leadscope.com
Scifinder	http://www.cas.org/products/scifinder
DiscoveryGate	http://accelrys.com/products/databases/database-access/discovery-gate.html

The category approach implies that comparisons are made between a very limited number of chemicals. The use of a category approach will mean that it is possible to identify properties which are common to at least some members of the category. Chemical categories can also be used to assign missing property values on the basis of the knowledge of the values of the other members of the group, which is referred to as 'read-across' (Schaafsma et al., 2009; Combes 2012).

> ***Read-across*** *Filling toxicity data gaps by adopting the hazard profile of another chemical with sufficiently similar physicochemical properties to categorize both chemicals together in the same group*

Endpoint information for one chemical is then used to predict the same endpoint for another chemical, which is considered to be similar in some way (usually on the basis of structural similarity and similar properties and/or activities). Information on chemicals may be derived from the tool that was applied to identify analogues (such as ACToR or the OECD Toolbox) or from databases such as the OECD's eChemPortal (http://www.echemportal.org/echemportal/page.action?pageID=0), the European chemical Substances Information System (ESIS;

http://www.cefic-lri.org/lri-toolbox/fedtex), or the European Chemicals Agency (ECHA) dissemination website (http://echa.europa.eu/web/guest/information-on-chemicals/registered-substances). For read-across, the overall category data and rationale need be adequate to support a screening-level hazard assessment for the untested endpoints. As described more in detail by Patlewicz et al. (2013), grouping strategies do not differ in terms of their scientific justification, but may differ with respect to grounds for confidence in the prediction. Confidence in the read-across prediction is higher when interpolation (data gap filling for closely related chemicals) can be applied rather than extrapolation (using data from category members at one end of the category to predict hazards for chemicals at the other end of the category) (**Figure 16**). Both methods should be periodically reviewed and updated, not only since category assessments are complex but also because the read-across approach is under on-going development and new data are constantly generated (ECETOC 2012).

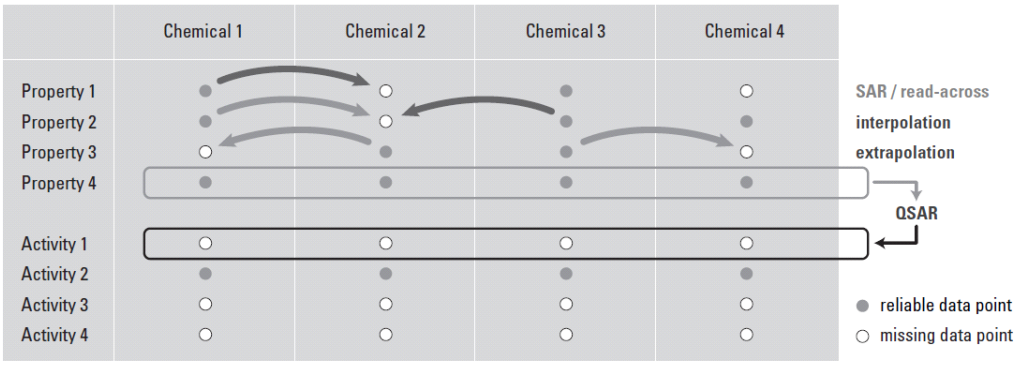

Figure 16. A chemical category can be represented graphically as a two-dimensional matrix in which different category members occupy different columns, and the different category endpoints occupy different rows. Data gaps may be filled by one or more of the following procedures: qualitative read-across, quantitative read-across, use of SARs and QSARs. From Van Leeuwen, C.J., G.Y. Patlewicz, and A.P. Worth (2007) Intelligent Testing Strategies In: Risk Assessment of Chemicals. An Introduction (2nd edition). Van Leeuwen, C.J. and T.G. Vermeire, eds. Springer Publishers, Dordrecht, The Netherlands, pp 467-509.

In principle, read-across can be applied for any property or endpoint, irrespective of whether it is a physicochemical property, environmental fate parameter, human health effect, or ecotoxicological effect. Read-across can be qualitative or quantitative. In qualitative read-across, the presence (or absence) of a property/activity for the target chemical is inferred from the presence (or absence) of the same property/activity for one or more source chemicals. Qualitative read-across gives a 'yes/no' answer. In quantitative read-across, the known value(s) of a property for one or more source chemicals is used to estimate the unknown value of the same property for the target chemical. In the case of a toxicological effect (human health or ecotoxicological), this assumption implies that the potency of an effect shared by the two chemicals is similar or follow a regular pattern. Quantitative read-across is used to obtain a quantitative value for an endpoint, such as a dose-response relationship (ECHA 2012).

An example of a tool allowing read-across is **Toxmatch**. Toxmatch has been shown to be able to define categories of chemicals specifically for aquatic toxicity, bio-accumulation and teratogenicity by using information from a number of different databases, and in some cases to make predictions about the toxicity of several query chemicals (Combes 2012).

For some category endpoints, the members may be related by a trend. With a sufficiently large number of compounds in a category, and where a trend is observed, the data can be

used to derive an internal QSAR model that describes the properties of a member of a category (ECETOC 2012).

3.3 PBTK modelling

Non-testing approaches that model toxicokinetic processes may predict whether chemicals may reach toxic internal dose levels. To quantitatively predict the toxicokinetics of a substance, it is necessary to model jointly their physiological determinants and the compound-specific properties. Physiologically based toxicokinetic (PBTK) modelling is currently the most advanced tool for that task.

> ***PBTK models*** *Mathematical descriptions of ADME processes that describe the organism as a set of compartments that are characterized physiologically or empirically*

PBTK models facilitate quantitative descriptions of the temporal change in the concentration of a chemical and/or its metabolites in biological matrices (e.g., blood, tissue, urine, alveolar air) of the exposed organism. Two categories of parameters are needed in order to simulate the toxicokinetics of a chemical in a PBTK model: (1) physiological parameters that are chemical-independent, such as cardiac output and organ blood flow (species-, sex- and age-specific and largely available in public literature) blood flow (species-, sex- and age-specific and largely available in public literature) and (2) chemical-specific parameters that have to be determined e.g. by *in vitro* test methods or predicted by *in silico* (data-based) methods for each chemical (Adler et al. 2011).

Published models range from simple compartmental to very sophisticated types. Between compartments, the transport of substances is dictated by various physiological flows (blood, bile, pulmonary ventilation, etc.) or by diffusions. Perfusion-rate-limited kinetics applies when the tissue membrane presents no barrier to distribution. Generally, this condition is likely to be met by small lipophilic substances. In contrast, permeability-rate kinetics applies when the distribution of the substance to a tissue is rate limited by the permeability of a compound across the tissue membrane. That condition is more common with polar compounds and large molecular structures. Consequently, PBTK models may exhibit different degrees of complexity. In the simplest and most commonly applied form, each tissue is considered to be a well-stirred compartment, in which the substance distribution is limited by blood flow (**Figure 17**). In such a model, any of the tissues can be a site of elimination. However, in this case it is assumed that the liver is the only metabolising organ and that excretion only happens in the kidney. Any route of exposure can be described either in isolation or in combination. For example, systemic toxicity may be studied following intravenous infusion, uptake via the gastrointestinal tract, dermal absorption and inhalation via the lungs (Adler et al. 2011).

Since a given exposure may induce different effects in the individuals of a population, and the same individual may respond differently to the same exposure at different times in his/her lifetime, inter-individual and intra-individual extrapolations of PBTK data are needed. These extrapolations are performed by setting parameter values to those of the subpopulation or individual of interest and are mainly used to predict the differential effects of chemicals on sensitive populations such as children, pregnant women, the elderly, the obese, and the sick, taking into account genetic variation of key metabolic enzymes, etc. The toxicokinetic behaviour of a compound can also be studied under special conditions, such as physical activity. Extrapolations between different doses are achieved by capturing both the linear and non-linear steps of the biological processes known to govern the kinetics of the chemical of interest, e.g. in the transport and metabolism (Adler et al. 2011).

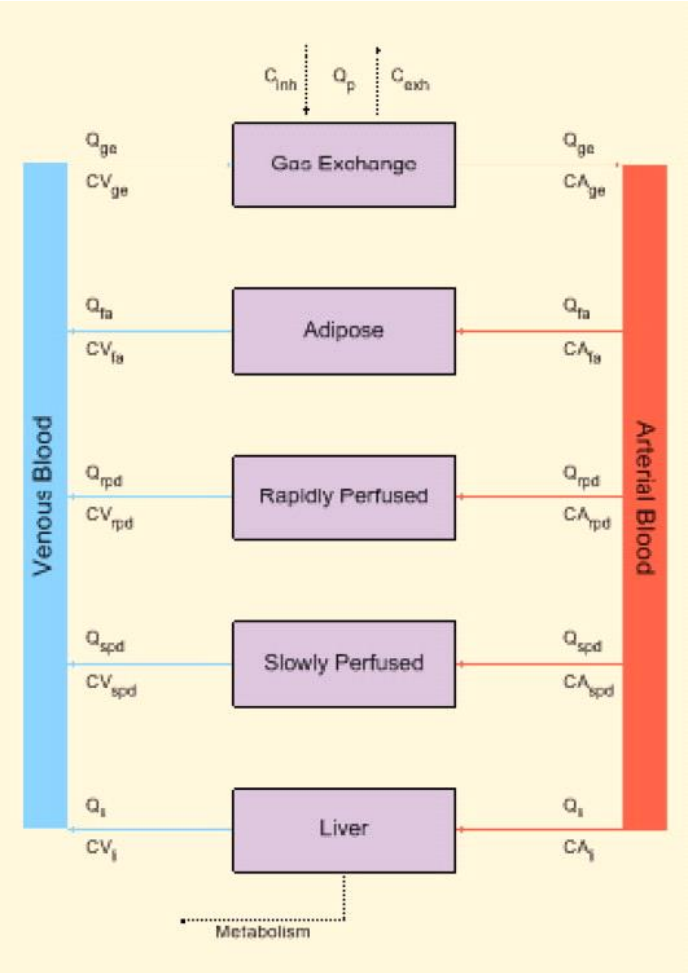

Figure 17. Example of a simple PBTK model. The arrows indicate pathways for transfer of substance between organs or tissues. Cinh is the inhaled concentration, Cexh is the exhaled concentration, Q, represents a physiological flow rate (alveolar ventilation or regional blood flow), CV is the venous blood concentration, CA is the arterial blood concentration, the subscripts represent the following: p, pulmonary, ge, gas exchange, fa, fat, rpd, rapidly perfused, spd, slowly perfused, li, liver. Physiological parameters (mainly volumes and flows) or substance-specific parameters (partition coefficients, permeability constants, metabolic rate constants etc.) are available from the published literature (Coecke et al. 2013).

The fact that PBTK models are mechanism-based, which allows them to be 'generic' to a certain extent (various extrapolations possible), has been critical for their success so far. The need for high-quality *in vitro* and *in silico* data on ADME as input for PBTK models to predict human dose–response curves is currently a bottleneck for risk assessment. In the absence of *in vitro* and/or *in silico* kinetic data that can parameterize a whole body PBTK model, concentration–response data generated at tissue/cell or sub-cellular level cannot be translated into *in vivo* dose response or potency information relating to the entire target organism, i.e. the human body (Coecke et al. 2013). Difficulties in predicting metabolism, renal excretion, and active transport have an impact on the predictive accuracy of PBTK models. The need remains to validate the QSAR sub-models or the *in vitro* assays used to assign a PBTK model's parameter values. *In silico* models to estimate oral bioavailability have been developed for the use in conjunction with PBTK models. The use of these organ-level *in silico* models is currently the best way to integrate the inputs from tests of bioaccessibility, absorption and metabolism, including hepatic clearance in the first-pass situation, because of the complex nature of bioavailability. Obviously, the quality of all input conditions the

validity of the PBTK model which uses it. Sensitivity and uncertainty analyses can also be performed to understand which are the critical aspects of the model that might require particular attention. Experimental or observational data are not always available to convincingly validate such complex models. 'Virtual' experiments simulated by varying parameters, as in sensitivity analysis, can point to important areas of future research needed to build confidence in *in silico* predictions. Formal optimal design techniques can also be used to that effect. In any case, the major challenge will probably be the coupling of PBTK models to predictive toxicity models, at the cellular and at the organ level. However, even if proof of concept has been provided for the strategy how to proceed there is presently not much experience, and hence, further development and refinement is necessary. One could envisage building a publicly available user friendly tool for PBTK modelling (Adler et al. 2011).

3.4 Expert systems

Recent developments in computing power, the ability to create extensive databases and the use of the internet to compile, organise and distribute information have increased the capability to investigate relationships between chemical structure and biological activity (IGHRC 2013). A range of software tools, or expert systems are available in which *in silico* models have been integrated. Some of those were already mentioned in this report. A range of available tools has also recently been presented in a workshop for the drinking water companies[1]. The following paragraphs describe a selection of freely and commercially available expert systems in more detail. The information presented is derived from the overview of expert systems provided by ECETOC (2012). The endpoints that are addressed by the tools evaluated by ECETOC, including metabolism, are listed in Table 1. Besides two tools that predict physicochemical properties, mainly tools that include relevant human health endpoints for drinking water contaminant exposure are highlighted below.

Table 1. Key (Q)SAR tools and the endpoints that they address (ECETOC 2012).

Endpoint	Tool
Actue oral toxcitity	T.E.S.T TOPKAT OECD (Q)SAR toolbox
Acute inhalation toxicity	TOPKAT OECD (Q)SAR toolbox
Skin irritation	TOPKAT BfR rulebase within Toxtree BfR rulebase within OECD (Q)SAR toolbox Derek Nexus
Eye irritation	BfR rulebase within Toxtree Bfr rulebase within OECD (Q)SAR toolbox TOPKAT Derek Nexus
Phototoxicity	TIMES
Endocrine disruption	TIMES(aromatase inhibition, estrogen, androgen receptor binding affinities) OECD (Q)SAR toolbox (E-receptor binding affinity profiler)
Skin sensitisation	CEASAR TIMES TOPKAT Derek Nexus MCASE

[1] Workshop Innovative testing strategies November 19th 2013; joint research projec 'Innovatieve risicobeoordeling van drinkwater prioritaire stoffen'

	OECD (Q)SAR toolbox (protein binding affinity profiler) SMARTS alerts within Toxtree
Mutageniticity: Ames	TOPKAT TIMES OECD (Q)SAR toolbox (DNA binding affinity profiler) Benigni-Bossa rulebase - Toxtree Benigni-Bossa rulebase - OECD (Q)SAR toolbox MCASE T.E.S.T. CEASAR
Mutagenicity: *in vitro* chromosomal aberration	TIMES Derek Nexus
Carcinogenicity	TOPKAT MCASE CAESAR ONCOLOGIC LAZAR
Developmental toxicity	CAESAR T.E.S.T Derek Nexus TOPKAT MCASE
Reproductive toxicity	Derek Nexus
Acute aquatic toxicity (fish)	Lazar (MRDD) TOPKAT (LOAEL, MTD) Derek Nexus OECD (Q)SAR toolbox (NEDO, Frauenhofer)
Acute aquatic toxicity (daphnid)	ECOSAR TOPKAT T.E.S.T. TIMES Verhaar rulebase – Toxtree Verhaar rulebase - OECD (Q)SAR toolbox OASIS MOA - OECD (Q)SAR toolbox Lazar
Acute aquatic toxicity (algae)	ECOSAR TIMES
Bioaccumulation (BCF)	Catlogic BCFBAF CAESAR T.E.S.T.
Biodegradation	Biowin TOPKAT Catalogic
Metabolism	Metaprint 2D METEOR TIMES META OECD (Q)SAR toolbox
Physical Chemical properties, log Kow	KOWWIN TOPKAT CAESAR SPARC ACD labs

3.4.1 Freely available tools

EPISuite (Estimation Programs Interface Suite) estimates a range of physicochemical properties, environmental fate parameters and ecotoxicity. It has been developed by the US Environmental Protection Agency (EPA) in collaboration with Syracuse Research Corporation (SRC). EPI Suite is available from the US EPA website: http://www.epa.gov/oppt/exposure/pubs/episuite.htm. The current version of EPI Suite is v4.1. One model that is incorporated is ECOSAR, an expert system of SARs for at least three aquatic species for over 120 different chemical classes. ECOSAR's current version v1.11 is available as a standalone tool. ECOSAR v1.0 is still integrated in EPI Suite v4.1. The standard profile typically contains three acute values, and three chronic values for fish, daphnid, and green algae. Models rely on log Kow for derivation of these acute or chronic values. The log Kow values are estimated using KOWWIN v1.68.

SPARC (SPARC Performs Automated Reasoning in Chemistry; supplied by ARChem), also developed by the US EPA, uses computational algorithms based on fundamental chemical structure theory to estimate a wide variety of reactivity parameters strictly from molecular structure. Parameters include pKa, vapour pressure, boiling point, diffusion coefficient (air and water) as a function of temperature and pressure, molecular volume as a function of temperature, polarisability, solubility, log Kow amongst others. Further information can be found on the website at http://archemcalc.com/sparc.php.

For prediction of toxicity endpoints, CAESAR/VEGA may be used. **CAESAR** was developed during an EU-funded project (http://www.caesar-project.eu) that sought to develop a series of models that would be specifically applicable for REACH. The project aimed at the derivation of five statistical models covering the following endpoints: mutagenicity (Ames), carcinogenicity, developmental toxicity, skin sensitisation, and the bio-concentration factor. These models were implemented into open-source software and made available for online use via the web, or in the case of Ames mutagenicity and developmental toxicity as standalone programs. After the CAESAR project was completed, a new platform called **VEGA** (Virtual models for property Evaluation of chemicals within a Global Architecture) was launched (http://www.vega-qsar.eu/index.php). The models that form the basis for VEGA have been taken from CAESAR or T.E.S.T. (see below), or have been developed later by the contributors to VEGA. VEGA can be interrogated on-line to make predictions of Ames mutagenicity, carcinogenicity, developmental toxicity, skin sensitisation, bio-concentration factor, 96hr LC50 in fathead minnow and log Kow. Alternatively, a standalone client called **VEGANIC** (VEGA Non-Interactive Client) can be downloaded and used off-line. The models may be interrogated to make predictions for the aforementioned endpoints or the inputs used in the application of these models may be used to support a quantitative/qualitative read-across. Functionality is also being developed to enable users to derive their own models based on categorical or continuous data in tools called SARpy and CORAL (correlation and logic models).

T.E.S.T., the Toxicity Estimation Software Tool, is an open-source application developed by the US EPA's National Risk Management Research Laboratory. T.E.S.T allows users to estimate a range of toxicological and physical properties. The toxicological endpoints that are currently in the software include 96-hrs fathead minnow LC50, 48-hrs daphnia magna LC50, Tetrahymena pyriformis 50% IGC50, oral rat LD50, bio-concentration factor, developmental toxicity, and Ames mutagenicity. The physical property endpoints include boiling point, flash point, surface tension, viscosity, density, water solubility, and thermal conductivity. Further information on the models, their training sets are described in the accompanying user guide

(http://www.epa.gov/nrmrl/std/qsar/testuserguide.pdf). The tool is freely downloadable from the EPA website at: http://www.epa.gov/nrmrl/std/qsar/qsar.html#TEST.

Toxtree is a flexible and user-friendly open-source platform that encodes a number of rulebases for the evaluation of toxicity. It was originally commissioned by the JRC to encode the Cramer structural classes that are routinely used as part of a thresholds of toxicological concern (TTC) approach. Since that time, Toxtree has been extended and further developed with other rulebases not only by its original developer, Ideaconsult Ltd (Sofia, Bulgaria) but by other consultants. It is freely available as a download from the JRC website: http://ihcp.jrc.ec.europa.eu/our_labs/computational_toxicology/qsar_tools/toxtree
The current version Toxtree v 2.6.0 (July 2013) includes a broad range of modules:

- Cramer scheme and an extended Cramer scheme;
- Mutagenicity and carcinogenicity rulebase known as the Benigni-Bossa rulebase as well as the ToxMic rulebase on the *in vivo* micronucleus assay;
- Skin sensitisation alerts based on the organic chemistry reaction principles;
- SMARTCyp, a two dimensional method for the prediction of cytochrome P450-mediated metabolism, SMARTCyp predicts which sites in a molecule are labile for metabolism by Cytochromes P450;
- Verhaar scheme for the MOA of aquatic acute toxicity to fish;
- BfR and SICRET rules to predict skin irritation and corrosion;
- BfR rules to predict eye irritation and corrosion;
- ILSI decision tree for the application of the Threshold of Toxicological Concern (TTC) approach;
- START biodegradability, a set of structural alerts compiled by the Canadian EPA for estimating the biodegradability potential of a chemical compound based on structural alerts;
- Michael acceptor profiler which identifies potential Michael acceptors based on structural alerts.

Some of the rulebases encoded in Toxtree have been implemented or re-encoded into the OECD (Q)SAR Toolbox (see below). In essence, Toxtree is effectively an expert system of SARs that can be useful to identify potential hazards but it also plays a useful role in providing the mechanistic information to substantiate read-across.

OpenTox (http://www.opentox.org) is an EU FP7 project on the development of Open Standards and an Open Source platform for predictive toxicology. The goal of OpenTox is to develop an interoperable predictive toxicology framework which may be used as an enabling platform for the creation of predictive toxicology applications. The overall project has many activities, but notable for this report are two applications that have been developed as prototypes, i.e. ToxPredict and ToxCreate. **ToxPredict** is an open-source web platform containing many different models. Some of the models are in common with those encoded in ToxTree v2.5 since one of the contributors and developers of OpenTox is Ideaconsult Ltd (Sofia, Bulgaria). Other tools include developers' own models for acute aquatic toxicity in fish, the OECD chemical categories, Caco-2 models for oral absorption, Lipinksi's rule of 5 as well as the toolbox of Lazar models (see below). **ToxCreate**, on the other hand, is an application for users to develop their own models using the suite of data mining tools and descriptor calculators made available on the website.

Lazy Structure-Activity Relationships (**Lazar**) is an open-source software programme that makes predictions of a range of endpoints using data published by DSSTox. Distributed Structure-Searchable Toxicity (**DSSTox**) Database Network (http://www.epa.gov/ncct/DSSTox) is a project of US EPA's National Center for Computational Toxicology, helping to build a public data foundation for improved structure-activity and predictive toxicology capabilities. The DSSTox website provides a public forum

for publishing downloadable, structure-searchable, standardised chemical structure files associated with chemical inventories or toxicity data sets of environmental relevance. The Lazar models include the following endpoints: Ames mutagenicity, rodent and hamster carcinogenicity, Maximum Recommended Daily Dose (MRDD) and fathead minnow acute aquatic toxicity. The models are based on the use of statistical algorithms for classification (k-nearest neighbours and kernel models) and regression (multi-linear regression and kernel models). Lazar performs an estimation of applicability domain, provides a confidence index for each prediction and presents the nearest neighbours within the training set with their similarity indices and experimental data to provide additional context and confidence in the prediction. A web-based prototype is accessible at http://lazar.in-silico.de. The models developed by Lazar are also available as part of the OpenTox project.

OncoLogic™ is an expert system that assesses the potential of chemicals to cause cancer. OncoLogic™ was developed by the US EPA in collaboration with LogiChem, Inc. It estimates the likely level of concern for carcinogenicity by applying the rules of SAR analysis and incorporating what is known about the mechanisms of action and human epidemiological studies. The Cancer Expert System is composed of four subsystems that evaluate fibres, metals, polymers, and organic chemicals of diverse chemical structures. Chemicals are entered individually and the user needs some knowledge of chemistry in order to select the appropriate structural class (of the 48 or so that are available) in order to make an assessment. There are six concern levels for carcinogenicity within OncoLogic™:

- Low - Unlikely to be carcinogenic
- Marginal - Likely to have equivocal carcinogenic activity or may be weakly carcinogenic at doses at or exceeding maximum tolerated doses
- Low-Moderate - Likely to be weakly carcinogenic, or carcinogenic toward a single target/species, or carcinogenic at relatively high doses
- Moderate - Likely to be a moderately active carcinogen toward one or more target/species
- High-Moderate - Highly likely to be a moderately active carcinogen toward one or more target/species
- High - Highly likely to be a potent carcinogen even at relatively low doses, or carcinogenic toward multiple targets/species

OncoLogic is freely downloadable from the US EPA website: http://www.epa.gov/oppt/sf/pubs/oncologic.htm. The structural classes have also been encoded as fully as possible into the OECD (Q)SAR Toolbox (see below) as one of endpoint profilers.

The OECD has released the **(Q)SAR Applications Toolbox** (http://www.oecd.org/chemicalsafety/risk-assessment/theoecdqsartoolbox.htm). The first version of the Toolbox, released in March 2008, was a proof-of-concept version. The most recent version 3.2 was released in December 2013. The objective of the toolbox is to provide a means for making QSAR technology readily accessible, transparent and less demanding in terms of infrastructure costs. The toolbox presents the process of *in silico* prediction of chemical toxicity as a logical, flexible workflow comprising six sequential steps: (a) chemical input; (b) chemical profiling; (c) selection of SIDS (Screening Information Data Sets; (d) categorization of chemicals; (e) data gap filling; and (f) use of templates for reporting (Combes 2012). The Toolbox also includes a range of profilers to quickly evaluate chemicals for common mechanisms or modes of action. Profiling capabilities include SAR rulebases for skin sensitisation, mutagenicity, aquatic toxicity, bioaccumulation, biodegradation, and other. Profilers also exist to simulate likely metabolites. In order to support read-across and trend analysis, the Toolbox contains numerous databases with results from experimental studies and several external (Q)SARs and thirth party tools.

3.4.2 Commercially available tools

TOPKAT (http://accelrys.com) is a commercial product of Accelrys Inc. that assesses the toxicity of chemicals from 2D molecular structure. (Q)SAR models (so called submodels) are available for different chemical classes and the program automatically selects the equation from the structural input. TOPKAT also makes visible experimental test data for similar analogues if available (taken from the respective training set). For each model, a model specific similarity distance between a query structure and a database compound can be calculated. TOPKAT is able to make predictions for a wide range of endpoints. It is available as a module within Pipeline Pilot and Discovery Studio as part of the portfolio of Accelrys tools. Endpoints covered include aerobic biodegradability, log Kow, acute aquatic toxicity to fathead minnow, acute aquatic toxicity to D. magna, FDA and NTP models for carcinogenicity in mice and rats (both sexes as well as single/multiple tumour sites), WoE carcinogenicity model, Ames mutagenicity, developmental toxicity, rat oral LD50, maximum tolerated dose (MTD) in rats based on feed/water and gavage dosing regimens, chronic LOAEL, skin irritation (negative/mild vs. severe/moderate), eye irritation (to discriminate between severe/moderate and mild/non irritating as well as to discriminate between severe/moderate or mild/non-irritating), skin sensitisation (presence/absence of sensitising potential as well as potency (severe vs. mild/moderate sensitisers). At first TOPKAT was only available as a standalone program. Currently its integration as a component in the Discovery Studio suite (also can be utilised within Pipeline Pilot) affords greater flexibility of the models themselves. Rather than the tools being fixed for predictions, users can now integrate their own data and information and refine and extend the scope of the modelled endpoints.

DEREK (Derek Nexus) (Deductive Estimation of Risk from Existing Knowledge) predicts a number of human toxicological endpoints (https://www.lhasalimited.org/index.php/derek_nexus). Derek Nexus contains 64 new Rapid Prototype alerts which are structural fragments for multiple endpoints and identify adverse effects and alerts the user to their presence. Derek Nexus presents the qualitative likelihood of an effect to occur. An effect may be certain, probable, plausible, equivocal, doubted, improbably, impossible, open, contradicted. In case these qualitative predictions are presented, it can be assumed that the query substance is in its applicability domain. The program may also give rise to predictions denoted as 'Nothing to report'. In these cases, the substance being in or outside the applicability domain is more difficult to evaluate. Derek Nexus' main focus lies in the areas of skin sensitisation, mutagenicity, carcinogenicity. Structural fragments and rules are also available for irritation endpoints, organ toxicity, photosensitisation and methaemoglobinia. For skin sensitisation, skin permeability parameters are included to refine the prediction. In general, effects are predicted for both humans and mammals if differences can be distinguished. Derek Nexus provides arguments on the MOA along with the predictions. Examples and references on which predictions are based are supplied as and when available .

MultiCASE Inc. (http://multicase.com), implements the so-called **CASE** (Computer Automated Structure Evaluation) approach, and is referred to in different ways (MCASE or MC4PC), depending on the software version and computer platform and its successor. The program automatically generates predictive models from datasets provided by the user. It is based on a fragment-based technology sometimes referred to as the CASE approach. The program performs a hierarchical statistical analysis of a database to discover substructures that appear mostly in active molecules thus being with high probability responsible for the observed activity. Initially, it identifies the statistically most significant substructure within the training set. This fragment, labelled the top biophore, is considered responsible for the activity of the largest possible number of active molecules. The active molecules containing this biophore are then removed from the database, and the remaining ones are submitted to

a new analysis for identification of the next biophore. The procedure is repeated until either the activity of all the molecules in the training set has been accounted for or no additional statistically significant substructure can be found. Then for each set of molecules containing a specific biophore, the program identifies additional parameters called modulators, which can be used to derive (Q)SAR within the reduced set of congeneric molecules. The modulators consist of certain substructures or physicochemical parameters that significantly enhance or diminish the activity attributable to the biophore. (Q)SAR are then derived by incorporating the biophores and the modulators into the model. The program includes modules to predict physicochemical properties and a range of toxicological endpoints, including carcinogenicity, mutagenicity, teratogenicity, irritation, developmental toxicity, and acute toxicity. For the endpoints, the software uses its own toxicity scale, from 0 to 100 CASE units, to cover the range from inactive, marginally active and active. In many cases, it is difficult to relate these CASE units to traditional measures of toxicity. Predictions generated from MCASE models have been collected by the Danish EPA and are available within the OECD (Q)SAR Toolbox (see below).

The Tissue MEtabolism Simulator (**TIMES**), developed by LMC (Bourgas University, Bulgaria; http://oasis-lmc.org) is a platform which encodes structure toxicity and structure metabolism relationships through a number of transformations simulating metabolism and interaction of the generated reactive metabolites with cellular nucleophiles. The metabolism simulators mimic metabolism using 2D structural information. Metabolic pathways are generated based on a set of hierarchically ordered principal transformations including spontaneous reactions and enzyme-catalysed reactions (phase I and II). The covalent reactions with nucleophiles are described by alerting groups (structural alerts). Some of these alerts are additionally underpinned by mechanistically based 3D-QSARs to refine the predictions. These 3D-QSAR models depend on both the structural alert and factors that influence its reactivity: steric effects, molecular size, shape, solubility, lipophilicity and electronic properties. The models within TIMES that are driven by metabolism simulators are those for Ames mutagenicity, *in vitro* chromosomal aberration, skin sensitisation and oestrogen receptor affinity. Work is on-going to develop and refine a model for the *in vivo* micronucleus assay. Other models that are incorporated within TIMES but which do not necessarily account for metabolic transformation include models for receptor-binding affinities (oestrogen, androgen and aryl hydrocarbon receptors), phototoxicity as well as acute aquatic toxicity to different species including fathead minnow, daphnia and microorganisms such as T.-pyriformis etc.

Catalogic is a sister platform to TIMES, also developed by LMC covers environmental fate properties notably biodegradation according to different OECD Test guidelines (301C, 301B, 301F). Models also exist for bio-concentration and half-life in fish. Catalogic models incorporate a microbial metabolism simulator. The metabolism simulators encoded within the TIMES and Catalogic tools are additionally available within the OECD (Q)SAR Toolbox (see below). Metabolic trees, quantities, probability or schemes are not presented but the predicted metabolites formed are shown.

The **ACD/Tox Suite** (formerly called **ToxBoxes**), provided by ACD/Labs (which merged with Pharma Algorithms in 2009), provides predictions of various toxicity endpoints including genotoxicity, acute toxicity, aquatic toxicity, eye/skin irritation and endocrine system disruption (http://www.acdlabs.com/products/pc_admet/tox/tox). The predictions are associated with confidence intervals and probabilities, thereby providing a numerical expression of prediction reliability. The software incorporates the ability to identify and visualise specific structural toxicophores, giving insight as to which parts of the molecule are responsible for the toxic effect. It also identifies analogues from its training set, which can

also increase confidence in the prediction. The algorithms and datasets are not disclosed. ACD/Tox Suite has now been replaced by **ACD/Percepta** (http://acdlabs.com/products/percepta) and will be gradually phased out by the end of 2014. ACD/Percepta unifies the software platforms for the physicochemical, ADME and toxicity prediction modules, and has better reporting abilities. It also offers the possibility to view the calculation protocols, easy integration with other software platforms, as well as unrestricted access to all information from individual prediction models and from the consensus model (for physicochemical predictions).

CompuDrug's **HazardExpert** (http://www.compudrug.com/hazardexpertpro) is a software tool for initial estimation of toxic symptoms of organic compounds in humans and in animals. HazardExpert accepts user input and is able to consider bioavailability of the compounds. HazardExpert predicts the toxicity of organic compounds based on toxic fragments, results are given for seven different toxicity classes: oncogenicity, mutagenicity, teratogenicity, membrane irritation, sensitivity, immunotoxicity, neurotoxicity. Also bioavailability calculation based on pKa and logP, bioaccumulation calculation, toxicity prediction for metabolites is available.

4 Opportunities and pitfalls of innovative testing strategies

Current applications of *in vitro* models and non-testing strategies for hazard evaluation are discussed in this chapter, along with their opportunities and pitfalls.

4.1 *In vitro* toxicity testing

Chapter 2 showed that a large array of *in vitro* assays have been and still are developed. The strengths and weaknesses of currently developed tests and expected improvements and implementations of *in vitro* assays are discussed below.

4.1.1 Predictive value

In vitro approaches are internationally regarded as the way forward in the area of human toxicity testing. Currently, many of the available *in vitro* toxicity tests are however still at the research and development stage. Only a small number of *in vitro* tests addressing specific endpoints have at present been formally validated and gained regulatory acceptance. A major drawback of primary cell cultures is their limited lifespan. Effects of long term exposure can thus not be measured. In addition, cells do not always provide stable phenotypes. Cell lines may undergo dedifferentiation and lose the specific functional properties of the cells *in vivo* or express uncharacteristic functional features. Biochemistry, gene expression, and responses to chemical challenges may differ to those of the target cell *in vivo*. Metabolism can be addressed to some extent by adding exogenous metabolizing fractions or by using metabolically competent cells. However, even if initially some metabolism is present, deterioration can occur as a function of culture time. *In vitro* cell culture systems may thus poorly resemble their *in vivo* equivalents (Adler et al. 2011).

The predictive value of *in vitro* models can be improved by extending metabolic capacity, realizing more physiologic culture conditions such as homeostasis, oxygen supply, and cell density, producing organotypic 3-dimensional (co)-cultures, using cells from human origin and stem cell-derived systems, better understanding of mechanism of action thereby refining and expanding endpoints measured, and appropriate statistics and prediction models. A substantial amount of research is currently being conducted on the potential use of stem cells as *in vitro* models for research and safety testing. Different types of differentiated cells can be obtained from embryonic and adult progenitor/stem cells of different species and also from induced pluripotent cells (iPSC). iPSC can be a source of potentially all tissues with a variety of genetic variability, and they renew themselves for a long period of time. iPSCs may thus have great potential for predicting toxicity. However, the culture of stem cells is not trivial, the differentiation rarely occurs in 100% of the population, and not all of the cells will be in the same stage of full differentiation (Basketter et al. 2012).

Existing *in vitro* assays may be useful for identifying the potential adverse effects of substances (hazard identification) to a limited number of targets or for obtaining mechanistic information. *In vitro* assays are in many cases used as a screening test.

> ***Screening test*** *An often rapid and simple test method conducted for the purpose of classifying substances into a general category of hazard*

None of the available *in vitro* tests is currently seen as appropriate for providing all information needed for quantitative safety assessment related to different classes of toxicity. Single cell based assays do not represent adequately the complex interplay between different cell types and the involvement of mediators released by other tissues (immune system or endocrine system). Possible feedback loops present at higher levels of biological complexity are thus not taken into account. Research efforts considering interactions between different biological tissues and systems, which would be more representative of the situation in the human body, have only recently been initiated (Adler et al. 2011). Besides, the target of a toxicant, its mechanism of action, and the toxic outcome may differ per organ or dose, duration and timing of exposure. A single *in vitro* test is therefore often a too much simplified picture of reality, especially when the pathology *in vivo* is complex such as is the case for developmental toxicity and carcinogenicity. Ideally, an array of complementary *in vitro* tests (a 'test battery'), in which each test evaluates a different component of a multi-factorial toxic effect, should be applied to indicate effects on the most sensitive and/or most relevant toxicity endpoints, which may be inferred from historic test results in databases.

> ***Test battery*** *A series of complementary test systems usually performed at the same time or in close sequence to predict of toxicity*

It should be emphasized that there are many more mechanisms of action and target organs than the ones assessed currently, for which unfortunately less or no *in vitro* methods are available. Moreover, the mode of action and most sensitive endpoint of toxicants are not always known (Adler et al. 2011).

4.1.2 Evaluation of molecular pathways

Many *in vitro* assays investigate a cellular mechanism which finally might or might not result in an adverse effect *in vivo*. Besides, it is not always clear whether effects that are detected are causal factors in the development of toxicity or rather associated with it. These limitations might be overcome as the understanding of pathways of toxicity (PoT) increases (t4 2013).

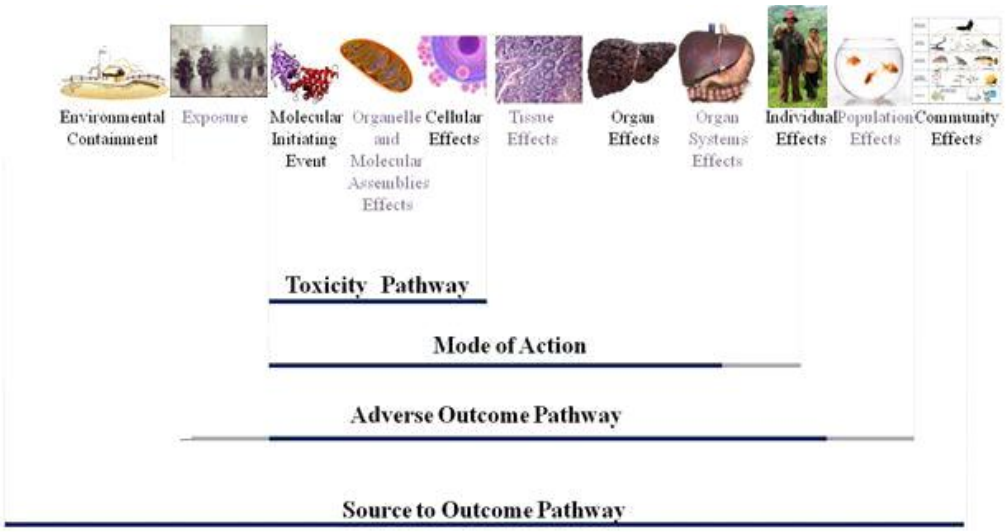

Figure 18. Representation of the relationships between Toxicity Pathways (or Pathways of Toxicity), Mode of Action Pathways, Adverse Outcome Pathways, and Source to Outcome Pathways. The black bars represent the breadth of research common to these concepts. The grey bars represent the theoretical extent of the concepts (OECD 2013).

> ***Pathway of toxicity*** *Molecular or cellular pathway that leads to adverse cellular effects when perturbed*

Mapping of PoT will show how chemicals cause toxicity, as opposed to merely what biological effects they might cause (Basketter et al. 2012). Adverse outcome pathways (AOPs) describe a broader pathway of events (**Figure 18**) (Ankley et al. 2010).

> ***Adverse outcome pathway*** *Series of events starting with a molecular initiating event (MIE) in which a chemical interacts with a biological target, leading to a sequential series of effects to produce an adverse outcome, and the associated biological responses that result in effects at the subcellular, cellular, tissue, organ, whole animal, and population levels*

The OECD has recently published a guidance document on developing and assessing AOPs, that intends to provide insight into which pieces of information are necessary to identify and document an AOP (**Figure 19**) and how to present them (OECD 2013). The intention of the use of AOPs is to integrate all available information and to link initial mechanistic knowledge to the prediction of hazard for humans. The AOPs concept however needs fine tuning. For instance, parallel cascades and crossing of pathways may be involved in toxicological processes, and classical kinetic determinants as well as more specific events, such as hormonal influences and adaptive responses, must be considered in AOP development (Vinken 2013).

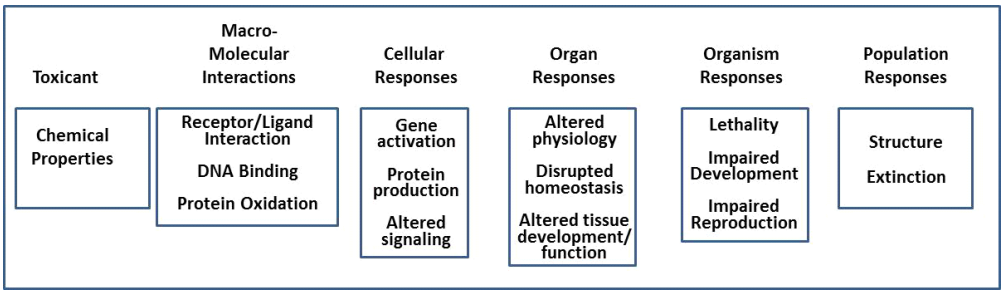

Figure 19. Components of the Adverse Outcome Pathway (AOP) (OECD 2013).

The increasing knowledge of biological systems and advances in molecular and computational tools, coupled with the concomitant development of high-throughput and high-content screening assays, is providing mechanistic insight into the mode of action in target cells, tissues, and organs This enables establishment and applications of PoT and AOPs. A number of PoT has been identified already; however, most PoT are only partially known, and no common annotation exists. The U.S. Environmental Protection Agency (EPA) ToxCast project focuses at identifying toxicity pathways for a large range of substances and substance classes using multiple biological high-throughput screening assays.

> ***ToxCast*** *A multi-year effort that uses automated chemical screening technologies to expose living cells or isolated proteins to chemicals in order to assess changes in biological activity that may suggest potential toxic effects and eventually potential adverse health effects*

ToxCast results are contributed to the federal agency collaboration called Toxicity Testing in the 21st Century (Tox21). Tox21 pools chemical research, data and screening tools from multiple agencies including the National Toxicology Program (NTP)/National Institute of

Environmental Health Science (NIEHS), National Center for Advancing Translational Sciences (NCATS), and the Food and Drug Administration (FDA). So far, Tox21 has compiled high-throughput screening data on nearly ten thousand chemicals. All data are publicly available. The approach is, however, limited by the use of known PoT and available tests. Unsupervised identification of PoT would be the logical complement to this approach.

In recent years, a range of 'omics tools (listed below), each with specific uses, have been introduced (Balls et al. 2012). Transcriptomics has been applied to *in vitro* models of human and rodent cells for the purpose of predicting toxicity. This so called toxicogenomics approach assesses the impact of chemical exposure on gene and protein expression and can indicate potential pathways by which a chemical may act. Compared to often used reporter gene assays, transcriptomics analysis can evaluate effects on a larger number of genes and more than one cellular pathway, thereby assessing multiple endpoints at the same time and allowing unbiased testing. When developing transcriptomics-based screens, gene expression data are derived from exposure of model systems (such as cellular models) to well-known toxicants belonging to specified classes of toxicity. On these data sets, bioinformatics tools are applied to identify gene expression profiles corresponding to different types of toxicity. These profiles are compared to a set of gene expression changes elicited by a suspected toxicant. If the characteristics match, a putative mechanism of action can be assigned to the unknown agent and hazard is predicted. Transcriptomics approaches have been shown to be able to accurately predict *in vivo* toxicity in rodents, although the number of chemicals evaluated is still limited and may not represent the full spectrum of toxicants. No formal validation of transcriptomics techniques has been performed. The technology has however been extensively evaluated by the MicroArray Quality Consortium (MAQC). For some tests based on gene expression analyses, standard protocols are being developed and optimised. Transcriptomics data are stored in databases such as ToxRefDB, Chemical Effects in Biological Systems (CEBS), and the Comparative Toxicogenomics Database (CTB). The Tox21 program aims to map all biochemical pathways implicated in toxic responses (the 'toxome') by transcriptomics analysis of thousands of compounds.

Varieties of 'omics tools

cellomics (about phenotype and functions at the cellular level)
cytomics (distinguishable from cellomics by its application at the single cell level)
epigenomics (about the parts of the genome, other than the DNA code, which modulate the operation of the genome)
interactomics (about interactions and their consequences among proteins and other molecules within a cell)
metabolomics (about the chemical processes involving metabolites), which is related to:
metabonomics (about quantitative, dynamic and multiparametric metabolic responses)
pharmacogenomics (a generic term, as referred to above)
phenomics (about the functional biochemical and physiological characterisation of cells, tissues and organisms in response to genetic changes and environmental influences)
proteomics (about the proteome, the entire complement of proteins and about their individual production, modification and functions)
toxicogenomics (about responses to toxic substances)
transcriptomics (about all the types of RNA, including mRNA, rRNA and tRNA, as applied to the total set of transcripts or a specific subset)

Some difficulties of the toxicogenomics approach are that the function of many genes is not yet understood, and activation of early response genes and repair processes are often detected, whereas these responses are associated with but do not actually cause toxicity and are therefore uninformative with regard to the mode of action. It may also be complicated to take into account dose and spatial and temporal effects when interpreting results. Scientifically established links between transcriptomics responses and adverse outcomes at a level of biological organization relevant to risk assessment need to be established, which is also referred to as phenotypic anchoring (ECETOC 2013).

> ***Phenotypic anchoring*** *Relating specific alterations in gene expression profiles to specific adverse effects of environmental stresses defined by conventional parameters of toxicity*

The integration of transcriptomics data with PoT analysis will improve the predictive capacity of assays by establishing the biological relevance of the pathways involved and their quantitative contribution to phenotypic changes (Adler et al. 2011). Besides, based on PoT information, predictive *in vitro* models could be designed to focus at key events in pathways that are mechanistically relevant to and predictive of adverse effect in humans (biomarkers of toxicity) (t4 2012). This could transform toxicity testing from a system based whole-animal testing approach to one founded primarily on *in vitro* methods that evaluate changes in molecular processes. In addition, proteomics and metabolomics analyses may provide insight into perturbations in protein structure and function and chemical processes within the cell, respectively. When added to transcriptomics tools, this allows a systems biology approach (ECETOC 2013).

> ***Systems biology*** *A multidisciplinary approach considering the interactions between the components of a biological system and combining this knowledge to provide a better, holistic understanding of the organism or of the phenomenon being considered*

Moreover, novel imaging technologies and physiological analyses provide the possibility for continuous observation of major cellular events such as migration, proliferation, cell morphology, cell–cell interactions, and colony formation. While these technologies hold considerable promise, further development is needed before they can be applied for quantitative risk assessment. Experience and inter-laboratory validations should identify the most sensitive and robust tests and platforms. These approaches do require specialized staff and expensive equipment. Besides, it must be noted that whatever innovative analysis is used, it can hardly overcome the limitations of the underlying *in vitro* model so the same considerations of the limits of cell models apply (Basketter et al. 2012).

4.1.3 Dose response characterization

Characterization of the free chemical concentration in cell-based assays is crucial for data interpretation. However, *in vitro* screenings often do not consider the actual (as opposed to nominal) *in vitro* concentration, bioavailability, and degradation of compounds. Frequently, synthesized compounds are not stable at 37°C and/or bind to plastics or media proteins (Basketter et al. 2012). Besides the assessment of the actual chemical concentration in the *in vitro* system, the level of response indicating adversity needs to be defined. It should be taken into account that often it is not a new pathway that is activated, but rather a normal signal that is disrupted by the toxicant, or that an adaptive response instead of a toxic one is detected. A proper quantification of both the dose and the response (e.g. the degree of perturbation of cell signalling) will help in determining an appropriate point of departure (PoD) for quantitative *in vitro* to *in vivo* extrapolations (QIVIVE).

Quantitative in vitro to in vivo extrapolation *The translation of a concentration-effect relationship in vitro to a dose-effect relationship in an intact organism*

This process, also referred to as "reverse dosimetry", implies the consideration of a compound's kinetic behaviour in the organism and results in the estimation of a dose (or any appropriate exposure scenario) *in vivo* that would result in the effective concentration at the site of toxic action. Depending on the features of the test system and the nature of the biomarker of toxicity chosen, the concentration used for QIVIVE may differ. For instance, the minimal significant effect concentrations corresponding to the lowest observed effect level (LOEL) of *in vivo* toxicity are relevant if mutations are chosen as endpoint of interest. In many cases the EC50 (the concentration that results in a half-maximal response) values may be a good choice as this parameter is the mathematically most robust data point to determine. Furthermore, the EC60 or EC90 value could be determined for instance in cases when cells have large reserve/buffering capacity, e.g., for glutathione depletion or ATP depletion as biomarker. Also, modelling the concentration-effect relationship by means of the benchmark approach may be considered. The benchmark dose (BMD) approach estimates a dose that produces some predetermined, and presumably biologically relevant, increase in the response over the control.

Benchmark approach *Mathematical dose-response modelling that takes sample size, confidence interval around data points, and shape of the curve into account*

As a PoD, the BMCL10 can for instance be used: the benchmark concentration-lower limit of confidence for 10% of the maximal response (t4 2012). **Figure 20** illustrates the different approaches to derive the PoD for quantitative *in vitro* to *in vivo* extrapolations.

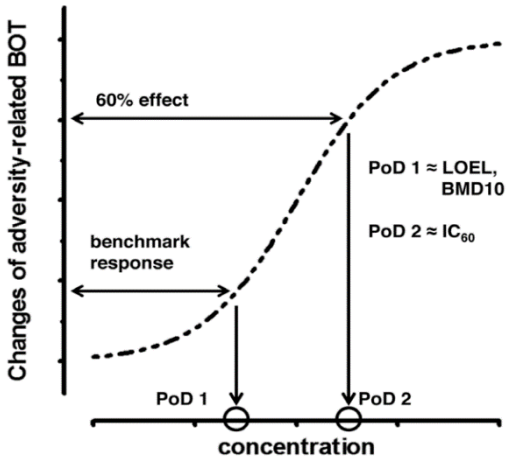

*Figure 20. Modelling and determination of the point of departure (PoD for quantitative in vitro to in vivo extrapolation). The concentration-dependent change of a biomarker of toxicity (BoT), which is related to the hazard of the respective chemical tested, is shown. It is not pre-defined which part of the curve is most appropriate for determination of the PoD. This information depends on the underlying biology and toxicological relevance. Two examples are given: PoD1 is defined by the BMD**10**, i.e., the lowest concentration that results in a 10% change of the baseline; PoD2 is defined as the IC**60** of the curve. Concentrations smaller than the PoD would be considered non-significant with respect to hazard-prediction. Different statistical and modelling approaches are available to define such PoD (t4 2012).*

Extrapolation of data produced in *in vitro* studies, and the PoD derived from these data, to *in vivo* conditions is a great obstacle to the use of *in vitro* systems. For quantitative risk assessment, better and more scientific knowledge on exposure, toxicokinetics, dose response and mechanisms of toxicity are needed. There are currently no reliable ways to quantitatively extrapolate dose–response data from concentrations of the test substance in *in vitro* experimental systems to toxic exposure levels in the whole body. Improved toxicokinetic models could allow correct estimation of the impact of the distribution of a substance over time and of the internal dose that results from the level of repeated external exposure. Such models may aid in translating doses of *in vitro* studies (in mM) into doses in *in vivo* tests (mg/kg body weight/day) and in the understanding of dose responses. This information is required to use *in vitro* data to estimate potency and perform risk assessment to identify safety thresholds. PBTK models may be applicable to predict concentrations of test substance and their metabolites in tissues, although these approaches are data-intensive (Coecke et al. 2013).

4.2 Non-testing approaches

In Chapter 3, examples of *in silico* hazard assessment approaches and related tools were shown. The predictive value of such models and ways to select the most appropriate tools are addressed below.

4.2.1 Predictive value

It should be possible to develop computer prediction systems for any type of (biological) activity provided suitable data for developing models are available and the chemical structure of the test material is known (Combes 2012). However, fundamental to all non-testing strategies (and sometimes quite limiting) are the size and quality of the databases on which these methods are based. The judgments made using chemical categorizing, read across and structure activity relationships rely heavily on the availability of existing high quality experimental toxicological data. At present, much more data are available on pharmaceuticals than on environmental contaminants (Basketter et al. 2012). In contrast to pharmaceuticals, environmental chemicals have a broad chemicals space, may not have discrete intended biological targets (with the exception of pesticides and biocides), usually do not exhibit high-affinity interactions with molecular targets, often have largely unknown metabolic patterns and can have highly variable patterns of use that may result in highly variable exposures (IGHRC 2013). Also, much more data are available on genotoxic carcinogens than on chemicals causing other types of toxicity. For complex biological endpoints, use of non-testing data currently represents an extreme challenge due to the multitude of possible mechanisms of action. The different models therefore are not overly realistic because they may ignore essential processes (Basketter et al. 2012). For ADME processes, there is a scarcity of high quality and relevant experimental data for use in model building (JRC 2010). Patlewicz et al. (2013) state that physicochemical parameters are crucial for the prediction of any biological activity. Key properties are Log Kow, vapour pressure, water solubility, molecular weight, as well as pKa and log D for ionisable substances. It is recommended that this type of data is always generated unless it is technically not feasible or not relevant for the specific substance under consideration.

With regard to data quality, training sets of chemicals are often biased toward toxic chemicals since negative data are difficult to interpret and/or to obtain. In addition, available data have not always been obtained from properly conducted studies with test samples of high purity. Besides, some models have been developed with information that is not in the public domain and that cannot be verified. It should also be kept in mind, that subtle changes in molecular structure have dramatic effects on activity. When new data are produced, *in silico* models can readily be updated and improved (Combes 2012).

Basing correlations between physicochemical properties and activity on mechanistic information maximizes the predictivity of models for novel chemicals (Combes 2012). For some endpoints, such as Ames mutagenicity, skin/eye irritation or skin sensitisation, the presumed mechanism of action has been reasonably established and structural rules/profilers have been encoded in (Q)SAR models or tools. For other endpoints, particularly repeated dose toxicity, adequate mechanistic information may be unavailable and/or too heterogeneous to predict by a single model (Patlewicz et al. 2013). Hence, (Q)SARs may need to be more focused on specific mechanisms, and a suite of (Q)SARs will probably be necessary for many endpoints. Information on other endpoints, ADME information, and AOP information can be helpful in substantiating the predictions made. The OECD Toolbox is pursuing an adverse outcome pathway (AOP) approach for repeated-dose toxicity by encoding a new repeated-dose profiler based on data received from Japan (HESS database). Knowledge of the mode or mechanism of action can provide also greater confidence in a read-across estimate (Patlewicz et al. 2013).

A limitation of all the non-testing approaches is that they can only be readily applied to discrete (groups of) substances (Basketter et al. 2012). In reviewing and comparing expert systems and QSAR models, it is essential that the applicability domain (AD) is defined. Methods for AD determination have been reviewed, but consensus on how ADs should be defined is still lacking. It is important that a model should only be used to predict the activities of chemicals within its AD, as it cannot be expected to correctly predict the activity of molecules outside its AD (Combes 2012).

While *in silico* toxicity prediction methods have been greatly improved in recent years, they have not yet become widely accepted by toxicologists and regulatory agencies. This is partly due to the fact that the validation of models is controversial and no models have been validated to internationally-agreed criteria in a fully independent and transparent way. Although few problems with reproducibility and interlaboratory transferability exist, it is difficult to find suitable sets of test chemicals for external validation, since new chemicals are easily and continually being added to the training set iteratively, in order to refine the model (Combs 2012). Concrete examples to advance acceptance of QSAR models under regulatory frameworks by establishing consistent approaches dependent on context for each chemical and endpoint under consideration are needed (Basketter et al. 2012). Read-across approaches are more commonly used than (Q)SARs and conceptually accepted by the ECHA and Member State Authorities but difficulties remain in applying them consistently in practice. Technical guidance is available and there are a plethora of models and tools that can assist in the development of categories and read-across, but guidance on how to practically apply categorisation approaches is still missing. Absence of toxicity remains a particular challenge to justify using read-across and other *in silico* tools (IGHRC 2013, Patlewicz et al. 2013). Besides, application of QSAR analysis on a routine basis may be hampered by lack of the necessary in-house modeling expertise and adequate familiarity with more than one software package. Predictions should always be interpreted by an expert with knowledge of the endpoint and an appreciation of the strengths and limitations of the specific model applied (JRC 2010).

Overall, it can be concluded that predictions for individual chemicals cannot replace experimental testing. The main value of *in silico* tools is in early screening of chemicals and in supporting information of different nature and from different sources. *In silico* models are complementary themselves as well: the combined positive predictions from different programs has been shown to have better overall performance than when any program was used alone (Comes 2012). Models with high specificity can be supplemented with models with high sensitivity and knowledge-based tools with statistical models, and models using

different molecular descriptors for the same toxicity endpoint can be combined. However, identification of optimal model combinations (batteries) still requires considerable research (JRC 2010). Non-testing information can aid in priority setting, mechanistic studies, and estimation of potency (Basketter et al. 2012, Combes 2012). Modelling the effects of mixtures of chemicals is however considered to be extremely complicated (Combes 2012).

4.2.2 Selection of tools

When selecting non-testing approaches for hazard assessment, the first consideration should be whether such approaches permit an accurate and credible assessment of the hazard for the substance in question. The use of a non-testing approach to address data gaps should be fit for purpose (e.g. include the endpoint of interest) and well supported with sound and robust justification (ECETOC 2012). The assessment of the usefulness of a model, and the adequacy of its predictions, is not a trivial exercise that requires a reasonable amount of QSAR knowledge, and needs to be performed on a case-by-case basis, making generalisations on model applicability difficult (JRC 2010). A model should mostly be scientifically valid and applicable to the chemical of interest with the necessary level of reliability. The OECD (2007) has determined that validated (Q)SAR models are defined by the following criteria:

1. A defined endpoint
2. An unambiguous algorithm
3. A defined domain of applicability
4. Appropriate measures of goodness-of-fit, robustness and predictivity
5. A mechanistic interpretation, if possible.

A second essential step is to demonstrate the applicability of the model to the chemical of interest. The evaluation of model applicability is related to the evaluation of the reliability of prediction for the chemical of interest. The assessment of whether a given model is applicable to a given chemical can be broken down into the following specific questions (JRC 2010):

1. Is the chemical of interest within the scope of the model, according to the defined applicability domain of the model?
2. Is the defined applicability domain suitable for the purpose of the predicition?
3. How well does the model predict chemicals that are similar to the chemical of interest?
4. Is the model estimate reasonable, taking into account other information?

The most useful models are thus commonly associated with transparent documentation on the model development and validation process, and in addition they should be implemented in software tools (JRC 2010). The ANTARES project (http://www.antares-life.eu) aimed to analyze the use of non-testing methods in accordance to REACH, and to identify suitable methods. Within this project, relevant criteria for comparing different QSAR methods were identified, divided into main and additional criteria (see Table 2) (Milan et al. 2010). Among several evaluations, we highlight the comparison of QSAR models for prediction of mutagenicity. The accuracy of the models with the best performance value was similar to the *in vitro* reproducibility of Ames test, which is estimated to be 0.85. The sensitivity of the best models was between 0.90 and 0.95. This suggests that the *in silico* models can achieve a high prediction performance. However, it was also concluded that the different *in silico* models are based on different approaches, and their results are therefore not identical. The user must assess the results closely, considering not only the predicted value but also all information provided, such as the selection of similar compounds and the applicability

domain. Thus, careful use of this information and selection of the most reliable results improves the overall reliability of the model (Antares, delivery report 23)

Table 2. Main and additional criteria for selection of QSAR tools within ANTARES (Milan et al. 2010).

Main criteria	Additional criteria
• Data quality • Number of chemicals • Definition of descriptors/fragments • Description of the algorithm • Description of the applicability domain • Performance • Internal validation • Output format • Costs	• Analysis of batch of chemicals possible • Explicit structure format • Verification of the presence of uncertainty • Further adequate and reliable documentation • Usability/user friendly • Comprehensiveness • Skill requested to interpret results • Access • Platform/software requirements • Connection problems • Time needed

In the context of the PESTISAR project, which studied the potential applicability of computational methods in the evaluation of the toxicological relevance of metabolites and degradates of pesticide active substances, the JRC selected a range of user-friendly freely and commercially available tools for non-testing strategies that are suitable for the non-specialist user (JRC 2010). Freely available tools to predict physiochemical properties include EpiSuite and SPARC. Commercial variants of these tools are Accord for Excel with ADME/Tox Add-on , ACD, and ADMEBoxes. TOPKAT (commercial), Lazar and Toxtree (for free) can be used for prediction of chronic toxicity. For genotoxicity, a vast range of commercial models are available, mostly for chemicals that are electrophilic and DNA-reactive. Freely available tools are limited: online ToxBoxes application, Lazar, Toxtree, CAESAR, and the OECD QSAR Toolbox. For carcinogenicity, commercial tools include Derek, TOPKAT, HazardExpert, and OASIS TIMES. Freely available tools are, again, limited: Lazar, Oncologic, and Toxtree. For developmental toxicity and reproductive effects, TOPKAT, Derek, and Leadscope are among the commercial tools, whereas only CEASAR is freely available. For endocrine disruption, various tools such as ToxBoxes and TIMES are commercially available. Simple and freely available decision tree approaches for screening and priority setting, especially the US EPA decision tree which is being implemented in the OECD QSAR Toolbox, have also been found useful. Commercially available tools for ADME predicitions are ACD/ADME Suite, MetabolExpert, Meteor, Accord for Excel with ADME/Tox Add-on, and Symcyp. Feely available tools for this endpoint are almost completely lacking. Many of the ADME software tools are not transparent in terms of their predictive algorithms or underlying datasets.

The advantage of using commercial systems is the provision of customer support and training. The most commonly used commercial software tool is TOPKAT. Despite the lack of transparency in its predictions, several studies have shown that it gives reasonable predictions for a range of chemicals (including pesticides and industrial chemicals) (JRC 2010). The tool is complex and requires some training, but is still considered to be user-friendly. This tool is used by Shell and AkzoNobel for prediction of human and ecotoxicology. AkzoNobel additionally uses DEREK (which is linked to METEOR) and ACD ToxSuite for human health endpoints, and ACD ToxSuite for prediction of physicochemical properties and ecotoxicity as well. These tools are, together with TOPKAT, at the moment considered to be the most widely applied commercial systems for the prediction of toxicological endpoints. However, ACD ToxSuite covers only a limited number of endpoints. Both AkzoNobel and

Shell also make use of the freely available VEGA tool, which is stated to be user-friendly and deliver reliable data, and OECD QSAR Toolbox for prediction of physicochemical properties and (eco)toxicity. AkzoNobel also applies EPISUITE and ALOGPS for physicochemical properties and EPiSUITE for ecotoxicity. The OECD QSAR Toolbox and EPISUITE are currently standard applications with high use in regulatory settings. According to AkzoNobel, of all these tools, the OECD QSAR Toolbox performs best with regard to data gap filling, incorporation of mechanism of action and experimental data, CAS number search, batch analysis, addition of own data, reporting format, and literature references. AkzoNobel however concludes that, since the systems are all different in their approaches, it cannot be stated that one is better or more appropriate than the other. The greatest benefit is expected to be acquired when using multiple tools, which leads to a higher validity and trust in the predictions when compared to each of them separately (personal communication).

When the hazard of the substance is misclassified, the assessment may be either too conservative or not conservative enough (ECETOC 2012). When evaluating data predicted *by in silico* tools, the principle of 'proportionality' and principle of 'caution' should therefore be taken into account in decision making (IGHRC 2013). The principle of proportionality means that the amount of information needed is dependent on the importance of the decision that will result from it. Thus, for screening purposes, less information may be sufficient compared to regulatory approaches. For drinking water applications, however, fast and reliable data are asked for to identity potential hazard and direct further specific testing. Here, the second principle of caution is important which states that the amount of information needed is dependent on the risk the substance poses: the more severe the consequences, the more conservative the approach.

4.3 Integration of testing strategies

The different testing strategies outlined in this report can benefit from each other (Bradbury et al., 2003). This is schematically represented in **Figure 21**. AOPs can for instance show which processes drive toxicity, and this information can be used to focus the development of *in vitro* test methods and to incorporate relevant mechanistic and biochemical factors contributing to toxicity. The disruption of AOPs by test substances can be used in priority setting for further testing. An AOP could moreover assist in determining what further information (and therefore, which test, if any), would increase the certainty of linking the initiating event and adverse effect(s). Moreover, a well-established AOP can be used for species-to-species comparisons. In return, a collection of *in vivo* and *in vitro* information, data from high-throughput screening (HTS) assays, endpoints from high-content screening (HCS) (such as 'omics) approaches, and outcomes of *in silico* methods, may be used to provide support and data to build and evaluate AOPs (t4 2012).

As the MIE in each AOP involves a rather specific interaction of chemicals with biological systems, AOPs can also be used as the basis for generating mechanistically based structure–activity relationships and categories of chemicals (Vinken 2013). In the future, AOPs will be included in the OECD Toolbox for this purpose (OECD 2013). In its turn, the presence or absence of structural features allowing direct reactivity or activation via metabolism can indicate a potential health hazard (Basketter et al. 2012). Since knowledge on chemical structure can be used to predict possible toxicological targets, structural and physicochemical properties of compounds can be the basis for selecting a proper *in vitro* test battery (t4 2012). Since QSAR tools are relatively inexpensive and quick as compared to toxicity testing and do not require expensive laboratory equipment and facilities, they are also very useful for compound prioritization for toxicity testing (Combes 20120).

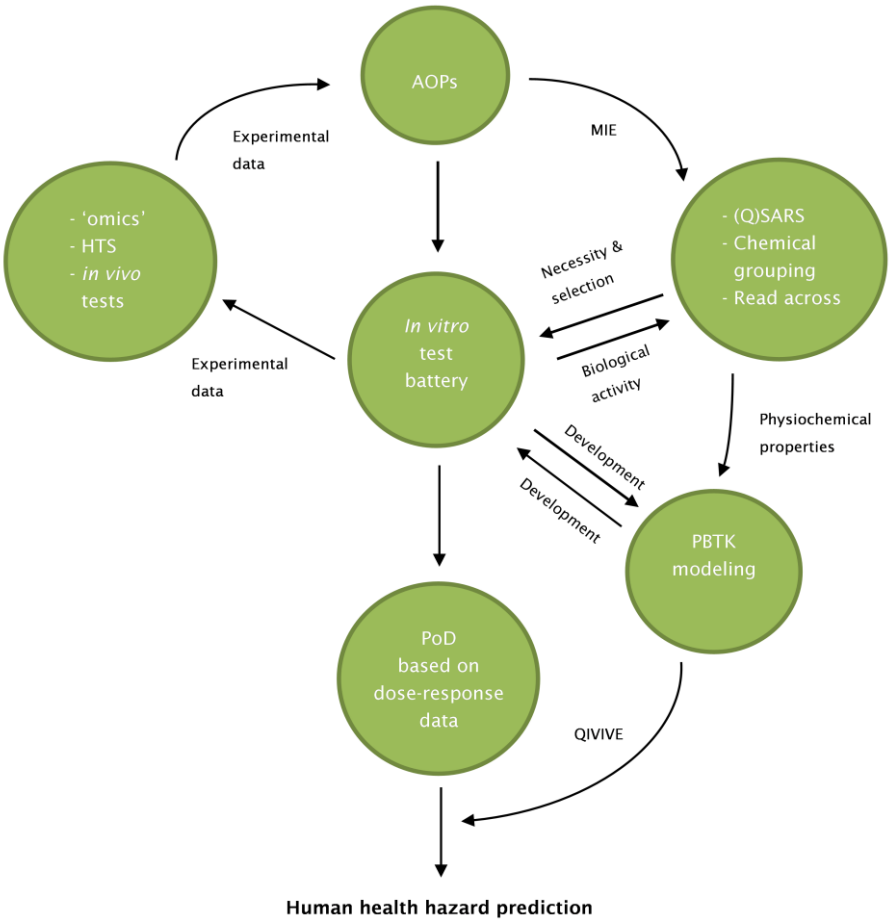

Figure 21. Combination of testing strategies outlined in this report to improve in vitro and non-testing approaches and human health hazard prediction.

Increasingly, kinetic information will also be integrated with other testing approaches. First, high-quality data are needed as input for PBTK models. These data should and can be generated with *in vitro* or *in silico* approaches that allow quantification of specific dose-response curves. PBTK models can be coupled to *in silico* models for partition coefficients, absorption or excretion rate constants and computer models of metabolism (assuming that such models are available for the chemical class of interest). Using expected exposure patterns, estimates of internal exposures, bioavailability, half-life, etc. can be obtained. Such results could either be sufficient to answer the question of interest (for instance, if an *in vivo* exposure level at the target site would be determined below which no adverse effect is expected, further testing could be avoided) or would provide estimates of concentration levels to be assayed *in vitro* (i.e. the dose or exposure concentration needed to achieve given target organ concentrations observed to produce toxic effects at the cell or organ level *in vivo*). This exercise aids in proper design of *in vitro* experiments. In further steps, PBTK models can incorporate the results of specific *in vitro* estimates of pharmacokinetic parameters (such as absorption rates, metabolic rate constants, etc.) to provide estimates of internal dose levels attained in predefined exposures scenarios, enabling a prediction of the most sensitive toxic endpoint, exposure-response relationships, no-effect levels (if the dynamic models provide toxicity thresholds), etc. Finally, toxicokinetic modelling can be used to extrapolate *in vitro* to *in vivo* internal doses.

As outlined above, the information produced by the different innovative hazard assessment approaches can be combined for the further development of the tools. Another way in which the different tools can be of added value, is to systematically combine multiple information sources in integrated testing strategies (ITS; van Leeuwen et al., 2007).

> ***Integrated testing strategy*** *Information-gathering and generating strategy, which in itself does not have to provide means of using the information to address a specific regulatory question*

This is seen as the most promising way forward for toxicity testing (Hartung et al. 2013). An ITS might include a strategic combination of non-testing methods (computer based tools such as QSARs, non animal-based ADME-models, etc.), *in vitro* assays based on human cells, toxicity pathways, and high-throughput techniques, and even *in vivo* models (Adler et al. 2011). Each tool in an ITS may provide support for, or form the basis of a challenge to, data from a non-standard investigation (Combes 2012). An example is given in **Figure 22**. ITS approaches are useful when not all possible outcomes of interest (e.g., modes of action and molecular and cellular interactions), classes of test substances (applicability domains), or severity classes of effect are covered in a single test. ITS are also valuable when the human predictivity of a single test is not satisfactory, for instance when a test produces is not sufficiently validated or yields many false positive results (Hartung et al. 2013). Predictions can then be substantiated by other available information. A tiered ITS approach provides the opportunity to combine existing data with new data. In addition, certain substances can be filtered out or prioritized before costly tests or animal tests are performed (Basketter et al. 2012). Sophisticated combinations of tools with interim decision points are emerging, but accepted concepts regarding how to construct and validate them are not yet established (Basketter et al. 2012).

The vision is that the new mode of action models, emerging technologies based on them, and their combination in ITS, allow the formulation of a 'systems toxicology' approach. These integrated and information-rich assessments require a shift to a more probabilistic evaluation, where each test changes to some extent the probability of a hazard and/or its uncertainty (t4 2013). By using a Weight of Evidence (WoE) approach, different pieces of evidence and test data be weighed and combined (Hartung et al. 2013).

> ***Weight of Evidence*** *The type of consideration made in a situation where there is uncertainty and which is used to ascertain whether the evidence or information supporting one side of a cause or argument is greater than that supporting the other side*

A value is assigned to each piece of information using either an objective method (formalised procedure) or by expert judgement. The weight is influenced by factors such as the data quality, consistency of results, nature and severity of effects, relevance, etc. In the example in **Figure 22**, tests are included that have not been formally validated according to internationally accepted criteria, even though most of these have standardised and optimised protocols. Since such non-validated test methods are able to produce data that can be used in WoE evaluations, particularly for classification and labelling purposes, they can still be applied in ITS approaches, even though the methods still know important limitations. Such methods are particularly useful for prioritising chemicals for further safety assessment (Balls et al. 2012). In order to make use of novel high-content, high-throughput, and pathway-based information, ways of distilling relevant information out of the large datasets that are produced need to be developed (Basketter et al. 2012).

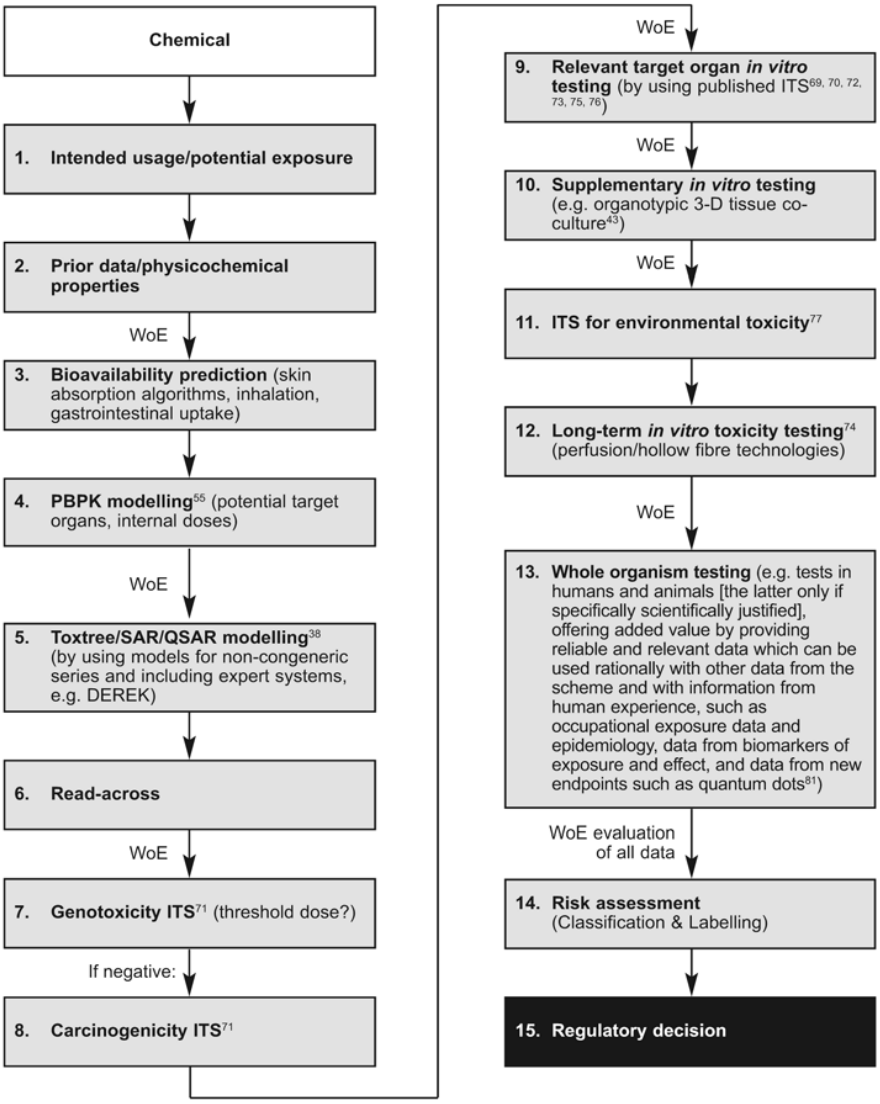

Figure 22. A general Integrated Testing Strategy (ITS) for chemicals. This ITS scheme could be used, for example, to provide the required comprehensive set of information for registration of new chemicals and so-called 'missing' information for existing chemicals (Balls et al. 2012).

Accepted WoE tools (that include for instance criteria for the acceptance and rejection of pre-existing data, ways to convert the individual method-specific levels of confidence to joint confidence for a consensus outcome, and guidelines for a statistically sound way to deal with conflicting results) allowing for objective judgments, are under development (Hartung et al. 2013). Within the EU-funded OSIRIS project (http://www.ufz.de/osiris), integrated testing strategies were developed for four human health endpoints fit for the EU chemicals legislation REACH in order to significantly increase the use of non-testing information for regulatory decision making. A formal, transparent, and scientifically sound decision webtool for ITS for four human toxicological endpoints (skin sensitisation, repeated dose toxicity, mutagenicity and carcinogenicity) and bioconcentration factor and aquatic toxicity was built. The tool allows for objective and reproducible decisions on whether or not additional data is needed by weighing the different and possibly contradictory information and by taking the respective uncertainties into account. The weighting process is incorporated in an interactive workflow that includes a comparison with a conditional WoE threshold (**Figure 23**). Experts

finally need to decide on further steps, e.g. (no) further testing, use of (Q)SARs or read-across or EBW, when this threshold is or is just not met (Vermeire et al,. 2013).

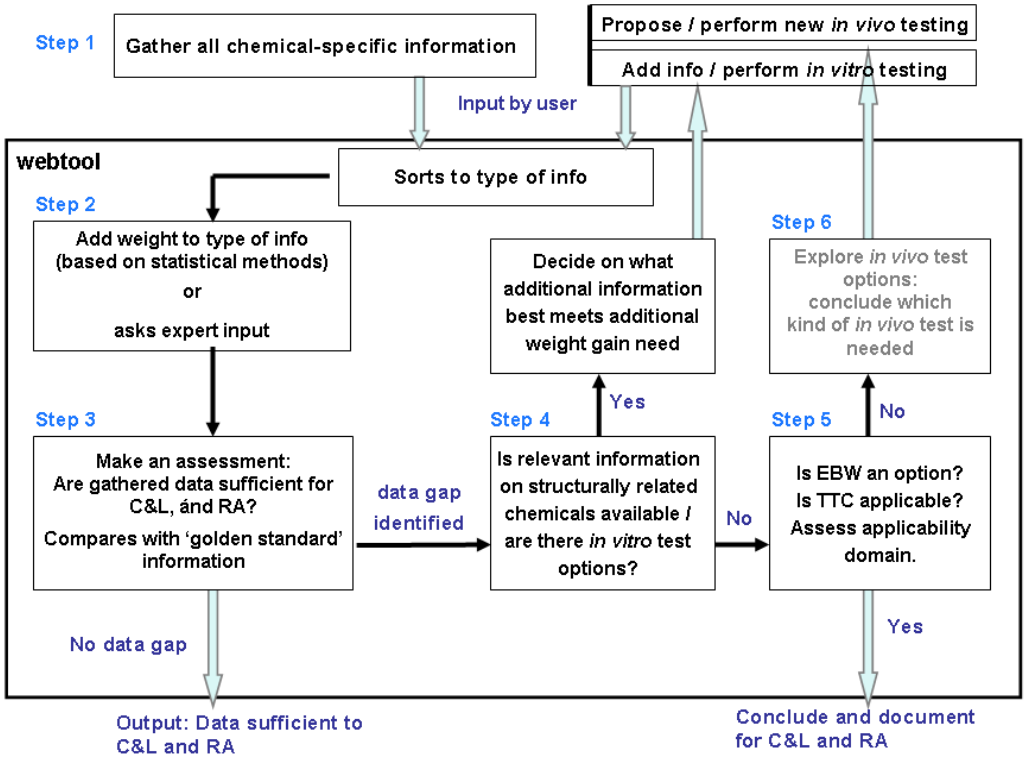

Figure 22. The user-interactive workflow of the Osiris webtool that weighs different types of data (http://osiris.simpple.com/OSIRIS-ITS/welcome.do).

5 Recommendations for assessment of water quality

The new tools described in this report are sophisticated and scientifically advanced, offering the prospect of a mechanistic understanding of the biological effect that is studied. The application of innovative experimental test methods has however been limited to date. Main reasons for this are not a lack of available methods, but a lack of consensus on their applicability domains and of international validation and regulatory acceptance. Scientifically justified and supported testing approaches nevertheless offer opportunities for human hazard assessment (Balls et al. 2012). As described in the previous chapter, different pieces of evidence and several methods can be strategically combined in integrated testing strategies, by breaking potential hazard down into its modes of action and combining those with physicochemical properties (including QSARs) and PBPK models (Hartung et al. 2013). Although it is recommended to use a battery of tests in ITS strategies, in which the different pieces of information substantiate each other, such strategies are aimed at thorough evaluation of health hazard of chemicals by fully covering the range of potential health outcomes, whereas testing of drinking water usually involves screening for potential hazard of low levels of (mixtures of) contaminants. For this purpose, application of a small selection of tools that assess a series of relevant health outcomes may be sufficient. It should be kept in mind that predictions made by these tools always imply a certain extent of uncertainty, and that a balance needs to be found between proper safety testing and feasibility with regard to time and costs invested.

Recommendation I
Combine a selection of complementary tools that assess a series of health outcomes relevant for exposure to drinking water contaminants. Development of guidelines for ITS and WOE should be closely watched.

Both non-testing strategies and *in vitro* approaches can be used to obtain an indication of potential human hazard when compounds emerge in drinking water or its resources for which no or little toxicological information is available. Non-testing approaches are relatively inexpensive and quick as compared to toxicity testing and do not require expensive laboratory equipment and facilities. The *in silico* tools can readily be used for high-throughput screening and for evaluating large numbers of chemicals (Combes 2010). This is useful in emergency situations, when hazard information is urgently needed. Non-testing information can indicate a possible threat and provide the basis for decision making aimed at reducing exposures. Non-testing approaches can also prioritize compounds for further toxicity testing. In addition, non-testing approaches can identify groups of similar chemicals, identify components in a mixture that present the greatest level of concern, and provide mechanistic information (ECETOC 2012, IGHRC 2013).

Recommendation II
Use non-testing approaches for quick hazard assessment, prioritization of compounds for further testing and to provide mechanistic information.

In order to perform quick hazard assessment, there should be immediate access to non-testing tools and familiarity with and confidence in the use of the tools (IGHRC 2013). Preferably, models that are integrated in software tools should be used. The choice of tools depends on the endpoint of interest and the applicability domain of the tool. Some level of expertise is required to select the proper tools, and also to perform and evaluate the predictions. Non-testing approaches involve, besides chemical and toxicological knowledge, the use of complex terminology as well as mathematical and statistical procedures that are not readily understood by non-experts or that have not been trained to use the tools (Combes 2012). In addition, there is general consensus that multiple non-testing approaches should be combined to produce a reliable prediction. One should however also be aware of the limitations of predictive approaches, such as the bias towards pharmaceuticals and the limited validation of prediction of human health endpoints (IGHRC 2013).

Recommendation III

Expertise is needed to select non-testing tools and perform and evaluate predictions. KWR should accommodate multiple non-testing tools and trained users that can be consulted by drinking water companies when needed. KWR currently holds a user licence for HazardExpert; the OECD QSAR Toolbox is a recommended and freely available addition to this.

In vitro assays may subsequently be used in a tiered approach for confirmation of biological activity that is predicted by non-testing approaches. Secondly, *in vitro* tools evaluating relevant health effects can be applied to demonstrate potential biological effects of (mixtures of) unidentified substances present in water samples. This may aid in identifying exposure to relevant concentrations of chemicals and prioritizing water sources that need attention with regard to human health protection. In addition, this may indicate the type of substances present in the water sample analysed by the type of biological response that is elicited or disturbed. When no effects are measured in *in vitro* tests, toxicological relevance may be considered to be low and further chemical analysis to identify contaminants becomes less urgent.

Recommendation IV

Use *in vitro* assays to confirm non-testing information and to demonstrate potential biological effects of unidentified substances; in both cases water sources that need further examination can be prioritized.

There are many *in vitro* assays already available or under development to choose from. All *in vitro* systems have limitations, and the choice of which one to use will depend on the question asked. Although only a few of the *in vitro* assays have been formally validated, validation of assays is as yet not required for application of *in vitro* testing since there are no regulatory requirements, and should therefore not necessarily limit the application of assays. Since only a number of *in vitro* assays are currently in use for evaluation of water quality, for some assays, sensitivity, robustness, specificity, and selectivity may only become evident during future use. In cases where international test guidelines are present, adherence to those guidelines is not a prerequisite for proper toxicity testing, but it facilitates comparability and acceptance of test data. When the identities of substances present in water samples are unknown, the most suitable tool cannot be selected on the basis of the most relevant mechanism of action. Assays that detect a large range of substances at low

dose levels would then be appropriate, and assays that assess health outcomes that have been shown to be relevant for drinking water contaminants. In a current joint research project[2], criteria for selection of *in vitro* assays for screening of drinking water quality are established, and these will subsequently be applied to select and optimise assays for detection of relevant health outcomes.

Correlation of biological effects detected using *in vitro* assays with *in vivo* health outcomes will remain a major challenge for risk assessors. Human relevance can be achieved by using cells from human origin in which metabolism and molecular pathways of importance are intact. The further development of 'omics tools, AOPs, and kinetic modelling are envisaged to improve the predictive value of *in vitro* tests with regard to human health hazard and will thereby facilitate risk assessment. A shift towards pathway-based approaches is taking place. At the same time, this will lead to an increasing complexity and variety of methods based on mechanisms of action at the molecular, cellular, tissue, organ, or system level. Quantitative use of *in vitro* data for risk assessment purposes by using for instance BMD modelling needs to be considered. Ongoing developments in these areas should be closely watched.

Recommendation V

The potential to evaluate a large range of substances at low dose levels and the predictive value for human health hazard should be taken into account when selecting *in vitro* assays.

In the near future, the evaluation of health hazard of compounds present in drinking water sources or added during treatment and distribution, is expected to benefit from the use of a suite of advanced *in vitro* and *in silico* tools. The combination of rapid assessment by carefully selected up-to-date tools and expert judgement will provide a more reliable and relevant estimation of the potential hazard of currently present and emerging compounds in the (aquatic) environment.

[2] '(In vitro) bioassays als screeningtools voor waterkwaliteit (fase I)'

References

Adler, S., Basketter, D., et al. (2011) Alternative (non-animal) methods for cosmetics testing: current status and future prospects. Arch Toxicol 85(5): 367-485.

Ankley, G. Bennett, T. R. S., et al. (2010) Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment. Environ Toxicol Chem 29(3): 730-741.

Bajpayee, M., Kumar, A., et al. (2013) The Comet Assay: Assessment of *In vitro* and *In vivo* DNA Damage. Genotoxicity Assessment 1044: 325-345.

Baken, K. A. (2013) Screening of water samples for developmental toxicity. An overview of relevant compounds and available assays. BTO 2013.226(s) Nieuwegein, The Netherlands, KWR Watercycle Research Institute.

Balls, M., Combes R., et al. (2012) The Use of Integrated and Intelligent Testing Strategies in the Prediction of Toxic Hazard and in Risk Assessment. Springer US. New Technologies for Toxicity Testing. Advances in Experimental Medicine and Biology 745, 221-253.

Basketter, D. Clewell, A. H., et al. (2012) A roadmap for the development of alternative (non-animal) methods for systemic toxicity testing - t4 report. ALTEX 29(1): 3-91.

Birrell, L., Cahill, P., et al. (2010) GADD45a-GFP GreenScreen HC assay results for the ECVAM recommended lists of genotoxic and non-genotoxic chemicals for assessment of new genotoxicity tests. Mutat Res 695(1-2): 87-95.

Bradbury, S, T. Feijtel and K. van Leeuwen. Meeting the scientific needs of ecological risk assessment in a regulatory context. Environm. Sci. Technol. 38 (2004) 463a-470a.

Burlinson, B. (2012) The *in vitro* and *in vivo* comet assays. Methods Mol Biol 817: 143-163.

Coecke, S., Pelkonen, O., et al. (2013) Toxicokinetics as a key to the integrated toxicity risk assessment based primarily on non-animal approaches. Toxicology *in vitro* 27(5): 1570-1577.

Combes, R.D. (2012) *In silico* methods for toxicity prediction. Springer US. New Technologies for Toxicity Testing. Advances in Experimental Medicine and Biology Volume 745, 96-116.

ECETOC (2012) Category approaches, Read-across, (Q)SAR. Technical Report No. 116. European Centre for Ecotoxicology and Toxicology of Chemicals.

ECETOC (2013) Omics and Risk Assessment Science. Workshop report no.25. European Centre for Ecotoxicology and Toxicology of Chemicals.

EPA (1991) Guidelines for Developmental Toxicity Risk Assessment (EPA/600/FR-91/001). Environmental Protection Agency Risk Assessment Forum. Washington, DC.

FDA (2011) Guidance for Industry Reproductive and Developmental Toxicities - Integrating Study Results to Assess Concerns (76 FR 59142). U.S. Food and Drug Administration, Department of Health and Human Services.

Harris, C. (2012) Overview of *In vitro* Models in Developmental Toxicology. Developmental Toxicology 889: 105-113.

Harris, C. (2012) Rodent Whole Embryo Culture. Developmental Toxicology 889: 215-237.

Hartung, T., Luechtefeld, T., et al. (2013) Integrated testing strategies for safety assessments. ALTEX 30(1)

Hastwell, P. W., Chai, L. L., et al. (2006) High-specificity and high-sensitivity genotoxicity assessment in a human cell line: validation of the GreenScreen HC GADD45a-GFP genotoxicity assay. Mutat Res 607(2): 160-175.

Hendriks, G., Atallah, M., et al. (2012) The ToxTracker assay: novel GFP reporter systems that provide mechanistic insight into the genotoxic properties of chemicals. Toxicol Sci 125(1): 285-298.

Hughes, C., Rabinowitz A., et al. (2012) Development of a high-throughput Gaussia luciferase reporter assay for the activation of the GADD45a gene by mutagens, promutagens, clastogens, and aneugens. J Biomol Screen 17(10): 1302-1315.

IGHRC (2013) Predictive approaches to chemical hazard identification and characterisation: Current use by UK Government Departments and Agencies (cr16). Institute of Environment and Health, Cranfield University, UK.

JRC (2010) Applicability of QSAR analysis to the evaluation of the toxicological relevance of metabolites and degradates of pesticide active substances for dietary risk assessment. Scientific report submitted to EFSA. Computational Toxicology Group, Institute for Health & Consumer Protection, European Commission - Joint Research Centre, Ispra, Italy.

Luilo GB1, Cabaniss SE. (2011) QSPR for predicting chloroform formation in drinking water disinfection. SAR QSAR Environ Res. 22(5-6):489-504.

Milan, C., Roncaglioni, A., Diaza, R.G., Cassano, A., Golbamaki, N., Benfenati, E. (2010) ANTARES LIFE08 ENV/IT/000435 Deliverable 2: Report on the identified criteria for non-testing methods, and their scores.

Mizota, T., Ohno, K., et al. (2011) Validation of a genotoxicity test based on p53R2 gene expression in human lymphoblastoid cells. Mutat Res 724(1-2): 76-85.

Nakazawa, K., Shinmura, Y., et al. (2011) Effects of Culture Conditions on a Micropatterned Co-culture of Rat Hepatocytes with 3T3 cells. J Bioprocess Biotechniq S3(002)

OECD (2006) Guidelines for the Testing of Chemicals, Section 4. Organisation for Economic Co-operation and Development.

OECD (2007) Guidance on grouping of chemicals. Series on testing and assessment No. 80 (ENV/JM/MONO(2007)28). Organisation for Economic Co-operation and Development.

OECD (2013) Guidance document on developing and assessing adverse outcome pathways. Series on Testing and Assessment No. 184 (ENV/JM/MONO(2013)6). Organisation for Economic Co-operation and Development.

Patlewicz, G., Ball, N., Booth, E.D., Hulzebos, E., Zvinavashe, E. and Hennes, C. (2013) Use of category approaches, read-across and (Q)SAR: general considerations. Regulatory toxicology and pharmacology 67, 1-12.

Reifferscheid, G., Arndt, C., et al. (2005) Further development of the beta-lactamase MutaGen assay and evaluation by comparison with Ames fluctuation tests and the umu test. Environ Mol Mutagen 46(2): 126-139.

Reifferscheid, G. and Heil, J. (1996) Validation of the SOS/umu test using test results of 486 chemicals and comparison with the Ames test and carcinogenicity data. Mutat Res 369(3-4): 129-145.

Reifferscheid, G., Maes, H. M., et al. (2012) International round-robin study on the Ames fluctuation test. Environ Mol Mutagen 53(3): 185-197.

Schaafsma, G, Kroese, D, Tielemans, E, van de Sandt, H and K. Van Leeuwen. (2009) REACH, non-testing approaches and the urgent need for a change in mind set. Reg Toxicol Pharmacol 53: 70-80.

Schulpen, S. W. and Piersma, A. (2013) The Embryonic Stem Cell Test. Teratogenicity Testing 947: 375-382.

Schulten H.-R., Leinweber P. (1996) Characterization of humic and soil particles by analytical pyrolysis and computer modelling. Journal of Analytical and Applied Pyrolysis, 38(1-2):1-53.

t4 (2012) The use of biomarkers of toxicity for integrating *in vitro* hazard estimates into risk assessment for humans. Workshop report transatlantic think tank for toxicology. ALTEX 29(4)

t4 (2013) Pathways of Toxicity. Workshop report transatlantic think tank for toxicology. ALTEX online first S1868696X1309261X.

Theunissen, P. T. and Piersma, A. H. (2012) Innovative approaches in the embryonic stem cell test (EST) Front Biosci 17: 1965-1975.

Uibel, F., Muhleisen, A., et al. (2010) ReProGlo: a new stem cell-based reporter assay aimed to predict embryotoxic potential of drugs and chemicals. Reprod Toxicol 30(1): 103-112.

van der Lelie, D., Regniers, L., et al. (1997) The VITOTOX test, an SOS bioluminescence Salmonella typhimurium test to measure genotoxicity kinetics. Mutat Res 389(2-3): 279-290.

van Leeuwen, C.J., G.Y. Patlewicz, and A.P. Worth (2007) Intelligent Testing Strategies In: Risk Assessment of Chemicals. An Introduction (2nd edition). van Leeuwen, C.J. and T.G. Vermeire, eds. Springer Publishers, Dordrecht, The Netherlands, pp 467-509.

Van Leeuwen, C.J., T. Schultz, T. Henry, G. Veith and B. Diderich (2009). Using chemical categories and QSAR to fill data gaps in hazard assessment. SAR QSAR in Environ Toxicol 20: 207-220.

Vermeire, T., Aldenberg, T., Buist, H., Escher, S., Mangelsdorf, I., Pauné, E., Rorije, E., Kroese, D. OSIRIS, a quest for proof of principle for integrated testing strategies of chemicals for four human health endpoints. Regul Toxicol Pharmacol. 2013 Nov;67(2):136-45

Vinken, M. (2013) The adverse outcome pathway concept: a pragmatic tool in toxicology. Toxicology 312: 158-165.

Westerink, W. M., Schirris, T. J. et al. (2011) Development and validation of a high-content screening *in vitro* micronucleus assay in CHO-k1 and HepG2 cells. Mutat Res 724(1-2): 7-21.

Westerink, W. M., Stevenson, J. C., et al. (2010) The development of RAD51C, Cystatin A, p53 and Nrf2 luciferase-reporter assays in metabolically competent HepG2 cells for the assessment of mechanism-based genotoxicity and of oxidative stress in the early research phase of drug development. Mutat Res 696(1): 21-40.

Westerink, W. M. A., Stevenson, J. C. R., et al. (2009) Evaluation of the Vitotox™ and RadarScreen assays for the rapid assessment of genotoxicity in the early research phase of drug development. Mutation Research/Genetic Toxicology and Environmental Mutagenesis 676(1–2): 113-130.

WHO/EEA (2002) Children's health and environment: A review of evidence. Environmental Issue Report No 29. Copenhagen, European Environment Agency and the WHO Regional Office for Europe.