



## Virtual screening prediction of new potential organocatalysts for direct aldol reactions

Xiang Hui Liu<sup>a</sup>, Hong Yan Song<sup>b,c</sup>, Xiao Hua Ma<sup>a</sup>,  
Martin J. Lear<sup>b</sup>, Yu Zong Chen<sup>a,\*</sup>

<sup>a</sup> *Bioinformatics and Drug Design Group, Centre for Computational Science and Engineering, National University of Singapore, Blk S16, Level 8, 3 Science Drive 2, Singapore 117543, Singapore<sup>1</sup>*

<sup>b</sup> *Department of Chemistry, Faculty of Science, and Medicinal Chemistry Group of the Life Sciences Institute, National University of Singapore, 3 Science Drive 3, Singapore 117543, Singapore*

<sup>c</sup> *Institute of Materials Research and Engineering, A\*STAR, 3 Research Link, Singapore 117602, Singapore*

### ARTICLE INFO

#### Article history:

Received 30 July 2009

Received in revised form 6 December 2009

Accepted 16 December 2009

Available online 23 December 2009

#### Keywords:

Aldol reaction

Asymmetric catalysis

Organocatalyst

Support vector machine

Virtual screening

### ABSTRACT

A support vector machine (SVM)-based virtual screening method is demonstrated as a rapid computational tool for the prediction of potential asymmetric organocatalysts for the direct aldol reaction. Our models show good accuracy at cross-validation and independent testing. Structure analyses of screening hits from the PubChem database revealed several new classes of compounds, including  $\beta$ -amino acids, diamines and hydrazides, as potential chiral organocatalysts.

© 2009 Elsevier B.V. All rights reserved.

### 1. Introduction

Organocatalysts have emerged as robust, inexpensive, readily available and non-toxic catalysts for asymmetric synthesis. Most organocatalysts have been discovered through serendipity or through empirical trial and error. Today computational methods are rapidly becoming a versatile tool for the rationalization and prediction of organocatalysts [1]. A computational model that is capable of evaluating large compound libraries at high speed and guiding the discovery of new types of organocatalysts, is highly desirable in the rational design of organocatalysts. Quantum mechanics (QM) is a well recognized computational method for organocatalyst design, which has successfully confirmed the enamine mechanism for proline catalysis of directed aldol reactions [2–9]. QM, however, is highly consuming of computational resources, requiring months to evaluate a compound library. This greatly limits its general use. To reduce computational times, several alternative methods have been developed. These include quantitative structure selectivity relationships (QSAR) [10–12], reverse docking [13], and the asymmetric catalyst evaluation (ACE)

program [14]. QSAR works well only for same series of compounds to the known classes of organocatalysts. The other two methods are based on molecular mechanic (MM) techniques and need additional assumptions or simplifications which might always be true [1].

Herein, we describe a new machine learning approach termed Support Vector Machine (SVM)-based virtual screening method. SVM is a ligand-based machine learning method based on statistical learning theory [15], which has consistently delivered good predictions in the area of drug design. In drug discovery and organocatalyst design, a common problem in SVM-based screening studies is the lack of negative compounds. This causes an imbalance between positive and negative compounds in the dataset and frequently leads to a high false positive rate for dataset screening [16–22]. As a solution to this problem, we have developed a new putative negative-generation process to specifically augment the negative datasets (the putative negatives) [23,24]. Studies have shown that SVM classification models derived from such putative negatives can perform reasonably well in virtual screening studies [23–28]. For this study, the organocatalytic aldol reaction was chosen as a model system because of its importance in asymmetric carbon–carbon bond formation in organic synthesis. Being well documented in the organocatalysis field, the asymmetric aldol reaction was anticipated to offer new insights into similar reactions that share common enamine-type catalytic intermedi-

\* Corresponding author. Tel.: +65 6874 6877; fax: +65 6774 6756.

E-mail address: [phacyz@nus.edu.sg](mailto:phacyz@nus.edu.sg) (Y.Z. Chen).

<sup>1</sup> <http://bidd.nus.edu.sg>.

ates. Herein, we tested the SVM-based virtual screening method as a rapid screening tool for the identification of new, potential organocatalysts for the direct aldol reaction.

Known organocatalysts that form asymmetric enamine intermediates in the direct, intermolecular aldol reaction may be classified into the following classes: proline and amide analogs (**1a** and **1b**), non-proline amino acid and amide analogs (**2a** and **2b**), diamines (**3a–g**), proline-tetrazoles (**4**), phosphoric-prolines (**5**),  $\beta$ -amino acids (**6**), binaphthyl analogs (**7**), simple amines (**8a–c**) and imidazolidinones (**9**) (Fig. 1). Our work was divided into two studies. In Study 1, we validated the SVM approach to organocatalyst design, whereby a model was built based on the known amino acid and amide analogs (Class 1 and Class 2 compounds). The predictive ability of this model was then tested against other classes of organocatalysts (Fig. 1). In Study 2 we next investigated what new classes could be identified with a model built on all currently known classes.

## 2. Materials and methods

In total, 195 unique organocatalysts that displayed an ee over 30% were collected from over 60 papers (see structures in supporting information 1). They were divided into a training set (156 organocatalysts reported before June 2007) and an independent testing set (39 organocatalysts reported after June 2007). Among the 156 organocatalysts in the training set, 117 belonged to Class 1 and Class 2. They were used as positives (organocatalysts) for the training set in Study 1, while all 156 organocatalysts were used as positives for the training set in Study 2. As anticipated, there

were few negatives (non-organocatalysts) being reported in the literature. Virtual negatives were thus generated using our putative negative-generation method [23,24]. The whole process can be divided into five main steps.

First, the PubChem 13.6 M compounds were calculated with 100 2D descriptors and clustered using the *K*-means method to represent the whole chemical space. The 100 2D descriptors include: molecular weight; the number of atoms, bonds, rings, H-bond donor/acceptors, rotatable bonds, N or O heterocyclic rings; the number of C, N and O atoms; the charge polarization; and the Kier shape index (see supporting information 2). The *K* value of the compound clusters was set as 9000. This *K* value is consistent with the 12,800 compound-occupying neurons (regions of topologically close structures) for 26.4 million compounds of up to 11 atoms of C, N, O, F [29], as well as the 2851 clusters for 171,045 natural products [30]. The final number of clusters was 8423 after the *K*-means clustering. Second, the known positives were mapped to the clusters. Those clusters containing known positives were defined as active families. Other clusters were defined as non-active families. The putative negatives were generated by taking eight representative samples from each of the non-active families. Third, the software LibSVM was chosen to perform the machine learning. Non-linear SVM separates the positives from the negatives with a hyperplane by mapping the input vectors to a higher dimensional feature space using a kernel function (Fig. 2). The Radial Basis Function (RBF) kernel, widely adopted to consistently give better performance, was used in this study. Optimally, the hard margin SVM ( $c=100,000$ ) was used with a  $\sigma$  scan between 0 and 15 for best performance, as determined from the fivefold cross-validation results. Fourth, a

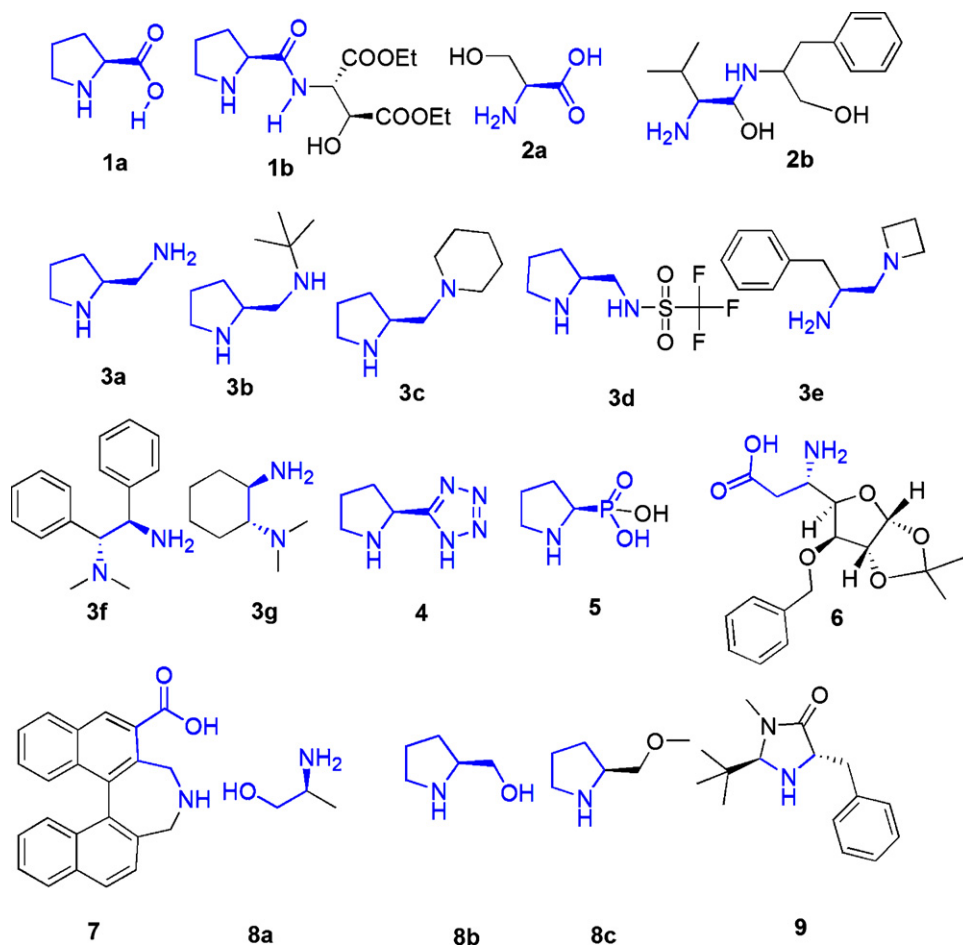


Fig. 1. Examples of known organocatalysts for direct intermolecular aldol reaction.

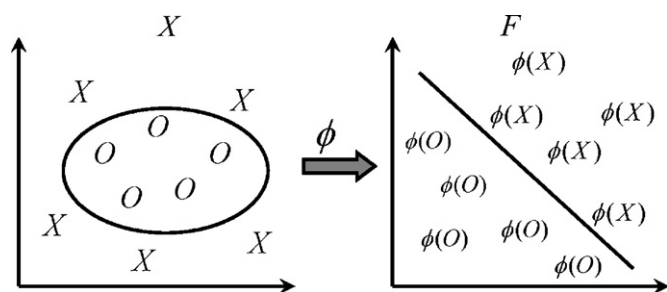


Fig. 2. SVM separation of positive and negative datasets.

model was built with all training compounds at this  $\sigma$ . The model was then tested using the independent testing set. Fifth and finally, the PubChem database was screened.

### 3. Results and discussion

The SVM results are summarized in Table 1. In the five-fold cross-validation test,  $\sigma$  values were found to be 0.2 and 0.7 for Studies 1 and 2, respectively, whereby the SVM models gave prediction accuracy values averaging 70.2% and 71.8% for organocatalysts and averaging 99.9% for non-organocatalysts (see supporting information 2). This confirmed the quality of our models for virtual screening. Moreover, the independent test, using 39 unique organocatalysts reported after June 2007, showed 41% and 64% accuracy. Screening of the 13.6 M PubChem compounds revealed 16,996 and 24,714 compound hits. This constituted 0.125% and 0.182% of the whole PubChem database, respectively. To undergo enamine catalysis, organocatalysts should be primary or secondary chiral amines capable of reacting with carbonyl groups. Thus, achiral compounds were first removed, which reduced the hit numbers to 10,847 and 13,688; compounds possessing no primary or secondary amine group were then removed, which reduced the hit numbers to 7284 and 10,471 for Studies 1 and 2, respectively.

To understand the virtual screening results, structure-class analysis was carried out to determine the structural types of the screening hits. The number of substructure query hits served as a rough evaluation of the potential of this structure-class analysis as organocatalysts. The screening results of various structure-classes (Fig. 1) are listed in Table 2. Although being built exclusively on Classes 1 and 2 organocatalysts, our structure analysis from model Study 1 showed a clear predictive ability for most other classes of organocatalysts. Hits of all these nine classes (Nos. 1–19) comprise up to 70% of the hits of Study 1. The diamine (Nos. 5–11),  $\beta$ -amino acid (No. 14) and amine (Nos. 16–18) classes presented the greatest number of hits, besides Classes 1 and 2. This result clearly validated our SVM-based virtual screening method as a useful tool that can reliably predict new classes of organocatalysts; however, our model failed to predict the proline-tetrazole (No. 12) and some diamines

(Nos. 7, 8 and 11) as organocatalysts. This latter observation may be due to the practical requirement of adding acid during the synthetic use of these organocatalysts, whereby ammonium species form strong hydrogen bonds to the aldehyde carbonyl group.

In Study 1, the model generated was further used to evaluate various other substructure types (Nos. 20–28) previously not studied or largely unstudied in practice. The detailed results are listed in Table 2. The majority of the predictions were consistent with practical experimental reports: the cyclohexane-1,2-diamines (No. 20) are active organocatalysts, while their sulphanamide (No. 21) and amide or substituted analogs (No. 22) are not [31]; small groups like methyl (No. 23) are allowed [32], yet large groups like naphthyl (No. 24) are not favored for  $\alpha$ -substituted proline analogs [33]; proline-tetrazole analogs like the proline benzimidazoles (No. 25) and amino acid tetrazoles are also active [34–36]. Importantly, our models predicted two new types of  $\alpha$ -amino acids (Nos. 26 and 27) as being potentially active in the catalytic asymmetric aldol reaction. Although there are no experimental reports of these two types for aldol reaction, they have been reported as organocatalysts of two similar reactions. The  $\beta$ -amino acids **2** (No. 26) catalyze the Hajos–Parrish–Eder–Sauer–Wiechert reaction [37], and the  $\beta$ -amino acids **3** (No. 27) act as organocatalysts of the Mannich reaction [38]. These two types of reactions share similar enamine catalysis mechanisms to the proline-catalyzed aldol reaction. Like for the proline-tetrazole and diamines, the proline-pyrimidine (No. 28) was falsely predicted [39], probably due to the requirement for ammonium hydrogen-bonding in the predominating catalytic step. Collectively, these structure-class analyses demonstrate our models as a useful means to guide the discovery of new organocatalysts.

In Study 2, our SVM-based virtual screening method was applied to build a model based on all currently known 9-classes of compounds. As shown in Table 2, about 60% of all hits belonged to the known classes. The remaining hits were structural new, but similar to the known classes listed in Fig. 1 (see structural examples in supporting information 2). Fig. 1 illustrates multiple types of substructures (as marked in blue). For enamine-based catalytic activity, an amine group is a clear prerequisite. To ensure asymmetric catalysis, however, there should be additional interactions (e.g. hydrophobic, hydrogen bonding or even steric interactions) between the homochiral organocatalyst and the reacting carbonyl compound to generate a predominant diastereomeric complex. We believe that the success of our models stemmed from the chosen descriptors and the reliability of our putative negative-generation method. Specifically, SVM classifies compounds based on the discriminative properties as represented by descriptors between organocatalysts and non-organocatalysts, rather than purely on structural similarities or substructure motifs. It is difficult to unambiguously define which particular descriptor sub-set of the 100 2D descriptors is exactly responsible for the asymmetric catalysis. However, several descriptors, including the number of hydrogen bond donors and acceptors, number of amine groups, hydrophobic effects, and molecular shape, all play an important role in describing

Table 1  
Results of SVM-based virtual screening models.

No.	Training dataset		Fivefold cross-validation		Virtual screening performance		
	P <sup>a</sup>	N <sup>b</sup>	SE <sup>c</sup> (%)	SP <sup>d</sup> (%)	Ind <sup>e</sup> (%)	Virtual hits <sup>f</sup>	Final hits <sup>g</sup>
1	117	66646	70.2	99.9	41	16,996	7284
2	156	66470	71.8	99.9	64	24,714	10,471

<sup>a</sup> Number of positives (organocatalysts).

<sup>b</sup> Number of putative negatives (non-organocatalysts).

<sup>c</sup> Average sensitivity (the prediction accuracy for organocatalyst) of fivefold cross-validation.


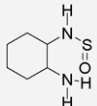
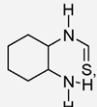

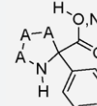




<sup>d</sup> Average specificity (the prediction accuracy for non-organocatalyst) of fivefold cross-validation.

<sup>e</sup> Prediction result of independent test set.

<sup>f</sup> Number of screening hits of PubChem 13.6 M compounds.

<sup>g</sup> Number of virtual hits after clean-up.

**Table 2**  
Structure-class analysis.

No.	Structure-class	Example structure <sup>a</sup>	Study	
			1	2
1	Proline-acid	<b>1a</b>	234	288
2	Proline-amide	<b>1b</b>	319	316
3	Amino acid	<b>2a</b>	2406	2172
4	Amino acid amide	<b>2b</b>	820	617
5	Diamine	<b>3a</b>	5	30
6	Diamine	<b>3b</b>	96	160
7	Diamine	<b>3c</b>	0	53
8	Diamine	<b>3d</b>	0	1
9	Diamine	<b>3e</b>	18	521
10	1,2-Diphenyl-ethane-1,2-diamine	<b>3f</b>	7	77
11	Cyclohexane-1,2-diamine	<b>3g</b>	0	5
12	Proline-tetrazole	<b>4</b>	0	3
13	Proline-phosphate	<b>5</b>	2	3
14	$\beta$ -Amino acid 1	<b>6</b>	328	371
15	Binaphthyl analogs	<b>7</b>	0	0
16	Simple amine 1	<b>8a</b>	703	873
17	Simple amine 2	<b>8b</b>	43	29
18	Simple amine 3	<b>8c</b>	9	20
19	Imidazolidinone	<b>9</b>	0	0
20	Cyclohexane-1,2-diamine		5	180
21	Cyclohexane-1,2-diamine sulphanamide		0	0
22	Cyclohexane-1,2-diamine amide		1	0
23	$\alpha$ -Substituted proline (methyl)		24	26
24	$\alpha$ -Substituted Proline (phenyl)		0	0
25	Proline-tetrazole analogs		2	13
26	$\beta$ -amino acid 2		13	15
27	$\beta$ -amino acid 3		1	32
28	Nicotine		0	0

<sup>a</sup> Example structures **1a–9** are shown in Fig. 1.

the discriminative features between organocatalysts and non-organocatalysts. From this study, we also know that descriptors defining chirality and amine-functionality are particularly useful for enamine-type catalysis and shall be added to our descriptor list in the future. Moreover, the putative negatives method of generation ensures that diverse types of non-organocatalysts will be enumerated across the whole chemistry space of PubChem, whereby our SVM model has great potential to remove the non-organocatalysts.

#### 4. Conclusions

We demonstrate a new SVM-based virtual screening method for the rapid screening of organocatalysts from a large compound library (PubChem). Our study validates this method as a convenient way to identify potential organocatalysts in a good hit-yield, low false hit rate, and in a timely, low-resource manner. While being capable of expanding current collections of organocatalysts, our method shows promise in the identification of new structural types of organocatalysts. Several types of screening hits including  $\beta$ -amino acids, diamines and hydrazides are anticipated to serve as good lead structures for future Aldol studies in both a computational (QM) and practical (synthetic) sense. Moreover, the identified screening hits also possess potential in reactions, including the Mannich,  $\alpha$ -amination,  $\alpha$ -aminoxylation, and the Morita Baylis-Hillman reactions, all of which share a common enamine-type catalysis mechanism to the aldol reaction. Since descriptors can be readily customized to known catalytic mechanisms and proposed catalytic complexes, we anticipate our SVM-based virtual screening method to find wide applicability in asymmetric synthesis and organocatalyst design.

#### Acknowledgement

This work was supported in part by R-148-000-081-112/101.

#### Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at [doi:10.1016/j.molcata.2009.12.008](https://doi.org/10.1016/j.molcata.2009.12.008).

#### References

- [1] K.N. Houk, P.H. Cheong, *Nature* 455 (2008) 309–313.
- [2] S. Bahmanyar, K.N. Houk, *J. Am. Chem. Soc.* 123 (2001) 11273–11283.
- [3] S. Bahmanyar, K.N. Houk, *J. Am. Chem. Soc.* 123 (2001) 12911–12912.
- [4] K.N. Rankin, J.W. Gauld, R.J. Boyd, *J. Phys. Chem. A* 106 (2002) 5155–5159.
- [5] Z. Tang, F. Jiang, L.T. Yu, X. Cui, L.Z. Gong, A.Q. Mi, Y.Z. Jiang, Y.D. Wu, *J. Am. Chem. Soc.* 125 (2003) 5262–5263.
- [6] C. Allemann, R. Gordillo, F.R. Clemente, P.H. Cheong, K.N. Houk, *Acc. Chem. Res.* 37 (2004) 558–569.
- [7] F.R. Clemente, K.N. Houk, *Angew. Chem. Int. Ed. Engl.* 43 (2004) 5765–5768.
- [8] F.R. Clemente, K.N. Houk, *J. Am. Chem. Soc.* 127 (2005) 11294–11302.
- [9] C.B. Shinisha, R.B. Sunoj, *Org. Biomol. Chem.* 5 (2007) 1287–1294.
- [10] S. Chavali, B. Lin, D.C. Miller, K.V. Camarda, *Comp. Chem. Eng.* 28 (2004) 605–611.
- [11] B. Lin, S. Chavali, D.C. Miller, K.V. Camarda, *Comp. Chem. Eng.* 29 (2005) 337–347.
- [12] S. Sciabola, A. Alex, P.D. Higginson, J.C. Mitchell, M.J. Snowden, I. Morao, *J. Org. Chem.* 70 (2005) 9025–9027.
- [13] D.J. Harriman, G. Deslongchamps, *J. Mol. Model* 12 (2006) 793–797.
- [14] C.R. Corbeil, S. Thielges, J.A. Schwartzentruber, N. Moitessier, *Angew. Chem. Int. Ed. Engl.* 47 (2008) 2635–2638.
- [15] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York/London, 1995.
- [16] G. Harper, J. Bradshaw, J.C. Gittins, D.V.S. Green, A.R. Leach, *J. Chem. Inform. Comput. Sci.* 41 (2001) 1295–1300.
- [17] L. Franke, E. Byvatov, O. Werz, D. Steinhilber, P. Schneider, G. Schneider, *J. Med. Chem.* 48 (2005) 6997–7004.
- [18] R.N. Jorissen, M.K. Gilson, *J. Chem. Inf. Model.* 45 (2005) 549–561.
- [19] M. Glick, J.L. Jenkins, J.H. Nettles, H. Hitchings, J.W. Davies, *J. Chem. Inf. Model.* 46 (2006) 193–200.
- [20] Z. Lepp, T. Kinoshita, H. Chuman, *J. Chem. Inf. Model.* 46 (2006) 158–167.
- [21] J. Hert, P. Willett, D.J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby, A. Schuffenhauer, *J. Chem. Inf. Model.* 46 (2006) 462–470.
- [22] B. Chen, R.F. Harrison, G. Papadatos, P. Willett, D.J. Wood, X.Q. Lewell, P. Greenidge, N. Stiefl, *J. Comput. Aided Mol. Design* 21 (2007) 53–62.
- [23] L.Y. Han, X.H. Ma, H.H. Lin, J. Jia, F. Zhu, Y. Xue, Z.R. Li, Z.W. Cao, Z.L. Ji, Y.Z. Chen, *J. Mol. Graph. Model.* 26 (2008) 1276–1286.
- [24] X.H. Ma, R. Wang, S.Y. Yang, Z.R. Li, Y. Xue, Y.C. Wei, B.C. Low, Y.Z. Chen, *J. Chem. Inf. Model.* 48 (2008) 1227–1237.
- [25] C.Z. Cai, L.Y. Han, Z.L. Ji, X. Chen, Y.Z. Chen, *Nucleic Acids Res.* 31 (2003) 3692–3697.
- [26] H.H. Lin, L.Y. Han, C.Z. Cai, Z.L. Ji, Y.Z. Chen, *Proteins Struct. Funct. Genet.* 62 (2006) 218–231.
- [27] L.Y. Han, C.Z. Cai, Z.L. Ji, Z.W. Cao, J. Cui, Y.Z. Chen, *Nucleic Acids Res.* 32 (2004) 6437–6444.
- [28] L.Y. Han, C.J. Zheng, B. Xie, J. Jia, X.H. Ma, F. Zhu, H.H. Lin, X. Chen, Y.Z. Chen, *Drug Discovery Today* 12 (2007) 304–313.
- [29] T. Fink, J.-L. Raymond, *J. Chem. Inf. Model.* 47 (2007) 342–353.
- [30] M.A. Koch, A. Schuffenhauer, M. Scheck, S. Wetzels, M. Casaulta, A. Odermatt, P. Ertl, H. Weldmann, *Proc. Natl. Acad. Sci. U.S.A.* 102 (2005) 17272–17277.
- [31] M. Kui, S.L. Zhang, S.T. He, P. Li, M. Jin, F. Xue, G.S. Luo, H.Y. Zhang, L.R. Song, W.H. Duan, W. Wang, *Tetrahedron Lett.* 49 (2008) 2681–2684.
- [32] K. Sakthivel, W. Notz, T. Bui, C.F. Barbas 3rd, *J. Am. Chem. Soc.* 123 (2001) 5260–5267.
- [33] L. Cheng, X. Wu, Y. Lu, *Org. Biomol. Chem.* 5 (2007) 1018–1020.
- [34] E. Lacoste, Y. Landais, K. Schenk, J.B. Verlhaca, J.M. Vincent, *Tetrahedron Lett.* 45 (2004) 8035–8038.
- [35] K.R. Reddy, G.G. Krishna, C.V. Rajasekhar, *Synth. Commun.* 37 (2007) 4289–4299.
- [36] E. Lacoste, E. Vaique, M. Berlande, I. Pianet, J.M. Vincent, Y. Landais, *Eur. J. Org. Chem.* 1 (2007) 167–177.
- [37] S.G. Davies, A.J. Russell, R.L. Sheppard, A.D. Smith, J.E. Thomson, *Org. Biomol. Chem.* 5 (2007) 3190–3200.
- [38] H. Pellissier, *Tetrahedron* 63 (2007) 9267–9331.
- [39] T.J. Dickerson, K.D. Janda, *J. Am. Chem. Soc.* 124 (2002) 3220–3221.