

# Characteristic Substructures and Properties in the Chemical Carcinogenicity Studied by the Cascade Model

Takashi Okada

Center for Information & Media Studies, Kwansei Gakuin University  
1-1-155 Uegahara, Nishinomiya, Japan  
okada@kwansei.ac.jp

**Abstract:** The cascade model is a rule induction methodology using the levelwise expansion of the lattice. An attribute-value pair is expressed as an item, and every node in the lattice is specified by an itemset and by its supporting instances. If the distribution of the class attribute values shows a large change along a link in the lattice, the link is represented as a rule "IF *added-item-along-link* added on *itemset-on-upper-node*, THEN *class-i*". The strength of the rule is measured by the *BSS* value of the link. In this study, we utilize linear substructure fragments and several physicochemical properties to describe a rule. A fragment leads to one of the two items [frag-i: y] and [frag-i: n] depending on whether or not the fragment exists in a molecule. Application of the cascade model to these items data set gives us rules about carcinogenicity. We could find several rules with large *BSS* values. Substructures and properties that appear in these rules are expected to provide a starting point for further chemical and biological study. Several rules with classification capability are used to predict the carcinogenicity for the compounds in the test set.

## 1 Introduction

The importance of SAR (structure activity relationship) study between chemical structures and biological activity is well established in the medicinal sciences. Moreover, the innovation in the high throughput screening technology resulted in the vast amount of SAR data, and a new data mining technology is expected to facilitate the drug development process. The principal aim of this paper is to acquire SAR rules from the Predictive Toxicology Challenge dataset [1] that lead to valuable hypothesis generation for further chemical and biological studies. The process employed here will be widely applicable to the qualitative SAR recognition problem.

We use the cascade model to extract valuable rules. The condition of a rule is represented by the combination of items. An item denotes the existence or the absence of a molecular fragment, or it may be a category of a numerical property.

Section 2 gives a brief introduction of the cascade model. The computation procedure for the challenge problem and its results are shown in Section 3. However, accurate classifications are not the aim of this paper. Rules with large *BSS* values do not always lead to accurate classifications, but they provide interesting viewpoints to analyze the data and to proceed to further research. Some of these rules are referred in Section 3.4. The last section gives a discussion on the preferable improvements in the mining process.

## 2 The Cascade Model

The model was originally proposed by the author [2]. It can be considered as an extension of the association rule mining. The method creates an itemset lattice where an [attribute: value] pair is employed as an item to constitute itemsets. Links in the lattice are selected and expressed as rules. That is, we watch the distribution of the RHS attribute values along all links, and if there appears a sudden change of distribution along some link, then we bring the two terminal nodes of the link into focus. Think that the itemset at the upper end of a link is [A: y], and an item [B: n] is added along the link. If a sharp activity decrease is found along this link, we can write a rule with the following expression,

IF [B: n] added on [A: y] THEN [Activity: low].

where the added item [B: n] is the main condition of the rule, and the items on the upper end of the link ([A: y]) are considered as preconditions. We can put any number of items in the RHS of a rule, if its distribution shows a strong interaction with the main condition.

In order to evaluate the strength of a rule, the within-group sum of squares (WSS) and between-group sum of squares (BSS) are defined by the following formulae [3, 4],

$$WSS_i = \frac{n}{2} \left( 1 - \sum_a p_i(a)^2 \right), \quad (1)$$

$$BSS_i = \frac{n^L}{2} \sum_a \left( p_i^L(a) - p_i^U(a) \right)^2, \quad (2)$$

where  $i$  designates an attribute; the superscripts U and L indicate the upper and lower nodes, respectively;  $n$  shows the number of supporting cases of a node; and  $p_i(a)$  is the probability of obtaining the value  $a$  for attribute  $i$ . The BSS takes a large value when the cases at the lower node show exceptional distribution compared to that of the upper node. Then, we can set our focus to the main condition part.

The formulation of the model was extended to cover the mining of classification rules and characteristic rules in a unified framework [5]. When we employ a mining method using the lattice expansion, there always appears the problem of combinatorial explosion in the number of nodes. A new pruning criterion opened a way to cope with this difficulty [6]. The cascade model was implemented as DISCAS, and it has already been applied to the analysis of chemical mutagenicity problem successfully [7].

### 3 Results and Discussion

#### 3.1 Computation by DISCAS

We need to provide an itemset expression of a molecular structure. The itemset does not need to restore the structural formula. However, an expert needs to understand the meaning of items. We employed all linear fragments in four class Blind\_0.10\_fragment\_table's by Kramer [8]. The number of fragments is categorized to  $y$  (presence) and 0 (absence). Also included are 9 physicochemical properties given by Treymers [9]. Their names and categorization thresholds are CLOGP (0 4), FLEX (0.05 0.25 0.50), VOLUME (150 300), SURF\_AREA (150 300), HBD (0 2), HBA (0 2), LUMO (-0.15 -0.10 -0.05), HOMO (-0.25 -0.20) and Dipole (2 4). We select them, as they are easy to understand. Categorizations are simply done by the visual inspection of histograms.

Four datasets for male (female) rat (mouse) were analyzed by DISCAS software, where the pruning conditions were set to  $minsup = 0.01$  and  $thres = 0.1$ ; their meanings are in [6]. DISCAS generated a lattice containing 40000 to 60000 nodes after 7 to 12 minutes using a PC with 450MHz Pentium III. A link was selected as a rule candidate if its BSS is larger than 1.7 (0.5% of cases). The rule selection process chose 134 - 919 candidate links, and they were represented as three rule sets with 10 to 30 rules.

#### 3.2 A sample rule

The strongest rule, the first rule in the first rule set, has the following expression in the application to the male mouse data set,

IF [HBA = 0] added on [FLEX > 0.5] THEN [MM = p]  
43.0% -> 94.7%; BSS: 5.08; Cases: 79 -> 19 .

The precondition means that a molecule is very flexible, while the main condition reveals the absence of hydrogen bond acceptors. The RHS denotes that a large change is observed in the percentage of compounds with [MM = p] (positive carcinogenicity for male mouse). The second line of this rule denotes that only 19 compounds satisfy the main condition among 79 compounds selected by the precondition. The percentage of [MM = p] increases from 43.0% to 94.7%, and the BSS value of this rule is 5.08.

Figure 1 shows a pie chart illustrated by Spotfire software. It shows the distribution of positive and negative cases for categorized HBA (y-axis) and FLEX (x-axis) values. Red and blue areas denote positive and negative cases, respectively. Y-axis categories are HBA=0, 1 HBA 2, and HBA 3 from the bottom. X-axis categories are FLEX 0.05, 0.05<FLEX 0.25, 0.25<FLEX 0.5, and FLEX>0.5 from the left.

We can recognize the characteristic increase of positive compounds in the solid box compared to that in the dotted box. This kind of visualization is effective to understand the nature of the distribution, and to detect nonsense rules coming from the accidental distribution changes.

DISCAS writes an optional RHS terms upon requests. That is, it denotes an attribute value pair if it has a high correlation to the main condition.

THEN O = n            34.2%→100.0%; BSS:8.23

THEN C-C-Cl = y    26.6%→ 78.9%; BSS:5.21

The absence of O is inferred directly from the absence of HBA. The presence of C-C-Cl is also highly correlated with the main condition, and it may be a point to be considered in further research.

### 3.3 Prediction of test data

Rules by the cascade model do not intend to give high classification accuracy, but their aim is to give valuable insights for further study. In fact, it is an interesting rule if the positive probability decreases from 0.9 to 0.5, but it does not work in the classification task. However, some rules possess enough accuracy to be used for classification. For example, the rule in the previous section can be interpreted in the following form.

IF [HBA = 0] and [FLEX > 0.5] THEN [MM = p]  
Cases: 19; Accuracy: 94.7%; BSS(root): 6.09

Here, *BSS*(root) is the *BSS* value when we employ the root node as the upper node of a rule. We calculated *BSS*(root) values for all rules. Five classification rules with the largest *BSS*(root) values are selected for positive and negative classes, respectively. If no new compounds in the test data set match the rule condition, we select a rule with the next largest *BSS*(root). The sample rule described above has a large *BSS*(root) value, but it could not find applicable compounds in the test data set.

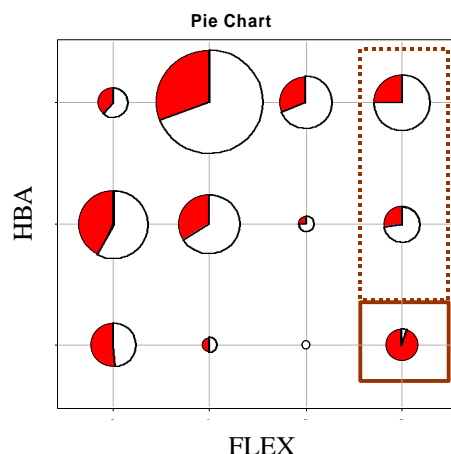
The results of classifications are shown in Table 1. "?" means that the test compound has not matched any of the rules condition. When rules give conflicting classifications, "pos", "equ", or "neg" are assigned depending on the accuracies of the applied rules.

**Table 1. Number of classifications in the four test data sets.**

	pos	equ	neg	?
MR	18	30	111	26
FR	29	2	61	93
MM	11	9	123	42
FM	38	6	71	70

### 3.4 Strong rules

Table 2 shows rules with *BSS* > 3.0. We could find no such strong rules in the rat data sets. When plural rules share the same main condition in an application to the same data set, weaker rules are shown as comments to the strongest rule. We think that this expression is useful to develop hypotheses for further study. Another interesting information comes from the strength comparison of the same rule among different species. For example, the distribution changes along FM-1 rule are MM: (.40 .60)/228 → (.69 .31)/49, FM: (.42 .58)/227 → (.73 .27)/48, MR: (.47 .53)/227 → (.61 .39)/44 and FR: (.35 .65)/234 → (.48 .52)/44.



**Figure 1. Distribution of pos/neg cases against categorized HBA and FLEX in male mouse data set.**

**Table 2. Strong rules (BSS>3.0) derived from four data sets.**

No.	Main condition	Preconditions	Changes in distribution <sup>†</sup>	BSS
MM1	[HBA = 0]	[FLEX > 0.5]	(.43 .57) / 79 → (.95 .05) / 19	5.08
	If no precondition is applied, (.38 .62) / 336 → (.65 .35) / 55, BSS=4.03.			
	If [C-c:c:c: n] is the precondition, (.40 .60) / 228 → (.69 .31) / 49, BSS=4.26.			
	If [c:c-N: n] is the precondition, (.39 .61) / 227 → (.65 .35) / 55, BSS=3.79.			
MM2	[C-Cl: y]	[C-O: n]	(.40 .60) / 214 → (.68 .32) / 44	3.56
	The percentage of [HBD = 0] also changes 52%→98%, [c:c-N = n] changes 62%→98%.			
MM3	[Dipole > 4]	[Cl: y]	(.54 .46) / 94 → (.07 .93) / 14	3.11
FM1	[HBA = 0]	[C-c:c:c:c: n]	(.42 .58) / 230 → (.73 .27) / 48	4.54
	The percentage of [c:c : n] also changes 58% → 90%.			
	If [C-c:c:c:c:c : n] is the precondition, (.42 .58) / 238 → (.73 .27) / 48, BSS=4.46.			
	If [C-c:c : n] is the precondition, (.43 .57) / 227 → (.73 .27) / 48, BSS=4.37.			
FM2	[N: y]	[c:c:c-N: n], [O: n]	(.57 .43) / 79 → (.16 .84) / 19	3.22
FM3	[HBA: y]	[C-O: n], [c:c : n]	(.57 .43) / 72 → (.16 .84) / 19	3.22
FM4	[C-C-O: y]	[C-Cl: y], [N: n]	(.60 .40) / 52 → (.00 .10) / 9	3.20

<sup>†</sup> In (*pos neg*)/#case, *pos* and *neg* show the probabilities of positive and negative cases, respectively. #Case denotes the number of cases. Distributions before and after the application of the main condition are shown.

## 4 Concluding Remarks

Detailed discussion of the individual rule in Table 2 is beyond the scope of this paper. But, useful insight for the carcinogenic mechanism is expected if we inspect the 2D and 3D structures of the molecules cited in the strong rules.

Lastly, we have to note that the pruning condition is too tight to mine all meaningful rules. Further developments of DISCAS system are necessary to scale up the computation size and to make things easier in the interpretation process of derived rules.

## References

- [1] Helma, C., King, R.D., Kramer, S., Srinivasan, A.: The Predictive Toxicology Challenge for 2000-2001, <http://www.informatik.uni-freiburg.de/~ml/ptc/index.html>.
- [2] Okada, T.: Finding Discrimination Rules Using the Cascade Model, *J. Jpn. Soc. Artificial Intelligence*, **15**, pp.321-330 (2000).
- [3] Gini, C.W.: Variability and Mutability, contribution to the study of statistical distributions and relations, *Studi Economico-Giuridici della R. Universita de Cagliari* (1912). Reviewed in Light, R.J., Margolin, B.H.: An Analysis of Variance for Categorical Data, *J. Amer. Stat. Assoc.* **66**, pp.534-544 (1971).
- [4] Okada, T.: Sum of Squares Decomposition for Categorical Data, *Kwansei Gakuin Studies in Computer Science*, **14**, pp.1-6 (1999). <http://www.media.kwansei.ac.jp/home/kiyou/kiyou99/kiyou99-e.html>.
- [5] Okada, T.: Rule Induction in Cascade Model based on Sum of Squares Decomposition, *Principles of Data Mining and Knowledge Discovery (Proc. PKDD'99)*, pp.468-475, *LNAI 1704*, Springer-Verlag (1999).
- [6] Okada, T.: Efficient Detection of Local Interactions in the Cascade Model, *Proc. PAKDD2000, LNAI*, Springer-Verlag (2000).
- [7] Okada, T.: SAR Discovery on the Mutagenicity of Aromatic Nitro Compounds Studied by the Cascade Model", *Proc. Int. Workshop KDD Challenge on Real-world Data*, pp.47-53, *PAKDD-2000* (2000).
- [8] Kramer, S.: <http://www.informatik.uni-freiburg.de/~ml/ptc/README.fragments>.
- [9] Treymers: <http://www.informatik.uni-freiburg.de/~ml/ptc/treymers.txt>.