MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **67** (2012) 773-786

ISSN 0340 - 6253

Prediction of Enantiomeric Excess in a Catalytic Process: A Chemoinformatics Approach Using Chirality Codes

Qing-You Zhang [†], Dan-Dan Zhang [†], Jing-Ya Li [†], Hai-Lin Long [†], Lu Xu [‡,*]

† *Institute of Environmental and Analytical Sciences, College of Chemistry and Chemical Engineering, Henan University, Kaifeng, 475004, PR China*

‡ *Changchun Institute of Applied Chemistry, Chinese Academy of Sciences, Changchun, 130022, PR China*

(Received September 1, 2011)

ABSTRACT: The enantiomeric excesses obtained with 296 chiral catalysts in the asymmetric hydrogen transfer to acetophenone are extracted from a combinatorial database constructed by Riant et al., and used to investigate their relationships with the molecular structures of the catalysts. These catalysts incorporate three components, namely a β-amino alcohol, an aldehyde or ketone, and a metal complex precursor. The structures of the catalysts are featured by molecular descriptors, including chirality codes, calculated for their three components. Selection of variables, from the initial pool of molecular descriptors, is performed by a genetic algorithm. These are fed to a random forest to predict the enantiomeric excess. A square correlation coefficient of R^2=0.82 and RMSE=9.96 are obtained for the test set, and the results for cross-validation of the whole dataset (in the out-of-bag procedure) are R^2=0.79 and RMSE=10.96. The method can be helpful for computer-aided design of enantioselective catalysts.

■ INTRODUCTION

Usually, a pair of enantiomers exhibit different physical, chemical and biological activity. In order to reduce known or unknown side effects of the non-functional enantiomer, the needs of pure chiral materials are growing [1, 2], e.g. for the pharmaceutical and agrochemical industries. The organic synthetic chemists continually try to design new methods for asymmetric synthesis, and catalytic asymmetric reactions are much desired to pursue high enantioselectivity [3].

Development of new chiral catalysts has an extremely high cost, although the combination

of high-throughput experiments and the combinatorial chemistry can accelerate the process of catalyst synthesis and catalyst evaluation [4]. However, a large number of experiments are still required. The integration of computer-aided virtual screening in catalyst development can reduce the number of experiments and the corresponding costs.

Raint and co-workers [5] tested a combinatorial database of chiral catalysts for asymmetric hydrogen transfer to acetophenone by high-throughput experiments and assessed the performance of a catalyst with a normalized performance factor (NPF), which was calculated on the basis of the experimental yield and enantiomeric excess of the catalyst. The catalysts with the top ten NPF were regarded as the best. In the second step, the genetic algorithm was used to screen experimentally the catalysts in vitro. At the beginning of the evolution, a population was composed of a small part of the library. From generation to generation, some catalysts with lower NPF were replaced by the other catalysts. Finally, when ca. 10% of the library had been investigated by genetic algorithm, about 60% of the ten best catalysts could be obtained.

In a previous work, the authors' laboratory reported the application of chirality codes and neural networks to screen virtually the catalysts of the same library to improve the hit rate of catalysts [6]. Like in the work of Riant [5], the catalysts were then assessed by the value of NPF and the top ten catalysts were regarded as the best catalysts. The library was divided into ca. 10% catalysts as training set and ca. 90% catalysts as test set. The Counterpropagation Neural Network (CPG NN) was trained by the chirality codes and NPFs of the catalysts in the training set, and then the trained CPG NN was used to predict NPFs of the catalysts in the test set. The catalysts in the test set were ranked decreasingly based on the predicted NPFs. It was observed that a selection of ca. 20% of the virtual library enabled to identify up to 85.5% of the ten target catalysts.

The major goal of both methods above was to select catalysts possessing high performance with less experiments, but not to predict the quantitative values of the performance.

In the present article we report a study to build mathematical models to predict quantitatively enantiomeric excess of an asymmetric reaction from the structure of the catalyst.

Aires-de-Sousa and Gasteiger have predicted the enantiomeric excess for a combinatorial library of enantioselective reactions performed by addition of diethyl zinc to benzaldehyde [7]. The small library was composed of 5 chiral catalysts and 13 chiral additives. There are not

many reports available in the literature on quantitative structure-enantioselectivity relationships of catalysts [1].

■ METHODS

Data Set. The experimental data in this article were retrieved from the literature [5] and will be briefly introduced in this section. Catalysts with a yield lesser than 5% have no values of enantiomeric excess and were removed. Thus, the data set used in this article contained 296 chiral catalysts for the asymmetric hydrogen transfer to acetophenone by isopropanol in the presence of potassium hydroxide. The reaction is shown in Figure 1. (R)-1-phenylethanol and (S)-1-phenylenthanol are the products of the reaction.

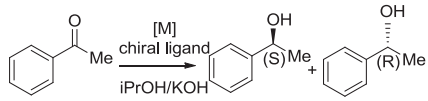

Figure 1. Metal-catalyzed hydride transfer to acetophenone.

The enantiomeric excess is the percentage of (R)-1- phenylethanol minus the percentage of (S)-1-phenylenthanol:

$$ee = ([R]-[S])/([R]+[S]) * 100$$

where [R] is the reaction yield of products with R configuration and the [S] is the reaction yield with S configuration.

The catalysts are synthesized from three components. Component A is an enantiopure β-amino alcohol; component B is an aldehyde or ketone; component M is a metal complex precursor as shown in Figure 2. The general scheme is displayed in Figure 3.

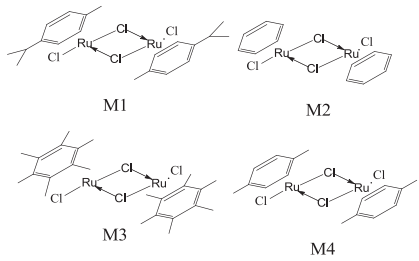

Figure 2. The set of metallic precursors M1 - M4.

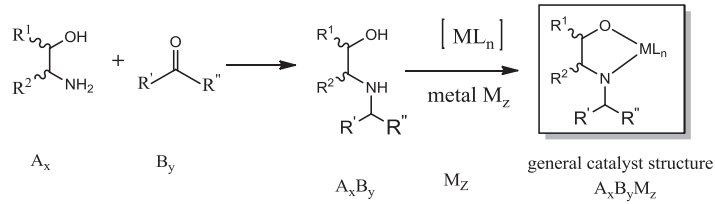

Figure 3. Synthesis of chiral catalysts AxByMz.

The 296 catalysts include 127 of type AxByM1, 83 AxByM2, 7 AxByM3 and 79 AxByM4. In order to build a stable prediction model, the 296 chiral catalysts were divided into 237 (80%) as training set and the remaining 59 (20%) catalysts as test set.

Radial Distribution Function Code. The carbonyl compounds (components B) were represented by a radial distribution function (RDF) code, which was originally developed by Gasteiger et al. and used to describe the 3D structure of the molecule [8]. The definition of RDF is briefly introduced as follows.

$$g(x) = f\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} a_i a_j e^{-b(x-r_{ij})^2} \qquad (1)$$

In this equation, a_i and a_j are the properties of atoms i and j such as atomic charge, r_{ij} is the distance between the atoms i and j, and b is a smoothing parameter. The value of x is a running variable for the function $g(x)$. The f is an optional scaling factor. The g (x) is calculated at a number of discrete points with defined intervals to obtain the same number of descriptors to any sample. The dimension number of RDF code is independent of molecular size.

A molecular encoding scheme has been developed on the basis of eq (1) and such a representation of the 3D structure of a molecule has been utilized for the simulation of infrared spectra [9].

Conformation-Independent Chirality Code. The amino-alcohols (component A) are enantiopure chiral compounds and were encoded by the conformation-independent chirality code (CICC) [10]. Here only a brief explanation of the CICC code is given. First, a value of e_{ijkl} is defined through eq (2) that considers atoms i, j, k, and l, each of them belonging to a different ligand of a chiral center.

$$e_{ijkl} = \frac{a_i a_j}{r_{ij}} + \frac{a_i a_k}{r_{ik}} + \frac{a_i a_l}{r_{il}} + \frac{a_j a_k}{r_{jk}} + \frac{a_j a_l}{r_{jl}} + \frac{a_k a_l}{r_{kl}} \qquad (2)$$

where a_i is a property of atom i, such as atomic charge, and r_{ij} is a distance between atoms i and j. In order to consider the 3D structure but make the chirality code independent of a specific conformer, r_{ij} is taken as the sum of the bond lengths between atoms i and j on the path with the minimum number of bond counts.

Furthermore, the chirality sign, s_{ijkl}, can be assigned a value of +1 or -1. For the computation of s_{ijkl}, atoms i, j, k, and l are ranked according to their decreasing atomic property a_i. The 3D coordinates of A, B, C, and D, which are the corresponding atoms directly bonded to the chiral center, are used for atom i, j, k, and l, respectively. A plane is defined by the first three atoms (in the order established by ranking above), and the fourth atom is behind the plane. If the order is clockwise, s_{ijkl} will take the value of +1; or else, s_{ijkl} will take the value of -1.

Then, the e_{ijkl} and s_{ijkl} are combined into eq (3) to generate the conformation-independent chirality code f_{CICC}.

$$f_{CICC}(x) = \sum_{i}^{n_i} \sum_{j}^{n_j} \sum_{k}^{n_k} \sum_{l}^{n_l} S_{ijkl} \cdot e^{-b\left(x - e_{ijkl}\right)^2} \qquad (3)$$

where n_i, n_j, n_k, and n_l are the number of atoms belonging to ligands i, j, k, and l, respectively; b is a smoothing factor. In practice, b controls the width of the peaks obtained by a graphical representation of $f_{CICC}(x)$ vs x.

The $f_{CICC}(x)$ is calculated at a number of discrete points with defined intervals. The number of descriptors is not dependent on the size of a molecule. The actual range of x used in an application is chosen according to the range of atomic properties related to the range of observed interatomic distances for the given molecules.

Molecular descriptors of AxByMz catalysts. Generation of molecular descriptors for components A, B and M was performed separately, and then combined into a representation for catalysts AxByMz.

The Cartesian coordinates of the atoms in a molecule were calculated from the connection tables of the molecules by the 3D structure generator CORINA [11, 12, 13, 14]. The physicochemical atomic properties were calculated using fast empirical methods included in the program package PETRA [15, 16]. The calculation of chirality codes was performed by using a program developed by Aires-de-Sousa and Gasteiger [10]. The process is as follows:

(1) Amino-alcohols (A), all of which were chiral compounds, were encoded by conformation-independent chirality code. A code possesses 51 values by using: (a) the partial atomic charge as atomic property; (b) all the atoms including hydrogen; (c) values of x in the interval [-0.03e2Å-1, 0.03e^2 $Å^{-1}$]; (e) the smoothing parameter b set to (code length/range of $x)^2$.

(2) Carbonyl compounds (B), most of which were achiral compounds, were encoded with RDF descriptors using the following parameters: (a) the partial atomic charge was the atomic property; (b) the range was [0, 15 Å] for *x*; (c) the length for a code was 63; (d) the smoothing parameter *b* was equal to 100.

(3) Metal precursors (M) included 4 possible metal complexes, and were encoded by a binary vector with four bits. The M1, M2, M3, and M4 were represented by “1,0,0,0”, “0,1,0,0”, “0,0,1,0” and “0,0,0,1”, respectively.

(4) A 51-dimensional CICC code for A, a 63-dimensional RDF vector for B, and a 4-dimensional vector derived from M were combined into a 118-dimensional vector to represent a catalyst. Twelve elements of the vector with constant zero values were deleted. Finally, a 106-dimensional vector was obtained for characterizing a catalyst AxByMz.

Classification and Regression Tree (CART). A regression tree was suitable for this research, in order to make quantitative predictions of the enantiomeric excess. A single regression tree [17] was sequentially constructed, partitioning objects from a parent node into two child nodes. Each node is produced by a logical rule, defined for a single variable (a variable of chirality code), where objects below a certain variable's value fall into one of the two child nodes, and objects above fall into the other child node. The prediction for an object reaching a given terminal node is obtained by the average of enantiomeric excesses of the catalysts (in the training set) reaching the same terminal node. The entire procedure comprises three main steps. First an entire tree is constructed by data splitting into smaller nodes; each produced split is evaluated by square errors of enantiomeric excess which decreases as long as the new split permits child node's square errors to be smaller than parent node. Second, a set of smaller, nested trees is obtained by obliteration (pruning) of certain nodes of the tree obtained in the first step. The selection of the weakest branches is based on sum of square errors of all terminal nodes that decides which subtree, from a set of subtrees with the same number of terminal nodes, has the lowest (within node) error. Finally, from the set of all nested subtrees, the tree giving the lowest value of error in cross-validation (where the set of objects used to

grow the tree is different from the prediction set) is selected as the optimal tree. In this study, a tree was grown with the R program version 2.10.1 [18] using the RPART library with the default parameters.

Random Forest. A random forest (RF) [18, 19] is an ensemble of unpruned classification and regression trees created by using bootstrap samples of the training data and random subsets of variables to define the best split at each node. Prediction is made by the average of the individual trees. Additionally, performance is internally assessed with the prediction error for the objects left out in the bootstrap procedure (internal cross-validation or out-of-bag estimation). Therefore, about 1/3 of training set was randomly selected as validation set and the other 2/3 or so were used to training a tree, and the prediction of validation set was used to assess the tree. Random Forests were grown with the R program, version 2.10.1, using the Random Forest library [18]. The RFs were trained to predict enantiomeric excess on the basis of chirality code of the catalyst. The number of trees in a random forest was set to 1000. It has been shown that the method is extremely accurate in a variety of applications [20].

Genetic Algorithms. Some variables of the chirality code may not be relevant for our purposes, and can even introduce noise. Models with few descriptors are usually preferred for increased robustness. Here we used genetic algorithms for the selection of 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 [21].

At the beginning of the evolution, the individuals are randomly generated. The probability of selecting a variable into subset (randomly for each subset) is between 0 and 0.4.

A population of individuals is allowed to evolve over a number of (300) 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. The offspring result from crossover of their parents, followed by random mutation. The population of the next generation consists of the new offspring and their parents.

The evaluation (scoring) of each individual is made by a Counter Propagation neural network (CPG NN) that uses the subset of molecular descriptors for predicting the enantiomeric excesses. The NN is trained with the training set, and the score of the subset of molecular descriptors is the root-mean-square of errors for the predictions obtained for the

training set. The individuals (subsets of descriptors) giving lower errors are considered to be fitter than those giving higher errors and are selected for mating.

■ RESULTS AND DISCUSSION

Prediction of enantiomeric excesses for the chiral catalysts by the regression tree. The molecular descriptors and enantiomeric excesses of the catalysts of the training set were used to train a classification and regression tree. The obtained tree is shown in Figure 4.

In Figure 4, variables, *X*(1)~*X*(106), are the elements of the 106-dimensional vector of molecular descriptors. Among them, *X*(1)~X(40) are the CICC codes of component A, *X*(41)-*X*(102) are the RDF codes of component B, and *X*(103)- *X*(106) are the indicator variables of component M. From Figure 4, it can be seen that nine variables were adopted to split the notes in the regression tree, among them *X*(2), *X*(3), *X*(6), and *X*(19) are the codes of CICC derived from A, *X*(57), *X*(59), and *X*(70) are the RDF codes from B and *X*(103) and *X*(104) are the indicator variables from M1 and M2. The nine selected variables show that all A, B and M play an important role to establish the prediction model. The indicator variable derived from M4 was not included, although 79 of 296 catalysts were synthesized from M4.

In the root node, *N*=237 means that the number of catalysts in the training set is 237, and *Y*=40 denotes the value of the node is 40, which is the average value of enantiomeric excesses of 237 catalysts. The root node was split by variable X(6). If *X*(6) of a catalyst is smaller than 0.09082, the catalyst falls into the left node, otherwise the catalyst falls into the right node. As a result, 237 catalysts in the training set were partitioned into 190 catalysts in the left node with *Y*=31.8 and 47 catalysts in the right node with *Y*=73.14.

When a catalyst is submitted to the trained classification and regression tree in Figure 4, the catalyst will finally fall into a terminal. The value of the terminal (Y) is the prediction value (enantiomeric excess in this article) of the catalyst. For training set and test set, the results were obtained with square correlation coefficients of R^2=0.71 and R^2=0.56, respectively.

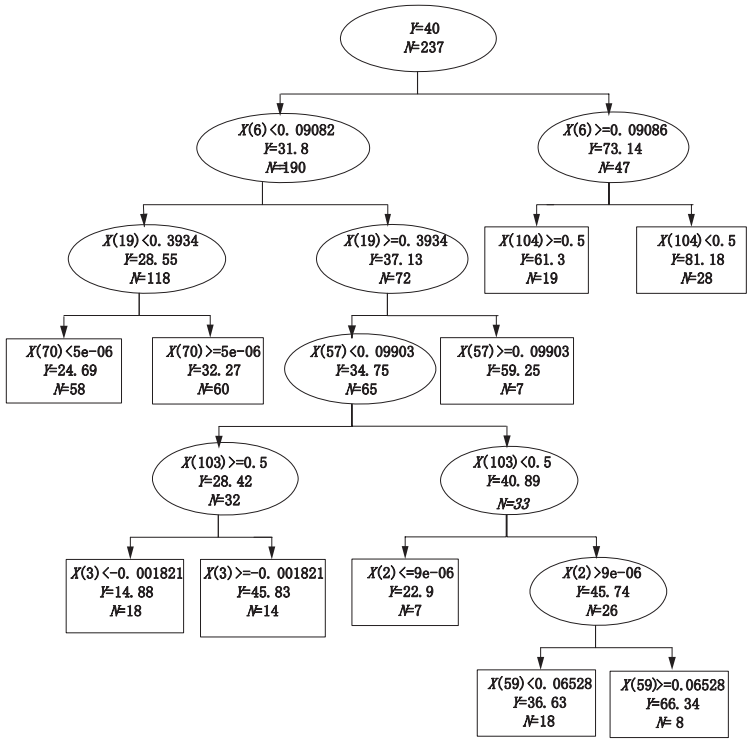

Figure 4. Graphical representation of the classification and regression tree for the prediction of enantiomeric excess of catalyst. Ovals represent the nodes and rectangles denote the terminals in the tree. The annotations in the nodes and in the terminals are the inequations of the conditions and the enantiomeric excess (*Y*), respectively.

Prediction of enantiomeric excesses for the chiral catalysts by random forest. In order to build more stable model, a random forest was trained with molecular descriptors and enantiomeric excesses of training set. The prediction results of the training set and the test set are shown in table 1.

Table 1. Square correlation coefficients (R^2) obtained by random forest

Methods [a]	OOB of training set (R^2)	Prediction of test set (R^2)	OOB of the whole data set (R^2)
RF	0.71	0.77	
GA + RF	0.77	0.82	

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

Table 1 shows that the result of the test set is R^2=0.77 and the result of cross-validation (OOB) of training set is R^2=0.71. The results were improved significantly comparing to CART.

If all the 296 catalysts were used to build prediction model, the result of OOB was R^2=0.74.

In addition, 5-fold cross-validation was also performed, so five prediction models were constructed with random forests. The result of cross-validation was the same as with OOB of the whole dataset (R^2=0.74).

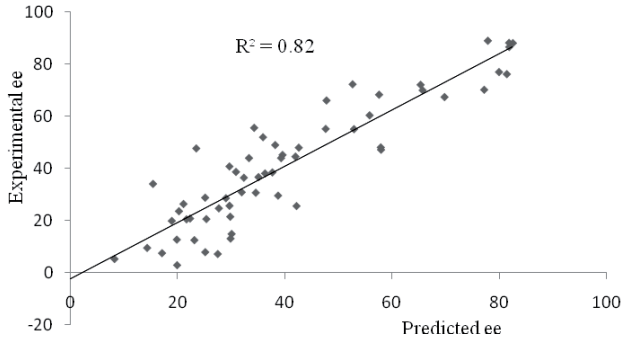

Figure 5. Experimental enantiomeric excesses vs. predicted enantiomeric excesses

In order to obtain more robust and accurate models, the variable selection of chirality codes was performed by genetic algorithm. After the selection, the subset of the molecular descriptors contained 18 variables of CICC codes of component A and 6 variables of RDF of component B. The M1-M4 were always added into subset, whatever they were selected or not by genetic algorithm. Finally, the subset was composed of 18 + 6 + 4 = 28 variables instead of 108. The result of OOB of the training set was R^2=0.77 and for the test set a value of R^2=0.82 and RMSE=9.96 were obtained (see figure 5). If the whole dataset were used to train random forest, the results of OOB were R^2 =0.79 and RMSE=10.96. These results are also presented in Table 1.

The data set of 296 catalysts included 127 AxByM1, 83 AxByM2, 7 AxByM3 and 79 AxByM4. Thus, separate RF prediction models were built for AxByM1, AxByM2, and 79 AxByM4 (the number of AxByM3 catalysts were not enough). The datasets were randomly partitioned into training set with 80% of the catalysts and test set with the remaining 20% of the catalysts. The results for the respective prediction models are displayed in Table 2.

Table 2. Square correlation coefficients of the predicted vs. observed enantiomeric excesses for 127 AxByM1, 83 AxByM2 and 79 AxByM4.

Methods [a]	Data set	OOB of training set (R^2)	^{2})	OOB of the whole data set (R^2)
RF	$AxByM_1$	0.78	0.87	0.83
	$AxByM_4$		0.66	0.61
GA + RF	$AxByM_1$	0.82	0.87	0.83
	$AxByM_2$	0.80	0.73	0.79
	$AxByM_4$	0.62	0.68	0.63

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

Table 2 shows that the best predictions were achieved by the data set of catalysts synthesized with M1. For the model constructed by random forest, the OOB predictions with the training set yielded R^2=0.87 and for the prediction of test set R^2=0.87 and RMSE=8.47. If all 127 catalysts were used as training set, the results of OOB were R^2=0.83 and RMSE=9.91. The worst prediction results, ca. R^2=0.60, were obtained with the data set of catalysts derived from M4.

Similarly, genetic algorithm was used to select the variables - 12 to 27 selected variables were used to construct the prediction model by using random forest. The same steps were performed as above and the results are also listed in Table 2. From Table 2, we can see that the results were slightly improved after reducing the number of the variables by using genetic algorithm.

In addition, the whole data set of 296 catalysts were repartitioned into a training set with 228 catalysts synthesized by cyclic aldehydes and ketones and a test set with 68 catalysts synthesized by acyclic aldehydes and ketones, i.e., the structures in the test set did not appear in the training set. The prediction models were constructed by random forest and genetic algorithm as mentioned above and the results are shown in Table 3. Even with the more challenging partition of the data, results are similar to those in Table 1.

Table 3. Square correlation coefficients (R^2) obtained using the 228 catalysts synthesized from cyclic aldehydes and ketones as the training set, and the 68 catalysts synthesized from acyclic aldehydes and ketones as the test set.

Methods [a]	OOB of training set (R^2)	Prediction of test set (R^2)
RF	0.72	
GA + RF	0.79	0.79

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

■ CONCLUSION

In this study, a library of 296 chiral catalysts for the asymmetric hydrogen transfer to acetophenone was investigated by using the methods of quantitative structure-activity relationships (QSAR). The molecular descriptors of the catalysts, including chirality codes, were submitted to classification and regression tree or random forest to construct the prediction models for the quantitative prediction of enantiomeric excesses, which is one of the most important factors to evaluate the performance of catalysts in asymmetric reactions. The results reveal that the molecular codes derived from the structure of catalysts can be applied to construct robust predictive models of the catalytic performance. It demonstrates a chemoinformatics method with the potential to screen virtually chiral catalysts in asymmetric reactions, which is valuable to reduce time and cost of the development of asymmetric catalysts.

■ ACKNOWLEDGRMENT:

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

■ REFERENCES

[1] D. Rio, Exploring enantioselective molecular recognition mechanisms with chemo-informatic techniques, *J. Sep. Sci.* **32** (2009) 1566–1584.

[2] O. Pàmies, J. E. Bäckvall, Combination of enzymes and metal catalysts. A powerful approach in asymmetric catalysis, *Chem. Rev.* **103** (2003) 3247–3262.

[3] C. Gennari, U. Piarulli, Combinatorial libraries of chiral ligands for enantioselective catalysis, *Chem. Rev.* **103** (2003) 3071–3100.

[4] J. G. de Vries, L. Lefort, The combinatorial approach to asymmetric hydrogenation: Phosphoramidite libraries, ruthenacycles, and artificial enzymes, *Chem. Eur. J.* **12** (2006) 4722–4734.

[5] N. Vriamont, B. Govaerts, P. Grenouillet, C. de Bellefon, O. Riant, Design of a genetic algorithm for the simulated evolution of a library of asymmetric transfer hydrogenation catalysts, *Chem. Eur. J.* **15** (2009) 6267–6278.

[6] 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, *Chemometr. Intell. Lab. Sys.* **109** (2011) 113–119.

[7] J. Aires de Sousa, J. Gasteiger, Prediction of enantiomeric excess in a combinatorial library of catalytic enantioselective reactions, *J. Comb. Chem.* **7** (2005) 298–301.

[8] J. Gasteiger, J. Sadowski, J. Schuur, P. Selzer, L. Steinhauer, V. Steinhauer, Chemical information in 3D space, *J. Chem. Inf. Comput. Sci.* **36** (1996) 1030–1037.

[9] M. C. Hemmer, V. Steinhauer, J. Gasteiger, Deriving the 3D structure of organic molecules from their infrared spectra, *Vib. Spectrosc.* **19** (1999) 151–164.

[10] J. Aires de Sousa, J. Gasteiger, New description of molecular chirality and its application to the prediction of the preferred enantiomer in stereoselective reactions, *J. Chem. Inf. Comput. Sci.* **41** (2001) 369–375.

[11] J. Sadowski, J. Gasteiger, From atoms and bonds to three-dimensional atomic coordinates: Automatic model builders, *Chem. Rev.* **93** (1993) 2567–2581.

[12] J. Gasteiger, C. Rudolph, J. Sadowski, Automatic generation of 3D-atomic coordinates for organic molecules, *Tetrahedron Comput. Method.* **3** (1990) 537–547.

[13] J. Sadowski, C. Rudolph, J. Gasteiger, The generation of 3D-models of host-guest complexes, *Anal. Chim. Acta*. **265** (1992) 233–241.

[14] J. Sadowski, J. Gasteiger, G. Klebe, Comparison of automatic three-dimensional model builders using 639 X-Ray structures, *J. Chem. Inf. Comput. Sci.* **34** (1994) 1000–1008.

[15] J. Gasteiger, M. Marsili, Iterative partial equalization of orbital electronegativity – A rapid access to atomic charges, *Tetrahedron* **36** (1980) 3219–3228.

[16] J. Gasteiger, H. Saller, Calculation of the charge distribution in conjugated systems by a quantification of the resonance concept, *Angew. Chem. Int. Ed. Engl.* **24** (1985) 687-689.

[17] L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, *Classification and Regression Trees*, Chapman & Hall/CRC, Boca Raton, 2000.

[18] R Development Core Team R: A language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria. ISBN 3-900051-07-0. URL http://www.R-project.org.

[19] L. Breiman, Random forests, *Machine Learning* **45** (2001) 5–32.

[20] 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, *J. Chem. Inf. Comput. Sci.* **43** (2003) 1947–1958.

[21] A. V. Homeyer, Evolutionary algorithms and their applications in chemistry, in: J. Gasteiger, J. Engel, (Eds.), *Handbook of Chemoinformatics*, Wiley-VCH, New York, 2003, pp. 1239–1280.