

A QSAR - Bayesian neural network model to identify molecular properties causing eye irritation in cationic surfactants

Grace Y Patlewicz¹ and Wael El-Deredy²

¹SEAC Toxicology Unit, Colworth, Unilever Research, Colworth House, Sharnbrook, Bedford, MK44 1LQ, UK

Grace.Patlewicz@unilever.com

²Adaptive Computation, Unilever Research Port Sunlight, Bebington,

L63 3JW, UK

Wael.El-Deredy@unilever.com

Abstract

QSAR models are frequently used to investigate and predict the toxicological effects of chemicals. Building QSAR models of the eye irritation potential of cationic surfactants is difficult, as the mechanism of action of these surfactants is still not fully understood.

This report describes a data driven QSAR model to predict Maximum Average Scores (MAS in accordance to Draize) for cationic surfactants from the calculated molecular properties Log P, Log CMC and molecular volume, and the surfactant concentration. We demonstrate that a Bayesian Neural Network, a statistical non-linear regression approach that estimates the noise in the modelling data and error bars on the predictions, provided the most robust and accurate representation of the relationship between the MAS score and the molecular properties. A dataset of 20 *in vivo* rabbit eye irritation tests on 19 different cationic surfactants, obtained from historic in-house data and the scientific literature, was used to train the Bayesian neural network. The model was then used to simulate a large number of molecules to explore the relationship between MAS score and molecular properties. MAS vs. Log P showed bell shaped curve as expected. A higher concentration (> 20%) was required in order to elicit the eye irritancy response of molecules with a wide range of Log P. The simulated results were used to identify the range of molecular properties of cationic surfactants most likely to cause more than mild irritancy. Within the parameter space of the model, defined by the training data, the probability of causing severe irritation is highest for molecules with molecular volume < 320 Å³, while -2 < Log P < 13 and -6 < Log CMC < 3. The simulated results were carried out at a concentration of 40%. For molecules with larger molecular volumes, the range of Log P and Log CMC for which these molecules would cause severe irritation is narrowed.

The model provides useful probabilistic predictions for the eye irritancy potential of new or untested cationic surfactants with physicochemical properties lying within the parameter space of the model.

Introduction

Eye irritancy potential *in vivo* is still based on the method described by Draize *et al* (1944). Chemical use solely on the basis of *in vitro* tests remains generally unacceptable (Speilmann *et al*, 1996) and there are neither complete replacements for the Draize test nor strategies which completely avoid the use of animals. Computer modelling based on a small number of carefully selected experiments can prove helpful in extending the knowledge domain whilst limiting the number of animals required for experimentation.

QSARs are often hindered by a lack of quality *in vivo* data and sufficient understanding of the mechanisms of action (Cronin *et al*, 1995). Understanding of the eye irritation potential of chemicals is complex. Here we demonstrate that the Bayesian neural network (BNN) produces a robust model due to its capability of predicting noise in the experimental data and providing error bars on the predictions. We use this model to simulate a large number of molecular properties for a range of cationic surfactants and predict their corresponding MAS scores. We then identify and isolate the region of the molecular space that is most likely to result in severe eye irritation.

Data and Methods

We used the dataset (Table 1) and mechanistic hypothesis as Rodford *et al* (1998). MAS scores were on a scale of 0 - 110 in accordance with the Draize scoring systems for *in vivo* eye tests. (Draize *et al*, 1944) Log P and molecular volume were chosen to model membrane permeability and Log CMC (critical micelle concentration) was our "reactivity" parameter.

The cationic surfactants were treated as pure chemicals despite their variation in alkyl chain length.

Choice of modelling technique

Both principle component analysis and linear regression methods were carried out but failed to produce a reasonable correlation coefficient.

Neural network analysis

A Bayesian neural network (BNN) is non-linear regression approach for training neural networks in which Bayes' theorem is used to estimate the model parameters. In a Bayesian model, all model parameters are explicitly expressed as probability distributions (MacKay, 1994). The posterior distributions of the model parameters are estimated from their priors given the training dataset. Training the network amounts to estimating the posterior distributions of the network parameters. These are then used to estimate the distribution of the model output $p(t)$ when given new data \mathbf{x} . The output distribution has the form (b)

$$p(t/\mathbf{x}, D) = \frac{1}{\sqrt{(2\pi\sigma_t^2)}} \exp\left\{-\frac{(t - y_{MP})^2}{2\sigma_t^2}\right\}$$

where the mean of the distribution y_{MP} is the most probable value and the variance σ_t is taken as a measure of the confidence in (or error bars on) the predictions. Using the Bayesian framework all the data can be utilised to estimate and tune the model parameters, hence there is no requirement for splitting the dataset. This is important when the dataset is small or difficult to obtain, as in the case of *in vivo* data.

The Bayesian framework for training neural networks controls the complexity of the model in order to avoid overfitting. The correlation plateau found suggested that a two hidden node model is best suited for the given dataset. Increasing the number of nodes beyond the simplest model does not improve the correlation, suggesting that it has eliminated the effect of the extra parameters. The 2-node BNN model was found to be reproducible and consistent using different random initial conditions for estimating noise and correlation coefficient.

Therefore, we chose to build our QSAR model and report the results based on BNN.

Results and Discussion

To integrate the predictions and the error bars in one measure, we present the results as a probability of MAS score exceeding the limit score for mild irritancy of 50. This value was chosen based on the qualitative interpretation of eye irritancy scale by Kay and Calandra (1962).

A good model was obtained with r^2 of 0.89. Despite this good fit, in some cases predicted values varied quite considerably from actual values. This is due in part to biological variability in the Draize test. The standard deviation of the noise estimate of the measured data in our predictions was ± 13 (Draize scale of 1 - 110). Cases where the predictions were poor seemed to be where tests had been conducted at relatively low concentrations. A 3D plot of MAS, concentration and Log P was drawn to identify the most suitable concentration with which to conduct our simulations. The plot showed that increasing the concentration allowed a wider range of Log Ps to be explored and thus a larger bandwidth of potential cationics to be investigated (Figure (i)).

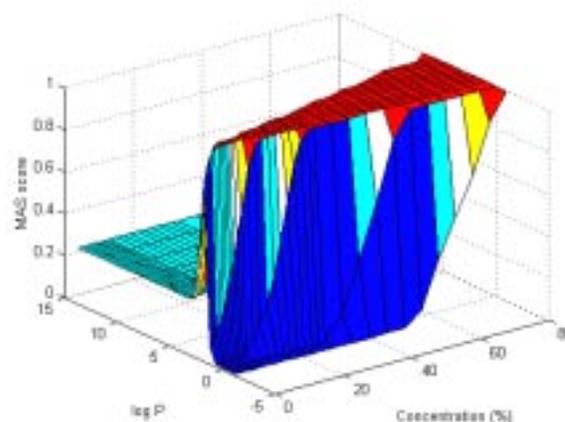
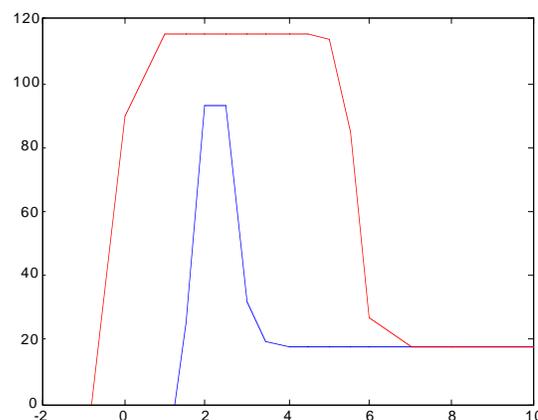


fig (i) Experiments varying Log P and concentration assuming a fixed Log CMC of -3 and a molecular volume of 275 depict a red area where the probability of the MAS score exceeding 50 is greater than 0.75. This illustrates the requirement for experiments to be conducted at a higher concentration to encompass a wider band of cationics.

Using the model, with a range of concentrations from 1% to 80%, a concentration of at least 40% was required to capture sufficient cationic surfactants in the range of Log P values. The latter being between -2 and 13 .

The relationship between Log P and MAS depicted a maximum at between 2 and 3 which then decreased and tended towards zero at high Log Ps (above 5) as expected. (Figure (ii))

fig (ii) The relationship between Log P and MAS depicts a maximum between 2 and 3. The plot also highlights that performing experiments at higher concentrations allows the exploration of a wider range of cationics.

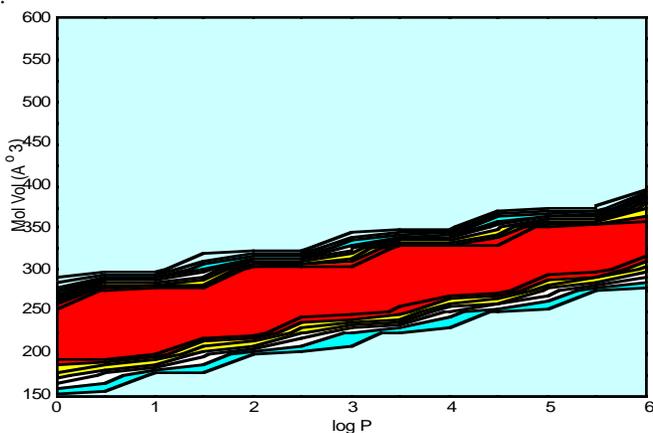


blue = 5 %

red = 50 %

The relationship between molecular volume and MAS was then explored through simulation experiments looking at a 3D plot of the Log P limited to range of 0 to 6 and concentration kept fixed at 40%. (Figure (iii))

fig (iii) The variation of MAS score (coloured contours) with Log CMC fixed at -3 and concentration constant at 40%. The optimum range of molecular volume for severe eye irritation (probability of MAS to exceed 50 is greater than 0.75) to occur is between 150 and 350. There is a weak association with Log P.



The plot indicates that at a fixed concentration and Log CMC at -3, the probability of the MAS score exceeding 50 is greater than 0.75 for molecular volumes between 150 to 350 A^3 . Above a volume of 350 to 500 A^3 the MAS score is likely to be low. There is a weak association of molecular volume with Log P within the range of Log P values simulated.

The last relationship explored was that between Log CMC, Log P and MAS. (Figure (iv))

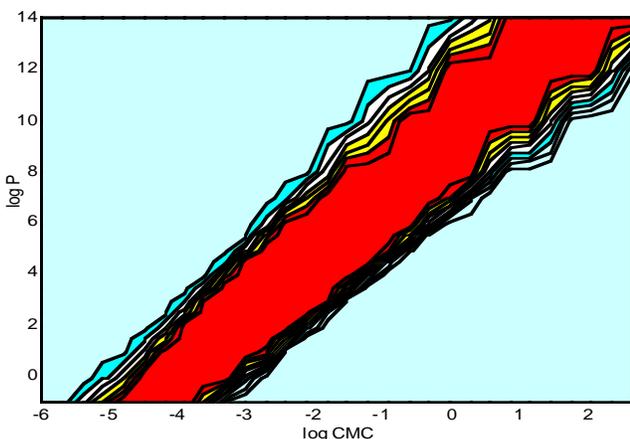


fig (iv) The variation of the MAS scores (coloured contours) with Log P and Log CMC at a fixed concentration (40%) and molecular volume (275).

It was observed as Log P increased, a higher Log CMC was required to cause a response. Or as the Log CMC was increased a molecule of sufficient high Log P would elicit a response indicating that there is a small bandwidth for an effect to take place. For a given molecular volume and concentration, the ridge of a probabilistically high MAS score draws a 45° line between Log CMC and Log P. The position of this “line” is dependent on molecular volume but for a given volume of 275, the band of effect for Log

CMC is between 2 and -5. As Log P increases, aqueous solubility and therefore Log CMC decreases resulting in reduced irritation potential.

We rationalised that the relationship for the three properties was so intricate that a 3D plot of all three variables at a constant concentration of 40% would depict the molecular surface where severe eye irritation existed. A higher concentration would reflect the same general trends though the window of effect for severe irritation would be narrowed. The representation reflects a skewed conelike figure. (Figure (v))

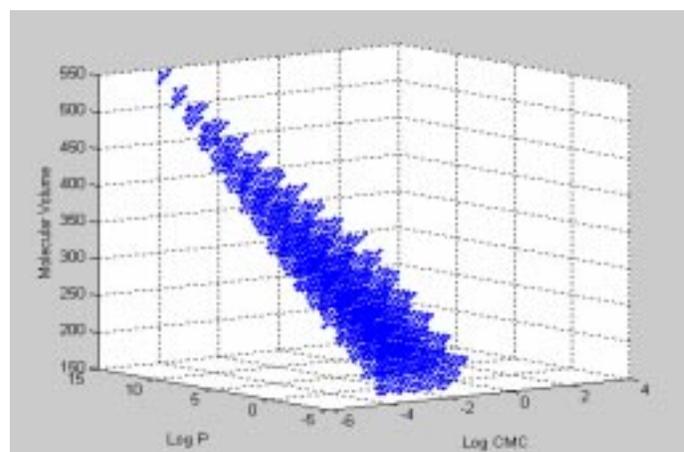


fig (v) The region of molecular space at a fixed concentration (40%) where the probability of MAS exceeding a score of 50 is greater than 0.75 represents a conelike structure. The entire boxgrid represents the region of available molecular space.

We simulated the entire range of molecular properties in the figure of the cube which represents the entire hypothetical molecular space (that is mathematically feasible). The cone highlights the region of that space where the probability of the MAS score exceeding 50 is greater than 0.75. In reality, some of the hypothetical molecules within the cone and outside will not be feasible practically to synthesise and the expert judgement of a physical chemist would be useful in such an instance.

Conclusions

The relationships between the chosen physicochemical properties have been found to be highly correlated in a non-linear manner. We have shown that the region of molecular space, which probabilistically exceeds an irritation score of 50, is in fact a skewed conelike representation. We feel this is of use in trying to assess the likelihood of eye irritation of new cationic surfactants within the molecular space without the need for testing.

References

Cronin M T D and Dearden J C. (1995) QSAR in Toxicology. Prediction of non-lethal mammalian toxicological endpoints

and expert systems for toxicity prediction. *QSAR* 14:(6) 518-523

Cutler R A and Drobeck H P. (1970). Toxicology of cationic surfactants. In *Cationic Surfactants*. Edited by Eric Jungermann pp. 527-615 Marcel Dekker Inc., New York.

Draize J H, Woodard G and Calvery H O. (1944). Methods for the study of irritation and toxicity of substances applied topically to the skin and mucous membrane. *J. Pharmacol* 82 377-390

ECETOC. (1992). Report No. 48. Eye irritation: Reference compounds data bank. Brussels.

MacKay D. (1994). Bayesian methods for backpropagation networks. In E.Domany, van Hemmen J L, and Schulden K. (Eds), *Models of Neural Networks III*, Chapter 6. New York: Springer-Verlag.

Kay J H and Calandra J C. (1962) Interpretation of eye irritation tests. *J.Soc Cosmet Chem* 13 281-289

Rodford R A, Patlewicz G, Ellis G and Barratt M D. (1998) A QSAR Model for the eye irritation of cationic surfactants. Submitted to *Toxicology in Vitro*

Spielmann H., Liebsch M, Sabibe K, Modenhauer F, Wirnsberger T, Holzhutter H-G, Schneider B, Glaser S, Gerner I, Paper W J W, Kreiling R, Krauser K, Miltenburger H G, Steiling W, Luepke N, Muller N, Kreuzer H, Murmann P, Spengler J, Bertram-Neis E, Siegemund B and Wiebel F J. (1996) Results of a validation study in Germany on two in vitro alternatives to the Draize eye irritation test, the HET-CAM test and the 3T3 NRU cytotoxicity test. *ATLA* 24 741-858

Waters J, Kleiser H H, How M J, Barratt M D, Birch R R., Fletcher R J, Haigh S D, Hales S G, Marshall S J and Pestell T C. (1991). A new rinse conditioner active with improved environmental properties. *Tenside Surf. Det.* 28, 460-468.

Table 1 Eye irritation dataset for cationic surfactants

Name	MAS Score	\$MAS	Conc.	Log CMC	Log P	Mol. Vol.	Literatureref
Benzalkoniumchloride	108	110331	10	-3.0498	2.506	27624	ECETOC 1992
Benzalkoniumchloride	83.8	82.1999	5	-3.0498	2.506	27624	" "
Benzalkoniumchloride	56.3	45.2386	1	-3.0498	2.506	27624	" "
Benzalkoniumchloride	34.3	45.2386	1	-3.0498	2.506	27624	" "
Cetylpyridiniumbromide	89.7	94.3172	10	-3.15	1.634	27799	" "
Cetylpyridiniumbromide	85.8	76.1521	6	-3.15	1.634	27799	" "
Cetylpyridiniumbromide	36	42.1274	1	-3.15	1.634	27799	" "
DMDHTAC	8.5	18.4782	5	-6.81	13.089	5192	In-house result
DMDHTAC	5.5	18.4782	1	-6.81	13.089	5192	" "
STMAC	47.33	18.4744	5	-3.696	4.068	29664	" "
STMAC	39.67	18.4700	1	-3.696	4.068	29664	" "
CTAC	36.67	19.5646	1	-3.062	2.996	2746	" "
CTAB	14.25	19.5646	1	-3.062	2.996	2746	" "
OAAP	9.67	8.6647	1	-3.696	2.581	30607	" "
MTATIMS	60.22	60.0232	75	-6.81	12.517	59925	" "
MTATIMS	5.75	5.57198	3	-6.81	12.517	59925	" "
Arquad 12	3.6	0.86453	1	-1.794	0.852	218.12	Cutler & Drobeck, 1970
Arquad 12	59.6	57.4891	10	-1.794	0.852	218.12	" "
Arquad 16	3.6	19.5646	1	-3.062	2.996	2746	" "
Arquad 16	47.5	47.1814	10	-3.062	2.996	2746	" "
Arquad C	17.2	22.4402	10	-1.794	0.852	222	" "
Arquad S	14.3	18.5313	10	-3.696	3.021	286.1	" "
Arquad T	30	18.4258	1	-3.696	3.604	29723	" "
Arquad 2C	16	17.2621	2	-3.75	6.657	368.03	" "
Arquad 2HT	11.7	18.4781	10	-6.81	13.089	5192	" "
Arquad 18	11.9	18.4800	10	-3.696	4.068	297.58	" "
Aliquat H226	16	18.4781	10	-6.81	13.089	517.95	" "
Atlas G263	91	90.2035	35	-3.062	3.02	318.35	" "
Hamburg Ester Quat	4	18.4782	50	-8.33	9.214	537.55	Waters et al, 1991

DMDHTAC : Dimethyl dihardened tallow ammonium chloride

STMAC : Stearyltrimethylammonium chloride

CTAC/ CTAB : Cetyltrimethylammonium chloride/bromide

OAAP : Oxyethyl alkylammonium phosphate

MTATIMS :Methyl-1-tallowamidoethyl-2-tallow

imidazolium methyl sulphate

\$MAS :predicted MAS