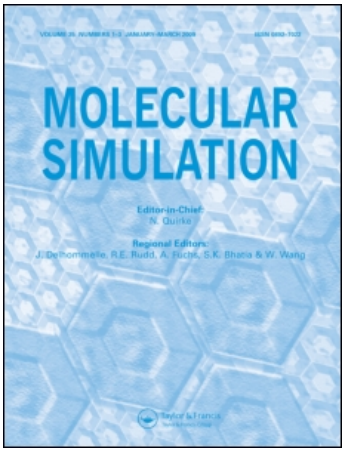

Molecular Simulation
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713644482

Comparative modeling of the conformational stability of chymotrypsin inhibitor 2 protein mutants using amino acid sequence autocorrelation (AASA) and amino acid 3D autocorrelation (AA3DA) vectors and ensembles of Bayesian-regularized genetic neural networks

Michael Fernández[a]; José Ignacio Abreu[ab]; Julio Caballero[c]; Miguel Garriga[d]; Leyden Fernández[a]
[a] Molecular Modeling Group, Center for Biotechnological Studies, Faculty of Agronomy, University of Matanzas, Matanzas, Cuba [b] Artificial Intelligence Laboratory, Faculty of Informatics, University of Matanzas, Matanzas, Cuba [c] Plant Biotechnology Group, Faculty of Agronomy, Center for Biotechnological Studies, University of Matanzas, Matanzas, Cuba [d] Centro de Bioinformática y Simulación Molecular, Universidad de Talca, 2 Norte, Casilla 721, Talca, Chile

To cite this Article Fernández, Michael , Abreu, José Ignacio , Caballero, Julio , Garriga, Miguel and Fernández, Leyden(2007) 'Comparative modeling of the conformational stability of chymotrypsin inhibitor 2 protein mutants using amino acid sequence autocorrelation (AASA) and amino acid 3D autocorrelation (AA3DA) vectors and ensembles of Bayesian-regularized genetic neural networks', Molecular Simulation, 33: 13, 1045 — 1056
To link to this Article: DOI: 10.1080/08927020701564479
URL: http://dx.doi.org/10.1080/08927020701564479

Molecular Simulation, Vol. 33, No. 13, November 2007, 1045–1056

Taylor & Francis
Taylor & Francis Group

Comparative modeling of the conformational stability of chymotrypsin inhibitor 2 protein mutants using amino acid sequence autocorrelation (AASA) and amino acid 3D autocorrelation (AA3DA) vectors and ensembles of Bayesian-regularized genetic neural networks

MICHAEL FERNÁNDEZ†, JOSÉ IGNACIO ABREU†‡, JULIO CABALLERO†¶, MIGUEL GARRIGA†§ and LEYDEN FERNÁNDEZ†*

†Molecular Modeling Group, Center for Biotechnological Studies, Faculty of Agronomy, University of Matanzas, 44740 Matanzas, Cuba
‡Artificial Intelligence Laboratory, Faculty of Informatics, University of Matanzas, 44740 Matanzas, Cuba
¶Centro de Bioinformática y Simulación Molecular, Universidad de Talca, 2 Norte 685, Casilla 721, Talca, Chile
§Plant Biotechnology Group, Faculty of Agronomy, Center for Biotechnological Studies, University of Matanzas, C.P. 44740 Matanzas, Cuba

(Received May 2007; in final form July 2007)

Predicting protein stability changes upon point mutation is important for understanding protein structure and designing new proteins. Autocorrelation vector formalism was extended to amino acid sequences and 3D conformations for encoding protein structural information with modeling purpose. Protein autocorrelation vectors were weighted by 48 amino acid/residue properties selected from the AAindex database. Ensembles of Bayesian-regularized genetic neural networks (BRGNNs) trained with amino acid sequence autocorrelation (AASA) vectors and amino acid 3D autocorrelation (AA3DA) vectors yielded predictive models of the change of unfolding Gibbs free energy change ($\Delta\Delta G$) of chymotrypsin Inhibitor 2 protein mutants. The ensemble predictor described about 58 and 72% of the data variances in test sets for AASA and AA3DA models, respectively. Optimum sequence and 3D-based ensembles exhibit high effects on relevant structural (volume, solvent-accessible surface area), physico-chemical (hydrophilicity/hydrophobicity-related) and thermodynamic (hydration parameters) properties.

Keywords: Point mutations; Artificial neural networks; Bayesian regularization; Protein stability

1. Introduction

Protein function is extremely dependent on that the nascent amino acid chain correctly folds into the biologically active, three-dimensional structure of the native state. In addition, such knowledge is critical for numerous biomedical applications, including but not limited to the preparation of stable protein-based therapeutics and the treatment of pathologies related to mutated, unstable proteins [1–4].

Predicting protein stability changes upon point mutations is important for understanding protein structure and designing new proteins [5–8]. By the way, the stability of protein is ruled by a number of interactions, such as, hydrophobic and electrostatic interactions and hydrogen bonding. Rational design of mutants in order to increase the stability of a protein essentially aims at either improving stabilizing interactions or reducing potentially destabilizing factors. Protein stability is quantitatively described by the standard Gibbs energy change (ΔG°). The energetic of mutants has been studied extensively both through theoretical and experimental approaches. The methods for predicting protein stability changes resulting from single amino acid mutations can be classified into four general categories: statistical potential approach [9–11], physical potential approach [12], and empirical potential approach [13,14].

Besides, there are other methods to predict stability, for instance those are based on correlations of free energy change with structural sequence information and amino acid properties. In this sense, Gromiha *et al.* [15,16] had

*Corresponding author. Tel.:+53-45-26-1251. Fax: +53-45-25-3101. Email: leyden.fernandez@umcc.cu; leydenfernandez@gmail.com

Molecular Simulation
ISSN 0892-7022 print/ISSN 1029-0435 online © 2007 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/08927020701564479

analyzed the correlation between the stability changes caused by buried and partially buried mutations and changes in physicochemical, energetic and conformational properties. They showed that changes in stability of buried mutation highly correlated with hydrophobicity but partially buried mutation stability also strongly correlated with hydrogen bonds and other polar interactions. In another work, they reported the importance of surrounding residues for protein stability of partially buried mutations finding that highest segment length effects for helical, strand and coil mutations are, respectively, 0, 9 and 4 residues on both sides of the mutant residues [17].

In other work, Takano and Yutani [18] derived a structure-based stability scale from mutation data of T4 lysozyme and human lysozyme. The structure-based scale contains many stabilizing factors such as accessible surface area, the number of hydrogen bonds, cavity volume and number of water molecules. More recently, Zhou and Zhou [19] found a scale of hydrophobic residues has an excellent correlation with the octanol-to-water transfer free energy corrected. 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 another work they incorporated the environmental effect of mutation sites. While some of these methods mentioned before are based on 3D information, Levin and Satir [20] used amino acid sequence information. In this context, they successfully evaluated the functional significance of mutations on hemoglobin using amino acid similarity matrixes. Recently, Frenz [21] reported an artificial neural network (ANN)-based model for predicting the stability of Staphylococcal Nuclease mutants using amino acid similarity scores as network inputs.

In this sense, machine learning algorithms have been also applied to the protein stability prediction paradigm, outstanding reports of Capriotti *et al.* [22] describe the implementation of neural network and support vector machine predictors 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 mutants. As network and vector machine inputs they used a combination of experimental condition data (pH and temperature), specific mutated residue and environmental residue information.

The present study considers amino acids sequences and protein 3D structure to model conformational stability of chymotrypsin inhibitor 2 mutants with ANNs. ANNs have been among the most successful pattern classification tools applied to problems in the field of biochemical sciences. ANNs usually overcome methods limited to linear regression models like MRA or Partial Least Square [23–31]. Contrary to these methods, ANNs can be used to model complex nonlinear relationships. Since biological phenomena are complex by nature, this ability has promoted the employment of ANNs in biological patron recognition problems. In this connection, we recently extend the concept of structural autocorrelation vectors in molecules to protein sequences and conformational stability of human lysozyme [29] and gene V protein [32] mutants, were successfully modelled using ensembles of Bayesian-regularized genetics neural networks (BRGNN). Optimum BRGNN predictive models of conformational stability were built with reduced subset of variables. In order to provide robust models, we employed data-diverse ensembles of BRGNN for calculating the conformational stability. In this way, the structure of the chymotrypsin inhibitor 2 is interesting because its small size and the large number of mutants available have made it a useful model for determining effects of amino acid substitutions on protein stability and function. We attempted to predict protein conformational stability by extending the concept of structural autocorrelation vectors [33–38] in molecules to protein primary and tertiary structures. Protein structure was encoded by means of amino acid sequence autocorrelation (AASA) vectors and AA3DA vectors both weighted by 48 physicochemical, energetic and conformational amino acid/residues properties extracted from the AAindex amino acid database [39].

2. Materials and methods

2.1 Protein autocorrelation vectors

2.1.1 Amino acid sequence autocorrelation (AASA) vectors approach. Structure-property/activity studies strategy to encoding structural information must, in some way, either explicitly or implicitly, account for hydrophobic, electrostatic, van der Waals and hydrogen bond interactions. Furthermore, usually data sets include structures of different size with different numbers of elements, so the structural encoding approaches must allow comparing them [38].

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, and so they are canonical. And thirdly, the length of the correlation vector is independent of the size of the molecule [38].

For the autocorrelation vectors in molecules, H-depleted molecular structure is represented as a graph and physicochemical 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. Two-dimensional spatial autocorrelations [33–36] has been successfully used in the last decades for modelling biological activities [35,36] and pharmaceutical research [37,38]. In recent works, our group has obtained outstanding results when such chemical code was used in combination with ANN approach in biological QSAR studies [25,30,31]. 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 protein conformational stability prediction.

Broto–Moreau's autocorrelation coefficient [35] is defined as follow:

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

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 δ(l, d_{ij}) is a Dirac-delta function defined as

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

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 amino acid residues. Autocorrelation approach in protein stability studies mainly differs from the Gromiha *et al.* [16] method in considering the whole amino acid sequence of the protein for calculation of the descriptors instead local sequence segments over the mutated point. In this way, the calculated autocorrelation vectors encode structural information concerning whole protein. Particularly, AASA vectors of lag l are calculated as follows:

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

where AASAlp_k is the AASA at spatial lag l weighted by the p_k property; L is the number of nonzero values 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 δ(l, d_{ij}) is a Dirac-delta function.

Autocorrelation measures the level of interdependence between properties and the nature and strength of that interdependence. It may be classified as either positive or negative. In a positive case all similar values appear together, while a negative spatial autocorrelation has dissimilar values appearing in close association [33,34]. 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 they were used 48 physicochemical, energetic and conformational amino acid/residues properties (table 1SM) selected by Gromiha *et al.* [16] from the AAindex data base [39] in a previous study concerning relationships between amino acid/residues properties and stability for a large set of proteins. In our work, spatial lag, l, was ranging from 1 to 15 with the aim of accessing to long-range interactions in the sequence due to tertiary structure arrangements. Computational code for AASA vector calculation was written in Matlab environment [40]. A data matrix of 720 AASA vectors, 48 properties × 15 different lags, were generated with the autocorrelation vectors calculated for each chymotrypsin inhibitor 2 mutant. Descriptors that stayed constant or almost constant were eliminated and pairs of variables with a square correlation coefficient (R^2) greater than 0.8 were classified as intercorrelated, and only one of these was included for building the model. Finally, 232 AASA descriptors were obtained. Afterwards, optimum predictive models were built with reduced subsets of variables by means of BRGNN algorithm.

2.1.2 Amino acid 3D autocorrelation (AA3DA) vectors approach. 2D Broto–Moreau's autocorrelation coefficient [35] was extended to 3D dimension [37]. 3D autocorrelation can be applied to protein 3D structure. Autocorrelations of amino acids/residues properties at 3D Euclidean distances over the tridimensional structure are computed according to equation (4).

$$\text{AA3DA}lp_k = \frac{1}{L}\sum_i \delta_{ij} p_{ki} p_{kj} \tag{4}$$

where AA3DAlp_k is the AA3DA at Euclidean lag l weighted by the p_i property; L is the number of nonzero values 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 δ (l, d_{ij}) is a Dirac-delta function defined as:

$$\delta(l, s, d_{ij}) = \left\{ \begin{array}{ll} 1 & \text{if } l - \frac{s}{2} < d_{ij} \leq l + \frac{s}{2}, \\ 0 & \text{otherwise} \end{array} \right\} \tag{5}$$

where the d_{ij} is the Euclidean distance between amino acid residues i and j in the Cα carbons 3D graph, l and s are the 3D Euclidean distance lag and the 3D Euclidean distance step, respectively.

As weights for amino acid residues they were also used the 48 physicochemical, energetic and conformational amino acid/residues properties (table 1SM) selected by Gromiha *et al.* [16] from the AAindex database [39]. AA3DA vectors were computed at distances ranging from 3 to 30 Å with a distance step of 3 Å using the protein descriptor calculations module that was written in Matlab environment [40].

A data matrix of 480 AA3DA vectors, 48 properties × 10 different Euclidean lags, were generated with the 3D autocorrelation vectors calculated for each chymotrypsin inhibitor 2 mutant. Descriptors that stayed constant or almost constant were eliminated and pairs of variables with a square correlation coefficient greater than 0.8 were classified as intercorrelated, and only one of these was included for building the model. Finally, 363 AA3DA descriptors were obtained. Afterwards, optimum classification and regression models were built with reduced subsets of variables by means GA optimization.

2.2 Bayesian-regularized genetic neural networks (BRGNN) approach

In the context of ANN-based modeling of biological interactions we introduced (BRGNNs) as a robust nonlinear modeling technique that combines GA and Bayesian regularization for neural network input selection and supervised network training, respectively. This approach attempts to solve the main weaknesses of neural network modeling: the selection of optimum input variables and the adjustment of network weights and biases to optimum values for yielding regularized neural network predictors [41–43].

By combining the concepts of BRANN and GA algorithms, BRGNNs are implemented in such a way that BRANN inputs are selected inside a GA framework. BRGNN approach is a version of the So and Karplus report [41] incorporating Bayesian regularization that has been successfully introduced by our group for modeling the inhibitory activity of several therapeutic target enzymes [23,27–31]. BRGNN was programmed within Matlab environment [40] using genetic algorithm and neural networks toolboxes. BRGNN technique leads to neural networks trained with optimum inputs selected from (1) the whole AASA vector data matrix (see Section 2.1.1) and (2) the whole AA3DA vector data matrix (see Section 2.1.2).

Bayesian networks are optimal devices for solving learning problems. They diminish the inherent complexity of ANNs, being governed by Occam's Razor, when complex models are automatically self-penalizing under Bayes's rule. The Bayesian approach to ANN modeling considers all possible values of network parameters weighted by the probability of each set of weights. The BRANN method was designed by Mackay [42,43] for overcoming the deficiencies of ANNs. Only a brief summary will be provided here. The Bayesian approach yields a posterior distribution of network parameters $P(w|D, H)$ from a prior probability distribution $P(w|H)$ according to updates provided by the training set D using the BRANN model H. Predictions are expressed in terms of expectations with respect to this posterior distribution. Bayesian methods can simultaneously optimize the regularization constants in ANNs, a process that is very laborious using crossvalidation. Instead of trying to find the global minimum, the Bayesian approach finds the (locally) most probable parameters. Bayesian approach produces predictors that are robust and well matched to the data. These properties become BRANNs in accurate predictors for QSAR analysis [44,45]. They give models, which are relatively independent of ANN architecture, above a minimum architecture, since the Bayesian regularization method estimates the number of effective parameters. The concerns about overfitting and overtraining are also eliminated by this method so that the production of a definitive and reproducible model is attained. The joining of BRANN and GA feature selection (BRGNN) increases the possibilities of BRANNs for modeling as we indicated in previous works [24,26–31]. This method is relatively fast and considers the whole data set in the training process. For other hybrids of ANN and GA the use of the MSE as fitness function could lead to undesirable well fitted but poor generalized networks as algorithm solutions.

Fully connected, three-layer BRANNs with back-propagation training were implemented in MATLAB environment [40]. In these nets, the transfer functions of input and output layers were linear and the hidden layer had neurons with a hyperbolic tangent transfer function. Inputs and targets took the values from independent variables selected by the GA and $\Delta\Delta G$ values respectively; both were normalized prior to network training. BRANN training was carried out according to the Levenberg-Marquardt optimization [46]. The initial value for μ was 0.005 with decrease and increase factors of 0.1 and 10, respectively. The training was stopped when μ became larger than 10^{10}.

The GA implemented in this paper keeps the same characteristics of the previously reported in earlier work [24,26–31]. Differently to other GA-based approach, the objective of our algorithm is not to obtain a sole optimum model but a reduced population of well fitted models, with MSE lower a threshold MSE value, which the Bayesian regularization guaranties to posses good generalization abilities. This is due to we used MSE of data training fitting instead crossvalidation or test set MSE values as cost function and therefore the optimum model can not be directly derived from the best fitted model yielded by the genetic search. However, from crossvalidation experiments over the subpopulation of well fitted models it can derive an optimum generalizable network with the highest predictive power. This process also assures to avoid chance correlations. This approach have shown to be highly efficient in comparison with crossvalidation-based GA approach since only optimum models, according to the Bayesian regularization, are crossvalidated at the end of the routine and not all the model generated throughout all the search process [31].

2.3 Artificial neural network ensembles

An artificial neural network ensemble (NNE) is a learning paradigm where many ANNs are jointly used to solve a problem. On the basis of this judgment, a collection of a finite number of neural networks is trained for the same task and the outputs can be combined to form one unified prediction. As a result, the generalization ability of the neural network system can be significantly improved [47].

An effective NNE should consist of a set of ANNs that not only are highly correct but make their errors on different parts of the input space as well. So, the combination of the output of several classifiers is only useful if they disagree on some inputs. Krogh and Vedelsby [48] later proved that the ensemble error can be divided into a term measuring the average generalization

Table 1. Experimental and calculated change of unfolding Gibbs free energy change (ΔΔG) at 25°C, pH = 6.3 in Gdn• or ·HCl for chymotrypsin inhibitor 2 wild-type and mutants according to 30 members neural network ensemble of optimum model AASA-BRGNN 3 and 20 members neural network ensemble of optimum model AA3DA-BRGNN 3.

	$\Delta\Delta G^{\dagger}$ (kcal/mol)				$\Delta\Delta G^{\dagger}$ (kcal/mol)		
Mutant	*Exp.*	*Cal.*$_{test}$ ‡	*Cal.*$_{test}$ ¶	*Mutant*	*Exp.*	*Cal.*$_{test}$ ‡	*Cal.*$_{test}$ ¶
Wild	0.00	−0.18	−0.33	L51A/V57/AF69L	−3.48	−2.68	−4.16
A35G	−1.09	−1.07	−1.68	L51I	−0.26	0.06	−0.86
A77G	−1.88	−0.88	−0.98	L51V	−0.50	−0.65	−1.31
D42A	−0.96	−1.37	−0.66	L51V/F69L	−2.42	−2.36	−2.61
D64A	−0.80	−0.72	−0.20	L51V/V57A	−1.85	−2.24	−1.38
D71A	−3.41	−1.61	−1.98	L51V/V57A/F69L	−2.72	−3.28	−3.08
E26A	−0.47	−0.99	−0.22	L68A	−3.82	−3.46	−3.66
E26Q	−0.62	−0.55	−0.52	N75A	−0.83	−1.11	−1.46
E33A/E34A	−0.76	−0.96	0.39	N75D	−1.21	−2.07	−0.38
E33D	−0.52	−1.00	−0.47	P25A	−1.57	−2.23	−1.61
E33N	−0.70	−0.33	−0.41	P25A/A35G	−2.65	−2.91	−2.84
E33Q	−0.29	−0.33	−0.22	P44A	−1.76	−2.47	−1.06
E34D	−0.74	−1.11	−0.61	P52A	−0.17	−1.11	−1.07
E34N	−1.07	−0.27	−0.33	P80A	−3.34	−0.94	−4.34
E34Q	−0.47	−0.43	−0.56	Q41A	−0.02	−0.65	−0.21
E45A	−0.32	−0.29	−0.35	Q41G	−0.60	−1.71	−0.81
E60A	−0.68	−0.11	−1.17	R62A	−0.58	−3.06	−0.34
F69A	−3.84	−2.91	−3.27	R62A/D64A	−1.22	−1.25	−0.74
F69L	−2.11	−1.42	−1.37	S31A	−0.89	−0.38	−0.93
F69V	−2.39	−2.47	−2.67	S31A/E33A/E34A	−1.67	−2.55	−1.90
I39V	−1.27	−0.69	−0.67	S31G	−0.80	−0.93	−0.48
I48A	−3.84	−2.16	−2.37	S31G/E33A/E34A	−1.63	−1.67	−2.04
I48A/I76V	−4.05	−3.82	−4.02	T22A	−0.85	−1.28	−1.15
I48V	−1.09	−0.78	−0.54	T22G	−1.16	−2.07	−1.33
I49A	−2.12	−2.71	−2.27	T22V	−0.32	−0.36	−0.55
I49G	−3.52	−3.21	−2.53	T55A	0.23	−0.72	−0.30
I49T	−1.34	−1.35	−1.26	T55S	−0.02	−0.81	−0.57
I49V	0.08	−0.80	−0.76	T55V	−0.76	−0.69	−1.01
I56A	−0.03	−3.10	−0.48	T58A	−0.69	−0.80	−0.27
I76A	−4.25	−2.71	−3.30	T58A/E60A	−0.87	−1.29	−0.34
I76V	0.21	−1.18	−1.71	T58D	0.04	−0.03	−0.35
K21A	−0.55	−0.49	−0.67	T58D/E60A	−0.25	−0.08	−0.57
K21A/E26A	−1.10	−0.69	−0.75	V38A	−0.46	−1.24	−2.10
K21M	−0.67	−0.08	−0.29	V53A	−0.64	−1.47	−1.49
K30A	0.42	−0.67	−0.25	V53G	−2.43	−1.62	−2.36
K36A	−0.49	−1.00	−0.39	V53T	−1.03	−0.80	−1.06
K36G	−2.32	−1.77	−2.60	V57A	−1.47	−1.54	−0.16
K37A	0.21	−0.62	−0.12	V57A/F69L	−2.58	−3.10	−2.85
K37G	−0.99	−1.52	−1.23	V57A/V79A	−4.37	−3.84	−2.45
K43A	−0.65	−1.19	−1.05	V66A	−4.88	−3.64	−3.95
K43G	−3.19	−1.93	−1.92	V70A	−1.95	−2.42	−2.00
K72N	0.00	1.19	−0.47	V79A	−1.51	−1.83	−1.59
L27A	−2.64	−2.80	−3.36	V79G	−3.24	−2.58	−2.26
L40A	−1.33	−1.35	−0.86	V79T	−0.38	−1.01	−0.62
L40G	−1.38	−1.73	−2.08	V82A	−1.45	−1.36	−2.64
L51A	−2.37	−2.92	−3.35	V82G	−3.50	−1.62	−2.79
L51A/F69L	−3.42	−3.21	−3.43	V82T	−1.15	−0.55	−0.64
L51A/V57A	−3.16	−3.60	−3.49				

† ΔΔG negative and positive values mean destabilizing and stabilizing mutations respectively. ‡ Calculated as average over test sets using a 30 members ensemble on the model AASA-BRGNN 3. ¶ Calculated as average over test sets using a 30 members ensemble on the model AA3DA-BRGNN 3.

error of each individual network and a term called diversity that measures the disagreement among the networks. Formally, they defined the diversity term d_i of network i on input j to be

$$d_i(j) = (o_i(j) - \bar{o}(j))^2 \tag{6}$$

where $oi(j)$ and $oj(j)$ are the ith classifier and the ensemble predictions, respectively. In other words, it is simply the variance of the ensemble around the mean. The quadratic errors of network i and those of the ensemble are, respectively,

$$\varepsilon(j) = (o_i(j) - f(j))^2 \tag{7}$$

and

$$e(j) = (\bar{o}(j) - f(j))^2 \tag{8}$$

where $f(j)$ is the target value for input j. If we define $\bar{E}$, E_i and D_i to be the averages, over the input distribution of $e(j)$, $\epsilon(j)$ and $d(j)$, respectively, then the ensemble's generalization error can be shown to consist of two distinct

portions:

$$\bar{E} = E - D \tag{9}$$

where $E = \sum_i w_i E_i$ is the weighted average of the individual networks' generalization error and $D = \sum_i w_i D_i$ is the weighted average of the diversity among these networks. What equation (9) shows, then, is that an ideal ensemble consists of highly correct ANNs that disagree as much as possible. In this way, the mean-square error (MSE $= \bar{E}$) of the ensemble estimator is guaranteed to be less than or equal to the averaged mean-square error of the component estimators.

2.4 Model's validation

In this work, we validated our regression model using a reasonable method employed by our group that consists into a robust validation process by means of NNE [24,29,31,32]. Recently Baumann [49] demonstrated that ensemble averaging significantly improve prediction accuracy by averaging the predictions of several models that are obtained in parallel with bootstrapped training sets and provide a more realistic meaning of the predictive capacity of any regression model.

For generating the predictors that will be averaged in the NNE, we partitioned the whole data into several training and test sets.

The assembled predictors aggregate their outputs to produce a single prediction. In this way, instead of predicting a sole randomly selected external set; we predict the result of averaging several ones. In this way, each mutant was predicted several times forming training and test sets and an average of both values were reported. The predictive power was measured accounting R^2 and root MSE (RMSE) mean values of the averaged test sets of BRGNN ensembles having an optimum number of members.

2.5 Chymotrypsin inhibitor 2 mutant dataset

Chymotrypsin inhibitor 2 was used in our study as a model protein to test the AASA and AA3DA approaches. Chymotrypsin inhibitor 2 (83 residues, PDB file: 2CI2) is a good model for protein stability studies because it is available a wide thermodynamic data of mutants in very homogeneous conditions. Chymotrypsin inhibitor 2 data (wild-type and 94 mutants) was collected from Protherm data base [50]. Table 1 shows change in unfolding Gibbs free energy change ($\Delta\Delta G$) at 25°C and pH = 6.3 in the presence of Gdn HCl for wild-type and mutants in comparison to wild-type enzyme.

Three-dimensional structure of chymotrypsin inhibitor 2 wild-type and mutants were taken from PDB files from Protein Data Bank [51] website when available, such as mutants I76V (PDB file: 1COA), S31A/E33A/E34A (PDB file: 1YPA), S31G/E33A/E34A (PDB file: 1YPB), E33A/E34A (PDB file: 1YPC). Mutants lacking 3D structure information were built by single residue substitution on the wild-type chymotrypsin inhibitor 2 PDB file 2CI2. Energy minimization steps were performed using CHARMM [52] computer software and the EEF1 energy function, [9,53] which is based on the polar hydrogen CHARMM energy (Charmm 19 parameter set) [54] and includes an implicit solvation term. In all cases, missing hydrogen coordinates were built with the HBUILD algorithm [55], followed by 300 steps of energy minimization with the ABNR method [52].

3. Results and discussion

Conformational stability of chymotrypsin inhibitor 2 was modelled by using two structure-encoding approaches. A sequence-based model with AASA vectors which are calculated over the protein primary structure and another 3D-based model with AA3DA vectors which are calculated over the protein tridimensional structure. Those vectors were computed weighted by a variety of physicochemical, energetic, and conformational properties that appear in (table 1SM). In this way, the structural information that can be relevant for modeling the conformational stability of chymotrypsin inhibitor 2 mutants was gathered in two pools of variables, containing sequential and three-dimensional information separately. Both models were developed using BRGNN. Inside the BRGNN framework, GA searches for the best fitted BRANN was achieved from one generation to another in order to minimize the MSE of the networks (fitness or cost function). By employing this approach instead a more complicated and time consuming cross-validation based fitness function, we gain in CPU time and simplicity of the routine on the two cases. Furthermore, we can devote the whole data set completely to train the networks. However, the use of the MSE fitness function could lead to undesirable well fitted but poor generalized networks as algorithm solutions. In this connection, we tried to avoid such results by two aspects: (1) keeping network architectures as simplest as possible (only three hidden nodes) inside the GA framework, and (2) implementing Bayesian regulation in the network training function (Section 2). The nonlinear subspaces in the data set were explored varying the number of network inputs from 3 to 10. As result of the algorithm a small population of well fitted models is obtained. Afterwards those models were tested in crossvalidation experiments in order to avoid chance correlations and the model with the best crossvalidation statistics was selected as optimum.

3.1 AASA vectors behavior

In table 2 are represented statistical parameters for the optimum AASA-based BRGNN (AASA-BRGNN). Optimum BRGNNs appear with eight inputs but varying the number of hidden nodes from 2 to 9. By inspection of table 2 it can be observed that Bayesian regularization yielded quite stable and reliable sequence-based networks. The

Table 2. Statistics of the optimum BRGNN predictors for the conformational stability of wild-type and mutant chymotrypsin inhibitor 2 proteins for *AASA* vectors. Optimum neural network predictor appears in bold letter.

	AASA vectors: AASA7ΔCp$_h$, AASA7Ht, AASA7N$_1$, AASA5V, AASA3H$_{nc}$, AASA1s, AASA12P, AASA8P						
AASA-BRGNN model	*hidd. nod.*	*num. par.*	*opt. par.*	R^2	S	R^2_{cv}	S_{cv}
1	2	21	17	0.711	0.677	0.548	0.861
2	3	31	25	0.806	0.554	0.593	0.822
3	**4**	**41**	**30**	**0.835**	**0.512**	**0.625**	**0.787**
4	5	51	34	0.843	0.503	0.611	0.806
5	6	61	37	0.853	0.483	0.610	0.800
6	7	71	36	0.844	0.499	0.612	0.800
7	8	81	36	0.844	0.499	0.606	0.804
8	9	91	36	0.844	0.499	0.609	0.805

hidd. nod. represents the number of hidden nodes, *num. par.* represents the number of neural network parameters, *opt. par.* represents the optimum number of neural network parameters yielded by the Bayesian regularization, R^2 and R^2_{cv} are square correlation coefficients of data set fitting and LOO crossvalidation, respectively, S and S_{cv} are standard deviations of data set fitting and LOO crossvalidation respectively.

behavior of the networks was asymptotic with respect to the number of hidden nodes with maximum number of optimum parameters about 35.5. However, considering the crossvalidation statistics, the optimum predictor was AASA-BRGNN 3 with 4 hidden nodes and 30 optimum parameters having highest values of the square correlation coefficients for data fitting (R^2) and leave-one-out (LOO) crossvalidation (R^2_{cv}) about 0.835 and 0.625, respectively.

AASA vectors in the optimum *AASA*-BRGNN mean: AASA7ΔCp$_h$ is the AASA of lag 7 weighted by unfolding hydration heat capacity change; AASA7H$_t$, is the AASA of lag 7 weighted by thermodynamic transfer hydrophobicity; AASA7N$_1$, is the AASA of lag 7 weighted by average long-range contacts; AASA5V, is the AASA of lag 5 weighted by volume (number of nonhydrogen side-chain atoms; AASA3H$_{nc}$ is the AASA of lag 3 weighted by normalized consensus hydrophobicity; AASA1s, is the AASA of lag 1 weighted by shape (position of branch point in a side-chain); AASA12P and AASA8P are the AASA of lag 12 and 8 weighted by polarity. In addition, there is not significant intercorrelation among selected descriptors so different information is brought to the model by each AASA descriptor. Interestingly, relevant amino acid/residue properties appear weighting the selected optimum AASA vectors: two thermodynamical (ΔCp$_h$, H_t), three structural (N_1, V, s) and three (H_t, H_{nc}, P) hydrofobicity\hydrofilicity related properties.

In order to build a robust model we used ensembles of BRGNNs instead a single network to calculate the ΔΔG values for chymotrypsin inhibitor 2, wild-type and mutant proteins. This approach applied by us [24,29,31,32] consists in training several BRGNNs with different randomly partitioned training sets of 71 proteins (75% of the data) and predicting the activity of the rest 24 proteins (25% of the data) in test sets. In this regard, the outputs of the trained networks were combined to form one unified prediction. In this sense, we reported in table 1 two calculated ΔΔG values for each mutant: one average over test sets to AASA-BRGNN 3 model and another over the test sets to AA3DA-BRGNN 3 model.

Figure 1A shows the behaviors of E, D and MSE of NNEs vs. the number of networks in the ensemble for the AASA-BRGNN model. Those statistics are stabilized for ensembles with 30 and more members at values about 1.00, 0.31 and 0.70, respectively. The test set E term is a measure of the likeness among external predictions and experimental activities. Ambiguity (D) is a measure of the difference among the members of an ensemble. High values of ambiguity mean disagreement among the networks whereby no redundancies on the data were found. In our case, the Bayesian regularization increased

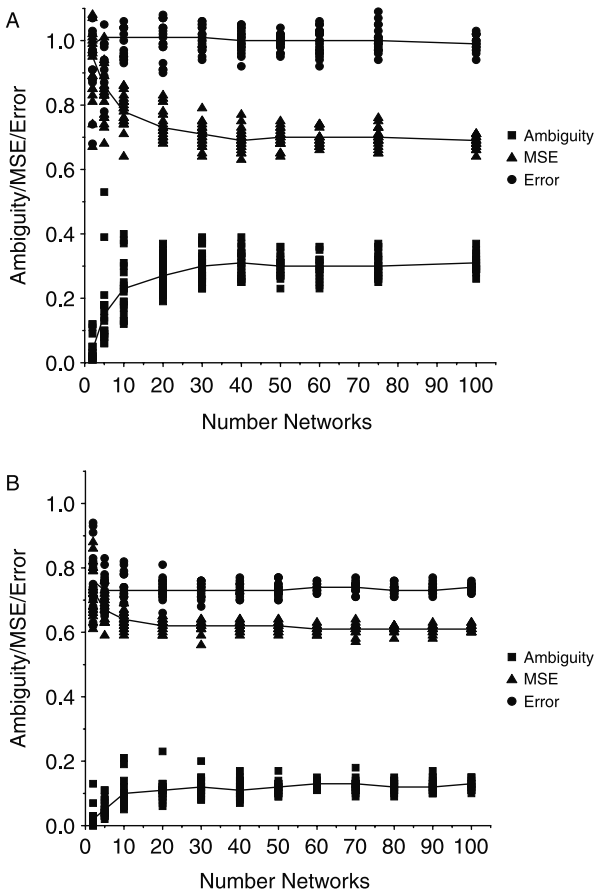

Figure 1. Plots of E, D, *RMSE* for ΔΔG average values vs. number of neural networks in each ensemble (a) for AASA vectors with 30 ensembles, (b) for AA3DA vectors with 20 ensembles.

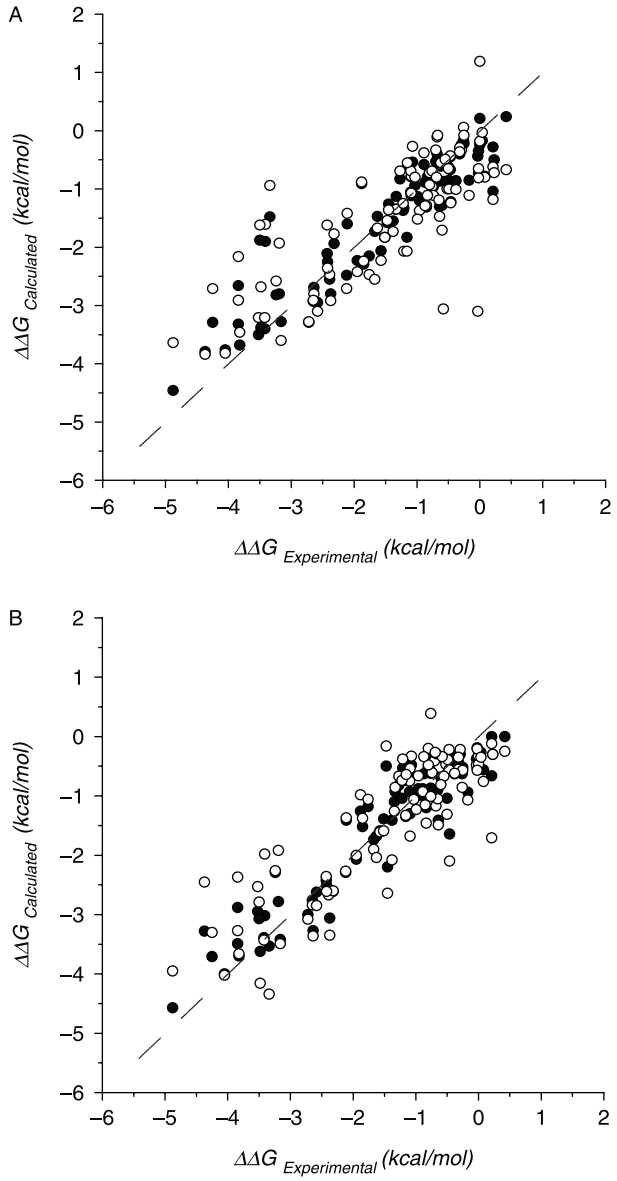

Figure 2. Plots of average calculated vs. experimental change of unfolding Gibbs free energy change (ΔΔG) of chymotrypsin inhibitor protein and mutants in training (•) and test (○) sets according to: (a) 30 member ensemble of the optimum sequence-based network AASA-BRGNN 3, (b) 30 member ensemble of the optimum 3D-based network AA3DA-BRGNN 3.

the stability of the predictors; therefore, no high values of ambiguity should be expected. However, MSE measures the difference between E term and D and it is noteworthy that with 30 members seem to be enough to get a precise statistic of the validation. Krogh and Vedelsby [48] reported that MSE value for NNEs should be smaller than the averaged MSE of the component predictors (equation (9)). In this sense, the MSE values in our optimum ensemble represent a decrease of 30% in comparison to the mean *E* value.

Figure 2A depicts plots of calculated vs. experimental unfolding ΔΔG values for each protein calculated as an average over training and test sets according to the ensemble predictor. The accuracy for data fitting was about 84% and 58% for proteins in training and test sets, respectively. AASA vectors approach well fit in a nonlinear way the ΔΔG by means a combination of sequence information and amino acid/residues properties. This model needs eight vectors to describe adequately the conformational stability pattern of chymotrypsin inhibitor 2 mutants. However, the predictive power about 0.6 is relatively low. This fact suggested that the interactions affecting conformational stability of chymotrypsin inhibitor 2 mutants are only partially assessed from a sequence framework.

3.2 AA3DA vectors behavior

Table 3 shows optimum AA3DA 3D structure-based BRGNN (AA3DA-BRGNN) vectors, with five inputs and hidden nodes varied from 2 to 6 and maximum number of optimum parameters about 30. Despite, AASA vectors yield acceptable prediction accuracy about 0.6; these networks overcome AASA-BRGNN model. The optimum predictor was AA3DA-BRGNN 3 with 4 hidden nodes, $R^2 = 0.851$ and $R^2_{cv} = 0.737$. The good behavior of this nonlinear model describing the conformational stability of the studied chymotrypsin inhibitor 2 mutants suggest that AA3DA vectors built a nonlinear vectorial space that well resembles chymotrypsin inhibitor 2 protein stability pattern. Furthermore, vectors using 3D structure information showed a better resembling of the stability of this protein with a crossvalidation accuracy > 0.7.

Table 3. Statistics of the optimum BRGNN predictors for the conformational stability of wild-type and mutant chymotrypsin inhibitor 2 proteins for *AA3DA* vectors. Optimum neural network predictor appears in bold letter.

	AA3DA vectors: AA3DA9V, AA3DA6R$_f$, AA3DA12N$_m$, AA3DA15ASA$_N$, AA3DA27ΔH$_h$						
AA3DA-BRGNN model	*hidd. nod.*	*num. par.*	*opt. par.*	R^2	S	R^2_{cv}	S_{cv}
1	2	15	13	0.788	0.579	0.672	0.729
2	3	22	18	0.833	0.514	0.710	0.692
3	**4**	**29**	**23**	**0.851**	**0**.486	**0**.737	**0**.648
4	5	36	25	0.852	0.485	0.680	0.733
5	6	43	30	0.877	0.440	0.686	0.721

hidd. nod. represents the number of hidden nodes, *num. par.* represents the number of neural network parameters, *opt. par.* represents the optimum number of neural network parameters yielded by the Bayesian regularization, R^2 and R^2_{cv} are square correlation coefficients of data set fitting and LOO crossvalidation, respectively, S and S_{cv} are standard deviations of data set fitting and LOO crossvalidation respectively.

Variables in the AA3DA-BRGNN 3 model mean: AA3DA9V, is the amino acid three-dimensional autocorrelation of lag 9 Å weighted by volume (number of nonhydrogen side-chain atoms; AA3DA6R$_f$, is the amino acid three-dimensional autocorrelation of lag 6 Å weighted by chromatographic index; AA3DA12N$_m$, is the amino acid three-dimensional autocorrelation of lag 12 Å weighted by average long-range contacts; AA3DA15ASA$_N$, is the amino acid three-dimensional autocorrelation of lag 15 Å weighted by solvent-accessible surface area for native; AA3DA27ΔH$_h$, is the amino acid three-dimensional autocorrelation of lag 27 Å weighted by unfolding enthalpy change of hydration.

Figure 1B shows the behavior of E, D and MSE terms for NNES using AA3DA vectors approach. Those values are stabilized about 0.74, 0.12 and 0.61, respectively, for ensembles with 20 members or more. In this case, a lower MSE decrease about 15% is observed in comparison to AASA approach that is related to the better accuracy for predicting stability by the 3D model. Besides ensemble needs fewer networks (only 20 members) for yielding a stable model with a minimum of MSE, therefore better results are obtained with a decrease on time consumption.

Figure 2B depicts plots of calculated vs. experimental unfolding $\Delta\Delta$G values for each protein calculated as an average over training and test sets according to the AA3DA-BRGNN 3 ensemble predictor. In addition, it can be observed AA3DA vectors ensemble was more robust than AASA vectors ensemble depicted in figure 2a. In this case, the accuracy for data fitting was about 89% and 72% for proteins in training and test sets, respectively and only needs five vectors to obtain this good result. AA3DA vectors approach well fits in a nonlinear way the $\Delta\Delta$G by means a combination of 3D information and amino acid/residue properties. As it was expected, the prediction accuracy of the 3D model was a 10% higher than the sequence model, even when it is only trained with five AA3DA vectors. Therefore, fewer AA3DA vectors allow to successfully resemble an optimum stability pattern to be learned by an ensemble of BRGNNs during supervised training.

3.3 Regression model's interpretation

In order to gain a deeper insight on the relative effects of each autocorrelation vector in the optimum BRGNN model, a recently reported weight-based input-ranking scheme was carried out. Black-box nature of three-layers ANNs has been "deciphered" in a recent report of Guha *et al.* [56]. Their method allows understanding how an input descriptor is correlated to the predicted output by the network and consists of two parts. First, the nonlinear transform for a given neuron is linearized. Afterward, the magnitude in which a given neuron affects the downstream output is determined. Next, a ranking scheme for neurons in the hidden layer is developed. Determining square contribution values (SCV) for each hidden neuron carries out ranking scheme. (see Ref. [56] for details). This method for ANN model interpretation is similar in manner to the partial least squares interpretation method for linear models described by Stanton [57].

Table 4. Effective weight matrix for the 8-4-1 AASA-BRGNN 3 model developed for conformational stability of Chymotrypsin Inhibitor 2 mutants.

	Hidden nodes			
	2	*4*	*1*	*3*
AASA7ΔCp$_h$	**− 1.21**	**1.38**	**0.19**	**0.38**
AASA7N$_t$	−0.54	1.58	−1.25	0.10
AASA7N$_l$	0.67	1.25	−0.43	0.23
AASA5V	**− 1.81**	**1.13**	**0.63**	**0.55**
AASA3H$_{nc}$	**2.83**	**0.12**	**− 1.11**	**− 0.94**
AASA1s	0.36	−0.09	0.65	−0.73
AASA12P	0.48	−1.42	−0.83	1.11
AASA8P	**0.73**	**− 2.42**	**1.13**	**− 0.65**
SCV	0.452	0.382	0.131	0.041

The results of the model interpretation analysis for AASA-BRGNN vectors appear in table 4. As can be observed, among the four hidden nodes in the predictor, the most ranked nodes are node 2 and node 4 having a SCV percent value about 40%, which is about 3.4-fold higher than the hidden node 1 and about 11-fold higher than node 3. According to the Guha's analysis [56] the most ranked node has the major impact in the overall output of the neural network. Consequently, the most weighted inputs in such node represent the most relevant descriptors for the regression problem under study. Specifically in table 4, descriptors having weights $>|1|$ in both nodes, such as, AASA5V and AASA7ΔCp_h and descriptor having weights $>|2|$ in any of the most ranked nodes such as, AASA3H$_{nc}$ and AASA8P are the most relevant descriptors. Those descriptors represent autocorrelations of thermodynamic (AASA7ΔCp_h), hydrophobicity/hydrofilicity (AASA3H$_{nc}$, AASA8P) and structural (AASA5V) related properties.

On the other hand, results of the model interpretation analysis for AA3DA-BRGNN 3 appear in table 5. In this case, it was used the same method describe above for AASA vectors, but the most relevant descriptors was selected according to their weights $>|1|$ on the most ranked node 2, which have a SCV percent value above 90%. Such descriptors are AA3DA6Rf, AA3DA15ASA$_N$, AA3DA9V and AA3DA27ΔH$_h$, which represent autocorrelations of thermodynamic (AA3DA27ΔH$_h$), hydrophobicity

Table 5. Effective weight matrix for the 5-4-1 AA3DA-BRGNN 3 model developed for conformational stability of Chymotrypsin Inhibitor 2 mutants.

	Hidden nodes			
	2	*1*	*4*	*3*
AA3DA9V	**1.58**	**0.74**	**− 1.26**	**− 0.68**
AA3DA6R$_f$	**3.58**	**− 0.59**	**− 0.80**	**− 0.96**
AA3DA12N$_m$	−0.75	−0.13	−0.41	1.39
AA3DA15ASA$_N$	**2.07**	**− 1.13**	**− 0.91**	**0.46**
AA3DA27ΔH$_h$	**− 1.41**	**0.82**	**− 2.12**	**1.74**
SCV	0.912	0.041	0.032	0.001

(AA3DA6Rf), and structural (AA3DA15ASA$_N$, AA3DA9V) related properties.

As result of the SCV ranking study a group of amino acid/residues properties appears as the most relevant for sequence and 3D models. Property volume weights both optimum AA3DA and AASA vectors. High significance of such property could be related to the fact that availability of volume to a side-chain at protein interior can produce energetic penalty for conformational alterations after mutation [58,59]. This effect is highly influenced by the size (*V*) of the substitute, added and also surrounding residues. Mutations may cause an unfavorable packing energy due to the rigidity of surrounding residues or, alternatively, the substituting residues themselves may be forced into unfavorable rotational isomers. Similarly, some surroundings of mutation positions may be readily deformable or there may be compensating effects that yield no net packing energy change [59]. This property (*V*) also appeared weighting optimum AASA vectors used for modeling the conformational stability of gene V protein in a previous work [32]. Another structural related property relevant on the 3D model is the solvent-accessible surface area (ASA$_N$), which is a measure of the number of amino acid atoms that can interact with solvent molecules.

The hydrophobicity/hydrophilicity related properties such as: polarity (P), normalized consensus hydrophobicity (Hnc) and chromatographic index (Rf) are so important to predict chymotrypsin inhibitor 2 conformational stability. Hydrophilic interactions between two amino acid residues at protein surface 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. On the contrary, hydrophobic interactions at protein core mainly appear at short-range in the sequence. In our sequence-based model polarity's lag is longer than hydrophobicity consensus's lag, 8 and 3, respectively, given in topological distance. These results point out the role of hydrophobic interactions on the core and hydrophilic interactions on the surface to maintain protein folding and stability.

However, on the protein surface frequently appear hydrophobic patches, defined as clusters of neighboring apolar atoms deemed accessible on a given protein surface [60]. 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-accessible surface area [61]. At the light of these facts, Rf and ASA$_N$ 3D autocorrelations could encode an hydrophobic patches pattern of chymotrypsin inhibitor 2 and mutants. Site directed mutagenesis studies have shown the importance of the Tyr 42 residue on the stability profile of the wild-type protein and six mutants. Despite its solvent exposure, the phenol ring of Tyr 42 makes hydrophobic contacts with the side-chains of Met 40, Glu 41, Arg 43, Ile 44 and Val 63 and it was expected the mutations of Y42A and Y42G are destabilizing, can be attributed mainly to the loss of packing interactions [62].

Thermodynamical properties such as: heat capacity change of hydration (ΔCp_h) and unfolding enthalpy change of hydration (ΔH_h) are relevant to the sequence-based and 3D-based models, respectively. ΔCp_h measurements in proteins mean the variation of Cp (heat capacity), which is consequence of the hydration of aminoacids groups. Taking into account that protein unfolding usually has a positive Cp (heat capacity), polar groups hydration is accompanied by a decrease in Cp meanwhile apolar groups hydration is accompanied by an increase on this magnitude [63]. In this sense Makhatadze and Privalov [64,65] found a good relation between Cp_h and surface area. On the other hand unfolding enthalpy change of hydration, is a direct measure of heat or energy of hydration upon unfolding, in addition the compact native state of a protein is stabilized by the enthalpic interactions between internal groups, while the hydration effects of all the groups, except the aliphatic ones, which are exposed upon unfolding destabilize this state [66].

In resume all of these properties analyzed above are in concordance with unfolding denaturation mechanism hypothesis. For denaturation process of globular proteins, Privalov and Gill [67] stated that hydration equilibrium, polar interactions between solvent and polar residues in the protein, is the main causes of unfolding meanwhile hydrophobic interactions in the protein core contribute to keep the folded state.

3.4 Comparison with previous conformational stability modeling studies

The statistical quality of our ensemble model is in concordance with the report of Marrero-Ponce *et al.* [68] in which they extended topological indexes to the study of biological macromolecules. In such report, protein linear indices of the "macromolecular pseudograph Cα-atom adjacency matrix" were applied to the prediction of actual melting points of Arc repressor mutants and a linear model was obtained using multilinear equation that described about 72% of crossvalidation data variance. Taking into account that conformational stability is a more complex protein property in comparison to other physical stability measurements such as protein melting point, the accuracy over 70% of our 3D approach for modeling the stability of chymotrypsin inhibitor 2 mutants is remarkably good.

Concerning to the prediction of Gibbs free energy change of proteins, our approach in the case of AA3DA vectors were able of resembling a 3D amino acid interaction pattern in chymotrypsin inhibitor 2 that was successfully learned by BRGNNs. Similar accuracy over 0.7 was observed when using protein radial distribution function (P-RDF) scores for solving the same problem, the conformational stability of chymotrypsin inhibitor 2 mutants [69]. Despite the disadvantage of some previous thermodynamic experimental data is required for generating a training set, our modeling technique is a viable alternative for stability prediction when some thermodynamic data exits. At the moment, the prediction approach presented here is protein-specific and then one needs to

obtain a model for each protein of interest. We gain in quality of predictions in comparison to more comprehensive models in Ref. [14, 19] and [22] but with lower generalization abilities. It is noteworthy that our predictor, differently to the most of the reported approaches, successfully encompasses single, double and triple mutants, as well as any kind of mutation.

Despite AASA vectors accuracy is lower than AA3DA vectors, sequence information can be able and useful to adequately predict protein stability when X-ray structural information is lacking but some thermodynamic data exits. The aim of our work was just to present reliable predictors for the conformational stability of a sole protein using both sequence and 3D frameworks and a wide thermodynamic data of their mutants.

4. Conclusions

Protein function are extremely depend on that the nascent amino acid chain correctly folds into the biologically active, three-dimensional structure of the native state. Biological active structure is stabilized by numerous intramolecular interactions such as hydrophobic, electrostatic, van der Waals and hydrogen bond. 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 modeling. We extended the concept of autocorrelation vectors in molecules to the amino acid sequence and 3D structure of proteins as a tool for encoding protein structural information for supervised training of ANNs. In this sense, AASA and AA3DA vectors were calculated measuring sequence and 3D autocorrelations of 48 amino acid/residue properties selected from the AAindex data base on the protein primary structure and 3D structure, respectively. Primary structure approach yielded an adequate eight-input ensemble model for the conformational stability of chymotrypsin inhibitor 2 protein and mutants Despite its relative low prediction accuracy about 0.6 this method is useful because do not require X-ray crystal structure of proteins for implementation. However, 3D approach is more robust and accurate describing about 89 and 72% (about a 10% higher than AASA) of training and test set variances.

The present work demonstrates the successful application of the AA3DA and AASA vectors for modeling protein conformational stability in combination with BRGNN approach. Optimum sequence and 3D-based ensembles exhibit high effects on relevant structural, physico-chemical and thermodynamic properties.

Acknowledgements

Authors would like to acknowledge Professor Akinori Sarai for providing useful information to prepare the manuscript. Financial support of this research by Cuban Ministerio de Ciencia, Tecnología y Medio Ambiente (CITMA) through a grant to M. Fernandez (Grant No. 20104102).

References

[1] B. Bishop, D.C. Koay, A.C. Sartorelli, L. Regan. Reengineering granulocyte colony-stimulating factor for enhanced stability. *J. Biol. Chem.*, **276**, 33465 (2001).
[2] A.N. Bullock, A. Fersht. Rescuing the function of mutant p53. *Nat. Rev. Cancer*, **1**, 68 (2001).
[3] T.J. Graddis, R.L. Remmele Jr., J.T. McGrew. Designing proteins that work using recombinant technologies. *Curr. Pharm. Biotechnol.*, **3**, 285 (2002).
[4] J. Buxbaum. Diseases of protein conformation: what do *in vitro* experiments tell us about *in vivo* diseases? *Trends Biochem. Sci.*, **28**, 585 (2003).
[5] J. Saven. Combinatorial protein design. *Curr. Opin. Struct. Biol.*, **12**, 453 (2002).
[6] J. Mendes, R. Guerois, L. Serrano. Energy estimation in protein design. *Curr. Opin. Struct. Biol.*, **12**, 441 (2002).
[7] D.N. Bolon, J.S. Marcus, S.A. Ross, S.L. Mayo. Prudent modeling of core polar residues in computational protein design. *J. Mol. Biol.*, **329**, 611 (2003).
[8] 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).
[9] T. Lazaridis, M. Karplus. Effective energy function for proteins in solution. *Proteins*, **35**, 133 (1999).
[10] D. Gilis, M. Rooman. Prediction of stability changes upon single site mutations using database-derived potentials. *Theor. Chem. Acc.*, **101**, 46 (1999).
[11] 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).
[12] C. Lee, M. Levitt. Accurate prediction of the stability and activity effects of site- directed mutagenesis on a protein core. *Nature*, **352**, 448 (1991).
[13] 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 (1998).
[14] 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).
[15] 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).
[16] 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).
[17] 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).
[18] K. Takano, K. Yutani. A new scale for side-chain contribution to protein stability based on the empirical stability analysis of mutant proteins. *Protein Eng.*, **14**, 525 (2001).
[19] H. Zhou, Y. Zhou. Stability scale and atomic solvation parameters extracted from 1023 mutation experiment. *Proteins*, **49**, 483 (2002).
[20] S. Levin, B.H. Satir. POLINA: detection and evaluation of single amino acid substitutions in protein superfamilies. *Bioinformatics*, **14**, 374 (1998).
[21] C.M. Frenz. Neural network-based prediction of mutation-induced protein stability changes in staphylococcal nuclease at 20 residue positions. *Proteins*, **59**, 147 (2005).
[22] E. Capriotti, P. Fariselli, R. Casadio. A neural-network-based method for predicting protein stability changes upon single mutations. *Bioinformatics*, **20**, 63 (2004) b) E. Capriotti, P. Fariselli, R. Calabrese, R. Casadio,. Prediction of protein stability changes from sequences using support vector machines. *Bioinformatics*, 2005, **21**, 54 (2005) c) E. Capriotti, P. Fariselli, R. Casadio.

I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucl. Acids Res.*, **33**, 306 (2005).
[23] M. Fernández, J. Caballero, A.M. Helguera, E.A. Castro, M.P. González. Quantitative structure-activity relationship to predict differential inhibition of aldose reductase by flavonoid compounds. *Bioorg. Med. Chem.*, **13**, 3269 (2005).
[24] M. Fernández, A. Tundidor-Camba, J. Caballero. 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**, 1884 (2005).
[25] M. Fernández, A. Tundidor-Camba, J. Caballero. 2D autocorrelation modeling of the activity of trihalobenzocycloheptapyridine analogues as farnesyl protein transferase inhibitors. *Mol. Simulat.*, **31**, 575 (2005).
[26] M.P. González, J. Caballero, A. Tundidor-Camba, A.M. Helguera, 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**, 200 (2006).
[27] M. Fernández, J. Caballero. Modeling of activity of cyclic urea hiv-1 protease inhibitors using regularized-artificial neural networks. *Bioorg. Med. Chem.*, **14**, 280 (2006).
[28] J. Caballero, 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**, 168 (2006).
[29] 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).
[30] M. Fernández, J. Caballero. Bayesian-regularized genetic neural networks applied to the modelling of non-peptide antagonists for the human luteinizing hormone-releasing hormone receptor. *J. Mol. Graph. Model.*, **25**, 410 (2006).
[31] J. Caballero, A. Tundidor, M. Fernandez. Modeling of the inhibition constant (Ki) of some cruzain ketone-based inhibitors using 2D spatial autocorrelation vectors and data-diverse ensembles of Bayesian-regularized genetic neural networks. *QSAR Comb. Sci.*, **26**, 27 (2007).
[32] L. Fernández, J. Caballero, J.I. Abreu, M. Fernández. Amino acid sequence autocorrelation vectors and bayesian-regularized genetic neural networks for modeling protein conformational stability: gene V protein mutants. *Proteins*, **67**, 834 (2007).
[33] P.A.P. Moran. Notes on continuous stochastic processes. *Biometrika*, **37**, 17 (1950).
[34] R.F. Geary. The contiguity ratio and statistical mapping. *Incorporated Statistician*, **5**, 115 (1954).
[35] G. Moreau, P. Broto. Autocorrelation of a topological structure: A new molecular descriptor. *Nouv. J. Chim.*, **4**, 359 (1980).
[36] G. Moreau, P. Broto. Autocorrelation of molecular structures: application to SAR studies. *Nouv. J. Chim.*, **4**, 757 (1980).
[37] M. Wagener, J. Sadowski, J. Gasteiger. Autocorrelation of molecular properties for modelling corticosteroid binding globulin and cytosolic ah receptor activity by neural networks. *J. Am. Chem. Soc.*, **117**, 7769 (1995).
[38] H. Bauknecht, A. Zell, H. Bayer, P. Levi, M. Wagener, J. Sadowski, 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**, 1205 (1996).
[39] (a) K. Nakai, A. Kidera, M. Kanehisa. Cluster analysis of amino acid indices for prediction of protein structure and function. *Protein Eng.*, **2**, 93 (1988) (b) K.M. Tomii, Kanehisa. Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein. Eng.*, **9**, 27 (1996). (c) S. Kawashima, M. Kanehisa. AAindex: amino acid index database. *Nucleic. Acids Res.*, **28**, 374 (2000)..
[40] MATLAB 7.0. program, MATLAB 7.0. program, available from The Mathworks Inc., Natick, MA. http://www.mathworks.com
[41] S. So, M. Karplus. Evolutionary optimization in quantitative structure-activity relationship: an application of genetic neural networks. *J. Med. Chem.*, **39**, 1530 (1996).
[42] D.J.C. Mackay. Bayesian interpolation. *Neural Comput.*, **4**, 415 (1992).
[43] D.J.C. Mackay. A practical Bayesian framework for backprop networks. *Neural Comput.*, **4**, 448 (1992).
[44] F.R. Burden, D.A. Winkler. Robust QSAR models using Bayesian-regularized neural networks. *J. Med. Chem.*, **42**, 3183 (1999).
[45] D.A. Winkler, F.R. Burden. Bayesian neural nets for modeling in drug discovery. *Biosilico*, **2**, 104 (2004).
[46] F.D. Foresee, M.T. Hagan. Gauss–Newton approximation to Bayesian learning. *Proceedings of the 1997 International Joint Conference on Neural Networks*, p. 1997, IEEE, Houston (1997).
[47] L.K. Hansen, P. Salamon. Neural Network Ensembles. *IEEE Trans. Pattern. Anal. Mach. Intell.*, **12**, 993 (1990).
[48] A. Krogh, J. Vedelsby. Neural network ensembles, cross-validation and active learning. In *Advances in Neural Information Processing Systems 7*, G. Tesauro, D. Touretzky, T. Lean (Eds.), p. 231, MIT Press, Cambridge (1995).
[49] K. Baumann. Chance correlation in variable subset regression: influence of the objective function, the selection mechanism, and ensemble averaging. *QSAR Comb. Sci.*, **24**, 1033 (2005).
[50] 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 http://gibk26.bse.kyutech.ac.jp/jouhou/protherm/protherm.html (2004).
[51] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne. The protein data bank. *Nucl. Acids Res.*, **28**, 235 (2000).
[52] B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. Status, S. Swaminathan, M. Karplus. CHARMM: a program for macromolecular energy minimization and dynamics calculations. *J. Comput. Chem.*, **4**, 217 (1983).
[53] T. Lazaridis, M. Karplus. "New view" of protein folding reconciled with the old through multiple unfolding simulations. *Science*, **278**, 1928 (1997).
[54] E. Neria, S. Fischer, M. Karplus. Simulation of activation free energies in molecular systems. *J. Chem. Phys.*, **105**, 1902 (1996).
[55] A.T. Brunger, M. Karplus. Polar hydrogen positions in proteins: empirical energy placement and neutron diffraction comparison. *Proteins*, **4**, 148 (1988).
[56] R. Guha, D.T. Stanton, P.C. Jurs. Interpreting computational neural network QSAR models: a detailed interpretation of the weights and biases. *J. Chem. Inf. Model.*, **45**, 1109 (2005).
[57] D.T. Stanton. On the physical interpretation of QSAR models. *J. Chem. Inf. Comput. Sci.*, **43**, 1423 (2003).
[58] W.S. Sandberg, T.C. Terwilliger. Energetics of repacking a protein interior. *Proc. Natl. Acad. Sci. USA*, **88**, 1706 (1991).
[59] W.S. Sandberg, T.C. Terwilliger. Engineering multiple properties of a protein by combinatorial mutagenesis. *Proc. Natl. Acad. Sci. USA*, **90**, 8367 (1993).
[60] P. Lijnzaad, P. Argos. Hydrophobic patches on protein subunit interfaces: characteristics and prediction. *Proteins*, **28**, 333 (1997).
[61] F. Eisenhaber, P. Argos. Hydrophobic regions on protein surfaces: definition based on hydration shell structure and a quick method for their computation. *Protein Eng.*, **9**, 1121 (1996).
[62] F. Eisenhaber, P. Argos. Analysis of protein–protein interactions by mutagenesis: direct versus indirect effects. *Protein Eng.*, **12**, 41 (1999).
[63] N.V. Prabhu, K.A. Sharp. Heat capacity in proteins. *Annu. Rev. Phys. Chem.*, **56**, 521 (2005).
[64] G.I. Makhatadze, P.L. Privalov. Heat capacity of proteins. I. Partial molar heat capacity of individual amino acid residues in aqueous solution: hydration effect. *J. Mol. Biol.*, **213**, 375 (1990).
[65] P.L. Privalov, G.I. Makhatadze. Heat capacity of proteins. II. Partial molar heat capacity of the unfolded polypeptide chain of proteins: protein unfolding effects. *J. Mol. Biol.*, **213**, 385 (1990).
[66] G.I. Makhatadze, P.L. Privalov. Hydration effects in protein unfolding. *Biophy. Chem.*, **51**, 291 (1994).
[67] P.L. Privalov, S.J. Gill. Stability of protein structure and hydrophobic interaction. *Adv. Prot. Chem.*, **39**, 191 (1988).
[68] 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).
[69] M. Fernández, J. Caballero, L. Fernández, J.I. Abreu, M. Garriga. Protein radial distribution function (P-RDF) and Bayesian-regularized genetic neural networks for modeling protein conformational stability: chymotrypsin inhibitor 2 mutants. *J. Mol. Graph. Model.* (2007), doi: 10.1016/j.jmgm.2007.04.011.