J Comput Aided Mol Des (2008) 22:339–343
DOI 10.1007/s10822-008-9202-y

Predicting allergic contact dermatitis: a hierarchical structure–activity relationship (SAR) approach to chemical classification using topological and quantum chemical descriptors

Subhash C. Basak · Denise Mills · Douglas M. Hawkins

Received: 25 September 2007 / Accepted: 20 February 2008 / Published online: 13 March 2008

Abstract A hierarchical classification study was carried out based on a set of 70 chemicals—35 which produce allergic contact dermatitis (ACD) and 35 which do not. This approach was implemented using a regular ridge regression computer code, followed by conversion of regression output to binary data values. The hierarchical descriptor classes used in the modeling include topostructural (TS), topochemical (TC), and quantum chemical (QC), all of which are based solely on chemical structure. The concordance, sensitivity, and specificity are reported. The model based on the TC descriptors was found to be the best, while the TS model was extremely poor.

Keywords Allergic contact dermatitis · Hierarchical structure–activity relationship · Model validation · Ridge linear discriminant analysis · Ridge regression · Theoretical molecular descriptors

This paper is dedicated to the memory of Dr. Phil Magee and was presented at the Phil Magee Memorial Symposium at the American Chemical Society 234th National Meeting in Boston, MA, USA, August 19–23, 2007.

Electronic supplementary material The online version of this article (doi:10.1007/s10822-008-9202-y) contains supplementary material, which is available to authorized users.

S. C. Basak (✉) · D. Mills
Natural Resources Research Institute, Center for Water and Environment, University of Minnesota, Duluth, 5013 Miller Trunk Hwy, Duluth, MN 55811, USA
e-mail: sbasak@nrri.umn.edu

D. M. Hawkins
School of Statistics, University of Minnesota Twin Cities, 224 Church Street SE, Minneapolis, MN 55455, USA

Abbreviations

ACD	Allergic contact dermatitis
HiSAR	Hierarchical structure–activity relationship
LOO	Leave-one-out (cross-validation)
OLS	Ordinary least squares
RR	Ridge regression
TS	Topostructural
TC	Topochemical
QC	Quantum chemical
STO-3G	Approximates a Slater-type orbital (STO) by combining 3 Gaussian functions (3G)

Introduction

One important toxic property of chemicals arising out of their interaction with the living skin is allergic contact dermatitis (ACD). This is a cell mediated immune reaction either to the chemical itself (hapten) or its metabolites, i.e., breakdown products, which react with cellular materials such as proteins after presumably passing through the stratum cornium. In the latter case, the chemical agent may be called a prohapten, the precursor of the hapten that precipitates ACD directly.

Prediction of ACD potential is important for industry and regulatory agencies [1]. Since carrying out bioassays, such as the local lymph node assay, exhaustively for all industrial chemicals is very costly, various authors have attempted to understand the fundamental molecular and physicochemical factors behind the precipitation of the ACD reaction. It has been proposed that the chemicals must cross the stratum cornium in order to produce the effect in deeper tissues such as the epidermis and dermis. Electronic aspects of xenobiotics and their metabolites

have also been indicated to be responsible for their reaction with proteins leading to ACD [2]. Based on the latter idea, Rosenkranz et al. [3] carried out QSARs for a lager set of ACD positive and negative chemicals using the Multicase methodology [3]. They considered predicted Salmonella mutagenicity as the factor representing electrophilic character and used it in discriminating between ACD and non-ACD chemicals. The result was not very encouraging. The authors concluded that not only electronic, but also non-electronic structural factors were involved in the causation of ACD. They also found that when one looks at biophores influential in determining ACD and non-ACD properties, almost the very same subset of indicators appears in both the cases. This indicates that the phenomenon of ACD is a subtle one which needs to be analyzed in terms of a diverse set of structural descriptors which encode information about various aspects of molecular architecture.

Therefore, in this study we have carried out a hierarchical classification study of a set of 70 chemicals, 35 positive for ACD and 35 negative, using a diverse set of molecular descriptors comprising topostructural, topochemical, as well as quantum chemical (QC) descriptors.

Materials and methods

Experimental data

The experimental database was obtained from an earlier publication by Magee et al. [4]. The dependent variable is binary in nature, classified either as a substance that produces an ACD response or that which does not. Although Magee et al. [4] report 36 allergens and 36 non-allergens, we excluded one pair of cis/trans isomers with differing activity values, namely Nerol and Gerniol. Thus our models are based on 35 allergens and 35 non-allergens, which are identified in Table 1.

Theoretical molecular descriptors and hierarchical SAR

Software programs including *POLLY v2.3* [5], *Triplet* [6], *Molconn-Z v3.5* [7], and *Gaussian 03W* [8] were used for the calculation of more than 350 molecular descriptors, each of which is derived solely from chemical structure without the need for any additional experimental data. The descriptors used in the current study can be classified into three hierarchical subsets based on level of complexity and demand for computational resources:

$$\mathrm{TS} < \mathrm{TC} < \mathrm{QC}$$

At the low end of the hierarchy are the topostructural (TS) descriptors. This is the simplest of the three classes in that molecular structure is viewed only in terms of connectivity, not as a chemical entity, and thus no chemical information is encoded. Examples include path length descriptors [9], path or cluster connectivity indices [9, 10], and number of circuits. The topochemical (TC) descriptors are more complex in that they encode chemical information, such as atom and bond type, in addition to encoding information about how the atoms are connected within the molecule. Examples of TC descriptors include neighborhood complexity indices [11], valence path connectivity indices [9], and electrotopological state indices [12]. The TS and TC are two-dimensional descriptors which are collectively referred to as topological indices. They are straightforward in their derivation and can be calculated very quickly. Since they do not encode conformational information, they are not complicated by the uncertainty which arises from conformational assumptions. At the upper end of the hierarchy are the QC descriptors which encode electronic aspects of chemical structure. QC descriptors may be obtained using either semi-empirical or ab initio calculation methods. The latter can be prohibitive in terms of time required for calculation, especially for large molecules. In this study, ab initio descriptors were calculated at the STO-3G level. A complete list of the descriptors calculated for use in the current study is provided as supplementary material. The hierarchical SAR approach used in this study involves first creating a predictive model using only the simple TS descriptors. This is followed by the development of models based on the TS + TC descriptors, and finally the TS + TC + QC descriptors. In this way, by examining the quality of the resulting models, we are able to examine the relative importance of the various descriptor classes in predicting the property or activity of interest. If we find that the more resource intensive descriptors do not result in models of greater predictive ability, we can save time and avoid their calculation. In addition to the hierarchical models, we have also developed models based on single descriptor classes.

Statistical methods

Any descriptor with a constant value for all chemicals in the data set was omitted, as was one descriptor of each perfectly correlated descriptor pair ($r = 1.0$), as identified by the CORR procedure of the *SAS* statistical package [13]. Prior to analysis, all remaining molecular descriptors were transformed by $\ln(x + c)$, where x represents the original value and c is a constant added to prevent arithmetic error. In most cases, $c = 1$, as the original values are generally greater than -1. A small number of descriptors, however, have minimum values less than or equal to -1, in which case the constant added was the smallest natural number that would provide a positive sum for $(x + c)$. The log transformation was motivated by the fact that several of the

Table 1 Chemicals used to develop ridge linear discriminant models

Name	CAS	Name	CAS
Allergens			
Butanediol diacrylate	1070-70-8	Cinnamaldehyde	104-55-2
Dihydrocoumarin	119-84-6	2,4-Dinitro-1-chlorobenzene	97-00-7
Hydroquinone	123-31-9	Isoeugenol	97-54-1
Kelthane	115-32-2	α-Methylcinnamaldehyde	101-39-3
p-Phenylenediamine	106-50-3	*N*-Phenyl-α-naphthylamine	90-30-2
3,5,3′,4′-Tetrachlorosalicylanilide	1154-59-2	4-Chloro-3-methylphenol	59-50-7
Ethylenethiourea	96-45-7	Ethylene glycol dimethacrylate	97-90-5
Eugenol	97-53-0	Hexachlorophene	70-30-4
Glycidyl benzoate	13443-29-3	Methyl 4-hydroxybenzoate	99-76-3
1,2,6-Hexanetriol	106-69-4	2-Hydroxyethyl methacrylate	868-77-9
4-Hydroxybenzoic acid	99-96-7	*N*-*i*-pr-*N*′-ph-*p*-phenylenediamine	101-72-4
Indomethacin	53-86-1	Penicillin G	61-33-6
Griseofulvin	126-07-8	*N*-Methylolchloroacetamide	2832-19-1
Phthalic anhydride	85-44-9	3-Aminophenol	591-27-5
Piperonal	120-57-0	Captan	133-06-2
Benomyl	17804-35-2	Kitazin P	26087-47-8
1,3-Dodecanesultone	58568-60-8	2-Methoxy-4-propenylphenol	97-54-1
Picryl chloride	88-88-0		
Non-allergens			
Indole	120-72-9	1-Bromo-2-phenylethylene	103-64-0
3,5-Dimethylcyclohex-3-en-1-yl CHO	68039-48-5	4-Methoxybenzyl alcohol	105-13-5
6-Isopropylquinoline	135-79-5	2-Phenylethanol	60-12-8
β-Naphthyl methyl ketone	93-08-3	Methyl-2-nonenoate	111-79-5
g-Nonalactone	104-61-0	Linalool	78-70-6
Methyl 2-aminobenzoate	134-20-3	4-Methoxyacetophenone	100-06-1
2-Methoxy-4-propenylphenyl acetate	93-29-8	5-Methyl-3-heptanone oxime	22457-23-4
4-Methoxybenzaldehyde	123-11-5	10-Undecenal	112-45-8
g-Undecalactone	104-67-6	α,α-Dimethylphenethyl alcohol	100-86-7
2-Isobutylquinoline	93-19-6	Diethyl phthalate	84-66-2
Serine	302-84-1	Tyrosine	556-03-6
Dimethyl isophthalate	1459-93-4	Arachidonic acid	506-32-1
3-Phenylpropionaldehyde	104-53-0	10-Undecenol	112-43-6
p-Cresyl acetate	140-39-6	Benzoic acid	65-85-0
Cholesterol	57-88-5	Glycerin	56-81-5
Leucine	328-39-2	Estradiol	50-28-2
Testosterone	58-22-0	Pregnane	24909-91-9
α,α -Dimethyl phenylethylacetate	151-05-3		

predictors have severely right skew distributions [14] calling for a symmetrizing transformation, while other predictors were generally neither helped nor harmed by the log transformation. Subsequently, the independent variables were standardized by way of autoscaling (mean = 0, standard deviation = 1).

A ridge linear discriminant analysis [15] was used to model the binary (+/−) data. This approach was implemented using a regular ridge regression (RR) computer code. Ridge regression [16, 17] is an appropriate methodology when the number of independent variables exceeds the number of observations and when the independent variables are intercorrelated, and it is designed to utilize all available descriptors without requiring variable selection. RR gives an estimated coefficient vector that is linear in Y, thus it is considered a 'linear smoother'. The RR coefficient vector **b** is given by:

$$\mathbf{b} = (\mathbf{X}^{\mathbf{T}}\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{Y} \quad (1)$$

where k is a non-negative constant known as the 'ridge constant'. k is selected using a generalized cross-validation method, and it controls the amount of 'smoothing'

done in RR. A value of $k = 0$ would correspond to ordinary least squares (OLS) regression. In numerous comparative studies, RR has outperformed alternative regression methods such as partial least squares (PLS) or principal components regression (PCR) [18–22].

To turn the regression into a classification, a cutoff of 0.5 was selected, i.e., those compounds with predicted values greater than 0.5 were considered to be positive for ACD, and those with predicted values of 0.5 or less were considered to be negative for ACD. The analysis was done both with and without leave-one-out (LOO) cross-validation, for comparative purposes. The concordance, sensitivity, and specificity, as defined below, were obtained from the modeling results.

$$\text{Concordance} = (\#\,\text{correct predictions}/\#\,\text{total chemicals}) \times 100\% \quad (2)$$

$$\text{Sensitivity} = (\#\,\text{correct positive predictions}/\#\,\text{total positive chemicals}) \times 100\% \quad (3)$$

$$\text{Specificity} = (\#\,\text{correct negative predictions}/\#\,\text{total negative chemicals}) \times 100\% \quad (4)$$

One additional statistic measure that is useful in identifying important descriptors, thereby providing some insight into the mechanistic basis, is the t value, which is defined as the model coefficient divided by its standard error. Those descriptors with high $|t|$ values are known to be important in predicting the property or activity being modeled. It should be clearly noted, however, that no conclusions can be drawn with respect to descriptors with small $|t|$ values.

Results and discussion

The concordance, sensitivity, and specificity associated with each of the hierarchical classification models are provided in Tables 2 and 3, based on the non-validated and LOO cross-validated methods, respectively. The k values for each of the hierarchical models are also provided in Tables 2 and 3.

The TS model is extremely poor. With no detectable relationship between ACD and the descriptors, an 'infinitely' large ridge constant is obtained which has the effect of shrinking all of the coefficients to zero, ultimately producing the TS classification results reported in Tables 2 and 3.

It is interesting to note that the TC descriptors gave the best statistics, being equal to or better than the QC descriptors. It is also interesting to note that the addition of STO-3G QC descriptors did not improve the predictive ability over and above that which was derived from the TC indices. It is tempting to speculate that the indices in the TC group take care of the stereo-electronic aspects behind ACD which are hypothesized to be responsible for the ACD reaction of these chemicals.

We examined those compounds which were predicted incorrectly by the cross-validated TC concordance model. Of the 70 compounds included in the study, 23 were predicted incorrectly. However, the majority had RR values near the arbitrary 0.5 classification cutoff. If we look at those compounds which were incorrectly classified with RR values far above or far below the 0.5 cutoff, say below 0.3 or above 0.7, five compounds are identified; namely, 1,2,6-hexanetriol, indomethacin, 1,3-dodecanesultone, α-methylcinnamaldehyde, and ethylene glycol dimethacrylate.

Table 2 Resubstitution models

Descriptors	k	Concordance (%)	Sensitivity (%)	Specificity (%)
TS	0.180E + 10	50	0	100
TS + TC	171	89	86	91
TS + TC + STO-3G	316	86	83	89
TS	0.180E + 10	50	0	100
TC	117	91	91	91
STO-3G	52.3	73	71	74

Table 3 LOO cross-validated models

Descriptors	k	Concordance (%)	Sensitivity (%)	Specificity (%)
TS	0.180E + 10	0	0	0
TS + TC	171	66	60	71
TS + TC + STO-3G	316	59	51	66
TS	0.180E + 10	0	0	0
TC	117	67	63	71
STO-3G	52.3	60	63	57

Since cis/trans isomers can be discriminated by QC descriptors, it was of interest to use the STO-3G model in order to predict the activity values for the two isomers omitted from the study, namely Nerol and Geraniol. The activity values were not accurately predicted, with ridge regression values of 0.45308 and 0.69145 corresponding to activity values of 0 and 1 for Nerol and Geraniol, respectively, based on our classification cut-off value of 0.5.

A perusal of the indices which are significant in terms of high $|t|$ value can provide some idea about the mechanistic basis of the derived SARs. Some of the more influential descriptors include the lower-order information theoretic descriptors (IC_0, IC_1, SIC_0, SIC_1, CIC_0, CIC_1), Triplet descriptors (AZN_4, AZV_4), path-length descriptor of order 1 (P_1), and a hydrogen bonding descriptor (NumHBa). AZN_4, AZV_4, and P_1 are related to the size of the molecules. Various authors have reported in the past the role of molecular size in ACD. The second class of influential descriptors are IC, SIC, and CIC indices defined by the application of information theory in the chemical and bonding neighborhood of atoms in the molecule. Since polarity and polarizability of molecules have been speculated to have an important role in ACD, the IC, SIC, and CIC indices might reflect these features of the chemicals. We also found that hydrogen bonding descriptors are important, as found by Magee et al. [4]. So, we found that the intuition of Magee et al. [4] is vindicated by our current studies using purely calculated molecular descriptors.

We also examined the $|t|$ values of the model based on STO-3G descriptors alone, as it was reported previously that the HOMO–LUMO gap is important in predicting skin sensitization [23]. Of the six STO-3G descriptors included in the model, $LUMO_1$ was found to have the highest $|t|$ value while HOMO–LUMO gap had the lowest. It must be noted again, however, that no conclusions can be drawn with respect to the importance of descriptors with low $|t|$ values.

It is essential that the modeling approach selected is appropriate for the data situation [24], and that model validation be performed in a statistically sound manner [25, 26]. A properly cross-validated model provides a realistic representation of predictive ability, in contrast to a resubstituted model which often provides an overly optimistic view (Tables 2 and 3). It is encouraging that the TC descriptors were found to be adequate to predict ACD, as they can be calculated quickly and inexpensively for any chemical, including those not yet synthesized. Future work will include chemical classification using recursive partitioning [27], which is an attractive methodology applicable for larger data sets.

Acknowledgements This is contribution number 482 from the Center for Water and the Environment of the Natural Resources Research Institute. Research reported in this paper was supported in part by Grant F49620-02-1-0138 from the United States Air Force and Cooperative Agreement Number 572112 from the Agency for Toxic Substances and Disease Registry. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Office of Scientific Research, the U.S. Government, or the Agency for Toxic Substances and Disease Registry.

References

1. Walker JD, Gerner I, Hulzebos E, Schlegel K (2004) QSAR Comb Sci 23:721
2. Ashby J, Basketter DA, Paton D, Kimber I (1995) Toxicol 103:177
3. Rosenkranz HS, Klopman G, Zhang YP, Graham C, Karol MH (1999) Environ Health Perspect 107:129
4. Magee PS, Hostynek JJ, Maibach HI (1994) In: Rougier A, Goldberg AM, Maibach HI (eds) Alternative methods in toxicology, vol 10. In vitro skin toxicology. Mary Ann Liebert, Inc., NewYork, pp 281–291
5. POLLY version 2.3 (1988) Copyright of the University of Minnesota
6. Filip PA, Balaban TS, Balaban AT (1987) J Math Chem 1:61
7. Molconn-Z Version 3.5 (2000) Hall associates consulting, Quincy, MA
8. Gausssian 03W, Version 6.0 (Revision C.02) (2004) Gaussian, Inc., Wallingford CT
9. Kier LB, Hall LH (1986) Molecular connectivity in structure–activity analysis. Research Studies Press, Letchworth, Hertfordshire, UK
10. Randic M (1975) J Am Chem Soc 97:6609
11. Roy AB, Basak SC, Harriss DK, Magnuson VR (1983) In: Avula XJR, Kalman RE, Liapis AI, Rodin EY (eds) Mathematical modelling in science and technology. Pergamon Press, New York, pp 745–750
12. Kier LB, Hall LH (1999) Molecular structure description: the electrotopological state. Academic Press, San Diego, CA
13. SAS Institute, Inc. (1988) In SAS/STAT User Guide, Release 6.03 edn. Cary, NC
14. Weisberg S (2005) Applied linear regression. Wiley, New York
15. Campbell NA (1980) Appl Stat 29:5
16. Hoerl AE, Kennard RW (1970) Technometrics 12:55
17. Hoerl AE, Kennard RW (2005) Technometrics 12:69
18. Basak SC, Mills D (2005) ARKIVOC 2005:308
19. Frank IE, Friedman JH (1993) Technometrics 35:109
20. Basak SC, Mills D, Mumtaz MM, Balasubramanian K (2003) Indian J Chem 42A:1385
21. Basak SC, Mills D, Hawkins DM, El-Masri H (2003) Risk Anal 23:1173
22. Basak SC, Mills D, Gute BD (2006) SAR QSAR Environ Res 17:515
23. Cronin MTD, Basketter DA (1994) SAR QSAR Environ Res 2:159
24. Hawkins D, Basak S, Shi X (2001) J Chem Inf Comput Sci 41:663
25. Hawkins DM, Basak SC, Mills D (2003) J Chem Inf Comput Sci 43:579
26. Kraker JJ, Hawkins DM, Basak SC, Natarajan R, Mills D (2007) Chemometr Intell Lab Syst 87:33
27. Hawkins DM, Young SS, Rusinko A (1997) Quant Struct–Act Relat 16:296