Quantification of the Influence of Single-Point Mutations on Haloalkane Dehalogenase Activity: A Molecular Quantum Similarity Study

David Robert, Xavier Gironés, and Ramon Carbó-Dorca*

Institute of Computational Chemistry, University of Girona, Campus Montilivi, 17071 Girona, Catalonia, Spain

Received September 28, 1999

Controlled modifications in certain protein amino acid residues can lead to changes in their function and stability. Amino acid structural features and their relation to these changes were examined by using quantum molecular similarity techniques. The effect of deliberate mutations in position 172 of the haloalkane dehalogenase enzyme, yielding to variations on the dehalogenation of 1,2-dibromoethane, was studied qualitatively and quantitatively using molecular quantum similarity techniques. A valuable classification of the residues according to their effect on activiy was obtained by representing the optimal two-dimensional classical scaling solution. In addition, satisfactory quantitative relationships were found, comparable to those attained by previous studies on this same data set using other techniques. Molecular quantum similarity analysis provides a consistent, unbiased, and homogeneous set of molecular descriptors and is a feasible alternative to the use of physicochemical properties.

INTRODUCTION

One of the most promising subjects in protein engineering is the design of enzymes with desired properties for certain chemical reactions.[1] Site-directed mutagenesis experiments have become a useful tool in evaluating the importance of a particular amino acid residue for protein activity, specificity, or stability.[2] The results of these experiments are usually interpreted qualitatively and empirically. As an alternative to this, quantitative structure−function (QSFR) or structure−stability (QSSR) relationship models have been tested on different proteins using various approaches.[3−7] These methodologies are in fact the analogue of the well-known QSAR framework. So far, QSFR and QSSR analyses have been performed using a combination of physicochemical properties as descriptors. In the present study, molecular quantum similarity techniques are proposed as an alternative procedure.

QSFR/QSSR models must be clearly differentiated from the usual QSAR models. There are some analogies between them, but they are, in certain aspects, opposite processes. In QSFR/QSSR, the ligand is fixed and the enzyme is temporarily modified by changing an amino acid residue at a relevant position for a given activity. Alternatively, in QSAR the receptor is fixed and the activity of a ligand family is analyzed. In the latter case, the different ligands are usually designed by modifying a small number of substituents.

In this study, the effect of single-point mutations in haloalkane dehalogenase (DHLA) on the dehalogenation of the 1,2-dibromoethane is examined. DHLA isolated from *Xanthobacter Autotrophicus* GJ10 is a single polypeptide chain enzyme of 310 amino acids (molecular weight about 36 000 D) that participates in the degradation of halogenated aliphatic compounds. The primary substrate of DHLA is 1,2-dichloroethane. The enzyme converts halogenated aliphatic compounds to their corresponding alcohols via the formation of a covalent alkyl−enzyme intermediate, which is subsequently hydrolyzed. Crystallographic analysis[8−12] has provided essential information on the structure and the reaction mechanism of this enzyme. DHLA is composed of two domains: domain I has an α/β type structure with a central eight-stranded mainly parallel β sheet. Domain II lies such as a cap on top of domain I and consists of α-helices connected by loops. The active site is completely buried in an internal hydrophobic cavity, which is located between the two domains.

Site-directed mutagenesis searches for an improvement in the properties of a protein by modifying single amino acid residues that are relevant to the enzymatic reaction.[13] Several site-directed mutagenesis experiments have already been performed on DHLA.[14−19] In particular, the amino acid residue Phe172 was one of the selected targets for mutations.[17,18] Phe172 is relevant because it is located inside the enzyme's active-site cavity and interacts with the substrate in the Michaelis−Menten complex during the catalysis. Schanstra et al.[18] suggested that Phe172 was involved in the stabilization of the helix−loop−helix structure that covers the active site of the enzyme and creates a rigid hydrophobic cavity for small apolar halogenated alkanes.

In the present study, the QSFR for the wild-type and 15 different mutations of DHLA are built using molecular quantum similarity techniques.[20−26] This methodology is based upon quantitative comparative measures between molecular density functions and has been applied successfully to QSAR within pharmacological[27−32] and toxicological[33,34] environments. Quantum similarity measures provide a suitable quantification of the resemblance between two molecular electron distributions, and the similarity matrices can then be manipulated to generate QSAR parameters. Quantum similarity methodology is particularly appropriate when

* Corresponding author. Phone: 34 972 418359. Fax: 34 972 418356. E-mail: director@iqc.udg.es.

10.1021/ci9903408 CCC: $19.00
Published on Web 02/01/2000

homogeneous molecular series are examined and when steric effects participate in the interaction mechanisms.

MATERIALS AND METHODS

Molecular Quantum Similarity Measures (MQSM) and Carbó Indices. The practical basis of MQSM is the use of first-order molecular density functions. First-order density functions are quantum-mechanical observable elements that produce information on the electron distribution of the molecules. Within this framework, two molecules are considered to be similar if their electron distributions also are. Thus, among other possibilities,[22] a quantitative measure of the resemblance between two molecules can be defined as the volume integral between their density functions, weighted by the Coulomb operator $|\mathbf{r}_1 - \mathbf{r}_2|^{-1}$:

$$Z_{AB} = \int \frac{\rho_A(\mathbf{r}_1)\,\rho_B(\mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|}\,d\mathbf{r}_1\,d\mathbf{r}_2 \qquad (1)$$

where $\rho_A(\mathbf{r}_1)$ and $\rho_B(\mathbf{r}_2)$ are the electron density functions of molecules *A* and *B*, respectively. The Coulomb operator lessens the density peaks and makes the MQSM less sensitive to the relative atomic positions. The overall set of MQSM for a series of molecules can be collected in a similarity matrix, which can be used to extract information and build QSAR descriptors. A possible scaling of the MQSM is done by means of the transformation:

$$C_{AB} = Z_{AB}/[Z_{AA}Z_{BB}]^{1/2} \qquad (2)$$

This gives the so-called Carbó index.[20] Carbó indices are comprised within the interval [0,1]. The closer to one the index is, the more similar to each other the compounds are.

To avoid expensive theoretical calculations, the *promolecular atomic shell approximation* (ASA)[35−37] was used to construct molecular electron density. It views molecular density as a sum of discrete atomic density contributions, which are taken as 1S Gaussian functions and fitted to atomic ab initio ones. Once the overall atomic densities are built, each molecular density function can be constructed by adding as appropriate these elementary pieces. Since it has been proved that the MQSM from molecular densities built in this way differ by less than 1% from the ab initio densities,[36] their use is clearly justified.

Treatment of the Quantum Similarity Matrices. Common chemometric tools were applied here to deal with quantum similarity matrices. Classical scaling[38] is an appropriate technique to reduce the dimensionality of (dis)-similarity data. This method considers the objects as points in a multidimensional space and then finds coordinates such that the interpoint distances match as well as possible the original (dis)similarities. Supposing that a coordinate matrix **X** is known for a set of *n* points in an Euclidean space. Then, a squared distance matrix $\mathbf{D}^{(2)}$ can be constructed as

$$\mathbf{D}^{(2)} = \mathbf{c1}^T + \mathbf{1c}^T - 2\mathbf{XX}^T \qquad (3)$$

being **1** an *n*-dimensional vector of ones and **c** a vector possessing the diagonal elements of $\mathbf{XX}^T$ as components. Classical scaling solutions are invariant under translations of the whole data, and therefore a center of coordinates needs to be chosen. The center of coordinates is usually placed at the centroid of the data, by using the centering matrix $\mathbf{J} = \mathbf{I} - n^{-1}\mathbf{11}^T$. Centering and multiplying both sides by $-1/2$ transforms eq 3 into

$$-1/2\,\mathbf{JD}^{(2)}\,\mathbf{J} = \mathbf{XX}^T \qquad (4)$$

The first two terms of eq 3 vanish because centering a vector of ones yields a vector of zeros. Furthermore, the centering on $\mathbf{XX}^T$ has no influence at all, because it is considered to be centered. Once $\mathbf{XX}^T$ is constructed, recovering coordinate matrix **X** can be done by means of the spectral decomposition:

$$\mathbf{XX}^T = \mathbf{VAV}^T \qquad (5)$$

where **V** contains the eigenvectors of $\mathbf{XX}^T$ and **A** is a diagonal matrix which has the eigenvalues of $\mathbf{XX}^T$ as no-null elements. Then, coordinate matrix **X** is simply

$$\mathbf{X} = \mathbf{VA}^{1/2} \qquad (6)$$

Taking the Carbó index matrix **C** as the starting point, the precedent steps allow one to derive the coordinate matrix **X** in the new multidimensional space. The coordinates in this space are called the *principal coordinates* (PCs) of the system and constitute the molecular descriptors used in this work. Each PC is associated with an eigenvalue, indicating the explained variance of the axis. All those PCs accounting for less than 3% variance were a priori neglected, so the dimension of the problem was effectively reduced from 16 to 10 variables. Furthermore, the optimal variables to describe the desired property are not necessarily those that account for the maximal variance, so a simple variable selection technique was adopted. The most predictive variables method (MPVM)[39] quantifies the importance of the descriptors simply by projecting each PC on the external data.

Quantitative models are built by means of multilinear regression using the selected PCs as independent *X*-variables. The goodness-of-fit is assessed with the conventional r^2 and standard deviation σ_N coefficients. The models were validated by leave-one-out cross-validation.[40] The cross-validation technique consists of removing one object from the set and then recalculating the predictive model with the remaining objects. This new model is then used to predict the property value for the extracted element. This process is repeated for all the objects of the set, and then a coefficient of prediction q^2 can be defined from the squared cross-validation residuals (PRESS). The coefficient of prediction measures the robustness of the models, and a value $q^2 > 0.5$ is usually accepted as satisfactory.

Finally, the *randomization test*[41] was adopted to detect possible chance correlations. In this statistical technique, the dependent *Y*-variables are randomly permuted in their positions, and new predictive models are built with the altered vectors. If a real structure−property exists, the only satisfactory results should be obtained for the correctly ordered *Y*-variables. Otherwise, even if the obtained model seems to correctly describe the system, it cannot be considered significant, because it is built up with an excess of parameters, able to correlate any data set.

RESULTS AND DISCUSSION

The aim of quantitative structure−function relationships (QSFR) is to establish a simple mathematical relation

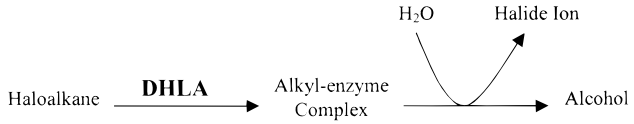

Figure 1. General diagram of dehalogenation.

Table 1. Amino acids, Acronyms, and Activity for DHLA-172 Mutants

amino acid	acronym	activity[a] (*k*)	−log *k*
alanine	Ala	0.1	1.000
arginine	Arg	0.01	2.000
asparagine	Asn	0.07	1.155
aspartate	Asp		
cysteine	Cys		
glutamine	Gln		
glutamic acid	Glu	0.03	1.523
glycine	Gly	0.04	1.398
histidine	His	0.88	0.056
isoleucine	Ile	0.05	1.301
leucine	Leu	0.04	1.398
lysine	Lys		
methionine	Met	1.57	−0.196
phenylalanine	Phe	2.32	−0.365
proline	Pro	0.02	1.699
serine	Ser	0.09	1.046
threonine	Thr	0.01	2.000
tryptophan	Trp	5.46	−0.737
tyrosine	Tyr	6.19	−0.792
valine	Val	0.08	1.097

[a] Data from ref 18.

between the function (activity, specificity) and a few molecular descriptors related to the modifications of the protein structure. As has been stated, the protein analyzed in the current study is haloalkane dehalogenase (DHLA). DHLA converts haloalkanes into their corresponding alcohols and halides. The reaction mechanism involves the formation of a covalent alkyl−enzyme complex which is hydrolyzed. This process is shown in Figure 1.

DHLA suffered controlled single-point mutations at position 172, where the wild-type amino acid residue, phenylalanine, was deliberately changed by 15 other ones.[18] An attempt is made to describe computationally the effect of these changes in the dehalogenation of the 1,2-dibromoethane molecule.

This section is structured as follows: first, a preliminary comparative structural study, to determine which computational method best models the mutant geometries. Then, a quantitative analysis was performed, validated with the usual methods and compared with the previous studies on this same data set. Finally, a qualitative data analysis was carried out, in which discrete classes were differentiated according to protein function.

Activity Data. In the present case, the analyzed property (*Y*-variables) is the DHLA activity of 15 mutants of the protein in position 172, including the original structure. According to Schanstra et al.,[17,18] mutants were built using the site-specific mutagenesis method of Kunkel,[42] and their activities were referred to 1,2-dibromoethane at a 5 mM substrate concentration.[18] The released bromine ions[43] were spectrophotometrically measured during the reaction and allowed for the quantification of the dehalogenation activity. Table 1 shows the molecular set, where the amino acid acronyms and their activities are also given. Only mutant

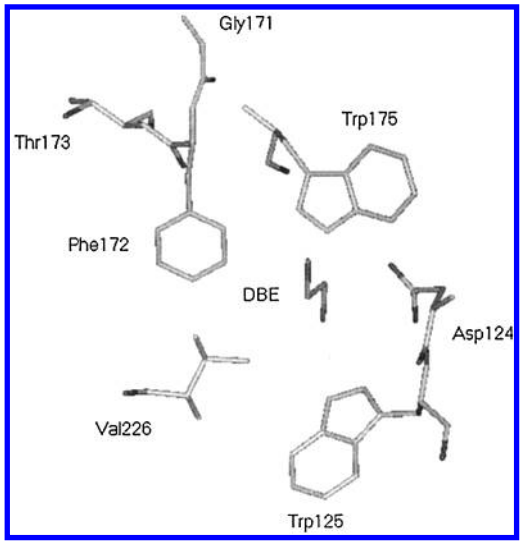

Figure 2. Active site model of DHLA used in the computational site-directed mutagenesis study. Hydrogens have been suppressed for clarity.

enzymes in which Phe172 was substituted by Tyr, Trp, Met and His led to any appreciable activity. The rest of the mutations were not favorable for dehalogenation. The minus logarithmic form of the activity was adopted to properly scale the data. In this way, any large differences in the activity data were reduced. Amino acid residues with undetermined activity were not analyzed.

Preliminary Structural Study: Modeling of Single-Point Mutants. The determination of the bioactive conformations is one of the most important problems in 3D-QSAR studies. In this particular case, the experimental structures for all the mutants in DHLA are not available, and X-ray crystallographic data exist only for the wild-type.[8,9] This structure is denoted by 2DHC in the Brookhaven Protein Database reference code,[44] and was used as a template to determine which computational method should be employed in order to best represent the bioactive geometries for the mutant residues. MQSM will be used to compare experimental and computational structures and will establish a criterion for selecting a particular methodology. Other studies using this strategy have been reported elsewhere.[30,37]

Starting from this X-ray structure, the wild-type Phe172 residue was optimized geometrically with different computational approaches. In order to reproduce as well as possible the real conditions, the active site of DHLA was computationally simulated. The active site is composed of four amino acids: Asp124, Trp125, Trp175, and Val226; together with a substrate molecule, 1,2-dibromoethane (DBE).[18] The crystal structure which served as a template for the optimization process contained 1,2-dichloroethane (DCE) as a substrate molecule. DCE was transformed simply into DBE by replacing the chlorine atoms by bromine atoms. The new C−Br bond lengths were optimized. The adjacent amino acids Gly171 and Thr173 were also included in the study, in order to restrict the spatial position of Phe172 and the remaining mutants. Uncompleted valences were filled with hydrogen atoms. Figure 2 shows all the structures considered in the computational model.

The optimization procedure was restricted to the variable substructure Phe172, while the remaining residues, which act as a force field for the targets, were kept frozen. The

Table 2. Upper Triangle of the Carbó Index Matrix for the Phe172 Residue Used to Compare the Different Computational Optimization Methodologies

	MM+	AMBER	PM3	AM1	X-ray
MM+	1.000	0.942	0.958	0.935	0.950
AMBER		1.000	0.946	0.958	0.916
PM3			1.000	0.902	0.954
AM1				1.000	0.909
X-ray					1.000

quantum-chemical methodologies employed here were as follows: molecular mechanics with a MM+ force field (an extension of Allinger's MM2 force field,[45] including high-order terms in bonds and angles description); molecular mechanics with AMBER electrostatic interactions;[46,47] a semiempirical calculation with the AM1 Hamiltonian[48] and its variant PM3,[49,50] parametrized with ab initio data rather than experimental data. The calculations were performed using HyperChem[51] for molecular mechanics and AMPAC 6.01[52] for semiempirical calculations. Ab initio calculations were a priori discarded because molecular structures with such a large number of atoms cannot be computed in a reasonable time.

Differences between structures were not quantified with the usual average root mean square errors, but employing MQSM. The fragment chosen for comparison was only made up of the Phe172 amino acid residue, calculated by means of the different theoretical procedures. These structures were aligned in such a way that the adjacent peptides were exactly superimposed. The overall pairwise Coulomb MQSM were then evaluated and transformed into Carbó indices. By definition, the closer to one that the Carbó index is, the more similar will be the electron distribution of the two molecules. The Carbó index matrix comparing the different computational methodologies for Phe172 is given in Table 2.

In general, all the selected methods produce acceptable geometries. Closeness between the experimental geometry and the energy minimum for the different computational approaches seems to indicate that the most relevant molecular interactions are well-reproduced and, therefore, that the active site has been correctly modeled. In more detail, AMBER and AM1 methods produce the poorest results. AMBER was originally developed for macromolecules, and it might fail for lack of parameters in other systems. Furthermore, it is accepted that AM1 does not describe satisfactorily long-range interactions. On the other hand, MM+ and PM3 generate valuable geometries. PM3 produces the smallest distortion in the structure, as indicated by the Carbó index value: $C_{\text{Xray}-\text{PM3}} = 0.954$. Even though AM1 and PM3 are both semiempirical methods, they produce very different results. PM3 is able to treat hydrogen bonding in a more realistic manner, as was proved with water dimer geometry prediction.[53] A slightly worse geometry, $C_{\text{Xray}-\text{MM+}} = 0.950$, is derived from molecular mechanics methods with the MM+ force field. MM+ was developed to produce very accurate results for certain classes of molecules, namely, small organic chemicals, rather than providing a generic approach. Since Lii et al.[54] proved its suitability for peptides too, this satisfactory description is not surprising. MM+ was taken here as the optimal computational source of mutant geometries, because the increment of computation time produced when using the semiempirical PM3 method does not justify the small improvement in the description, given by the difference in the Carbó indices. This choice is supported by the similarity between both descriptions, $C_{\text{MM+}-\text{PM3}} = 0.958$. Figure 3 shows the complex Gly171-Phe172-Thr173 obtained from both crystallographic analysis and theoretical MM+ calculation. As the figure and the similarity index reflect, the energy minimum for this parametrization lies very close to the X-ray template structure.

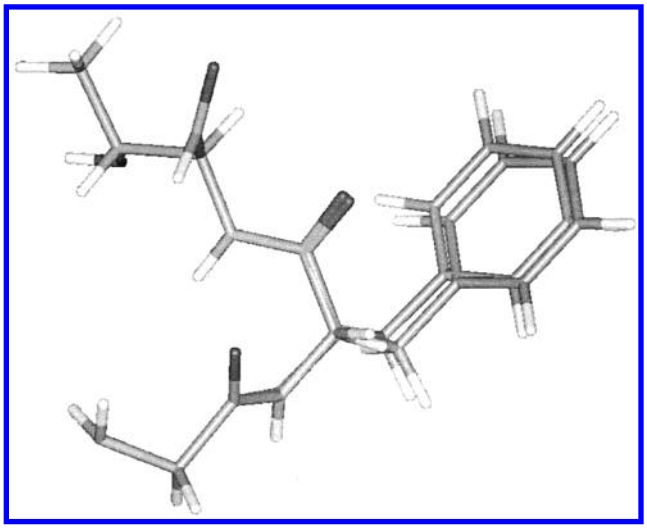

Figure 3. Gly171-Phe172-Thr173 tripeptide obtained from crystallographic analysis and computationally using the MM+ Hamiltonian.

The series of mutants were constructed using the Hyperchem "mutate" option over the selected residue. Afterward, a restricted MM+ optimization was carried out for each mutant.

Another computational site-directed mutagenesis study on DHLA was reported, using other homology modeling strategies.[55]

Molecular Alignment. The MQSM depend on the relative orientation of the two molecules, and therefore the QSAR results are sensitive to the alignment chosen. Even though only the amino acid side chains change from one mutant to another, here the whole the amino acid residue was taken to construct the molecular descriptor. This is because a set of common backbone atoms is required to properly align the compounds. As discussed in the previous section, the position of the neighboring amino acids determines the molecular alignment. Assuming that the nearest neighbors Gly171 and Thr173 exactly match for all the mutations, the different substitutions are then oriented in such a way that this condition is fulfilled. This procedure seeks to simulate the location of the amino acids within the cavity and avoids any type of optimization in the overlaying procedure. Figure 4 shows the alignment chosen, once the adjacent residues are removed. The figure shows how the common CO−CH−NH group is almost exactly superimposed for all the residues.

This common skeleton contributes in a similar way to the quantum similarity measure, and then the differences between the MQSM matrix elements provide information on the noncommon substructures, namely, the side chain. These noncommon regions are those considered to be responsible for the differences in protein function.

Quantitative Structure−Function Relationships. To quantify the influence on protein activity of deliberate

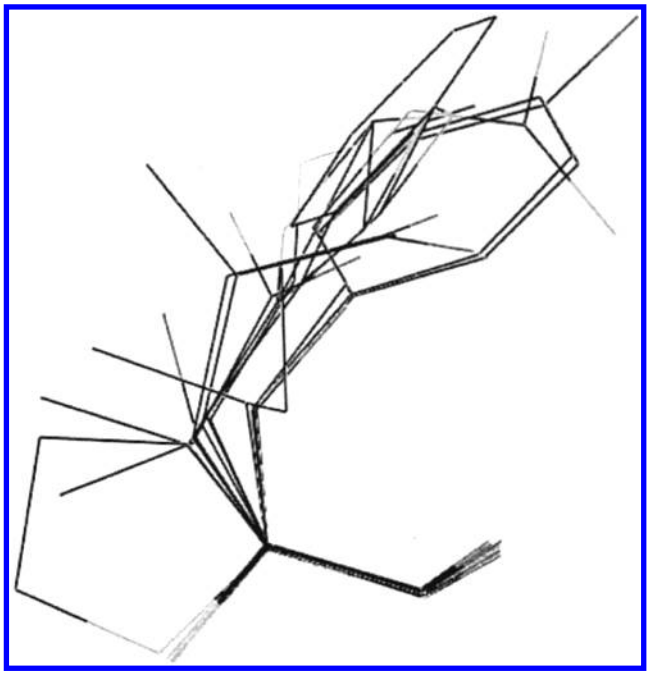

Figure 4. Alignment of the 16 amino acid residues studied. Hydrogens have been suppressed for clarity.

changes in its structure, the quantum similarity matrix deals with multivariate analysis techniques as detailed previously. Simple mathematical manipulations yield to reliable models that relate the molecular descriptors to the protein function.

Hydrophobicity descriptors have provided valuable QSFR/QSSR results in site-directed mutagenesis studies,[56] but in this particular case they appear to be not very descriptive. For instance, Phe and Leu possess equal hydrophobicity, log $P = -1.52$,[57] but these mutations lead to a very different catalytic activity: 2.32 and 0.04, respectively. As will be discussed below, DHLA mutations were satisfactorily predicted using shape descriptors. MQSM provide information on molecular shape, and therefore they could correctly estimate the property.

A quantitative relationship between the structural descriptors and the property can be established by means of a multilinear regression, as pointed out in the theoretical section. Variance distribution in quantum similarity data is not usually concentrated in a few PCs, as occurs in common data tables. Here, since MQSM are quite sensitive to small structural differences, a larger-dimensional subspace is necessary to explain a high percentage of variance. Figure 5 shows the evolution of the cumulated variance against the number of PCs. There are no elbows in the plot, usually present when using classical descriptors. A large number of PCs are necessary to reach a high value of the explained variance. This behavior justifies the use of variable selection techniques.

The selection of the optimal PCs, able to explain the differences in the activity, will include the actual data. The MPVM methodology sacrifices explained variance in order to get a better description of the property. Variable selection is necessary to choose those axes whose information best contributes to explaining the activity differences, and therefore the *X*-variables in the multilinear regression will be the PCs chosen according to the MPVM technique. The predictive models obtained are given in Table 3.

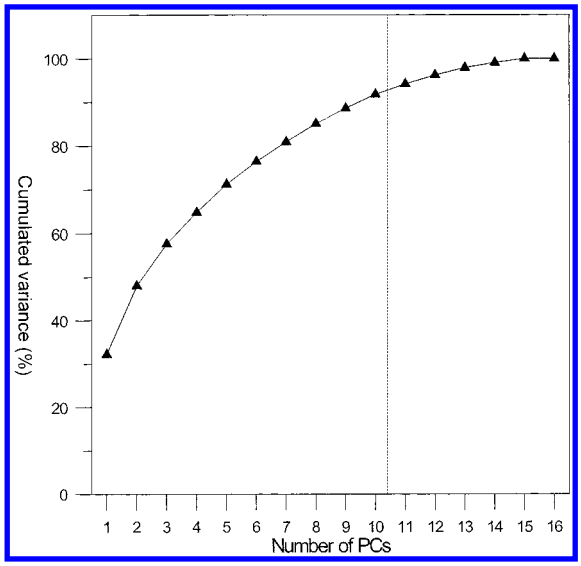

Figure 5. Cumulated variance (%) versus the number of PCs. Dashed line shows the variance threshold chosen.

Table 3. QSFR Models[a]

no. of PCs	selected PCs	r^2	q^2	σ_N
1	1	0.410	0.194	0.699
2	1, 8	0.688	0.426	0.508
3	1, 8, 6	0.809	0.563	0.398
4	*1, 8, 6, 9*	*0.887*	*0.746*	*0.306*
5	1, 8, 6, 9, 4	0.905	0.744	0.280

[a] The optimal model has been marked in italic.

Valuable models are achieved using few descriptors. Optimal values of $r^2 = 0.887$ and $q^2 = 0.746$ were obtained when using four PCs. As can be seen, the most descriptive PCs are not those accounting for maximal variance. This indicates that an intuitive shape−function relationship does not exist. The equation of the optimal model is

$$-\log k = -2.537\mathbf{x}_1 - 5.809\mathbf{x}_8 - 3.412\mathbf{x}_6 - 3.354\mathbf{x}_9 + 0.849 \quad (7)$$

Figure 6 shows the plot of the cross-validated activities versus the experimental ones for the four PCs model. Even though the properties are scaled, two isolated regions are still present in the data. This suggests that the correlation results can be artificially overrated. Separation into two well-differentiated classes recommends qualitative discrimination studies, as will be discussed below.

The present results can be compared with those obtained by J. Damborský in previous studies.[56,58] Two QSFR models were proposed by this author, using physicochemical properties as descriptors in a classical Hansch−Fujita analysis. This approach assumes that biological activity in a compound is produced by a combined effect of different factors, which act independently. These factors are modeled using physicochemical properties, quantum-chemical magnitudes, or indicator variables. The first model[58] took 33 properties from the AAindex amino acid database[59−61] which were optimally combined to generate satisfactory QSFR models. $r^2 = 0.843$ and $q^2 = 0.745$ values were obtained using four descriptors. The second model[56] used the full set of 402 variables from the AAindex, and the results were considerably improved: $r^2 = 0.827$ and $q^2 = 0.765$ with only three descriptors. In

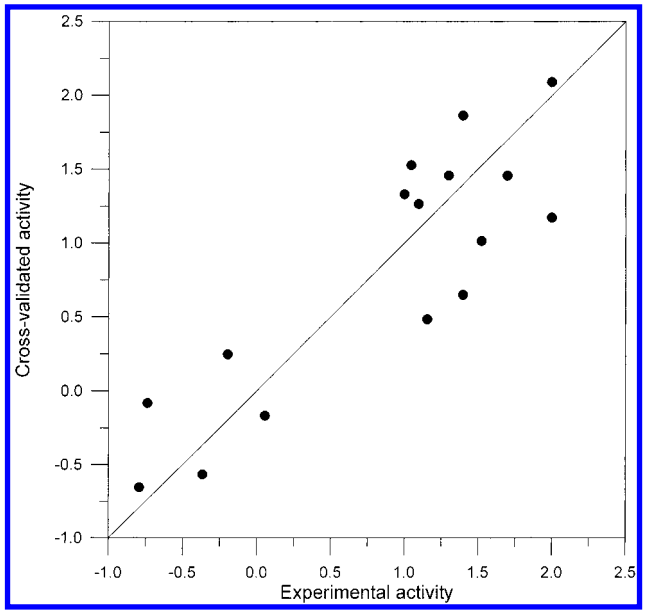

Figure 6. Experimental versus cross-validated activities for the 16 amino acids studied. Optimal QSFR model has been used.

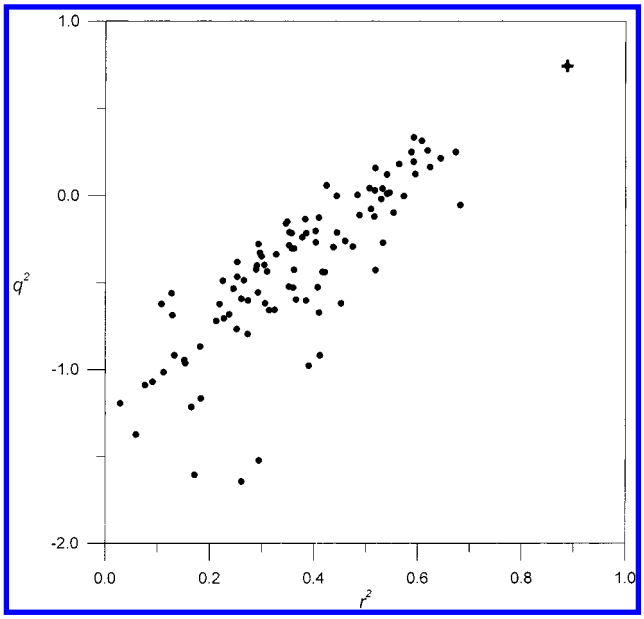

Figure 7. Randomization test for the optimal model. The randomized responses (100) have been marked with circles, and the correctly ordered activity has been marked with a cross.

Damborský's study, the property was not scaled, so the differences between the experimental values were really large and the correlation indices had even less significance. In both papers, a mechanistic interpretation of the model was discussed. The current approach provides comparable results regarding the number of descriptors of the database. This is one of the advantages of molecular quantum similarity techniques: they provide a consistent and homogeneous set of variables, in contrast to the classical approach. Nevertheless, the connection between these descriptors and molecular features, which is a posteriori in the classical approach, is less intuitive. It must be emphasized that some molecular features or properties that were used as optimal descriptors in the classical approach are in fact contained in the molecular electron distribution. For instance, the first Damborský QSFR model[58] used a discrete indicator descriptor which represented molecular aromaticity. Aromaticity is present in the molecular shape, and therefore it is included in the density functions. Another property used, the refractive index, correlates with molecular volume.[62] Volume information is clearly contained in the density function, and MQSM is able to extract it. The most important parameter in the second Damborský model[56] was the relative hydrophilicity R_f, descriptor *P383* in the AAindex database.[59,60] R_f is the relative hydrophilicity determined with a mixture of saturated ammonium sulfate and 1 M ammonium acetate.[63] This descriptor played the role of aromaticity of the previous model, in the sense that it generated the same molecular groups. Bulkiness,[64] *P399* in the AAindex database, is another descriptor used in both models. It is defined as the ratio of the side-chain volume to its length. This is another steric property, presumably contained in the density function information.

Finally, the randomization test was performed to assess that the relationships found were not due to an overparametrization of the model or background noise correlation. Thus, a hundred new models were built in the same optimal conditions (MPVM, four PCs), but the *Y*-variables were randomized permutations of the actual activities. The results of this test are compiled in Figure 7, where the axes are the multiple determination (r^2) and prediction (q^2) coefficients. As can be observed, a clear separation between the two situations exists. The real activity produces the best results, whereas none of the altered responses produce satisfactory models. All the randomized activities yield to $q^2 < 0.5$ values, and most of them are even negative. It can be concluded that a real QSFR has been discovered, and that no fortuitous correlations or overparametrization exist.

Qualitative Structure−Function Relationships: Analysis of the Similarity Subspaces. As commented on in the previous section, the particular distribution of the activity data recommends a qualitative structure−function study, in order to classify the mutations according to their effect on DHLA dehalogenation activity. This qualitative analysis is based simply on the observation of the grouping of the molecules once projected on the adequate similarity subspace. Since a structure−function relationship has been found, the molecules should be located in such a way that visual discrimination between the high and low active mutations could occur. The selection of the optimal similarity subspace was carried out by means of the MPVM method. As was seen in the previous analysis (see Table 3), the first and eighth PCs are the two optimal ones for activity. A plot showing the classical scaling solution in this subspace is presented in Figure 8. In this diagram, closeness of points may indicate similar activity, and in fact a clear separation between high and low active mutations is observed. The first PC (*X*-axis) gives a good approximation to the behavior of the property, except for the Phe172Arg mutant. The eighth PC (*Y*-axis) amends this misclassification, leading to a satisfactory discrimination. In addition, the two less favorable mutations, Phe172Arg and Phe172Thr, possess the lowest *Y*-projections. To better illustrate this fact, a discriminating line has been drawn, which separates the high ($\log k \geq 1.000$) and low ($\log k < 0.100$) active mutants. Even when the shape of the residues seems not to be straightforwardly connected with its activity, MQSM contains in some degree this information. Sometimes this simple binary discrimination is enough for the interests of protein engineers: a lot of work

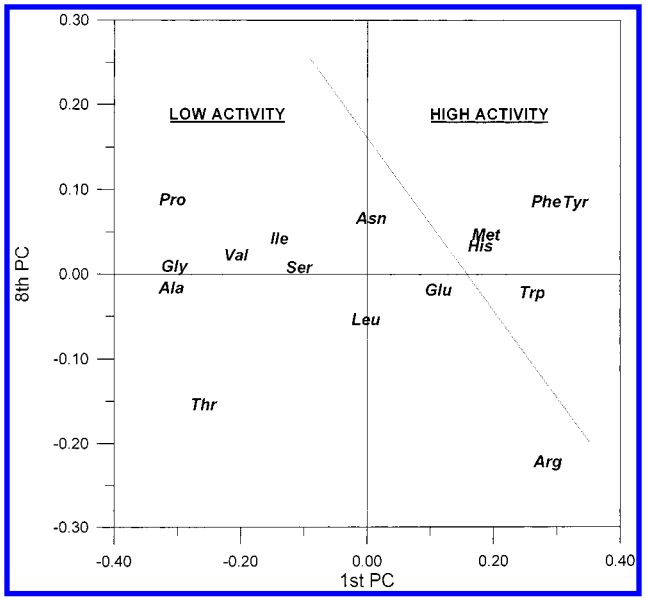

Figure 8. Two-dimensional classical scaling solution. PCs 1 and 8 have been used as axes. The dashed line discriminates between high and low active mutations.

could be saved if computational methodology was able a priori to classify a mutant into a discrete class, according to its effect on protein function.

CONCLUSIONS

In the present study, molecular quantum similarity was applied to correlate systematically the single-point mutations in haloalkane dehalogenase with its dehalogenation activity. Quantitative structure−function relationships (QSFR) provide valuable information on the role of amino acid residues in protein activity, specificity, or stability and might aid protein engineering experiments as potentially predictive tools.

Satisfactory qualitative discrimination between the high and low active mutations was achieved. Quantitative models were also constructed using principal coordinate regressions, achieving comparable results to previous QSFR studies on this same data set. The models passed the validation tests: leave-one-out cross-validation and randomization.

Quantum similarity is a suitable alternative to the use of physicochemical properties as descriptors in QSFR analysis when steric effects play a relevant role in the biological reaction. However, this approach has not to be considered as a completely general procedure: other factors, namely, hydrophobic, electrostatic, or orbital, are not described by this picture, but can intervene in the biochemical process and constitute the most relevant descriptors in the QSFR models.

ACKNOWLEDGMENT

This research was partially supported by the CICYT grant SAF 96-0158 and the European Commission Project No. ENV4-CT97-0508. X.G. benefited from a predoctoral fellowship from the University of Girona. Financial support from the *Fundació María Francisca de Roviralta* is also acknowledged. Thanks also to Prof. Jiří Damborský (Masaryk University, Czech Republic) for his kindness and enlightening comments, and to Lluís Amat (University of Girona) for discussion on the computational amino acid geometry modeling.

REFERENCES AND NOTES

(1) Oxender, D. L.; Fox, C. F., Eds. *Protein Engineering*., Alan R. Liss: New York, 1987; pp 221−224.
(2) Leatherbarrow, R. J.; Fersht, A. J. Protein Engineering. *Protein Eng*. **1986**, *1*, 7−16.
(3) Hellberg, S.; Sjostrom, M.; Skagerberg, B.; Wold, S. Peptide quantitative structure−activity relationships, a multivariate approach. *J. Med. Chem*. **1987**, *30*, 1126−1135.
(4) Kato, A.; Yutani, K. Correlation of surface properties with conformational stabilities of wild-type and six mutant tryptophan synthase alpha-subunits substituted at the same position. *Protein Eng*. **1988**, *2*, 153−156.
(5) Lee, C.; Levitt, M. Accurate prediction of the stability and activity effects of site-directed mutagenesis on a protein core. *Nature* **1991**, *352*, 448−451.
(6) Collantes, E. R.; Dunn, W. J., III. Amino acid side-chain descriptors for quantitative structure−activity relationship studies of peptide analogues. *J. Med. Chem*. **1995**, *38*, 2705−2713.
(7) Zbilut, J. P.; Giuliani, A.; Webber, C. L., Jr.; Colosimo, A. Recurrence quantification analysis in structure−function relationships of proteins: An overview of a general methodology applied to the case of TEM-1 beta-lactamase. *Protein Eng*. **1998**, *11*, 87−93.
(8) Rozeboom, H. J.; Kingma, J.; Janssen, D. B.; Dijkstra, B. W. Crystallization of haloalkane dehalogenase from Xanthobacter autotrophicus GJ10. *J. Mol. Biol*. **1988**, *200*, 611−612.
(9) Franken, S. M.; Rozeboom, H. J.; Kalk, K. H.; Dijkstra, B. W. Crystal structure of haloalkane dehalogenase: An enzyme to detoxify halogenated alkanes. *EMBO J*. **1991**, *10*, 1297−1302.
(10) Verschueren, K. H. G.; Franken, S. M.; Rozeboom, H. J.; Kalk, K. H.; Dijkstra, B. W. Refined X-ray structures of haloalkane dehalogenase at pH 6.2 and pH 8.2 and implications for the reaction mechanism. *J. Mol. Biol*. **1993**, *232*, 856−872.
(11) Verschueren, K. H. G.; Seljee, F.; Rozeboom, H. J.; Kalk, K. H.; Dijkstra, B. W. Crystallographic analysis of the catalytic mechanism of haloalkane dehalogenase. *Nature* **1993**, *363*, 693−698.
(12) Ridder, I. S.; Rozeboom, H. J.; Dijkstra, B. W. Haloalkane dehalogenase from *Xanthobacter autotrophicus* GJ10 refined at 1.15A. *Acta Crystallogr*. **1999**, *D55*, 1273−1290.
(13) Hutchison, C. A., III; Phillips, S.; Edgell, M. H.; Gillam, S.; Jahnke, P.; Smith, M. Mutagenesis at a specific position in a DNA sequence. *J. Biol. Chem*. **1978**, *253*, 6551−6560.
(14) Pries, F.; Kingma, J.; Pentega, M.; Van Pouderoyen, G.; Jeronimus-Stratingh, C. M.; Bruins, A. P.; Janssen, D. B. Site-directed mutagenesis and oxygen isotope incorporation studies of the nucleophilic aspartate of haloalkane dehalogenase. *Biochemistry* **1994**, *33*, 1242−1247.
(15) Pries, F.; Kingma, J.; Janssen, D. B. Activation of an Asp-124→Asn mutant of haloalkane dehalogenase by hydrolytic deamidation of asparagine. *FEBS Lett*. **1995**, *358*, 171−174.
(16) Pries, F.; Kingma, J.; Krooshof, G. H.; Jeronimus-Stratingh, C. M.; Bruins, A. P.; Janssen, D. B. Histidine 289 is essential for hydrolysis of the alkyl-enzyme intermediate of haloalkane dehalogenase. *J. Biol. Chem*. **1995**, *270*, 10405−10411.
(17) Schanstra, J. P.; Janssen, D. B. Kinetics of halide release of haloalkane dehalogenase: Evidence for a slow conformational change. *Biochemistry* **1996**, *35*, 5624−5632.
(18) Schanstra, J. P.; Ridder, I. S.; Heimericks, G. J.; Rink, R.; Poelarends, G. J.; Kalk, K. H.; Dijkstra, B. W.; Janssen, D. B. Kinetic characterization and X-ray structure of a mutant of haloalkane dehalogenase with higher catalytic activity and modified substrate range. *Biochemistry* **1996**, *35*, 13186−13195.
(19) Schanstra, J. P.; Ridder, A.; Kingma, J.; Janssen, D. B. Influence of mutation of Val226 on the catalytic rate of haloalkane dehalogenase. *Protein Eng*. **1997**, *10*, 53−61.
(20) Carbó, R.; Arnau, J.; Leyda, L. How similar is a molecule to another? An electron density measure of similarity between two molecular structures. *Int. J. Quantum Chem*. **1985**, *17*, 1185−1189.
(21) Carbó-Dorca, R.; Besalú, E. A general survey of molecular quantum similarity. *THEOCHEM* **1998**, *451*, 11−23.
(22) Carbó-Dorca, R.; Amat, L.; Besalú, E.; Lobato, M. Quantum Similarity. In *Advances in Molecular Similarity*; Carbó-Dorca, R.; Mezey, P. G., Eds.; JAI Press: Greenwich, CT, 1998; Vol. 2, pp 1−42.
(23) Carbó-Dorca, R. Fuzzy sets and boolean tagged sets; vector semispaces and convex sets; quantum similarity measures and ASA density functions; diagonal vector spaces and quantum chemistry. In *Advances in Molecular Similarity*; Carbó-Dorca, R., Mezey, P. G., Eds.; JAI Press: Greenwich, CT, 1998; Vol. 2, pp 43−72.

(24) Besalú, E.; Carbó, R.; Mestres, J.; Solà, M. Foundations and Recent Developments on Molecular Quantum Similarity. *Top. Curr. Chem.* **1995**, *173*, 31−62.
(25) Carbó, R.; Besalú, E. Theoretical Foundations of Quantum Molecular Similarity. In *Molecular Similarity and Reactivity: From Quantum Chemical to Phenomenological Approaches*; Carbó, R., Ed.; Kluwer: Amsterdam, 1995; pp 3−30.
(26) Carbó-Dorca, R.; Besalú, E.; Amat, L.; Fradera, X. Quantum Molecular Similarity Measures: Concepts, Definitions, and Applications to Quantitative Structure−Property Relationships. In *Advances in Molecular Similarity*; Carbó-Dorca, R., Mezey, P. G., Eds.; JAI Press: Greenwich, CT, 1996; Vol. 1, pp 1−42.
(27) Carbó, R.; Besalú, E.; Amat, L.; Fradera, X. Quantum molecular similarity measures (QMSM) as a natural way leading towards a theoretical foundation of quantitative structure-properties relationships. *J. Math. Chem.* **1995**, *18*, 237−246.
(28) Fradera, X.; Amat, L.; Besalú, E.; Carbó-Dorca, R. Application of Molecular Quantum Similarity to QSAR. *Quant. Struct.-Act. Relat.* **1997**, *16*, 25−32.
(29) Lobato, M.; Amat, L.; Carbó-Dorca, R. Structure−Activity Relationships of a Steroid Family using Quantum Similarity Measures and Topological Quantum Similarity Indices. *Quant. Struct.-Act. Relat.* **1997**, *16*, 465−472.
(30) Amat, L.; Robert, D.; Besalú, E.; Carbó-Dorca, R. Molecular Quantum Similarity Measures Tuned 3D QSAR: An Antitumoral Family Validation Study. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 624−631.
(31) Robert, D.; Amat, L.; Carbó-Dorca, R. Three-Dimensional Quantitative Structure−Activity Relationships from Tuned Molecular Quantum Similarity Measures. Prediction of the Corticosteroid-Binding Globulin Binding Affinity for a Steroid Family. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 333−344.
(32) Robert, D.; Gironés, X.; Carbó-Dorca, R. Facet Diagrams for Quantum Similarity Data. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 597−610.
(33) Robert, D.; Carbó-Dorca, R. Aromatic Compounds Aquatic Toxicity QSAR using Quantum Similarity Measures. *SAR QSAR Environ. Res.* **1999**, *10*, 401−422.
(34) Gironés, X.; Amat, L.; Carbó-Dorca, R. Molecular Quantum Similarity Measures Used to Characterize Molecular Toxicology. Institute of Computational Chemistry, Technical Report IT-IQC-98-29. Also: *SAR QSAR Environ. Res.*, in press.
(35) Constans, P.; Carbó, R. Atomic Shell Approximation: Electron Density Fitting Algorithm Restricting Coefficients to Positive Values. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1046−1053.
(36) Amat, L.; Carbó-Dorca, R. Quantum Similarity Measures under Atomic Shell Approximation: First-Order Density Fitting using Elementary Jacobi Rotations. *J. Comput. Chem.* **1997**, *18*, 2023−2039.
(37) Amat, L.; Carbó-Dorca, R. Fitted Electronic Density Functions from H to Rn for Use in Quantum Similarity Measures: *cis*-Diammine-Dichloroplatinum(II) Complex as an Application Example. *J. Comput. Chem.* **1999**, *20*, 911−920.
(38) Cox, T. F.; Cox, M. A. A. *Multidimensional Scaling*; Chapman & Hall: London, 1994.
(39) Cuadras, C. M.; Arenas, C. A distance based regression model for prediction with mixed data. *Commun. Stat. Theor. Method.* **1990**, *19*, 2261−2279.
(40) Allen, D. M. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* **1974**, *16*, 125−127.
(41) Wold, S.; Eriksson, L. Statistical validation of QSAR results. In *Chemometric Methods in Molecular Design;* Van der Waterbeemd, H., Ed.; VCH: New York, 1995; pp 309−318.
(42) Kunkel, T. A. Rapid and efficient site-specific mutagenesis without phenotypic selection. *Proc. Natl. Acad. Sci. U.S.A.* **1985**, *82*, 488−492.
(43) Keuning, S.; Janssen, D. B.; Witholt, B. Purification and characterization of hydrolytic haloalkane dehalogenase from *Xanthobacter autotrophicus* GJ10. *J. Bacteriol.* **1985**, *163*, 635−639.
(44) Allen, F. H.; Bellard, S.; Brice, M. D.; Cartwright, B. A.; Doubleday: A.; Higgs, H.; Hummelink, T.; Hummelink-Peters, B. G.; Kennard, O.; Motherwell, W. D. S.; Rodgers, J. R.; Watson, D. G. The Cambridge Crystal Data Centre: Computer-Based Search, Retrieval, Analysis and Display of Information. *Acta Crystallogr.* **1979**, *B35*, 2331−2339.
(45) Allinger, N. L. Conformational Analysis 130. MM2. A Hydrocarbon Force Field Utilizing V1 and V2 Torsional Terms. *J. Am. Chem. Soc.* **1977**, *99*, 8127−8134.
(46) Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta S., Jr.; Weiner, P. A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.* **1984**, *106*, 765−784.
(47) Weiner, S. J.; Kollman, P. A.; Nguyen, D. T.; Case, D. A. An all atom force field for simulations of proteins and nucleic acids. *J. Comput. Chem.* **1986**, *7*, 230−252.
(48) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A New General Purpose Quantum Chemical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.
(49) Stewart, J. J. P. Optimization of parameters for semiempirical methods. I. Method. *J. Comput. Chem.* **1989**, *10*, 209−220.
(50) Stewart, J. J. P. Optimization of parameters for semiempirical methods. II. Applications. *J. Comput. Chem.* **1989**, *10*, 221−264.
(51) *Hyperchem*, Release 3 for Windows. Molecular Modeling System; Hypercube, Inc. and Autodesk, Inc: Sausalito, CA, 1993.
(52) *AMPAC 6.01*, Semichem: Shawnee, KS, 1994.
(53) Coitiño, E. L.; Irving, K.; Rama, J.; Ventura, O. N. Theoretical Studies of Hydrogen-Bonded Complexes Using Semiempirical Methods. *J. Mol. Struct. (THEOCHEM)* **1990**, *69*, 405.
(54) Lii, J.-H.; Gallion, S.; Bender, C.; Wikström, H.; Allinger, N. L.; Flurchick, K. M.; Teeter, M. M. Molecular Mechanics (MM2) Calculations on Peptides and on the Protein Crambin Using the Cyber 205. *J. Comput. Chem.* **1989**, *10*, 503−513.
(55) Damborský, J.; Boháč, M.; Prokop, M.; Kutý, M.; Koča, J. Computational site-directed mutagenesis of haloalkane dehalogenase in position 172. *Protein Eng.* **1998**, *11*, 901−907.
(56) Damborský, J. Quantitative structure−function and structure−stability relationships of purposely modified proteins. *Protein Eng.* **1998**, *11*, 21−30.
(57) Hansch, C.; Leo, A.; Hoekman, D. *Exploring QSAR: Hydrophobic, Electronic, and Steric Constants*; American Chemical Society: Washington, D.C., 1995.
(58) Damborský, J. Quantitative structure-function relationships of the single-point mutants of haloalkane dehalogenase: A multivariate approach. *Quant. Struct.-Act. Relat.* **1997**, *16*, 126−135.
(59) Nakai, K.; Kidera A.; Kanehisa, M. Cluster analysis of amino acid indices for prediction of protein structure and function. *Protein Eng.* **1988**, *2*, 93−100.
(60) Tomii, K.; Kanehisa, M. Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Eng.* **1996**, *9*, 27−36.
(61) Kawashima, S.; Ogata, H.; Kanehisa, M. AAindex: Amino acid index database. *Nucleic Acids Res.* **1999**, *27*, 368−369.
(62) Jones, D. D. Amino acid properties and side-chain orientation in proteins: A cross correlation approach. *J. Theor. Biol.* **1975**, *50*, 167−183.
(63) Weber, A. L.; Lacey, J. C. Genetic code correlations: Amino acids and their anticodon nucleotides. *J. Mol. Evol.* **1978**, *11*, 199−210.
(64) Zimmerman, J. M.; Eliezer, N.; Simha, R. The characterization of amino acid sequences in proteins by statistical methods. *J. Theor. Biol.* **1968**, *21*, 170−201.

CI9903408