

RESEARCH ARTICLE

Open Access

# The development of a knowledge base for basic active structures: an example case of dopamine agonists

Takashi Okada<sup>1\*</sup>, Masumi Yamakawa<sup>1</sup>, Norihito Ohmori<sup>1</sup>, Sachio Mori<sup>1</sup>, Hiroshi Horikawa<sup>1</sup>, Taketo Hayashi<sup>1</sup>, Satoshi Fujishima<sup>1,2</sup>

## Abstract

**Background:** Chemical compounds affecting a bioactivity can usually be classified into several groups, each of which shares a characteristic substructure. We call these substructures "basic active structures" or BASs. The extraction of BASs is challenging when the database of compounds contains a variety of skeletons. Data mining technology, associated with the work of chemists, has enabled the systematic elaboration of BASs.

**Results:** This paper presents a BAS knowledge base, BASiC, which currently covers 46 activities and is available on the Internet. We use the dopamine agonists D1, D2, and Dauto as examples and illustrate the process of BAS extraction. The resulting BASs were reasonably interpreted after proposing a few template structures.

**Conclusions:** The knowledge base is useful for drug design. Proposed BASs and their supporting structures in the knowledge base will facilitate the development of new template structures for other activities, and will be useful in the design of new lead compounds via reasonable interpretations of active structures.

## Background

The bioactivity of chemical compounds can usually be explained in terms of common structural features, and the recognition of these features is the starting point in drug design processes. For example, if a group of compounds shares a large carbon skeleton, then recognition of this aspect is relatively straightforward at the structural formula and three-dimensional (3D) levels. However, structural aspects of drugs are very diverse, and no catalog or database of characteristic substructures exists that includes most known active compounds.

The concept of a basic active structure (BAS) is described herein and applied at the structural formula level to express characteristic features shared by a group of active compounds. This expression is similar to a combination of the pharmacophore, or the component indispensable to binding and biological activity, and the scaffold, which consists of secondary components that assist in ligand coordination. However, a BAS cannot

always be divided into these two components. Active compounds are usually categorized into groups with a BAS.

Many papers and reviews have been published that propose relationships among characteristic substructures, structural modifications, and associated biological activity. One typical approach is employed in the commercial system LHASA [1], which compiles empirical knowledge with QSAR results in toxicology and uses an expert system framework to predict the toxic effects of a new compound. Characteristic substructures of dopamine agonists have been extracted based on purely empirical considerations [2]. The present study was performed to elaborate on the characteristic substructures using dopamine agonists as an example.

Several reports have described automatic clustering and classification of chemical substructures using various types of descriptor and fingerprint analyses to select candidates for scaffold hopping and second-stage screening of pharmaceutical compounds from vast amounts of high-throughput screening (HTS) data [3]. Clustering, however, can result in mixing of different structural classes within a cluster, or conversely, the

\* Correspondence: okada-office@kwansei.ac.jp

<sup>1</sup>Department of Informatics, School of Science & Technology, Kwansai Gakuin University, 2-1 Gakuen, Sanda, Hyogo 669-1337, Japan

appearance of many clusters containing only a single compound. For example, Böcker attempted to improve cluster purity first by the formation of a 2-means cluster [4], followed by application of PCA scores [5] and the introduction of MCS [6]. However, a mixture of compounds with different structural features is essential to the clustering technique as compounds are detected from different structural classes to enable scaffold hopping. Therefore, a user must always refer to a list of supporting structures within a cluster to recognize the structural characteristics of the cluster.

Another approach incorporates descriptors of predefined graphs independent of the data set to be analyzed. For example, data-driven clustering reported by Harper [7] used reduced graph descriptors during the clustering process and displayed those clusters that exhibited a high activity ratio. Medina-Franco [8] used a similar approach and isolated classes of active chemotypes. As these methods retrieve database information through a predefined series of scaffolds, the resulting cluster is structurally homogeneous. However, the descriptors are typically ring systems with side chains omitted. When the side chain of the compound plays an important role in its biological activity, the resulting cluster often contains compounds of mixed activity and requires further interpretation on the part of the chemist. Structurally homogeneous clusters are familiar to organic chemists, and some commercial software packages employ this approach [9].

The clustering applied in these approaches depends essentially on fingerprints and descriptors, and does not take into account the activity of the compound itself. A conceptual clustering method based on chemical structures and incorporating activity ratios was proposed 20 years ago, but it was an experiment in artificial intelligence with no real practical use [10].

A different approach to extract structural characteristics from active compounds can be seen in CASE [11] and MultiCASE [12] developed by Klopmann. These methods use linear fragments and search through combinations of conditions for expression of activity, and were used to extract common substructures in dopamine agonists [13]. The same problem was analyzed to a greater depth in the present study. Klopmann's methods are now applied primarily for automated prediction of toxicity, whereby common substructures are relatively small compared to those found in pharmaceuticals.

Many computer scientists have focused their attention on graph mining and have extended Klopmann's work [14]. One of the research streams in this field uses inductive logic programming, which in its first application was used to recognize common structures of aromatic amines responsible for mutagenicity [15]. The flexibility of this method is attractive, but the vast search

space represents a problem for graph mining. Recently, the method was applied to scaffold hopping in 3D space [16]. Another research stream that tackles SAR problems can be found in the work of DeRaedt. Molfea constructed a lattice of SMILES fragments and succeeded in detecting meaningful fragments from a large data set [17]. Later, a more sophisticated, rule-learning algorithm was introduced in SMIREP [18]. The framework for association rule mining was first extended to graphs in AGM [19,20]. Recent work has included application to large data sets [21]. All of these studies represent important progress in computer science with attractive algorithms, and some valuable common structural features have been recognized from large data sets. However, most of this work was done without strict evaluation by medicinal chemists. Therefore, mixtures of trivial substructures often prevent further chemical assessment.

Despite the large database of drugs available, no systematic method exists to extract BASs that are specific for a desired activity. A collection of such BASs would provide medicinal chemists with a means to generate a general picture of their desired active compounds.

A successful method to extract BASs according to their activities is described here. The method combines data mining and the recognition of characteristic structures by medicinal chemists. Approximately 46 agonist/antagonist activities have been characterized in terms of their BAS components, as well as some repeated dose toxicities for rats caused by chemical substances. To date, the focus has been on biogenic amines, such as dopamine, serotonin (5-HT), and adrenaline, and on toxicities related to hemolytic anemia. The results have been made available on the Internet as the BAS knowledge base, Basic Active Structures in Chemicals (BASiC) [22]. BAS extraction from dopamine D1 agonists is provided as an example, and the resulting BASs from D1, D2, and Dauto agonists are described using template structures. The results provide a more detailed understanding of active structures compared to previous studies [2,13].

## Experimental

### Overview of the analysis process

The MDL Drug Data Report (MDDR) database (version 2003.1) [23] was used as the data source, which contains approximately 120,000 active compound records, including those of dopamine agonists D1 (77 records), D2 (164 records), and Dauto (198 records). As the dopamine molecule itself is an agonist of these three receptors, these molecules were expected to share some common structural features. The first step in the extraction process was to create a model with which to discriminate and characterize these three agonists. As some

of the compounds were expressed as salts, these were converted into neutral molecules by omitting a counter ion and an ammonium proton as well as adding a hydrogen atom. Tautomers were converted to the most common structure from the viewpoint of organic chemists. Duplicate structures were omitted, but optical isomers were retained in the data set. The remaining 407 structures (D1, 74; D2, 162; Dauto, 194) were evaluated, of which some affected multiple receptors.

The process of BAS extraction is described in steps (i) through (iv) (see Additional file 1). (i) Linear fragments were first extracted from the molecules and a table was constructed in which each row represented a single compound and the columns showed whether the corresponding linear fragment exists in the compound. (ii) The cascade model, a data mining method, was applied to create characteristic rules for each activity. For convenience, rules were arranged on a self-organizing map (SOM) based on their coverage of compounds (see Additional file 1). (iii) Each rule was then examined using a structural refinement system. Examination of a rule begins with optimization of the core characteristic substructure to which the rule applies. The substructure incorporates the surrounding atoms and bonds, and increases the discriminating capability between active and inactive compounds. The result is a larger, more chemically meaningful substructure and a supporting structures chart (SSC). (iv) Experienced chemists will recognize BAS candidates in the SSC, as the structural diversity and the number of compounds become relatively limited. Finally, a BAS candidate is run through the refinement system to confirm its ability to perform the desired activity. Steps (iii) to (iv) are repeated until the extracted BASs are those found in most of the active compounds. Each step of BAS extraction is shown for D1 agonists in the following subsections.

#### Generation and selection of linear fragments

Figure 1 shows a sample fragment [24] in a style similar to the Simplified Molecular Input Line Entry System (SMILES) notation. A fragment consists of two terminal atoms and the shortest connecting path of atoms. An atom symbol has qualifiers that show the number of bonding neighbors and the presence of hydrogen atoms. No notations are included to discriminate stereo- and optical isomeric substructures. A carbonyl group is treated as a single, united atom. We generated all possible fragments of length less than or equal to 10, but those containing no heteroatoms were omitted. The number of unique fragments generated was 4626 for the dopamine agonist, and 660 fragments that appeared in 3%-97% of all compounds were selected. Many cases occurred in which two fragments were highly correlated. We omitted one fragment if the correlation coefficient

of a pair was greater than 0.9 and if we could infer the existence of a fragment from the presence of another [25]. The final number of fragments was 335.

#### Rule generation by the cascade model

The cascade model is a mining method for generating characteristic rules [26,27]. It detects links in the itemset lattice where the activity ratio changes sharply in many compounds. Figure 2 shows a typical example of a link and its rule expression. Here, the problem contains four explanatory variables—OH, CO, Me, COO—and a target variable acid, which takes  $(y, n)$  values. The itemset at the upper end of the link contains item [OH:  $y$ ], and another item, [CO:  $y$ ], is added along the link. The two small tables at the side of the nodes show the frequencies of the items. We can see that a large distribution change in acid (60/40 to 54/6) occurs with addition of the [CO:  $y$ ] item. The distribution of COO also changes sharply. The cascade model searches for such links with large distribution changes, and expresses them as rules.

The textbox at the right in Figure 2 shows the derived rule. The added item [CO:  $y$ ] appears as the main condition of the rule, while the items on the upper node ([OH:  $y$ ] in this case) are preconditions. The *acid* distributions before and after application of the main condition are shown in the *THEN* clause as well as [COO:  $y$ ] distributions denoted in a *then* clause at the bottom line.

The cascade model employs the between-groups sum of squares (BSS) as a measure of rule strength to select these characteristic rules. This BSS is also used in the refinement process. Equation (1) is the sum of squares definition for categorical data as given by Gini [28]. It is a simple extension of the numerical sum of squares concept.

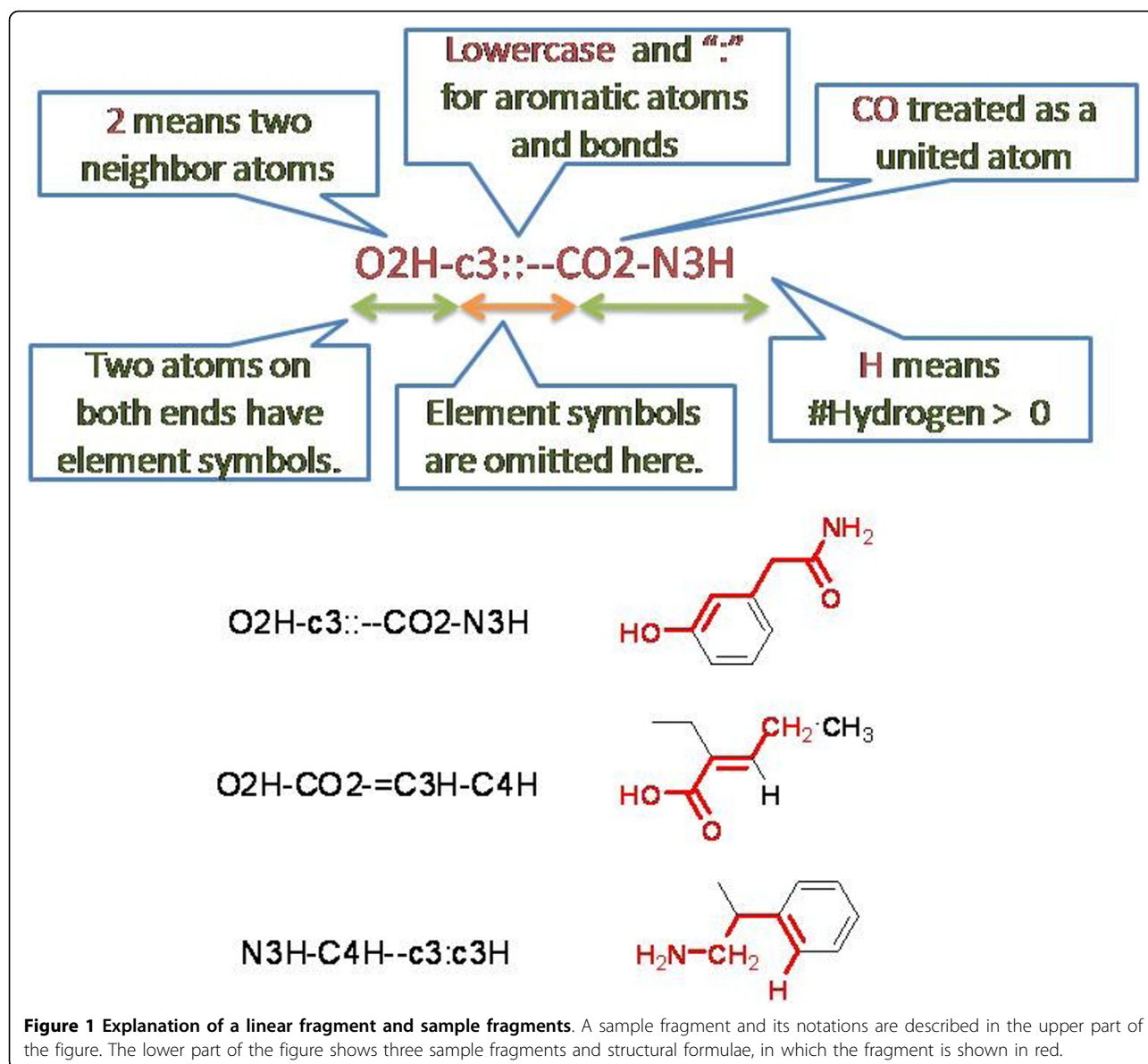
$$TSS = \frac{n}{2} \left( 1 - \sum_i p_i^2 \right) \quad (1)$$

$$WSS^g = \frac{n^g}{2} \left( 1 - \sum_i p_i^{g^2} \right) \quad (2)$$

$$BSS^g = \frac{n^g}{2} \sum_i (p_i^g - p_i)^2 \quad (3)$$

$$TSS = \sum_g (WSS^g + BSS^g) \quad (4)$$

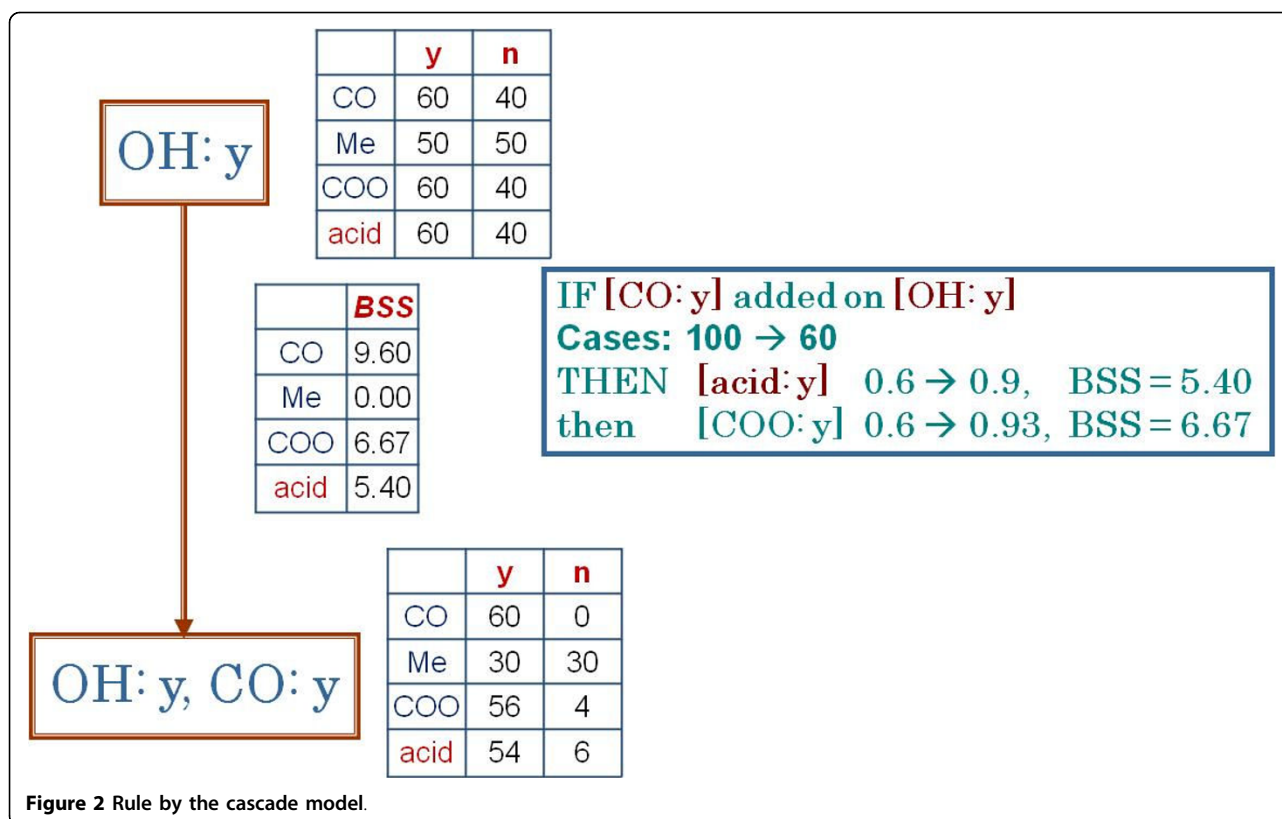
The total sum of squares (TSS) can be decomposed into the within-group sum of squares (WSS) and BSS contributions, as shown in Equations (2)-(4), where  $n$



and  $p_i$  denote the number of cases and the probability of taking value  $i$ , respectively, and the superscript  $g$  is attached to designate the group. The tree in Figure 3 shows an example of TSS decomposition. The top of the tree represents 1000 cases consisting of 800 positives and 200 negatives. The corresponding TSS value is 160, as calculated by Equation (1). The cases are then divided into two groups at the bottom according to some criterion. The distributions in these groups are 760/40 on the lower left and 40/160 on the lower right, with BSS values of 18 and 72, respectively. The distribution in the right group is inversely related to that at the top of the tree. Hence, the right link has a higher BSS value despite the smaller number of cases. We can also see that the equality of Equation (4) holds for this sample.

We considered D2 and Dauto agonists as inactive, when finding rules for D1 agonists. The application of the cascade model yielded 21 rules (see Additional file 2 for details of rule selection). Figure 4 shows the strongest rule that selects 60 D1 agonists from among 407 compounds.

The first line shows that this rule selects 60 of 407 compounds, and the associated BSS value is 35.41. The second line indicates that the main condition of this rule is [O2H-c3:c3-O2H: y], the presence of a catechol with no preconditions. The first *THEN* clause reveals that the activity ratio rises from 18% in 407 compounds to 95% in 60 compounds. The succeeding lines show variables the distributions of which change sharply before and after application of the main condition. For example, the



first *then* clause denotes that the Dauto agonist activity ratio decreases from 0.48 to 0.0, and the following *then* clauses show that these fragments often appear with catechol, suggesting the existence of a phenethylamine *meta* and *para* to the OH group. In this way, one may conceive of a dopamine structure by combining catechol and phenethylamine substructures. However, the recognition of a BAS has yet to be accomplished. For this, the characteristic environment of the catechol substructure must be considered.

#### Structural refinement

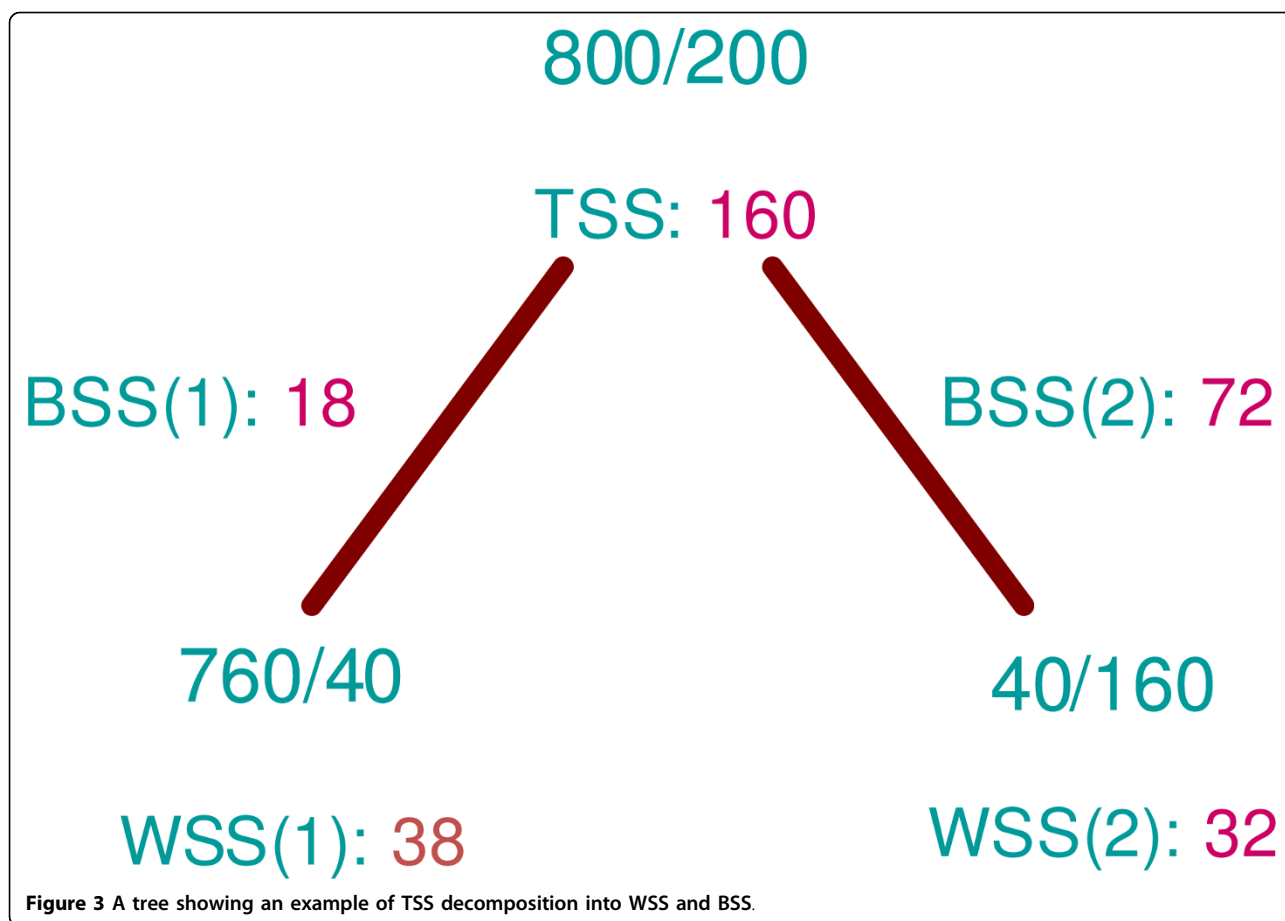
A structural refinement system was developed that accounts for the surroundings of the substructure, while maintaining the capability of the substructure to discriminate active compounds [29]. For the rule given in Figure 4, catechol was provided as an initial seed pattern, and various other structures were generated by inserting every type of bond atom pair into every position in the seed. The generated structures were then entered as a query to the database and a BSS value was computed. Query structures with the highest BSS values were used as the next seed pattern. The process was repeated until every query structure resulted in a decrease in the BSS value. The structures employed in this refinement

process are shown at the top of Figure 5, where an increase in BSS values occurred only at step 3 in this example.

The SSC displays active and inactive structures separately at each step of the refinement. The lower section of Figure 5 shows part of the SSC obtained from the last refinement step. Components in red indicate the common substructure obtained by refinement.

#### Designation of BASs and their knowledge base

The structure reached from a rule by the refinement system sometimes gives a reasonable BAS. That is, the supporting compounds of this BAS are those obtained by substituent modifications from the BAS. However, a SSC often contains a mixture of different skeletons from the viewpoint of lead compound classification. The SSC shown in Figure 5 is an example. In such cases, an experienced medicinal chemist is able to recognize BAS candidates in the SSC even though it contains a mixture of compounds with differing structural features. A chemist makes up a larger BAS candidate structure, and gives it to the refinement system again to confirm whether the BAS should be refined further. A BAS candidate query sometimes retrieves compounds with a low active/inactive ratio, and the user must then modify the candidate structure.



Cases: 407 → 60;      BSS = 35.41

IF [O2H-c3:c3-O2H: y] added on [ ]

THEN      D1Ag:      0.18 → 0.95

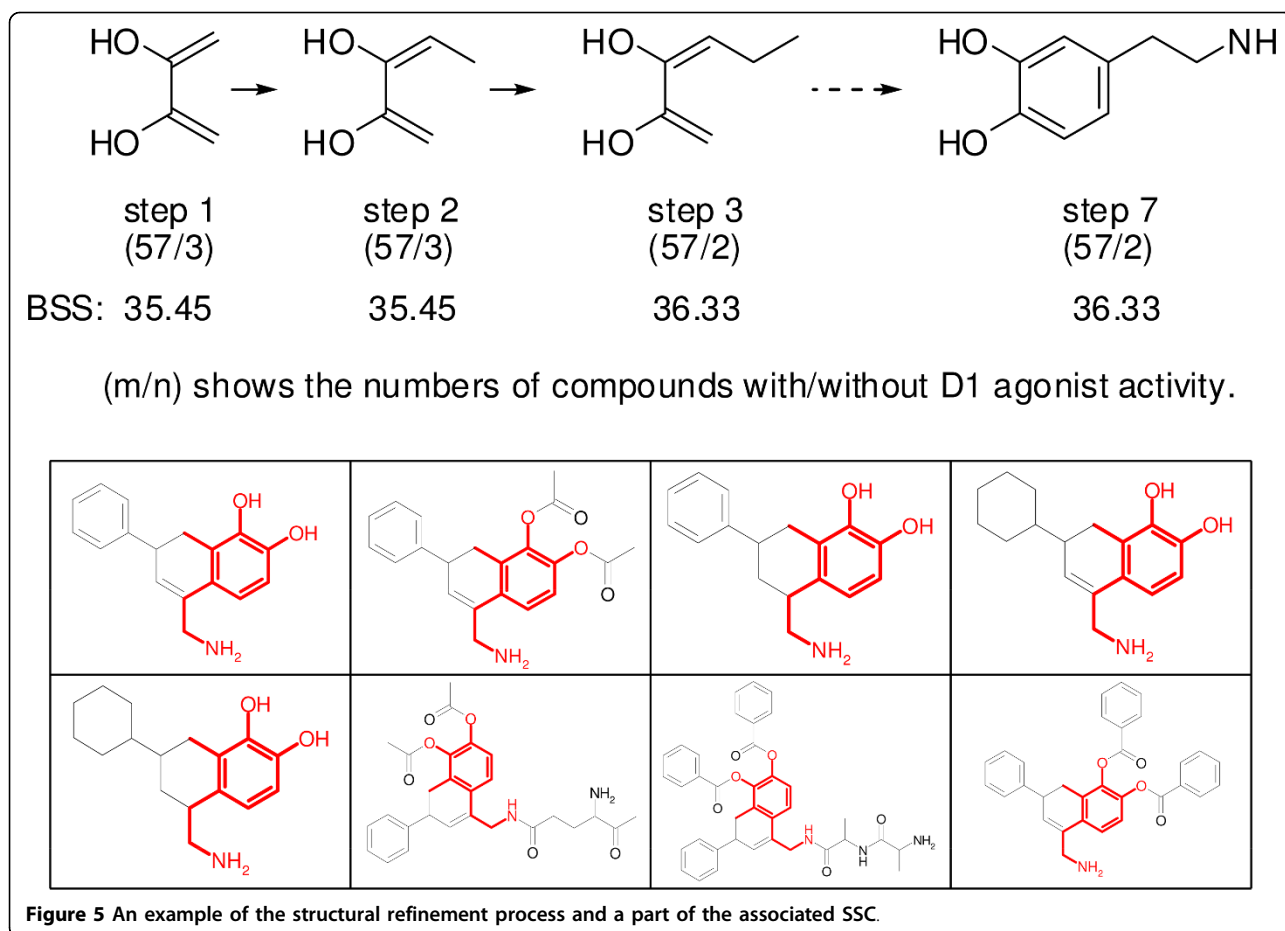
then      DAuAg:      0.48 → 0.00

then      N3H-C4H--:::c3-O2H:      0.14 → 0.67

then      N3H-C4H--:::c3-O2H:      0.10 → 0.68

then      C4H-N3H--C4H-c3:      0.15 → 0.62

Figure 4 The strongest rule from the D1 agonist problem.



The criteria used to select a structure as a BAS depends on the user's subjective judgment, and mechanizing this selection process is difficult. In this study, chemists with a great deal of experience in drug design for pharmaceutical companies selected a BAS proactively if it could inspire a chemist with the design of a new scaffold and even if the number of supporting compounds is limited.

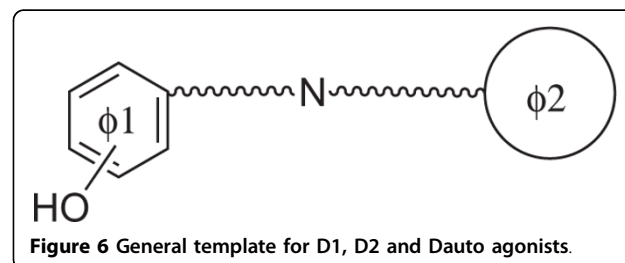
Five BASs were extracted from the SSC in Figure 5 and are presented in 57 active compounds. However, 17 active materials did not contain these BASs. The user must then examine an additional rule that extends to these 17 uncovered compounds. The extraction process is then repeated until the resulting BAS list covers the majority, if not all, of the active compounds.

The final knowledge base then consists of a list of BASs, each of which consists of a structural formula, its SMARTS expression, the number of covered active/inactive compounds, and a SSC. Miscellaneous compounds not covered by the BASs are also listed. The license of the source database prohibits users from viewing the SSC, but users can use compound ID lists to depict chemical structures if they have the right to access the source database.

## Results

### BASs and their roles in D1 agonists

BASs extracted from D1 agonists are discussed in this section. The introduction of a structural template is useful in BAS interpretations. A template depicts a typical structural pattern and serves as a standard for BAS comparisons. A common template is shown in Figure 6 that applies to all D1, D2, and Dauto agonists. The nitrogen atom at the center depicts the crucial amine group, and  $\phi 1$  is an aromatic group substituted with an OH group and separated by a certain distance (usually 2-3 atoms) from the amine.  $\phi 2$  is a lipophilic group that includes aromatic functional groups. Differences in



structural arrangements among these components are considered to be the discriminating factors determining the three agonist activities.

The following two templates shown in Figure 7 are proposed for D1 agonists, where  $\phi 1$  is a catechol and  $\phi 2$  is an aromatic ring. Figure 8 shows eight BASs extracted from D1 agonists, which cover 70 of the 74 D1 active compounds. Hereafter, the characteristic features in these templates will be discussed as they relate to these eight BASs.

The D1Ag-template1 is representative of BASs D1Ag-A to -E. A typical BAS that expresses the features of this template is D1Ag-A, which exhibits a rigid tetracyclic system in which the steric arrangements of the catechol OH, amine, and  $\phi 2$  groups are specified. Similarly, D1Ag-B and -C have rigid carbon skeletons and are well described by the template, although D1Ag-B does include a few compounds without catechol groups and D1Ag-C possesses rotational flexibility in the  $\phi 2$  arrangement.

D1Ag-D has more freedom with regard to the amine location because the amine is not relegated to a ring in most supporting compounds and the  $\phi 2$  group includes nonaromatic moieties, such as cyclohexyl groups. Esters of D1Ag-E have virtually the same structure as template1, at least on the  $\phi 1$  side, because they can be hydrolyzed to yield catechol. The amine groups of template1 are primary or secondary in most compounds except D1Ag-B. Although the arrangement of  $\phi 2$  is coded less rigorously relative to that of  $\phi 1$ , it is located 1-2 atoms from the amine group in most compounds.

D1Ag-template2 matches 18 compounds under D1Ag-F, all of which also exhibit D2 agonist activity. In addition, they share two remarkable structural features: (i)  $\phi 2$  is connected to the amine group via a long flexible chain, and (ii) another amine appears near  $\phi 2$ . Two compounds do not have catechol groups in template2, but rather in the aromatic ring on the right side of D1Ag-F, which can be interpreted via template1. In this BAS, bending of the interconnecting six or eight atoms may bring the two amine groups closer to each other and the amine on  $\phi 2$ , which then has the possibility of becoming the active center.

BASs of D1Ag-G and -H cannot be explained by these templates because they do not contain OH groups. Indole nitrogen and oxo, oxime groups likely play the role of the OH group. Three compounds of D1Ag-G show D2 agonist activity. The availability of more data in this category would enable a more detailed analysis. Four compounds (see Additional file 3) that were not covered by the eight BASs can be explained with D1Ag-template1.

#### BASs and their roles in D2 agonists

There were 162 compounds that exhibited D2 agonist activity, and their chemical structures showed markedly richer variety than those of compounds exhibiting D1 activity. The four templates shown in Figure 9 are proposed to represent the nine BASs shown in Figure 10, where D2Ag-A, -B, and -C- are subcategorized, and D2Ag-A' represents an additional BAS very similar to D2Ag-A5.

D2Ag-template1 is representative of four subBASs: D2Ag-A1 to D2Ag-A4. This template would be similar to D1Ag-template1 if the X in -XR were substituted for an oxygen atom. However, for this template, only one OH at the *meta* position of phenethylamine is required, as opposed to the two OH groups that comprised the catechol in D1Ag-template1. The positions of the amine and OH groups are fixed in these subBASs, and the OH group is positioned to one side of the line connecting the center of  $\phi 1$  to the amine, while the two OH groups are located much closer to that line in D1Ag-template1. The  $\phi 2$  group is often lacking in A2 and A4 compounds, and it is located in the flexible environment of A1 compounds, which includes a mostly aromatic lipophilic group. D2Ag-D, -E, -F, and -G are also represented by D2Ag-template1, where X in -XR is replaced by X = N. However, the  $\phi 2$  group in the template is absent in most supporting compounds of this category.

D2Ag-template2 is most representative of D2Ag-B, with a characteristic -O-C-C- link between  $\phi 1$  and the amine group. The NH in the fused ring of  $\phi 1$  plays the role of -XR in template1. Compounds with 1-3 atoms between the amine and  $\phi 2$  are prevalent.

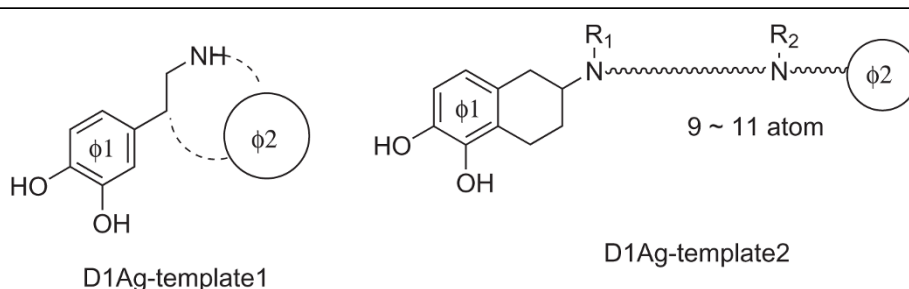
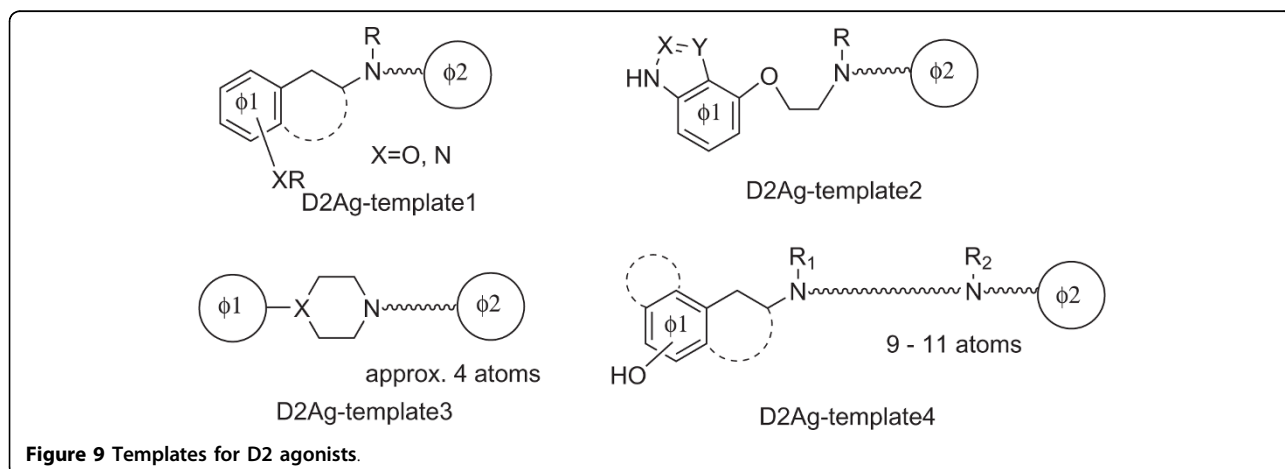
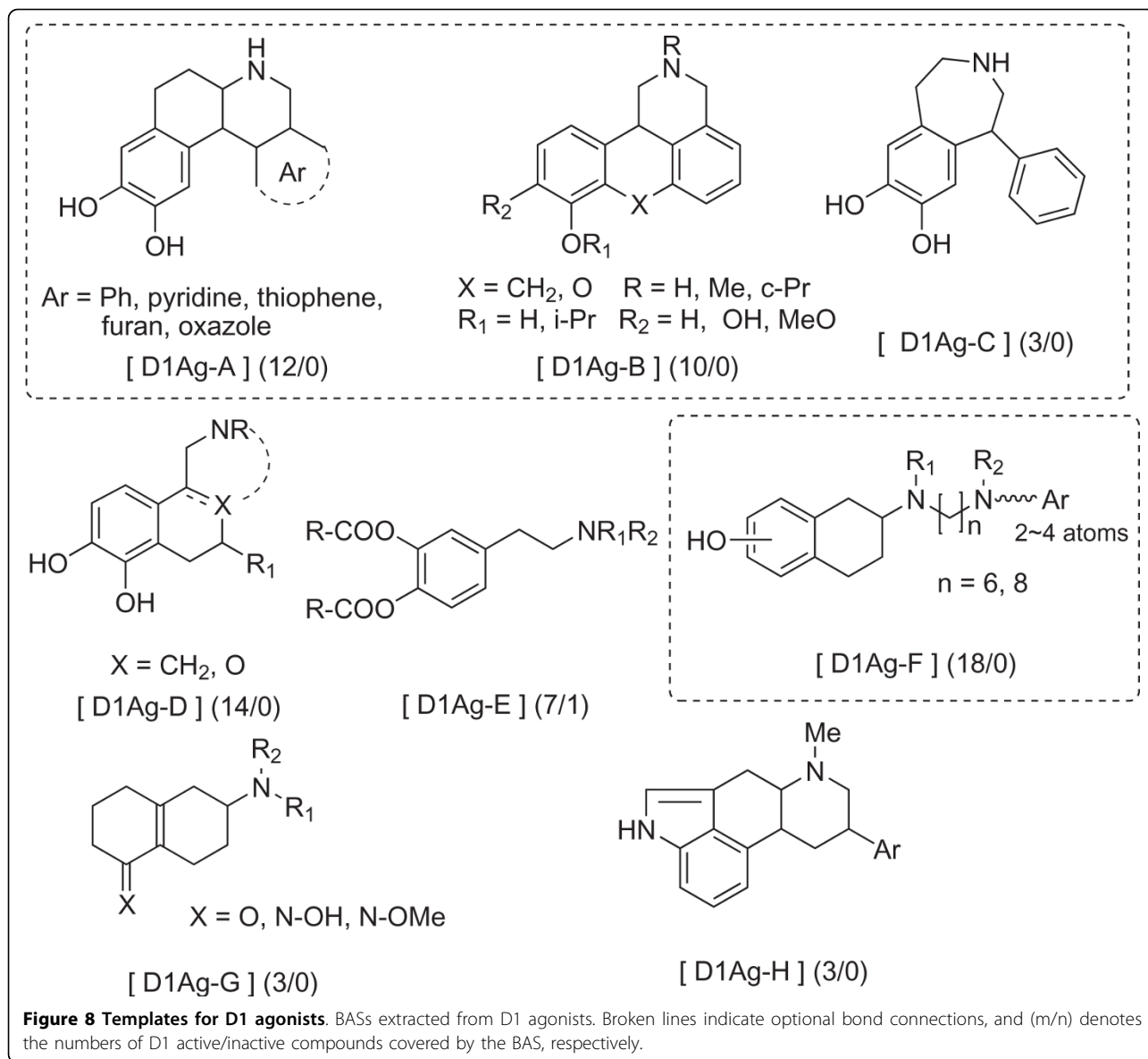
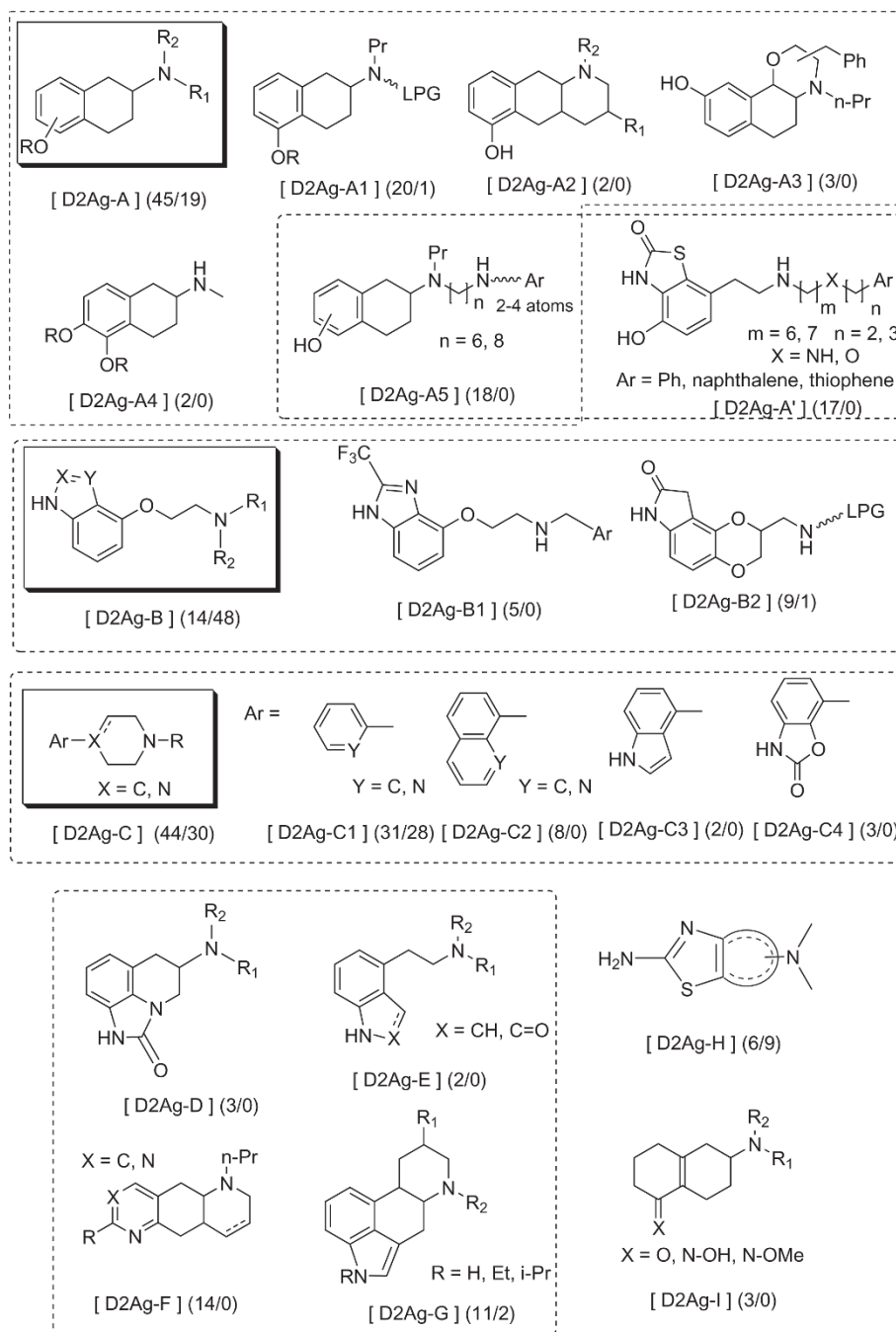


Figure 7 Templates for D1 agonists.







**Figure 10** BASs extracted from D2 agonists. LPG refers to a lipophilic group.

D2Ag-templete3 is representative of D2Ag-C, with a large diversity of  $\phi_1$  moieties, as shown by subBASs C1 through C4. The relative positions of the amine group and the benzene ring in  $\phi_1$  are similar to those in templete1. The -XR group in templete1 corresponds to the NH of the 5-membered rings of C3 and C4, but has no counterpart in C1 or C2.  $\phi_2$  is connected to the amine via a long flexible chain (approximately 4 atoms in

length), and it usually has a fused heterocyclic ring or substituents with hydrogen-bonding capability. Nearly half of the C1-supporting compounds are Dauto agonists. The  $\phi_2$  group in many D2 agonists of this category has a N-hetero ring connected to a flexible chain at the N atom. If the  $\phi_1$ -O-C-C-N chain in templete2 adopts the same conformation as that in templete3, then these three templates exhibit approximately the

same distance between the center of the  $\phi 1$  benzene ring and the amine N atom.

The last template, D2Ag-template4, is representative of BASs, D2Ag-A5 and D2Ag-A'. The former is identical to D1Ag-F, and its supporting compounds show activity with both D1 and D2. The  $\phi 1$  component of these BASs is the same as that in template1, with  $\phi 2$  connected to it via a long flexible chain that incorporates another amine en route. Most of these compounds have a benzene ring in the  $\phi 2$  component.

The amine groups at the active center in most D2 agonists are tertiary, but secondary amines prevail when  $\phi 1$  is fused to a hetero 5-membered ring as shown in D2Ag-A' and B. At present, no adequate templates exist that describe the remaining BASs: D2Ag-H with a 2-aminothiazole ring, and D2Ag-I (identical to D1Ag-G) without the aromatic ring. Four compounds are not covered by these BASs (see Additional file 3).

#### BASs and their roles in Dauto agonists

There were 194 compounds exhibiting Dauto agonist activity. Evolutionarily, the Dauto receptor is closely related to D2, and hence some of the five templates proposed in Figure 11 resemble those of D2. Close inspection reveals that DAuAg-template1 has a similar structure to D2Ag-template1 and 2, and that DAuAg-template3 is almost identical to D2Ag-template3. Figure 12 shows the BASs extracted from Dauto agonists, some of which are categorized as subBASs. All but seven of the compounds with Dauto agonist activity contain one or more of these BASs.

DAuAg-template1 can be associated with DAuAg-A and -B. The  $\phi 2$  component typically consists of a  $-\text{CH}_2-\text{Ar}$  unit in DAuAg-A, and a  $-(\text{CH}_2)_3-\text{Ar}$  or  $-(\text{CH}_2)_3-\text{O}-\text{Ar}$  unit in DAuAg-B. The large differences in linker lengths suggest different docking modes. Four subBASs of DAuAg-A contain a hydrophilic group in  $\phi 1$ . D2Ag-B, which is structurally similar to DAuAg-A, also

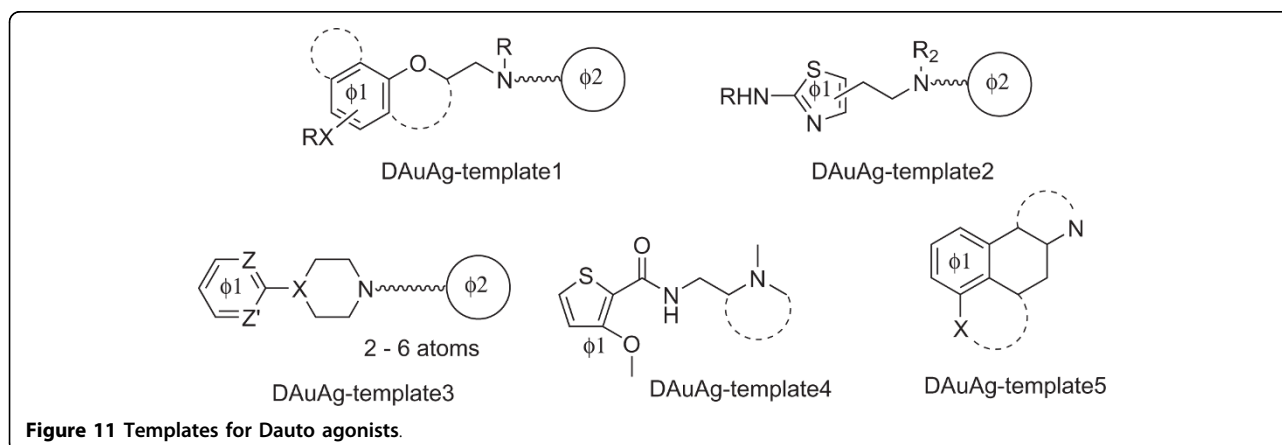
contains a hydrophilic group, but no corresponding identical  $\phi 1$  structure is present in either of these BASs. Another interesting observation is that both DAuAg-A and B are specific to Dauto activity, although the superposition of these structures results in D2Ag-B2, which exhibits activity specific to D2.

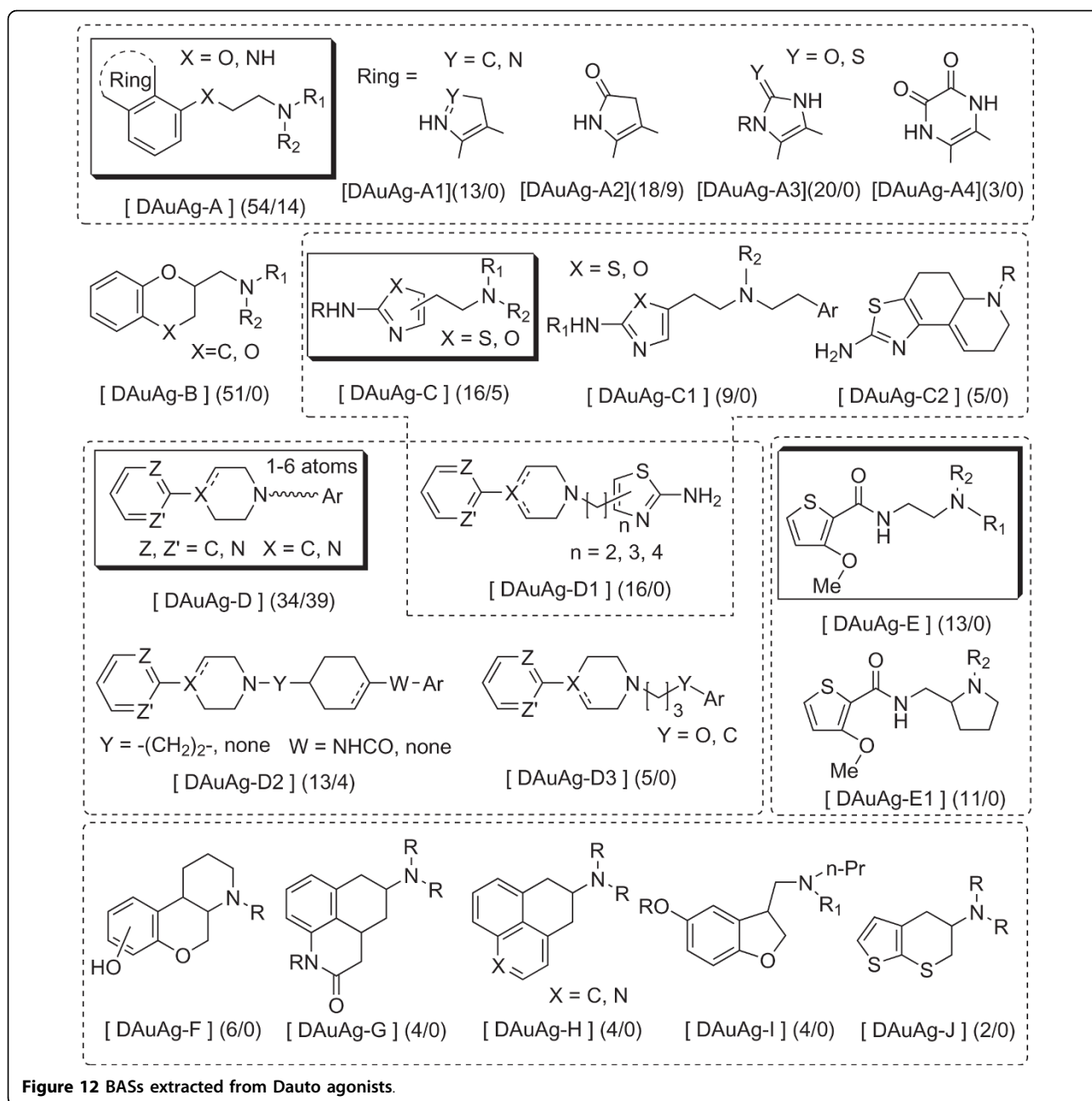
The DAuAg-template2 is representative of DAuAg-C, of which most associated compounds can be described by subBASs C1 and C2. The  $\phi 1$  component of this template consists of a 2-aminothiazole ring. This functional group is also present in D2Ag-H, but the number of associated compounds was not sufficient to warrant a discussion of activity differences. The BAS of DAuAg-C2 lacks a  $\phi 2$  component.

DAuAg-template3 is representative of DAuAg-D. The  $\phi 2$  component of subBAS DAuAg-D1 is the 2-aminothiazole ring that also appears in DAuAg-template2. DAuAg-template3 is structurally similar to D2Ag-template3, representing D2Ag-C. There are two likely reasons for the observed activity differences: (i) the  $\phi 1$  of D2Ag-C compounds often consists of a bicyclic system, but is only present as such in one DAuAg-D compound, and (ii) the distance to  $\phi 2$  is usually longer in DAuAg than that in D2.

DAuAg-template4 is representative of DAuAg-E, and most of the supporting compounds are expressed by subBAS E1. This structure contains a unique thiophene ring in  $\phi 1$  and a hydrophilic amido group between  $\phi 1$  and the amine. The formation of a hydrogen bond between the oxygen atom in the methoxy group and the NH group of the amide creates a pseudo 6-membered ring that may act in place of  $\phi 1$  when combined with the thiophene ring.

The last template, DAuAg-template5, is associated with DAuAg-F, -G, and -H, all of which consist of a rigid tricyclic system without a  $\phi 2$  component. Other BASs, such as DAuAg-I and -J do not contain a tricyclic  $\phi 1$  and the amine of DAuAg-I is not in a rigid





**Figure 12** BASs extracted from Dauto agonists.

environment. Their supporting compounds, however, lack a  $\phi$ 2 component and can therefore be categorized under this template. The amine group at the active center of the BAS is secondary in most compounds categorized under template1 (with Ar-O-C-C-N), whereas it appears as a tertiary amine in the other templates, with the exception of a single compound. Seven compounds are not covered by the BASs (see Additional file 3).

## Conclusions

BASs were successfully extracted from chemical compounds exhibiting a diverse array of structural features.

The key factor has been the incorporation of information from the chemical environment surrounding the BAS in addition to the empirical deduction of medicinal chemists. That is, the discrimination of D1 agonists from D2 and Dauto agonists can be achieved with the catechol structure alone, and the refinement to aggregate the phenethylamine group was useful. Designation of a BAS was impossible without the empirical knowledge of chemists.

The BASiC knowledge base now contains BASs for a variety of activities related to dopamine, 5-HT, adrenaline, adenosine, histamine, benzodiazepine, and

endothelin. BASs that exhibit various toxic and adverse effects are also being extracted. Although BASiC does not describe structural templates, the collection of BASs, which includes supporting chemical structures, would greatly aid in the design of new drugs and the prediction of adverse effects.

### List of abbreviations

BAS: Basic Active Structure; SSC: Supporting Structures Chart; BSS: Between-groups Sum of Squares.

**Additional file 1: BAS extraction procedure.** Figures of the flowchart of analysis and SOM display are shown.

**Additional file 2: Rule derivation process by the cascade model.** Detailed data in the rule set selection process is described.

**Additional file 3: Miscellaneous compounds.** The compound structures not covered by BASs are shown for D1, D2 and Dauto agonists.

### Acknowledgements

This work was supported in part by the Ministry of Education, Culture, Sports, Science and Technology, Japan, through a Grant-in-Aid for Scientific Research on Priority Areas from 2001 to 2004 (13131210) and Grants-in-Aid for Scientific Research (A) from 2002 to 2005 (14208032) and 2006 to 2008 (18200010), and through NEDO project "Development of hazard assessment techniques using structure-activity relationship methods" 2007-2010. We thank Y. Nakano, D. Kitajima, and N. Kamiguchi for contributions to programming and helpful conversations.

### Author details

<sup>1</sup>Department of Informatics, School of Science & Technology, Kwansei Gakuin University, 2-1 Gakuen, Sanda, Hyogo 669-1337, Japan. <sup>2</sup>Department of Electrical and Electronic Engineering, Kanazawa Technical College, 2-270 Hisayasu, Kanazawa, Ishikawa 921-8601, Japan.

### Authors' contributions

MY, SM, HH, and TH carried out the chemical structure analysis, NO and SF developed software and carried out computations, and TO planned and participated in all aspects of the work.

### Competing interests

The authors declare that they have no competing interests.

Received: 18 June 2009 Accepted: 23 January 2010

Published: 23 January 2010

### References

1. Russom CL, Bradbury SP, Broderius SJ, Hammermeister DE, Drummond RA: Prediction modes of toxic action from chemical structure: acute toxicity in the fathead minnow. *Environ Toxicol Chem* 1997, **16**(5):948-967.
2. Ohtaka H, Fujita T: Structural modification patterns from agonists to antagonists and their application to drug design; a new serotonin (5-HT3) antagonist series. *Prog Drug Res*. 1993, **41**:313-357.
3. Harper G, Pickett D: Methods for mining HTS data. *Drug Discovery Today* 2006, **11**:694-699.
4. Böcker A, Derksen S, Schmidt S, Teckentrup A, Schneider G: A hierarchical clustering approach for large compound libraries. *J Chem Inf Model* 2005, **45**:807-815.
5. Böcker A, Schneider G, Teckentrup A: NIPALSTREE: a new hierarchical clustering approach for large compound libraries and its application to virtual screening. *J Chem Inf Model* 2006, **46**:2220-2229.
6. Böcker A: Toward an improved clustering of large data sets using maximum common substructures and topological fingerprints. *J Chem Inf Model* 2008, **48**:2097-2107.

7. Harper G, Bravi GS, Pickett SD, Hussain J, Green DVS: The reduced graph descriptor in virtual screening and data-driven clustering of high-throughput screening data. *J Chem Inf Comput Sci* 2004, **44**:2145-2156.
8. Medina-Franco JL, Petit J, Maggiora M: Hierarchical strategy for identifying active chemotype classes in compound databases. *Chem Biol Drug Des* 2006, **67**:395-408.
9. ALTORIS: SARvision Plus. [http://www.altoris.com/Products.html].
10. Okada T, Wipke WT: CLUSMOL: A System for the Conceptual Clustering of Molecules. *Tetrahedron Comput Method* 1989, **2**:249-264.
11. Klopman G: Artificial intelligence approach to structure-activity studies. Computer automated structure evaluation of biological activity of organic molecules. *J Am Chem Soc* 1984, **106**:7315-7321.
12. Klopman G: MULTICASE 1. A Hierarchical Computer Automated Structure Evaluation Program. *Quant Struct-Act Relat* 1992, **11**:176-184.
13. Klopman G, Sedykh A: An MCASE approach to the search of a cure for Parkinson's disease. *BMC Pharmacol* 2002, **2**:8.
14. Okada T: Mining from Chemical Graphs. *Mining Graph Data* Wiley & Sons/Cook DJ, Holder LB 2007, 347-379.
15. King RD, Muggleton SH, Srinivasan A, Sternberg MJ: Structure-Activity Relationships Derived by Machine Learning. *Proc Nat Acad Sci USA* 1996, **93**:438-442.
16. Tsunoyama K, Amini A, Sternberg MJE, Muggleton SH: Scaffold hopping in drug discovery using inductive logic programming. *J Chem Inf Model* 2008, **48**(5):949-957.
17. Kramer S, De Raedt L, Helma C: Molecular Feature Mining in HIV Data. *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining: San Francisco* ACM Press 2001, 136-143[http://delivery.acm.org/10.1145/510000/502533/p136-kramer.pdf?key1=502533&key2=0334025621&coll=GUIDE&dl=GUIDE&CFID=76266136&CFTOKEN=85729598].
18. Karwath A, De Raedt L: SMIREP: Predicting Chemical Activity from SMILES. *J Chem Inf Model* 2006, **46**:2432-2444.
19. Inokuchi A, Washio T, Motoda H: An Apriori-Based Algorithm for Mining Frequent Substructures from Graph Data. *Principles of Data Mining and Knowledge Discovery, PKDD* Springer; BerlinZighed DA, Komorowski J, Zytkow J 2000, **1910**:13-23[http://www.springerlink.com/content/a7f7yk7mh6n5dnvm/].
20. Inokuchi A, Washio T, Motoda H: Complete Mining of Frequent Patterns from Graphs: Mining Graph Data. *Mach Learn* 2003, **50**:321-354.
21. Inokuchi A, Washio T, Motoda H: A General Framework for Mining Frequent Subgraphs from Labeled Graphs. *Fundamenta Informaticae* 2005, **66**(1-2):53-82.
22. BASiC:[http://www.dm-lab.ws/BASIC/].
23. Symyx: MDDR.[http://www.symyx.com/products/databases/bioactivity/mddr/index.jsp].
24. Okada T: Discovery of Structure Activity Relationships using the Cascade Model: The Mutagenicity of Aromatic Nitro Compounds. *J Comput Aided Chem* 2002, **2**:79-86.
25. Okada T: A Correlation-Based Approach to Attribute Selection in Chemical Graph Mining. *New Frontiers in Artificial Intelligence* Springer; BerlinSakurai A, Hasida K, Nitta K 2007, **3609**:517-526.
26. Okada T: Rule Induction in Cascade Model based on Sum of Squares Decomposition. *Principles of Data Mining and Knowledge Discovery, PKDD'99* Springer; BerlinDjamel A, Zighed J, Komorowski J, Zytkow M 1999, **1704**:468-474.
27. Okada T: Efficient Detection of Local Interactions in the Cascade Model. *Knowledge Discovery and Data Mining, Current Issues and New Applications, PAKDD-2000* Springer; BerlinTerano T, Liu H, Chen ALP 2000, **1805**:193-203.
28. Gini CW: *Studi Economico-Giuridici della R. Università de Cagliari* 1912. This article is reviewed in Light RJ, Margolin BH: An analysis of variance for categorical data. *J Amer Stat Assoc* 1971, **66**:534-544.
29. Fujishima S, Takahashi Y, Okada T: Pharmacophore Refinement in the Chemical Structure Space. *J Comp Chem Jpn* 2008, **7**:63-70.

doi:10.1186/1752-153X-4-1

**Cite this article as:** Okada et al.: The development of a knowledge base for basic active structures: an example case of dopamine agonists. *Chemistry Central Journal* 2010 **4**:1.