社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2003－ＩＣＳ－133 (8)
2003／9／14

Knowledge Discovery using Attributes from 3D Molecular Structures
－Importance of Active User's Response－

Takashi OKADA† Masumi YAMAKAWA† Hirotaka NIITSUMA† and Naomi KAMIGUCHI‡

† Department of Informatics, Kwansei Gakuin University 2-1 Gakuen, Sanda-shi, Hyogo, 669-1337 Japan
‡ Osaka Research Ctr., Takeda Chemical Industries, 2-17-85 Juso-honmachi, Yodogawa-ku, Osaka, 532-8686 Japan
E-mail: † {okada-office, abz81166}@ksc.kwansei.ac.jp, ‡ kamiguchi_naomi@takeda.co.jp

Abstract Active responses from experts plays an essential role in the knowledge discovery of SAR (structure activity relationships) from drug data. Experts often think of hypotheses, and they want to reflect these ideas to the attribute generation and selection process. Authors have analyzed SAR of dopamine antagonists using the cascade model. In this paper, we generated attributes indicating the presence of hydrogen-bonded fragments from 3D coordinates of molecules, which were suggested by experts. The selection of attributes by experts has been shown to be useful in obtaining valuable knowledge.

Keyword Attribute Generation and Selection, Cascade Model, Dopamine, 3D Molecular Structure, Hydrogen-bond

３次元分子構造からの属性生成による知識発見
－利用者による積極的関与の重要性－

岡田 孝† 山川 真透† 新妻 弘崇† 上口 尚美‡

†関西学院大学理工学部 〒669-1337 兵庫県三田市学園 2-1
‡武田薬品工業薬物機能研究所 〒532-8686 大阪市淀川区十三本町 2-17-85
E-mail: † {okada-office, abz81166}@ksc.kwansei.ac.jp, ‡ kamiguchi_naomi@takeda.co.jp

あらまし 薬品の化学構造と生理活性間の相関データをマイニングしようとする場合，利用者である専門家の積極的な関与が必要である．専門家はマイニング過程で多くの仮説をアドホックに思いつくが，その内容を属性として取り込む作業を効率化することにより，質の高い知識発見を行うことが期待できる．著者らはこれまでからドーパミンのアンタゴニスト作用を有する薬品群について，グラフ表現された構造式を対象として、カスケードモデルによる解析を行ってきた．本報告では分子の３次元構造から水素結合のフラグメントを生成し，利用者自身が属性選択を繰り返し行うことにより，知識発見の過程が効率的に遂行できることを示す．

キーワード 属性生成，属性選択，カスケードモデル，ドーパミン，３次元分子構造，水素結合

1. Introduction

The importance of SAR (structure-activity relationship) studies relating chemical structures and biological activity is well recognized. Early studies used statistical techniques, and concentrated on establishing quantitative structure activity relationships involving compounds sharing a common skeleton. However, it is more natural to treat a variety of structures together, and to identify the characteristic substructures responsible for a given biological activity. Recent innovations in high throughput screening technology have produced vast amounts of SAR data, and the demand for a new data mining method to facilitate drug development has increased.

The author has already analyzed SAR's [1, 2] using the cascade model that we developed [3]. Later, I pointed out the importance of the "datascape survey" in the mining process in order to obtain valuable knowledge. We added several functions to the mining software of the cascade model (DISCAS) to facilitate the datascape survey [4, 5].

This new method was recently used in a preliminary study of the SAR for the antagonist activity of dopamine receptors [6]. The resulting interpretations of the rules were highly regarded by experts of drug design. However, the interpretation process of rules employed the visual inspection of supporting chemical structures as an essential step, and a user had to be very careful so that he/she did not miss characteristic substructures.

Fruitful mining results will never be obtained unless an active user's response is not expected. This paper reports an attempt to reflect expert's ideas to attribute generation

and selection process. Attributes indicating the hydrogen-bonded fragments are created using 3D coordinates of molecular structures. The attribute selection process provides a framework for active user's responses. The next section briefly describes the aims of mining as well as the basic introduction to the mining method employed. The attribute generation and selection methods are described in Section 3. Typical rules and their interpretations are discussed in Section 4.

2. Aims and Basic Methods

2.1. Aims and Data Source for the Analysis of Dopamine Antagonist Activity

Dopamine is a neurotransmitter in the brain. Neural signals are transmitted via the interaction between dopamine and proteins known as dopamine receptors. There are five different receptor proteins, D1 – D5, each of which has a different biological function. Their amino acid sequences are known, but their 3D structures are not yet established.

Certain chemicals act as antagonists for these receptors. An antagonist binds to a receptor, but does not function as a neurotransmitter. Therefore, it blocks the function of the dopamine molecule. Antagonists for these receptors might be used to treat schizophrenic patients. The structural characterization of these antagonists is an important problem in developing new schizophrenia drugs.

We used the MDDR database of MDL Inc. as the data source. It contains 1,349 records that describe dopamine (D1, D2, D3, and D4) antagonist activity. Some of the compounds affected multiple receptors. The problem is to discover the structural characteristics responsible for each type of antagonist activity.

2.2. The Cascade Model

The cascade model can be considered an extension of association rule mining [3]. The method creates an itemset lattice in which an [attribute: value] pair is used as an item to constitute itemsets. Links in the lattice are selected and interpreted as rules. That is, we observe the distribution of the RHS (right hand side) attribute values along all links, and if a distinct change in the distribution appears along some link, then we focus on the two terminal nodes of the link. Consider that the itemset at the upper end of a link is [A: y] and item [B: n] is added along the link. If a marked activity change occurs along this link, we can write the rule:

```
IF [B: n] added on [A: y]
            Cases: 200 ==> 50 BSS=12.5
THEN [Activity]:.80 .20 ==> .30 .70 (y n)
THEN [C]:       .50 .50 ==> .94 .06 (y n)
Ridge
[A: n]: .70 .30/100  ==> .70 .30/50 (y n)
```

where the added item [B: n] is the main condition of the rule, and the items at the upper end of the link ([A: y]) are considered preconditions. The main condition changes the ratio of the active compounds from 0.8 to 0.3, while the number of supporting instances decreases from 200 to 50. *BSS* means the between-groups sum of squares, which is derived from the decomposition of the sum of squares for a categorical variable. Its value can be used as a measure of the strength of a rule. The second "THEN" clause indicates that the distribution of the values of attribute [C] also changes sharply with the application of the main condition. This description is called the *collateral correlation*.

2.3. Functions for the Datascape Survey

New facilities introduced to DISCAS (mining software for the cascade model) consist of three points. Decreasing the number of resulting rules is the main subject of the first two points [4]. A rule candidate link found in the lattice is first greedily optimized in order to give the rule with the local maximum *BSS* value, changing the main and preconditions. Let us consider two candidate links, (M added on P) and (M added on P'). Here, their main conditions, M, are the same. If the difference between preconditions P and P' is the presence/absence of one precondition clause, the rules starting from these links converge on the same rule expression, and it is useful for decreasing the number of resulting rules.

The second point is the facility to organize rules into principal and relative rules. In the association rule system, a pair of rules, R and R', are always considered independent entities, even if their supporting instances overlap completely. We think that these rules show two different aspects of a single phenomenon. Therefore, a group of rules sharing a considerable amount of supporting instances are expressed as a principal rule with the largest *BSS* value and its relative rules. This function is useful for decreasing the number of principal rules to be inspected, and to indicate the relationships among rules.

The last point is to provide ridge information of a rule [5]. The last line of the aforementioned rule shows ridge information. This example describes [A: n], the ridge region detected, and the change in the distribution of

"Activity" in this region. Compared to the large change in the activity distribution for the instances with [A: y], the distribution does not change on this ridge. This means that the *BSS* value decreases sharply if we expand the rule region to include this ridge region. This ridge information is expected to guide the survey of the datascape.

3. Attribute Generation and Selection

3.1. Basic Scheme

We used two kinds of explanation attributes generated from the structural formulae of chemical compounds. The first group consists of four physicochemical estimates: the HOMO and LUMO energy levels, the dipole moment, and LogP. The first three values were estimated by the molecular mechanics and molecular orbital calculations using MM-AM1-Geo method provided by *Cache*. LogP values were calculated by ClogP program in *Chemoffice*.

The other group is the presence/absence of various structural fragments. Obviously, the number of all possible fragments is too large. We generated linear fragments with lengths shorter than 10. One of the terminal atoms of a fragment was restricted to be a heteroatom or a carbon constituting a double or triple bond.

Linear fragments were expressed by constituent elements and bond types. The number of coordinating atoms and the presence/absence of attached hydrogens are added to the terminal and its adjacent atoms. C3H:C3-C-N-C3=O1 is a sample expression, where "C3H" means a three-coordinated carbon atom with at least one hydrogen atom attached, and ":" denotes an aromatic bond.

Number of fragments generated from dopamine antagonist data was more than 120,000, but most of them are useless as they appear only a few times among 1349 molecules. On the other hand, the upper limit of the attributes is about 150 in the current implementation of DISCAS. Therefore, we selected 73 fragments, of which probability of appearance is in the range: 0.15 – 0.85.

3.2. Hydrogen-bonded Fragments

When we visualize chemical structures that satisfy rule conditions, we sometimes see a group of compounds that might be characterized by an intramolecular hydrogen-bond, XH...Y, where X and Y are usually oxygen or nitrogen. However, the fragment generation scheme above-mentioned utilizes only the graph topology of the structure, and we cannot recognize the hydrogen-bond.

The results of MM-AM1-Geo calculation used for the estimation of physicochemical properties provide 3D coordinates of atoms. Therefore, we can detect the existence of a hydrogen-bond using 3D coordinates. We judged the existence of a hydrogen-bond, XH...Y, when the following conditions were satisfied.

1. Atom X is O, N, S or 4 coordinated C with at least one hydrogen atom.
2. Atom Y is O, N, S, F, Cl or Br.
3. The distance between X and Y is less than 3.7 Å if Y is O, N or F; and it is less than 4.2 Å otherwise.
4. Structural formula does not contain fragments X-Y or X-Z-Y, where any bond type will do.

When these conditions are satisfied, we generate fragments: Xh.Y, V-Xh.Y, Xh.Y-W, and V-Xh.Y-W, where "h." denotes a hydrogen-bond, and neighboring atoms V and W are included. Other notations follow the basic scheme.

Application to the dopamine antagonists dataset resulted in 431 fragments, but the probability of the most frequent fragment was less than 0.1. Therefore, all hydrogen-bonded fragments are not employed in the standard mining process.

3.3. Attributes Selection and Spiral Mining

When experts think of a characteristic substructure for the appearance of some biological activity, it can be expressed by few linear fragments. Some fragments might lead to clear and strong rules even if its probability of appearance in the data set is out of the specified range: 0.15 – 0.85.

We provided a mechanism to add specified fragments as the attribute used in the mining. Consequently, a user can repeat the following steps, in order to discover better characteristic substructures.

1. Prepare fragment attributes by the basic scheme.
2. Compute rules.
3. Read resulting rules and make hypotheses by the language of chemistry.
4. Confirm the hypothesis by browsing supporting structural formulae.
5. If one notices a characteristic fragment that does not appear in the rule, add the fragment as an attribute. Go to step 2.
6. Repeat until satisfactory results are obtained.

Since experts can put his/her ideas in the mining process, adding fragments and reading rules are now an interesting exploration. This spiral mining process is not limited to the incorporation of hydrogen-bonded fragments, but it is applicable to all kinds of fragments.

4. Results and Discussion

For the calculations with the DISCAS ver.3 software the parameters were set at *minsup*=0.01, *thres*=0.1, *thr-BSS*=0.01, *min-rlv*=0.7. These parameters are defined elsewhere [3, 4, 5]. We added 32 fragments after reading rules and inspecting chemical structures. They consist of 27 hydrogen-bonded fragments with probability_of _appearance > 0.02, and 5 fragments (N3-C3:C3-O2, N3H-C3:C3-O2, N3-C3:C3-O2H, N3H-C3:C3-O2H, O1=S4). The inspection process is not complete yet, but we can depict two examples that show the effectiveness of the current method.

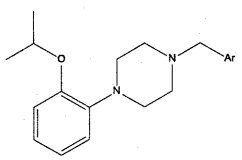
4.1. D2 antagonist activity

The analysis of this activity was hard in the former study. That is, the strongest rule indicating active compounds takes the following form when we use the basic scheme for the attribute selection.

```
IF [C4H-C4H-O2: y]  added on   [ ]
THEN D2AN: 0.32 0.68 ==> 0.62 0.38(on off)
THEN C3-O2:      0.42 ==> 0.89 (y)
THEN C3H:C3-O2: 0.33 ==> 0.70 (y)
Ridge [C3H:C3H:C:C3-N3: y]
  D2AN: 0.49 0.51/ 246 --> 0.92 0.08 / 71
```

There appear no preconditions, and the main condition shows that an oxygen atom bonded to alkyl carbon is important. However, this finding is so different from the common sense of experts, and it will never be accepted as a useful suggestion. In fact, the ratio of active compounds is only 62%. Collateral correlations suggest that the oxygen atom constitutes aromatic ethers, and the ridge information indicates the relevance of aromatic amines. But, it has been difficult even for a expert to make a reasonable hypothesis.

Experts found a group of compounds sharing the skeleton shown below, when they browse the supporting structures. So, they added fragments consisting of two aromatic carbons bonded to N3 and O2. This addition did



not change the strongest rule, but there appeared a new relative rule shown below.

```
IF [N3-C3:C3-O2: y]  added on   [ ]
THEN D2AN: 0.31 0.69 ==> 0.83 0.17(on off)
THEN HOMO: .16 .51 .33 ==> .00 .19 .81
        (low medium high)
THEN C3H:C3-N-C-C4H-N3: 0.24 ==> 0.83 (y)
```

This rule has a higher accuracy and it explains about 20% of active compounds. The tendency observed in HOMO value also gives us a useful insight. However, the collateral correlation on the last line shows that most compounds supporting this rule have the skeleton show above. Therefore, we cannot exclude the possibility that other part of this skeleton is responsible for the activity. Further inspection is necessary to reach satisfactory hypotheses.

4.2. D3 antagonist activity

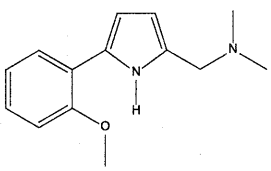
The analysis of this activity is also complex, because there appears more than 10 principal rules. We suggested 5 characteristic substructures in the former study. The strongest rule leading to this activity has the following form.

```
IF [O1: y] added on [C3-N3: n] [C3=O1: n]
THEN D3AN: 0.79 0.21 ==> 0.06 0.94(off on)
THEN C3:N2:     0.33 ==> 0.01 (y)
THEN N3H:       0.58 ==> 0.91 (y)
THEN O2:        0.45 ==> 0.76 (y)
THEN N3Hh.O2:         0.09 ==> 0.54 (y)
```

The main condition of this principal rule is an oxygen atom, and the preconditions employed are the absence of two short fragments. Therefore, its interpretation is very difficult. After the inclusion of hydrogen-bonded fragments, there appeared the last line in the collateral correlations, where a steep increase of N3Hh.O2 is observed. The relevance of this fragment was confirmed by the appearance of a relative rule shown below.

```
IF [N3Hh.O2: y] added on   [C3-N3H: n]
THEN D3AN: 0.88 0.12 ==> 0.05 0.95(off on)
```

In fact, this rule accounts for about half of the active compounds supported by the principal rule. Visual inspection of the supporting structures has shown that the following skeleton leads to this activity. We have to note that a hydrogen-bond itself is not responsible for the activity.



Another advantage of adding this hydrogen-bonded fragment is illustrated by the next principal rule.

```
IF [C4H-N3H-C3=O1: y] added on
  [LUMO: 0 - 2] [HOMO: 0 - 1] [N3Hh.O2: n]
THEN D3AN: 0.77 0.23 ==> 0.27 0.73(off on)
```

Here, the absence of the fragment appears as a

precondition. Since the compounds with this skeleton are excluded, most compounds at the upper node are inactive, giving a higher *BSS* value to the rule. After we recognize the above skeleton as active compounds, this type of rule expression is easy to understand the existence of distinct lead structures for the activity.

5. Conclusion

The proposed mining process has succeeded to evoke active responses from experts. They can put their ideas in the mining task, and step up the mining spiral by themselves. Now, experts can make rough hypotheses by reading rules. Browsing the supporting structures is still necessary. However, they do not need to be nervous when they inspect structures, since they can add many fragments and judge their importance by the resulting rules.

We have to make a note on the interpretation of resulting rules after adding attributes. If an expert adds some fragments that were found in the preliminary step of analysis, they will often appear as a rule condition. But the appearance of a rule is no guarantee of truth. The fragment might be a part of large skeleton shared by the supporting structures. He/she has to check the collateral correlations carefully. Visual inspection of structures is also necessary before he/she reaches final hypotheses.

The comprehensive analysis of ligands for dopamine receptor proteins are now under progress using the proposed system. They include not only discriminations of antagonists, but also those among agonists. Also under investigation are factors that distinguish antagonists and agonists. The results will be a model work in the field of SAR analysis.

References

[1] T. Okada, Discovery of structure activity relationships using the cascade model: the mutagenicity of aromatic nitro compounds, J. Computer Aided Chemistry, vol.2 pp.79-86, 2001.

[2] T. Okada, Characteristic substructures and properties in chemical carcinogens studied by the cascade model, Bioinformatics, vol.19, no.10, pp.1208-1215, 2003.

[3] T. Okada, Efficient detection of local interactions in the cascade model, in Knowledge Discovery and Data Mining PAKDD-2000, ed. T. Terano et al. ed., pp.193-203, LNAI 1805, Springer-Verlag, 2000.

[4] T. Okada, Datascape survey using the cascade model, in Discovery Science 2002, K. Satoh et al. ed., pp.233-246, LNCS 2534, Springer-Verlag, 2002.

[5] T. Okada, Topographical expression of a rule for active mining, in Active Mining, H. Motoda ed., pp.247-257, IOS Press, 2002.

[6] T. Okada, and M. Yamakawa, "Mining characteristics of lead compounds using the cascade model," Proc. 30th Symposium on Structure-Activity Relationships, pp.49-52, 2002 (in Japanese).