Journal of Molecular Catalysis B: Enzymatic 101 (2014) 7–15

Contents lists available at ScienceDirect

Journal of Molecular Catalysis B: Enzymatic

journal homepage: www.elsevier.com/locate/molcatb

An integrated platform for automatic design and screening of virtual mutants based on 3D-QSAR analysis

Valerio Ferrario[a], Cynthia Ebert[a], Allan Svendsen[b], Werner Besenmatter[b], Lucia Gardossi[a,*]

[a] Dipartimento di Scienze Chimiche e Farmaceutiche, Università degli Studi di Trieste, Piazzale Europa 1, 34127 Trieste, Italy
[b] Novozymes A/S, Krogshoejvej 36, DK-2880 Bagsvaerd, Denmark

ARTICLE INFO

Article history:
Received 8 October 2013
Received in revised form
19 November 2013
Accepted 9 December 2013
Available online 18 December 2013

Keywords:
Rational enzyme engineering
3D-QSAR
In silico automatic screening
Multivariate, Statistical analysis

ABSTRACT

An innovative application of 3D-QSAR methodology to the rational design of enzymes is here reported. The introduction of amidase activity inside the scaffold of lipase B from *Candida antarctica* (CaLB) was studied and 3D-QSAR models were constructed to correlate the structures of a set of CaLB mutants with their experimentally measured activities. Properties, like hydrophilicity, hydrophobicity and hydrogen bonding capability of the enzyme active site were computed by means of the GRID method and the output was used as molecular descriptors. Correlations with experimental behavior of the catalysts were calculated by means of partial least square regression (PLS). The analysis of the QSAR model fully exploits fundamental knowledge while avoiding conceptual biases. Rationales for driving enzyme engineering are disclosed and *a priori* evaluation of new virtual candidate mutants becomes feasible. On that respect, the whole procedure for production of virtual mutants and scoring of their activity was automated within a workflow constructed by means of the modeFRONTIER package. The method allows for the automated construction and scoring of each mutant in 2 h on a normal workstation.

1. Introduction

Natural biocatalysts are often not optimally suited for industrial applications. The development of efficient strategies for improving enzyme properties and expanding the range of reactions catalyzed by enzymes is of crucial importance for boosting the use of biocatalysts at industrial level.

Over the past twenty years, enzyme properties have been tailored through *in vitro* directed evolution of enzymatic proteins using random genetic mutation and recombination, followed by screening for a desired enzymatic activity. This technique does not rely on *a priori* knowledge of the relationship between protein structure and function [1].

As an alternative approach, enzyme properties have been optimized by means of rational design, usually assisted by molecular modeling and bioinformatic analysis [2,3]. Ideally, rational design of enzyme mutants should be driven by fundamental knowledge of the whole array of structural, electronic and functional factors that affect enzyme properties. Current computational studies of enzyme activity as measured by the activation free energy generally restrict their focus to the wild type enzyme and a limited number of mutants, which have been described with a comparatively high [4–6] or moderately high [7–9] level of theory. However, the computational demand of these methods makes it difficult to apply them to the actual design of new enzymatic catalysts where the activity of hundreds of mutants has to be evaluated. Hediger et al. have recently published a computational method for high-throughput computational screening of mutant activity [10] and the method was benchmarked against experimentally measured amidase activity for mutants of *Candida antarctica* lipase B (CaLB) with the aim of identifying promising mutants.

Here we report on an alternative computational strategy for approaching the same engineering problem. Molecular modeling, experimental studies and multivariate statistical analysis are merged to develop low computational demanding models for *in silico* screening and analysis of mutants.

The main concept at the basis of this hybrid strategy is that catalytic efficiency of enzymes depends on multiple contributions of factors that are not independent but rather interact among them, thus resulting in a complex behavior. Consequently, it is mandatory to follow a strategy that allows identifying and analyzing not only the effect of each single variable but also their interactions. Multivariate statistical analysis is the tool able to account for variable interactions and to extract the relevant information contained in huge matrixes of data [11]. Principal Component Analysis (PCA) and Partial Least Square analysis (PLS) are suitable for the interpretation and modeling of complex systems described by exceeding number of variables. Variables that are correlated with one another

* Corresponding author.
E-mail address: gardossi@units.it (L. Gardossi).

1381-1177/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.molcatb.2013.12.004

are combined into a 'principal component' (or latent variable) so that objects are projected on a space of reduced dimensionality (*i.e.* a reduced number of independent variables corresponding to the new components) [12]. In particular, when a certain response (*Y* variable) of the system must be modeled and optimized, the new variables or components are extracted to give the best fit of both *Y* and *X* variable matrices. This is accomplished by applying the PLS analysis, where the *Y* latent variables are correlated to *X* latent variables [13].

In principle, data calculated using molecular simulation methods could be analyzed by PLS to reveal the inherent structure of the data and to find three-dimensional quantitative structure–activity relationships (3D-QSAR). Suitable molecular descriptors are necessary for extracting relevant structural information from target molecules and correlate them with chemical or physical properties [14]. Although QSAR methods are well established in drug design strategies [15], they have been applied to biocatalysis only recently for the study of enzyme thermostability [16], enzyme specificity [17,18] and enantioselectivity [19].

The present study represents, to the best of our knowledge, the first attempt to apply 3D-QSAR analysis for correlating structures of mutants and their activity within an enzyme engineering strategy. In order to fully exploit fundamental knowledge while avoiding conceptual biases, computational and statistical techniques were evaluated in the context of enhancing amidase activity in *Candida antarctica* lipase B (CaLB). As is widely known, most proteases, amidases and lipases contain the same type of catalytic machinery as serine hydrolases. Nonetheless, esterases/lipases have very low or undetectable amidase activity and this fact has been extensively discussed [20]. The QSAR study here reported allowed obtaining quantitative information on variables affecting amidase activity and their interactions, thus providing guidelines for mutagenesis strategies. Overall, by "learning" from a model that correlates structural information with kinetic data, the methodology allows to extract objectively information that actually affects the activation energy of the reaction of interest. Of course, QSAR predictive models are reliant on the data on which they are based and on the overall quality of the information, including the item to be modeled. Therefore, the validity and extendibility of the method strongly depends on the data set used for training the model. The 3D-QSAR mathematical analysis here reported provides a semi-quantitative tool for *in silico* screening of virtual mutants and the perspectives for the exploitation of these computational approaches to rational enzyme engineering are also discussed herein.

2. Experimental

2.1. Production of CaLB mutants

The experimental CaLB variants were produced according the procedure previously reported by Suplatov et al. [21].

Variants of CaLB were generated by polymerase chain reaction-based (PCR) site-directed mutagenesis. The PCR was set up with the proof-reading KOD DNA polymerase (from Novagen, Toyobo) and a 7467 basepair *Escherichia coli–Aspergillus* plasmid that was previously methylated with CpG methyltransferase (from NEB). The PCR products were used to transform competent *E. coli* DH5a cells (from TaKaRa). Plasmid DNA was recovered and sequenced to verify the presence of the desired substitution. Confirmed plasmid variants were used to transform an *Aspergillus oryzae* strain that is negative in pyrG (orotidine-5-O-phosphate decarboxylase) and that is also negative in the proteases pepC (a serine protease homologous to yscB), alp (an alkaline protease) NpI (a neutral metalloprotease I) to avoid degradation of the lipase variants during and after fermentation. The transformed *Aspergillus* strains were fermented as submerged culture in shake flasks and the lipase variants secreted into the fermentation medium. After the fermentation, the lipase variants were purified from the sterile filtered fermentation medium in a three-step procedure with (i) hydrophobic interaction chromatography on decylamine-agarose, (ii) buffer exchange by gel filtration, and (iii) ion exchange chromatography with cation exchange on SP-sepharose at pH 4.5. The lipase variant solutions were stored frozen.

2.2. Experimental screening of mutants

The amidase activity was determined with a fluorimetric assay similar to the one described previously [21].

The aqueous reaction mixture contained 0.03 mg/ml CaLB variant, 5 mM substrate (benzyl chloroacetamide), 25 mM phosphate (potassium salt) pH 7.0, 10% (w/v) tetrahydrofuran and was set up in 96-well microtiter plates. The concentration of the enzyme stock solution was determined by measuring the absorbance at 280 nm and calculated based on the extinction coefficient of 1.21 for CaLB. The microtiter plate was covered with parafilm and incubated for 18 h at 37 °C and 300 rpm in the MTP Thermomixer Comfort (form Eppendorf AG). Afterwards 50 ml of 20 mM 4-chloro-7-nitrobenzofurazan (NBDCl) in 1-hexanol was pipetted to 200 ml reaction mixture and incubated for 1 h at 37 °C and 500 rpm. Afterwards the fluorescence intensity was measured with the fluorimeter Fluostar Optima (from BMG Labtech GmbH) with excitation filter at 485 nm and emission filter at 540 nm. The specific activity of CaLB wild type is $1.27 \pm 0.16 \times 10^{-2}$ μmol/mg/h. The activity value for each enzyme variant is the average of three replications, in three wells. In parallel, each enzyme variant was also set up in three different wells without substrate, to measure the small background fluorescents from each variant. Further three wells were set up without substrate and without enzyme, thus only aqueous medium with buffer and solvent, to measure the small background fluorescents from the medium. The average value from the measurements with enzyme and with substrate is named 'es'; the value with enzyme and without substrate is 'e'; the value without enzyme and with substrate is 's'; the value without enzyme and without substrate is 'o'. To correct for fluorescents from enzyme, medium and autohydrolysis the following subtractions were calculated: (es-e)-(s-o) = (es-s)-(e-o) = es-e-s + o = corrected enzyme variant activity value. S-o = autohydrolysis of substrate. Because these measurements are in arbitrary fluorescents units, these values were normalized to percentage of substrate that reacted. Zero percent means no reaction of substrate, 100% means complete reaction of substrate to products. For this normalization two wells for each enzyme variant were set up in parallel. These wells contained the same composition as the other wells, except that they contained instead of substrate the products at a concentration that corresponds to 10% reaction. Thus, the concentrations in these two wells were 0.03 mg/ml enzyme, 0.5 mM hydrolysis products (*i.e.* 0.5 mM benzylamine and 0.5 mM chloroacetic acid), 25 mM phosphate pH 7.0, 10% (w/v) tetrahydrofuran. The average value from the measurements with enzyme and with product is named 'ep'. To correct for background fluorescents from the enzyme variants, the average value from the measurements with enzyme and without product (and without substrate), which was named 'e', was subtracted. In short, the normalized and corrected enzyme variant activity was calculated as 10((es-e)-(s-o))/(ep-e).

2.3. Modeling CaLB WT and mutants

All the calculations were performed using GROMACS software [22] version 4 compiled on a virtual ROKCS linux cluster with 4 nodes.

2.4. Pre-equilibration of CaLB

The structure 1TCA [23] (Protein Data Bank, PDB) [24] of CaLB was used. The structure was pre-processed by deleting the crystal water molecules, ions and sugar moieties in correspondence of glycosylation sites. The force-field employed was GROMOS-96 53a6 [25]; the enzyme was put in the center of a cubic space of 343 $Å^3$ and solvated with explicit water at experimental density. The charges were counter-balanced using Na^+ or Cl^- ions. The system was minimised (10.000 steps of steepest descendent gradient) and dynamised for 6 ns at 300 K with the software GROMACS [22] version 4.

2.5. Computation of the mutants structures

Mutations were performed starting from the pre-equilibrated CaLB structure and using the mutagenesis tool of the software PyMOL [26]. Each of the 30 generated mutants along with the wild type structure were defined in the GROMOS-96 53a6 force-field [25] and then embedded in a cubic space of 343 $Å^3$. Each system was solvated with explicit water and charges were neutralized with Na^+ or Cl^- ions. Each structure was minimised (10,000 steps of steepest descendent gradient) and subjected to a 500 ps molecular dynamic (MD) simulation performed with the software GROMACS [22] version 4. The outcome of each MD simulation (mutant structure) was superposed to the WT CaLB and used for the calculation of the descriptors.

2.6. Computation of molecular descriptors

The molecular descriptors (differential Molecular Interaction Fields, dMIFs) were calculated on the outcome of the MD simulations for the 30 mutants and the wild type CaLB, The MIF descriptors were calculated on each protein structure by the GRID program [27] on a three dimensional grid with a resolution of 2 planes per Å, extended over a volume of 100 nm^3 that comprises the opening of the active site and its floor. Four different probes were used for the calculation: WATER probe, which describes and quantifies the dipolar interactions and the hydrogen bond formation; DRY probe, which describes hydrophobic interactions; O:: probe, which represents a H-bond acceptor carbonylic oxygen and therefore accounts for interactions with H-bond donors; N1 probe, which represents a H-bond donor group and describes interactions with H-bond acceptors. The final dMIFs descriptors were obtained by subtracting the corresponding MIF computed for the wild type CaLB from the MIF of each mutant. This procedure led to the calculation of 479,700 variables for each object (mutant). The variables were divided into four different blocks, belonging to the original probes (WATER, DRY, O:: and N1) used for the MIF calculation.

2.7. 3D-QSAR model

The variables of the dMIFs descriptors were used in order to generate a QSAR model using the software GOLPE version 4.5 [28] a different weight factor was assigned to each block of variables, therefore, the initial 479,700 variables were selected by using D-Optimal preselection and FFD variable selection algorithm, finally 1650 variables were selected for the statistical model.

The predictivity q^2 of the model was calculated by using both the leave-one-out (LOO) and the leave-group-out (LGO) methods. The latter was applied by using five different groups with random object selection.

2.8. Construction of the automatic workflow for virtual screening with ModeFRONTIER

modeFRONTIER version 4 was used [29]; a number of different software were integrated into the modeFRONTIER workflow (PyMOL, GRID, GROMACS and the QSAR equation). The first generation of mutants was randomly produced and the next mutant generations were created using NGSA II genetic algorithm. The improvement of the mutant activity, as defined by the QSAR score, was set as the optimization objective.

3. Results and discussion

Enzyme catalysis is a matter of different and complex factors such as, for instance, enthalpic contribution, entropic factors and dynamic behavior. The present did not aim at computing each contribution but rather at finding an empiric equation suitable to correlate enzyme structure with its ability to catalyze the hydrolysis of a model amide (*N*-Benzyl-2-chloroacetamide). Therefore, GRID mapping of the active site and multivariate statistical analysis were used to record variables describing the active site, both in terms of electrostatic and geometrical features. Through the selection of these variables, aminoacids were identified relevant for the catalytic activity of the considered substrate. The correlation between the selected variables and the activity of the mutants was described through the construction of a Three-Dimensional Quantitative Structure Activity Relationship (3D-QSAR) mathematical model.

Three-Dimensional Quantitative Structure Activity Relationships (3D-QSAR) were established for a set of mutants (training set) by means of PLS analysis that extracts the latent *X* (structural properties) and *Y* variables (amidase activity). The new variables or components are extracted to give the best fit of both *Y* and *X* variable matrices.

The PLS method has been conceived to allow the use of a number of independent variables extremely high in relation to the number of objects. Using the principal components, only the variables directly correlated to the dependent variable contribute to the model. It is also possible to quantify the extent of this contribution, thus identifying the variables of interest in the description of the biological activity [12].

The strategy for the construction of the 3D-QSAR model can be schematized in the following 4 main steps:

1. Generation and selection of mutants constituting the “training set”.
2. Modeling the mutants.
3. GRID analysis of each mutant and extraction of molecular descriptors.
4. Statistical analysis.

Finally, the overall procedure for the generation of mutant models was automated within a workflow where all software and computational procedurer are integrated. The validated 3D-QSAR model was then integrated as well inside the work-flow, as scoring function for the screening of new generations of virtual mutants.

3.1. Generation and selection of mutants that form the “training set”

In the present work the objective of the engineering strategy was to introduce amidase activity inside the scaffold of lipase B from *Candida antarctica* (CaLB) and a training set of 30 CaLB mutants was used for constructing the mathematical model (Table 1). As a general rule, the training set should include mutants that present

Table 1
Training set used for the construction of the 3D-QSAR model. Experimental activity of mutants is expressed as improvement factor as compared to the wild type CaLB (wild type = 1).

Mutant	Mutation	Activity (improvement factor)
D01	A225V	1.22
D02	D145S	0.69
D03	D265P	0.57
D04	G39A	2.76
D05	I189H	1.11
D06	I189Q	0.41
D07	L140E	0.55
D08	L278A	2.50
D09	Q191R	1.28
D10	R249Q	0.74
D11	T103G	1.10
D12	T256K	0.75
D13	T40A	0.12
D14	T40S	0.50
D15	W104F	2.00
D16	A225I	1.55
D17	D223G	1.26
D18	D223K	0.86
D19	39A278A281G	1.67
D20	39A191R	1.49
D21	39A104F	4.15
D22	T42G	1.87
D23	40G278G104F225M157V189A134A	0.10
D24	39A103G	0.80
D25	39A278A	3.20
D26	39A103G278A	2.90
D27	39A104F278A	5.75
D28	39A103G104F278A	10.60
D29	39A103G104Q278A	1.90
D30	39A104F225K278A	0.80

sufficient variability in terms of enzymatic activity and structural properties. Models trained on structurally homogeneous objects are expected to have an excellent predictivity but only within a restricted range of structural properties because regression models are able to predict the effect of a variable *X* (*i.e.* a specific structural feature of an enzyme/mutant) on amidase activity (*Y* variable) as long as such a variable is somehow represented in the training set.

The activity of mutants was evaluated in terms of hydrolytic activity towards *N*-Benzyl-2-chloroacetamide (Fig. 1) that was chosen as reference substrate for the amidase activity (see materials and methods). The measured activity was expressed as improvement factor as compared to the wild type CaLB (wild type amidases activity = 1).

The 30 variants were selected because of their heterogeneous distribution of activity values within a wider data base containing CaLB mutants presenting mutations focused in different active site regions. Globally the mutations involve 17 residues, determining 19 monovariants and 11 multivariants, which result from their combinations.

3.2. Modeling the mutants

The wild type (WT) CaLB was modeled starting from the structure available on the Protein Data Bank (PDB) [24] with the code 1TCA [23], which was also used as template for modeling the mutants. The wild type CaLB was previously subjected to 6 ns of molecular dynamic simulation in explicit water with the aim to stabilize the flexible lid domain [30,31] (see Experimental for details) because such flexibility might interfere in the computation of structural differences coming from mutations.

The dynamised WT CaLB structure was used as template for the generation of mutants structures. Therefore, starting from that equilibrated structure, mutations were introduced and the resulting protein was minimised as well as dynamised for 500 ps.

The dynamised mutant structures were superposed using the backbone of the wild type CaLB as reference, thus obtaining an average RMSD of 1.206 Å (Figure S1 in Supplementary material).

3.3. GRID analysis of CaLB and mutants: computation of molecular descriptors

A molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into useful numbers [32]. Any characteristic of a given molecule can be translated into numbers that constitute the molecular descriptors. Accordingly, the information contained in the descriptor influences the quality of the QSAR model, its ability to describe the data set quantitatively and to correlate molecular properties with the dependent variable of interest (*Y* variable).

Taking this into account, the analysis of the structure of CaLB and its mutants was performed by means of molecular descriptors generated by the GRID program [27], a computational procedure that calculates the interaction energies between any target molecule (in the present case represented by the active site enzyme) and a small chemical probe (*i.e.* a chemical functional group). Energies are calculated in all the nodes of a three-dimensional grid that spans the structure of interest, namely the enzyme active site. The computational output is called Molecular Interaction Field (MIF), which can be visualized as an isopotential surface. The calculated MIFs were used as input for the multivariate analysis to generate QSAR models [33].

MIFs descriptors [19] were computed using four different probes: WATER probe, which describes and quantifies the dipolar interactions and the hydrogen bond formation; DRY probe which describes hydrophobic interactions; O:: probe, which represents a H-bond acceptor carbonylic oxygen and therefore accounts for interactions with H-bond donors; N1 probe, which represents a H-bond donor group and describes interactions with H-bond acceptors. The selection of these four probes was motivated by the consideration that the ability of a certain enzyme to catalyze the hydrolysis of the amide bond in *N*-Benzyl-2-chloroacetamide (Fig. 1) is not only the result of the geometric and electronic features of the residues located in the close proximity of the catalytic triad and the oxyanion hole. Indeed, this "catalytic machinery" is embedded in an active site that acts as a micro-environment where the reaction takes place. In any chemical reaction, the medium can contribute to the stabilization/destabilization of the transition state. In analogy, the active site of an enzyme is "pre-organized" also for stabilizing the transition state through the establishment of a favourable micro-environment, for instance by enabling H-bonds or hydrophobic interactions. Therefore, the aim of the present computational method is to identify the array of physical–chemical factors that contribute to stabilize the transition state of the hydrolytic reaction of an amide inside the active sites of different CaLB mutants.

GRID descriptors were calculated for each mutant by computing MIFs with a resolution of 2 planes per Å, extended over a volume of 100 nm^3, which comprises the opening of the active site and its floor. The conformations of the mutants superposed to the WT enzyme are reported in Figure S1 of Supplementary materials. In order to focus the attention on the differences between the wild type CaLB and the mutants, differential MIFs were calculated (dMIFs) [19] before computing the 3D-QSAR model. dMIFs were obtained by subtracting the corresponding MIF computed for the wild type CaLB from the MIF of each mutant. This procedure led to the calculation of 479,700 variables for each object (mutant). The variables were divided into four different blocks, each belonging to

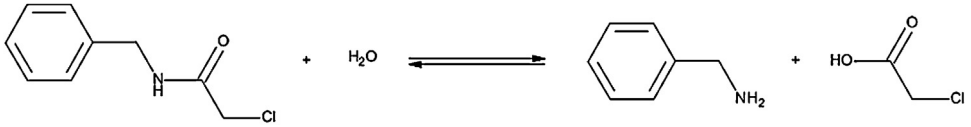

Fig. 1. Scheme of the hydrolysis of *N*-Benzyl-2-chloroacetamide catalyzed by CaLB and used for the screening of mutants.

the different probes used for the MIF calculation (WATER, DRY, O:: and N1).

3.4. PLS analysis and construction of the QSAR model

A different weight factor was assigned to each block of variables to confer them the same initial importance in the model, thus overcoming the differences in absolute value of hydrophobic and polar interactions. Variables were selected by using D-Optimal pre-selection and FFD variable selection algorithm, which allowed retaining only 1650 variables [15].

A PLS model was built up by correlating the dMIFs descriptors with the experimental activity values of the training set. The mathematical model, on the second Principal Component (PC), explains 38% of the whole variance (r^2 = 0.98). The predictivity coefficient was calculated on the second principal component by means of the leave-one-out cross-validation (LOO) and by using the "leave-group-out" (LGO) procedure, obtaining q^2 values of 0.79 and 0.74 respectively.

For the first PC a q^2 of 0.43 was obtained using both methods (Fig. 2).

The graphical representation of the model points out how the PLS analysis was performed using a limited number of highly active mutants, whereas most of the CaLB variants had activity in the range of 0.5 and 2.0. Therefore, the model is expected to describe and predict more accurately new virtual mutants within that range of relative activity (between 0.5 and 2.0) and this explains why the activity of the best mutant (D28) is underestimated.

The predictivity of the model was validated by computing the descriptors for 7 mutants external to the training set and by comparing the experimentally determined value with the activity predicted by the 3D-QSAR model (Table 2). Results confirm that the model is able to discriminate between poor and good mutants, thus allowing a pre-selection *in silico*.

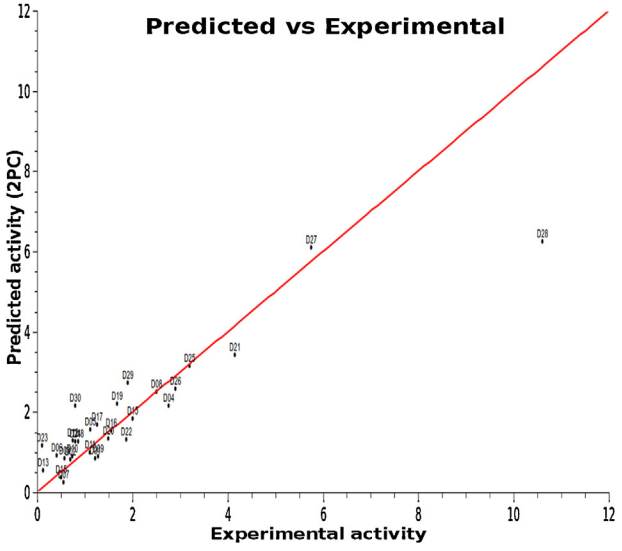

Fig. 2. QSAR model: predicted *vs* experimental graph. Two principal components were considered in the PLS analysis.

Table 2
External validation of the QSAR model reported in Fig. 2 by using mutants external to the training set.

Mutation	Predicted activity	Experimental activity
I189N	0.82	0.95
D223N	2.08	1.30
T42S	1.47	1.21
G39S	0.71	0.55
A225L	0.95	1.19
A225F	0.99	0.75
A225G	1.13	1.26

3.5. Statistical analysis of the effect of the variables

The PLS analysis selected the variables that correlate with the amidase activity, which were ultimately projected on the three-dimensional structure of CaLB to visualize promising hot spots. Fig. 3 illustrates the spatial location of all selected variables for all probes used in the GRID analysis (water, H-bond donor, H-bond acceptor, hydrophobic interactions). By positioning a certain variable on the 3D model it is possible to identify aminoacids, within a radius of 1.5 Å, which contribute at a different extent to the effect of the selected variable.

It must be underlined that the amidase activity depends not only on the effect of the single variables but also it is the complex result of the interaction between them. This is mathematically accounted by the 3D-QSAR model. Accordingly, any interpretation of the effect of the selected variables cannot rely on a simple visual inspection of

Fig. 3. Selected variables represented as dots. Colours are indicative of the value of variable loadings: green = negative; from yellow to red = increasingly positive. (For interpretation of the references to color in figure legend, the reader is referred to the web version of the article.)

Table 3
List of the aminoacids that contribute to determine the effect of the selected variables of the QSAR model. Residues not included in the list of mutations (see Table 1) are highlighted in bold.

Probe	Aminoacids
Water	Gly39, Thr42, Trp104, **Leu109**, Asp134, Leu140, **Ala141**, Gln157, Ile189, **Leu228**, **Ile285**
O::	Thr42, Leu109, Ala141
N1	Thr42, Leu109, Asp134, Leu140, **Val190**, Ile285, **Lys290**

the enzyme structure but, rather, must be based on a rigorous statistical analysis of the data. Conversely, the analysis of the loadings of the selected variables provides quantitative information on the contribution of each variable to the second principal component of the 3D-QSAR model.

It is important to underline that each variable comes from the differential MIF obtained by subtracting each MIF calculated for the wild type CaLB from the corresponding MIF of the mutant.

Once identified a "hot spot" residues, the mutagenesis strategy must go in the direction suggested by the loading values; hot-spot residues related to highly positive or highly negative loadings will have the highest impact on the mutant activity. For instance, a highly positive loading suggests a direct correlation with the mutant activity, therefore the difference between the values calculated for the mutant and the wild type must be positive in order to have a positive effect on the mutant activity.

All the selected variables were sorted out and represented visually on the basis of the probe used for their computation, so that each variable block was analyzed separately. Table 3 reports a schematic overview of the residues identified as relevant by the statistical analysis and representing promising hot spots for mutagenesis (Fig. 4).

It must be underlined that out of the 17 residues addressed by the mutations of the training set (Table 1), 7 were recognized as correlated to the observed variations in the amidase activity. On the other hand, 6 new residues were pointed out despite the fact they were not mutated within the training set.

From the results reported in Table 3 water interactions or "hydratability" of the active site emerge as the key factors for the conversion of the CaLB scaffold into an amidase enzyme. Notably, the model does not select any variable coming from interactions with the DRY probe. In order to verify the criteria used for the selection of the variable a model was constructed using only the DRY variables and its mathematical output confirms that DRY variables do not correlate with the activity. Nevertheless, this result might have two explanations: either the training set contains mutants displaying negligible differences in terms of hydrophobicity or such differences are not correlated with the amidase activity.

Notably, the statistical analysis identified five new hydrophobic residues as hot spots located in the core of the protein and only one polar residue (Lys 290) located on an external flexible domain (Fig. 4B). Globally the results suggest that an engineering strategy aiming at introducing amidase activity in CaLB scaffold should address those residues that affect the ability of the active site to be hydrated. It must be underlined that this concept is consistent with the ability of amidases to accept polar substrates able to establish H-bonds and to drag water inside the active site. Moreover, it is widely known that lipases are active in non-aqueous media even at very low water activity values, whereas amidases need to be hydrated in order to maintain their activity in organic media and this observation suggests a specific role of water in amidase activity of hydrolases [34].

The role of H-bonds in amidase mechanism was also discussed recently by Hult and co-workers [35]. Their study pointed out how, in the mechanism of enzymatic hydrolysis of amides, nitrogen inversion or rotation needs to take place in order to prepare

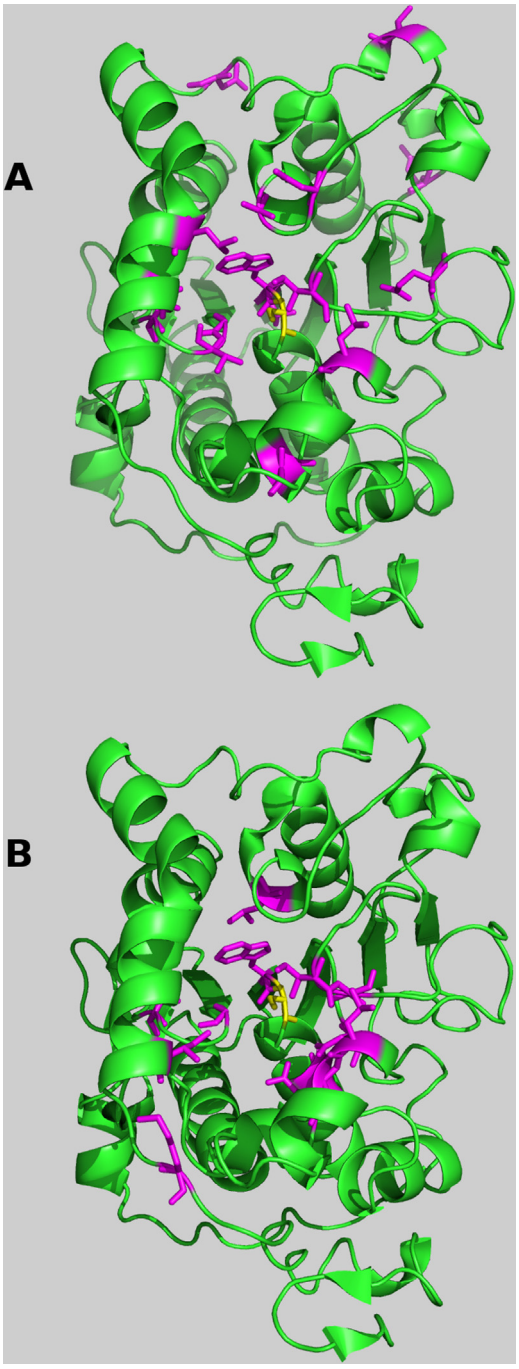

Fig. 4. Structure of CaLB (wild type). (A) Highlights (violet) the 17 residues object of mutations (see Table 1) whereas (B) shows the hot spots identified through the statistical analysis of significant variables (Fig. 3). (For interpretation of the references to color in figure legend, the reader is referred to the web version of the article.)

for proton transfer between the catalytic base (the Histidine Nε atom for serine proteases) and the scissile nitrogen atom (the P1′ nitrogen). This is because after the attack of the catalytic Ser Oγ on the carbonyl carbon of the amide bond, the scissile nitrogen lone pair will be situated antiperiplanar to the formed Ser Oγ-C bond in the tetrahedral intermediate and will point away from the Histidine in accordance with Deslongchamps' stereoelectronic theory. Results reported by the group of Hult were based on the study of the hydrolysis of *p*-nitroanilide: data supported the hypothesis that enzymatic catalysis occurs when the transition state is stabilized thanks to a hydrogen bond donated by the scissile amide bond and

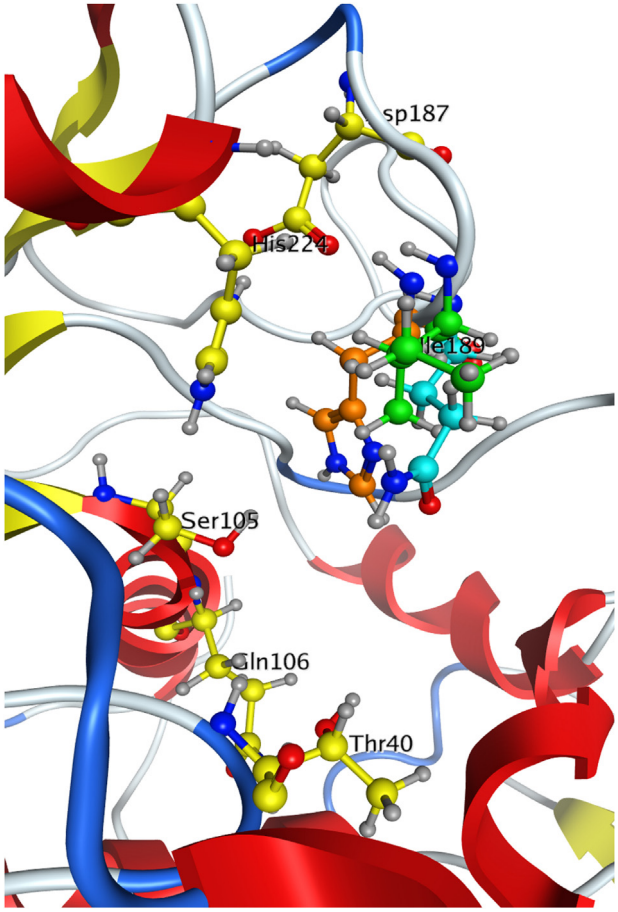

Fig. 5. A comparison of variants at residue 189. Ile189 (wild type) in green, His189 (D05) in orange, Gln189 (D06) in cyan. Residues forming the catalytic triad (Ser 105, His 224, Asp 187) and the oxyanion hole (Gln 106, Thr 40) are in yellow. (For interpretation of the references to color in figure legend, the reader is referred to the web version of the article.)

accepted by the enzyme. The same study identified Ile 189 as a hot spot for mutagenesis that should go in the direction of introducing a residue able to establish an extra H-bond necessary for the stabilization of the transition state of the reaction. Data obtained in our study shows that mutant I189Q (D06) displays a reduced amidase activity, while the mutation I189H (D05) (Fig. 5) has a negligible influence (Table 1). However, in the present study the screening for amidase activity was based on the hydrolysis of an amide and not an anilide, so that structural and electronic differences of the two substrates must be taken into account and no direct comparison with the results of the previous study can be done.

In conclusion, from the statistical analysis (Table 3) it appears that the structural differences that correlate most with the observed variations in amidase activity depend on the ability of the active site to bind water. Of course, WATER probe provides information that includes also the general ability of the target to establish H bonds but this property is not related to one single residue of the active site but it is rather the complex combination of contributions coming from different aminoacids that cannot be analyzed separately.

The possibility of revealing the structural and electronic features responsible for a change of activity from a simple visual inspection results prohibitive even for single point mutants. As an example, Fig. 5 focuses the attention on position 189 and presents a comparison between Ile (CaLB wild type), His (mutant D05) and Gln (mutant D06). A simple visual inspection would not be sufficient for predicting the negative effect of mutation I189Q, since Gln causes

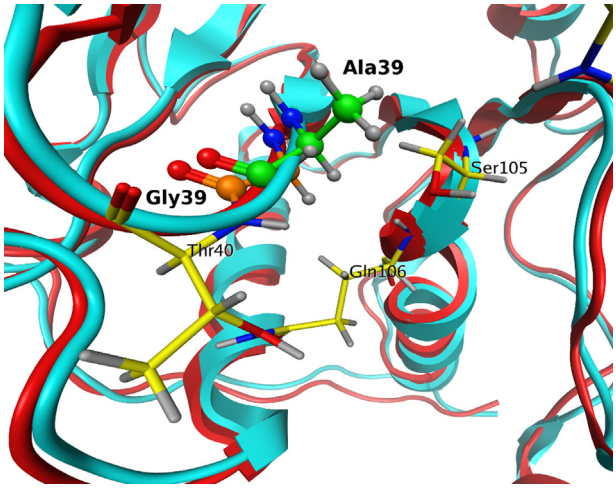

Fig. 6. Detail of the wild type CaLB (red) superposed to mutant D04 (G39A) in cyan. G39 is in orange, A39 is in green. Residues forming the catalytic triad and the oxyanion hole are in yellow. (For interpretation of the references to color in figure legend, the reader is referred to the web version of the article.)

not only a variation of the electrostatic factors in the surrounding region but also it produces steric effects due to its bulkiness.

In the case of the single mutation of Gly39 into Ala there is apparently a simpler interpretation of the increase of amidase activity: Fig. 6 shows how the bulkier Ala reduces the space available to the substrate.

As a consequence, the G39A mutation forces the substrate to assume a productive conformation, stabilized by a H-bond formed with Gln106 (Fig. 7), which favours the nucleophilic attack of Ser 105 to the acyl carbon. The distance between the nucleophilic oxygen of Ser105 and the amidic nitrogen is reduced from 3.1 Å to 2.9 Å. Moreover, the oxyanion stabilization is increased by the reduced distance between the H-bond donors and the substrate oxygen (from 3.4 and 1.9 Å to 2.1 and 1.6 Å; Fig. 7). However, it would be difficult to predict and explain why the combination of this mutation with others (for example D24, D30) produces a negative result as emerged from Table 1.

3.6. Automatic procedure for in-silico virtual screening

The credibility of any *in silico* engineering strategy at industrial scale depends strongly on the time scale of its application. Ideally, *in silico* design and screening should be feasible by using a moderate computational power and within a highthroughput frame. On that respect, the automatic generation of virtual mutants requires effective integration of all necessary software into a coherent work-flow and efficient *in silico* screening depends on reliable scoring functions. In order to explore the possibility of developing a comprehensive computational tool for automatic *in silico* design and screening of mutants, the modeFRONTIER software [29] was used. It provides an environment for automation of different procedures or software. In the present case modeFRONTIER integrates in a single workflow the following programs: PyMOL [26] (generation of 3D-structure of mutants), GROMACS [22] (molecular dynamics, structure minimisation and equilibration) and GRID [27] (calculation of molecular descriptors); finally, also the mathematical equation of the 3D-QSAR model was integrated. A scheme of the workflow used for the process automation in modeFRONTIER is showed in Fig. 8.

The workflow used as an input a number of hot spots residues identified by the previous QSAR analysis and the corresponding virtual mutants were firstly generated and then automatically scored by means of the 3D-QSAR model.

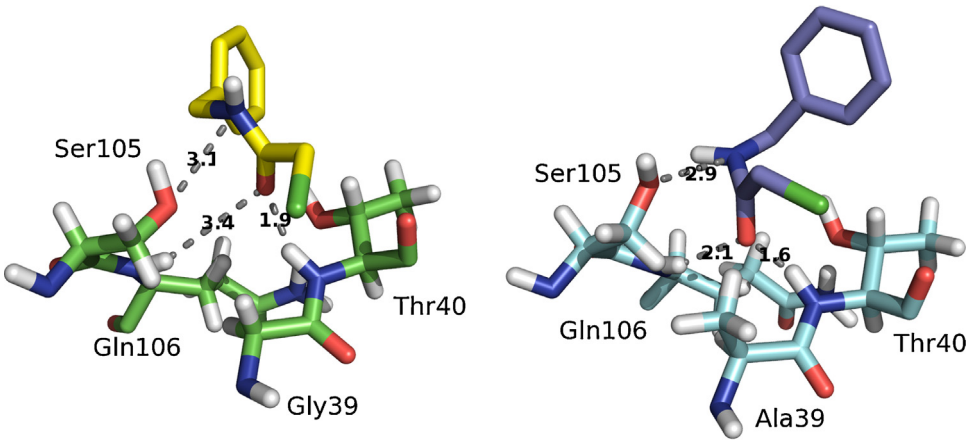

Fig. 7. Docking of *N*-Benzyl-2-chloroacetamide in wild type CaLB (left) and in D04 (right).

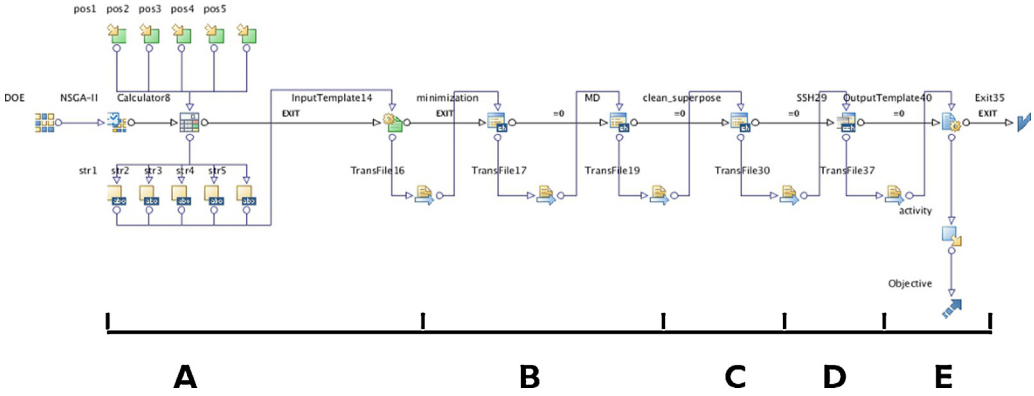

Fig. 8. Workflow used for the automatic *in-silico* generation and screening of mutants. From left to right: selection of mutations and mutant generation, mutant minimisation and mutant scoring.

The workflow can be subdivided into the following steps (Fig. 8):

A. Selection of mutations and generation of the virtual mutants using the PyMOL software and CaLB structure as a template (see Experimental).
B. Equilibration of each mutant by a minimisation and MD simulation procedure performed with GROMACS.
C. Superimposition of each mutant to the WT CaLB structure (PyMOL).
D. Automatic calculation of the molecular descriptors ($dMIF_S$) using the software GRID.
E. Scoring of each virtually generated mutant by applying the mathematical 3D-QSAR model to the matrix of the descriptors.

The modeFRONTIER software computes randomly a first generation of 20 mutants (the number of mutants for each generation can be modified by the user), which were scored. The scoring results of the first generation were then exploited for the calculation of the next 20 mutants using a genetic algorithm that is already integrated into the modeFRONTIER software (NSGA II). The use of the genetic algorithm makes modeFRONTIER able to learn generation after generation which are the empiric rules necessary to increase the amidases activity of CaLB (hydrolysis of *N*-Benzyl-2-chloroacetamide).

The automatic workflow generates and scores each virtual mutant in 2 h on a normal workstation and, in principle, the modeFRONTIER software can compute generations of mutants until the established convergence criteria are achieved.

Therefore, although preliminary, these results support the possibility to develop computational tools able to rationally design, generate and screen virtual mutants within a hightroughput scheme.

4. Conclusions

This study represents, at the best of our knowledge, the first example of 3D-QSAR strategy applied to enzyme rational engineering. The 3D-QSAR analysis provides a tool for approaching protein engineering through the comparisons of three-dimensional structures rather than sequences. The problem of engineering amidase activity inside a lipase (CaLB) scaffold was faced by taking into account that each mutation affects the properties of a mutant at multiple levels, leading to a complex, not linear, combination of factors. Consequently, simple visual inspection was considered inadequate for identifying structural motifs responsible for a certain property. Instead, multivariate statistical analysis (PLS) was applied because it has no conceptual bias in the interpretation of data and it simplifies the representation of results. Remarkably, the model takes into account also the interactions between variables, which are of major importance in such complex systems.

The methodology identified variables affecting the amidase activity of a set of CaLB mutants towards *N*-Benzyl-2-chloroacetamide and in particular the major role of hydratability of the active site. Specific hot spots were pointed out and they should be addressed for further improvement of mutants in terms of amidase activity.

Regarding the applicability of the 3D-QSAR approach to *in silico* screening of mutants the model here reported is able to discriminate between poor and good mutants. More importantly, by integrating all computational and statistical procedures inside a modeFRONTIER workflow it was demonstrated that it is possible to automate the whole methodology. This is of major importance in the perspective of a realistic application of the virtual screening process at industrial scale.

Overall, the present study underlines also some of the constraints of the regression approach: the model can predict a property as long as it has been trained with the information relevant for describing such property. Accordingly, highly active mutants can be predicted and selected only by starting from a training set of mutants with a broad range of activity values. Therefore, the methodology appears particularly suited for optimizing and tuning engineering strategies once an appropriate library of mutants is made available.

Acknowledgments

This work has received funding from the European Community's Seventh Framework Programme under the FP7-KBBE-2008-2B grant agreement no. 227279.

We are grateful to Dr. Danilo Di Stefano and Dr. Lorena Knapic for useful discussions.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.molcatb.2013.12.004.

References

[1] M.T. Reetz, J.D. Carballeira, Nature Protocols 2 (2007) 891–903.
[2] T. Nagai, A. Sawano, E.S. Park, A. Miyawaki, Proceedings of the Natural Academy of Sciences of the United States of America 98 (2001) 3197–3202.
[3] C.A. Ouzounis, A. Valencia, Bioinformatics 19 (2002) 2176–2190.
[4] F. Claeyssens, J. Harvey, F. Manby, R. Mata, A. Mulholland, K.E. Ranaghan, M. Schutz, S. Thiel, W. Thiel, H.J. Werner, Angewandte Chemie International Edition 118 (2006) 7010–7013.
[5] J. Parks, H. Hu, J. Rudolph, W. Yang, The Journal of Physical Chemistry B 113 (2009) 5217–5224.
[6] J. Hermann, J. Pradon, J. Harvey, A. Mulholland, The Journal of Physical Chemistry A 113 (2009) 11984–11994.
[7] L. Noodleman, T. Lovell, W. Han, J. Li, F. Himo, Chemical Reviews 104 (2004) 459–508.
[8] P. Sirén, K. Hult, ChemCatChem 3 (2011) 853–860.
[9] L. Tian, R. Friesner, Journal of Chemical Theory and Computation 5 (2009) 1421–1431.
[10] M.R. Hediger, L. De Vico, A. Svendsen, W. Besenmatter, J.H. Jensen, PLoS ONE 7 (2012) e49849, http://dx.doi.org/10.1371/journal.pone.0049849.
[11] P. Braiuca, C. Ebert, A. Basso, P. Linda, L. Gardossi, Trends in Biotechnology 24 (2009) 419–425.
[12] S. Wold, K. Esbensen, P. Geladi, Chemometrics and Intelligent Laboratory Systems 2 (1987) 37–52.
[13] S. Wold, PLS for linear modelling, in: D. van de Waterbeem (Ed.), Chemometric Methods in Molecular Design, vol. 2, VCH Verlagsgesellschaft, Weinheim, 1995, pp. 195–218.
[14] S. Wold, Chemometrics and Intelligent Laboratory Systems 58 (2001) 109–130.
[15] G. Cruciani, S. Clementi, M. Baroni, 3D QSAR in Drug Design: Theory, Methods and Applications, ESCOM, Leiden, 1993, pp. 567–582.
[16] P. Braiuca, A. Buthe, C. Ebert, P. Linda, L. Gardossi, Biotechnology Journal 2 (2007) 214–220.
[17] V. Ferrario, P. Braiuca, P. Tessaro, L. Knapic, C. Gruber, J. Pleiss, C. Ebert, E. Eichhorn, L. Gardossi, Journal of Biomolecular Structure and Dynamics 30 (2012) 74–88.
[18] P. Braiuca, L. Boscarol, C. Ebert, P. Linda, L. Gardossi, Advanced Synthesis and Catalysis 348 (2006) 773–780.
[19] P. Braiuca, L. Knapic, V. Ferrario, C. Ebert, L. Gardossi, Advanced Synthesis and Catalysis 351 (2009) 1293–1302.
[20] Y. Nakagawa, A. Hasegawa, J. Hiratake, K. Sakata, Protein Engineering Design & Selection 20 (2007) 339–346.
[21] D.A. Suplatov, W. Besenmatter, V.K. Svedas, A. Svendsen, Protein Engineering, Design & Selection 25 (2012) 689–697.
[22] H.J.C. Berendsen, D. van der Spoel, R. van Drunen, Computer Physics Communications 91 (1995) 43–56.
[23] J. Uppenberg, M.T. Hansen, S. Patkar, T.A. Jones, Structure 2 (1994) 293–308.
[24] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, Nucleic Acids Research 28 (2000) 235–242.
[25] C. Oostenbrink, A. Villa, A.E. Mark, W.F. van Gunsteren, Journal of Computational Chemistry 25 (2004) 1656–1676.
[26] The PyMOL Molecular Graphics System, Version 1.5.0.3 Schrödinger, LLC.
[27] P.J. Goodford, Journal of Medicinal Chemistry 28 (1985) 849–857.
[28] Golpe, MIA s.r.l. (www.miasrl.com).
[29] modeFRONTIER, ESTECO s.p.a. (www.esteco.com).
[30] M. Skjot, L. De Maria, R. Chatterjee, A. Svendsen, S.A. Patkar, P.R. Ostergaard, J. Brask, ChemBioChem 10 (2009) 520–527.
[31] V. Ferrario, C. Ebert, L. Knapic, D. Fattor, A. Basso, P. Spizzo, L. Gardossi, Advanced Synthesis and Catalysis 253 (2011) 2466–2480.
[32] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley-VCH, Weinheim, 2000.
[33] M. Pastor, G. Cruciani, I. McLay, S. Pikett, S. Clementi, Journal of Medicinal Chemistry 43 (2000) 3233–3243.
[34] A. Basso, L. De Martin, C. Ebert, L. Gardossi, P. Linda, V. Zlatev, Journal of Molecular Catalysis B: Enzymatic 11 (2001) 851–855.
[35] P. Syrén, P. Hendil-Forssell, L. Aumailley, W. Besenmatter, F. Gounine, A. Svendsen, M. Martinelle, K. Hult, ChemBioChem 13 (2012) 645–648.