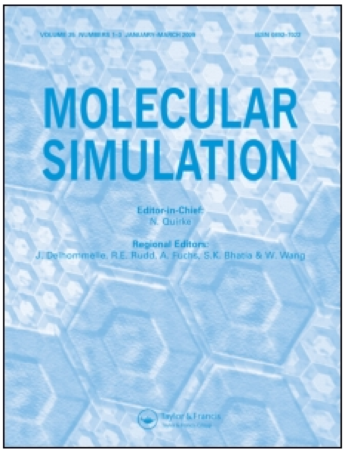

Molecular Simulation
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713644482

Classification of conformational stability of protein mutants from 2D graph representation of protein sequences using support vector machines

M. Fernández[a]; J. Caballero[ab]; L. Fernández[a]; J. I. Abreu[c]; G. Acosta[d]
[a] Molecular Modelling Group, Faculty of Agronomy, Center for Biotechnological Studies, University of Matanzas, Matanzas, Cuba [b] Centro de Bioinformática y Simulación Molecular, Universidad de Talca, Talca, Chile [c] Artificial Intelligence Lab, Faculty of Informatics, University of Matanzas, Matanzas, Cuba [d] National Bioinformatics Center, Havana, Cuba

Online publication date: 27 July 2010

To cite this Article Fernández, M. , Caballero, J. , Fernández, L. , Abreu, J. I. and Acosta, G.(2007) 'Classification of conformational stability of protein mutants from 2D graph representation of protein sequences using support vector machines', Molecular Simulation, 33: 11, 889 — 896

To link to this Article: DOI: 10.1080/08927020701377070

URL: http://dx.doi.org/10.1080/08927020701377070

Molecular Simulation, Vol. 33, No. 11, September 2007, 889–896

Taylor & Francis
Taylor & Francis Group

Classification of conformational stability of protein mutants from 2D graph representation of protein sequences using support vector machines

M. FERNÁNDEZ[†]*, J. CABALLERO[†][‡], L. FERNÁNDEZ[†], J. I. ABREU[†][¶] and G. ACOSTA[§]

[†]Molecular Modelling Group, Faculty of Agronomy, Center for Biotechnological Studies, University of Matanzas, 44740 Matanzas, Cuba
[‡]Centro de Bioinformática y Simulación Molecular, Universidad de Talca, 2 Norte 685, Casilla 721, Talca, Chile
[¶]Artificial Intelligence Lab, Faculty of Informatics, University of Matanzas, 44740 Matanzas, Cuba
[§]National Bioinformatics Center, 10200 Havana, Cuba

(Received March 2007; in final form April 2007)

Euclidean distance counts derived from the protein 2D graphs were used for encoding protein structural information. A total of 35 amino acid 2D distance count (*AA2DC*) descriptors were calculated from the Euclidean distance matrices (EDM) derived from the 2D graphs at distances ranging from 0.05 to 1.8 units with a lag of 0.05 units. *AA2DC* descriptors were tested for building predictive classification model of the signs of the change of thermal unfolding Gibbs free energy change ($\Delta\Delta G$) of a large data set of 2048 single point mutations on 64 proteins. A support vector machine (SVM) classifier with a Radial Basis Function kernel was implemented for classifying the conformational stability of protein mutants. Temperature and pH of the $\Delta\Delta G$ experimental measurements were also conveniently used for SVM training in addition to calculated *AA2DC* descriptors. The optimum SVM model correctly predicted about 72% of $\Delta\Delta G$ signs in crossvalidation test for all the dataset and also for stable and unstable mutant separately. To the best of our knowledge, this level of accuracy for stable mutant recognition is the highest ever reported for a predictor using sequence information. Furthermore, the classifier adequately recognized unstable mutants of human prion protein and human transthyretin associated to diseases.

Keywords: Protein stability prediction; Point mutations; Kernel-based methods; Graph similarity

1. Introduction

Evidence is accumulated that many disease-causing mutations exert their effects by altering protein folding. Predicting proteins structures and stabilities are fundamental goals in molecular biology, therefore, predicting changes in structure and stability induced by point mutations has immediate application in computational protein design [1–4]. Despite free energy simulations have accurately predicted relative stabilities of point mutants [5], the computational costs that most of the methods actually demand, are extremely high to test the large number of mutations studied in protein design applications.

Fast algorithms for protein energy calculations are being developed today for intending the translation of structural data into energetic parameters. However, the development of fast and reliable protein force-fields is a complex task due to the delicate balance between the different energy terms that contribute to protein stability. Force-fields for predicting protein stability can be divided in three main groups: physical effective energy function (PEEF), statistical potential-based effective energy function (SEEF) [6] and empirical data-based effective energy function (EEEF).

A simplified energy function with only Van der Waals and side chain torsion potentials [7], a PEEF approach, has been used to predict the stabilities of the λ repressor protein for mutations involving only hydrophobic residues. In addition, an improved optimization method including continuously flexible side chain angles also demonstrated better prediction accuracy as compared to discrete side chain angles from a rotamer library [8]. In turn SEEF method includes statistical potentials derived from geometric and environmental propensities and correlations of residues in X-ray crystal structures [6,9,10]. On the othear hand, EEEF approach combines a physical description of the interactions with some data

*Corresponding author. Tel.: +53 45 26 1251. Fax: +53 45 25 3101. Email: michael.fernandez@umcc.cu; Email: michael_llamosa@yahoo.com

Molecular Simulation
ISSN 0892-7022 print/ISSN 1029-0435 online © 2007 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/08927020701377070

obtained from experiments previously ran on proteins. Examples of such algorithms are the helix/coil transition algorithm AGADIR [11] or FOLDEF, a fast and accurate EEEF approach based on AGADIR algorithm that uses a full atomic description of the structure of the proteins was reported by Guerois *et al.* [12] for predicting conformational stability of more than 1000 mutants.

Gromiha *et al.* [13–15] reported stability prediction studies not based on protein force-field calculations but focused on correlations of free energy change with structural, sequence information and amino acid properties such as hydrophobicity, accessible surface area, etc. On the other hand, empirical equations involving physical properties calculated from mutant structures have been reported. Zhou and Zhou [16] developed a broad study regarding 35 proteins and 1023 mutants from which they derived a new stability scale.

Machine learning algorithms have been also applied to the protein stability prediction problem. In this connection, outstanding reports of Capriotti *et al.* [17,18] describe the implementation of neural network and support vector machine (SVM) classifiers of change of protein free energy change upon mutations by using sequence and 3D structure information. This approach allows to qualitative and quantitative predict stability change using a data set of more than 2000 mutations for training and testing the models. As network and vector machine inputs they used a combination of experimental condition data (pH and temperature), specific mutated residue information and environmental residues information.

Furthermore, recent reports refer the novel extensions of different structure/property relationships approaches to the prediction of protein stability using both sequence [19–22] and 3D structure based information [23]. In such reports, descriptors are calculated over the protein sequences or 3D structures in such a way that several variables are computed considering the protein structure as a simplified molecular pseudo-graph of Cα atoms.

In this work, an optimum classification model for the conformational stability of protein mutants was successfully built from protein sequences. A novel 2D graph representation of protein primary sequence, recently reported by Randić *et al.* [24], was used in order to enrich the information derived from the protein primary structure. A total of 35 amino acid 2D distance count (*AA2DC*) descriptors were calculated from the Euclidean distance matrices (EDM) computed for the residues on the protein 2D graphs. A large set of 2048 mutations on 64 proteins, previously collected by Capriotti *et al.* [17], was used for deriving a general protein predictor for recognition of stable and unstable protein mutants. Prediction of the signs of change of thermal unfolding Gibbs free energy change ($\Delta\Delta G$) upon single mutations was accomplished by SVM classification. Temperature and pH of the $\Delta\Delta G$ experimental measurements were also used as SVM inputs in addition to computed *AA2DC* descriptors.

2. Materials and methods

2.1 Single point mutant dataset

The dataset was the same previously collected for Capriotti *et al.* [17] for deriving predictive models for the signs and the actual values of $\Delta\Delta G$ by neural networks and SVMs. Those authors used sequence and amino acid substitution information in conjunction with the experimental conditions of the experimental thermodynamic determination for training the predictors. Capriotti *et al.* collected the data from the Protherm data base [25] according to the following constrains:

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

After the filtering they gather 2048 single point mutations obtained from 64 proteins. The dataset is available at http://gpcr2.biocomp.unibo.it/~emidio/I-Mutant2.0/dbMutSeq.html

2.2 Randić 2D graph representation of protein sequences and amino acid distance count (*AA2DC*) descriptors

In addition to the intensive computation required by the free energy function based methods for predicting protein stability, another limitation arises when considering that X-ray crystal structures of the mutants under study are needed [1–12]. Despite protein crystallographic data base continuously grows, crystal structures are not always available for proteins of interest. In this regard, some X-ray structural-independent protein stability prediction methods have gained attention. The main advantages of such methods are the use of amino acids sequence information for predicting protein stability and their extremely less computational cost in comparison with free energy function based methods.

Exploitation of protein sequences for prediction and similarity studies have been extended by developing 2D [24,26], 3D [27,28] graph representations of sequences as well as weighted linear graph representations [19–22,29]. The first of these approaches intended to enrich the structural information derived from the protein primary structure by mapping sequence residues in a higher dimensional space. Among these reports, Randić *et al.* [25] proposed a highly compact graphical representation for proteins, a version of the “chaos game representation” of DNA and protein sequences [30], which offers graphical and numerical characterization of individual proteins. Protein representations are constructed in the interior of a unit “magic circle” (figure 1A), on the circumference of which at equal distances are positioned 20 amino acids. Graphical

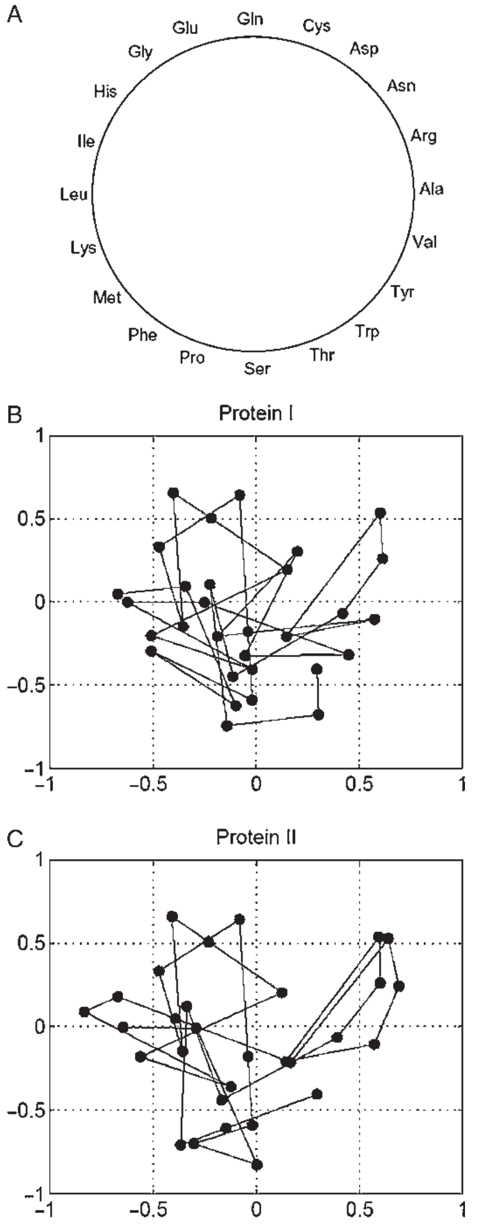

Figure 1. Graphical representation of the "magic circle" proposed by Randić *et al.* [24] for the 2D representation of protein sequence (A). 2D graph representation of example proteins: Protein I WTFESRNDPAKDPVILWLNGGPGCSSLTGL (B) Protein II WFFESRNDPANDPIILWLNGGPGCSSFTGL (C).

representation of proteins is depicted by starting at the center of the circle following amino acids sequence and moving half way towards the corresponding residue, analogously to the scheme of Jeffrey for graphical representation of DNA [30]. Graphical form of depicted protein depends of the amino acids ordering, however, the relative variations between proteins sequences remain to large extent independent of such ordering along the circle periphery. As example, figure 1B and 1C, depict graphical representations of two shorter protein segments of yeast *Saccharomyces cerevisiae* already used for Randić *et al.* [25] for illustrating their method:

Protein I WTFESRNDPAKDPVILWLNGGPGCS-SLTGL
Protein II WFFESRNDPANDPIILWLNGGPGCS-SFTGL

From the protein 2D representations in figure 1, the amino acid EDM of the graph can be computed. The EDM represent the relative distance among nodes (amino acid residues) in the Randić 2D graphs representation of the sequence. Afterwards, similarity between such graphs can be assessed by calculating Euclidean Distance Count in such a way that pairs of nodes (amino acid residues) at discrete distances are counted. *AA2DC* descriptors are then computed using equation (1).

$$AA2DC_l = \frac{1}{L}\sum_i \delta_{ij} \tag{1}$$

where L is the length of the protein sequence used for normalizing according to the size of the sequence and $\delta(l, d_{ij})$ is a Dirac-delta function defined as:

$$\delta(l, s, d_{ij}) = \begin{Bmatrix} 1 & \text{if } l - \frac{s}{2} < d_{ij} \leq l + \frac{s}{2}, \\ 0 & \text{otherwise} \end{Bmatrix} \tag{2}$$

where the d_{ij} is the Euclidean distance between amino acid residues i and j in the 2D protein graph, l and s are the Euclidean distance and the step used for the distance count, respectively.

Node pair summations were carried out from an initial distance l of 0.05–1.8 units at distance steps s of 0.05 units. Finally, a total of 35 *AA2DC* descriptors were computed for discriminating among mutant sequences. Computational codes for protein sequence 2D graphs generation and *AA2DC* descriptors calculation were written in Matlab environment [31] and the M-file is available from the authors upon request. Before classification, the calculated variables for the 2048 mutations used in this study were scaled to the [−1, 1] range.

2.3 Support vector machine (SVM)

The SVM, a new machine learning method, has been used for many kinds of pattern recognition problems. Since excellent introductions to SVM appear in Ref. [32], only the main idea of SVM 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 minimize structural risk whereas previous techniques were usually based on minimization of empirical risk. So SVM is usually 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 regularization parameter. The kernel function and its specific parameters, together with regularization parameter, can not be set from the optimization problem but have to be tuned by the user. These can be optimized by the use of Vapnik-Chervonenkis bounds, crossvalidation, an independent optimization set, or Bayesian learning. In this paper, the radial basic function (RBF) was used as kernel function. A crossvalidation was developed for setting the optimized values of the two parameters: the regularization parameter and the width of the RBF kernel. For implementing the SVM it was used the toolbox LIBSVM for Matlab by Chang and Lin [33] that can be downloaded from: http://www.csie.ntu.edu.tw/cjlin/libsvm/.

2.4 Model's validation

The efficiency of the SVM predictor for protein mutant classification was evaluated by the same set of statistics used by Capriotti *et al.* in Ref. [17] and listed below.

The overall accuracy is

$$Q^2 = \frac{p}{N} \tag{3}$$

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 follow:

$$C(s) = \frac{[p(s)n(s) - u(s)o(s)]}{D} \tag{4}$$

where D is the normalization factor

$$D = [(p(s) + u(s))(p(s) + o(s))(n(s) + u(s))(n(s) + o(s))]^{1/2} \tag{5}$$

for each class s ($+$ and $-$ for positive and negative $\Delta\Delta G$ values); $p(s)$ and $n(s)$ are the number of correct predictions and correctly rejected assignments, respectively, and $u(s)$ and $o(s)$ are the number or under- and over-predictions.

The coverage for each discriminant structure s is evaluated as

$$Q_S = \frac{p(s)}{p(s) + u(s)} \tag{6}$$

The accuracy for s is computed as

$$P_S = \frac{p(s)}{p(s) + o(s)} \tag{7}$$

where $p(s)$ and $u(s)$ are the same as in equation (4).

In order to avoid over estimation of the model's predictive power, similar mutants where kept in the same set during crossvalidation even when reported at different experimental Temperature and pH conditions.

3. Results and discussion

2D graph representation of proteins attempts to discriminate among protein sequences according to a differentiated 2D spatial distribution of residues in a bidimensional map. Thus, the unidimensional space of protein primary structure is translated to a bidimensional one. Consequently, sequence information changes into a more easily readable format in order to compute some descriptors for statistical pattern recognition studies. After 2D graph representation of protein sequences, we applied graph similarity measurements for obtaining a feature data matrix for SVM training. SVM predictor was trained with 35 *AA2DC* descriptors calculated over the 2D graph representations of the 2048 mutations studied (see Section 2.3 Randić 2D graph representation of protein sequences and Amino Acid Distance Count (*AA2DC*) descriptors). Temperature and pH values of the experimental $\Delta\Delta G$ measurements were also used for training the predictor. Since the 2048 mutation dataset is three-fold unbalanced towards unstable mutants, we implemented a 3-fold higher penalty for stable mutant misclassification inside the SVM framework. We optimized the SVM classifier by adjusting the value of the RBF kernel width and regularization parameter throughout a 20-fold out crossvalidation test. Optimum values of $\sigma^2 = 0.05$ and regularization parameter about 10 yielded the training and crossvalidation results that appear in table 1. An overall 20-fold out crossvalidation accuracy of 72% for the classification of mutant's $\Delta\Delta G$ signs was achieved with a correlation coefficient $C = 0.41$. It is noteworthy, that crossvalidation accuracies for recognizing stable mutants $Q(+) = 0.72$ and unstable mutants $Q(-) = 0.72$ are both equal to the overall accuracy achieved.

Table 2 shows crossvalidation results for models developed using other machine learning classifiers. A SVM predictor with a linear kernel (SVM-Linear), that was build also using LIBSVM toolbox for Matlab, showed lower predictive power in comparison to the SVM model with RBF kernel reported in table 1. This linear predictor exhibited lower crossvalidation accuracies about 0.65. In addition, we implemented Bayesian regularized artificial neural networks (BRANN) and probabilistic neural networks (PNN) classifiers within Matlab environment. These nonlinear predictors exhibited overall Q^2 accuracies slightly higher than the optimum nonlinear SVM predictor in table 1. However, correct predictions were completely shifted towards unstable mutants with $Q(-)$ values about 0.90 meanwhile stable mutants were poorly recognized with

Table 1. Training and crossvalidation statistics for the *AA2DC*-SVM model with a RBF kernel for the classification of protein mutant $\Delta\Delta G$ signs.

	Q^2	$P(+)$	$P(-)$	$Q(+)$	$Q(-)$	C
Training	0.80	0.61	0.93	0.86	0.77	0.59
Crossvalidation	0.72	0.52	0.86	0.72	0.72	0.41

$+$ and$-$, the indexes were evaluated for positive and negative $\Delta\Delta G$ signs.

Table 2. Crossvalidation statistics for the classification of protein mutant stability according to other machine learning classifiers: SVM with a linear kernel (SVM-Linear), Bayesian regularized artificial neural networks (BRANN) and probabilistic neural networks (PNN).

	Q^2	$P(+)$	$P(-)$	$Q(+)$	$Q(-)$	C
SVM-linear	0.66	0.44	0.82	0.64	0.66	0.28
BRANN	0.77	0.62	0.82	0.55	0.86	0.43
PNN	0.78	0.69	0.80	0.46	0.91	0.43

$+$ and$-$, the indexes were evaluated for positive and negative $\Delta\Delta G$ sign.

$Q(-)$ values about 0.5. By analyzing the behaviors of such models we stated that the SVM with a RBF kernel in table 1 is superior in comparison to the other tested classifiers for the protein conformational stability problem studied here.

Since it has been reported that the effects caused by a specific mutation depend on the type of substituted and new added residues, it is interesting to analyze the performance of the predictor regarding the nature of the mutations [34]. Classification results according to the chemical–physical properties of the mutations appear in table 3. By analyzing the SVM predictor accuracy as a function of the mutation type, we found that mutations involving substitutions of charged residues, mainly placed on the protein surface, by other charged or apolar residues, exhibit the lower classification accuracy. This fact suggests that sequence derived information does not properly describe the effect of salt–bridge interactions at the protein surface that should be better attempted using 3D structure details. At the protein core interactions among residues often occur at shorter range in the sequence whilst at the protein surface interactions can occur at larger ranges.

Some other classification models for protein mutant's stability have been reported. Classifiers using just sequence information were described by Gonzalez-Diaz *et al.* [19] dealing with the linear discriminant analysis (LDA) classification of Arc repressor mutants, but according to its melting point value. Similarly, this set of Arc repressor mutants was employed by Marrero-Ponce *et al.* [20] for deriving a LDA classification describing both reports more than a 80% of validation data variance using protein linear indices of the "macromolecular pseudograph Cα-atom adjacency matrix", a sequence-exploiting approach. However, above mentioned models have poor utility because they are protein-specific and use a thermodynamic parameter (melting point) not directly related to protein conformational stability. In addition, other single-protein models, previously developed by us, built with amino acids sequence autocorrelation vectors and Bayesian regularized neural networks, successfully mapped lysozymes [21] and gene V protein [22] mutants in self-organizing maps according to mutant's $\Delta\Delta G$ levels.

Table 3. Crossvalidation accuracies (Q^2) of the *AA2DC*-SVM model with a RBF kernel for the classification of $\Delta\Delta G$ signs upon mutations according to the mutation types.

		New	
Native	*Charged*	*Polar*	*Apolar*
Charged	0.66 (6%)	0.77 (9%)	0.69 (11%)
Polar	0.75 (6%)	0.72 (8%)	0.71 (15%)
Apolar	0.80 (5%)	0.72 (11%)	0.71 (29%)

On the other hand, taking into account classification models of protein stability change upon mutations with large and diverse mutant data, our classification model overcomes the optimum reported by Capriotti *et al.* [17] for the same dataset including only sequence information. Despite they reported an overall accuracy about 77%, the correct predictions were drastically shifted towards unstable mutants with a value about 91% whilst for recognizing stable mutants it was reported a very low value about 46%. Such statistics reflect that their model almost recognized all mutants as unstable yielding and overall adequate accuracy but with very low discriminating ability. In turn, our predictor recognized both stable and unstable mutants with identical good accuracy about 72% due to the higher penalty imposed to stable mutant misclassification during SVM training. In this regards, to the best of our knowledge, this is the highest accuracy reported for positive $\Delta\Delta G$ signs recognition by a model with more than 2000 mutations and only exploiting sequence information. Interestingly, when Capriotti *et al.* [18] used 3D structure information the highest overall classification accuracy achieved was about 80% but stable mutants were poorly recognized with and accuracy of 56% supporting the fact that our *AA2DC*-SVM predictor, although its primary sequence nature, is more adequate for the mutant stability recognition task. Similarly, Gonzalez-Diaz *et al.* [23] reported a LDA model for recognizing stable mutants using 3D stochastic average electrostatic potentials derived from protein 3D structure with validation accuracy nearly 90%. 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.

Since we used a whole sequence representation for studying mutation effect on stability rather than a mutation location-dependent representation such as other reports [13–15,17], it is interesting to analyze the classification results appearing in table 4 for each set of mutants from a particular protein. As can be observed, bad results ($Q^2 < 0.5$) only appeared in one case for one protein (chicken lysozyme) with > 50 mutations and ten proteins with less than 20 mutants, but the later proteins have very low statistical weight on the whole classification model. This analysis reflects the robustness of our predictor.

An interesting application of our classifier is to recognize unstable mutations when mutations have been reported to correlate to diseases. Indeed, many disease-causing mutations exert their effects by altering protein folding. In table 5 appears the classification for a series of 22 mutations for human prion protein [35] and human transthyretin [36]. Our classifier correctly recognized as unstable 13 of the 15 $\Delta\Delta G$-known mutants reported in table 5, this result represents an accuracy about 0.87 that is

Table 4. Percent crossvalidation accuracies (Q^2) of the *AA2DC*-SVM model with a RBF kernel for the classification of $\Delta\Delta G$ signs upon mutations according to mutants per protein.

Protein	*Q^2 (%)*	*Number of mutants*	*Protein*	*Q^2 (%)*	*Number of mutants*
DsbA	100	7	Ribonuclease T1	64	67
Bovine ubiquitin	60	10	Rop	67	21
Murine cellular prion	100	9	Ribonuclease A	65	17
Chicken spectrin	0	1	Dihydrofolate reductase	62	13
Cytochrome c	100	1	Ribonuclease SA	100	5
Adenylate kinase	100	4	Alpha spectrin (SH3 domain)	89	9
Arc repressor	0	3	Staphylococcal nuclease	70	44
Adrenodoxin	0	11	Subtilisin BPN’	100	6
Barnase	92	221	Plasminogen activator kringle-2 domain	83	6
Trypsin inhibitor	96	51	Tumor suppressor P53	100	5
Barstar	0	1	Gene V	84	114
Myoglobin	75	84	Iso-1 cytochrome c	62	39
Cytochrome c2	40	5	Acyl-coenzyme a binding protein	85	27
Cold shock protein *Bacillus caldolyticus*	72	50	Acidic fibroblast growth factor	0	3
Cold shock protein *Bacillus subtilis*	50	6	Adenylate kinase	100	4
Chitosanase	100	9	Chymotrypsin inhibitor 2	85	117
Carbonic anhydrase II	71	14	HPr protein	67	3
Cytochrome b5	85	20	Lysozyme bacteriophage T4	61	426
Canine lysozyme	100	4	Ribonuclease HI	69	122
Streptococcal protein G helix variant	100	12	Thioredoxin	100	2
Catabolite activator protein	0	2	Beta lactamase	100	1
Alpha lactalbumin	100	4	Maltose binding protein	83	6
HU DNA-binding protein	20	5	Subtilisin inhibitor	68	50
Calbindin D9k	100	3	Cytochrome c551	86	7
Interleukin 1 beta	71	7	Chicken Lysozyme	45	51
Human lysozyme	61	145	Lambda cro repressor	91	11
C-Myb	67	3	Phage lambda endolysin	100	2
Glycogen phosphorylase	100	1	Plasminogen activator inhibitor 1 precursor	0	2
Onconase	100	9	Fibroblast growth factor 12	100	1
Ubiquitin	87	30	Phage lambda repressor protein cl	42	12
Streptococcus protein G	87	23	Phage P22 tail protein	100	7
Histidine-containing phosphocarrier protein	76	17	Tryptophan synthase	53	76

in the range of the validation result for all mutant dataset. It should be noticed that the only two positive signs correspond to very low values of $\Delta\Delta G$ (<1.0 kJ/mol). The rest 7 $\Delta\Delta G$-unknown mutants were also predicted as unstable when some have been reported as disease provoking mutations. However, all the mutants with $\Delta\Delta G \geq \pm 2.1$ (kJ/mol) (in bold face letter in table 5) that correspond to diseases were well recognized as unstable

Table 5. Stability classification of human prion and human transthyretin mutants according to the *AA2DC*-SVM model with a RBF kernel.

Protein	*Mutation*	*Disease phenotype*	*$\Delta\Delta G$ (kJ/mol)*	*Classification*
Human prion	P102L[a]	GSD	0.8 ± 2.5	Unstable
	M129V[b]	Polymorphism	− 1.4 ± 2.0	Unstable
	D178N/M129[b]	**FFI**	**− 7.2 ± 1.7**	**Unstable**
	D178N/V129[b]	**CJD**	**− 8.0 ± 1.8**	**Unstable**
	V180I[b]	**GSD**	**− 2.1 ± 1.7**	**Unstable**
	T183A[b]	**CJD**	**− 19.3 ± 3.1**	**Unstable**
	T190V[b]	Polymorphism	0.7 ± 2.4	Unstable
	F198S[b]	**GSD**	**− 10.3 ± 1.7**	**Unstable**
	E200K[b]	CJD	− 0.6 ± 2.4	Unstable
	R208H[b]	**CJD**	**− 6.0 ± 2.5**	**Unstable**
	V210I[b]	CJD	− 1.1 ± 2.6	Unstable
	Q217R[b]	**GSD**	**− 8.9 ± 1.7**	**Unstable**
	M166V[c]	Polymorphism	SC (1E1J)	Unstable
	S170N[c]	Polymorphism	SC (1E1P)	Unstable
	R220K[c]	Polymorphism	SC (1FKC)	Unstable
Human transthyretin	**V50M**[d]	**Amyloidosis**	**− 9.2 ± 10.0**	**Unstable**
	L75P[d]	**Amyloidosis**	**− 6.3 ± 9.6**	**Unstable**
	T139M[d]	Unclassified	− 0.42 ± 11.7	Unstable
	T80A[e]	Amyloidosis	SC (1TSH)	Unstable
	S97Y[e]	Amyloidosis	SC (2TRY)	Unstable
	Y134C[e]	Amyloidosis	SC (1IIK)	Unstable
	V142I[e]	Unclassified	SC (1TTR)	Unstable

Values of $\Delta\Delta G \geq \pm 2.1$ (kJ/mol) appear in bold face letter. GSD, Gerstmann–Straussler disease; FFI, fatal familial insomnia; CJD, Creutzfeldt–Jakob disease; SC, structural conformational change determined by comparing native (1QLX, human prion protein; 1BM7 human transthyretin) with mutated 3D structures (PDB codes are reported within parenthesis).[a] From Ref. [35]. [b] From Ref. [36]. [c] From Ref. [37]. [d] From Ref. [38]. [e] From Ref. [17].

mutants. This analysis and the fact that defective protein folding is main cause of mutation related disease suggest that our model could be useful for correlating nucleotide polymorphisms and stability-related diseases.

AA2DC descriptors were able of resembling an amino acid interaction pattern that was learned by the SVM without having to be explicitly fed with residue proximities or other structural information. In this regard, conformational stability, a 3D dependent feature, was successfully modeled employing scarce information derived from protein primary sequence. The reported model improves over previous one exploiting only sequence information. Furthermore, our approach based on the similarity measurements among 2D graph representation of whole protein sequence seems to be superior to mutation point-centered methods [13–15,17]. Consequently, unlike previous reports [13–15,17], our method can handle any mutant as well as any protein polymorphism since it is not restricted to single point mutations. Despite we used only single point mutants for training the classifier in order to compare with previous reports, it should be noticed in table 5 two mutants of human prion protein (D178N/M129 and D178N/V129) involving double point changes correctly recognized as unstable mutants.

Our work intends to demonstrate the utility of the 2D graph representation of protein sequence in combination with SVMs. By using protein sequence information and a wide thermodynamic data, we built a predictor that recognizes between stable and unstable protein mutants. Despite the disadvantage of some previous thermodynamic experimental data for generating a training set, our modelling technique is an alternative stability prediction approach for proteins which lack X-ray structural information but protein sequence is known.

4. Conclusions

Protein primary structure-based methods are less computational intense and do not require X-ray crystal structure of proteins for implementation. Due to the availability of an enormous amount of thermodynamic data on protein stability, it is possible to use structure-properties relationship approach for protein stability modelling. We used a recently reported 2D graph representation of protein sequence for calculating 2D protein descriptors for training a SVM classifier of mutant stability. In this sense, novel *AA2DC* descriptors were obtained from the protein 2D graphs. This approach yielded an adequate classification model for the conformational stability of protein mutants describing about 72% of correct classifications in crossvalidation test for all the dataset and stable and unstable mutants separately. To the best of our knowledge, this is the most robust and best balanced predictor ever reported for a large mutant dataset model exploiting only sequence information for stability classification. Finally, the classifier adequately recognized some disease-related unstable mutants of human prion and human transthyretin. The present work demonstrates the successful application of the *AA2DC* descriptors as well as 2D graph representation of protein sequence, in combination with SVMs, for modelling protein conformational stability.

Acknowledgements

Financial supports of this project by Cuban Ministerio de Ciencia, Tecnología y Medio Ambiente (CITMA) through a grant to M. Fernandez (Grant No. 20104102).

References

[1] J. Saven. Combinatorial protein design. *Curr. Opin. Struct. Biol.*, **12**, 453 (2002).

[2] J. Mendes, R. Guerois, L. Serrano. Energy estimation in protein design. *Curr. Opin. Struct. Biol.*, **12**, 441 (2002).

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

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

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

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

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

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

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

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

[11] E. Lacroix, A.R. Viguera, 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**, 173 b) V. Munoz, 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**, 495 (1997). (1998).

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

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

[14] M.M. Gromiha, M. Oobatake, H. Kono, H. Uedaira, A. Sarai. Role of structural and sequence information in the prediction of protein stability changes: comparison between buried and partially buried mutations. *Protein Eng.*, **12**, 549 (1999).

[15] M.M. Gromiha, M. Oobatake, H. Kono, H. Uedaira, A. Sarai. Importance of surrounding residues for protein stability of partially buried mutations. *J. Biomol. Struct. Dyn.*, **18**, 1 (2000).

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

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

[18] E. Capriotti, P. Fariselli, R. Casadio. A neural-network-based method for predicting protein stability changes upon single

mutations. *Bioinformatics*, **20**, 63 b) E. Capriotti, P. Fariselli, R. Calabrese, R. Casadio. Prediction of protein stability changes from sequences using support vector machines. *Bioinformatics*, **21**, 54 (2005). c) E. Capriotti, P. Fariselli, R. Casadio. I-Mutant 2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucl. Acids Res.*, **33**, 306 (2005). (2004).

[19] R. Ramos de Armas, H. González-Díaz, R. Molina, E. Uriarte. Markovian backbone negentropies: molecular descriptors for protein research. I. Predicting protein stability in arc repressor mutants. *Proteins*, **56**, 715 (2004).

[20] Y. Marrero-Ponce, R. Medina-Marrero, J.A. Castillo-Garit, V. Romero-Zaldivar, F. Torrens, 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**, 3003 (2005).

[21] J. Caballero, L. Fernández, J.I. Abreu, 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**, 1255 (2006).

[22] L. Fernández, J. Caballero, J.I. Abreu, M. Fernández. Amino acid sequence autocorrelation vectors and Bayesian-regularized genetic neural networks for modelling protein conformational stability: gene V protein mutants. *Proteins*, **67**, 834 (2006).

[23] H. González-Díaz, R. Molina, E. Uriarte. Recognition of stable protein mutants with 3D stochastic average electrostatic potentials. *FEBS Lett.*, **579**, 4297 (2005).

[24] M. Randić, D. Butina. Novel 2-D graphical representation of proteins. *Chem. Phys. Lett.*, **419**, 528 (2006).

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

[26] G. Agüero-Chapin, H. González-Díaz, R. Molina, J. Varona-Santos, E. Uriarte, Y. González-Díaz. Novel 2D maps and coupling numbers for protein sequences. The first QSAR study of polygalacturonases; isolation and prediction of a novel sequence from *Psidium guajava* L. *FEBS Lett.*, **580**, 723 (2006).

[27] M. Randić, G. Krilov. Characterization of 3-D sequences of proteins. *Chem. Phys. Lett.*, **272**, 115 b) G. Krilov and M. Randić Quantitative characterization of protein structure: application to a novel a/b fold. *New J. Chem.*, **28**, 1608 (2004). (1997).

[28] F. Bai1, T. Wang. On graphical and numerical representation of protein sequences. *J. Biomol. Struct. Dyn.*, **23**, 537 (2006).

[29] J. Caballero, L. Fernández, M. Garriga, J.I. Abreu, S. Collina, 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*, (2006) doi: 10.1016/j.jmgm.2006.11.002..

[30] a) H.I. Jeffrey. Chaos game representation of gene structure. *Nucl. Acid Res.*, **18**, 2163 (1990). b) A. Fiser, G.E. Tusnády, I. Simon. Chaos game representation of protein structures. *J. Mol. Graph.*, **12**, 302 (1994).

[31] MATLAB 7.0. program, available from The Mathworks Inc., Natick, MA. http://www.mathworks.com,

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

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

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

[35] A.C. Apetri, K. Surewicz, W.K. Surewicz. The effect of disease-associated mutations on the folding pathway of human prion protein. *J. Biol. Chem.*, **279**, 18008 (2004).

[36] S. Liemann, R. Glockshuber. Influence of amino acid substitutions related to inherited human prion diseases on the thermodynamic stability of the cellular prion protein. *Biochemistry*, **38**, 3258 (1999).

[37] L. Calzolai, D.A. Lysek, P. Güntert, C. Schroetter, R. Riek, R. Zahn, K. Wüthrich. NMR structures of three single-residue variants of the human prion protein. *Proc. Natl Acad. Sci. USA*, **97**, 8340 (2000).

[38] V.L. Shnyrova, E. Villar, G.G. Zhadana, J.M. Sanchez-Ruiz, A. Quintas, M.J.M. Saraiva, R.M.M. Brito. Comparative calorimetric study of non-amyloidogenic and amyloidogenic variants of the homotetrameric protein transthyretin. *Biophys. Chem.*, **88**, 61 (2000).