Journal of Molecular Graphics and Modelling 43 (2013) 11–20

Contents lists available at SciVerse ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

The derivation of a chiral substituent code for secondary alcohols and its application to the prediction of enantioselectivity

Jing-Jie Suo[a], Qing-You Zhang[a,*], Jing-Ya Li[a], Yan-Mei Zhou[a], Lu Xu[b]

[a] Institute of Environmental and Analytical Sciences, College of Chemistry and Chemical Engineering, Henan University, Kaifeng 475004, PR China
[b] Changchun Institute of Applied Chemistry, Chinese Academy of Sciences, Changchun 130022, PR China

ARTICLE INFO

Article history:
Accepted 23 March 2013
Available online 17 April 2013

Keywords:
Chiral substituent code
Secondary alcohol
Asymmetric reaction
Structure–enantioselectivity relationship
Random forest

ABSTRACT

A chiral substituent code was proposed based on the features of secondary alcohols, in which a chiral center is attached to two substituents in addition to —OH and —H substituents. The new chirality code, which was generated by predefining positional information of four substituents attached to stereocenter, was applied to two datasets composed of secondary alcohols as the enantioselective products of asymmetric reactions. In the first dataset, the chemical reaction was catalyzed by a biocatalyst, lipase from *Candida rugosa*. The catalyst for the second dataset was (−)-diisopinocampheylchloroborane. The structure–enantioselectivity relationship models were constructed using random forests with the chiral substituent code as the input. The resulting models were assessed both in terms of single enantiomers and pairs of enantiomers. Satisfactory results were obtained for both datasets. Although the chiral substituent code was specifically developed for secondary alcohols, it can easily be extended to represent chiral compounds possessing a specific chiral center bonded to two variable substituents.

1. Introduction

Two enantiomers are mirror images of each other, just like one's left and right hands that are the same except for the 3D configuration. Enantiomers often possess different biological activity. For example, the (S)-enantiomer of penicillamine has antiarthritic activity, but the (R)-enantiomer is extremely toxic. Thus, enantioselectivity is a key issue in the preparation of chiral compounds, such as chiral drug. As the need for enantiopure compounds grows, the ability to prepare compounds enantioselectively is becoming increasingly important [1]. One of the most efficient strategies for the synthesis of chiral molecules is asymmetric catalysis. However, the number of experiments performed to screen for a catalyst that can carry out a specific transformation with the desired enantioselectivity is typically rather limited. Therefore, it is critical to extract knowledge from previous experiments and use this knowledge when designing additional experiments [2].

Hypotheses and empirical rules regarding the enantioselectivities of asymmetric reactions of secondary alcohols have been suggested in the literatures, but these hypotheses and rules are generally only suitable for a specific subset of substrates, catalysts, or types of chemical reactions [3–7]. For example, Kazlauskas et al. proposed an empirical rule based on the size of the substituent at the stereocenter to predict which enantiomer of a secondary alcohol reacts faster in reactions catalyzed by lipase from *Candida rugosa* [7]. However, this rule is only suitable for cyclic secondary alcohols.

Chemoinformatic techniques can take advantage of available experimental data to make predictions in new situation and avoid complicated calculations, such as molecular dynamics and quantum chemistry [8]. Herein, chemoinformatic models were built based on information concerning the two substituents of secondary alcohols, without explicitly encoding the whole structure of the catalysts. Therefore, we built models of structure–enantioselectivity relationships based on a chirality code derived from multiple properties of the substituents attached to the chiral center to automatically predict the enantioselectivity for the secondary alcohols in two datasets.

Studies of structure–enantioselectivity relationships rely on chirality codes to distinguish between enantiomers [9]. Among the many diverse molecular descriptors currently available, only a few are capable of discriminating between enantiomers. Among the molecular descriptors, some chiral descriptors [10–12] have been suggested to extend the 2D topological descriptors based on the Cahn–Ingold–Prelog (CIP) rules [13–15]. Although the CIP descriptors, which are derived from the atomic number of an atom, are excellent for labeling and identifying the configurations of chiral centers, they have a fundamental weakness for developing chemoinformatic models because they are not designed to bear any intrinsic chemical meaning [16].

* Corresponding author at: College of Chemistry and Chemical Engineering, Henan University, Jinming Street, Kaifeng 475004, PR China, Tel.: +86 15993351143.
E-mail addresses: zhqingyou@henu.edu.cn, zhqingyou@yahoo.com.cn (Q.-Y. Zhang).

1093-3263/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jmgm.2013.03.005

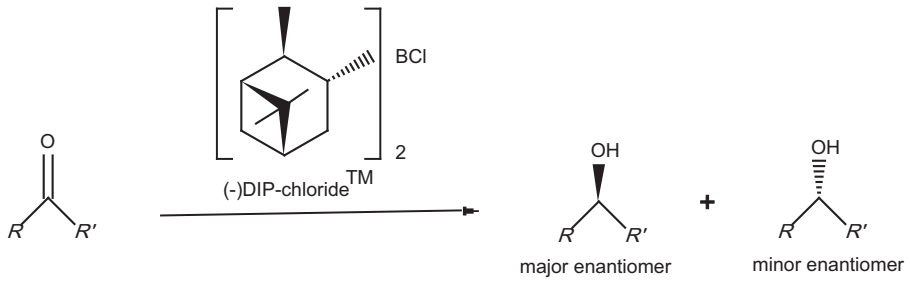

Fig. 1. Enantioselective reduction of ketones by (−)-diisopinocampheylchloroborane as catalyst.

Instead of atomic number, several chirality codes based on atomic or bond properties have also been developed [17–19]. These chirality codes include conformation-independent and conformation-dependent chirality codes that were derived from the radial distribution function and applied to the prediction of NMR chemical shift [20], the elution order of each enantiomer in a pair when separated by chiral chromatography [21], and the enantioselectivity of catalysts in organic synthesis [22,23]. However, the conformation-independent and conformation-dependent chirality codes can only be derived from a single property of an atom.

We have proposed a physicochemical atomic stereodescriptor (PAS), which was derived from the multiple topological and physicochemical properties of the four substituents attached to a chiral center [16]. The PAS is a general descriptor for a chiral center and depends on configuration assignment. PAS could be used to represent secondary alcohols by using multiple properties. However, PAS does not take advantage of the specific structure features of secondary alcohols.

All of the secondary alcohols studied in this paper possess an —OH and —H substituent at their chiral center. Thus, the codes for secondary alcohols need only to consider the other two substituents. In the present study the other two substituents, designated as "left substituent" and "right substituent", were described by multiple properties, such as the size of the substituent, the electronegativity of the atom directly bonded to the chiral center. Although chirality is three-dimensional, the process of overlapping the structures, as in comparative molecular field analysis (CoMFA), is not required to generate a chiral substituent code.

2. Datasets

2.1. The (−)-diisopinocampheylchloroborane catalysis dataset

The first dataset was retrieved from the literature and was composed of 50 pairs of secondary alcohols that were the products of reductions of prochiral ketones using (−)-diisopinocampheylchloroborane as a catalyst [17]. The catalytic asymmetric reaction is shown in Fig. 1, and the predominant enantiomeric product for each pair (**1**–**50** in Fig. 2) was assigned as class A. The opposite enantiomers (not shown in Fig. 2) were assigned as class B.

All of the secondary alcohols in this dataset contain only one chiral center, and all of the chiral centers have —OH and —H substituents. Because the catalyst is constant, the identities of the substituents attached to the chiral center of the secondary alcohols can be regarded as the crucial factors affecting the configuration of the enantiopreferred products. The 50 enantiomeric pairs were divided into a training set including 80 secondary alcohols (40 pairs of enantiomers) and a test set including 20 secondary alcohols (**1**, **5**, **8**, **13**, **20**, **23**, **27**, **35**, **37**, **39** and their enantiomers).

2.2. The lipase catalysis dataset

The second dataset, which was gathered by Kazlauskas et al., included 134 secondary alcohols (67 pairs of enantiomers) as the chiral products of racemic resolutions by transesterification, esterification, or hydrolysis catalyzed by lipase from *C. rugosa* (CRL) [7]. Only one diastereomeric product for each reaction was included in the dataset. For enantiomeric pairs of secondary alcohol substrates, only the 67 secondary alcohols that reacted faster than their enantiomers are shown in Fig. 3 and were assigned as class A, while the opposite enantiomers were assigned as class B.

Kazlauskas et al. proposed a rule for predicting the enantioselectivity of this secondary alcohol dataset based only on the sizes of the substituents connected to the chiral secondary alcohol center. However, the results of this rule are not reliable for acyclic alcohols. Similarly, in this article the prediction model is constructed based on only secondary alcohol stereocenter but takes multiple properties of substituents into consideration.

To test the accuracy of the constructed prediction model, the dataset, which was composed of 134 compounds (67 pairs of enantiomers), was randomly partitioned into a training set of 104 compounds (52 pairs of enantiomers) and a test set including the remaining 30 compounds (**3**, **8**, **11**, **14**, **19**, **23**, **26**, **29**, **31**, **38**, **45**, **49**, **54**, **59**, **66** and their enantiomers).

3. Methodologies

3.1. Chiral substituent code

As mentioned above, for all secondary alcohols there is a chiral center attached to —OH and —H substituents. The nomenclature used in the paper takes advantage of this constant structural feature and is as follows: when the molecule is oriented such that the —OH is behind the plane of the paper and the —H is front of the plane of the paper, the other two substituents attached to the chiral center are labeled as "Left" and "Right" as shown in Fig. 4.

Based on this specific feature of secondary alcohols, the chiral substituent code was derived from the properties of the chiral center, as well as the left and right substituents. The procedure used to generate the chirality code is as follows.

First, the Cartesian coordinates and physicochemical properties of the atoms were calculated by the Jchem and MarvinBean programs in ChemAxon [24]. Next, 13 topological and physicochemical properties were extracted by in-house software to describe the chiral secondary alcohol (see Table 1).

In Table 1, the first four properties are topological and relate to the size of the substituents at the chiral center. The size of substituent is an important spatial factor for enantioselectivity. Other properties in this table are atomic charge and polarizability, also known to potentially influence the enantioselectivity [18,21,25].

Atomic partial charges represent the charge distribution in molecules. When an electrically neutral atom bonds to another neutral atom, the electrons of the former atom are partially drawn

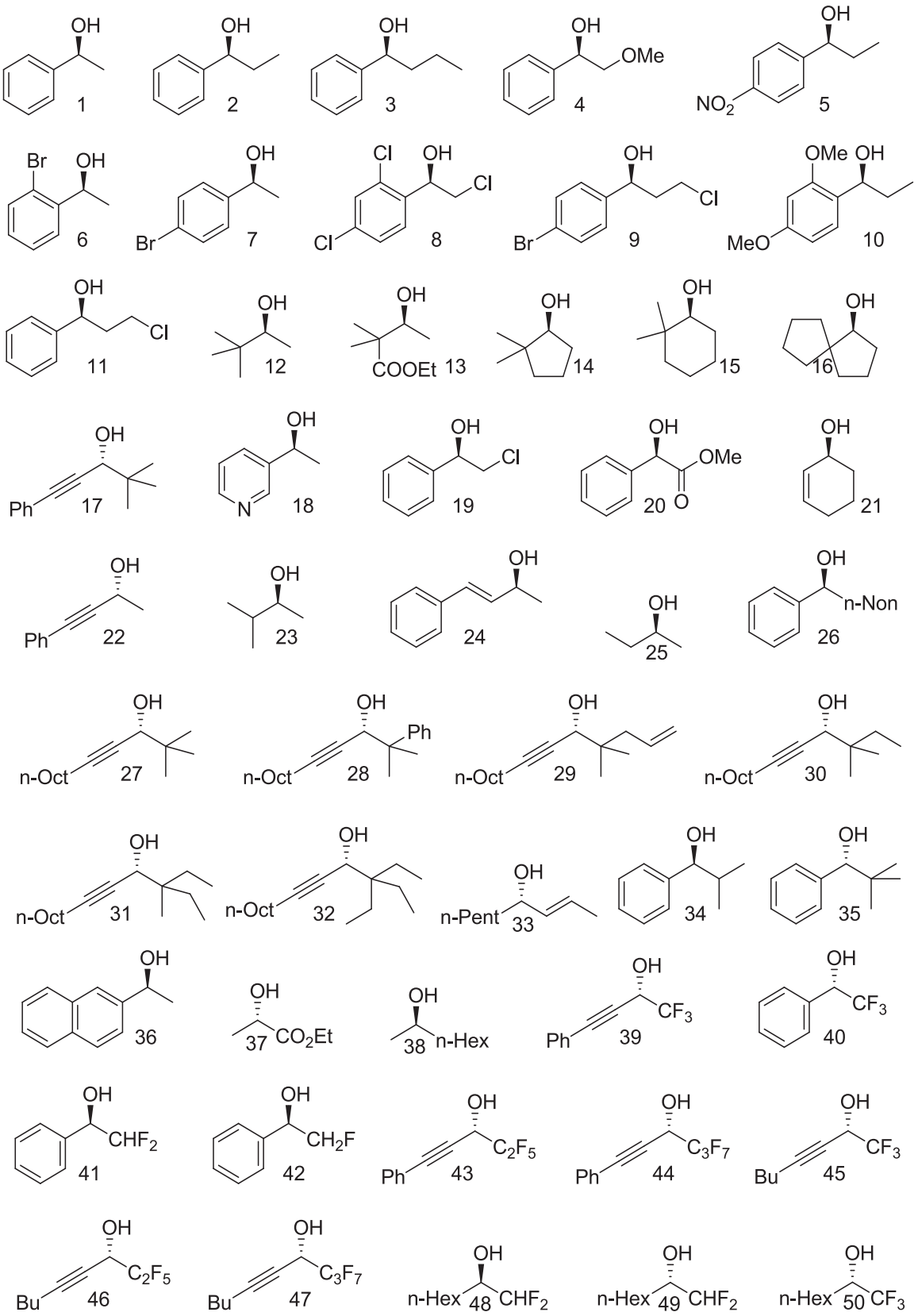

Fig. 2. The 50 preferred chiral alcohols obtained via the reduction of the correspondent ketones.

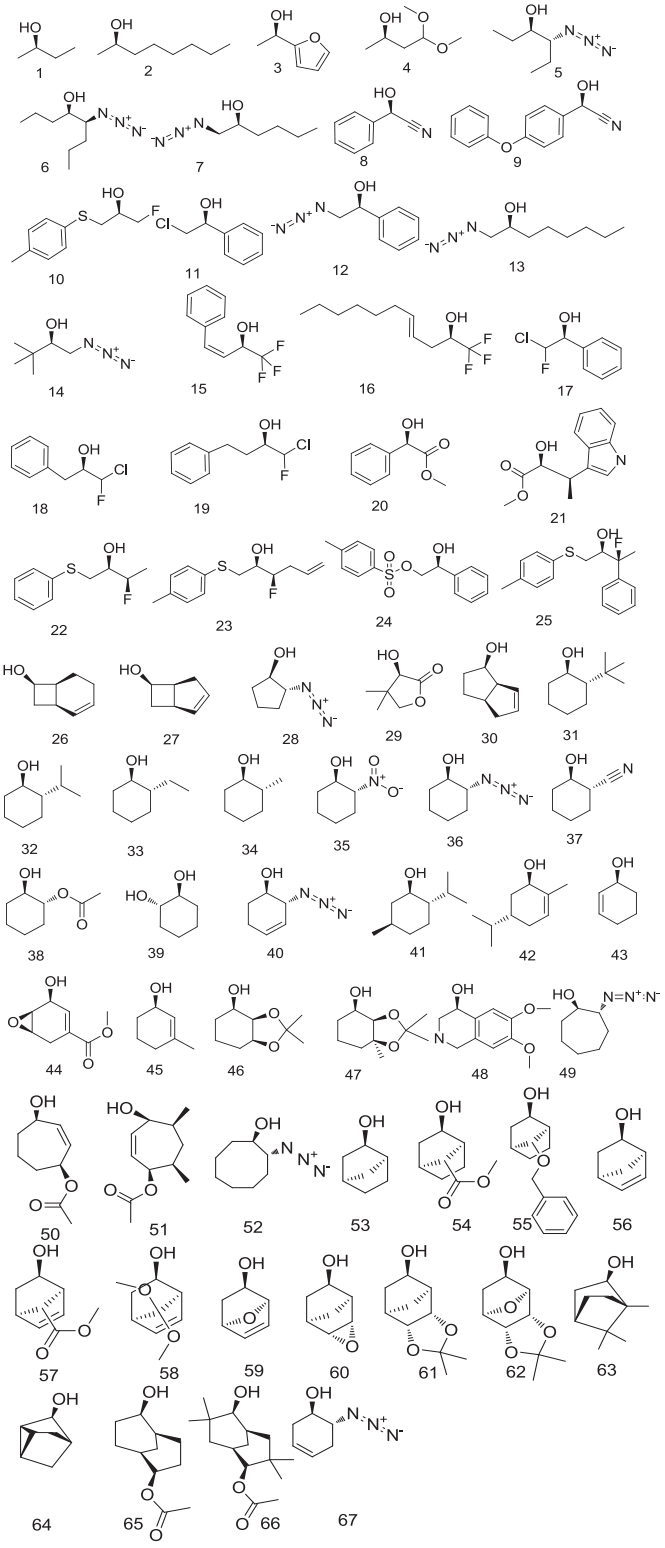

Fig. 3. The 67 secondary alcohols preferentially produced in reactions catalyzed by CRL.

away. This atom will become more positive to its initial state. On the contrary, the atomic charge of its bonded atom is negative. The π atomic charge, the δ atomic charge and the sum of them were selected and listed as the fifth, the sixth and the seventh properties in Table 1, individually.

The electric field generated by partial charges of a molecule spread through intermolecular cavities and the solvent. The induced partial charge (induced dipole) has a tendency to diminish the external electric field and this is termed polarizability. The calculation takes into account the effect of partial charges upon atomic polarizability as well as 2D and 3D geometries. Atomic polarizability is listed as the eighth property in Table 1.

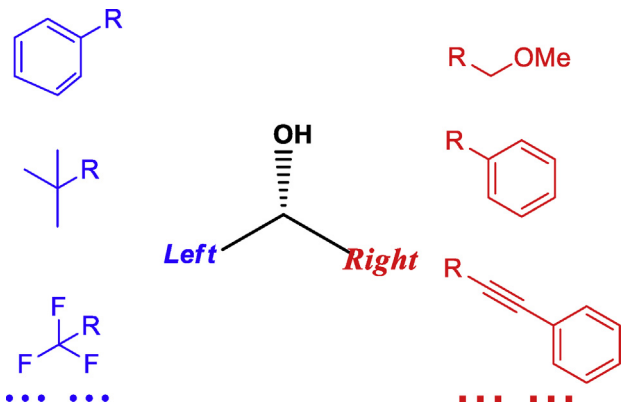

Fig. 4. The universal structure of secondary alcohol.

Electronegativity of an atom is the average of its ionization potential and the electron affinity. The π orbital electronegativity and the δ orbital electronegativity are the ninth and the tenth properties in Table 1.

Charge density is a measure of electric charge per unit volume of space. The π charge density and the total charge density are shown as the eleventh and the twelfth properties in Table 1.

The steric hindrance of an atom is calculated from the covalent radii values and geometrical distances. The hindrance is the thirteenth property in Table 1.

Ten types of variables were derived from the 13 properties shown in Table 1. The "i" subscript denotes the number of the property in Table 1. For the ten types of variables, in addition to the properties of the atoms attached to the chiral center, the properties of some atoms that are up to three bonds away from the chiral center have also been considered, e.g., the maximum and minimum values of a property of the atoms in the substituent. The properties of the atoms nearby the chiral center should be essential factor to determine the enantioselectivity.

The ten types of variables are as following.

(a) L_i (i = 1–13) represents the properties of the left substituent (for the first four properties in Table 1) or the properties of the atom directly bonded to chiral center in the left substituent (for the last nine properties).

Table 1
The 13 topological and physicochemical properties.

No.	Properties
1	Number of atoms
2	Number of atoms that are up to three bonds away from the chiral center
3	Distance (in number of bonds) from the chiral center to the farthest atom in the substitute
4	Maximum distance (in number of bonds) between two atoms in the substitute
5	π atomic charge
6	δ atomic charge
7	The sum of π charge and δ atomic charge
8	Atomic polarizability
9	π orbital electronegative
10	δ orbital electronegative
11	π charge density
12	Total charge density
13	Steric hindrance

(b) R_i (i = 1–13) represents the properties of the right substituent (for the first four properties) or the properties of the atom directly bonded to chiral center in the right substituent (for the last nine properties).
(c) C_i (i = 5–13) represents the properties of chiral secondary alcohol carbon atom.
(d) D_i (i = 1–13) represents the difference between the value of L_i and R_i. If $L_i > R_i$, $D_i = 1$; if $L_i < R_i$, $D_i = -1$; and if $L_i = R_i$, $D_i = 0$.
(e) $minL_i$ (i = 5–13) represents the minimum value of the properties of the atoms that are up to three bonds away from the chiral center for the left substituent.
(f) $minR_i$ (i = 5–13) represents the minimum value of properties of the atoms that are up to three bonds away from the chiral center for the right substituent.
(g) $minD_i$ (i = 5–13) represents the difference between the value of $minL_i$ and $minR_i$. If $minL_i > minR_i$, $minD_i = 1$; if $minL_i < minR_i$, $minD_i = -1$; and if $minL_i = minR_i$, $minD_i = 0$.
(h) $maxL_i$ (i = 5–13) represents the maximum value of the properties of the atoms that are up to three bonds away from the chiral center for the left substituent.
(i) $maxR_i$ (i = 5–13) represents the maximum value of the properties of the atoms that are up to three bonds away from the chiral center for the left substituent.
(j) $maxD_i$ (i = 5–13) represents the difference between the value of $maxL_i$ and $maxR_i$,. If $maxL_i > maxR_i$, $maxD_i = 1$; if $maxL_i < maxR_i$, $maxD_i = -1$; and if $maxL_i = maxR_i$, $maxD_i = 0$.

Only three types of variables were generated from the first four topological properties, type (a), (b), and (d). The remaining nine physicochemical properties contributed to each of the ten types of variables. As a result, a vector of $4 \times 3 + 9 \times 10 = 102$ variables was obtained for each chiral secondary alcohol.

Alcohol **4** in Fig. 3 is used as an example to illustrate the generation of the 102-dimensional chiral substituent code. The structure and atom-numbering scheme of alcohol **4** are shown in Fig. 5. If the hydrogen atoms are omitted, atom 1 is the only atom in the left substituent, and atoms 3, 4, 6 and 8 are the atoms that are up to three bonds away from the chiral center in the right substituent. The physicochemical properties of these atoms are listed in Table 2, and 102 variables for the chiral substituent code of alcohol **4** are displayed in Table 3.

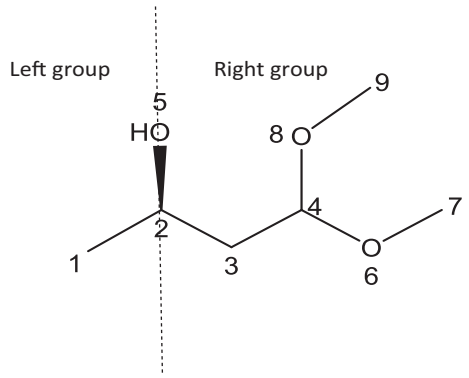

Fig. 5. An illustration of chiral secondary alcohol with atom number.

3.2. Chiral molecular connectivity indices

The molecular connectivity indices were initially suggested by Randic [26] and were extended by Kier and Hall [27]. The definition of these indices ${}^{m}\chi_t$ is as follow:

$${}^{m}\chi_t = \sum (\delta_1 \delta_2 \ldots \delta_n)^{-0.5}$$

where m is the order (the highest order is 6 herein), t is the type of a subgraph which can be path, cluster, and chain, and the δ_i is the degree of an atom which can be derived from topological structures or from structures with type of bonds and heteroatom considered. These indices are based on the 2D connection between atoms and cannot distinguish stereoisomers, such as enantiomers. For studies

Table 2
Physicochemical properties of the atoms involved in the calculation of chiral substituent code.

Atom no.	Property no.					
	1	2	3	4	6	8
5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
6	−0.0388	0.0561	0.0242	0.1590	−0.3557	−0.3557
7	−0.0388	0.0561	0.0242	0.1590	−0.3557	−0.3557
8	1.1163	1.1163	1.1163	1.1163	0.8321	0.8321
9	0.0000	0.0000	0.0000	0.0000	3.5269	3.5269
10	7.6270	8.5010	8.2031	9.4870	9.7606	9.7606
11	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
12	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
13	0.6160	0.9867	0.9267	0.9831	0.9188	0.9106

Table 3
The 102-dimensional chiral substituent code of alcohol **4**.

Property no.	*L*	*R*	*D*	*C*	*minL*	*minR*	*minD*	*maxL*	*maxR*	*maxD*
1	1	6	−1	–	–	–	–	–	–	–
2	1	4	−1	–	–	–	–	–	–	–
3	1	4	−1	–	–	–	–	–	–	–
4	0	4	−1	–	–	–	–	–	–	–
5	0.0000	0.0000	0	0.0000	0.0000	0.0000	0	0.0000	0.0000	0
6	−0.0388	0.0242	−1	0.0561	−0.0388	−0.3557	1	−0.0388	0.1590	−1
7	−0.0388	0.0242	−1	0.0561	−0.0388	−0.3557	1	−0.0388	0.0590	−1
8	1.1163	1.1163	0	1.1163	1.1163	0.8321	1	1.1163	1.1163	0
9	0.0000	0.0000	0	0.0000	0.0000	0.0000	0	0.0000	3.5269	−1
10	7.6270	8.2031	−1	8.5010	7.6270	8.2031	−1	7.6270	9.7606	−1
11	0.0000	0.0000	0	0.0000	0.0000	0.0000	0	0.0000	0.0000	0
12	0.0000	0.0000	0	0.0000	0.0000	0.0000	0	0.0000	0.0000	0
13	0.6160	0.9267	−1	0.9867	0.6160	0.9106	-1	0.6160	0.9831	-1

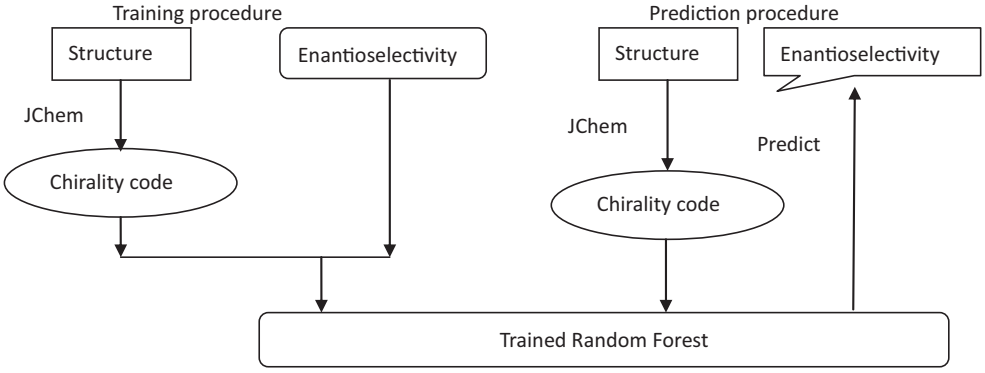

Fig. 6. The training and the test procedures.

on chiral compounds, the indices were extended [10] by changing the value of δ_i. For asymmetric atoms in R-configurations, the δ_i is substituents with $\delta_i + c$; for asymmetric atoms in S-configurations, the δ_i is substituents with $\delta_i - c$. The correction c is less than 3. The details of the chiral extension of molecular connectivity indices are in Ref. [10].

3.3. Chiral topological charge indices

The procedure for generating the topological charge index [28,29] is briefly introduced as follows:

(1) The distance matrix D is constructed. Its element d_{ij} is the number of bonds between atoms, i and j;
(2) The Coulombic matrix Q is constructed. Its element is q_{ij}. If $i = j$, $q_{ij} = 0$; if $i \neq j$, $q_{ij} = 1/d_{ij}^2$;
(3) The connectivity matrix A is constructed;
(4) The matrix M, $M = AQ$, is constructed. Its element is m_{ij};
(5) Let $g_{ij} = m_{ij} - m_{ji}$, and k is the distance between atom i and j, the indices G_k and J_k are defined as follows:

$$G_k = \sum_{i=1}^{N-1} \sum_{j=1}^{N} |g_{ij}| \delta_{k,d_{ij}}$$

$$J_k = \frac{G_k}{N-1}$$

where k is the order, δ is Kronecker's delta:

$$\delta_{i,j} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

If $k = 1$, G_1 and J_1 indices will be obtained. In this case, a pair of atoms at a distance of 1 will be described. Indices G_2 and J_2 represent the pairs of atoms between which the distance is 2, and so on. Here k will take a number which is less or equal to 6.

Topological charge indices are also 2D topological indices. The extension for a chiral compound is to replace the diagonal elements of the connectivity matrix by $a_{ii} \pm c$. In which, R-configuration will take "+"; S-configuration will take "−". As the $a_{ii} = 0$ for topological charge index, thus $a_{ii} \pm c = \pm c$.

3.4. Random forest

A random forest [30] (RF) is an ensemble of unpruned classification trees created by using bootstrap samples of the training set and random subsets of variables to define the best split at each node. It is a high-dimensional nonparametric method that works well for datasets containing large numbers of variables. Prediction is made by a majority vote for the individual trees. This method has been demonstrated to be extremely accurate in a variety of applications [31]. Additionally, performance is internally assessed using the prediction error for the objects excluded during the bootstrap procedure. The random forest method quantifies the importance of a variable by determining the increase in misclassification that occurs when the values of the variable are randomly permuted, or by the decrease in a node's impurity every time the variable is used for splitting. In this study, random forests were grown with version 2.10.1 of the R software using the random forest library [32]. The number of descriptors was set to the default values for each random selection, and 1000 trees were grown for each forest. The RFs were used to classify the enantiomers by using the aforementioned chiral substituent code.

The out-of-bag (OOB) estimation is a reliable indication of the robustness of the model. In the OOB estimation, every tree of the forest is grown with a random subset of the training set, and predictions are obtained for the excluded objects. All of the predictions for each tree are combined to calculate the out-of-bag estimation (OOB).

Random forest calculates a classification probability value between 0 and 1 for each predicted object (secondary alcohol). For example, if a compound is predicted by 900 trees to belong to class A and by 100 trees to belong to class B, the probability that the compound is class A (P_A) is 0.9 and the probability that the compound is class B (P_B) is 0.1. If the P_A given by random forest for a compound is higher than 0.5, the compound is predicted to belong to class A; otherwise, the compound is predicted to belong to class B.

3.5. Counterpropagation neural network

Counterpropagation neural networks (CPG NN) [33] are useful to model complex and non-linear relationships. They were used to investigate the relationship between the chiral substituent code of secondary alcohol and the enantioselectivity in an asymmetric reaction. If a secondary alcohol belongs to class A, a value of +1 was assigned as the output to represent enantioselectivity of this product, and an output of −1 was assigned to products of class B.

The input data for a CPG network are stored in a two dimensional grid of neurons, each containing as many elements (weights) as there are input variables. In the investigations described in this paper the input variables are chiral substituent codes. This part of the CPG network is basically a Kohonen network, or self-organizing map (SOM). The output data (in this case the classification of enantioselectivity) are stored in a second layer that acts as a look-up table.

Before the training of a CPG network starts, random weights are generated. During the training, each individual object (chiral substituent code) is mapped into that neuron of the Kohonen layer (central neuron or winning neuron) that contains the most similar

Table 4
The prediction results of enantioselectivity of secondary alcohols obtained via the reduction of the correspondent ketones.

Method[a]	Number of variable	OOB of the training set (80 alcohols)	Correct prediction of the test set (20 alcohols)	OOB of the whole data set (100 alcohols)
RF	102	86%	80%	84%
GA + RF	24	91%	85%	89%

[a] RF denotes random forest; GA denotes genetic algorithm.

weights compared to the input data (chiral substituent codes). The weights of the winning neuron are then adjusted to make them even more similar to the presented data, and the weight of the corresponding output neuron is adjusted to become closer to the value of enantioselectivity (+1 or −1). The neurons in the neighborhood of the winning neuron are also corrected, the extent of adjustment depending on the topological distance to the central neuron. The network is trained iteratively, i.e., all the objects of the training set are presented several times, and the weights are corrected, until the network stabilizes.

After the training, the CPG NN is able to predict the enantioselectivity on input of an object represented by its chiral substituent codes. The winning neuron is chosen and the corresponding weight in the output layer is used for prediction. When a trained network was applied to make predictions, a positive value of the output was interpreted as a prediction of class A for an alcohol, and vice versa.

In order to reduce the impact of fluctuations derived from the random values of the weights at the outset of the training and the random order by which examples were presented during the training, five CPG networks were trained independently with the training set, and the average value of the five outputs was used for the prediction.

3.6. Genetic algorithms

Some variables of the chiral substituent code may not be relevant to the enantioselectivity of the secondary alcohol for our purposes and can even introduce noise. Models with few descriptors are usually preferred to make predictions with increased robustness. In this paper, we used a genetic algorithm for the selection of relevant variables. Genetic algorithms simulate the evolution of a population, where each individual of the population represents a subset of descriptors and its fitness is assessed by the ability to generated accurate models.

At the beginning of the evolution, a population that consisted of 24 individuals was generated. The probability of randomly selecting a variable into a subset was between 0 and 0.4.

A population of individuals was allowed to evolve over 100 generations. In each generation, half of the population die, and the other half survive (the fittest individuals). Each of the surviving individuals mates with another (randomly chosen) surviving individual, and two new offspring are generated. These new individuals result from crossover of their parents chromosomes, followed by random mutation. The population of the next generation consists of the new offspring and their parents.

Crossover occurs at a randomly chosen single point. Mutation is allowed to occur at every gene of the new offspring with a random probability. The probability of mutation $0 \rightarrow 1$ is set for each individual (randomly) between 0 and 0.05. The probability of mutation $1 \rightarrow 0$ is set for each individual (randomly) between 4 and 6 times higher than the probability of mutation $0 \rightarrow 1$.

The evaluation (scoring) of each individual was made by a CPG NN that uses the subset of chiral substituent codes to make predictions. The NN is trained with the training set, and the score of the subset of molecular descriptors is based on the root-mean-square of errors for the predictions obtained for the training set. The individuals (subsets of the code) resulting in lower errors are considered to be fitter than those resulting in higher errors and are selected for mating.

4. Results and discussion

4.1. Prediction of enantioselectivity for (−)-diisopinocampheylchloroborane-catalyzed reductions

A model establishing relationships between the chiral substituent codes of secondary alcohols and their classification (A and B) was built by random forests. First, a random forest was trained with the 102-dimensional chiral substituent codes of 80 secondary alcohols in the training set. Next, the trained random forest was used to make classifications for the test set. The procedure above was shown in Fig. 6. Finally, the percentage of correct predictions was calculated to assess the prediction ability of the built model (ratio of the number of correctly classified compounds to the number of compounds submitted to the model). If a preferred secondary alcohol is predicted to class A, or a secondary alcohol which react slower is predicted to class B, i.e. the predict result is in accordance with the experimental classification, the prediction for the secondary alcohol is considered to be correct. The results for the training set (OOB cross-validation) and for the test set are shown in Table 4.

The results in Table 4 indicate that, in a cross-validation test, 86% of the training set results and 80% of the test set results were predicted correctly. If the entire dataset was used to train the random forest, 84% of the secondary alcohols were predicted correctly in the cross-validation test.

To obtain a more robust and accurate model, selection of variables from the (102-dimensional) chiral substituent code was performed using a genetic algorithm (GA) that was described in the METHODOLOGIES section. Instead of the 102-dmensional chirality code, a subset of 24 variables selected by genetic algorithm was used to train a random forest, and classify the test set. The results are displayed in Table 4. The correct stereochemical outcome was predicted for 91% of the secondary alcohols in the training set and 85% of the secondary alcohols in the test set. If the entire dataset was used to train the random forest, the OOB cross-validation reached 89%. This demonstrates that the results were significantly improved after a subset of the variables was chosen for inclusion in the prediction model.

The above procedure is not perfect because it does not apply to situations in which the compounds in class A and those in class B are unrelated to each other. In general, the two classes of compounds being classified are not related each other. However, in this work, if one enantiomer of a secondary alcohol is assigned as A, the opposite enantiomer will be assigned as B. Compounds in class A and class B have an obvious relationship to each other. Thus, it is not quite appropriate to consider A and B independently for their classifications. Consequently, we also assessed the results taking into account the predictions obtained for both enantiomers of each secondary alcohol. There are three possible outcomes: (1) if enantiomer A is recognized as A and enantiomer B is recognized as B, the pair of enantiomers is classified correctly; (2) if enantiomer A is recognized as B and enantiomer B is recognized as A, the pair of enantiomers is classified incorrectly; (3) if enantiomer A is

Table 5
The prediction results of enantioselectivity in term of pair of enantiomers for (−)-diisopinocampheylchloroborane-catalyzed reductions.

Method [a]	OOB of the training set (40 pairs)	Correct prediction of the test set (10 pairs)	OOB of the whole data set (50 pairs)
GA + RF	90%	80%	88%
GA + RF + Pr	90%	90%	90%

[a] RF denotes random forest; GA is genetic algorithm; Pr represents probability given by random forest.

recognized as A, and enantiomer B is also recognized as A, or enantiomer A is recognized as B, and enantiomer B is recognized as B, the pair of enantiomers cannot be predicted, that is, the prediction result is undecided. The results obtained using this procedure are shown in line 2 of Table 5, where a pair of enantiomers was only designated as correctly predicted when both of the enantiomers were predicted correctly. When these results are compared with the results in Table 4, we can see that they are slightly different.

To avoid the undecided prediction result, the probabilities that each enantiomer belonged to either class A or class B given by random forest (P_A or P_B, see Section 3) were used to further determine to which class an enantiomer should be assigned. In case (3) as mentioned above, although the two enantiomers of a secondary alcohol can be misclassified as both A or both B, the probabilities obtained by using random forest for the classification of each enantiomer are not the same. In this case, whichever probability (P_A or P_B) was higher was given higher priority for the classification of an enantiomer as class A or class B. This means that the enantiomer with the highest P_A (or P_B) would be assigned as class A (or B) and the opposite enantiomer would be assigned to the opposite class. For example, when the whole dataset was used to train the random forest, both secondary alcohol **20** and its enantiomer were assigned as class A according to the results of the OOB cross-validation. The value of P_A for secondary alcohol **20** was 0.80 and the P_A for its enantiomer was 0.69. The former was larger than the later, and thus compound **20** was assigned as class A and its enantiomer was assigned as class B, i.e., the pair of enantiomers was correctly predicted. The results obtained by using this procedure are also displayed in Table 5. For the test set, this procedure increased the percentage of correct predictions from 80% to 90%. The results were clearly improved using the rule of higher probability possessing a higher priority in the assignment process.

The relative importance of every variable was determined by random forest. The ranking of the 24 variables in order of decreasing importance is as follows: $maxR_{10}$, L_2, R_2, $maxD_8$, $maxL_7$, $minR_{13}$, L_8, $minR_8$, L_{11}, D_{10}, $maxL_{11}$, $minL_8$, $maxL_5$, $maxD_{12}$, $minD_8$, D_5, $maxD_{10}$, $maxD_7$, C_{11}, $minD_{11}$, D_1, $minR_{11}$, $maxL_{12}$, and C_5. Several properties of each substituent were used by the model, including, e.g., the size of substituent (L_2, R_2 and D_1), atomic polarizability ($maxD_8$, L_8, $minR_8$, $minL_8$ and $minD_8$).

To confirm the prediction ability of this method, a leave-one-pair-out cross-validation was also performed. 4 pairs of enantiomers and 2 single enantiomers were predicted incorrectly, i.e., the correct prediction rate was 90% in terms of single enantiomers. If the random forest probability was utilized, 5 pairs of enantiomers were predicted incorrectly, resulting in a correct prediction percentage of 90% for pair of enantiomers. The results were the same for both methods.

In addition, a leave-one-pair-out cross-validation was performed using a CPG NN of size 9×9, and the correct prediction rate was 87% in terms of single enantiomers – slightly worse than the result obtained using the random forest.

In order to assess the performance of the approached chiral substituent code, the comparison between chiral substituent code and the combination of chiral molecular connectivity indices and chiral topological charge indices was performed. The chiral molecular connectivity indices and topological charge indices of 50 pairs of secondary alcohols were calculated by in-house program. The chirality codes were derived from three corrections $c = 1$, 2, or 2.5, individually. The variable selection was also performed by genetic algorithm. Each chirality code was used to build prediction model by random forest. The result of corresponding OOB cross-validation of the whole dataset was up to 71%, which is obviously worse than the result in this paper.

4.2. Prediction of enantioselectivity in lipase-catalyzed reactions

Similarly, the relationship model between the chiral characteristics of secondary alcohols represented by the chiral substituent codes and their classification (class A or class B) was built by random forest. The results obtained by random forest are shown in Table 6. The result for the internal cross-validation (out-of-bag estimation, OOB) of the training set was 84% correct and the prediction result for test set was 83% correct. If the whole dataset was used to train random forest, the OOB of the cross-validation result was 86%.

The selection of variables from the 102-dimensional chiral substituent code was performed using a genetic algorithm. A variable subset composed of a 14-dimensional chiral substituent code was obtained. The obtained subset was used to construct the prediction models as mentioned above, and the results are displayed in Table 6.

The OOB cross-validation for the training set and the correct prediction percentage for the test set were 93% and 87%, respectively. The OOB cross-validation for the whole dataset reached 91%. It is clear that better results were obtained after variable selection, in which the number of the variables was decreased to 14. The results obtained using this method are more reliable when compared with the result of 78% (52 out of the 67 pairs of enantiomers) predicted correctly by Kazlauskas' empirical rules [7].

Inspection of the OOB cross-validation for the whole dataset indicated that 4 pairs of enantiomers were undecided and 4 pairs of enantiomers were predicted incorrectly. Therefore, the correct prediction was 59 out of 67 (88%, see row 2 of Table 7). The results for the training set and the test set are also displayed in row 2 of Table 7.

Table 6
The prediction results of enantioselectivity of secondary alcohols produced in the reactions catalyzed by CRL.

Method[a]	Number of variable	OOB of the train set (104 alcohols)	Correct prediction of the test set (30 alcohols)	OOB of the whole data set (134 alcohols)
RF	102	84%	83%	86%
GA + RF	14	93%	87%	91%

[a] RF denotes random forest; GA denotes genetic algorithm.

Table 7
The prediction results of enantioselectivity in term of pair of enantiomers for lipase-catalyzed reactions.

Method[a]	OOB of the training set (52 pairs)	Prediction of the test set (15 pairs)	OOB of the whole data set (67 pairs)
GA + RF	90%	87%	88%
GA + RF + Pr	94%	87%	91%

[a] RF denotes random forest; GA denotes genetic algorithm; Pr represents probability given by random forest.

The probability (P_A or P_B) given by random forest was used to further assess the 4 pairs of enantiomers that were predicted as undecided in the same way as in the case of the (−)-diisopinocampheylchloroborane dataset. The results of this method were that 2 pairs of enantiomers were predicted correctly and the other 2 pairs of enantiomers were predicted incorrectly, i.e., 61 out of 67 (91%, see row 3 of Table 7) were predicted correctly. The corresponding results for the training set and the test set are also listed in row 3 of Table 7. For this dataset, the results are similar to those in row 3 of Table 6, but the results in terms of pairs of enantiomer are more reasonable for this dataset because the enantioselectivity for any pair of enantiomers can be obtained clearly without any pairs remaining undecided.

Leave-one-pair-out cross-validation was also performed. Each pair of enantiomers was selected as test set once and the remaining 66 pairs of enantiomers were used to construct a prediction model by random forest. 5 pairs of enantiomers out of 67 pairs were predicted incorrectly, resulting in a correct prediction percentage of 92.5%, whether the probability (P_A or P_B) was used or not.

The relative importance of each variable was revealed by the random forest: the 14 variables selected by the genetic algorithm ranked in order of decreasing importance are as follows: L_2, $maxD_{10}$, R_2, D_{12}, C_{12}, L_{12}, C_9, $maxD_6$, $minR_6$, C_6, D_1, $minL_{10}$, $maxL_{13}$, $minR_9$, which were derived from the size of substituent, electronegativity, charge density, etc. The most important variable L_2 represents the number of atoms up to three bonds away from the chiral center for the left substituent; the second most important variable is $maxD_{10}$ which represents the difference between the δ orbital electronegative of the atom directly bonded to chiral center in the left substituent and in the right substituent; the third most important variable, R_2, represents the number of atoms up to three bonds away from the chiral center in the right substituent. Obviously, variables L_2 and R_2 embody the effect of the size of substituent, but, it is different from Kazlauskas' empirical rules based on the numbers of atoms in the substituents. In addition, variables, C_9, $minL_{10}$ and $minR_9$ derived from orbital electronegativity were also selected by genetic algorithm, which indicates that atomic orbital electronegativity is an important factor for enantioselectivity.

The results suggest, as possible research targets in the future, the effects of some physicochemical properties and size of substituents in the neighborhood of the chiral center. That is, the method has the potential to assist in the design of experiment to discover new products with interesting enantioselectivity features.

The combination of chiral molecular connectivity indices and topological charge indices was also used to train a random forest using the dataset. The results in the OOB cross-validation of the whole dataset were up to 68% of correct predictions, which are inferior to those reported in this paper.

5. Conclusions

A two-substituents-based chirality code, which was suggested based on the specific structural features of secondary alcohols, was successfully applied to predictions of enantioselectivity for two datasets by using random forests to build prediction models. Although the chiral substituent code used here was derived for secondary alcohols, it is possible to extend this code to apply to chiral compounds possessing a chiral center bonded to two common substituents, like —OH and —H in the case of secondary alcohols. For example, a dataset we are currently studying is composed of chiral primary alcohols bearing two common substituents, i.e., —CH_2OH and —H, attached to a chiral center [34]. The chiral substituent code also has the potential to be extended to other classes of compounds by using different atomic properties, bond properties, etc. to describe the substituents.

We proposed a method for resolving the problem of undecided predictions by introducing the probability generated by random forest into the result statistics for a pair of enantiomers. This method can be applied to similar types of chemoinformatics investigations in which the two classes of compounds being classified are related each other.

Acknowledgements

The authors thank the National Natural Science Foundation of China (No. 20875022) for financial support. The authors also acknowledge the International Science and Technology Cooperation of Henan Province (No. 114300510009) and our collaborators in the research group of Prof. João Aires-de-Sousa (Universidade Nova de Lisboa, Portugal). This project was also sponsored by the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jmgm.2013.03.005

References

[1] J.F. Traverse, M.L. Snapper, High-throughput methods for the development of new catalytic asymmetric reactions, Drug Discovery Today 7 (2002) 1002–1012.
[2] Q.Y. Zhang, J. Aires-De-Sousa, Structure-based classification of chemical reactions without assignment of reaction centers, Journal of Chemical Information and Modeling 45 (2005) 1775–1783.
[3] H.C. Brown, J. Chandrasekharan, P. Ramachandran, Chiral synthesis via organoboranes. 14. Selective reductions. 41. Diisopinocampheylchloroborane, an exceptionally efficient chiral reducing agent, Journal of the American Chemical Society 110 (1988) 1539–1546.
[4] H.C. Brown, P.V. Ramachandran, Asymmetric reduction with chiral organoboranes based on alpha-pinene, Accounts of Chemical Research 25 (1992) 16–24.
[5] P.V. Ramachandran, B. Gong, A.V. Teodorovic, H.C. Brown, Selective reductions. 52. Efficient asymmetric reduction of α-acetylenic α′-fluoroalkyl ketones with either B-chlorodiisopinocampheylborane or B-isopinocampheyl-9-borabicyclo [3.3.1] nonane in high enantiomeric purity. The influence of fluoro groups in such reductions, Tetrahedron: Asymmetry 5 (1994) 1061–1074.
[6] P.V. Ramachandran, A.V. Teodorovic, B. Gong, H.C. Brown, Selective reductions. 53. Asymmetric reduction of [alpha]-fluoromethyl ketones with B-chlorodiisopinocampheylborane and B-isopinocampheyl-9-borabicyclo [3.3.1] nonane. Combined electronic and steric contributions to the enantiocontrol process, Tetrahedron: Asymmetry 5 (1994) 1075–1086.
[7] R.J.W. Kazlauskas, N.E. Alexandra, T. Rappaport Aviva, A. Cuccia Louis, A rule to predict which enantiomer of a secondary alcohol reacts faster in reactions catalyzed by cholesterol esterase, lipase from pseudomonas cepacia, and lipase from *Candida rugosa*, Journal of Organic Chemistry 56 (1991) 2656–2665.
[8] A. Del Rio, Exploring enantioselective molecular recognition mechanisms with chemoinformatic techniques, Journal of Separation Science 32 (2009) 1566–1584.

[9] Q.Y. Zhang, L.Z. Xu, J.Y. Li, D.D. Zhang, H.L. Long, J.Y. Leng, L. Xu, Methods of studies on quantitative structure–activity relationships for chiral compounds, Journal of Chemometrics 26 (2012) 497–508.
[10] A. Golbraikh, D. Bonchev, A. Tropsha, Novel chirality descriptors derived from molecular topology, Journal of Chemical Information and Computer Sciences 41 (2001) 147–158.
[11] R. Natarajan, S.C. Basak, T.S. Neumann, Novel approach for the numerical characterization of molecular chirality, Journal of Chemical Information and Modeling 47 (2007) 771–775.
[12] A. Golbraikh, A. Tropsha, QSAR modeling using chirality descriptors derived from molecular topology, Journal of Chemical Information and Computer Sciences 43 (2003) 144–154.
[13] R.S. Cahn, C. Ingold, V. Prelog, Specification of molecular chirality, Angewandte Chemie International Edition in English 5 (1966) 385–415.
[14] V. Prelog, G. Helmchen, Basic principles of the cip system and proposals for a revision, Angewandte Chemie International Edition in English 21 (1982) 567–583.
[15] P. Mata, A.M. Lobo, C. Marshall, A.P. Johnson, The CIP sequence rules: analysis and proposal for a revision, Tetrahedron: Asymmetry 4 (1993) 657–668.
[16] Q.Y. Zhang, J. Aires-De-Sousa, Physicochemical stereodescriptors of atomic chiral centers, Journal of Chemical Information and Modeling 46 (2006) 2278–2287.
[17] J. Aires-De-Sousa, J. Gasteiger, New description of molecular chirality and its application to the prediction of the preferred enantiomer in stereoselective reactions, Journal of Chemical Information and Computer Sciences 41 (2001) 369–375.
[18] J. Aires-De-Sousa, J. Gasteiger, Prediction of enantiomeric selectivity in chromatography: application of conformation-dependent and conformation-independent descriptors of molecular chirality, Journal of Molecular Graphics and Modelling 20 (2002) 373–388.
[19] Q.Y. Zhang, J.Y. Li, J. Aires-De-Sousa, L. Xu, J. Leng, Conformation-dependent chirality code based on electronegativity and its applications, Chinese Journal of Analytical Chemistry 39 (2011) 257–260.
[20] Q.Y. Zhang, G. Carrera, M.J.S. Gomes, J. Aires-De-Sousa, Automatic assignment of absolute configuration from ^{1}D NMR data, The Journal of Organic Chemistry 70 (2005) 2120–2130.
[21] S. Caetano, J. Aires-De-Sousa, M. Daszykowski, Y.V. Heyden, Prediction of enantioselectivity using chirality codes and classification and regression trees, Analytica Chimica Acta 544 (2005) 315–326.
[22] Q.Y. Zhang, D.D. Zhang, J.Y. Li, Y.M. Zhou, L. Xu, Virtual screening of a combinatorial library of enantioselective catalysts with chirality codes and counterpropagation neural networks, Chemometrics and Intelligent Laboratory Systems 109 (2011) 113–119.
[23] Q.Y. Zhang, D.D. Zhang, J.Y. Li, H.L. Long, L. Xu, Prediction of enantiomeric excess in a catalytic process: a chemoinformatics approach using chirality codes, Match-Communications in Mathematical and in Computer Chemistry 67 (2012) 773–786.
[24] http://www.chemaxon.com
[25] J. Aires-De-Sousa, J. Gasteiger, Prediction of enantiomeric excess in a combinatorial library of catalytic enantioselective reactions, Journal of Combinatorial Chemistry 7 (2005) 298–301.
[26] M. Randic, Characterization of molecular branching, Journal of the American Chemical Society 97 (1975) 6609–6615.
[27] L.B. Kier, L.H. Hall, Molecular Connectivity in Chemistry and Drug Research, vol. 2, Academic Press, New York, 1976.
[28] J. Galvez, R. Garcia-Domenech, J. De Julian-Ortiz, R. Soler, Topological approach to drug design, Journal of Chemical Information and Computer Sciences 35 (1995) 272–284.
[29] J.D. De Julián-Ortiz, C. De Gregorio Alapont, I. Rıos-Santamarina, R. Garcıa-Domenech, J. Gálvez, Prediction of properties of chiral compounds by molecular topology, Journal of Molecular Graphics and Modelling 16 (1998) 14–18.
[30] B. Leo, E. Schapire, Random forests, Machine Learning 45 (2001) 5–32.
[31] V. Svetnik, A. Liaw, C. Tong, J.C. Culberson, R.P. Sheridan, B.P. Feuston, Random forest: a classification and regression tool for compound classification and QSAR modeling, Journal of Chemical Information and Computer Sciences 43 (2003) 1947–1958.
[32] http://www.r-project.org/
[33] J. Gasteiger, J. Zupan, Neural networks in chemistry, Angewandte Chemie International Edition in English 32 (1993) 503–527.
[34] A.N.E. Weissfloch, R.J. Kazlauskas, Enantiopreference of lipase from *Pseudomonas cepacia* toward primary alcohols, Journal of Organic Chemistry 60 (1995) 6959–6969.