

## 4. Filling Data Gaps: Introduction to Predictive Models

4 Filling Data Gaps: Introduction to Predictive Models .....	4-1
4.1 SARs and QSARs .....	4-1
4.1.1 Developing Predictive Models .....	4-1
4.1.2 Mathematical Algorithms .....	4-1
4.2 Fragment-based Approaches .....	4-3
4.3 Expert System Models .....	4-4
4.4 Combination Models .....	4-5
4.5 Potential Structural Entry Formats for Predictive Models .....	4-6

## 4 Filling Data Gaps: Introduction to Predictive Models

The P2 Framework models and methods that have been incorporated into the Sustainable Futures Initiative contain three general types of predictive models:

1. Mathematical Models such as SARs and QSARs which use descriptors and mathematical relationships to derive predictions. Examples of SAR / QSAR models are ECOSAR and select modules contained in EPISuite™ like WSKOWWIN.
2. Fragment-based models such as the BOWIN module within EPISuite™ evaluate the features of molecular fragments present on the molecule to make predictions.
3. Expert Systems, like OncoLogic™, use rule-based decision trees to mimic an expert's judgment. Other Expert Systems utilize Artificial Neural Networks and Molecular Models.

### 4.1 SARs and QSARs

#### 4.1.1 Developing Predictive Models

##### Structure Activity Relationships (SARs)

A Structure Activity Relationship (SAR) is the relationship between the molecular structure of a chemical and its activity. The goal of developing SARs is being able to use the SAR to predict the activity of a chemical lacking data by comparing its structure to a chemical with experimental data that is associated with specific activity. The SARs contained in the Sustainable Futures / P2 Framework methods most often give us “flags” to identify potential activity and to identify chemicals that require additional testing. These are screening level estimations that give qualitative predictions. An example of a SAR method that gives qualitative predictions is the biodegradation estimation method BOWIN™ which estimates aerobic and anaerobic biodegradability of organic chemicals using seven different models. BOWIN™ is incorporated into the EPI Suite™ of methods and provides qualitative estimation such as “likely” or “not likely” to biodegrade rapidly.

##### Quantitative Structure Activity Relationships (QSARs)

Quantitative Structure Activity Relationship (QSAR) is a Structure Activity Relationship for which a numeric value can be determined. The QSAR describes the relationship between descriptors of chemical structure (e.g., molecular fragments, physical-chemical properties, etc.) and biological activity usually based on mathematical algorithms such as linear regression.

#### 4.1.2 Mathematical Algorithms

##### Designing a Mathematical QSAR (EPISuite™ and ECOSAR Method Basis)

1. Begin with a set of chemicals with experimental data for your endpoint of interest, referred to as “training set” for the model.

## Sustainable Futures / P2 Framework Manual 2012 EPA-748-B12-001

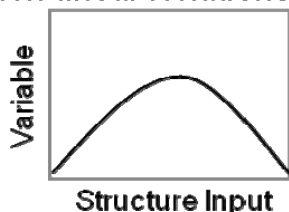
### 4. Filling Data Gaps: Introduction to Predictive Models

- Establish Variables for each property of interest and create a “data matrix” (shown below) using the experimental data.

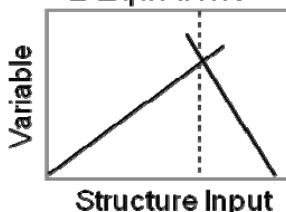
Chemical	Fish LC50	Variable 1 KOW	Variable 2 Water Sol	Variable 3 MW
82-54-7	0.32 mg/L	3.8	12	110
95-82-4	2.65 mg/L	2.3	108	94
654-86-7	12.3 mg/L	0.32	125	45

- Run statistical regression on the data matrix to identify the MOST highly correlated variables. In this example Variable 1 and Variable 2 have higher correlations than Variable 3.
- Graph the data sets to see if there are mathematical relationships. Generally the relationships have characteristic appearances, shown below.

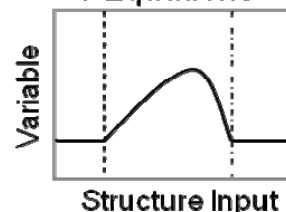
**Non-linear Relationship**



**2 Equations**



**3 Equations**



Here is a sample of a training set, taken from the ECOSAR on-line Help QSAR Class Reference Document for Acrylates.

CAS No.	Chemical Name	MW.	log Kow (CLogP)	log Kow (EPI)	log Kow (M)	Fish 96-h LC50 (mg/L)	Log Fish 96-h LC50 (mmol/L)	Reference (Meas. Kow)	Reference (Fish 96-h LC50)
CBI	CBI	626	1.8	-1.2		100	-0.80		P 96-_____
818-61-1	2-Hydroxyethyl acrylate	116	-0.01	-0.25	-0.21	4.8	-1.38	Hansch & Leo, 1985	DUL
999-61-1	2-Hydroxypropyl acrylate	130	0.3	0.17	0.35	3.26	-1.60	Hansch & Leo, 1985	U.S. EPA, 1991
999-61-1	2-Hydroxypropyl acrylate	130	0.3	0.17	0.35	3.61	-1.56	Hansch & Leo, 1985	DUL
999-61-1	2-Hydroxypropyl acrylate	130	0.3	0.17	0.35	3.1	-1.62	Hansch & Leo, 1985	DUL
CBI	CBI	186	1.1	0.53	1.71	6.6	-1.45		P 04-_____
CBI	CBI	144	0.83	0.66		4.2	-1.54		P 87-_____
96-33-3	Methyl acrylate	86	0.75	0.73	0.8,0.739	3.4	-1.40	Hansch C. et al., 1995	Datasheet
140-88-5	Ethyl acrylate	100	1.3	1.2	1.32	2.5	-1.60	Hansch & Leo, 1985	DUL
106-63-8	Isobutyl acrylate	128	2.3	2.1	2.22	2.09	-1.79	Hansch & Leo, 1985	DUL
106-63-8	Isobutyl acrylate	128	2.3	2.1	2.22	2.11	-1.78	Hansch & Leo, 1985	DUL
CBI	CBI	258	3.5	2.8		2.3	-2.05		P 99-_____
3066-71-5	Cyclohexyl acrylate	154	2.8	3		1.48	-2.02		DUL
2499-95-8	Hexyl acrylate	156	3.2	3.4		1.09	-2.16		DUL
2499-95-8	Hexyl acrylate	156	3.2	3.4		1.14	-2.14		DUL
CBI	CBI	805	MF	4.3		4.5	-2.25		P 00-_____

**Sustainable Futures / P2 Framework Manual 2012 EPA-748-B12-001**  
**4. Filling Data Gaps: Introduction to Predictive Models**

**Example QSAR Algorithm – Developing a Linear Equation in ECOSAR**

This example shows how SAR equations are developed in ECOSAR to predict Acute Fish Toxicity (LC50).

The equation is  $Y = mc+b$  (linear relationship)

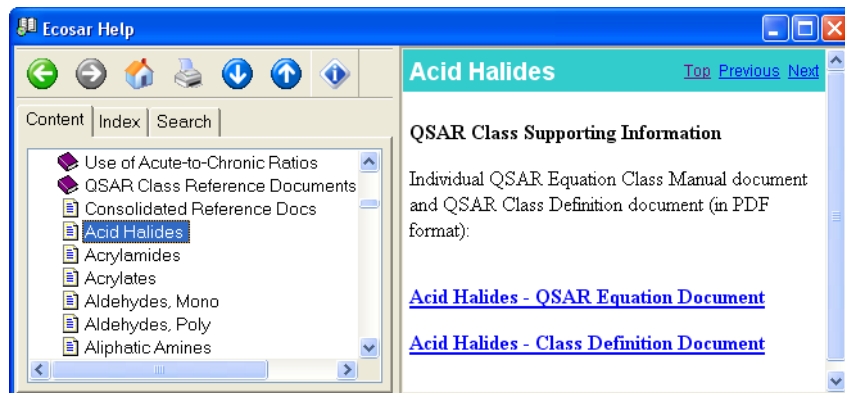
$m$  = slope of line ( $y/x$ )

$c$  = correlated variable for prediction (Log Kow)

$b$  = equation constant ( $y$ -intercept)



ECOSAR Equation:  $\text{Log } 96\text{-h LC}_{50} \text{ (fish, millimoles/L)} = -1.46 - 0.18 \text{ log Kow}$



ECOSAR's On-line Help has QSAR equations for each SAR contained within the model (shown left).

## 4.2 Fragment-based Approaches

Designing a Fragment-based QSAR (EPISuite™ Method Basis)

1. Begin with a set of chemicals with experimental data for your endpoint of interest. This is referred to as the "training set" for the model.
2. Create a matrix showing molecular fragments and fragment counts for each training set chemical (example shown below).

Chemical	Endpoint	Molecular Fragment 1	Molecular Fragment 2	Molecular Fragment 3
82-54-7	Degrades Fast	0	1	0
95-82-4	Degrades Slowly	2	0	0
654-86-7	Degrades Fast	1	0	1
Fragment Contribution →		$f_i = -1.1$	$f_i = +4.4$	$f_i = +2.2$

3. Run statistical regression on matrix to identify coefficient for each variable which will correlate to the "weight" of the fragment to the endpoint of interest.

**Sustainable Futures / P2 Framework Manual 2012 EPA-748-B12-001**  
**4. Filling Data Gaps: Introduction to Predictive Models**

Some fragments will have a positive contribution and some will have a negative contribution.

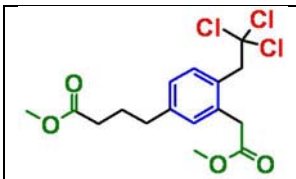
Using molecular descriptors

$$\text{Prediction} = \sum a_i f_i + c$$

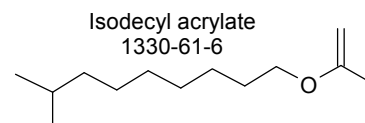
$f_i$  = molecular descriptor coefficient and

$a_i$  = number of time descriptor appears

$$\text{Prediction} = [(1 * 2.2) + (2 * 4.4) + (3 * -1.1)] + 0.52$$

	Fragment	Number Present	Value ( $f_i$ )
	Aromatic	1	2.2
	Ester	2	4.4
	Chlorine	3	-1.1

This figure below, KOWWIN results for Isodecyl acrylate, CAS 1330-61-6, shows how the model combines the fragments to estimate a KOW value. KOWWIN is one of the estimation modules within EPISuite™.



Kowwin Results					
Print Save Results Copy Remove Window Help					
<b>Log Kow(version 1.68 estimate): 5.07</b>					
SMILES : O=C(C=C)OCCCCCCCC(C)C					
CHEM : 2-Propenoic acid, isodecyl ester					
MOL FOR: C13 H24 O2					
MOL WT : 212.34					
TYPE	NUM	LOGKOW FRAGMENT DESCRIPTION		COEFF	VALUE
Frag	2	-CH3	[aliphatic carbon]	0.5473	1.0946
Frag	7	-CH2-	[aliphatic carbon]	0.4911	3.4377
Frag	1	-CH	[aliphatic carbon]	0.3614	0.3614
Frag	1	=CH2	[olefinic carbon]	0.5184	0.5184
Frag	1	=CH- or =C<	[olefinic carbon]	0.3836	0.3836
Frag	1	-C(=O)O	[ester, aliphatic attach]	-0.9505	-0.9505
Const		Equation Constant			0.2290
				Log Kow =	5.0742

### 4.3 Expert System Models

The Cancer Expert System, OncoLogic™, which is one of the Sustainable Futures / P2 Framework methods, is described in detail in chapter 10 of this document. This brief overview is presented here to describe how an Expert System model works.

An Expert System is a computerized system that mimics the thinking and reasoning of human experts. They are rules-based and contain a formalized, codified, and organized system of structure-activity relationships that also incorporates Mechanisms of Action (MOAs). An Expert System typically provides qualitative results such as “Low, Moderate, or High” or “Certain, Probable, or Improbable”.

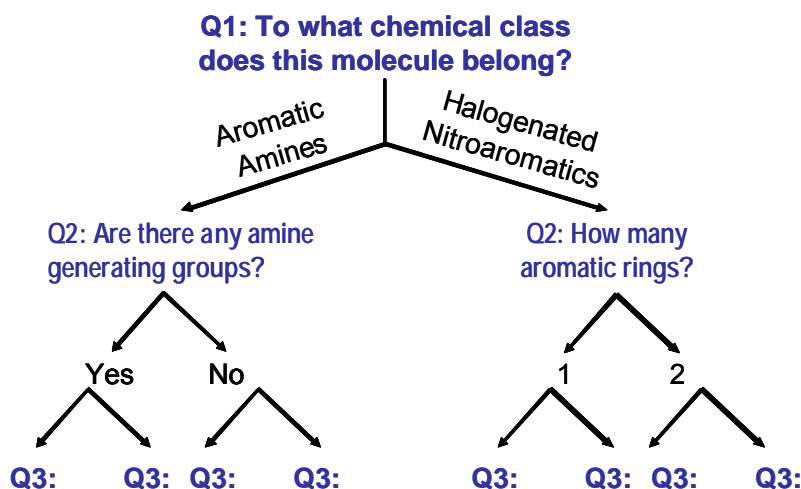
Here are the definitions of the OncoLogic™ Concern Levels:

## 4. Filling Data Gaps: Introduction to Predictive Models

Low	Unlikely to be carcinogenic
Marginal	Likely to have equivocal carcinogenic activity
Low – Moderate	Likely to be weakly carcinogenic
Moderate	Likely to be a moderately active carcinogen
Moderate – High	Highly likely to be a moderately active carcinogen
High	Highly likely to be a potent carcinogen

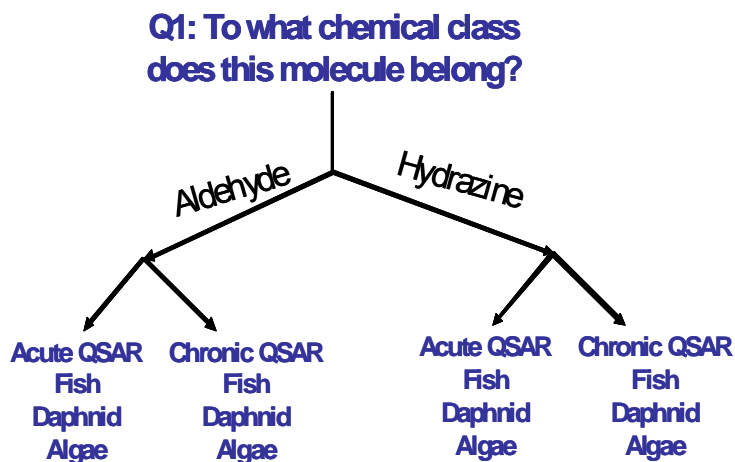
This figure shows a simplified “decision tree”. The decision trees contained within Expert Systems like OncoLogic™ are very complex.

## Example of a Decision Tree



## 4.4 Combination Models

Some predictive models use a combination of approaches. An example is ECOSAR which has both “rules” for selecting chemical classes (shown in this figure below) and class-specific mathematical QSARs. .



## 4. Filling Data Gaps: Introduction to Predictive Models

### Scientific Validity of (Q)SAR Predictions

According to OECD Principles\*, a valid (Q)SAR should be associated with the following information:

1. defined endpoint (universal vs. local),
2. unambiguous algorithm (transparent method),
3. defined applicability domain (with descriptors),
4. appropriate measures of goodness-of-fit, robustness and predictivity (internal and external validation), and
5. mechanistic interpretation, if possible because it enhances scientific support.

\*Guidance Document On The Validation Of (Quantitative) Structure-Activity Relationships [(Q)SAR] Models. (2007). OECD Environment Health and Safety Publications, Series on Testing and Assessment, No. 69. ENV/JM/MONO(2007)2.

([http://www.oecd.org/officialdocuments/displaydocumentpdf/?cote=env/jm/mono\(2007\)2&doclanguage=en](http://www.oecd.org/officialdocuments/displaydocumentpdf/?cote=env/jm/mono(2007)2&doclanguage=en))

### 4.5 Potential Structural Entry Formats for Predictive Models

There are various methods used to translate a three-dimension chemical structure into a format that a computer program will understand. Some of those methods are described here.

**SMILES Notation** is described in Appendix F of this document. EPISuite™ and ECOSAR allow for entering structures using their incorporated SMILECAS databases which contain pre-drawn SMILES Notations for many discrete chemicals.

**MDL mol files** Some programs like ECOSAR allow users to import .mol files which can be developed by most of the chemical drawing software programs.

**Structure Data Format**, described at (<http://www.epa.gov/ncct/dsstox/MoreonSDF.html>), also known as SD Files, are ASCII text files that contain multiple chemicals in a single file. When doing batch runs of numerous chemicals in EPISuite™ and ECOSAR, SD files can be used to import structures. Most Chemical Relational Database (CRD) applications used for structure-searching of chemical information are capable of importing and exporting SD files.