COMPOUND ANALYSIS VIA GRAPH KERNELS INCORPORATING CHIRALITY

J.B. BROWN, TAKASHI URATA, TAKEYUKI TAMURA,
MIDORI A ARAI, TAKEO KAWABATA, TATSUYA AKUTSU

Institute for Chemical Research, Kyoto University
Gokasho, Uji, Kyoto 611-0011, Japan
{jbbrown, urata, tamura}@sunflower.kuicr.kyoto-u.ac.jp
kawabata@scl.kyoto-u.ac.jp, takutsu@kuicr.kyoto-u.ac.jp

Natural Products Chemistry,
Graduate School of Pharmaceutical Sciences, Chiba University
1-33 Yayoi-cho, Inage-ku, Chiba 263-8522, Japan
marai@p.chiba-u.ac.jp

Abstract

High accuracy is paramount when predicting biochemical characteristics using Quantitative Structural-Property Relationships (QSPRs). Although existing graph-theoretic kernel methods combined with machine learning techniques are efficient for QSPR model construction, they cannot distinguish topologically identical chiral compounds which often exhibit different biological characteristics. In this paper, we propose a new method that extends the recently developed tree pattern graph kernel to accommodate stereoisomers. We show that Support Vector Regression (SVR) with a chiral graph kernel is useful for target property prediction by demonstrating its application to a set of human vitamin D receptor ligands currently under consideration for their potential anti-cancer effects.

Keywords: Kernel method; graph kernel; QSAR; QSPR; support vector machine.

1 Introduction

Drug design is one of the main practical and industrial targets of bioinformatics and chemoinformatics. When searching for lead compounds in the development of new pharmaceuticals, researchers must often select a small subset of compounds from a vastly larger set that satisfies design requirements. After compound selection, high-throughput screening is a method for the synthesis and evaluation of compounds, though it requires considerable time and cost. Hence, it is advantageous to reduce screening to only those candidates which have been filtered by computational prediction.

A common approach to property prediction is to quantitatively analyze the structural features of a compound and find a connection between the target property and features analyzed, i.e., to derive

$$\text{target property} = f(\text{structural features}).$$

This methodology is known as a Quantitative Structure-Activity/Property Relationship (QSAR or QSPR). Spatial QSPRs such as Comparative Molecular Field Analysis[1] (CoMFA) and 4D-QSAR[2] can produce predictions with high accuracy, but at the expense of a large increase in

computational time. In contrast, two-dimensional (2D) analysis techniques such as graph kernels[3] or topological descriptors[4] offer QSPRs that generally are easier to implement and are faster, while still maintaining good accuracy.[4]

A great number of compounds feature "handedness", which emerges when topologically identical molecules contain carbon atoms bound to four different substituents. The carbon atoms are referred to as stereocenters. In the most simple case where a structure contains a single stereocenter, two *chiral* forms of the molecule are possible. Much like human hands, the chiral forms are mirror images of each other but cannot be perfectly superimposed on each other because they are different molecules. Chemists refer to the "right-handed" version of the pair as the R-form and the "left-handed" version as the S-form, and below we will give an algorithm for discriminating the handedness of a compound (Section 2.5). An example where basic two-dimensional topology alone cannot separate the R/S forms of a chiral molecule is simple bromochlorofluoromethane (CHBrClF).

Separation of chiral molecules is important because it is well known that enantiomers can have completely different biological responses. The S enantiomer of ibuprofen is an active analgesic, while the R enantiomer is inactive.[5] The analgesic $(+)$-$3S,4R$-stereoisomer of picenadol has the opposite effect of the $(-)$-$3R,4S$-stereoisomer which is an antagonist.[6] Without correct stereochemistry, cyclic urea HIV protease inhibitors are ineffective.[7]

Improved planar analysis must specify not only the connectivity relationships amongst atoms (topology or configuration) in their 2D structure, but also spatial relationships (conformation). This paper addresses this issue. We formulate a new graph kernel method that includes additional stereo configuration information in topology analysis. Though it is an extension of a previous kernel development, the proposed kernel actually constrains the amount of computation to perform. Classification and regression experiments show that the additional considerations for stereochemistry produce an improvement, suggesting a new alternative for drug screening and design. The work herein is an important step in the advancement of kernel methods for QSPR research.

2 Methods

In Sections 2.1 and 2.2 we describe tree patterns and graph kernels, and thereafter in remaining sections we explain the stereoisomer extensions.

2.1 Tree-patterns

Let us assume that we are given a graph $G = (V_G, E_G, label_G)$ and a tree $t = (V_t, E_t, label_t)$, where $label_G$ is the mapping of labels to each vertex V_G and edge E_G of the graph, and $label_t$ is analogous for the tree.[1] For chemical compounds, $V_{(\cdot)}$ and $E_{(\cdot)}$ correspond to atoms and inter-atomic bonds, respectively. Let t's size be $|t|$, and $(n_1, n_2, ..., n_{|t|})$ be an enumeration of its vertices. If there is a sequence $(v_1, v_2, ..., v_{|t|})$ of vertices in G that satisfy the requirements below, then the sequence $(v_1, v_2, ..., v_{|t|})$ is a *tree-pattern* of the graph G with respect to tree t:

$$\begin{cases} \forall i \in [1, |t|], & label_G(v_i) = label_t(n_i) \\ \forall (n_i, n_j) \in E_t, & (v_i, v_j) \in E_G \quad \wedge \quad label_G(v_i, v_j) = label_t(n_i, n_j) \\ \forall (n_i, n_j), (n_i, n_k) \in E_t, & j \neq k \iff v_j \neq v_k \end{cases} . \tag{1}$$

In other words, this definition means that a structural pattern (tree) t corresponds to an identical pattern in the graph G, including identical labels. The third condition in (1) enforces that sibling nodes in t must correspond to different vertices in G.

[1] $label(\cdot)$'s behavior is clear from input context.

2.2 Tree-pattern graph kernels

We represent the collection of trees to use as patterns, the tree space, by $T = \{t_1, t_2, \ldots\}$, let weight $w(t)$ be a weighting for each tree t, and define the function $\psi_t(G)$ which counts the frequency of tree-pattern t in graph G. The Tree-Pattern Graph Kernel[3], is then defined as:

$$K(G_1, G_2) = \sum_{t \in T} w(t)\psi_t(G_1)\psi_t(G_2) \quad . \tag{2}$$

If $|T| \neq \infty$, theoretically we can perform explicit calculation of the Tree-Pattern Graph Kernel by creating a feature vector

$$\phi(G) = \left(\sqrt{w(t_1)}\psi_{t_1}(G), \sqrt{w(t_2)}\psi_{t_2}(G), \ldots, \sqrt{w(t_{|T|})}\psi_{t_{|T|}}(G) \right) \quad , \tag{3}$$

whose inner product between two graphs gives us $K(G_1, G_2) = \langle \phi(G_1), \phi(G_2) \rangle$, and hence K is a valid kernel.

Mahé and Vert have defined several types of tree spaces, with appropriate weightings described here.[3] The balanced tree-pattern kernels require that all leaves in the tree space have the same depth, where the depth of a node is defined as the number of edges connecting it to the root plus one.[2] Then, for determining weight $w(t) = \lambda^\mu$ in Eq. (2), the size-based kernel takes the general size of the tree and its depth[3] into account, $\mu = |t| - \text{depth}(t)$, whereas the branching-based kernel considers only the number of leaf nodes[4], $\mu = \text{branch}(t) = |\text{ leaves}(t)\ | - 1$. The Until-N Branching extension removes the restriction that all trees be balanced. Kernels (2) are calculated efficiently by dynamic programming rather than by using feature vectors (3).

2.3 Adding support for chirality

Based on convolution kernels, the graph kernels are defined by incorporating a more fundamental kernel (similarity) $K_S(s_1, s_2)$ between substructures $s_1, s_2 \in S$ existing inside of graphs. Removing coefficients and constants, essentially $K(G_1, G_2) = \sum_{s_1, s_2 \in G_1, G_2} K_S(s_1, s_2)$.

Our proposed kernel is also achieved via convolution, by defining a more fundamental tree kernel which additionally includes spatial information in parent-child relationships. The previous kernel for tree-patterns [3] was defined such that if trees $t_1, t_2 \in T$ have matching substructures, then $K_T(t_1, t_2) > 0$; otherwise, $K_T(t_1, t_2) = 0$. At the most basic level, t_1 and t_2 are simply two leaves (atoms); with identical labels, $K_T(t_1, t_2) > 0$, but with different labels, $K_T(t_1, t_2) = 0$.

Here, we set $K_T(t_1, t_2) > 0$ only when t_1 and t_2 have matching substructures *with identical stereo bonding information*; otherwise 0 is analogously assigned. Figure 1 shows an example of a stereocenter with four different substituents a, b, c, and d. On the left of the chiral pair (center) is the existing tree-pattern graph kernel computation, where two tree patterns t_1 and t_2 are given, resulting in the same feature vector $(\sqrt{w(t_1)} \times 1, \sqrt{w(t_2)} \times 1)$ for the enantiomer pair in the center of the figure, which prevents a learning method from distinguishing the two compounds. The right side of Figure 1 shows how we have included stereo information by expanding the tree space. The extension produces two different feature vectors for the pair of enantiomers, hence allowing them to be differentiated. The application of the extension to a dataset with many stereoisomers is demonstrated in Figure 2.

2.4 Simultaneous enantiomerism and *cis-trans* isomerism

If we were interested in distinguishing compounds containing only *cis-trans* isomers, then it would be sufficient to simply add E/Z labels to the chemical structures in question. For double bonds, the E notation indicates that the higher ordered substituents of each carbon atom are on opposite sides (thus forming a diagonal between them and the double bond), and the Z notation indicates

[2] $\text{depth}(\text{root}(t)) = 1$

[3] $depth(t)$ is the maximum number of edges from the root of t to one of its leaf nodes plus one.

[4] $leaves(t)$ is the set of leaves in tree t.

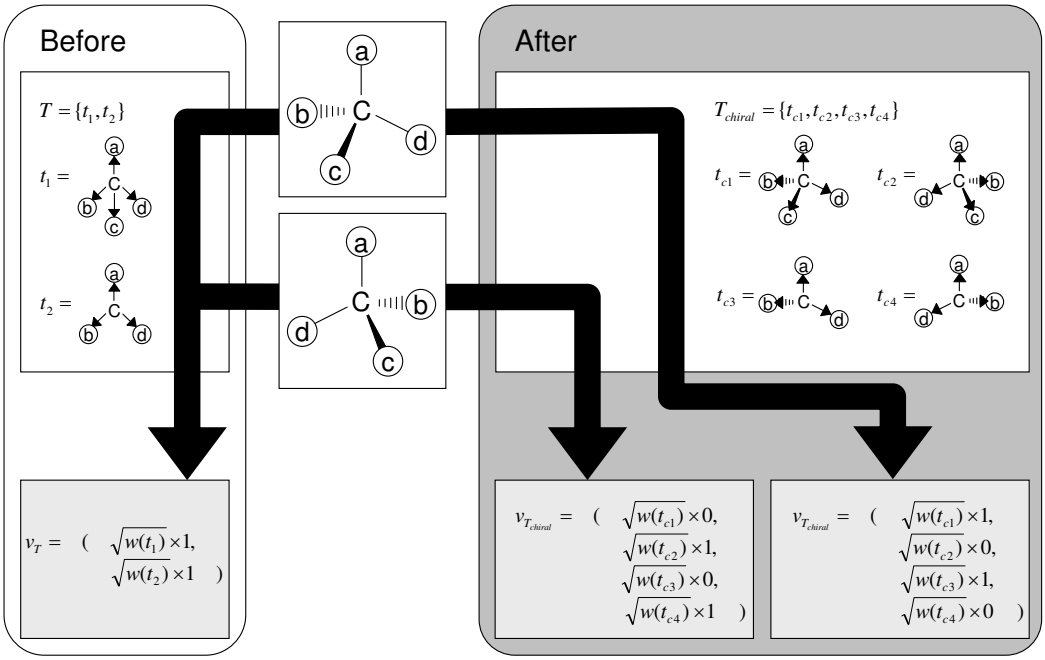

Figure 1: Difference in feature vectors by accounting for chirality. Atoms correspond to labeled graph nodes, and a bond corresponds to two oppositely directed graph edges with identical labels indicating the bond order. $w(t_i)$ is the weighting given to a specific tree pattern, as defined in Section 2.2.

that the substituents are on the same side (thus forming an arc shape). However, substituents of an alkene may be stereocenters (e.g., Cl(OH)CHCH=CHCH(OH)Cl) or additional alkenes (e.g., Hex-2,4-diene), forcing the need to also consider *cis-trans* isomers in a systematic way. To handle this possibility, we process *cis-trans* isomers with the same concept as that for processing of chiral stereocenters. Our approach is to add planar configuration information to the area around the double bond. As a result, $K_T(t_1, t_2) > 0$ for tree patterns t_1, t_2 that include *cis-trans* isomerism only when both t_1 and t_2 have matching substructures including matching planar configuration. Details for *cis-trans* isomerism can be found in the references.[8] [5]

2.5 Computational algorithm including chirality

Kernel function values are computed by a dynamic programming ("DP") algorithm that accomplishes the calculation of equation (2). The details of the DP are already published[3]; therefore, we limit this section to the extensions to accommodate stereoisomers.

For each vertex pair ($u \in G_1, v \in G_2$), the neighborhood matching set (NMS)[3] $M(u, v)$ lists all of the label-matching parent-child relationships that are possible from (u, v).[6] Denote the set of children of vertex u by $\delta^+(u)$, and let $label(\cdot)$ be defined as before.[7] The non-chiral NMS is:

$$\begin{aligned} M(u,v) \quad = \quad & \big\{ R \subseteq \delta^+(u) \times \delta^+(v) \mid R \neq \emptyset \\ & \wedge \ \forall (a,b),(c,d) \in R \ : \ a = c \Leftrightarrow b = d \\ & \wedge \ \ [\forall (a,b) \in R \ : \ label(a) = label(b) \wedge label(u,a) = label(v,b)] \ \big\} \end{aligned} .$$

[5] Ref. 8 is a Ph.D. dissertation, and not a peer-reviewed journal paper.

[6] G_1 and G_2 are general graphs, but since our pattern space is limited to tree structures, we can define a local parent-child relationship.

[7] We abbreviate $label_G(u)$ to $label(u)$ to simplify notation.

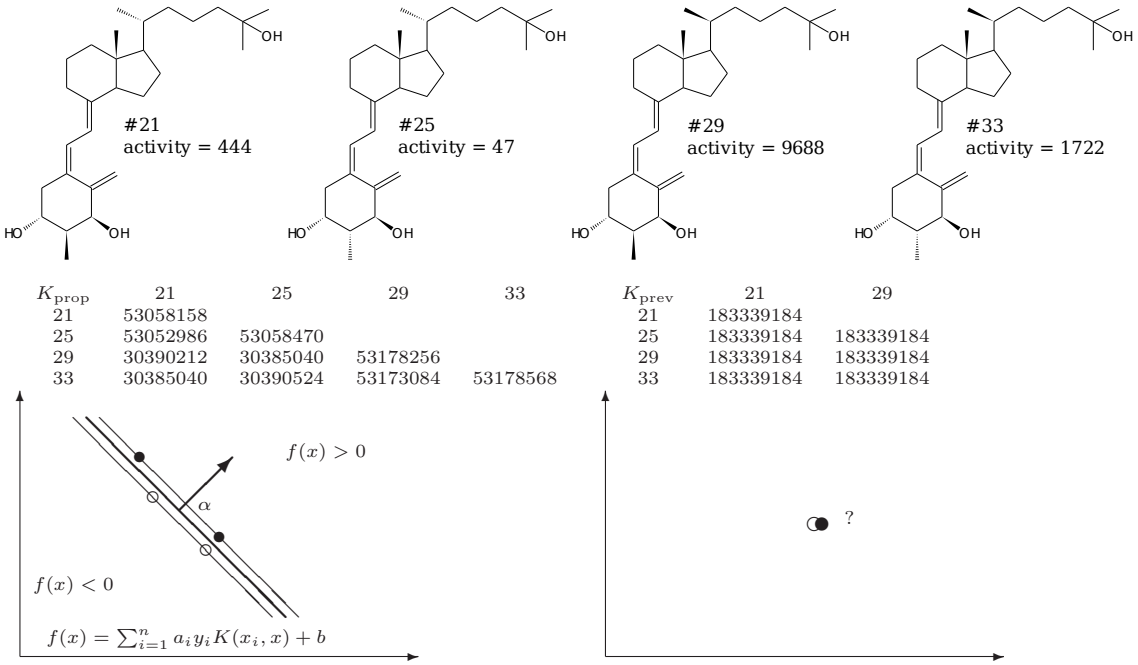

Figure 2: The concepts of kernel methods and support vector machines (SVMs), as applied to the human vitamin D receptor ligand dataset used in this paper. Four ligands are used to generate a similarity matrix. The existing graph kernel method K_{prev} produces the same similarity value for any two ligands (kernel matrix abbreviated). Our proposed method, K_{prop}, solves this problem by using stereo information when calculating similarity. (Bottom left) A support vector machine can use our kernel matrix to derive a hyperplane for separating the data into two classes, for example to isolate stereoisomers 29 and 33 which have relative activity > 1000. (Bottom right) Because of identical kernel values, a hyperplane cannot be derived from K_{prop} to separate the hVDR ligands into two classes. The kernel matrices were generated by a size-based tree kernel with pattern tree depth set to 4 and λ set to 1 (parameter details in Graph Kernel Definition), and thus each kernel value corresponds to the number of matching substructures in a pair of hVDR ligands.

R denotes a set of matching pairs of descendants of u and v.[8] For example, let $\delta^+(u) = \{a, c\}$, $\delta^+(v) = \{b, d\}$, and assume that all atoms are carbons that are singly bonded. Then, $M(u, v) = \Big\{\{(a, b)\}, \{(a, d)\}$,{(c,b)}, {(c,d)}, {(a,b),(c,d)}, $\{(a, d), (c, b)\}\Big\}$, and R refers to any one of the singleton pair sets or 2-pair sets.

We introduce stereo information into the formulation. Let v be a carbon atom singly-bonded to v_1, v_2, v_3, and v_4. Define $CH(v) = 1$ when stereo bond information is present in one of v's bonds and $CH(v) = 0$ otherwise. Also, define $chiral(v_1, v_2, v_3, v_4) = 1$ if v_2, v_3, and v_4 are arranged in clockwise order when looking from child v_1 to parent v (R-configuration). Assign $chiral(v_1, v_2, v_3, v_4) = -1$ when the arrangement is counter-clockwise (S-configuration).[9] Then, when $CH(u) = CH(v) = 1$ and $|\delta^+(u)| = |\delta^+(v)| = 3$,[10] both parent atoms u, v contain stereo bonding information each with four substituents - they possibly are stereocenters. The NMS is constrained as follows:

$$
\begin{aligned}
M(u,v) = \Big\{ & R \subseteq \delta^+(u) \times \delta^+(v) \mid R \neq \emptyset \\
& \wedge \forall (a,b),(c,d) \in R \ : \ a = c \Leftrightarrow b = d \\
& \wedge \Big[\ \forall (a,b) \in R \ : \ label(a) = label(b) \ \wedge \ label(u,a) = label(v,b) \ \Big] \\
& \wedge \Big[\ |R| \neq 3 \vee CH(u) = 0 \vee CH(v) = 0 \ \vee \\
& \quad (\ chiral(u_0, a, c, e) = chiral(v_0, b, d, f) \ ) \ \Big] \ \Big\}
\end{aligned}
\tag{4}
$$

where u_0 and v_0 correspond to parent vertices of u and v, respectively.

If the NMS is created from R using our extended definition, then the DP algorithm[3] can be applied in its original design, and the chiral constraint reduces the size of $M(u, v)$ to lower computation time.

The neighborhood matching set for the *cis-trans* extension is slightly more complex, with details in ref. 8. Finally, to incorporate both the chiral and *cis-trans* formulations simultaneously, graphs are extended to $G_1 = (V_{G_1}, E_{G_1}, label_{G_1}, CH_{G_1}, CT_{G_1})$ and $G_2 = (V_{G_2}, E_{G_2}, label_{G_2}, CH_{G_2}, CT_{G_2})$, with a NMS that incorporates Eq. (4).

3 Experimental Setup and Results

3.1 Datasets

Three datasets are used to evaluate the proposed kernel methods.

The ecdysteroid dataset is a collection of 20-hydroxyecdysone agonists that are involved in the control of ecdysis (shedding) and metamorphosis in arthropods. The EC_{50} value, that is, the effective concentration necessary of an ecdysteroid to bind to the ecdysteroid receptor and trigger the biological response 50% of the time, is expressed as a numerical value. Dinan *et al.*[9] and Hormann *et al.*[10] have provided the EC_{50} values of 108 ecdysis hormones. There are a total of 11 stereoisomer groups, where a group contains two or more stereoisomers. Using the negative decadic logarithm, the dataset average value is 6.40, with a standard deviation of 1.34.

Cramer's steroids is a dataset of 31 steroids where EC_{50} values represent concentrations required for a steroid to bind to corticosteroid binding globulin (CBG).[1] There are two small stereoisomer groups. The dataset average value is 6.15, with a standard deviation of 1.17.

The final dataset used is synthetic vitamin D derivatives. 1α,25-dihydroxyvitamin D_3 [1α,25$(OH)_2D_3$] is an endogenous cellular ligand of the human vitamin D receptor (hVDR) in the nucleus. Among the many important roles of 1α,25$(OH)_2D_3$ are HL-60 (human promyelocytic leukemia cell) differentiation induction, and antiproliferation or apoptosis induction in malignant cancer cells. Many

[8] It is common to misinterpret R as containing only 2-tuples of direct children of u, v. Up to n-tuples are possible, where $n = \min(\delta^+(u), \delta^+(v))$.

[9] If $chiral(v_1, v_2, v_3, v_4) = 1$, then $chiral(v_1, v_2, v_4, v_3) = -1$ and $chiral(v_1, v_3, v_4, v_2) = 1$.

[10] u is a child of a parent in a tree structure, which is why $|\delta^+(u)| \neq 4$.

derivatives of 1α,25$(OH)_2D_3$ with different stereochemistry and/or substituents have been synthesized and evaluated for their biological activities. The majority of this dataset can be found in a textbook.[11] To the 56 unique structures with differentiation data available in the textbook, we have added 13 other published stereochemical modifications to the 1α,25$(OH)_2D_3$ structure.[12, 13] There are 18 stereoisomer groups (see ref. 8 for group details). This dataset is highly valuable for QSPR research because it provides an abundant number of stereoisomers with large differences in activity levels. The dataset is available upon request from the authors.

3.2 Graph kernel and support vector learning parameters

Here, we group the various parameters that affect the system performance of chiral tree pattern graph kernels.

1. Tree-pattern kernel types : { Size-based, Branching-based, Until-N Branching-based }
   A description of these different tree types is in the "Tree-Pattern Graph Kernels" section (2.2).

2. Kernel extensions : {CH, CT}
   The kernel option which adds the chiral requirement to the neighborhood matching set is labeled CH, and similar accounting for *cis-trans* isomerism is an option we label CT.

3. Depth of tree : $h \in \{3, 4, 5, 6\}$
   h specifies the maximum depth of subtrees constituting the tree pattern space. As a simple example, we may consider methanol as an unbalanced tree of depth $h = 3$, where the carbon atom is the root of the tree, and the hydroxyl group extends the depth of the tree from 2 to 3. Our range is chosen by considering results in the literature [3].

4. Weight factor : $\lambda \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$
   This parameter has a large influence on performance. For each tree t, its weight $w(t)$ in (2) is calculated as $w(t) = \lambda^{\mu}$, where λ is a free parameter such that larger values emphasize similarity of complex substructures and de-emphasize linear chains of atoms. Calculation of μ is done as explained in Section 2.2. This range of values tested is also influenced by previous results [3].

5. Normalization : { on, off } Before input to SVM/SVR, we can normalize all of the kernel matrix values to the unit ball: $K_{norm}(a, b) = \frac{K(a,b)}{\sqrt{K(a,a)K(b,b)}}$.

6. RBF kernel parameter : $\gamma \in \{\text{off}, 0.01, 0.05, 0.1, 0.5, 1, 2\}$
   The RBF kernel measures the similarity of two feature vectors $\mathbf{a}$ and $\mathbf{b}$ using the function $K_{RBF}(\mathbf{a}, \mathbf{b}) = exp\Big(-\frac{\|\mathbf{a}-\mathbf{b}\|^2}{2\gamma^2}\Big)$, where the free parameter γ controls the learning balance between possibly over-fitting (low γ) and over-generalizing (high γ). The RBF kernel can effectively describe similarity of points in feature space, and can be highly useful when the decision boundary to be learned is non-linear (an extreme case would be a colored chess board). To apply the RBF kernel to a matrix of real values, we can use the RBF kernel in the rewritten form: $K_{RBF}(a, b) = \exp\Big(-\frac{\|a-b\|^2}{2\gamma^2}\Big) = \exp\Big(-\frac{K(a,a)-2K(a,b)+K(b,b)}{2\gamma^2}\Big)$. It is straightforward to show that $K_{norm}\big(\ K_{RBF}(a, b)\ \big) \neq K_{RBF}\big(\ K_{norm}(a, b)\ \big)$, and as a result, when normalization and RBF kernels are both applied, the normalization must be done before applying the RBF kernel. The parameter range selected is appropriate for dealing with normalized kernels. The "off" option means that we alternatively do not apply the RBF kernel to a kernel matrix as a post-processing step.

7. SVM/SVR trade-off parameter : $C \in \{0.1, 1, 10, 100, 1000, 10000\}$
   The soft-margin formulations of the support vector learning algorithm include a specified maximum tolerance for mistakes during learning in exchange for maximizing the learning margin, which can considerably impact performance.

8. SVR tube width : $\epsilon \in \{0.1, 1, \sigma_{tr}/5, \sigma_{tr}/10\}$
   The tube width is an important setting for preventing overfitting of training data when the variance of target properties is large. Therefore, we expanded on SVM$^{\text{light}}$'s default value of 0.1 because the variance of the hVDR dataset is much larger than this. By using the standard deviation of the target property in the known training data σ_{tr} and scaling it appropriately, we can set a tube width appropriate to the application at hand.

3.2.1 Support vector learning implementations

Since the explicit feature vector representation of our proposed kernel may be in an infinite-dimensional space, it is *necessary* to use a SVM implementation that can take the kernel matrix directly as input. The GIST [14] SVM implementation allows us to do such. For 2-class SVM experiments, we normalize the kernel matrix, after which GIST uses heuristics to set the γ and C parameters. GIST also features tools for automatic calculation of metrics (sensitivity, etc.) to gauge how well the support vector algorithm learned from the training data.

Unfortunately, the GIST software used for SVM experiments does not support SVR, so we use the SVM$^{\text{light}}$ package[15] for SVR experiments. Though SVM$^{\text{light}}$ requires vectorial input, we have built a small extension that enables direct input of the kernel matrices produced by our method. In SVR experiments using SVM$^{\text{light}}$, we test the grid of γ, C, and ϵ values above both with and without normalization.

3.3 Performance metrics

3.3.1 SVM experiments

For each classification experiment, we use leave-one-out cross-validation (LOO-CV) on the training set. We generate ROC curves by using the range of scores output by the SVM algorithm, and calculate each parameter set's ROC curve area (AUC).

3.3.2 SVR experiments

Two measures are used for assessing QSPR prediction performance. The first of these, q^2, is the cross-validated version of the standard residual R^2 for the training set. Let y_i be sample (compound) i's known experimental value (activity level or target property), and let $\hat{y}_i$ be its value output by a predictor during cross-validation. Using the known experimental average value $\bar{y}$, q^2 is calculated as: $q^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$.

The second metric is the correlation R between the predicted and known experimental values for a test dataset after a model has been constructed using the full training dataset. Labeling the average of the predicted values as $\bar{\hat{y}}$, R is defined as: $R = \frac{\sum (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (\hat{y}_i - \bar{\hat{y}})^2}}$.

Good prediction performance is signaled when the values of both metrics are close to 1. A perfect predictor would have a $[q^2, R]$ vector with length $\left|[1,1]\right|_2 = \sqrt{2}$.

3.4 SVM experiments using steroid datasets

For SVM experiments with ecdysteroids, the first 71 compounds are used for training data.[9] Four of the 37 test set ecdysteroids (77, 78, 83, 90) have inexact EC_{50} values reported, and were excluded in a previous analysis[10]; we accordingly exclude them for SVM experiments using the train-test split in ref. 10. For Cramer steroid SVM experiments, compounds 1-21 are for training and compounds 22-31 are for testing.[1] The activity levels of the hVDR ligands cover a large range $[0.1, 9688]$, and dividing this data into two classes based on a single value for SVM experiments is nonsensical. For both steroid datasets, we separate test and training datasets into two classes using EC_{50} thresholds of 6, 6.5, and 7, based on EC_{50} averages.

Detailed tabular results of the different kernel extensions are deferred to ref. 8, where our proposed methods show clear improvements for the ecdysteroid dataset. Calculated averages of

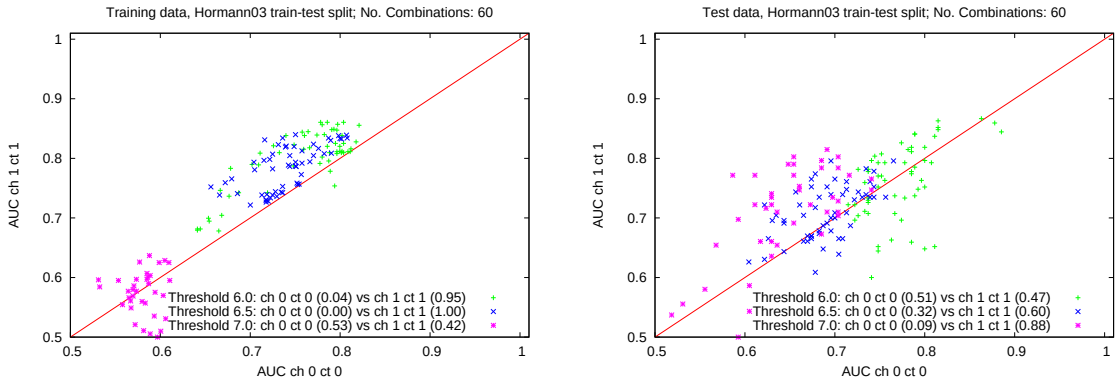

Figure 3: Ecdysteroid classification performance, where each point is the per-parameter set AUC pair of the previous graph kernel (horizontal axis) and our proposed extension (vertical axis). Values in parentheses indicate the percentage of superior performance over all parameter sets for a particular threshold. Left: training data; Right: test data.

optimal tree depth and weight validate our belief that complex subtree patterns are more useful in compound recognition. We plot the ecdysteroid AUC scores for each of $3 * 4 * 5 = 60$ (tree space/depth/weight) parameter sets along non-chiral (horizontal) and chiral-*cis-trans* (vertical) axes in Figure 3. A majority of points located above the even performance line confirms the improvement that chirality contributes to tree pattern graph kernels.

3.5 Results with SVR

3.5.1 SVR experiment datasets

The hVDR dataset has no defined test dataset, and hence no value of R can be calculated. To circumvent this problem, we created five experiment datasets by selecting 30% of the data to serve as a test set. The test set is uniformly drawn at random independently for each experiment. We also created randomized training and test sets of the Cramer steroid and ecdysteroid datasets, to evaluate our proposed methods more completely. Including train-test splits from the literature, we created 18 experiments to perform (6 Cramer[1], 7 ecdysteroid[9, 10], 5 hVDR).

If a particular kernel matrix contains little or no variance, time will be wasted trying to learn from data. We transform each kernel matrix value by dividing it by the matrix mean, and then calculate the matrix variance $\sigma_{\bar{x}}^2$. SVR experiments are aborted if $\sigma_{\bar{x}}^2 < 0.1$ for a parameter set matrix; otherwise the grid of (ϵ, C) pairs are then applied to SVR experiments for each matrix.

To compensate for the large range of hVDR dataset values, an adjustment

$$\text{activity}^{adj} = \begin{cases} \lfloor 100 * \ln(\text{activity}^{orig}) \rfloor & \text{activity}^{orig} \geq 1 \\ 0 & \text{activity}^{orig} < 1 \end{cases} \tag{5}$$

is used to scale down the range of activity values.[11] The bottom case of Eq. (5) is necessary to prevent activity values less than 1 from stretching the scaled range. The result is a reduction in standard deviation from 1773.11 to 236.71, and this standard deviation is useful as a heuristic for the SVR tube width ϵ.

3.5.2 Assessing results using q^2 and R

We compare our proposed extensions to the results published using CoMFA and a 2D topological descriptor extension. We also use the fingerprints from PubChem, which can be extracted for any

[11] activity^{orig} refers to the unscaled hVDR differentiation induction activity value in reference 11.

Table 1: Comparison of the graph kernels. For train-test splits as defined in the literature, references are given. For random train-test splits, the split number is given after the dataset name (e.g., hVDR ligand-1). Format: $v_{opt} = \left|v_{opt}\right|_2$, where $v_{opt} = [q^2, R]$ is as selected by Eq. (6). The top performing method per dataset is highlighted in bold.

Methodology	Cramer steroid[1]	Ecdysteroid[9]	hVDR ligand-1
Reference 4	**[0.830,0.940]=1.254**	**[0.750,0.900]=1.172**	-
Reference 8-B	-	[0.690,0.350]=0.774	-
PubChem SVR	[0.258,0.398]=0.475	[0.055,0.718]=0.720	[0.028,0.443]=0.444
No CH/CT	[0.848,0.621]=1.051	[0.297,0.731]=0.789	[0.019,0.335]=0.336
CH	[0.722,0.707]=1.011	[0.323,0.900]=0.956	[0.721,0.610]=0.944
CH+CT	[0.722,0.707]=1.011	[0.320,0.902]=0.957	**[0.706,0.676]=0.977**
	Cramer steroid-4	Ecdysteroid[10]	hVDR ligand-3
No CH/CT	**[0.822,0.835]=1.172**	[0.311,0.614]=0.688	[0.000,0.223]=0.223
CH	[0.784,0.863]=1.166	[0.378,0.660]=0.761	[0.632,0.805]=1.023
CH+CT	[0.784,0.863]=1.166	**[0.373,0.675]=0.771**	**[0.565,0.868]=1.036**

compound by using the CACTVS software package.[16] QSPRs that are useful for real application should meet or exceed $m_{q^2} = 0.50$ and $m_R = \sqrt{0.60} = 0.774$. The results of the comparison are in Table 1. In the table, "optimal" models $v_{opt} = [q^2, R]$ per extension are given (along with resulting L_2 norm), where optimal is defined by the following criteria[12]:

$$\begin{aligned} v_{opt} \quad &= \underset{v}{\operatorname{argmin}} \|v - w\| \\ \text{s.t.} \quad &\begin{cases} w = [1,1] & \exists\ v \mid v_{q^2} \geq m_{q^2} \text{ or } v_R \geq m_R \\ w = [m_{q^2}, m_R] & \text{otherwise} \end{cases} \end{aligned} \quad . \tag{6}$$

The goal of Eq. (6) is to select the model which represents the best balance between training and test results.

3.5.3 Assessing improvement graphically using q^2 and R

In Figure 4, we visualize the performances of the QSPRs by plotting q^2 values on the horizontal axis and R values on the vertical axis. Vertical and horizontal bars are placed at m_{q^2} and m_R.

In each plot of Figure 4, the number of parameter set models ($q^2 > 0, R > 0$) is listed per extension. On average, approximately 25% (10000/40000) of the parameter set models (using all three extensions) met this requirement, meaning that 75% of models were discarded for the figures. Though this may seem rather high, it is important to remember that the large majority of optimal models are from parameter sets with larger values for the tree depth setting, and most of the parameter sets with smaller values for the tree depth are discarded. For the hVDR dataset, many more models were discarded because of the original graph kernel's formulation that did not consider stereocenters.

The results in Figure 4 show again that chiral graph kernels are an important and useful extension for 2D-QSPR research, especially considering the hVDR ligand data. A graphical example of experiments using PubChem fingerprints is shown in Figure 5, where comparison with Figure 4 demonstrates that the proposed methods outperform by a considerable margin. We emphasize the importance of randomizing the dataset, as without such our methods might appear insufficient for the Cramer steroid reference set (Table 1 and Figure 4). Also, the use of heuristics for setting the value of ϵ is validated, as most of the optimal models in Figure 4 and ref. 8 use one of the two heuristic values.

[12] $\|x - y\|$ is the standard Euclidean distance of x and y.

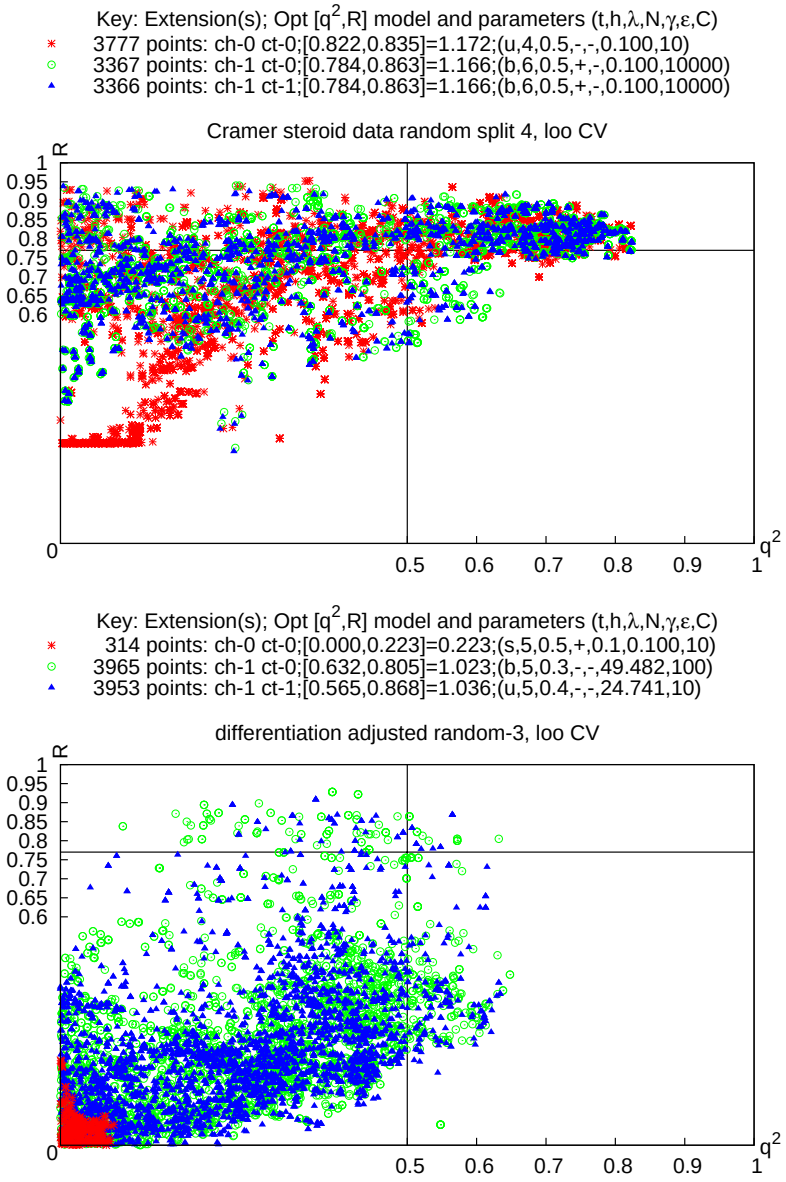

Figure 4: Plots of q^2 and R performance for selected experiments. Each plot includes all possible parameter combinations and is trimmed to the region $q^2 \geq 0$, $R \geq 0$. For each extension, the model selected by criteria (6) is shown along with its L_2-norm and parameter set. "ch/ct-0/1" refers to disabling(0)/enabling(1) of each chiral/*cis-trans* extension. Top : A randomized Cramer steroid dataset; Bottom : A randomized hVDR ligand dataset.

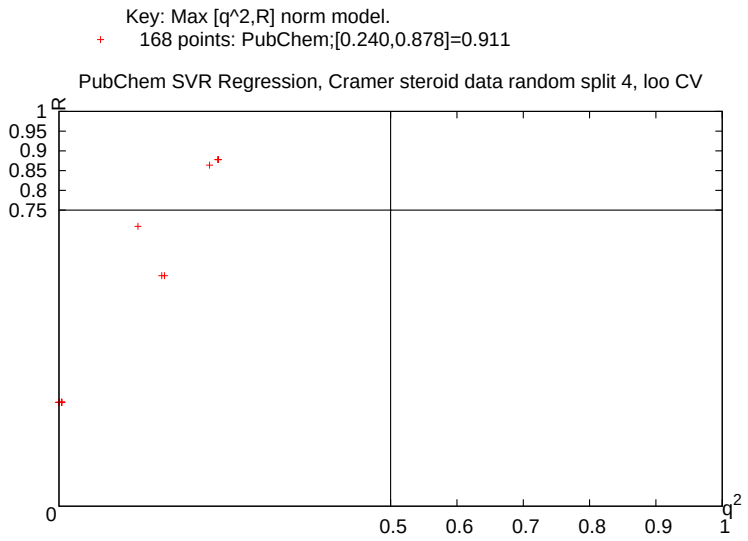

Figure 5: A q^2/R plot similar to Figure 4 using the PubChem fingerprints as the features for QSPR model construction. The graph kernels are considerably more effective than the PubChem fingerprints.

3.5.4 Comparison to other models

As Table 1 shows, the proposed extensions produce an improvement in QSPR accuracy over the existing non-chiral kernel. This is extremely clear for the two hVDR datasets shown in Table 1; the conclusion is the same for other hVDR random train-test splits not shown in the table.

For the Cramer steroid dataset, ref. 4 produces better results for the published training-test split[1], though we found that by shuffling the training and test datasets, we could produce comparable performance. Notice in Table 1 that the shuffled performance compared to the original Cramer steroid train-test split is improved. Though q^2 performance for the Cramer steroid datasets without the proposed extensions was marginally higher than the chiral kernel, this is simply due to a lack of stereoisomer pairs in the dataset. For the ecdysteroid dataset, we could achieve performance close to but unfortunately not quite as good as ref. 4. Like the hVDR dataset, the extensions produce an improvement in ecdysteroid QSPR performance when ample stereoisomers exist. There are a number of points which merit further discussion.

First, the method herein using kernels is easier to understand than the highly complex topological extension in ref. 4. We have included no specific chemical knowledge in our methodology yet we have provided comparable performance using only the information contained in the datasets.

Second, the proposed method is a new alternative available when other QSPR methods are unsuccessful. When tested with cross validation (CV), the new kernel's performance improved as more data became available. For example, in the hVDR 5th random train-test split experiment (not shown in Table 1), the chiral methodology developed increasingly better models of the dataset, as the 2-fold to 5-fold to LOO-CV performance transited $\left|[0.238, 0.673]\right|_2 = 0.714 \rightarrow \left|[0.359, 0.832]\right|_2 = 0.906 \rightarrow \left|[0.515, 0.854]\right|_2 = 0.997$, satisfying m_{q^2} and m_R.[13] The CH+CT extension had similar results for the experiment using the 5th shuffling of ecdysteroid data (results not shown in Table 1). Future datasets with large numbers of enantiomers should exhibit similar performance improvements. This is particularly relevant in situations such as drug design and refinement where more data becomes available over time.

Third, because the proposed method is a kernel method, it is not restricted to use in Sup-

[13]Though the 2-fold results are for a preliminary version of the kernel, results with the current kernel should be almost the same.

port Vector Machines. It can be conveniently inserted into other kernel-based pattern analysis techniques, a considerable merit.[18]

Fourth, the results show that complex tree patterns are important for property prediction. The optimal tree depths both with and without chirality are in the larger range of values tested. Mahé and Vert's non-chiral tree-pattern kernel has already been sufficiently shown to be effective[3], and our results enhance performance.

Fifth, additional assessment using y-randomization[19], where the target properties of the training dataset were randomized by shuffling, concluded that the proposed methodology is learning from the input examples. Abbreviating detailed results, in a number of experiments, no model such that $[q^2 > 0, R > 0]$ could be derived using the full range of parameters tested, even with the stereoisomer extensions. Especially in the case of the hVDR dataset, the greatly boosted performance as a result of the proposed extensions is not attributable to chance correlation.

4 Conclusion

In this paper, we have extended the tree-pattern graph kernel method to account for stereoisomerism, such that stereoisomers with multiple stereocenters and/or multiple *cis-trans* double bonds can be systematically accounted for. If the number of stereocenters was limited to a small number, manual assignment of R/S and E/Z labels would be sufficient, but in reality there are many cases in which a compound contains a non-trivial number of stereocenters. Drug design is very dependent on chirality[6], and being able to accommodate this situation is one of the major advantages of our method. As far as the authors are aware, this is the first method that does such using a portable kernel method. By judiciously selecting chemical knowledge to design further customized kernel functions, we anticipate that we can boost performance beyond what we have achieved thus far.

Acknowledgements

The authors thank Professor Toshio Okano from Kobe Pharmaceutical University, who kindly allowed us to use his 1α,25$(OH)_2D_3$ structure-activity data. This work was partially supported by Grant-in-Aid #19200022 from the Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT). J.B. Brown gratefully acknowledges partial support from MEXT. The authors additionally thank Mr. Yoshiyuki Naitou and Mr. Tobias Schmidt-Goenner for discussions on computer representations of chemicals, Professor Atsushi Kittaka for discussion on the hVDR dataset, and Professor J.P. Vert for discussions on support vector learning.

Availability

A software implementation (binary executable file) of the chiral graph kernel is provided online at *http://sunflower.kuicr.kyoto-u.ac.jp/∼jbbrown*.

References

[1] Cramer RD, Patterson DE, Bunce JD, Comparative Molecular Field Analysis (CoMFA) 1. Effect of shape on binding of steroids to carrier proteins, *J Am Chem Soc* **110**:5959–5967, 1988.

[2] Hopfinger AJ, Reaka A, Venkatarangan P, Duca JS, Wang S, Construction of a virtual high throughput screen by 4D-QSAR analysis: application to a combinatorial library of glucose inhibitors of glycogen phosphorylase b, *J Chem Inf Comput Sci* **39**:1151–1160, 1999.

[3] Mahé P, Vert JP, Graph kernels based on tree patterns for molecules, *Mach Learn* **75**:3–35, 2009.

[4] Golbraikh A, Tropsha A, QSAR modeling using chirality descriptors derived from molecular topology, *J Chem Inf Comput Sci* **43**:144–154, 2003.

[5] McMurry J, Simanek E, *Fundamentals of Organic Chemistry*, 6th edition, Thomson Brooks/Cole: Belmont, California, pp. 192–193, 2007.

[6] Hutt AJ, Drug chirality and its pharmacological consequences, in Smith HJ (ed.), *Introduction to the Principles of Drug Design and Action*, 4th edition, Taylor and Francis Group: Boca Raton, Florida, pp. 117–183, 2006.

[7] Lam PYS et al., Cyclic HIV protease inhibitors: Synthesis, conformational analysis, P2/P2' structure-activity relationship, and molecular recognition of cyclic ureas, *J Med Chem* **39**:3514–3525, 1996.

[8] Brown J, *Kernel Methods in Biochemical Informatics and Applications to DNA Repair Research*, Ph. D. Thesis, Kyoto University, March 2010.

[9] Dinan L, Hormann RE, Fujimoto T, An extensive ecdysteroid CoMFA, *J Comput-Aided Mol Des* **13**:185–207, 1999.

[10] Hormann RE, Dinan L, Whiting P, Superimposition evaluation of ecdysteroid agonist chemotypes through multidimensional QSAR, *J Comput-Aided Mol Des* **17**:135–153, 2003.

[11] Okano T, Kubodera N, Vitamin D inductive form pharmacology and clinical effect, in Akizawa T, Kato S, Nakamura T, Matsumoto T (eds.), *Vitamin D Update 2001*, Chugai Pharma Co, Ltd: Tokyo, Japan, pp. 213–227, 2001. Content in Japanese, titles translated to English by J.B. Brown.

[12] Ono K, Yoshida A, Saito N, Fujishima T, Honzawa S, Suhara Y, Kishimoto S, Sugiura T, Waku K, Takayama H, Kittaka A, Efficient synthesis of 2-modified 1α,25-dihydroxy-19-norvitamin D_3 with julia olefination: High potency in induction of differentiation on HL-60 cells, *J Org Chem* **68**:7407–7415, 2003.

[13] Saito N, Suhara Y, Kurihara M, Fujishima T, Honzawa S, Takayanagi H, Kozono T, Matsumoto M, Ohmori M, Miyata N, Takayama H, Kittaka A, Design and efficient synthesis of 2α-(γ-hydroxyalkoxy)-1α,25-dihydroxyvitamin D_3 analogues, including 2-epi-ED-71 and their 20-epimers with HL-60 cell differentiation activity, *J Org Chem* **69**:7463–7471, 2004.

[14] Pavlidis P, Wapinski I, Noble W S, Support vector machine classification on the web, *Bioinformatics* **20**:586–587, 2004.

[15] Joachims T, Making large-scale SVM learning practical, in *Advances in Kernel Methods - Support Vector Learning*, Scholkopf B and Burges C and Smola A (eds.), 169–184, 1999.

[16] Ihlenfeldt WD, Takahashi Y, Abe H, Sasaki S, CACTVS: A chemistry algorithm development environment, in Machida K, Nishioka T (eds.), *Abstracts of the 20th QSAR Symposium*, Kyoto University Press: Kyoto, Japan, pp. 102–105, 1992.

[17] A Golbraikh, A Tropsha, Beware of q2! *J Mol Graphics and Modell* **20**:269–276, 2002.

[18] Shawe-Taylor J, Cristianini N, *Kernel Methods for Pattern Analysis*, Cambridge University Press, Cambridge, U.K., 2004.

[19] Rücker C, Rücker G, Meringer M, y-Randomization and its variants in QSPR/QSAR, *J Chem Inf Model* **47**:2345–2357, 2007.

J.B. Brown received B.S. degrees in Mathematics and Computer Science from the University of Evansville, USA in 2003, and M.S. and Ph.D. degrees in Informatics from Kyoto University in 2007 and 2010, respectively. He was a postdoctoral researcher in the group of Hiroshi Ishikita at the Kyoto University Career-Path Promotion Unit for Young Life Scientists from April 2010 to August 2010, and is a special postdoctoral researcher at the Department of Systems Bioscience for Drug Discovery, Graduate School of Pharmaceutical Sciences, Kyoto University since September 2010. His research interests include computational drug design, DNA repair, and development of scientific software.

Takashi Urata received B.S. and M.S. degrees in Informatics from Kyoto University in 2006 and 2008, respectively. He has been a systems designer since 2008. His research interest is in QSAR/QSPR development.

Takeyuki Tamura received B.E., M.E. and Ph.D. degrees in informatics from Kyoto University, Japan, in 2001, 2003, and 2006, respectively. He joined the Bioinformatics Center, Institute for Chemical Research, Kyoto University as a postdoctoral fellow in April 2006. He works as an assistant professor since December 2007. His research interests are bioinformatics and the theory of combinatorial optimization.

Midori A. Arai received her Ph. D (2000) from the University of Tokyo, Japan. She was a JSPS fellow at Harvard University (2001), a special postdoctoral researcher at RIKEN (2003), and an assistant professor in Teikyo University (2004). Since 2006, she has been an associate professor in the Graduate School of Pharmaceutical Sciences, Chiba University, Japan. She received the Pharmaceutical Society of Japan Award for Young Scientists in 2010. Her research interests include organic chemistry and chemical biology.

Takeo Kawabata received his D. Pharm. Sc. degree from Kyoto University, Japan, in 1983. He worked as a postdoctoral fellow in Indiana University, USA, from 1983 to 1985, and as a researcher in Sagami Chemical Research Center, Japan from 1985 to 1989. He joined the Institute for Chemical Research, Kyoto University as an assistant professor in 1989 and has been a professor since 2004. His research interests include synthetic organic chemistry and asymmetric synthesis.

Tatsuya Akutsu received M.Eng in Aeronautics (1986) and Dr. Eng. in Information Engineering (1989) degrees both from the University of Tokyo, Japan. From 1989 to 1994, he was with Mechanical Engineering Laboratory, Japan. He was an associate professor in Gumma University from 1994 to 1996 and in Human Genome Center, University of Tokyo from 1996 to 2001 respectively. He joined the Bioinformatics Center, Institute for Chemical Research, Kyoto University, Japan as a professor in Oct. 2001. His research interests include bioinformatics and discrete algorithms.