COMMUNICATION

ChemComm
www.rsc.org/chemcomm

Computational screening of combinatorial catalyst libraries†

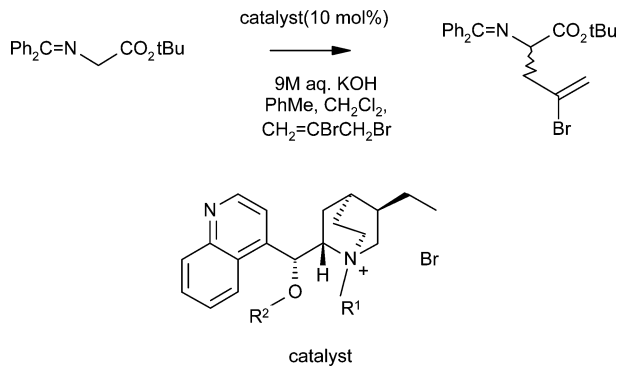
James L. Melville, Benjamin I. Andrews, Barry Lygo and Jonathan D. Hirst*
School of Chemistry, University of Nottingham, University Park, Nottingham, UK NG7 2RD.
E-mail: jonathan.hirst@nottingham.ac.uk; Fax: 44 115 951 3562; Tel: 44 115 951 3478

Received (in Cambridge, UK) 17th February 2004, Accepted 28th April 2004
First published as an Advance Article on the web 17th May 2004

A catalyst design methodology, utilizing combinatorial synthesis in parallel with chemometric analysis, is presented, which considers the 3D steric and electrostatic properties of substituents about a constant core structure.

High-throughput screening has revolutionized drug discovery;[1] the field of catalyst discovery and optimization is poised to undergo an analogous upheaval. However, due to the large number of possible compounds that can be synthesized, computational approaches to guide synthetic efforts are needed. In the area of asymmetric catalysis, some first steps in this direction have recently been published, but these have concentrated on building and analysing the entire catalyst structure.[2] An advantage of combinatorial chemistry is that a small number of reactants can be combined to form a large number of products. In this report, we describe a method of constructing a three-dimensional Quantitative Structure-Selectivity Relationship (QSSR), based around the Comparative Molecular Field Analysis (CoMFA) methodology,[3] that is focused on the substituents of a common catalytic core. By modelling substituents rather than the individual catalysts as a whole, very large reductions in complexity and computational effort can be achieved. These enhancements in efficiency increase with the size of the libraries studied, thereby allowing the assessment of larger catalyst libraries than is possible with the techniques used to date.

For a test case we examine the asymmetric alkylation shown in Scheme 1. This process has wide utility in the preparation of α-amino acid derivatives,[4,5] and high levels of enantioselectivity have been reported using readily-available dihydrocinchonidine-derived catalysts.[6] Recently it has been established that these catalysts can be generated *in situ* through sequential *N*- and *O*-alkylation of the parent alkaloid and in this way diversity can be introduced at the points marked R^1 and R^2.[7] Using this approach, a library of 88 catalysts was synthesized in parallel *via* combination of the 13 substituents specified in Scheme 2. Substituents **1**, **3–7**, **12** and **13**



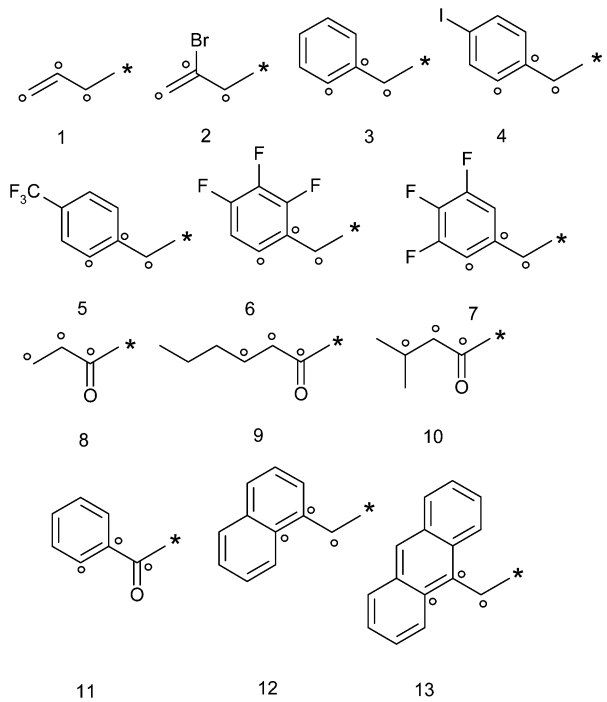
Scheme 1 The phase-transfer catalyzed synthesis of 2-amino-4-bromopent-4-enoic acid. Diversity is introduced to the dihydrocinchonidine-based catalyst core at the sites marked R^1 and R^2.

† Electronic supplementary information (ESI) available: predicted and observed e.e. values for the 88 catalysts in the library; full CoMFA parameters; aligned molecular coordinates. See http://www.rsc.org/suppdata/cc/b4/b402378a/

DOI: 10.1039/b402378a

could be used to alkylate the quaternary N atom and substituents **1–11** were added to the O atom. Enantioselectivity was assessed by HPLC. The measured selectivities of the catalysts are presented in Tables S1 and S2 of the Supplementary Information†. For the purposes of constructing the QSSR, the enantioselectivities were converted to the ratio of the amount of *S* and *R* enantiomer formed. The natural log of this value was taken, which is linearly proportional to ΔG, and thus suitable for fitting with a linear regression technique. In a conventional CoMFA, the 3D structure of each molecule must be built and aligned. However, in this case, there is a constant molecular core for all the catalysts. As only the *differences* in the structures leads to a difference in selectivity, we consider only the substituents. This has two main benefits: the number of structures to be built and aligned is reduced from 88 to 13 and we do not include any descriptors related to the core structure, which could introduce "noise" to the model, masking the effect of the ligands and reducing the quality of the predictions. Such an approach has been advocated in the field of drug design,[8] but harnessing combinatorial chemistry with substituent-focused CoMFA has yet to be applied to catalyst design.

3D structures of the substituents given in Scheme 2 were built in SPARTAN PC Pro version 1.0.8 (Wavefunction, Inc. Irvine, CA). The open valence where the substituents connect to the core was capped with an H atom. These structures were optimized using the MMFF force-field.[9] Repeating optimizations with an *ab initio* calculation at the BLYP/6-31G* level gave a very similar resulting QSSR, indicating that a molecular mechanics optimization is



Scheme 2 The 13 substituents that make up the 88-member catalyst library. The asterisk shows the point of attachment to either the O or quaternary N of the cinchona-based core. Atoms marked with a circle were used for the alignment step of the CoMFA.

adequate. The structures must then be superposed. We chose a simple alignment rule: a least squares fit of the three C atoms on the substituent that were closest to the core of the catalyst, as marked in Scheme 2. For the substituents attaching to the O atom, the alignment template was substituent **3**; for the N atom, it was substituent **13**. These templates were chosen because the presence of those substituents resulted in the most selective catalysts, on average. Each of the aligned substituents was then placed in a rectilinear lattice, large enough to contain the largest substituent. An sp^3 C probe atom with a single positive charge was placed at each lattice point and the electrostatic and steric interactions between the probe and the substituent were computed by means of the Coulomb and Lennard-Jones equations, respectively. Partial charges were assigned to the atoms of the substituents using the Gasteiger-Hückel method[10] and the Lennard-Jones parameters were taken from the TRIPOS force-field.[11] Complete CoMFA parameters are given in Table S3 of the Supporting Information†.

Each catalyst is then represented by the electrostatic and steric descriptors of its two substituents. These are correlated with the observed log selectivity by building a regression model using partial least squares (PLS).[12] The dimensionality of the model is determined by leave-one-out (LOO) cross-validation,[13] which also assesses the quality of the predictions. The discrepancy between the observed and predicted values is used to define a standard error of cross-validation, SE_{CV}. The number of components corresponding to the first minimum of SE_{CV} is chosen as the final model dimensionality, which is built using all observations. The quality of the predictions is measured by the cross-validated coefficient of determination, q^2. For a dataset of this size, a value of $q^2 > 0.6$ is indicative of a model with good predictive properties. The conventional coefficient of determination, R^2, is measured using the fitted, rather than cross-validated values. If the value of R^2 is much higher than that of q^2, over-fitting of the model must be suspected, which is likely to result in poor predictions. The PLS regression and statistics were calculated using an in-house program written in C++. We constructed a virtual catalyst ("training") library using all the substituents in Scheme 2, except substituent **4**, which we removed to create a test set. Principal component analysis of the catalyst descriptors showed that none of the compounds made with substituent **4** represents an extrapolation for a regression model made with the other 12 substituents, so it represents a reasonable external test of the model, as dictated by statistical best practice. However, in practical application for molecular design, the obtained model will be robust to mild extrapolation from the model, in order to discover catalysts that are more selective than those included in the training set.

CoMFA was applied to the 70 catalyst training library, resulting in a five component model with a conventional R^2 of 0.82 and a LOO q^2 of 0.72. In terms of %ee, these values represent a root mean square (RMS) error of fit and cross-validation of 10% and 13%, respectively. These values indicate that the model has good predictive qualities and the proximity of the q^2 and R^2 values gives confidence that no over-fitting has occurred. Fig. 1 shows cross-validated *versus* observed selectivities, converted back into %ee. To test further that the observed correlation is not due to chance, the selectivities were scrambled, so that the observed selectivities are associated with the wrong catalyst substituents. A PLS model was built and the q^2 recorded. 100 of these scramble sets were carried out. The resultant q^2 values (mean: −0.11; standard deviation: 0.08; maximum: 0.14; minimum: −0.31) are clearly inferior to the results obtained with the real model. The utility of this method as a tool for screening the selectivity of possible catalysts *in silico* is also demonstrated by prediction of the selectivity of the test set. The introduction of substituent **4**, in combination with the other substituents, allows the synthesis of a further 18 catalysts and we used the training model to predict their selectivity. Comparing the predicted and experimentally observed selectivities, we obtained an external q^2, $Q^2_X = 0.69$, corresponding to a %ee RMS error of prediction of 13%. These predictions, comparable to the cross-validated results, are also plotted in Fig. 1 (see also Table S2 in the Supplementary Information†).

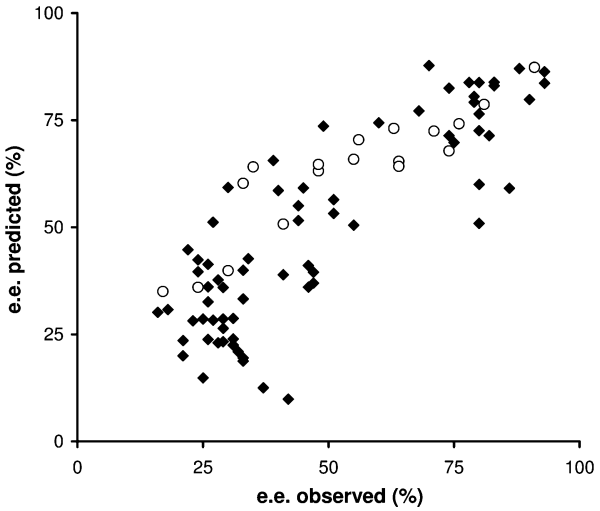


Fig. 1 Plots of the predicted enantiomeric excesses of the catalysts against the observed experimental values. Diamonds represent the predictions of the 70-member training library using substituents **1–3** and **5–13**. The circles show the predicted selectivity of the catalysts formed by adding substituent **4** to the library at the R^1 and R^2 position.

To summarize, good 3D-QSSR models can be created for catalysts with a constant core, by considering only the substituents. We built an accurate model for 70 catalysts (based on only 12 structures), which was used to predict the selectivity of 18 new catalysts by the introduction of only one new substituent to the library. The versatility of the cinchonidine system augurs well for the applicability of this technique to other reaction mechanisms. In general, the method shows promise for use with asymmetric catalysis studies, particularly when combined with other techniques from chemoinformatics (*e.g.* statistical design and database searching).

We thank the Gatsby Foundation for funding for JM and the EPSRC for an equipment grant (GR/R62052/01) for computers.

Notes and references

1 V. Murphy, A. F. Volpe Jr and W. H. Weinberg, *Curr. Opin. Chem. Biol.*, 2003, **7**, 427–433.
2 K. B. Lipkowitz and M. C. Kozlowski, *Synlett*, 2003, 1547–1565 and references therein.
3 R. D. Cramer III, D. E. Patterson and J. D. Bunce, *J. Am. Chem. Soc.*, 1988, **110**, 5959–5967.
4 M. J. O'Donnell, *Aldrichimica Acta*, 2001, **34**, 3–15.
5 B. Lygo and B. I. Andrews, *Tetrahedron Lett.*, 2003, **44**, 4499–4502.
6 See for example: B. Lygo, J. Crosby, T. Lowdon, J. A. Peterson and P. G. Wainwright, *Tetrahedron*, 2001, **57**, 2403–2409; H. G. Park, B. S. Jeong, M. S. Yoo, J. H. Lee, B. S. Park, M. G. Kim and S. S. Jew, *Tetrahedron Lett.*, 2003, **44**, 3497–3500 and references therein.
7 B. Lygo, B. I. Andrews, J. Crosby and J. A. Peterson, *Tetrahedron Lett.*, 2002, **43**, 8015–8018.
8 R. D. Cramer, *J. Med. Chem.*, 2003, **46**, 374–388.
9 T. A. Halgren, *J. Comput. Chem.*, 1996, **17**, 490–519.
10 J. Gasteiger and M. Marsili, *Tetrahedron*, 1980, **36**, 3219–3228; J. Gasteiger and H. Saller, *Angew. Chem., Int. Ed. Engl.*, 1985, **24**, 687–689.
11 M. Clark, R. D. Cramer III and N. Van Opdenbosch, *J. Comput. Chem.*, 1989, **10**, 982–1012.
12 S. Wold, M. Sjöström and L. Eriksson, *Chemom. Intell. Lab. Syst.*, 2001, **58**, 109–130.
13 S. Wold, *Technometrics*, 1978, **20**, 397–405.