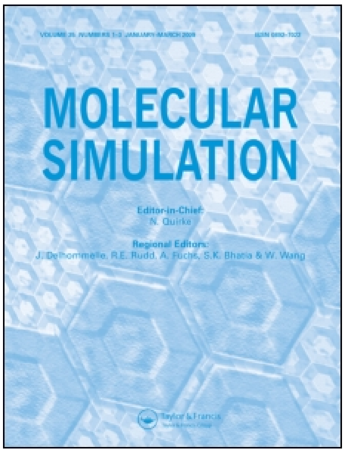

Molecular Simulation
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713644482

Proteometric modelling of protein conformational stability using amino acid sequence autocorrelation vectors and genetic algorithm-optimised support vector machines

Michael Fernández[ab]; Leyden Fernández[a]; Pedro Sánchez[ac]; Julio Caballero[d]; Jose Ignacio Abreu[ac]
[a] Molecular Modeling Group, Faculty of Agronomy, Center for Biotechnological Studies, University of Matanzas, Matanzas, Cuba [b] Department of Bioscience and Bioinformatics, Kyushu Institute of Technology (KIT), Iizuka, Fukuoka, Japan [c] Artificial Intelligence Lab, Faculty of Informatics, University of Matanzas, Matanzas, Cuba [d] Centro de Bioinformática y Simulación Molecular, Universidad de Talca, Talca, Chile

To cite this Article Fernández, Michael , Fernández, Leyden , Sánchez, Pedro , Caballero, Julio and Abreu, Jose Ignacio(2008) 'Proteometric modelling of protein conformational stability using amino acid sequence autocorrelation vectors and genetic algorithm-optimised support vector machines', Molecular Simulation, 34: 9, 857 — 872

To link to this Article: DOI: 10.1080/08927020802301920

URL: http://dx.doi.org/10.1080/08927020802301920

Molecular Simulation
Vol. 34, No. 9, August 2008, 857–872

Taylor & Francis
Taylor & Francis Group

Proteometric modelling of protein conformational stability using amino acid sequence autocorrelation vectors and genetic algorithm-optimised support vector machines

Michael Fernández[ab]*, Leyden Fernández[a], Pedro Sánchez[ac], Julio Caballero[d] and Jose Ignacio Abreu[ac]

[a]*Molecular Modeling Group, Faculty of Agronomy, Center for Biotechnological Studies, University of Matanzas, Matanzas, Cuba;* [b]*Department of Bioscience and Bioinformatics, Kyushu Institute of Technology (KIT), Iizuka, Fukuoka, Japan;* [c]*Artificial Intelligence Lab, Faculty of Informatics, University of Matanzas, Matanzas, Cuba;* [d]*Centro de Bioinformática y Simulación Molecular, Universidad de Talca, Talca, Chile*

(*Received 8 May 2008; final version received 23 June 2008*)

The conformational stability of more than 1500 protein mutants was modelled by a proteometric approach using amino acid sequence autocorrelation vector (AASA) formalism. 48 amino acid/residue properties selected from the AAindex database weighted the AASA vectors. Genetic algorithm-optimised support vector machine (GA-SVM), trained with subset of AASA descriptors, yielded predictive classification and regression models of unfolding Gibbs free energy change ($\Delta\Delta G$). Function mapping and binary SVM models correctly predicted about 50 and 80% of $\Delta\Delta G$ variances and signs in crossvalidation experiments, respectively. Test set prediction showed adequate accuracies about 70% for stable single and double point mutants. Conformational stability depended on autocorrelations at medium and long ranges in the mutant sequences of general structural, physico-chemical and thermodynamical properties relative to protein hydration process. A preliminary version of the predictor is available online at http://gibk21.bse.kyutech.ac.jp/llamosa/ddG-AASA/ddG_AASA.html.

Keywords: protein stability prediction; point mutations; kernel methods; structure–property relationship

1. Introduction

Current computational methods of assessing protein functions and properties are based to a large extent on prediction of sequence similarity of proteins with other proteins having known functions.

The accuracy of such predictions depends on the ability of the computational methods to extend sequence similarity to functional similarity [1]. Conventional approaches to molecular recognition have until now required determining protein 3D structures, which is resource-demanding, error-prone and generally requires prior knowledge such as 3D structure of a homologous protein. But the great gap between the amount of known protein sequences and the elucidated structures is a large drawback for the 3D-based protein function modelling [1].

Predicting protein structures and stability is a fundamental goal in molecular biology. Even predicting changes induced by point mutations has immediate application in computational protein design [2–4]. Although free energy simulations have accurately predicted relative stabilities of point mutants [5], the computational cost of the methods actually is extremely high to test the large number of mutations studied in protein design applications. Translation of structural data into energetic parameters is intended today by developing fast algorithms for protein energy calculations. However, the development of fast and reliable protein force-fields is a complex task due to the delicate balance between the different energy terms which contribute to protein stability. Force-fields for predicting protein stability can be divided in three main groups: physical effective energy function [7,8], statistical potential-based effective energy function [6,9,10] and empirical data-based energy function [11,12].

Furthermore, stability prediction studies not based on protein force-field calculations have been focused on correlations of free energy change with structural, sequence information and amino acid properties such as hydrophobicity, accessible surface area (ASA), etc. In this sense, Gromiha et al. had reported some of the seminal works in this topic [13–15]. On the other hand, empirical equations involving physical properties, have been calculated from mutant structures. Zhou and Zhou [16] reported a broad study regarding 35 proteins and 1023 mutants, from which they derived a new stability scale. A ‘transfer free energy’ scale was extracted assuming that the mutation-induced stability change is equal to the change in transfer free energy without needing any structural information.

In addition, some X-ray structural-independent stability prediction methods have gained attention. The advantages of such methods are that they just employ amino acid sequence information for predicting protein stability and that they are much less computationally intensive than free energy function methods [17,18]. Levin

*Corresponding author. Emails: michael.fernandez@umcc.cu; michael_llamosa@yahoo.com

ISSN 0892-7022 print/ISSN 1029-0435 online

DOI: 10.1080/08927020802301920
http://www.informaworld.com

and Satir [17] successfully evaluated the functional significance of mutations on hemoglobin by amino acid similarity matrixes. Frenz [18] reported a nonlinear model for predicting the stability of staphylococcal nuclease mutants by amino acid similarity scores. Outstanding reports of Capriotti et al. [19–21] describe predictors of the change of protein Gibbs free energy change ($\Delta\Delta G$) upon mutations by sequences and 3D structures from a dataset of more than 2000 mutants.

More recently, new predictors have been published using sequence and/or 3D structure information. iPTREE-STAB server [22] discriminates the stability of proteins and predicting their changes upon single amino acid substitutions from amino acid sequence. Similarly, Cheng et al. [23] developed sequence and 3D structure-based support vector machine (SVM) predictors. In addition, the prediction of protein mutant stability from distance and torsion potentials were also published by Parthiban [24].

In chemistry and related fields, chemometrics has been developed in the last 30 years, consisting of the use of mathematical, statistical and symbolic methods to improve the understanding of chemical information [1]. Chemometrics has been most successfully applied in four areas: multivariate calibration, quantitative structure–activity/property relationship (QSAR/QSPR) studies, pattern recognition, classification and discriminate analysis and multivariate modelling and monitoring process [1]. But recently the development of the bioinformatics has brought up chemometrics studies focused on proteins, known as proteometrics studies.

QSAR/QSPR studies of proteins have been developed by extending conventional graph-theoretical representation of chemical structures to sequences and 3D structures in combination with statistical methods for regression and/or classification analysis [25–27]. This approach has been applied to protein stability prediction. Similarly, the new proteochemometric approach includes the mapping of molecular recognition without needing knowledge of the 3D structure of biological macromolecules and it has successfully been applied to the study of protein–ligand interactions [1].

Nonlinear computing, artificial neural networks (ANNs) and SVMs have grown up rapidly in research fields as biochemistry, chemical engineering and pharmacy. In this regard, such frameworks have encountered successful applications in chemometric and bioinformatic studies, overcoming methods to linear regression models like Multilinear Regression Analysis (MRA) or partial least square [20,21,23,26–35]. Unlike these methods, ANNs and SVMs can be used to model complex nonlinear relationships. Since biological phenomena are complex by nature, this ability has encouraged the employment of nonlinear techniques in biological patron recognition problems.

In this work, protein conformational stability was successfully modelled from their amino acid sequences. We predicted the conformational stability by extending the concept of structural autocorrelation vectors [36–41] in molecules to protein primary structure. Protein sequence was encoded by means of amino acid sequence autocorrelation (*AASA*) vectors weighted by 48 physico-chemical, energetic and conformational amino acid/residue properties extracted from the AAindex amino acid database [42]. *AASA* vectors were previously reported by us for developing predictive models of the conformational stability of human lysozyme and gene V protein mutants [26,27]. Here in, conformational stability of a large dataset of more than 1500 mutants of 64 proteins, previously collected by Capriotti et al. [20], will be used for deriving general protein predictive models not biased to particular proteins. Genetic algorithm-optimised SVMs (GA-SVMs) trained with a reduced subset of *AASA* vectors yielded optimum nonlinear regression and classification models of $\Delta\Delta G$ upon single mutations. Temperature and pH of the $\Delta\Delta G$ experimental determinations were also conveniently added as extra SVM inputs, in order to improve predictors' performance.

2. Materials and methods

2.1 AASA vector approach

Conformational stability of a protein depends on a variety of intramolecular interactions such as hydrophobic, electrostatic, van der Waals and hydrogen bond that are ruled by the amino acid sequence. Therefore, in structure–property/activity studies the strategy for encoding structural information must, either explicitly or implicitly, account for these interactions. Furthermore, usually datasets include structures with different size and numbers of elements, so the structural encoding approaches must allow comparing such structures [36].

Autocorrelation vectors have several useful properties. Firstly, a substantial reduction in data can be achieved by limiting the topological distance, *l*. Secondly, the autocorrelation coefficients are independent of the original atom numberings, so they are canonical. And thirdly, the length of the correlation vector is independent of the size of the molecule [36].

For the autocorrelation vectors in molecules, H-depleted molecular structure is represented as a graph and physico-chemical properties of atoms as real values assigned to the graph vertices. These descriptors can be obtained by summing up the products of certain properties of two atoms, located at given topological distances or spatial lag in the graph. 2D spatial autocorrelations [37–41] has been successfully applied in the last decades for modelling biological activities [38,39] and pharmaceutical research [37,41]. In recent works, our group has obtained outstanding results when such chemical code was used in combination with ANN approach in biological QSAR studies [29,32]. Such results have inspired us to extend the application of the autocorrelation vector formalism to the study of other biological phenomena, particularly to encode protein structural information for conformational stability prediction. *AASA* vectors were previously reported by us for separately developing predictive models for the conformational stability of human lysozyme and gene V protein mutants [26,27].

Broto–Moreau's autocorrelation coefficient [39] is defined as:

$$A(p_k, l) = \sum_i \delta_{ij} p_{ki} p_{kj}, \tag{1}$$

where $A(p_k,l)$ is Broto–Moreau's autocorrelation coefficient at spatial lag l; p_{ki} and p_{kj} are the values of property k of atom i and j, respectively, and $\delta(l,d_{ij})$ is a Dirac-delta function:

$$\delta(l, d_{ij}) = \begin{Bmatrix} 1 & \text{if} & d_{ij} = l \\ 0 & \text{if} & d_{ij} \neq l \end{Bmatrix}, \tag{2}$$

where d_{ij} is the topological distance or spatial lag between atoms i and j.

The autocorrelation vector formalism can be easily extended to amino acid sequences considering protein primary structure as a linear graph with nodes formed by residues. Autocorrelation approach in protein stability mainly differs from the Gromiha et al. [15] method, when considering the whole sequence of the protein to calculate descriptors instead of local segments over the mutated point. In this way, the calculated autocorrelation vectors encode structural information of whole protein. Particularly, *AASA* vectors of lag l are calculated:

$$AASAlp_k = \frac{1}{L} \sum_i \delta_{ij} p_{ki} p_{kj}, \tag{3}$$

where $AASAlp_k$ is the *AASA* at spatial lag l weigthed by the p_i property; L is the number of terms in the sum; p_{ki} and p_{kj} are the values of property; k of amino acids; i and j in the sequence, respectively and $\delta(l, d_{ij})$ is a Dirac-delta function.

For example, if we consider the decapeptide ASTCGFHCSD, *AASA* vectors at spatial lag 1 and 5 are calculated as follows:

$$\begin{aligned} AASA1p_k = \frac{1}{9}(&p_{kA} \cdot p_{kS} + p_{kS} \cdot p_{kT} + p_{kT} \cdot p_{kC} + p_{kC} \cdot p_{kG} \\ &+ p_{kG} \cdot p_{kF} + p_{kF} \cdot p_{kH} + p_{kH} \cdot p_{kC} \\ &+ p_{kC} \cdot p_{kS} + p_{kS} \cdot p_{kD}), \end{aligned} \tag{4}$$

$$\begin{aligned} AASA5p_k = \frac{1}{5}(&p_{kA} \cdot p_{kF} + p_{kS} \cdot p_{kH} + p_{kT} \cdot p_{kC} + p_{kC} \cdot p_{kS} \\ &+ p_{kG} \cdot p_{kD}). \end{aligned} \tag{5}$$

Autocorrelation measures the level of interdependence between properties, and the nature and strength of that interdependence. It may be classified either as positive or negative. In a positive case all similar values appear together, while a negative spatial autocorrelation has dissimilar values appearing in close association [37,41]. In a protein, autocorrelation analysis tests whether the value of a property at one residue is independent of the values of the property at neighbouring residues. If dependence exists, the property is said to exhibit spatial autocorrelation. *AASA* vectors represent the degree of similarity between amino acid sequences.

As weights for sequence residues we used 48 physicochemical, energetic, and conformational amino acid/residue properties (Table S1 in the Supplementary Material, available online). These were collected by Gromiha et al. [13] from the AAindex database [42] in a previous study regarding relationships between amino acid/residue properties and protein stability for a large set of proteins. In our work, spatial lag l, was ranging from 1 to 15 with the aim of accessing long range interactions in the sequence, due to tertiary structure arrangements. Computational code for *AASA* vector calculation was written in Matlab environment (MATLAB 7.0 Program, available from The Mathworks Inc., Natick, MA, USA. http://www.mathworks.com).

A data matrix of 720 *AASA* vectors, 48 properties × 15 different lags, was generated with the sequence autocorrelation vectors for each mutant. Descriptors, which remained constant or almost constant, were eliminated. Pairs of variables with square correlation coefficients greater than 0.8 were classified as intercorrelated and only one of these was included for building the model. Finally, 302 descriptors were obtained. Afterwards, optimum classification and regression models were built with reduced subsets of normalised variables by means of a hybrid approach combining GA optimisation and SVM training.

2.2 SVM

SVM is a new machine learning method, which has been used for many kinds of pattern recognition problems. Since there are excellent introductions to SVMs [44], only the main idea of SVMs applied to pattern classification problems is stated here. Firstly, the input vectors are mapped into one feature space (possible with a higher dimension). Secondly, a hyperplane, which can separate two classes, is constructed within this feature space. Only relatively low-dimensional vectors in the input space and dot products in the feature space will involve by a mapping function. SVM was designed to minimise structural risk, whereas previous techniques were usually based on minimisation of empirical risk. SVM is less vulnerable to the overfitting problem, so it can deal with a large number of features.

The mapping into the feature space is performed by a kernel function. There are several parameters in the SVM, including the kernel function and regularisation parameter. The kernel function and its specific parameters, together with regularisation parameter, cannot be set from the optimisation problem but have to be tuned by the user. These can be set by the use of Vapnik–Chervonenkis bounds, crossvalidation, an independent optimisation set or Bayesian learning. In this paper, the radial basic function (RBF) was used as kernel function. GA-based SVM (GA-SVM) algorithm was implemented for choosing the optimum subset of input training vectors and setting the two SVM parameters, regularisation parameter and width of the RBF kernel. The optimisation inside the GA framework was driven by crossvalidation. The toolbox used to implement the SVM with RBF kernel (RBF–SVM) was LIBSVM for Matlab by Chang and Lin [45] that can be downloaded from: http://www.csie.ntu.edu.tw/cjlin/libsvm/.

2.3 GA-based feature selection and hyperparameter optimisation

The applications of SVMs for solving classification and function mapping problems in biological QSAR studies have vastly developed in the last years [46,47]. However, it is difficult to choose the adequate descriptors for predictor training, due to lack of absolute rules that govern this choice. Evolutionary algorithms and specifically GA have been used for variable selection problems [26,27,30–33,47]. Since 302 *AASA* vectors were available for modelling and only a subset of them is statistically significant in terms of correlation with the mutants stability, it was necessary to implement an optimal model by variable selection.

GA was applied at the same time for selection of the optimum subset of variables and also to the optimisation of regularisation parameter and width of an RBF kernel, according to Fröhlich et al. [48]. We can simply concatenate a representation of the parameter to a chromosome representing subset of variables used for SVM training. That means, we are trying to select an optimal feature subset and an regularisation parameter at the same time. This is reasonable because the choice of the parameter is influenced by the feature subset taken into account and vice versa. Usually, it is not necessary to consider any arbitrary value except certain discrete values with the form: $n \times 10^k$, where $n = 1$–9 and $k = -4, \ldots, 4$. So, these values can be calculated randomly, generating n and k values as integers between 1–9 and $-4, \ldots, 4$, respectively. In a similar way, we used GA to optimise the width of an RBF kernel. Then, our chromosome was concatenated with another gene with discrete values in the interval 0.001–90,000 for encoding the regularisation parameter and the width of the RBF kernel.

A five-fold-out (FFO) crossvalidation assessed model quality throughout the GA search. Five data subsets were created, four subsets are generated in the crossvalidation process for training the SVM and another subset is then predicted. This process is repeated until all subsets have been predicted. A ‘venetian-blind’ method was used for creating the data subsets. In the first place, dataset is ordered according to the dependent variable and in the second step, the cases are added consecutively to each subset, in such a way that they become representative samples of the whole dataset. In order to avoid overestimation of the model’s predictive power, similar mutants were kept in the same set during crossvalidation even when they reported under different experimental temperature and pH values. The GA routine minimised the regression mean squared error ($\mathrm{MSE}_{\mathrm{FFO}}$) of FFO crossvalidation experiment.

Afterwards, the same subset of optimum variables selected by the regression GA-SVM was used for training a SVM classifier. Nevertheless, regularisation parameter and width of RBF kernel for the SVM binary classifier were set by a bidimensional grid search around optimum GA-selected parameters, which minimised the percent of misclassifications ($\mathrm{MC}_{\mathrm{FFO}}$) of FFO crossvalidation.

A version of the GA implemented in this paper was recently reported by our group [33] and applied to SVM hyperparameters optimisation. GlibSVM [49] toolbox for Matlab was programmed within Matlab environment (MATLAB 7.0 Program) using GA [50] and libSVM Toolboxes [45].

2.4 Model’s validation

The quality of the regression SVM models was evaluated by the squared correlation coefficient of FFO crossvalidation (R^2_{FFO}) and the root $\mathrm{MSE}_{\mathrm{FFO}}$ ($\mathrm{RMSE}_{\mathrm{FFO}}$) and also calculated classification statistics of test set.

The efficiency of the SVM predictor for the classification problem was accomplished using the set of statistics listed below.

The overall accuracy is

$$Q^2 = \frac{p}{N}, \tag{6}$$

where p is the total number of correct predicted mutations and N is the total number of mutations.

The correlation coefficient C is defined as follows:

$$C(s) = \frac{[p(s)n(s) - u(s)o(s)]}{D}, \tag{7}$$

where D is the normalisation factor,

$$D = [(p(s) + u(s))(p(s) + o(s))(n(s) + u(s)) \times (n(s) + o(s))]^{1/2}, \tag{8}$$

for each class s (+ and −, for stable and unstable mutant); $p(s)$ and $n(s)$ are the number of correct predictions and correctly rejected assignments, respectively, and $u(s)$ and $o(s)$ are the number of under- and over-predictions.

The coverage for each discriminant structure s is evaluated as

$$Q_S = \frac{p(s)}{p(s) + u(s)}, \tag{9}$$

where $p(s)$ and $u(s)$ are the same as in Equation (8)

The accuracy for s is computed as

$$P_S = \frac{p(s)}{p(s) + o(s)}, \tag{10}$$

where $p(s)$ and $u(s)$ are the same as in Equation (8).

2.5 Mutant training and test datasets

Our training dataset was a non-redundant version of the mutant dataset previously collected for Capriotti et al. [20] for deriving a predictive model for the signs and the actual values of $\Delta\Delta G$ using ANNs and SVMs predictors. They collected the data from the Protherm database [43] according to the following constrains:

(1) $\Delta\Delta G$ values have been experimentally determined and reported in the database.
(2) The data is related to single point mutations (non-multiple mutations were taken into account).

After the filtering, they gathered 2048 single point mutants, obtained from 64 proteins. But, we removed the redundant mutants from Capriotti et al.'s dataset and a total of 1383 non-redundant single point mutants was used as training set.

We collected a test set including non-redundant mutations, with single, double and multiple characteristics, according to Protherm's database and Cappriotti et al.'s selection until September, 2007. The test dataset was collected according to the following constraint:

(1) $\Delta\Delta G$ values have been experimentally determined and reported in the database.
(2) Included single mutants in Capriotti's dataset were not considered in the test set.

After filtering, we gathered a test set including non-redundant 222 single, 277 double and 144 multiple point mutations corresponding to 22, 43 and 18 proteins, respectively.

Table S2 in the Supplementary Material reveals the values of calculated and experimental $\Delta\Delta G$ for training's and tested dataset, the secondary structure, solvent accessible surface details of the mutated sites, as well as pH and temperature values of the experimental determinations.

3. Results and discussion

We trained new robust classification and regression predictors of the conformational stability of protein mutants, using just sequence information from a dataset with 1383 non-redundant from the 2048 mutations collected by Capriotti et al. [20]. In addition to internal crossvalidation, the stability prediction of a test set with non-redundant single, double and multiple mutations were also predicted. We computed a large set of 720 *AASA* vectors (see Section 2.1, *AASA* vector approach) calculated on the mutant sequences. Since autocorrelation vectors have some intercorrelations [26,27,47], redundancy on the *AASA* data matrix was reduced by eliminating intercorrelated variables. Finally, a total of 302 autocorrelation vectors were available for building the models. Then, function mapping of conformational stability was accomplished by training regression GA-SVMs. In the GA-SVM framework, optimum subset of training *AASA* inputs and SVM parameter values were set using GA rules. The optimum SVM was also trained for solving the classification problem regarding the recognition of stable and unstable mutants, but the regularisation parameter and width of the RBF kernel were set by grid search. Afterwards, normalised Temperature and pH values of the $\Delta\Delta G$ experimental determinations, were also passed to the SVM in order to improve predictor performance.

3.1 Function mapping of mutants conformational stability

Firstly, GA-SVM approach was applied for yielding optimum nonlinear regression models of protein conformational stability using RBF kernel inside the SVM framework. Nonlinearity was obtained by using a RBF kernel inside the SVM framework. Nonlinear subspace in the dataset was searched varying problem dimension from 5 to 30. From one generation to another GA minimised $\mathrm{MSE_{FFO}}$. FFO crossvalidation subsets were selected according to 'venetian blind' method and the $\mathrm{MSE_{FFO}}$ was minimised throughout the GA search. As it was pointed out, in addition to selecting the optimum input subset, GA also optimised the kernel regularisation parameter and the width of an RBF kernel (see Section 2.3, GA-based feature selection and hyperparameter optimisation).

Table 1 shows inputs, parameters and statistical quantities for data fitting and crossvalidation experiment

Table 1. Crossvalidation statistics of the RBF-SVM model for prediction of protein mutant $\Delta\Delta G$ actual values.

Regression RBF−SVM inputs	R^2_{FFO}	$\mathrm{RMSE}_{\mathrm{FFO}}$
$AASA11N_{\mathrm{m}}$, $AASA8P$, $AASA7P_{\mathrm{B}}$, $AASA7G_{\mathrm{hN}}$, $AASA10Ht$, $AASA14f$,$AASA12\Delta G_{\mathrm{C}}$, $AASA15\Delta ASA$, $AASA15\Delta Cp_{\mathrm{h}}$, $AASA14\Delta Cp_{\mathrm{h}}$	0.42	0.139
$AASA11N_{\mathrm{m}}$, $AASA8P$, $AASA7P_{\mathrm{B}}$, $AASA7G_{\mathrm{hN}}$, $AASA10Ht$, $AASA14f$,$AASA12\Delta G_{\mathrm{C}}$, $AASA15\Delta ASA$, $AASA15\Delta Cp_{\mathrm{h}}$, $AASA14\Delta Cp_{\mathrm{h}}$, temperature, pH	0.45	0.136

Note: SVM regularisation parameter was 1 and width of the RBF kernel was 0.071. R^2_{FFO} and $\mathrm{RMSE}_{\mathrm{FFO}}$ are the square correlation coefficient and the root mean square error of FFO crossvalidation.

of the optimum RBF-SVM predictor. Optimum regularisation parameter and width of the RBF kernel were 1 and 0.071, respectively. Input *AASA* vectors in Table 1 mean: $AASA11N_{\mathrm{m}}$ is the amino acids sequence autocorrelation vector at lag 11 weighted by average medium-range contacts; *AASA8P* is the amino acids sequence autocorrelation vector at lag 8 weighted by polarity; $AASA7P_{\mathrm{B}}$ is the amino acids sequence autocorrelation vector at lag 7 weighted by β-structure tendency; $AASA7G_{\mathrm{hN}}$ is the amino acids sequence autocorrelation vector at lag 7 weighted by Gibbs free energy change of hydration for native protein; *AASA10Ht* is the amino acids sequence autocorrelation vector at lag 10 weighted by thermodynamic transfer hydrophobicity; *AASA14f* is the amino acids sequence autocorrelation vector at lag 14 weighted by flexibility; $AASA12\Delta G_{\mathrm{C}}$ is the amino acids sequence autocorrelation vector at lag 12 weighted by unfolding Gibbs free energy change of side-chain; $AASA15ASA_{\mathrm{N}}$ is the amino acids sequence autocorrelation vector at lag 15 weighted by solvent-ASA for native protein; $AASA15\Delta Cp_{\mathrm{h}}$ and $AASA14\Delta Cp_{\mathrm{h}}$ are the amino acids sequence autocorrelation vectors at lag 15 weighted by hydration heat capacity change. The optimum autocorrelation vector subset contains only two significant pair correlations ($R^2 > 0.7$): *AASA14f* vs. $AASA15ASA_{\mathrm{N}}$ and $AASA15\Delta Cp_{\mathrm{h}}$ vs. $AASA14\Delta Cp_{\mathrm{h}}$ (Table S3 in the Supplementary Material). Despite this little intercorrelation, the adequate fitting of the dataset and the FFO crossvalidation obtained by such descriptor subset, reflect that relevant structural information is brought into the model by each *AASA* descriptor.

Figure 1(a) depicts plots of calculated vs. experimental $\Delta\Delta G$ values in crossvalidation experiment according to the optimum RBF-SVM with 10 *AASA* vectors. The maximum correlation coefficient of 0.65 in Figure 1(a) is higher than the value of 0.62 previously reported by Capriotti et al. [20], despite they trained a RBF−SVM regressor with temperature and pH values, in addition to sequence-derived variables. In order to increase the predictive accuracy of the model, we then passed temperature and pH values of the experimental determinations as extra inputs to the regression SVM. Figure 1(b) depicts calculated FFO crossvalidation vs. experimental plot of the regression SVM using 10 optimum *AASA* vectors plus temperature and pH values as extra inputs. The correlation coefficient

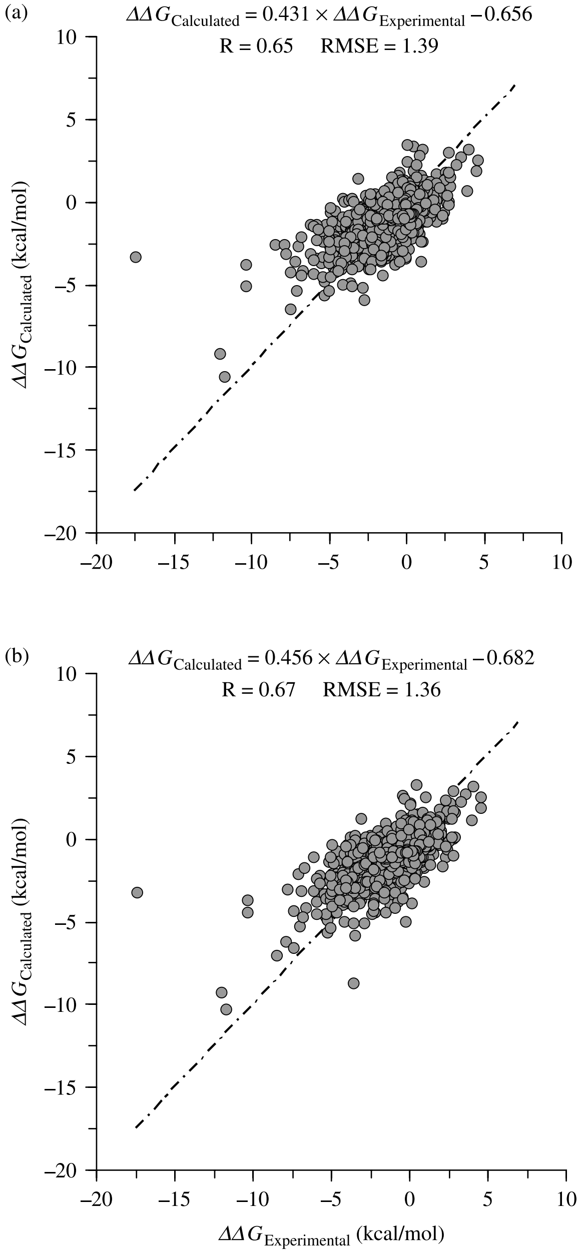

Figure 1. Plots of crossvalidation calculated vs. experimental change of unfolding Gibbs free energy change ($\Delta\Delta G$) of protein mutants according to regression SVM models without including experimental condition data (a) and including experimental condition data (b) as SVM inputs. Dotted lines are an ideal fit with the respective intercept and slope equal to zero and one.

was increased up to 0.67 representing nearly a 50% of explained crossvalidation data variance. This value is remarkably higher than 38% of the Capriotti et al. [20] regression model. Meanwhile, *AASA* vector approach encodes sequence information considering quasi-sequence order effect by calculating autocorrelation vectors at different lags all over the protein sequence and also by using as weighting values a set of 48 amino acid/residue properties [13].

Since it has been reported that the effects caused for a specific mutation depend on the type of substituted and new added residues [20,51], it is interesting to analyse the performance of the predictor regarding the nature of the mutations. Mutations were classified according to the physico-chemical properties of the substituted and new residues. Prediction accuracies of the optimum regression predictor depending on the mutation type appear in Figure 2 that depicts plots of calculated FFO vs. experimental $\Delta\Delta G$ values for mutants according to mutation types. The worst predictions were achieved for charged/charged, polar/charged and apolar/charged. Otherwise, the specific effects of residue substitutions on the actual $\Delta\Delta G$ values are better predicted for polar/polar and polar/apolar mutations with crossvalidation accuracy over 50%. The prediction accuracies of the $\Delta\Delta G$ actual values varied for each mutation type. This fact suggests that the regression SVM model can better learn the effects produced by some mutation types, in comparison to others. The effects caused

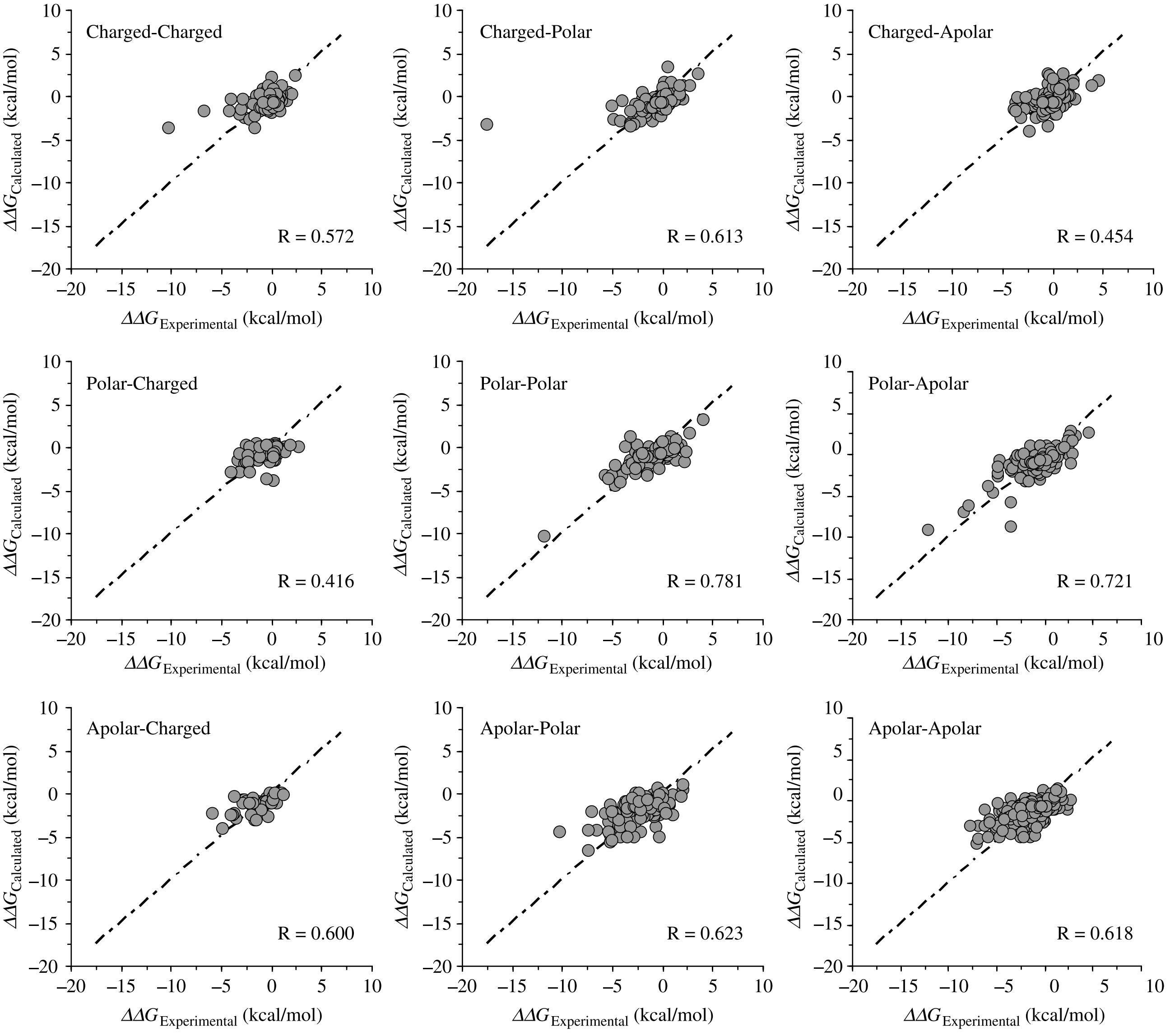

Figure 2. Plots of crossvalidation calculated vs. experimental change of unfolding Gibbs free energy change ($\Delta\Delta G$) of protein mutants for each mutation (physico-chemical properties of the substituted and new residues) type according to regression SVM models including experimental condition data as SVM inputs. Dotted lines are an ideal fit with the respective intercept and slope equal to zero and one.

for charged/charged, charged/apolar and polar/charged mutations seem more complex. Therefore, such interactions are not sufficiently contained in the mutant dataset or they can not be successfully encompassed by the sequence descriptors used.

Salt-bridge and hydrogen-bridge interactions at protein surface of charged and polar residues usually appear at long-ranges. Despite being separated by long stretches of polypeptide in the primary sequence, surface groups lie next to each other in space. Consequently, these interactions are very difficult to model from a sequence framework. On the contrary, hydrophobic interactions at protein core mainly appear at short-range in the sequence. But mutations in the protein core, even residue size rather than polarity (apolar/apolar mutations), may cause an unfavourable packing energy due to the rigidity of surrounding residues or, alternatively, the substituting residues themselves may be forced into unfavourable rotational isomers. Similarly, some surroundings of mutation positions may be readily deformable or compensate effects, if no net packing energy change occurs [51]. In the light of this fact, complex and 3D environment-dependent interactions take place in the protein core that can be also somehow inaccessible from a primary structure approximation.

In addition, we compared the accuracy of the prediction according to the type of secondary structure found in the mutation site. The type of secondary structure was assigned to each mutation from the database Protherm [43], in which residues are classified in four different secondary structures: helix, sheet, turn and coil. Figure 3 depicts calculated vs. experimental $\Delta\Delta G$ values for each secondary structure type. The lower correlations were found for mutations allocated at helix and coil structures. In the case of helix mutations, the deficient predictor performance might be related to the fact that destabilisation of helix structures was caused by variations of very

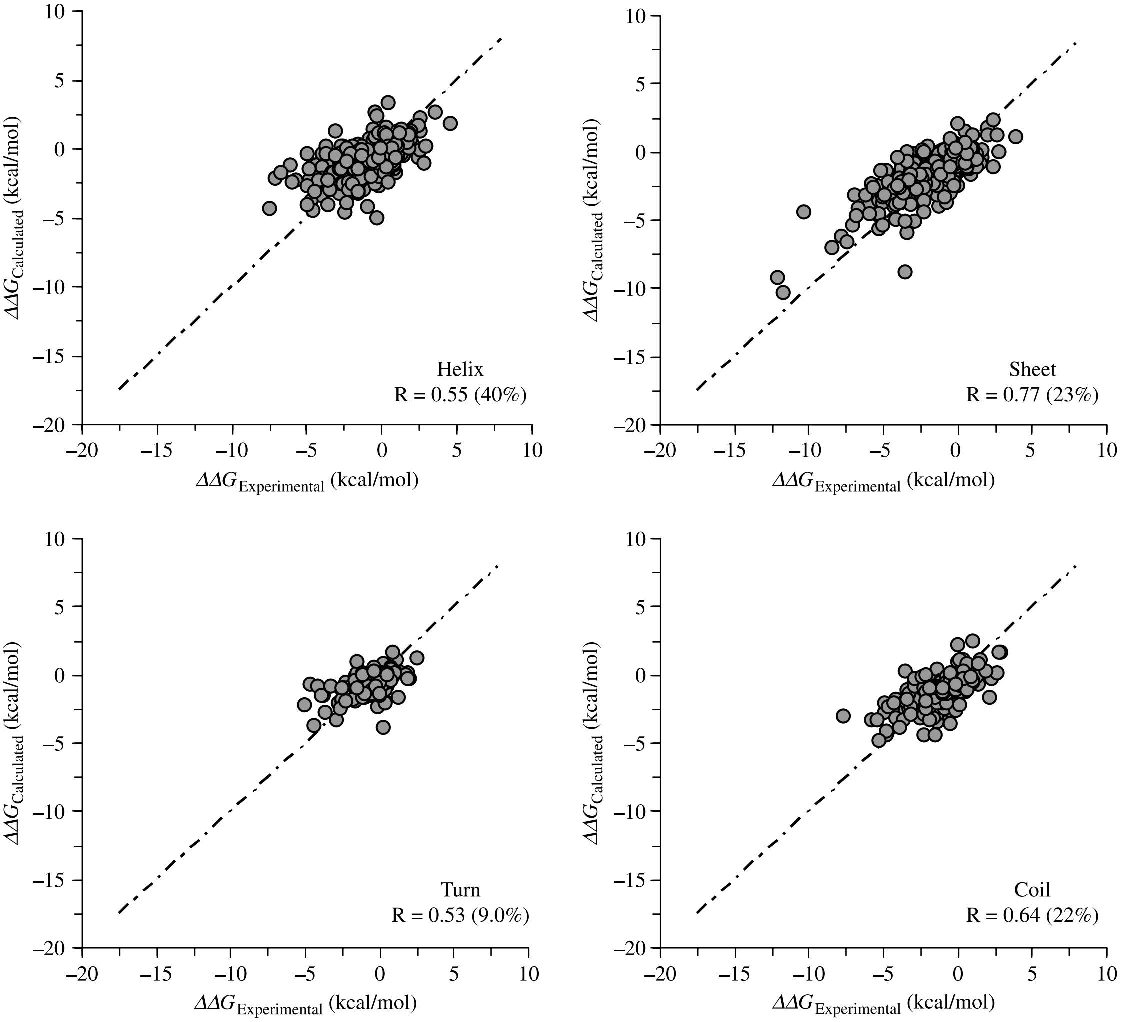

Figure 3. Plots of crossvalidation calculated vs. experimental change of unfolding Gibbs free energy change ($\Delta\Delta G$) of protein mutants for each mutation type (secondary structure found in the mutation site) according to regression SVM models including experimental condition data as SVM inputs. Dotted lines are an ideal fit with the respective intercept and slope equal to zero and one.

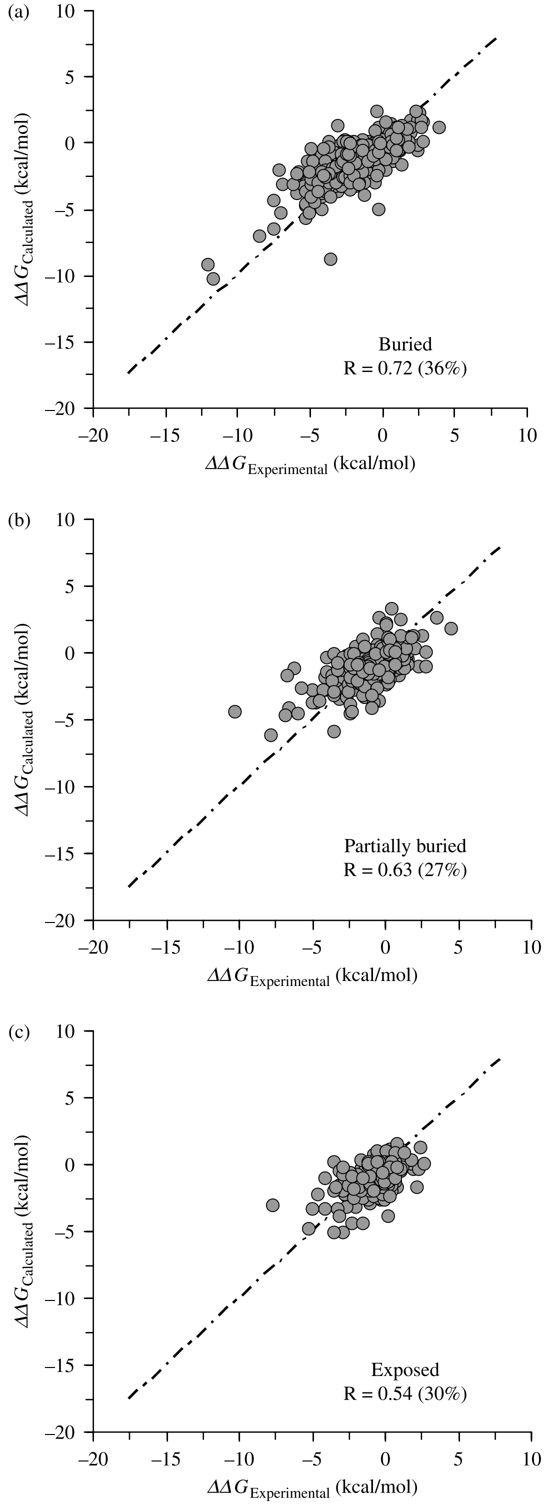

Figure 4. Plots of crossvalidation calculated vs. experimental change of unfolding Gibbs free energy change ($\Delta\Delta G$) of protein mutants for each mutation type (accessible surface area (ASA) values of the mutation sites) according to regression SVM models including experimental condition data as SVM inputs. Dotted lines are an ideal fit with the respective intercept and slope equal to zero and one.

complex residue–residue and residue–solvent interaction patterns at the protein surface. It is noteworthy that mutations of charged residues, mainly allocated at protein surface, were also the worst predicted when analysing the prediction results according to the physic-chemical nature of the mutation site (Figure 2). In the case of mutation allocated at turn structures, the lower correlation could be related, in addition to the variability and complex nature of turn regions in proteins, to the low statistical significance of this group of mutations, taking into account that it represents only 9% of the training dataset.

Similarly, an analysis of the regression accuracy was carried out taking into account the ASA values of the mutation sites. According to Protherm [43] database, mutations were grouped as buried (ASA $<$ 20%), partially buried (50% $<$ ASA $<$ 20%) and exposed (ASA $>$ 50%). Results in Figure 4 support the previous findings in Figures 2 and 3. The correlation coefficients decrease with the decrements of ASA values of the mutated site. The lower regression accuracy was found for the exposed mutations allocated at the protein's surface and the highest correlation value was reported for mutations in buried sites at the protein's core.

Concerning the prediction of actual values of Gibbs free energy change of proteins, in previous reports with large dataset ($>$1000) no more than 60% of validation data variances were described; although, such models used protein 3D structure information [12,16,49,50]. AGADIR [11] or FOLDEF was reported by Guerois et al. [12] for predicting conformational stability of more than 1000 mutants with a crossvalidation accuracy about 60%. Zhou and Zhou's method D-FIRE [52] is based on distance-scaled, finite ideal-gas reference state that improved structure-derived potentials of mean force for structure selection and stability prediction. Their model, with 3D protein structures from a database of 895 large-to-small mutations, described 0.45% of crossvalidation data variance. In addition, Borner and Abagyan [53] developed a model to predict both geometry and relative stability of point mutants and it may be used for arbitrary mutations. An empirical energy function, which includes energy contributions of the folded and denatured proteins, and the prediction of a side chain mutant, was fitted to a training set consisting in a half of a diverse set of nearly 2000 experimental stability values for single point mutations. The prediction method was then tested on the remaining half of the experimental data, giving a covariance of 0.66 for 97% of the test set.

On the other hand, machine learning algorithms in combination with sequence and 3D information were also applied to solve the protein conformational stability problem [18–23,26,27]. Capriotti et al. [19–21] described the implementation of ANNs and SVMs predictors of $\Delta\Delta G$ upon mutations using sequences and 3D structures of more than 1000 mutants. As predictor inputs they used

Table 2. Crossvalidation statistics of the RBF–SVM model for the classification of protein mutant $\Delta\Delta G$ signs.

RBF–SVM inputs	Q^2	$P(+)$	$P(-)$	$Q(+)$	$Q(-)$	C
$AASA11N_m$, $AASA8P$, $AASA7P_B$, $AASA7G_{hN}$, $AASA10Ht$, $AASA14f$, $AASA12\Delta G_C$, $AASA15ASA_N$, $AASA15\Delta Cp_h$, $AASA14\Delta Cp_h$	0.78	0.59	0.87	0.68	0.82	0.48
$AASA11N_m$, $AASA8P$, $AASA7P_B$, $AASA7G_{hN}$, $AASA10Ht$, $AASA14f$, $AASA12\Delta G_C$, $AASA15ASA_N$, $AASA15\Delta Cp_h$, $AASA14\Delta Cp_h$, temperature, pH	0.77	0.57	0.88	0.71	0.80	0.48

Note: SVM regularisation parameter was 7 and width of the RBF kernel was 0.167. + and −, the indexes were evaluated for positive and negative $\Delta\Delta G$ signs; Q^2 is the number of correct predictions/number of examples; $P(s)$ is the number of correct prediction for class s/all prediction made for s; $Q(s)$ is the number of correct prediction for class s/observed in class s; C is Matthews's correlation coefficient.

a combination of experimental condition data (pH and temperature), specific mutated residue and sequence environment information. Their 'best' sequence-based model explained a discrete 0.38% value of crossvalidation data variance, whilst the 3D structure-dependent predictor exhibited an acceptable value of 50%. It is noteworthy that, despite its sequence nature, our optimum *AASA*-SVM overcomes the previous sequence-based models of Capriotti et al. in [20] and nearly reaches the result obtained by their 3D structure-based predictor in [21] by yielding a 0.45% of crossvalidation accuracy. However, it should also be taken into account that the dataset they used contained some redundant mutations that contribute to over-estimate the predictor performance.

Recently, Huang et al. [22] reported the iPTREE-STAB server to discriminating the stability of proteins (stabilising or destabilising) and predicting their stability changes upon single amino acid substitutions from amino acid sequence. The predictor was trained with a dataset of 1859 non-redundant single point mutations of 64 proteins. The prediction of actual $\Delta\Delta G$ values is mainly based on regression tree using three neighbouring residues of the mutant site along N- and C-terminals. Their method showed a crossvalidation correlation of 0.70 for predicting protein stability changes upon mutations, which is similar to our results. Other recent report by Cheng et al. [23] referred to the prediction of single mutant actual $\Delta\Delta G$ values and signs by SVM predictors trained with information from three different encoding schemes: sequence, structure and combined sequence and structure. In this case, the prediction of values $\Delta\Delta G$ actual was higher than our results, having crossvalidation correlation coefficients of 0.75 and 0.76 for the sequence- and structure-based predictors, respectively.

Table 3. Percent of crossvalidation correct classifications of the optimum RBF–SVM model for the $\Delta\Delta G$ signs upon mutations according to the mutations type.

	New		
Native	Charged (%)	Polar (%)	Apolar (%)
Charged	66 (6)	79 (9)	72 (10)
Polar	70 (5)	84 (9)	72 (16)
Apolar	71 (4)	88 (13)	80 (28)

Note: In brackets the relative fraction of each mutation type in the training dataset of 1383 single point mutants.

3.2 *Classification of mutants conformational stability*

In addition to the regression SVM model, we also built a classifier for the recognition of stable and unstable mutants. The optimum *AASA* vector subset was used for training a binary SVM classifier. Parameters of the binary SVM were set by minimising $\mathrm{MC_{FFO}}$ in a grid search. Since the mutant dataset was three fold unbalanced towards unstable mutants, we used a three-fold higher penalty for stable mutant misclassification inside the SVM framework. This allows obtaining a classifier not biased to one class. Optimum values of regularisation parameter and RBF kernel width were 7 and 0.167, respectively, which yield training and crossvalidation results in Table 2. As can be observed, when SVM was only trained with optimum *AASA* subset overall FFO the crossvalidation accuracy was 78% for mutant's $\Delta\Delta G$ signs and the correlation coefficient $C = 0.48$. It is noteworthy that the crossvalidation statistics for recognising stable mutants $Q(+) = 0.68$ and unstable mutants $Q(-) = 0.77$ are in the range of the overall accuracy achieved. This result is quite interesting since the predictor just used sequence

Table 4. Crossvalidation classification accuracies of the optimum RBF–SVM model for the $\Delta\Delta G$ signs upon mutations according to secondary structure allocation and ASA of the mutated residue.

Mutation type (%)	Q^2	$Q(+)$	$Q(-)$
Helix (40)	0.76	0.62	0.82
Sheet (23)	0.82	0.79	0.82
Turn (9)	0.68	0.73	0.65
Coil (22)	0.83	0.69	0.86
Buried (36)	0.80	0.71	0.81
Partially buried (27)	0.81	0.75	0.84
Exposed (30)	0.73	0.60	0.80

Note: In brackets the relative fraction of each mutation type in the training dataset of 1383 single point mutants. + and −, the indexes were evaluated for positive and negative $\Delta\Delta G$ signs; Q^2 is the number of correct predictions/number of examples; $Q(s)$ is the number of correct prediction for class s/observed in class s.

information encoded in 10 *AASA* vectors. Afterwards, temperature and pH normalised values were passed to the classification SVM as extra inputs and Q^2, $Q(+)$ and $Q(-)$ values increased up to 0.77, 0.71 and 0.80, respectively, with a correlation coefficient $C = 0.48$. Similar to the regression model, the classification SVM tuned its prediction accuracies, while it was trained combining experimental conditions of the $\Delta\Delta G$ measurements and the quasi-sequence order information, gathered by the autocorrelation vector and weighted by amino acid/residue properties.

In Table 3, the classification results according to the physico-chemical properties of the mutations are shown. Analysing the SVM predictor accuracy as a function of the mutations classification, we found that mutations of charged residues, mainly placed at the protein surface by other charged residues, exhibit the lower classification accuracy. This fact suggests that sequence information incorrectly described the effect of Salt-bridge and polar–polar interactions at the protein surface that should be better using 3D structure details. As we previously stated, in the protein core interactions often occur among residues at shorter range in comparison to the protein surface where interactions can be at larger ranges.

We also compared the accuracy of the classifier according to the type of secondary structure found in the mutation site. In Table 4, it can be observed that, similar to the regression model, the lower accuracies were for mutations allocated at helix and coil structures. However, overall classification accuracies, about or higher than 70%, were observed for all types of mutations. On the other hand, the results of the classification accuracy, taking into account the ASA values of the mutation site, appear also in Table 4. The overall classification accuracies for the buried and partially buried mutations were about 80%, but again the classification accuracy for mutations in the protein surface was adequately lower, about 70%. These results support the fact that protein properties which depend on interactions at the protein surface are more difficult to predict.

Some other classification models for protein mutant's stability have been reported. Two classifiers used sequence information for Linear Discriminant Analysis (LDA) classification of Arc repressor mutants, but according to its melting point value. Both reports explained more than 80% of validation data variance using the 'macromolecular pseudograph Cα-atom adjacency matrix', a sequence approach [25]. However, such models lack utility because they are protein-specific and use a thermodynamic parameter (melting point) not directly related to the protein conformational stability. In addition, other single-protein models, previously developed by us which were built with *AASA* vectors and Bayesian regularised neural networks, successfully mapped lysozymes [26] and gene V protein [27] mutants in self-organising maps according to mutant's $\Delta\Delta G$ levels.

In turn, taking into account classification models of protein stability change upon mutations, using large and diverse mutant data, our classification model overcomes the optimum reported by Capriotti et al. [20] using only sequence information. Despite the fact that they reported an overall accuracy about 77%, the correct predictions were drastically shifted towards unstable mutants with a value about 91%, whilst for stable mutants, a very low value about 46% was reported. Such statistics reflect that their model nearly recognised all mutants as unstable, yielding an overall adequate accuracy, but an inefficient discriminating ability. However, our classification model overcomes Capriotti's classifier by predicting unstable and stable mutants with accuracies over 70%. In this connection, we recently reported two SVM predictors based on 2D and 3D graph representations of protein sequences [34,35], which identified both stable and unstable mutants with identical good accuracy over 70%. In comparison with these own reports, the presented *AASA*-SVM revealed a similar predictive power, but the correlation coefficient to classification (about 0.48) was higher the reported values 0.41 and 0.39, through our 2D and 3D graphs-based models. Interestingly, when Capriotti et al. [21] used 3D structure information, the highest overall classification accuracy achieved was about 80%, but stable mutants were poorly recognised with an accuracy of 56%. That is the fact that evaluates our *AASA*–SVM predictor, despite its primary sequence nature, as more adequate for the mutant stability recognition task.

Huang et al. [22] recently reported that iPTREE-STAB server was able to recognise unstable and stable mutants with overall crossvalidation accuracy about 82%, and sensitivity and specificity were about 75.3 and 84.5%, respectively. The best sequence-based SVM classifier reported by Cheng et al. in [23] had an overall accuracy about 84%, but it discriminated between unstable and stable mutants with accuracies about 90 and 71%, respectively, in crossvalidation experiment. This result shows that the predictor is somewhat unbalanced and tends

Table 5. Test set classification accuracies according to RBF-SVM model for the classification of protein mutant $\Delta\Delta G$ signs. SVM regularisation parameter was 7 and width of the RBF kernel was 0.167.

Test set	Q^2	$Q(+)$	$Q(-)$
Single point mutations	0.51	0.66	0.45
Double point mutations	0.50	0.70	0.34
Multiple point mutations	0.31	0.31	0.31
Single and double point mutations	0.50	0.68	0.40
Single, double and multiple point mutations	0.46	0.57	0.38

Note: + and −, the indexes were evaluated for positive and negative $\Delta\Delta G$ signs; Q^2 is the number of correct predictions/number of examples; $Q(s)$ is the number of correct prediction for class *s*/observed in class *s*.

to recognise stable mutants with lower accuracy ($Q(+) = 0.71$), which is equal to the reported value by our SVM classifier. In addition, when Cheng et al. used structural information for training the predictors, the results were very similar.

Other approaches used 3D structural information for predicting stability change upon single point mutations. Parthiban et al. [24] implemented a distance-dependant pair potential through-space interactions and torsion angle potential neighbouring effects a basic statistical mechanical set-up to compare theoretically the predicted stabilising energy values with experimental values from thermal and chemical denaturation experiments. The derived force fields yielded a correlation of 0.77 and more than 80% classification accuracy in crossvalidation for chemical denaturation. For thermal denaturation the force field yielded a correlation of 0.78 with a prediction efficiency of 84.65%. In addition, Gonzalez-Diaz et al. [54] reported a LDA model for stable mutants using 3D stochastic average electrostatic potentials from protein 3D structure with validation accuracy nearly 90% for all the dataset and for each class separately. However, instead of discriminating between stable or unstable mutants according to wild-type protein, they classified the mutants in higher stable and near-wild-type stable.

3.3 New mutants classification

In addition to the crossvalidation experiment, we predicted the sign of $\Delta\Delta G$ values for a test set with new single point mutants in Protherm database [43], all double and multiple mutations on this database (see Section 2.5, Mutant training and test datasets). Prediction of actual $\Delta\Delta G$ values was very inaccurate. Results of stability classification of mutations in the test set appear in Table 5. As can be observed, the performance of the predictor on the single point mutation test set was poor. A lower overall accuracy about 51% was achieved with an adequate recognition of about a 66% of the stable mutations and low 45% of the unstable ones. Similarly, the double point mutation test set showed $Q^2 = 0.50$ with $Q(+) = 0.70$ and $Q(-) = 0.34$. Despite the discrete results, the predictor can account for single point mutation effects and can generalise them to some extent to double point mutations. It should be pointed out that unbalanced classification results had been also achieved by Cappriotti et al. [20,21] in crossvalidation experiments. Multiple mutations exhibited the worst prediction results with very low accuracies around 30%. On the contrary, the accuracies for single and double mutant test set were about 50, 68 and 40% for the recognition of all mutants, stable and unstable mutations, respectively. Meanwhile, these statistical quantities for the whole test set (single, double and multiple mutants) were lower with values of 46, 57 and 38%, respectively. When considering only single and double point mutant test set the overall accuracies were about 50% and for recognition of the stable single and double point mutants, the classifier exhibited a higher accuracy, near 70%. Finally, Table 6 shows classification accuracies for the single point mutants in the test set according to the secondary structure allocation and the ASA of the mutation site. The low prediction result for the test set suggests that our sequence-based approach could be somehow limited to achieve better results due to the complexity of the stabilising–destabilising interactions in proteins. In addition, supervised learning of a predictor should have a training dataset with a complete description of the modelled phenomena. Probably more experimental measurements on protein conformational stability are needed for a more successful machine learning approach. Nevertheless, our approach provided a useful and fast prediction of the stability of protein sequences.

Table 6. Classification accuracies of the optimum RBF–SVM model for the $\Delta\Delta G$ signs upon mutations of the single mutants in the test set according to the secondary structure allocation and the ASA of the mutated residue.

Mutation type (%)	Q^2	$Q(+)$	$Q(-)$
Helix (42)	0.61	0.78	0.57
Sheet (23)	0.50	0.82	0.25
Turn (13)	0.44	0.71	0.35
Coil (12)	0.50	0.50	0.50
Buried (51)	0.59	0.84	0.52
Partially buried (24)	0.46	0.56	0.42
Exposed (14)	0.50	0.90	0.30

Note: The relative fraction of each mutation type in the single point mutation test set of 222 single point mutants are in brackets. + and −, the indexes were evaluated for positive and negative $\Delta\Delta G$ signs; Q^2 is the number of correct predictions/number of examples; $Q(s)$ is the number of correct prediction for class s/observed in class s.

Training single point dataset should be complemented as more experimental conformational stability studies are published and collected in the Protherm database [43]. Beyond the agreeable results obtained in crossvalidation experiments and the poor results for the test set, the major advantage of our approach is the capability of our classifier to predict stability's changes upon double or multiple mutations. A preliminary version of the predictor is available online at http://gibk21.bse.kyutech.ac.jp/llamosa/ddG-AASA/ddG_AASA.html.

3.4 Model analysis

Interestingly, relevant amino acid/residue properties appear weighting the optimum *AASA* vectors: three structural (N_m, f and ASA_N), one secondary structure-related (P_B), two physico-chemical (P and H_t) and three thermodynamic (G_{hN}, ΔG_C and ΔCp_h) properties. These relevant autocorrelations were found at lags from 7 to 15, medium to large range interactions on the sequence. The occurrence in our models of structural, secondary structure-related, physico-chemical and thermodynamic properties reveals the complexity of the interactions ruling protein stability, which has better accessibility by a multifactor approach.

Distributions of structural properties at lags ranging from 11 to 15 reflect the significance of an adequate amino acid frame at large ranges in the primary structure, resembling certain polypeptidic structural pattern. Number of medium range contacts (N_m) is a property that contains information of tridimensional proximities of residues in space. The property N_m also appeared weighting optimum autocorrelation vectors in a neural network implemented for modelling the conformational stability of chymotrypsin inhibitor two mutants. Shape-related amino acid property, flexibility (f), appears relevant at autocorrelations of large range encoding the distribution of freedom degrees on the sequence. Re-accommodation of residue side-chains is a critical step in protein folding after amino acid substitution. Mutations may cause an unfavourable packing energy due to the rigidity of surrounding residues or, alternatively, the substituting residues themselves may be forced into unfavourable rotational isomers. Similarly, some surroundings of mutation positions may be deformable or compensated effects exist that do not yield net packing energy change [48]. The property f also appeared weighting optimum *AASA* vectors for modelling conformational stability and functional variations upon mutations of gene V protein [24] and ghrelin receptor [44], respectively. Another relevant structural property is the solvent-ASA for native protein (ASA_N), which is a measure of the number of amino acid atoms interacting with solvent molecules in the native state [55]. Interestingly, solvent-ASA was reported by Gromiha et al. [13] among the properties most linearly correlated with the changes of unfolding Gibbs free energy change for a diverse set of protein mutants. In this connection, one of the simplest and most widely used models for calculating hydration heat capacity in proteins is the solvent-ASA model [56].

In turn, the relevance β-structure tendency strongly suggested that optimum secondary structure pattern is another key factor for a stable tertiary conformation. Point mutations studies have highlighted the role of secondary structure propensities in protein stability. By manipulating favourable and unfavourable secondary structure propensities at certain positions in a protein can produce significant variations in protein stability [12]. In addition, we recently reported that secondary structure propensities also appeared relevant in a neural network model of the conformational stability of gene V protein mutants [27].

Hydrophilicity/hydrophobicity related properties such as polarity (P) and thermodynamic transfer hydrophobicity (H_t) are also important for predicting protein conformational stability according to our optimum SVM models. The relevant autocorrelations of such properties appear at medium lags. Hydrophilic interactions between amino acid residues at protein surface usually appear at medium and long ranges. Despite being separated by long stretches of polypeptide in the primary sequence, surface groups lie next to each other in space. On the contrary, hydrophobic interactions at protein core mainly appear at shorter range in the sequence. The P and H_t properties were previously found relevant for neural network modelling of conformational stability of human lysozyme mutants [26]. The autocorrelation vectors weighed by these properties encoded the role of hydrophylic interactions on the surface and hydrophobic interactions on the core to maintain protein folding and stability. Furthermore, on the protein surface, frequently appear hydrophobic patches are defined as clusters of neighbouring apolar atoms accessible on a given protein surface [56]. In addition, the hydrophobic part of the solvent-accessible surface of a typical monomeric globular protein consists of a single, large interconnected region formed from faces of apolar atoms and constituting approximately 60% of the solvent-ASA [57]. At the light of these facts, the combination of hydrophilicity/hydrophobicity and solvent-accessible surface properties could also encode hydrophobic patches patterns of protein mutants.

Thermodynamical properties for quantifying unfolding and hydration processes of proteins [unfolding Gibbs free energy change of side-chain (ΔG_C), Gibbs free energy change of hydration for native protein (G_{hN}) and hydration heat capacity change (ΔCp_h)] are relevant to model protein conformational stability. In previous reports, the property G_{hN} was found relevant in SVM models of the functional variations upon mutations of ghrelin receptor [47], meanwhile ΔG_C and ΔCp_h properties were important in neural network modelling of human lysozymes conformational stability [26]. ΔCp_h measurements in proteins mean the variation of heat capacity (Cp), which is consequence of the hydration of amino acid groups. Considering that protein unfolding usually has a positive ΔCp, polar groups hydration is accompanied by a decrease in Cp; meanwhile, apolar groups hydration increases this magnitude [56]. In this sense, Makhatadze and Privalov [58] found a good relation between ΔCp_h and surface area. On the other hand, G_{hN}, is a measure of spontaneity of the hydration process. Free energy has a direct relationship to a primary observable, the equilibrium constant K, through $\Delta G = -kT \ln K$, which describes the balance between enthalpy and entropy. In addition, Makhatadze and Privalov [58] showed that the compact native state of a protein is stabilised by the enthalpic interactions between internal groups. Hydration effects are clearly significant for protein unfolding; the evidence showed that hydration is the major effect [59]. The strongest current evidence is that it can be accounted for heat capacity change of unfolding for many proteins by adding up hydration contributions from individual residues [56]. In summary all those thermodynamical properties are related to the unfolding denaturation mechanism hypothesis. For denaturation process of globular proteins, Privalov and Gill [59,60] pointed out the hydration equilibrium, polar interactions between solvent and polar residues in the protein, as the main causes of unfolding while hydrophobic interactions in the protein core contribute to keep the folded state.

4. Conclusions

Protein structures are stabilised by numerous intramolecular interactions such as hydrophobic, electrostatic, van der Waals and hydrogen bond. Stability changes induced by mutations have been analysed by various computational methods but most of them require X-ray structural analysis and also they have a limited prediction accuracy.

Protein primary structure-based methods are less computationally intense and do not require X-ray crystal structure of proteins for implementation. Due to the availability of some thermodynamic data on protein stability, it is possible to use a structure–property relationship approach to modelling its properties. We extended the concept of autocorrelation vectors in molecules to the amino acid sequence of proteins as a tool for encoding protein structural information. In this sense, novel (*AASA*) vectors were obtained calculating autocorrelations on the protein primary structure of 48 amino acid/residue properties from the AAindex database. GA-SVMs proves to be a powerful technique to determine the relevant factors in the modelling phenomena of classification and function mapping. This approach yielded an adequate classification model for the conformational stability of protein mutants describing nearly 80% of correct classifications in cross-validation experiment. Similarly, the regression model described nearly 50% of crossvalidation data variance. Our approach also allows us to study multiple mutations. Despite low test set prediction accuracy, stable single and double point mutants were recognised with adequate accuracies about 70%. Optimum *AASA* vectors, selected by GA-SVM approach, showed that conformational stability model depends on a combination of structural, secondary structure-related, physico-chemical and thermodynamical properties mainly associated with protein hydration process.

This work demonstrates the successful application of the *AASA* vectors to modelling protein conformational stability in combination with SVMs. Encoding amino acid properties and protein primary structure information on a same pool of descriptors are more appropriate than those which exclusively consider amino acid substitution information. This approach leads to a powerful method for the scientific community interested in protein prediction studies.

Acknowledgements

The authors would like to acknowledge Professor Akinori Sarai, who provided useful information to prepare the revised manuscript. The authors also would like to thank the anonymous referees because their useful comments improved the quality of the manuscript. Financial supports of this research by Cuban Ministerio de Ciencia, Tecnología y Medio Ambiente (CITMA) through a grant to M. Fernandez (Grant No. 20104102).

References

[1] J.E.S. Wikberg, M. Lapinsh, and P. Prusis, *Proteochemometrics: A tool for modelling the molecular interaction space*, in *Chemogenomics in Drug Discovery – A Medicinal Chemistry Perspective*, H. Kubinyi and G. Müller, eds., Wiley-VCH, Weinheim, 2004, pp. 289–309.

[2] (a) J. Saven, *Combinatorial protein design*, Curr. Opin. Struct. Biol. 12 (2002), p. 453. (b) J. Mendes, R. Guerois, and L. Serrano, *Energy estimation in protein design*. Curr. Opin. Struct. Biol. 12 (2002), p. 441.

[3] D.N. Bolon, J.S. Marcus, S.A. Ross, and S.L. Mayo, *Prudent modeling of core polar residues in computational protein design*, J. Mol. Biol. 329 (2003), p. 611.

[4] L.L. Looger, M.A. Dwyer, J.J. Smith, and H.W. Helling, *Computational design of receptor and sensor proteins with novel functions*, Nature 423 (2003), p. 185.

[5] L.X. Dang, K.M. Merz, and P.A. Kollman, *Free-energy calculations on protein stability: Thr-1573Val-157 mutation of T4 lysozyme*, J. Am. Chem. Soc. 111 (1989), p. 8505.

[6] T. Lazaridis and M. Karplus, *Effective energy functions for protein structure prediction*, Curr. Opin. Struct. Biol. 10 (2000), p. 139.

[7] C. Lee and M. Levitt, *Accurate prediction of the stability and activity effects of site-directed mutagenesis on a protein core*, Nature 352 (1991), p. 448.

[8] C. Lee, *Testing homology modeling on mutant proteins: Predicting structural and thermodynamic effects in the Ala98–Val mutants of T4 lysozyme*, Fold. Des. 1 (1995), p. 1.

[9] C.M. Topham, N. Srinivasan, and T.L. Blundell, *Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables*, Protein Eng. 10 (1997), p. 7.

[10] D. Gilis and M. Rooman, *Prediction of stability changes upon single site mutations using database-derived potentials*, Theor. Chem. Acc. 101 (1999), p. 46.

[11] (a) E. Lacroix, A.R. Viguera, and L. Serrano, *Elucidating the folding problem of alpha-helices: Local motifs, long-range electrostatics, ionic-strength dependence and prediction of NMR parameters*, J. Mol. Biol. 284 (1998), p. 173. (b) V. Munoz and L. Serrano, *Development of the multiple sequence approximation within the AGADIR model of alpha-helix formation: Comparison with Zimm–Bragg and Lifson–Roig formalisms*. Biopolymers 41 (1997), p. 495.

[12] R. Guerois, J.E. Nielsen, and L. Serrano, *Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations*, J. Mol. Biol. 320 (2002), p. 369.

[13] M.M. Gromiha, M. Oobatake, H. Kono, H. Uedaira, and A. Sarai, *Relationship between amino acid properties and protein stability: Buried mutations*, J. Protein Chem. 18 (1999), p. 565.

[14] ———, *Role of structural and sequence information in the prediction of protein stability changes: Comparison between buried and partially buried mutations*, Protein Eng. 12 (1999), p. 549.

[15] ———, *Importance of surrounding residues for protein stability of partially buried mutations*, J. Biomol. Struct. Dyn. 18 (2000), p. 1.

[16] H. Zhou and Y. Zhou, *Stability scale and atomic solvation parameters extracted from 1023 mutation experiment*, Proteins 49 (2002), p. 483.

[17] S. Levin and B.H. Satir, *POLINA: Detection and evaluation of single amino acid substitutions in protein superfamilies*, Bioinformatics 14 (1998), p. 374.

[18] C.M. Frenz, *Neural network-based prediction of mutation-induced protein stability changes in staphylococcal nuclease at 20 residue positions*, Proteins 59 (2005), p. 147.

[19] E. Capriotti, P. Fariselli, and R. Casadio, *A neural-network-based method for predicting protein stability changes upon single mutations*, Bioinformatics 20 (2004), p. 63.

[20] E. Capriotti, P. Fariselli, R. Calabrese, and R. Casadio, *Prediction of protein stability changes from sequences using support vector machines*, Bioinformatics 21 (2005), p. 54.

[21] E. Capriotti, P. Fariselli, and R. Casadio, *I-Mutant2.0: Predicting stability changes upon mutation from the protein sequence or structure*, Nucleic Acids Res. 33 (2005), p. 306.

[22] L.-T. Huang, M.M. Gromiha, and S.-Y. Ho, *iPTREE-STAB: Interpretable decision tree based method for predicting protein stability changes upon mutations*, Bioinformatics 23 (2007), p. 1292.

[23] J. Cheng, A. Randall, and P. Baldi, *Prediction of protein stability changes for single-site mutations using support vector machines*, Proteins 62 (2006), p. 1125.

[24] V. Parthiban, M.M. Gromiha, C. Hoppe, and D. Schomburg, *Structural analysis and prediction of protein mutant stability using distance and torsion potentials: Role of secondary structure and solvent accessibility*, Proteins 64 (2006), p. 41.

[25] (a) R. Ramos de Armas, H. González-Díaz, R. Molina and E. Uriarte, *Markovian backbone negentropies: Molecular descriptors for protein research. I. Predicting protein stability in arc repressor mutants*, Proteins 56 (2004), p. 715. (b) H. González-Díaz, R. Molina, and E. Uriarte, *Recognition of stable protein mutants with 3D stochastic average electrostatic potentials*, FEBS Lett. 579 (2005), p. 4297. (c) Y. Marrero-Ponce, R. Medina-Marrero, J.A. Castillo-Garit, V. Romero-Zaldivar, F. Torrens, and E.A. Castro, *Protein linear indices of the 'macromolecular pseudograph α-carbon atom adjacency matrix' in bioinformatics. Part 1: Prediction of protein stability effects of a complete set of alanine substitutions in Arc represor*, Bioorg. Med. Chem. 13 (2005), p. 3003. (d) G. Agüero-Chapin, H. González-Díaz, R. Molina, J. Varona Santos, E. Uriarte, and Y. González-Díaz, *Novel 2D maps and coupling numbers for protein sequences. The first QSAR study of polygalacturonase; isolation and prediction of a novel sequence from* Psidium guajava *L.*, FEBS Lett. 580 (2006), p. 723.

[26] J. Caballero, L. Fernández, J.I. Abreu, and M. Fernández, *Amino acid sequence autocorrelation vectors and ensembles of Bayesian-regularized genetic neural networks for prediction of conformational stability of human lysozyme mutants*, J. Chem. Inf. Model. 46 (2006), p. 1255.

[27] L. Fernández, J. Caballero, J.I. Abreu, and M. Fernández, *Amino acid sequence autocorrelation vectors and Bayesian-regularized genetic neural networks for modeling protein conformational stability: Gene V protein mutants*, Proteins 67 (2007), p. 834.

[28] M. Fernández, J. Caballero, A.M. Helguera, E.A. Castro, and M.P. González, *Quantitative structure–activity relationship to predict differential inhibition of aldose reductase by flavonoid compounds*, Bioorg. Med. Chem. 13 (2005), p. 3269.

[29] M. Fernández, A. Tundidor-Camba, and J. Caballero, *2D autocorrelation modeling of the activity of trihalobenzocycloheptapyridine analogues as farnesyl protein transferase inhibitors*, Mol. Simulat. 31 (2005), p. 575.

[30] ———, *Modeling of cyclin-dependent kinase inhibition by 1H-pyrazolo [3,4-d] pyrimidine derivatives using artificial neural networks ensembles*, J. Chem. Inf. Comput. Sci. 45 (2005), p. 1884.

[31] M.P. González, J. Caballero, A. Tundidor-Camba, A.M. Helguera, and M. Fernández, *Modeling of farnesyltransferase inhibition by some thiol and non-thiol peptidomimetic inhibitors using genetic neural networks and RDF approaches*, Bioorg. Med. Chem. 14 (2006), p. 200.

[32] M. Fernández and J. Caballero, *Modeling of activity of cyclic urea HIV-1 protease inhibitors using regularized-artificial neural networks*, Bioorg. Med. Chem. 14 (2006), p. 280.

[33] J. Caballero and M. Fernández, *Linear and nonlinear modeling of antifungal activity of some heterocyclic ring derivatives using multiple linear regression and Bayesian-regularized neural networks*, J. Mol. Model. 12 (2006), p. 168.

[34] M. Fernández, J. Caballero, L. Fernández, J.I. Abreu, and G. Acosta, *Classification of conformational stability of protein mutants from 2D graph representation of protein sequences using support vector machines*, Mol. Simulat. 33 (2007), p. 889.

[35] ———, *Classification of conformational stability of protein mutants from 3D pseudo-folding graph representation of protein sequences using support vector machines*, Proteins 70 (2008), p. 167.

[36] H. Bauknecht, A. Zell, H. Bayer, P. Levi, M. Wagener, J. Sadowski, and J. Gasteiger, *Locating biologically active compounds in medium-sized heterogeneous datasets by topological autocorrelation vectors: Dopamine and benzodiazepine agonists*, J. Chem. Inf. Comput. Sci. 36 (1996), p. 1205.

[37] P.A.P. Moran, *Notes on continuous stochastic processes*, Biometrika 37 (1950), p. 17.

[38] R.F. Geary, *The contiguity ratio and statistical mapping*, Incorporated Statistician 5 (1954), p. 115.

[39] G. Moreau and P. Broto, *Autocorrelation of a topological structure: A new molecular descriptor*, Nouv. J. Chim. 4 (1980), p. 359.

[40] ———, *Autocorrelation of molecular structures: Application to SAR studies*, Nouv. J. Chim. 4 (1980), p. 757.

[41] M. Wagener, J. Sadowski, and J. Gasteiger, *Autocorrelation of molecular properties for modelling corticosteroid binding globulin and cytosolic Ah receptor activity by neural networks*, J. Am. Chem. Soc. 117 (1995), p. 7769.

[42] (a) K. Nakai, A. Kidera, and M. Kanehisa, *Cluster analysis of amino acid indices for prediction of protein structure and function*, Protein Eng. 2 (1988), p. 93. (b) K. Tomii and M. Kanehisa, *Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins*, Protein Eng. 9 (1996), p. 27. (c) S. Kawashima and M. Kanehisa, *AAindex: Amino acid index database*, Nucleic Acids Res. 28 (2000), p. 374.

[43] K.A. Bava, M.M. Gromiha, H. Uedaira, K. Kitajima, and A. Sarai, *ProTherm, version 4.0: Thermodynamic database for proteins and mutants*, Nucleic Acids Res. 32 (2004), p. 120. Available at http://gibk26.bse.kyutech.ac.jp/jouhou/protherm/protherm.html.

[44] (a) C. Cortes and V. Vapnik, *Support-vector networks*, Mach. Learn. 20 (1995), p. 273. (b) C.J.C. Burges, *A tutorial on support vector machines for pattern recognition*, Data Min. Knowledge Discov 2 (1998), p. 1. (c) V. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998.

[45] C. Chih-Chung and L. Chih-Jen, *LIBSVM: A library for support vector machines*, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm (2001).

[46] (a) W. Lua, N. Donga, and G. Náray-Szabó, *Predicting anti-HIV-1 activities of HEPT-analog compounds by using support vector classification*, QSAR Comb. Sci. 24 (2005), p. 1021. (b) X. Yao, H. Liu, R. Zhang, M. Liu, Z. Hu, A. Panaye, J.P. Doucet, and B. Fan, *QSAR and classification study of 1,4-dihydropyridine calcium channel antagonists based on least squares support vector machines*, Mol. Pharm. 2 (2005), p. 348. (c) H. Fröhlich, J.K. Wegner, and A. Zell, *Towards optimal descriptor subset selection with support vector machines in classification and regression*, QSAR Comb. Sci. 23 (2004), p. 311.

[47] J. Caballero, L. Fernández, M. Garriga, J.I. Abreu, S. Collina, and M. Fernández, *Proteometric study of ghrelin receptor function variations upon mutations using amino acid sequence autocorrelation vectors and genetic algorithm-based least square support vector machines*, J. Mol. Graph. Model. 26 (2007), p. 166.

[48] H. Fröhlich, O. Chapelle, and B. Schölkopf, *Feature selection for support vector machines by means of genetic algorithms*, in *Proceedings of 15th IEEE International Conference on Tools with AI*, 2003, pp. 142–148.

[49] M. Fernandez, *GlibSVM toolbox for Matlab version 1.0*, Molecular Modeling Group, University of Matanzas, 2007.

[50] The MathWorks Inc. *Genetic algorithm and direct search toolbox user's guide for use with MATLAB*, The Mathworks Inc., Massachusetts (2004).

[51] (a) W.S. Sandberg and T.C. Terwilliger, *Energetics of repacking a protein interior*, Proc. Natl Acad. Sci. U S A 88 (1991), p. 1706. (b) W.S. Sandberg and T.C. Terwilliger, *Engineering multiple properties of a protein by combinatorial mutagenesis*, Proc. Natl Acad. Sci. USA 90 (1993), p. 8367.

[52] H. Zhou and Y. Zhou, *Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction*, Protein Sci. 11 (2002), p. 2714.

[53] A.J. Bordner and R.A. Abagyan, *Large-scale prediction of protein geometry and stability changes for arbitrary single point mutations*, Proteins 57 (2004), p. 400.

[54] H. Gonzalez-Diaz, R. Molina, and E. Uriarte, *Recognition of stable protein mutants with 3D stochastic average electrostatic potentials*, FEBS Lett. 579 (2005), p. 4297.

[55] N.V. Prabhu and K.A. Sharp, *Heat capacity of proteins*, Annu. Rev. Phys. Chem. 56 (2005), p. 521.

[56] P. Lijnzaad and P. Argos, *Hydrophobic patches on protein subunit interfaces: Characteristics and prediction*, Proteins 28 (1997), p. 333.

[57] F. Eisenhaber and P. Argos, *Hydrophobic regions on protein surfaces: Definition based on hydration shell structure and a quick method for their computation*, Protein Eng. 9 (1996), p. 1121.

[58] (a) G.I. Makhatadze and P.L. Privalov, *Heat capacity of proteins. I. Partial molar heat capacity of individual amino acid residues in aqueous solutions: Hydration effect*, J. Mol. Biol. 213 (1990), p. 375. (b) P.L. Privalov and G.I. Makhatadze, *Partial molar heat capacity of the unfolded polypeptide chain of proteins: Protein unfolding effects*, J. Mol. Biol. 213 (1990), p. 385.

[59] ———, *Hydration effects in protein unfolding*, Biophys. Chem. 51 (1994), p. 291.

[60] P.L. Privalov and S.J. Gill, *Stability of protein structure and hydrophobic interaction*, Adv. Protein Chem. 39 (1988), p. 191.