2004－ICS－138　(22)
2004／12／5

Characteristic Ligand Substructures to Dopamine Receptors

Takashi Okada and Masumi Yamakawa

The structure activity relationship studies of ligands to dopamine receptor proteins have been set to one of the main target in the active mining project. Authors started to solve this problem using the cascade model and linear fragments extracted from structural formulae. The original method of analysis was found to be not sufficient to capture the characteristic substructures, and a variety of improvements are incorporated into rule derivation process and into fragment expressions. This paper reports the final results obtained in D1 agonist analysis using the current methodology. The obtained results are evaluated to provide rational hypotheses of active sites and binding sites from a viewpoint of pharmaceutical research.

1 Introduction

The importance of SAR (structure-activity relationship) studies relating chemical structures and biological activity is well recognized. Active mining project in Japan (2000-2005) selected it as one of the target area, and a data set of chemicals that act as agonists and antagonists to dopamine receptor proteins was provided as the common data. There are 5 receptor proteins, and a chemical compound works as an agonist or an antagonist to some of these receptors. The problem was to extract specific substructures for each type receptor, and to build a model that discriminate these biological activities. Since the compounds possess diverse chemical structures, conventional technology that assumes a common skeleton among structures does not work.

The authors have already analyzed SAR's in a mutagenicity and a carcinogenicity data set using the cascade model [1, 2]. Linear fragments extracted from structural formulae were used as attributes. The early results could show some characteristic substructures, but the resulting knowledge was not sufficient from the viewpoint of drug design. In order to overcome this difficulty, we incorporated various facilities to the cascade model, and we also improved the expressions of linear fragments. These new facilities have enabled to provide a better set of rules, and now we can extract valuable knowledge from these rules.

This paper shows the latest results for the characteristic substructures obtained for D1 agonist activity. The summary of the results for every activity will be published in some expert journal. The next section briefly describes the data source and its preprocessing scheme. A brief introduction to the mining method is described in Section 3. Typical rules and their interpretations are discussed in Section 4.

2 Source Data and Linear Fragments Generation

2.1 Data Source for the Dopamine Agonists Analysis

Dopamine is a neurotransmitter in the brain. Neural signals are transmitted via the interaction between dopamine and proteins known as dopamine receptors. There are six different receptor proteins, D1 – D5 and Dauto, 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 agonists for these receptors. An agonist binds to the receptor, and it

Department of Informatics, Kwansei Gakuin University

stimulates a neuron. On the contrary, an antagonist binds to the receptor, but its function is to occupy the binding site and to block the neurotransmitter function of a dopamine molecule.

We used the MDDR database (version 2001.1) of MDL Inc. as the data source. It has about 120,000 drug records, and contains 400 records that describe dopamine (D1, D2, and Dauto) agonist activities. Some of the compounds affected multiple receptors. Some compound structures contained salts, which were omitted from the structural formulae.

2.2 Physicochemical Properties

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* software. The initial geometries were taken from those of structural formulae. LogP values were calculated by ClogP program in *Chemoffice* software. In some compounds, MO calculations reached unreasonable geometries. In other cases, ClogP calculations failed because of the lack of parameters. We omit these compounds from the data set. The final numbers for compounds are 63, 143, and 186 compounds for D1, D2 and Dauto agonists. The total number of compounds was 369, as some compounds show 2 activities. In the mining computation by the cascade model, we employed categorized physicochemical properties: their split values are (2.0, 4.0, 6.0) for LogP and dipole, (-1.0, -0.5, 0.0) for LUMO and (-9.0, -8.5) for HOMO.

2.3 Linear Fragments

The other attributes group is the presence/absence of linear structural fragments. We limited the length of linear fragments within 10. One of the terminal atoms of a fragment was restricted to be a heteroatom or a carbon constituting a double or a triple bond.

Linear fragments are expressed by constituent atoms and bonds. The terminal and its adjacent atom of a fragment are expressed by its element symbol attached by the number of coordinating atoms and by the presence/absence of attached hydrogens. Atom symbols are omitted in the intermediary part of the fragment. Lowercase letters are used for atoms in the aromatic ring. Carbonyl group was treated as a unified atom, CO, with 2 coordinating atoms. We use ":" to denote an aromatic bond. A sample expression is "c3H:c3--N-CO2", where "c3H" means a three-coordinated aromatic carbon with at least one hydrogen atom attached, "N" denotes a *tertiary* amine, and "CO2" shows a carbonyl group without attached hydrogens. The atom between "c3" and "N" can be any element.

We also generated hydrogen-bonded fragments. 3D coordinates resulting from MO calculations are used to judge the existence of hydrogen bonds. The details of all fragment generation scheme are found in [3].

Number of fragments generated from dopamine agonists data was 4,626. We omit a fragment from the attribute set, unless the probability of its appearance satisfies the condition: $0.03 < P(\text{fragment}) < 0.97$, giving 660 fragments as candidates of attributes. If there are many highly correlated fragment pairs, the computation of the lattice in the cascade model often results in the combinatorial explosion. Therefore, we omit a fragment in a correlated pair, if their correlation coefficient is greater than 0.9. The number of fragments decreased to 306 at this stage. The scheme of attribute selection using correlations is shown in [4]. However, the omission of some fragments was supposed to cause some difficulty in the rule interpretation process, and we designated some fragments to be included in the attribute set. The final number of fragments we used for the analysis was 345 in the dopamine agonists study.

3 Mining Method and Computation

3.1 The Cascade Model

The cascade model can be considered an extension of association rule mining [5]. 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 [6, 7]. That is, we observe the distribution of the activity (y n) 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: y] is added along the link. If a marked activity change occurs along this link, we can write a rule:

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

where the added item [B: y] 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*.

The number of detected links with high BSS values, however, is usually too numerous to be interpreted by human analysts. Therefore, we introduced two schemes in order to decrease the number of rules [8]. 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. This optimization of rule conditions is useful for decreasing the number of resulting rules. Some precondition clause might raise *BSS* value slightly leading to a too complex rule, and we add a new precondition clause only when its *BSS* value increases by more than 20%.

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. We used 0.6 as the value of *min-rlv* defined in [8].

We also added a ridge information to a rule that is shown at the last line of the aforementioned rule [9]. This example describes [A: n] as the ridge region detected at the outside of the current precondition. The distribution change of "Activity" in this ridge region is denoted. 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*.

3.2 Computation of Rules

We used 345 fragments and 4 physicochemical properties to construct the itemset lattice. *thres* parameter value was set to 0.125, which controls the breadth of the lattice search [7]. The resulting lattice contained 9,508 nodes, and we selected 1,762 links (*BSS* > 2.583 = 0.007*#compounds) as rule candidates. Greedy optimization of these links resulted in 407 rules (*BSS* > 5.535 = 0.015*#compounds). Organization of these rules gave us 14 principal rules and 53 relative rules. Many rules indicate characteristics leading to inactive compounds or have few supporting compounds, and we omitted those rules with activity ratio < 0.8 and those with #compounds < 10 after the application of the main condition. The final number of rules we inspected was 2 principal and 14 relative rules.

4 Results and Discussion

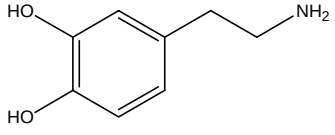
We inspected all rules in the final rule set, and we often needed to browse the supporting chemical structures using the structure visualization feature of *Spotfire* software. Table 1 summarizes important features of valuable rules guiding us to characteristic substructures for the D1 agonist activity.

R1 is the strongest rule derived from D1 agonist study. There appear no preconditions, and the activity ratio increases from 17% in 369 compounds to 96% in 52 compounds by the inclusion of catechol structure (O2H-c3:c3-O2H). Dauto activity depresses to 0% by this main condition. Other collateral correlations suggest the existence of N3H-C4H-C4H-c3, and OH groups are supposed to exist at the *meta* and *para* positions to this ethylamine substituent. However, this ethylamine group is not the indispensable substructure as N3H-C4H-C4H-c3 exists in 81% of compounds. This observation is also supported by the ridge information. That is, the selection of a region inside the rule by the introduction of a new condition [N3H-C4H--:::c3-O2H: y] results in 53 (35 actives, 18 inactives) and 35 (34 actives, 1 inactive) compounds before and after the application of the main condition, respectively. It means that there are 1 active and 17 inactive compounds when catechol structure does not exist, and we can say that N3H-C4H--:::c3-O2H cannot show D1Ag activity without the aid of catechol. Therefore, we can draw a hypothesis for D1Ag activity that catechol is the active site and that it is supported by the ethylamine substituent at *meta* and *para* positions as the binding site.

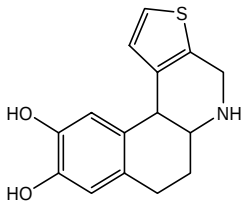
The catechol supported by ethylamine substituent is just the structure of dopamine molecule (I), and it covers 50 compounds out of 63 D1 agonists. Therefore, this hypothesis can be evaluated to be a rational one.

All 14 relative rules are associated to the principal rule: R1. Some of them employ [N3H-C4H--::c3-O2H: y] and [O2H-c3:c3: y] as the main condition with a variety of preconditions. Another relative rule has the main condition: [N3: n] and preconditions: [C4H-O2-c3: n], [C4H-C4H----:::c3-O2H: n]; characterization depending on the absence of substructures are hard. But, the supporting compounds of these relative rules overlap to those of R1, and the inspection of their structures supported the above hypothesis drawn from R1.

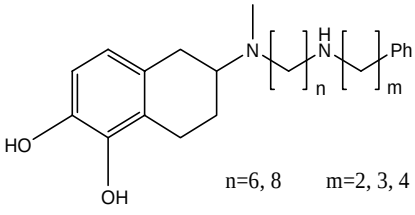
The third entry in Table 1, R1-UL12, seems to provide new substructures. Its collateral correlations indicate an N3H group at the position separated by 1, 2 and 3 atoms from an aromatic ring as well as a diphenylmethane structure. The supporting structures are found to have skeleton (II), where the thiophene ring can be benzene or furan rings. Therefore, we do not need to change the hypothesis since it contains the dopamine structure.



I



II



n=6, 8 m=2, 3, 4

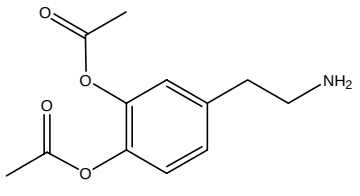
III

Table 1. Rules suggesting the characteristics of D1 agonist activity

Rule ID	Number of compounds and conditions of a rule		Distribution changes in D1Ag and collateral correlations			
			Descriptor	before		after
R1	#compounds	369 → 52	D1Ag:	17%	→	96%
			DAuAg:	50%	→	0%
	Main condition	[O2H-c3:c3-O2H: y]	C4H-C4H-:::c3-O2H	19%	→	92%
			C4H-C4H-::c3-O2H:	18%	→	98%
			N3H-C4H--:::c3-O2H:	14%	→	67%
	Preconditions	none	N3H-C4H--::c3-O2H:	10%	→	69%
			N3H-C4H-C4H-c3:	18%	→	81%
			C4H-N3H--C4H-c3	15%	→	62%
	Ridge 1: new inside		[N3H-C4H--:::c3-O2H: y]	66% in 53	→	97% in 35
R1-UL9	#compounds	288 → 16	D1Ag:	19%	→	100%
			DAuAg	51%	→	0%
	Main condition	[C4H-N3----:c3-O2H: y]	D2Ag:	36%	→	100%
			C4H-N3---:::c3-O2H:	5%	→	87%
			C4H-N3---C4H-c3:	11%	→	100%
	Preconditions	[N3H: y]	O2H-c3:c3-O2H:	17%	→	100%
R1-UL12	#compounds	170 → 12	D1Ag:	14%	→	100%
			DAuAg	40%	→	0%
	Main condition	[N3H-C4H---::c3-O2H: y]	D2Ag:	50%	→	0%
			C4H-N3H---:::c3-O2H:	8%	→	100%
	Preconditions	[N3H-C4H--:c3H:c3H: n] [O2-c3:c3: n]	C4H-N3H---C4H-c3:	7%	→	100%
			C4H-N3H-C4H-c3:	8%	→	100%
			O2H-c3:c3-O2H	12%	→	100%
			O2H-c3:::-C4H-c3:	7%	→	100%
			O2H-c3::-C4H-c3:	7%	→	100%
R14	#compounds	72 → 11	D1Ag:	15%	→	9%
	Main condition	[C4H-C4H-O2-c3: n]	DAuAg:	72%	→	0%
			CO2-O2-c3:c3H:	31%	→	64%
	Preconditions	[LUMO: 1-3] [C4H-N3---c3:c3: n] [O2-c3:c3:c3H: y]	O2-c3:c3-O2:	22%	→	64%
			N3H-C4H---:c3H:c3H:	7%	→	45%

The only exceptional relative rule was R1-UL9, which is shown as the second entry in Table1. The interesting points of this rule are the 100% co-appearance of the D2 agonist activity as well as the *tert*-amine structure in the main condition. These points make a sharp contrast to those found in R1, where *prim*- and *sec*-amines aid the appearance of D1Ag activity and the D2Ag activity was found to appear in 38% of 52 compounds. The importance of *prim*- or *sec*-amines, ethylamine substituent and catechol structure are also suggested by the precondition and collateral correlations. Inspection of the supporting structures showed that this rule was derived from compounds with skeleton (III). We could find a dopamine structure around the phenyl ring at the right in some compounds, but it could not explain D1Ag activity for all supporting compounds. Therefore, we propose a new hypothesis that the active site is the catechol at the left ring, but the binding site is the *sec*-amine at the middle of the long chain. This *sec*-amine can locate itself close to the catechol ring by the folding of $(CH4)_n$ (n=6, 8) chain.

R14 is the second and the last principal rule leading to D1Ag activity. Contrary to the R1 group rules, there appear no OH's substituted to an aromatic ring that played an essential role in the above hypothesis. It is hard to interpret this rule as the main condition and the second precondition are designated by the absence of ether and *tert*-amine substructures. But, we could find that 6 out of 11 compounds have the skeleton (IV), where vicinal OH's are transformed to esters. These esters are supposed to be hydrolyzed to OH's after absorbed to cells, and act as prodrugs.



IV

5 Conclusion

The development of the cascade model has enabled the characterization of the active site and the binding site for the D1 agonist activity. The proposed model bears a close resemblance to the dopamine molecule, and is not a striking one. But, the hypothesis is rational and it has not been published elsewhere.

Results of D2 and Dauto agonists analysis is not included in this paper, but they have yielded other reasonable models. The analysis of antagonists is being carried out successfully, too. And the work is making steady progress toward the publication in some international journal. All these results indicate that the research direction employed in the active mining project has been a fruitful one.

Acknowledgement

The authors thank Ms. Naomi Kamiguchi for her preliminary analysis.

References

[1] Takashi Okada: Discovery of Structure Activity Relationships using the Cascade Model: The Mutagenicity of Aromatic Nitro Compounds, Journal of Computer Aided Chemistry, Vol.2, 79-86 (2002).

[2] Takashi Okada: Characteristic Substructures and Properties in Chemical Carcinogens Studied by the Cascade Model, Bioinformatics, Vol.19, pp.1208-1215 (2003).

[3] Takashi Okada, Masumi Yamakawa and Hirotaka Niitsuma: Spiral Mining using Attributes From 3D Molecular Structures, ISMIS2003, Second International Workshop on Active Mining, Maebashi, pp.103-107 (2003).

[4] Takashi Okada: A Correlation-Based Approach to Attribute Selection in Chemical Graph Mining, JSAI2004, Thrid International Workshop on Active Mining, Kanazawa, pp.73-82 (2004).

[5] Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules, Proc. VLDB, pp.487-499 (1994).

[6] Takashi Okada: Rule Induction in Cascade Model based on Sum of Squares Decomposition, J.M.Zytkow and J. Rauch (Eds.): Principles of Data Mining and Knowledge Discovery , PKDD'99, pp.468-474, LNAI 1704, Springer-Verlag (1999).

[7] Takashi Okada: Efficient Detection of Local Interactions in the Cascade Model, T.Terano , H. Liu, and A.L.P. Chen (Eds.) Knowledge Discovery and Data Mining, Current Issues and New Applications, PAKDD-2000, LNAI 1805, pp.193-203 (2000).

[8] Takashi Okada: Datascape Survey using the Cascade Model, Discovery Science 2002, pp.233-246, LNCS 2534 , Springer-Verlag (2002).

[9] Takashi Okada: Topographical Expression of a Rule for Active Mining, Hiroshi Motoda (eds.) "Active Mining", pp.247-257, IOS Press (2002).