

3D-QSAR Applied to the Quantitative Prediction of Penicillin G Amidase Selectivity

Paolo Braiuca,* Luca Boscarol, Cynthia Ebert, Paolo Linda, Lucia Gardossi

Laboratory of Applied and Computational Biocatalysis, Dipartimento di Scienze Farmaceutiche, Università degli Studi, Piazzale Europa 1, 34127 Trieste, Italy
Fax: (+39)-040-52572, e-mail: braiuca@units.it

Received: September 6, 2005; Revised: January 31, 2006; Accepted: February 3, 2006

 Supporting Information for this article is available on the WWW under <http://asc.wiley-vch.de/home/>.

Abstract: A new approach for predicting the selectivity of penicillin G amidase (PGA) – expressed as $k_{\text{cat}}/K_{\text{M}}$ – is here described. Regression models were constructed correlating the experimentally determined $k_{\text{cat}}/K_{\text{M}}$ of a limited number of substrates to molecular descriptors calculated by using methods generally employed in drug discovery for quantitative structure-activity relationship (3D-QSAR methods). Two different methods for the calculation of molecular descriptors have been tested, namely GRIND and Volsurf. The real predictions, made on molecules not used for constructing the models, had an accuracy sufficient for being useful in the experimental practice.

Both approaches led to models able to predict substrate selectivity even without modelling the enzyme-substrate complex, whereas the prediction of enantioselectivity was feasible only by combining the GRIND approach with the conformational analysis of the substrates inside the enzyme's active site. The present approach represents an actual alternative to screening procedures since it allows one to develop a whole predicting model in a few hours, once a small set of experimental data is made available.

Keywords: biocatalysis; 3D-QSAR; enzyme selectivity; molecular modelling; penicillin amidase

Introduction

Substrate selectivity, along with regio-, chemo- and stereoselectivity, are the most valuable properties of enzymes and the full exploitation of biocatalysts in productive processes relies on the possibility to identify and select the proper enzyme for a reaction of interest. Although high-throughput screening methods are becoming routine practice, their cost makes them usually affordable only to some industrial realities.

As an alternative strategy, the most suitable biocatalyst can be identified by applying models able to predict the selectivity of a known enzyme, that means predicting the $k_{\text{cat}}/K_{\text{M}}$ constant (selectivity constant) of the enzyme for a given substrate. In this view it appears clear that the possibility to predict enzyme kinetics by means of fast *in silico* methods would be of major utility for avoiding expensive and time-consuming unsuccessful experiments.

Molecular modelling approaches have been extensively used in biocatalysis research with the aim to understand and ultimately to predict enzyme selectivity. The $k_{\text{cat}}/K_{\text{M}}$ ratio depends on the free energy of the transition state of the reaction, which is generally calculated either by simplified methods based on molecular me-

chanics^[1] or more refined methods, such as QM/MM and free energy perturbation.^[1,2] 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, often they still provide unsatisfactory quantitative accuracy.

Recently, Tomić et al.^[3] described an example of an alternative strategy for the development of quantitative predicting models while avoiding direct free energy calculations. This was accomplished by using 3D-QSAR (three-dimensional quantitative structure-activity relationships) to regress descriptors of the chemical system to the experimentally determined enantioselectivity (E value) of *Bulkholderia cepacia* lipase. 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 rationale behind the present research is to try to construct multiple regression models able to correlate $k_{\text{cat}}/K_{\text{M}}$ to suitable molecular descriptors of the substrates considered, thus avoiding the calculation of the

free energy of the transition state. This allows us to predict enzyme selectivity *in silico* by applying an empirical mathematical model constructed on the basis of a set of experimental data.

Results and Discussion

Selection of the Experimental Data Set (Training Set)

The hydrolysis of different amides and of one ester, catalyzed by penicillin amidase (PGA), was chosen as model reaction and the values of $k_{\text{cat}}/K_{\text{M}}$, experimentally measured and previously published by the group of Švedas,^[4,5] were used as the training data set. Multiple regression models were constructed on the basis of molecular descriptors calculated by two different GRID-derived QSAR methods – namely GRIND^[6] and Volsurf.^[7] Unlike the classical 3D-QSAR approach, such as the CoMFA method,^[8] the GRIND and the Volsurf methods present the major advantage of being independent from any alignment of molecules, which is usually a primary source of error in this kind of studies.^[9,10]

The experimental data set (DS19) includes the $k_{\text{cat}}/K_{\text{M}}$ values for the hydrolysis of 19 substrates presenting a quite broad structural variability comprising seven β -lactam compounds, ten *N*-phenylacetyl amino acids, one anilide and one ester (Table 1). Within these substrates, a sub-set made of only 10 structures (DS10) was defined (Table 1), in order to verify the possibility to build up the models on the basis of a restricted training data set.

Conformation of the Training Set Molecules

The GRID analysis was performed on conformations of the substrate molecules selected according to two strategies. In the first case, an accurate conformational analysis of each substrate within the active site of the enzyme was performed. The molecules were docked into the active site of PGA, following a set of guidelines previously reported by our group,^[11] then each enzyme-substrate complex was solvated, by adding explicit water molecules, and subjected to 100 ps of a molecular dynamics simulation. In the first 10 ps the system was heated, then the temperature was kept constant at 300 K. During the simulation no drastic structural variation was observed (rms for heavy atoms < 1.2 Å in all cases), demonstrating a general structural stability of the complexes. All the simulations reached the equilibrium within the first 60 ps, as demonstrated by the fact that the rms atom deviation was always < 0.9 Å and the average temperature was between 299.98 and 300.01 K.

Ten different conformations were sampled from the last 10 ps of the simulation for each molecule of the

Table 1. Molecules of the training set. Data Set 19 is made up by all the molecules of Data Set 10 and by the molecules reported in the table. The experimental $k_{\text{cat}}/K_{\text{M}}$ values were taken from the literature.^[4,5]

Data Set 10	
Molecules	$k_{\text{cat}}/K_{\text{M}}$
Ampicillin	$2.20 \cdot 10^3$
Cephalexin	$2.60 \cdot 10^4$
Benzilpenicillin	$1.70 \cdot 10^7$
Phenylacetyl-ADCA	$5.00 \cdot 10^6$
<i>N</i> -Phenylacetylalanine	$2.1 \cdot 10^7$
<i>N</i> -Phenylacetylphenylglycine	$4.7 \cdot 10^6$
<i>N</i> -Phenylacetylphenylalanine	$2.1 \cdot 10^7$
<i>N</i> -Phenylacetylglutamine	$1.9 \cdot 10^7$
Ethyl phenylacetate	$3.80 \cdot 10^6$
2-Amino- <i>N</i> -(4-nitrophenyl)-2-phenylacetamide	$1.57 \cdot 10^2$
Data Set 19 = Data Set 10 + the following	
Benzilpenicilloic acid	$2.00 \cdot 10^4$
Cefalothin	$6.00 \cdot 10^5$
Cefaloridin	$3.30 \cdot 10^5$
<i>N</i> -Phenylacetylaspargine	$1.20 \cdot 10^4$
<i>N</i> -Phenylacetylaspargate	$1.70 \cdot 10^6$
<i>N</i> -Phenylacetylisoleucine	$2.00 \cdot 10^6$
<i>N</i> -Phenylacetylleucine	$3.80 \cdot 10^5$
<i>N</i> -Phenylacetylvaline	$5.00 \cdot 10^5$
<i>N</i> -Phenylacetyllysine	$2.00 \cdot 10^6$

data set and a mean structure was calculated from them and used for the construction of the two training sets (DS19 and DS10).

The second strategy was developed with the aim of verifying whether the predictivity of the models relies strictly on the knowledge of the correct conformation of each substrate upon its binding into the active site. Therefore, in this second case, conformations of the molecules were randomly generated (training sets DS19_{random} and DS10_{random}) and the information obtained by the molecular dynamics was completely discarded. The conformation of each molecule was entirely regenerated by using the Cartesian Perturbation function of the MOE program. This function randomly changes the spatial coordinates of all the atoms of the molecule, preserving bonds and chirality, so that the structure is literally “scrambled”. For each molecule one conformer was generated and its energy was minimised.

The procedure leading to the construction of the DS10_{random} and DS19_{random} was repeated three times so that three different and independent models, in which each molecule has a different randomly chosen conformation, were obtained.

Calculation of the Molecular Descriptors: the GRIND Method

Both the GRIND and the Volsurf methods stem from the chemical description of the substrates by means of the GRID program,^[12] which is a computational procedure that calculates the interaction energies between the substrate and a small chemical probe (e.g., a functional group). Energies are calculated in all the nodes of a three-dimensional grid, which spans the structure of the substrates considered. The output is called Molecular Interaction Field (MIF), which can be visualised as an isopotential surface and it can be used as input for a subsequent multivariate analysis to generate QSAR models.^[10]

The GRIND method transforms the information included in the MIF into alignment independent descriptors (correlograms) able to describe the different chemical groups in the molecule and their relative spatial position. The calculation of GRID Independent Descriptors (GRIND) is a two-step procedure. Firstly the hundreds of thousands variables, which constitute the original MIF, are filtered in order to select the most relevant groups of nodes and to discard redundant variables.

The chosen nodes must fulfil the requirements of having low energy values (corresponding to favourable interactions with a given probe) and being as distant as possible from each other. The second step is the so-called Maximum Auto and Cross Covariance (MACC) transformation. It is an autocorrelation procedure,^[13] in which the nodes, selected in the first step, are screened by identifying couples of nodes that are localised at a defined distance. The algorithm uses vectors of length going from 1 Å to n Å, in dependence on the size of the molecule under investigation, to localise the couple and then the energy product of the two nodes is calculated. When more than one couple of nodes fulfils the distance requirement, only the vector representing the maximum energy product is conserved. The correlation can be performed between nodes belonging to the same MIF (generated by the same probe), or to different MIFs (generated by two different probes), resulting in Auto- or Cross-correlation, respectively. Therefore, at the end of this procedure each molecule of the data set is described by *i*) a number of vectors, which link couples of original MIF nodes, and *ii*) their energy products. A graphical example can be found in Figure S1 in the Supporting Information. The descriptors can be plotted in a correlogram profile, where the distances (the lengths of the vectors) appear on the x axis and the energy product on the y axis (Figure S2 in the Supporting Information). Each correlogram constitutes a sort of fingerprint of the molecule and represents the molecule independently from its position in the space. Ultimately, the correlograms form the input matrix for the multivariate analysis and the construction of the regression models.

Calculation of the Molecular Descriptors: the Volsurf Method

In the case of the Volsurf method, the molecular descriptors are calculated from the GRID molecular interaction fields with the aim to describe chemical-physical properties of the molecules. As a matter of fact, the Volsurf method was originally developed for the prediction of pharmacokinetic properties of drugs and especially the interactions between drug molecules and biological membranes. The method has demonstrated impressive performances in the prediction of drug solubility, membrane permeation and intestinal adsorption.^[6] Exhaustive information on the molecular descriptors used by the Volsurf methods can be found in the original work of Cruciani et al.,^[6] whereas Table 2 reports a schematic description of some of the most relevant ones which were used in the present investigation.

Regression Models

The multivariate statistical strategy for the calculation of regression models was the same for both the GRIND and the Volsurf methods. The standard NIPALS algorithm, in its GRIND and Volsurf implementation, was used to calculate the Partial Least Squares (PLS) models: five principal components were calculated and all the descriptors were used.^[14] The predictivity of the model was evaluated by cross-validation performed by means of the Leave-One-Out method, because of the limited number of molecules in the data set. The predictive correlation coefficient (q^2) provided a quantitative evaluation of the capacity of the model to predict the k_{cat}/K_M value for molecules external to the training set.

The FFD algorithm for variable selection was then applied in order to reduce the number of variables and to maximise the information that can be extracted from them. After this variable selection procedure, the cross-validation was repeated, obtaining an increase of q^2 of at least 20% for all the models. The complexity of the model (the number of principal components conserved) was chosen in order to maximise the explained variance and the predictivity. In most of the cases two principal components represented the best compromise.

Results for GRIND and Volsurf are summarised in Table 3, while a schematic representation of the whole strategy can be found in Figure 1.

Predictivity of the Models

Both GRIND and Volsurf demonstrated to be appropriate methods for the quantitative prediction of PGA substrate selectivity. The correlation coefficient (r^2) – which indicates the ability of the model to explain the variance of the original variables – and the predictive correlation

Table 2. A schematic description of some of the most relevant Volsurf descriptors.^[7] The complete list can be retrieved from the Moldiscovery web site (www.moldiscovery.com).

Molecular Surface	The surface generated by a water probe interacting at 0.20 kcal/mol.
Molecular Volume	It represents the volume contained within the water accessible surface computed at 0.20 kcal/mol.
Rugosity	Ratio volume/surface. The smaller the ratio, the larger the rugosity
Molecular Globularity	Globularity is defined as $S/S_{\text{equivalent}}$, in which $S_{\text{equivalent}}$ is the surface area of a sphere of volume V. It assumes values greater than 1.0 for real spheroidal molecules.
Hydrophilic Regions (W1–W8)	The volume of MIF generated by water probe at 8 different energy levels (–0.2, –0.5, –1.0, –2.0, –3.0, –4.0, –5.0, –6.0 kcal/mol)
Hydrophobic Regions (D1–D8)	The same, but for the hydrophobic probe.
Integy Moments (Iw1–Iw8)	Vectors pointing from the centre of mass to the centre of W1–W8, respectively. If the integy moment is high, the hydrated regions are clearly concentrated only in one part of the molecular surface. If the integy moment is small, the polar moieties are either close to the centre of mass or uniformly distributed around the molecule.
Hydrophobic Integy Moments	The same but considering the hydrophobic regions.
Capacity Factors	The ratio between the hydrophilic regions and the molecular surface
Hydrophilic-Lipophilic Balance	The ratio between the hydrophilic and the hydrophobic regions' surfaces
Amphiphilic Moment	Vector pointing from the centre of the hydrophobic domain to the centre of the hydrophilic domain
Critical Packing Parameter	Ratio between the hydrophobic and lipophilic parts of a molecule. Conversely to HL balance, CP refers just to the molecular shape
Hydrogen Bonding	Differences between the hydrophilic volumes (W1–W8) generated by the water probe and any other polar probe included, at the same energy level. Since water can donate and accept H-bonds, this descriptor measures the nature of H-bonding capabilities of the molecule.

Table 3. Statistical results of the GRIND and Volsurf models. The models with highest statistical significance are reported in bold characters.

GRIND								
Number of PCs	DS10		DS10 _{random}		DS19		DS19 _{random}	
	R^2	$q^{2[a]}$	r^2	$q^{2[a]}$	R^2	$q^{2[a]}$	r^2	$q^{2[a]}$
1	0.86	0.54	0.88	0.53	0.71	0.45	0.57	0.36
2	0.98	0.76	0.96	0.78	0.83	0.52	0.72	0.48
3	0.99	0.87	0.98	0.85	0.94	0.48	0.85	0.21
Volsurf								
Number of PCs	DS10		DS10 _{random}		DS19		DS19 _{random}	
	R^2	$q^{2[a]}$	r^2	$q^{2[a]}$	R^2	$q^{2[a]}$	r^2	$q^{2[a]}$
1	0.92	0.66	0.88	0.44	0.59	0.12	0.65	0.33
2	0.96	0.93	0.94	0.82	0.80	0.49	0.76	0.48
3	0.99	0.95	0.95	0.85	0.85	0.69	0.81	0.40

^[a] The predictive correlation coefficient is calculated by the Leave-One-Out (LOO) method and it is based on this equation:

$$q^2 = 1 - \left(\frac{\sum (Y - Y')^2}{\sum (Y - \bar{Y})^2} \right)^{[14]}$$

A $q^2 > 0.4$ indicates that the predictivity of the model is adequate. All the models are acceptable and predictive, since the explained variance is $> 80\%$ and q^2 is > 0.4 .

coefficient (q^2) show that the models are robust and effectively predictive (Table 3).

Concerning the models based on the randomly generated conformations, the procedure leading to the construction of the DS10_{random} and DS19_{random} was repeated three times in order to have three different independent models and to exclude chance results. The differences among the three resulting models were extremely limited (standard deviation of r^2 and $q^2 < 0.18$ in the case of GRIND, < 0.08 in the case of Volsurf). Therefore, the three models are discussed as a single one and q^2 and r^2 in Table 3 are the mean of the three values.

It must be underlined from the inherent nature of the regression models it derives that they are able to predict the effect of a variable (e.g., a specific structural feature of a molecule) as long as such a variable is somehow represented in the training set. As a consequence, models that are built up on the basis of a training set of structurally homogeneous molecules are expected to have an excellent predictivity but only with respect to molecules that fall within the structural features accounted for by the training set. This explains the lower predictivity of the enlarged model (DS19), which comprises a wider structural variability, although its lower predictivity is largely compensated by the applicability of the model to molecules having a broader structural diversity.

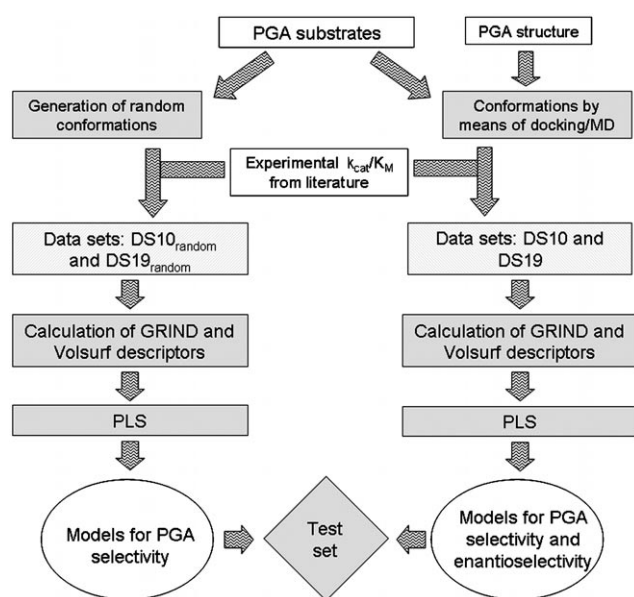


Figure 1. Workflow of the process for the construction of 3D-QSAR models. Conformations of the molecules can be derived from the interaction with the enzyme or they can be randomly generated, then any information about the enzyme structure is discarded and MIFs are calculated considering the substrates as target molecules. Molecular descriptors, either derived from GRIND or Volsurf are calculated, then their values are used as variables for the regression with the experimental $k_{\text{cat}}/K_{\text{M}}$. Finally, an external prediction *versus* a test set of molecules, which are not included in the training sets, is performed. Model derived from data set generated by taking into account the enzyme-substrate interaction can predict also enantioselectivity while, if conformations are generated randomly, only the substrate selectivity can be predicted.

Some real predictions, using molecules not included in the original training set, were also performed, in order to have a further “independent” validation criterion. Table 4 reports $k_{\text{cat}}/K_{\text{M}}$ values calculated by using the DS19 model for a test set of substrates (molecules not originally included in the training set) and, for comparison, also the experimentally determined data. It must be noted that also the experimental data used for the test set were taken from the work of Švedas^[4,5] to minimise errors coming from the variability of experimental conditions.

The conformations of molecules of Table 4 were generated simply by means of an automated docking simulation into the active site of PGA. Afterwards, the GRIND and Volsurf descriptors for the resulting conformations were calculated and $k_{\text{cat}}/K_{\text{M}}$ values were predicted by using the PLS models previously constructed. The whole procedure was completed in the minutes time-scale.

Both GRIND and Volsurf demonstrated to be able to correctly predict $k_{\text{cat}}/K_{\text{M}}$ with a satisfactory accuracy (Table 4).

The limited differences between DS10 and DS10_{random} and between DS19 and DS19_{random} demonstrate that the validity of the GRIND method is quite independent from the calculation of the precise conformation acquired by the molecules inside the active site. Of course this observation cannot be automatically extended to other biocatalysed systems and, especially, it is valid as long as the aim of the model is confined to the prediction of substrate selectivity. As a matter of fact, when the predictivity of the GRIND models was verified against a test set (Table 4) the prediction of enantioselectivity was satisfactory only in the case of the GRIND models constructed on the basis of the DS10 and DS19 training sets, while the models based on the data set with random conformations (DS19_{random} and DS10_{random}) were not predictive. This expected result is a consequence of the fact that enzyme enantiodiscrimination is mainly a consequence of the different conformations acquired by the two enantiomers upon binding inside the active site.

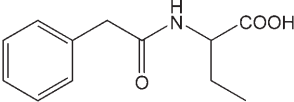
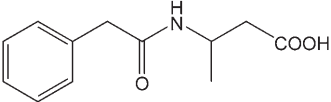
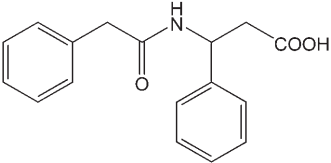
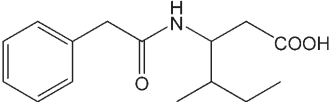
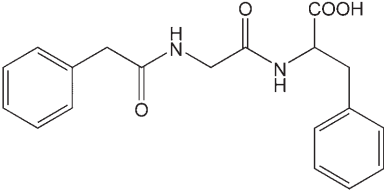
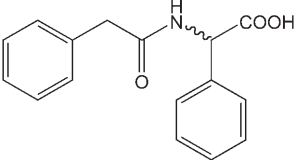
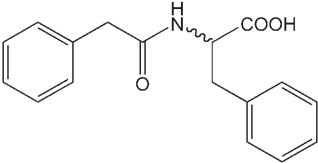
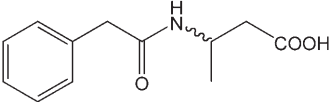
The good results obtained with the Volsurf models are somehow surprising. Volsurf descriptors are appropriate for a detailed description of physical-chemical properties, whereas the information accounting for the steric features of the molecules is very limited.

Nevertheless, the predictivity is high both for the random and for the conformation-based models (q^2 higher than 0.8 for DS10, higher than 0.6 in the case of DS19, see Table 3). The unexpected result can be interpreted in terms of the original ability of the Volsurf descriptors to describe interactions between molecules and biological membranes and physical-chemical phenomena more in general. On this basis, the Volsurf analysis should be able to provide information on the contribution to $k_{\text{cat}}/K_{\text{M}}$ of physical-chemical phenomena occurring in the biocatalysed system such as, for instance, solvation and desolvation of the substrates. Actually, it is widely recognised that the solvent can change enzyme selectivity as a result of variations that are largely ascribable to differences of solvation/desolvation energies of the substrates.^[15]

As expected, the quantitative prediction of enantioselectivity becomes unfeasible by using the Volsurf models. Volsurf descriptors are designed for the predictions of pharmacokinetic properties of drugs, which are mainly passive diffusion phenomena and quite independent of stereoisomerism. Therefore, the molecular descriptors calculated for the two enantiomers by Volsurf are rather similar, and some minor differences emerge only in the case of enantiomers acquiring very different conformations after binding in the enzyme active site.

Finally, it must be noted that the same data sets were used also for building 3D-QSAR models by the CoMFA method, but the results were completely unsatisfactory (q^2 of 0.20 in the best case, -0.37 in the worst one: the model was unable to predict the $k_{\text{cat}}/K_{\text{M}}$; procedure not discussed). The limits of the CoMFA method emerge also from its inadequacy to include in the regression

Table 4. Predictions for the test set calculated by GRIND and Volsurf DS19 models. Experimental data were taken from refs.^[4,5]

Molecules	Experimental ($k_{\text{cat}}/K_{\text{M}}$)	GRIND ($k_{\text{cat}}/K_{\text{M}}$)	Volsurf ($k_{\text{cat}}/K_{\text{M}}$)
	$4.6 \cdot 10^7$	$2.9 \cdot 10^7$	$4.1 \cdot 10^7$
	$2.8 \cdot 10^6$	$3.9 \cdot 10^6$	$1.8 \cdot 10^6$
	$1.1 \cdot 10^6$	$1.4 \cdot 10^6$	$1.3 \cdot 10^6$
	$3.0 \cdot 10^4$	$1.2 \cdot 10^4$	$4.2 \cdot 10^4$
	$0.9 \cdot 10^7$	$1.7 \cdot 10^7$	$2.0 \cdot 10^7$
Enantioselectivity [$(k_{\text{cat}}/K_{\text{M}})^{\text{L}}/(k_{\text{cat}}/K_{\text{M}})^{\text{D}}$]	Experimental	GRIND	
	4300	3700	
	5700	4600	
	220	190	

model the only ester molecule present in the DS19 data set, which resulted to be an outlier in the CoMFA model. On the contrary, the GRIND and Volsurf models were able to predict the $k_{\text{cat}}/K_{\text{M}}$ for the ester with the same level of accuracy as for the amides.

These results demonstrate the general validity of GRIND and Volsurf QSAR approaches for the prediction of PGA selectivity. The methods can easily be ap-

plied to different enzymes and reactions, by calculating new models as long as a limited number of experimental measurements ($k_{\text{cat}}/K_{\text{M}}$) is gathered or data from the literature are available.

Of course, the construction of any 3D-QSAR model is strictly founded on the availability of experimental data. Moreover, the models will be endowed with satisfactory predictivity towards the molecules of interest only when

a compromise between the size of the data set available and the structural variability of the structures will be found.

Therefore, if the experimental data regarding a specific enzyme are limited or severely affected by experimental errors, the quality of prediction will be lower, although information coming from this approach will be still useful for orienting the experimental planning.

The construction of a new predictive model for an enzyme of interest can be completed in a time scale that is considerably limited. In that sense, it can be envisaged that a database for different enzymes can be constructed, thus enabling the possibility to perform *in silico* the choice of the proper enzyme for a specific experimental need. In that sense QSAR tools can be seen as a complementary approach to screening procedures, which would be greatly simplified and accelerated.

Conclusions

The results here described demonstrate for the first time that enzyme selectivity can be predicted using 3D-QSAR regression models which correlate $k_{\text{cat}}/K_{\text{M}}$ to descriptors of the substrates. More important, the calculation and the analysis either of the transition states or of their stable analogues are not necessary, if descriptors contain the necessary information. In this respect, while GRIND approach proved to be successful in predicting both substrate selectivity and enantioselectivity, the Volsurf method showed excellent performances only if applied to substrate selectivity predictions.

Moreover, the two methods demonstrated to be able to predict substrate selectivity independently from the conformational analysis of the substrates docked into the enzyme. This translates into an enormous reduction of the operational time (from a few days, to a few hours) which makes the methods definitely competitive as compared to experimental screening procedures. While this works well in the case of PGA, the extension of this finding to different biocatalytic systems should be verified. In principle, this would make it possible to create predictive models of substrate selectivity starting from a few experimental measurements, even when the structure of the enzyme is not known.

Finally, our results indicate that the GRIND and especially the Volsurf methods are very efficient in taking into account solvation effects and indirectly they demonstrate the great effect of these phenomena on enzyme selectivity.

Since 3D-QSAR techniques are not *ab initio* methods, they necessarily depend on the availability of accurate and reliable experimental data sets. Nevertheless, the great advantage of this approach is that, once a set of experimental data is made available, a whole predicting model can be built up in a few hours.

Experimental Section

All the calculations were performed by a dual-Xeon workstation running a Red Hat Linux operating system and an SGI Octane workstation.

Molecule Structures and Preparation of Enzyme-Substrate Complexes

The PGA structure used for the study was retrieved from the Protein Data Bank (Id: 1PNK). Crystallographic water molecules were removed and hydrogen atoms were added by means of the BIOPOLYMER tool of SYBYL 6.9, then their positions were optimised by means of an energy minimisation, in which all the other atoms were fixed, using the Amber 4.1 force field in its SYBYL implementation. Subsequently, the whole side-chains were minimised, while taking the backbone atoms as fixed. For all the minimisation calculations the Powell method and an rms gradient termination criterion of 0.001 kcal/mol were used.

The substrates were docked into the active site of PGA by means of the DOCKING module of MOE. The simulations were tuned in 25 runs of simulated annealing, with 8 cycles per run and an initial temperature of 1000 K, using a cubic docking box with 45 Å sides, centred on the substrate. The force field used for the docking was MMFF94, 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, by using the criteria previously reported.^[11] For each substrate, the conformation presenting the highest score and fulfilling the structural requirements for the initiation of the enzymatic catalysis,^[11] was chosen.

Each enzyme-substrate complex was solvated by a box of water molecules by using the SOLVATE function of SYBYL and their positions were optimised by means of an energy minimisation.

Molecular Dynamics

The molecular dynamic simulations were performed by using the DYNAMICS module of SYBYL. During the 100 ps simulation, the system was firstly heated for 10 ps, and the temperature was kept constant at 300 K till the end of the run. 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 and solvent molecules within a sphere of 15 Å radius, centred on the substrate, were allowed to move, all the rest was kept fixed. An integration time of 1 fs was used and a frame of the trajectory was saved every 10 fs.

Data Sets

Each substrate conformation, for the construction of the data set for QSAR analysis, was calculated from the mean of ten sampled conformations, taken from the last 10 ps of the MD trajectory.

The data sets DS10 and DS19 included the molecules reported in Table 1.

Data sets DS10_{random} and DS19_{random} were built up by using random conformations for the substrates, discarding all the information coming from the MD simulation. Each substrate was subjected to 10 steps of the “Cartesian perturbation” function in the MOE program and the random conformation generated was energy minimised. The procedure was repeated for three times, thus generating three different versions of the DS10_{random} and DS19_{random} data sets.

GRIND

For the construction of the GRIND model the ALMOND program version 3.3.0 was used. No alignment of the substrates was performed. The GRID probes used for the calculation of descriptors were DRY (hydrophobic probe), O (carbonyl oxygen probe), N1 (amidic nitrogen probe), TIP (“shape” probe). All the possible auto- and cross-correlograms were calculated and used for PLS analysis. The FFD algorithm for refinement of the predictive model was applied to each model by using the number of components which led to the highest predictivity.

The same procedure was applied for the construction of four different regression models, using the four different datasets (DS10, DS10_{random}, DS19, DS19_{random}). The results reported in Table 3 for DS10_{random} and DS19_{random} are the mean of three different models, since three different random conformation data sets were generated (see previous paragraph). It has to be noted that standard deviations of r^2 and q^2 were always <0.18 .

Volsurf

The Volsurf program version 3.0 was used. The probes utilized for the calculation of the descriptors are: DRY (hydrophobic probe), H2O (water probe), O (carbonyl probe) and N: (= sp^2 nitrogen).

All the descriptors were calculated and only the most significant ones were selected by applying the FFD selection algorithm, following the same protocol applied in the construction of GRIND models.

The same procedure was applied for the construction of four different regression models, using the four different data sets (DS10, DS10_{random}, DS19, DS19_{random}). The results reported in Table 3 for DS10_{random} and DS19_{random} are the mean of three different models, as in the case of GRIND. The standard deviations of r^2 and q^2 were always <0.08 , thus indicating the low dependence of the Volsurf descriptors on the substrate conformation.

External Predictions

The molecules reported in Table 4 were sketched and docked into the active site of PGA, following the procedure described in the section “Molecule Structures and Preparation of Enzyme-Substrate Complexes”. The extracted conformations were imported into GRIND and Volsurf for the calculation of the molecular descriptors and for the prediction.

Acknowledgements

The present work has been made possible by the contribution of the Centre of Excellence for Biocrystallography (C.E.B.) of the University of Trieste. Thanks are due to Consorzio Italiano per le Biotecnologie (C.I.B.) for the financial support. Two anonymous referees are gratefully acknowledged for useful comments and constructive criticisms. We are grateful to Moldiscovery, for providing the software.

References

- [1] R. J. Kazlauskas, *Curr. Opin. Chem. Biol.* **2000**, *4*, 81–88.
- [2] F. Felluga, G. Pitacco, E. Valentin, A. Coslanich, M. Ferrmeglia, M. Ferrone, S. Pricl, *Tetrahedron: Asymmetry* **2003**, *14*, 3385–3399.
- [3] S. Tomić, B. Kojić-Prodić, *J. Mol. Graph. Model.* **2002**, *21*, 241–252.
- [4] V. K. Švedas, M. V. Savchenko, A. I. Beltser, D. F. Guranda, *Ann. N.Y. Acad. Sci.* **1996**, 659–669.
- [5] A. L. Margolin, V. K. Švedas, I. V. Berezin, *Biochim. Biophys. Acta* **1980**, *6165*, 283–289.
- [6] M. Pastor, G. Cruciani, I. McLay, *J. Med. Chem.* **2000**, *43*, 3233–3243.
- [7] G. Cruciani, M. Pastor, W. Guba, *Eur. J. Pharm. Sci.* **2000**, *11*, 29–39.
- [8] R. D. Cramer III, D. E. Patterson, J. D. Bounce, *J. Am. Chem. Soc.* **1988**, *110*, 5959, 5967.
- [9] S. J. Cho, M. L. Serrano Garsia, J. Bier, A. Tropsha, *J. Med. Chem.* **1996**, *39*, 5064–5071.
- [10] M. Pastor, G. Cruciani, K. A. Watson, *J. Med. Chem.* **1997**, *40*, 4089–4102.
- [11] A. Basso, P. Braiuca, C. Ebert, L. Gardossi, P. Linda, *Biophys. Acta* **2002**, *1601*, 85–92.
- [12] P. J. Goodford, *J. Med. Chem.* **1985**, *28*, 849–857.
- [13] D. Riganelli, R. Valigi, G. Costantino, M. Baroni, S. Wold, *Pharm. Pharmacol. Lett.* **1993**, *3*, 5–8.
- [14] S. Wold, M. Sjöström, L. Eriksson, *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109–130.
- [15] A. M. Klibanov, *Nature* **2001**, *409*, 241–246.