DOI: 10.1002/adsc.200900009

A Three-Dimensional Quanititative Structure-Activity Relationship (3D-QSAR) Model for Predicting the Enantioselectivity of *Candida antarctica* Lipase B

Paolo Braiuca,[a,b] Knapic Lorena,[a,b] Valerio Ferrario,[a] Cynthia Ebert,[a] and Lucia Gardossi[a,*]

[a] Department of Pharmaceutical Sciences, Universita' degli Studi di Trieste, P.le Europa 1, 34127 Trieste, Italy
Fax: (+39)-040-52572; e-mail: gardossi@units.it
[b] SPRIN Technologies for Sustainable Chemistry Srl, P.le Europa 1, 34127 Trieste, Italy

Received: January 8, 2009; Revised: April 20, 2009; Published online: June 3, 2009

Abstract: Computational techniques involving molecular modeling coupled with multivariate statistical analysis were used to evaluate and predict quantitatively the enantioselectivity of lipase B from *Candida antarctica* (CALB). In order to allow the mathematical and statistical processing of the experimental data largely available in the literature (namely enantiomeric ratio E), a novel class of GRID-based molecular descriptors was developed (differential molecular interaction fields or DMIFs). These descriptors proved to be efficient in providing the structural information needed for computing the regression model. Multivariate statistical methods based on PLS (partial least square – projection to latent structures), were used for the analysis of data available from the literature and for the construction of the first three-dimensional quanititative structure-activity relationship (3D-QSAR) model able to predict the enantioselectivity of CALB. Our results indicate that the model is statistically robust and predictive.

Keywords: biocatalysis; *Candida antractica* lipase B (CALB); differential molecular interaction fields; enantioselectivity; three-dimensional quanititative structure-activity relationship (3D-QSAR)

Introduction

Recent advances in computational sciences have led to novel sophisticated and refined methods that are able to describe the biocatalyst machinery in detail. Solutions of research problems in molecular modeling of enzymes can be found within different time frames and accuracy levels, which depend on the computational techniques used. Nonetheless, the *in silico* quantitative prediction of enzyme enantioselectivity is still a formidable goal to be reached.[1] A number of studies has faced this problem by searching for the structural differences inducing discrimination between the fast-reacting and the slow-reacting enantiomer,[2] also by means of approaches based on free energy differences between the two tetrahedral intermediates.[3]

Methods able to predict quantitatively the enantioselectivity would have great practical and theoretical impact and they would act as a rational tool, alternative to experimental screening procedures. However, despite the application of computationally intensive simulation protocols and state of the art molecular modeling techniques, very rarely has a real quantitative prediction of enzyme enantioselectivity been achieved so far.

In a pioneering work Tomić and co-workers developed a predictive model for the enantioselectivity of lipase from *Burkholderia cepacia* (BCL) by exploiting 3D-QSAR (quantitative structure-activity relationship) methodology.[4] The strategy was based on the calculation of the energies of interaction between the two enantiomers and selected crucial amino acid residues playing key roles in secondary alcohol enantiodifferentiation. The free energy of the tetrahedral intermediate was calculated by means of a linear combination of interaction energy, polar and non-polar solvent accessible surface, whose relative weights were evaluated by PLS (partial least square) analysis. The models were able to unambiguously predict the fast-reacting enantiomer and the approximate magnitude of the enantioselectivity.

In a recent work we demonstrated that 3D-QSAR can be applied for the quantitative prediction of penicillin G amidase (PGA) selectivity by elaborating statistical models able to correlate the 3D structure of substrates with the selectivity constants (k_{cat}/K_m) of

selected hydrolytic reactions.[5] As in the case of the work of Tomić, the predictive model was based on the idea of correlating the structural features of substrates with their selectivity but in the latter case no energy calculations were utilized. Instead, the correlation was based on the pure geometric interpretation of substrate conformations generated by the docking algorithm, subsequently refined by short molecular dynamics (MD) simulations.

The present work now focuses attention on the enantioselectivity of the lipase B from *Candida antarctica* (CALB): the original computational approach is expanded with the aim of enabling the mathematical and statistical processing of the experimental data largely available in the literature expressed as enantiomeric ratio "E". For this purpose, the predictive models were calculated on the basis of an experimental data set coming from three different works found in the literature, which are constituted by the values of enantiomeric ratio for the resolution of *sec*-alcohols and *sec*-amines catalyzed by CALB.

The resolution of racemic compounds by CALB is one of the most studied biocatalytic systems for which the molecular basis was therefore previously analyzed.[6] The stereochemical preference of this enzyme usually follows an empirical model generally referred to as Kazlauskas rule.[7] Other significant molecular modeling-based studies of the enantioselectivity of CALB have also been done by the groups of Pleiss et al.[8] and Hult et al.[9,10] which, among other things, contributed to elucidate the binding mode of the enantiomers in the stereospecificity pocket of CALB. Even though these results can be a valuable tool for the qualitative prediction of the resolving potential of CALB towards many compounds, they do not offer a quantitative prediction of the enzyme's enantioselectivity which is, in principle, hard to achieve.

The k_{cat}/K_m ratio depends on the free energy of the transition state of the reaction, which is generally calculated either by simplified methods based on molecular mechanics[1] or more refined methods, such as QM/MM and free energy perturbation.[1,11] While the over-simplification of the former methods makes quantitative predictions unfeasible, the latter are definitely much too time-consuming to be attractive as predicting tools and, above all, they often still provide unsatisfactory quantitative accuracy.

The calculation of the 3D-QSAR models by regression analysis implies the correlation between the E values and suitable molecular descriptors. A novel class of molecular descriptors was conceived and calculated to provide quantitative information on the diversity in enzyme-enantiomers interactions. Such descriptors are based on the GRID[12] computational method and were named "differential molecular interaction fields" (DMIFs) since they account for the different interactions established by the two enantiomers inside the active site of the enzyme. Although the overall computational protocol is based on low computational demanding algorithms, the quality of prediction obtained with the 3D-QSAR model resulted to be among the highest reported so far in the literature, thus confirming the potential of the hybrid approach coming from the combination of modeling and multivariate statistics.

Results and Discussion

General Strategy

The construction of the 3D-QSAR model involved a protocol that can generally be divided into four principal stages: a) definition of the training set and refinement of the structures of both the enzyme and the substrates; b) calculation of the active conformers of the substrates by molecular dynamics simulations; c) generation of the molecular descriptors for the couples of enantiomers by means of GRID analysis and DMIFs calculation; d) multivariate statistical analysis of the data and generation of the mathematical predictive model.

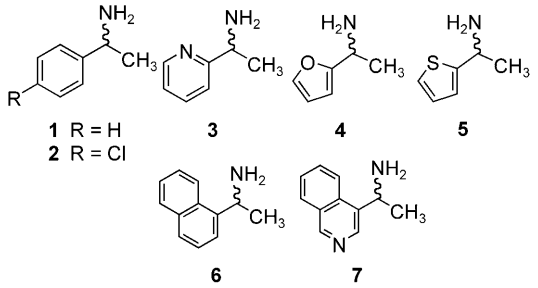
Stages a) and b) were performed by means of molecular mechanics simulations, whereas, for stages c) and d), the GRID analysis was employed in combination with chemometric tools. In particular, the PLS method (partial least squares – projection to latent structures) was used for the calculation of the regression model.

Definition of the Training Set and Refinement of the Structures

The model was constructed on the basis of a training set constituted by seven racemic amines and twelve racemic alcohols and the corresponding values of enantiometic ratio (E) measured experimentally in the enzymatic acylations as reported in three works taken from the literature (Table 1 and Table 2).[7–11] These reactions were selected, among the numerous examples reported in the literature, because they meet the necessary requirements in terms of variability of structures and E values, which are crucial for the generation of a consistent 3D-QSAR model.

Secondary alcohol resolution catalyszd by CALB is probably the most studied biocatalytic system, therefore this supports the reliability of the experimental data and ensures the robustness of the model. The distribution of enantiomeric ratio values throughout the data set is well balanced and structural diversity of nucleohpiles is significant. Some difficult cases are included, such as nucleophiles bearing halogen substi-

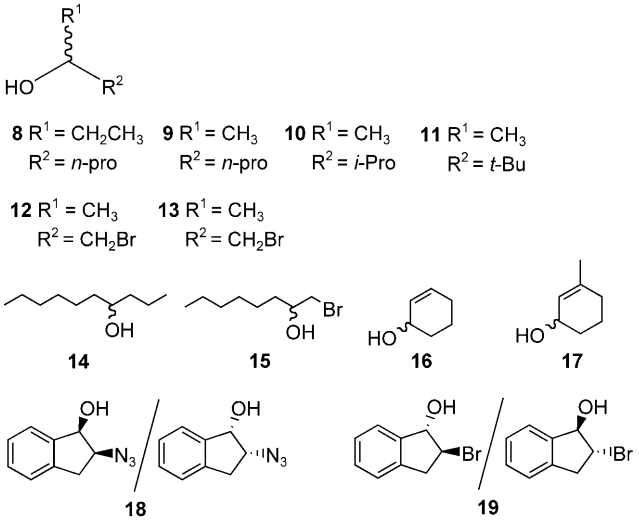
Table 1. Data set for the resolution of the chiral 1-arylethylamines.



Amine	Donor	E[24,25]/F.r.[a]
1	Ethyl acetate	110 (*R*)
2	Methyl methoxyacetate	232 (*R*)
3	Ethyl acetate	66 (*R*)
4	Ethyl acetate	>100 (*R*)
5	Ethyl acetate	32 (*R*)
6	Ethyl acetate	24 (*R*)
7	Ethyl acetate	120 (*R*)

[a] F.r. = fast reacting enantiomer.

Table 2. Data set for the resolution of *sec*-alcohols.



Alcohol	Donor	E[26,27,28]/F.r.[a]
8	Octanoyl acetate	340 (*R*)
9	Octanoyl acetate	8 (*R*)
10	Octanoyl acetate	760 (*R*)
11	Octanoyl acetate	430 (*R*)
12	Octanoyl acetate	100 (*R*)
13	Octanoyl acetate	370 (*R*)
14	Octanoyl acetate	10 (*R*)
15	Octanoyl acetate	7 (*S*)
16	Octanoyl acetate	1.6 (*R*)
17	Octanoyl acetate	62 (*R*)
18	Octanoyl acetate	1.3 (*S,R*)
19	Octanoyl acetate	90 (*S,S*)

[a] F.r. = fast reacting enantiomer.

tution in the medium-sized chain, which are not resolved by CALB because of polarity effects.[13]

The aim of this work was to quantify mostly the effect of the nucleophile structure on E, since it is known that it is generally predominant.[7] As a consequence, only three different acyl donors are considered in the data set. Moreover, the acyl donor acts also as solvent, so that the increase of its chemical diversity might introduce the risk to confound the effect of solvent and enzyme-substrate interactions on the enantioselectivity, rather than increasing the predictivity of the model.

It must also be noted that overextending the data set, namely taking data from many different works, made by different research groups under different experimental conditions, introduces the risk of data unconsistency, chancy correlations and increase of noise in the model. For the evaluation of a QSAR application on similar systems, the priority is to ensure the quality of the experimental data. Only in a second stage can an increase of the data set diversity and complexity be considered in order to expand the application field of the models.

The first and the most delicate step of the study involved the calculation and the assessment of the tetrahedral intermediates for each acylation reaction by molecular modeling techniques. For this purpose, the corresponding esters and amides were docked into the active site of the lipase and the best conformers were chosen on the basis of the results of the docking algorithm scoring function (London dG[14]) as well as by evaluating the geometric compatibility with the initiation of the catalytic mechanism. Different criteria were taken into account during the structural compatibility assessment: *i*) the correct orientation of the acylic and nucleophilic portion of the conformer inside the hydrophilic/hydrophobic pocket of the active site; *ii*) the distance of the catalytic Ser105 from the carbonyl carbon of the substrate, which must be compatible with the nucleophilic attack; *iii*) the correct orientation of the carbonyl oxygen toward the Thr40 and Gly106 that constitute the oxyanion hole.

The tetrahedral intermediates (TI) were then simulated by forming a covalent bond between the hydroxy group of the catalytic serine (Ser105) and the carbonyl carbon of the acylated substrate, thus resulting in the corresponding oxyanions (Figure 1).

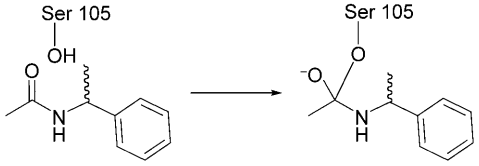


Figure 1. Generation of the tetrahedral intermediate.

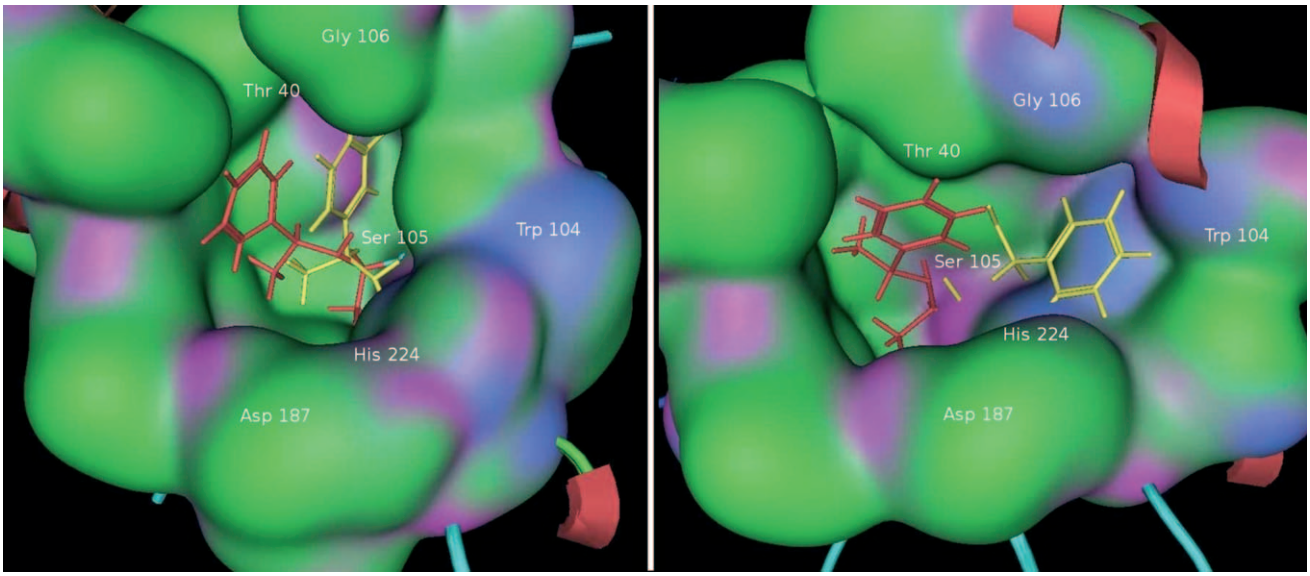


Figure 2. Initial (red) and final (yellow) conformation of the slow-reacting (*S*), on the left, and the fast-reacting (*R*) enantiomer, on the right, of amide **1** in Table 1.

Calculation of the Active Conformers of the Substrates by Molecular Dynamics Simulations

Each of the tetrahedral intermediates was subjected to 300 ps of a molecular dynamics simulation in which only amino acid residues within a 10 Å radius sphere from the catalytic serine (Ser105) were allowed to move. The rest of the protein was kept constrained. The MDs were performed in the NTV ensemble (see Experimental Section) and they generated energy-stable complexes within the first few tens of ps of the simulations. In the case of amides with high E values, important structural differences were observed between the TIs of the fast- and the slow-reacting enantiomers. As shown in Figure 2 for substrate **1**, the TI of the fast-reacting enantiomer is embraced inside the hydrophilic pocket on the right hand portion of the active site (the so-called alcoholic sub-site). On the other hand, the TI of the slow-reacting enantiomer remains at the outer region of the active site which makes the second nucleophilic attack unfeasible.

Another evident discriminating factor is illustrated in Figure 3, which represents the outcome of the MD-based conformational search of the two enantiomers of substrate **8**. In the case of the slow-reacting enantiomer, the minimum energy conformer is not able to perfectly place the oxyanion in the oxyanionic hole, with a consequent energy destabilization as compared to the fast-reacting enantiomer where stabilizing hydrogen bonds take place between the oxyanion and the Thr40 and Gly106 residues of the oxyanion hole.

In this case the MD causes the evolution of the slow-reacting enantiomer towards an unproductive conformation, as defined by the criteria used for the docking scoring. This means that the initiation of the reaction for that enantiomer is unfavorable and consequently the E values is very high. Although this leads to the comparison of productive and unproductive conformations in the QSAR, this dramatic conformational difference is certainly correlated to the high E, so that both the productive and unproductive conformers must be included in the model.

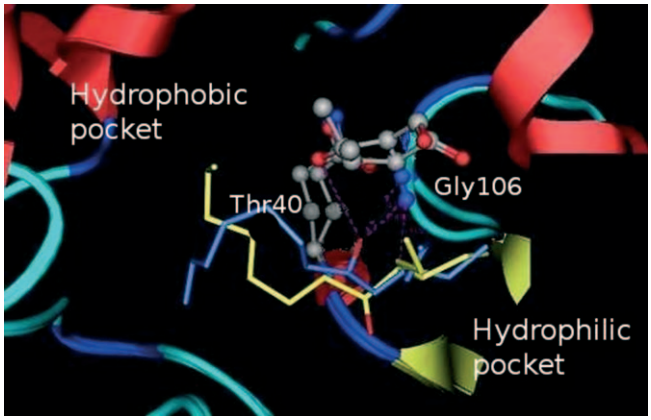


Figure 3. Energy minima conformations of the fast-reacting (blue) and the slow-reacting (yellow) enantiomers for the acylation of compound **8** obtained by MD simulations. The different orientation of the oxyanions (in red) is clearly visible: only the fast-reacting enantiomer is stabilized through the formation of hydrogen bonds (dashed lines) with Thr 40 and Gly 106.

Generation of the Molecular Descriptors for the Couples of Enantiomers by means of GRID Analysis and DMIFs Calculation

The outcome of every MD simulation was carefully analyzed and conformers with the lowest potential energy (as calculated by MD algorithm for the whole unconstrained part of the system, therefore within the active site region) were selected as the best simulations of the different TIs and they were used for the construction of the 3D-QSAR model. The enzyme-substrate complexes were superimposed by overlaying the catalytic triad and the oxyanion hole of all selected configurations. This was necessary because after the MD simulations the Cartesian coordinates of the systems were perturbed, although the conformational changes of the active site residues were always negligible. The protein structures were discarded after the removal of the covalent bond between the substrates and the catalytic serine, while the overall geometry of the substrate conformers was kept unaltered, to generate a so-called "supermolecule", which consisted in all the 38 active conformers (19 enantiomeric couples), both *R* and *S*.

GRID analysis was then performed by setting the dimensions of the grid to contain all the conformers and each of them was analyzed separately. In order to take into account the most important non-covalent interactions, two probes with diverse physico-chemical properties were used in the calculation of the molecular interaction fields, namely the *water* and the *dry* probe. The *water* probe describes and quantifies the dipolar interactions and the hydrogen bond formation, whereas the *dry* probe considers all the hydrophobic interactions.[6]

The basic concept in every QSAR analysis is to associate unambiguously an "activity" value (namely the property of interest) to each constituent of the data set. Since the "activity" considered in this study is the enantiomeric ratio, which is an intrinsic property of a couple of molecules, the molecular interaction fields (MIFs) calculated for every conformer separately miss the correspondence to the above-mentioned property. Therefore, the structural information regarding couples of enantiomers and contained in the respective MIFs had to be merged in order to become single entities. For this purpose, a new class of molecular descriptors was conceived and calculated, which were named "differential molecular interaction fields" (DMIFs). The calculation was performed in a matrix differential procedure where each variable of the MIF of the slow-reacting enantiomer was mathematically subtracted from the corresponding variable of the MIF of the fast-reacting enantiomer (Figure 4). It must be noted that the redundancy of the information contained in the calculated MIFs was cut by operating a "zeroing values pretreatment": all the positive/unfavorable interaction energies were set to zero, because every cluster of positive variables (nodes of the MIFs' grid) has a corresponding cluster of negative variables that contains information that is identical from the statistical point of view.

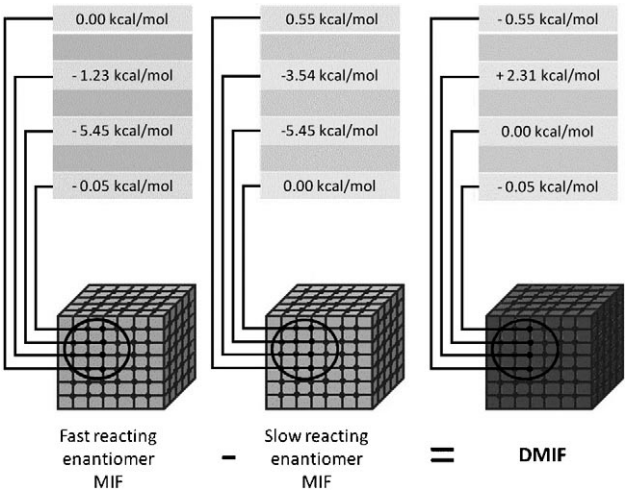


Figure 4. The procedure used for the calculation of the DMIFs taking as example the case of interaction energies between the water probe and the two enantiomers of compound **16**. The mathematical difference between matrices was calculated to generate a single "differential matrix".

This procedure led to the quantitative evaluation of the differences in interactions between the two enantiomers and both polar and hydrophobic regions of the active site. Consequently, DMIFs present null values in the correspondence of areas where the enantiomers establish identical interactions with the active site, whereas high absolute values indicate that the enantiomers establish different interaction patterns with the enzyme (Figure 4).

The DMIF concept is somehow inspired by the GRID/CPCA approach,[15] but while in the GRID/CPCA the original MIFs are used to generate the matrix for a PCA analysis, with the aim of pointing out the differences among the MIFs (corresponding to the analyzed objects), here the matrix is constructed by the differences of the MIFs because the aim is to perform a regression analysis for the prediction of a property (E) which is a function of a couple of objects (the enantiomers).

Multivariate Statistical Analysis of the Data and Generation of the Mathematical Predictive Model

The energy values contained in the "differential matrices" of the DMIFs were statistically analyzed to generate PLS[16] models able to correlate the quantitative differences between the two enantiomers with the experimental E values. The three-dimensional

DMIFs are unfolded to form the so called bi-dimensional X-matrix, where each row corresponds to an enantiomeric couple and each column to a MIF grid node, matching a specific three-dimensional position. Each column of the X-matrix (containing the values of the DMIFs) is an X variable and the enantiomeric ratio E is the Y variable (or the "dependent" variable).

As a matter of fact, molecular interaction fields describe the steric and electrostatic properties of substrates by sampling the interaction energies at all predefined gridpoints. The multitude of gridpoints and, therefore, the quantity of variables present in a DMIF can be extremely high even in the case of small molecules. Moreover, some of these variables are more informative than others. Although the procedure of DMIF calculation halved the number of objects, the number of independent X variables was still unvaried and it amounted to 24,950 for each couple of enantiomers. Therefore, the first stage in the statistical analysis was the choice of the most important variables and the discarding of the insignificant and redundant ones.

For this purpose the GOLPE program was used. GOLPE is a software package largely used for the construction, the validation and the interpretation of 3D-QSAR models. It is particularly adequate for models with large numbers of variables since it has a variety of tools for their selection. Once the DMIFs were calculated, all those variables having very low absolute values were discarded due to their negligible contribution to the quantification of the differences in enzyme-enantiomer interactions. Then, variables with a standard deviation close to zero were discarded as well because of their small variation through all of the DMIFs, that makes them useless in discriminating the objects in the data set. A last action was performed on the remaining active variables by using the "block unscaled weights" algorithm that attributes different weights to all blocks of variables giving them the same initial importance in the model without modifying the variable scale. This latter step was necessary because the polar interaction energies are significantly higher in absolute value than the hydrophobic interaction energies, therefore the statistical analysis would overestimate their importance in the model. Finally, the standard GOLPE procedure was applied on these pre-treated data, by employing the D-optimal pre-selection and the FFD variable selection algorithm which conserved only 568 active variables.[17]

The multivariate statistical analysis was performed on 16 of the initial 19 compounds, by performing the PLS regression and five principal components were calculated. Three compounds, fulfilling the requirement of having small, medium and high E values, were randomly chosen and excluded from the training set. The predictivity of the model was evaluated by means of the leave-one-out (LOO) cross-validation method as well as by performing an external validation using the GOLPE PLS external prediction on the three compounds not included in the training set (Table 3). The predictive correlation coefficient (q^2) provides the quantitative evaluation of the consistency of the model. The best q^2 value for the model is 0.76 on the third principal component and 99 percent of the variance of the model is explained by the first two principal components (expressed by the correlation coefficient r^2).

Table 3. Comparison between the measured experimental E values and the values calculated by the model in the LOO (leave-one-out) cross-validation procedure applied on the training set.

Compound	Experimental E value	Predicted E by LOO
1	110	50
2	232	98
3	66	46
4	100	80
5	32	42
6	24	35
7	120	67
10	760	326
11	430	253
12	100	60
13	370	146
14	10	8
15	7	11
16	1.6	13
18	1.3	7
19	90	73

Although the mathematical model was constructed on the basis of an experimental data set with a broad distribution of E values, the algorithm proved to be quite predictive and robust as illustrated in Figure 5. The worst predictions are represented by compounds **16** and **18** that are, however, characterized by extremely low E values (1.6 and 1.3. respectively). Because of these two compounds the model appears to be more predictive towards compounds having higher E values. It is an intrinsic property of any QSAR model to be more robust for the compounds in the middle of the activity range, simply because this zone is usually more populated. Nevertheless, in all cases the model is able to identify correctly the fast-reacting enantiomer and, more importantly, to recognize those couples of enantiomers characterized by poor enantiodiscrimination (for substrates **16** and **18** the calculated E values are <15 in both cases).

The external validation was performed on three additional compounds (**8**, **17, 9**) not originally included in the training set that were chosen due to their low, intermediate and high E values, respectively. For every compound the complete procedure was repeat-

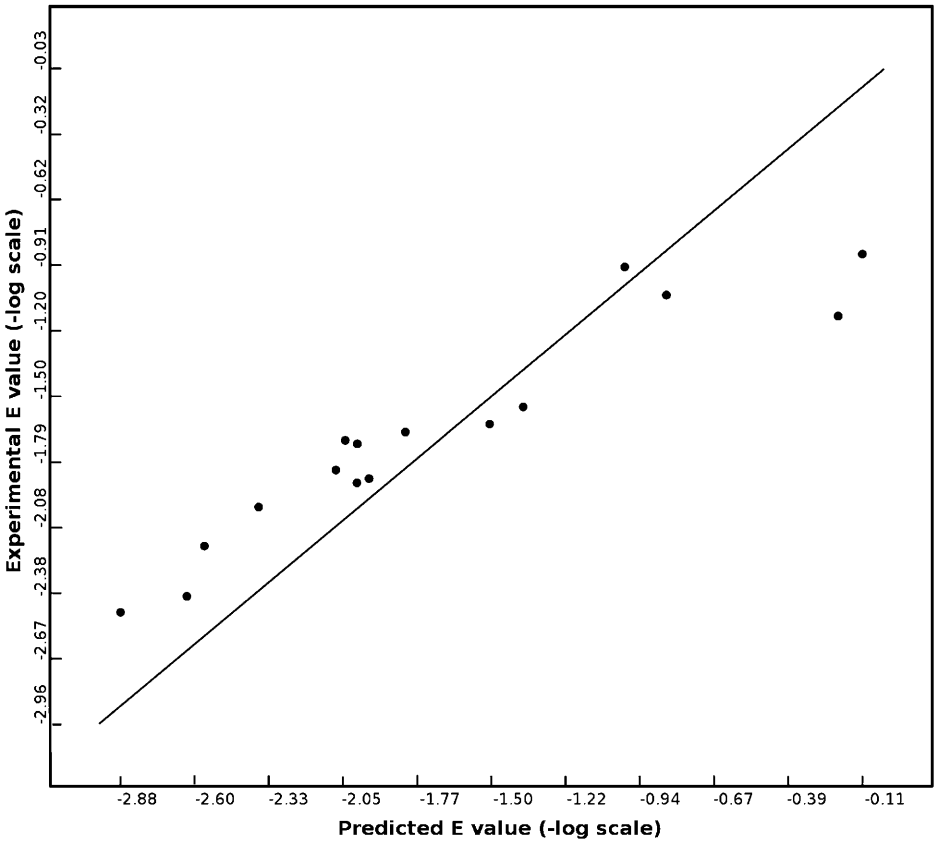


Figure 5. Predictivity of the model in terms of experimental versus predicted E values.

ed, as described above, in order to generate the molecular descriptors (DMIFs). Their E values were then predicted by applying the generated 3D-QSAR model. As it can be seen from Table 4, although the model predicts with good precision the ability of the enzyme to enantiodiscriminate within a couple of enantiomers (predictivity espressed as $q^2=0.78$), in the case of compound **8** the E value is underestimated. This underestimation was observed also for compounds **2**, **10**, **11**, **13** and it might suggest that the variables crucial for structural discrimination for substrates having low E values are different as compared to variables describing substrates with high E values. In other words, the 3D-QSAR model is trained on the basis of a pattern of interactions which are actually different as compared to those taking place in the case of substrates with high E values, and this might limit the predictivity of the PLS model. It should be noted that in the case of compounds **2** and **13** the underestimation given by the model could be ascribed to the presence of the halogen substituent, whose polar character might be measured with insufficient precision by the water probe. Even though this probe can adequately estimate polar interactions that are not correlated to hydrogen bonding, the halogen atoms might not be described comprehensively by the force field parameterization of the water probe.

Table 4. External validation of the calculated PLS model ($q^2=0.78$) towards compounds not included in the training set.

Compound	Experimental E value	E calculated by the model
8	340	105
17	62	46
9	8	8

Therefore, a second model, specifically trained for the prediction of high E values, was calculated in order to refine the quantitative predictivity of E values for those enantiomers that are efficiently enantiodiscriminated by the enzyme. The new data set was constructed by setting the value of $E=50$ as a threshold since E values lower than 50 correspond to enantiomers poorly enantiodiscriminated (examples are compounds **3**, **5**, **6**, **15** in Table 5).

Indeed, the predictivity of this second model improved ($q^2=0.88$) and, as expected, the same model was less efficient in predicting the E values for copules of enantiomers that are poorly enantiodiscriminated (data not shown).

This second "specialized" model is based on a larger number of variables as compared to the first

Table 5. Data set used for the calculation of the second model and for its validation in terms of predicted E values by the LOO (leave-one-out) cross-validation procedure ($q^2 = 0.88$).

Compound	Experimental E value	Predicted E by LOO
1	110	114
2	232	218
4	100	134
7	120	156
8	340	340
10	760	480
11	430	360
12	100	112
13	370	275
19	90	130

general model (618 instead of 568 of the general model) and its increased predictivity suggests that the variables involved in the two models are substantially different, not only quantitatively.

It must be noted that each variable corresponds to a specific grid point, therefore to a specific Cartesian coordinate in the active site of the enzyme. To understand the differences between the two models in deeper detail, each single variable was analyzed and its position in the space refolded. The two models share nearly 25% of the variables (125 variables), while they differ for the rest of them. A detail of the analyzed space with the spatial position of included variables is represented in Figure 6. It is evident that in the first general model the crucial variables are scattered throughout the active site, whereas in the second "specialized" model the crucial interactions are concentrated in the oxyanion hole and in the alcoholic subsite (central and the right-hand part of the molecule). The analysis confirms what emerged from the conformational analysis: for substrates characterized by high E value the slow-reacting enantiomer either cannot place the oxyanion into the oxyanionic hole or it cannot place the alcoholic moiety inside the corresponding subsite.

As a rule of thumb, when dealing with the prediction for a new substrate, the first general model should be used to obtain an initial classification of the CALB enantiodiscriminating potential. If the first model predicts a high E value, the second specialized model should be used for obtaining a more refined quantitative prediction.

Conclusions

The combination of molecular modeling with multivariate statistics constitutes a powerful tool for predicting and also interpreting the enantioselectivity of biocatalysts. The remarkable flexibility of this "hybrid" computational tool makes it adaptable to the solution of different problems as well as to the investigation of the molecular basis of enantiodiscrimination. By definition, the success of any 3D-QSAR strategy depends strongly on the experimental data set used for the training of the mathematical model. Moreover, the generation of the PLS model heavily relies on the selection of the most informative variables of the whole data set. This statistical procedure is of fundamental importance and constitutes one of the bases of the QSAR paradigm.

Concerning the time scale of the whole computational procedure, a whole PLS model including a data set of about 20 compounds, can be developed in about one week using standard low-end computational facilities. Once the model is available, screening of substrates requires approximately one hour per molecule. However, times can be heavily reduced by increasing the computational power, since the conformational analysis represents the most time-consuming step of the protocol.

In conclusion, optimal 3D-QSAR models for the *in silico* screening of a set of substrates must be devel-

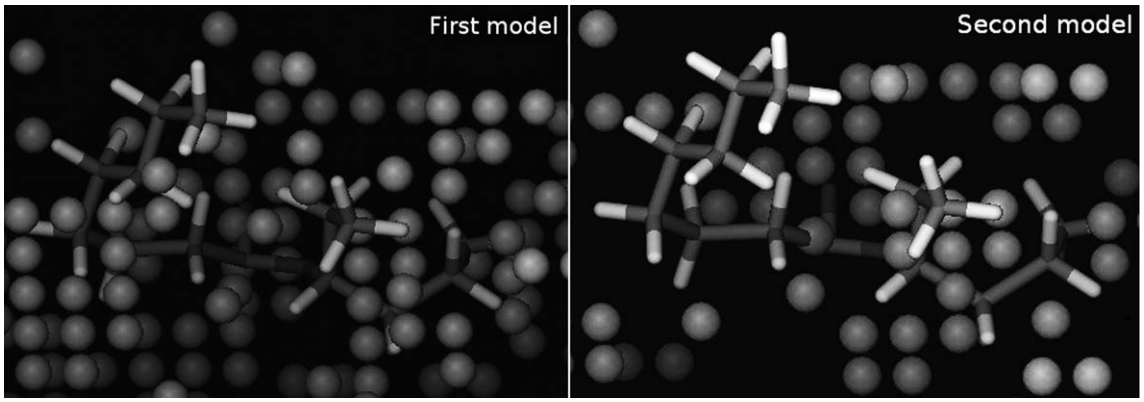


Figure 6. Representation of variables utilized for the construction of the first general predictive model (*left*) and the second "specialized" model (*right*). Substrate **8** is shown in both images, as an example.

oped taking into account such constraints and then by exploiting properly the flexibility of the statistical tools.

Experimental Section

The protein structure used for this study was retrieved from the Protein Data Bank (Id:1TCA). This initial structure was pretreated in MOE[13] by removing the crystallographic water molecules and two molecules of NAG (*N*-acetylglucosamine) present in the pdb file. Hydrogen atoms were added and their position was optimized with an energy minimization procedure in the Amber99 force field in its MOE implementation. Subsequently a minimization of the side chains was performed keeping the backbone atoms fixed.

The substrates were built, minimized and then docked into the active site of CALB by means of the DOCKING module of MOE. The docking was performed on a 10 Å radius selected area surrounding the active site. The force field used for the docking was MMFF94x, the charges of substrate atoms were calculated at the QM PM3 semi-empirical level, by means of the MOPAC7 program. The initial positions of the substrates were manually set in order to meet the criteria previously reported. For each substrate, the conformation presenting the highest score and fulfilling the structural requirements for the initiation of the enzymatic catalysis, was chosen.

Construction of Tetrahedral Intermediates

All tetrahedral intermediates were sketched bonding the hydroxy oxygen of the serine residue (Ser 105) and the reactive carbonyl carbon of the substrate. This carbon atom changes to a tetrahedral sp^3-hybridized configuration. The partial charges and geometry of this chemical species (the substrate and the serine) were calculated by an *ab initio* algorithm, based on DFT-TZVB, by Turbomole.[23] In the molecular mechanics calculations, the standard MMFF94x atom types were used for the atoms of the tetrahedral intermediates, while bond lengths, angles and torsions on tetrahedral carbon were constrained to the values obtained by the ab initio optimization.

Molecular Dynamics

The molecular dynamics simulations were performed using the DYNAMICS module of MOE. All the dynamics were performed in an NVT environment simulating the temperature of approximately 300 K. In order to reduce the calculation time, the attention was focused on the relevant part of the system: all the atoms of the substrate and the protein residues within a sphere of 10 Å radius, centered on the catalytic serine (Ser105) were allowed to move, all the rest of the system was kept constrained. An integration time of 2 fs was used and a frame of the trajectory was saved every 10 fs.

Each substrate conformation chosen for the construction of the data set for the QSAR analysis, was the one characterized by the lowest potential energy out of all the frames saved in the dynamics database. All enzyme structures chosen by these criteria were superimposed with the database viewer superpose implementation. The active conformers were than extracted.

GRID

The GRID analysis was performed on every constituent of the data set. The chosen dimensions of the cage were 14 Å × 24 Å × 21 Å with NPLA (number of grid planes per Ångström) set to 2 while the probes used were DRY (hydrophobic probe) and H_2O (water probe). Once the MIFs have been calculated, all the unfavorable interactions were set to zero. For the DMIFs calculation a specific algorithm was constructed which performs the matrix differential procedure for the subtraction of the two MIFs.

GOLPE

The pretreatment section of GOLPE was used to perform the variable selection. All the variables having an absolute value lower than 0.1 for the water probe and 0.03 for the dry probe were set to zero and those with standard deviation of less than 0.2 for the water probe and 0.06 for the dry probe were discarded. The pretreatment was eventually completed with the block unscaled weight application.

Both PLS models with 5 principal components were computed and validated with the LOO (leave-one-out) method. The prediction ability of the general model was tested on a test set by using the PLS predictions module of the GOLPE program.

References

[1] R. J Kazlauskas, *Curr. Opin. Chem. Biol.* **2000**, *4*, 81–88.
[2] E. Henke, U. Bornscheuer, R. Schmid, J. Pleiss, *ChemBioChem* **2003**, *4*, 485–493.
[3] G. Colombo, S. Toba, K. M. Merz, Jr. , *J. Am. Chem. Soc.* **1999**, *121*, 3486–3493.
[4] S. Tomić, B. Kojić-Prodić, *J. Mol. Graphics Modell.* **2002**, *21*, 241–252.
[5] P. Braiuca, L. Boscarol, C. Ebert, P. Linda, L. Gardossi, *Adv. Synth. Catal.* **2006**, *348*, 773–780.
[6] J. Pleiss, in: *Enzymes in Lipid Modification*, (Ed.: U. Bornsheuer), Wiley-VCH Verlag, Weinheim, **2005**, pp 85–99.
[7] R. J. Kazlauskas, A. N. E. Weissfloch, A. T. Rappaport, L. A. Cuccia, *J. Org. Chem.* **1991**, *56*, 2656–2665.
[8] U. H. Kahlow, R. D. Schmid, J. Pleiss, *Protein Sci.* **2001**, *10*, 1942–1952.
[9] F. Haeffner, T. Norin, K. Hult, *Biophys. J.* **1998**, *74*, 1251–1262.
[10] S. Raza, L. Fransson, K. Hult, *Protein Sci.* **2001**, *10*, 329–338.
[11] F. Felluga, G. Pitacco, E. Valentin, A. Coslanich, M. Fermeglia, M. Ferrone, S. Pricl, *Tetrahedron: Asymmetry* **2003**, *14*, 3385–3399.
[12] P. J. Goodford, *J. Med. Chem.* **1985**, *28*, 849–857.
[13] D. Rotticci, C. Orrenius, K. Hult, T. Norin, *Tetrahedron: Asymmetry* **1997**, *8*, 359–362.
[14] L. E. Iglesias, V. M. Sanchez, F. Rebolledo, V. Gotor, *Tetrahedron: Asymmetry* **1997**, *8*, 2675–2677.

[15] K. A. Skupinska, E. J. McEachern, I. R. Baird, R. T. Skerlj, G. J. Bridger, *J. Org. Chem.* **2003**, *68*, 3546–3551.
[16] J. Ottosson, L. Fransson, K. Hult, *Protein Sci.* **2002**, *11*, 1462–1471.
[17] S. Raza, L. Fransson, K. Hult, *Protein Sci.* **2001**, *10*, 329–338.
[18] D. Rotticci, J. Ottosson, T. Norin, K. Hult, in: *Enzymes in nonaqueous solvents methods and protocols*, (Eds.: E. N. Vulfson, P. J. Halling, H. L. Holland), Humana Press, Totowa, New Jersey, **2001**, pp 261–276.
[19] *MOE v.2006.08*, Chemcomp, Montreal, Canada.
[20] M. A. Kastenholz, M. Pastor, G. Cruciani, E. E. J. Haaksma, T. Fox, *J. Med. Chem.* **2000**, *43*, 3033–3044.
[21] S. Wold, M. Sjöström, L. Erikkson, *Chemom. Intell. Lab. Syst.* **2001**, *409*, 241–246.
[22] *GOLPE 4.5.* Multivariate Infometric Analysis Srl., Viale dei Castagni 16, Perugia, Italy, **1999**.
[23] *TURBOMOLE 5*, distributed by Cosmologic, Leverkusen, Germany.

