Amino Acid Sequence Autocorrelation Vectors and Ensembles of Bayesian-Regularized Genetic Neural Networks for Prediction of Conformational Stability of Human Lysozyme Mutants

Julio Caballero,[†] Leyden Fernández,[†] José Ignacio Abreu,[†,‡] and Michael Fernández*[,†]

Molecular Modeling Group, Center for Biotechnological Studies, Faculty of Agronomy, and Artificial Intelligence Lab, Faculty of Informatics, University of Matanzas, 44740 Matanzas, Cuba

Received November 22, 2005

Development of novel computational approaches for modeling protein properties from their primary structure is a main goal in applied proteomics. In this work, we reported the extension of the autocorrelation vector formalism to amino acid sequences for encoding protein structural information with modeling purposes. Amino Acid Sequence Autocorrelation (AASA) vectors were calculated by measuring the autocorrelations at sequence lags ranging from 1 to 15 on the protein primary structure of 48 amino acid/residue properties selected from the AAindex database. A total of 720 AASA descriptors were tested for building predictive models of the thermal unfolding Gibbs free energy change of human lysozyme mutants. In this sense, ensembles of Bayesian-Regularized Genetic Neural Networks (BRGNNs) were used for obtaining an optimum nonlinear model for the conformational stability. The ensemble predictor described about 88% and 68% variance of the data in training and test sets, respectively. Furthermore, the optimum AASA vector subset was shown not only to successfully model unfolding thermal stability but also to distribute wild-type and mutant lysozymes on a stability Self-organized Map (SOM) when used for unsupervised training of competitive neurons.

1. INTRODUCTION

Predicting protein structures and stability is a fundamental goal in molecular biology. Even predicting changes in structure and stability induced by point mutations has immediate application in computational protein design.[1−4] Although free energy simulations have accurately predicted relative stabilities of point mutants,[5] the computational cost that most of the methods actually demand are extremely high to test the large number of mutations studied in protein design applications.

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

Among the PEEF approach a simplified energy function with only van der Waals and side-chain torsion potentials[7] 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 the SEEF method includes statistical potentials derived from geometric and environmental propensities and correlations of residues in X-ray crystal structures. Potentials derived from substitution and occurrence frequencies for amino acids in different structural environment classes, such as main-chain conformations and solvent accessibilities, have also been used to calculate the stability differences induced by point mutations.[6,9,10] On the other hand, the EEEF approach combines a physical description of the interactions with some data obtained from experiments previously run on proteins. Examples of such algorithms are the helix/coil transition algorithm AGADIR[11] or FOLDEF, a fast and accurate EEEF approach based on the AGADIR algorithm that uses a full atomic description of the structure of the proteins reported by Guerois et al.[12] for predicting the conformational stability of more than 1000 mutants.

Furthermore, other stability prediction studies nonbased on protein force-field calculations have been focused on correlations of free energy change with structural, sequence information, and amino acid properties such as hydrophobicity, accessible surface area, etc. In this sense, Gromiha et al. had reported some of the seminal works in this topic.[13−15] Such authors had referred the linear relationship between physicochemical, energetic, and conformational amino acid/residues properties and mutation-induced stability for a large set of mutants.[13] The effect of local sequence on the mutations stability was evaluated by computed average properties at sequence positions at segment residues ranging from 3 to 5 about the mutated residue. Similarly surrounding structural effects were established by considering average

* Corresponding author phone: (53) (45) 26 1251; fax: (53) (45) 25 3101; e-mail: michael.fernandez@umcc.cu, michael_llamosa@yahoo.com.
[†] Faculty of Agronomy.
[‡] Faculty of Informatics.

10.1021/ci050507z CCC: $33.50
Published on Web 04/04/2006

properties but at 3D structure neighboring residues at specific radius from the mutation points. In such studies, it was reported the role of structural and sequence information in the prediction of protein stability changes by comparing buried and partially buried mutations.[14] They found that free energy changes 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 the highest segment length effects for helical, strand, and coil mutations are, respectively, 0, 9, and 4 residues on both sides of the mutant residues.[15]

On the other hand, empirical equations involving physical properties calculated from mutant structures have been reported. Several studies concerning mutations on human lysozymes[16] referred that the stability of each mutant can be represented by equations involving physical properties calculated from mutant structures such as hydrophobicity; in addition, hydrogen bond contributions were also important for inducing stability. More recently, Zhou and Zhou[17] reported a broad study regarding 35 proteins and 1023 mutants from which they derived a new stability scale. A "transfer free energy" scale was extracted assuming that the mutation-induced stability change is equal to the change in transfer free energy without needing any structural information. In a second method, the structures of wild-type proteins were used to incorporate the environmental effect of mutation sites.

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 the fact that the size of the protein crystallographic database is continuously growing, 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 they just use amino acid sequence information for predicting protein stability and are extremely less computational intensive in comparison with free energy function based methods.[18] In this context, Levin and Satir[18] successfully evaluated the functional significance of mutations on hemoglobin using amino acid similarity matrixes. Recently, Frenz[19] reported an artificial neural network-based model for predicting the stability of staphylococcal nuclease mutants using amino acid similarity scores as network inputs.

In this connection, outstanding reports of Capriotti et al.[20] 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 for quantitatively and qualitatively predicting stability change using a data set of more than 2000 mutants for training and testing the predictors. 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.[21,22] In such reports, topological molecular descriptor concepts are extended to protein amino acid sequences in such a way that several topological descriptors are computed considering the protein structure as a simplified molecular pseudograph of Cα atoms. In these reports, protein stability studies, specifically how alanine substitution mutation on an Arc repressor wild-type protein affects melting temperate, were accomplished by means of Multilinear Regression Analysis (MRA) and linear discriminant analysis.

In chemistry and related fields of research like biochemistry, chemical engineering, and pharmacy, interest in Artificial Neural Networks (ANNs) computing has grown rapidly. In this regard, ANNs have encountered successful applications in bioinformatic studies. ANNs usually overcome methods limited to linear regression models such as MRA or partial least square.[24−29] 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 pattern recognition problems.

In this work we reported our contribution to the modeling of protein conformational stability by extending the concepts of structural autocorrelation vectors[30−35] in molecules to protein primary structure. Conformational stability of human lysozymes,[16] wild-type and mutants, was nonlinear modeled using amino sequence information. Protein structure information was encoded by means of Amino Acid Sequence Autocorrelation (AASA) vectors weighted by 48 physicochemical, energetic, and conformational amino acid/residues properties extracted from the AAindex amino acid database.[36] In this way, a large set of descriptors was computed, and by employing a nonlinear modeling technique recently introduced by our group, Bayesian-Regularized Genetic Neural Networks (BRGNNs),[26−29] optimum ANN-based predictive models of lysozyme stability were built. To provide robust models, we employed ensembles of BRGNN for calculating the conformational stability instead of one single network. In addition to the quantitative predictive model, we built a Self-organizing Map (SOM) of lysozyme conformational stabilities using the inputs of the optimum BRGNN predictor for unsupervised training of competitive neurons.

2. METHODS AND EXPERIMENTAL PROCEDURE

Amino Acid Sequence Autocorrelation Vector (AASA) Approach. Conformational stability of a protein depends on a variety of intramolecular interactions such as hydrophobic, electrostatic, van der Waals, and hydrogen bond that are ruled by the amino acid sequence. Therefore, in structure−property/activity studies the strategy for encoding structural information must, in some way, either explicitly or implicitly, account for these interactions. Furthermore, usually data sets include structures of different size with different numbers of elements, so the structural encoding approaches must allow for comparing such structures.[35]

Autocorrelation vectors have several useful properties. First, a substantial reduction in data can be achieved by limiting the topological distance, *l*. Second, the autocorrelation coefficients are independent of the original atom numberings, so they are canonical. And third, the length of the correlation vector is independent of the size of the molecule.[35]

For the autocorrelation vectors in molecules, the 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. 2D spatial autocorrelations[30−32] have been successfully used in the past decades for modeling biological activities[32,33] and pharmaceutical research.[34,35] In recent works, our group has obtained outstanding results when such a chemical code was used in combination with the ANN approach in biological QSAR studies.[25,28] 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[32] is defined as follows

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

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

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

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.[15] method in considering the whole amino acid sequence of the protein for calculation of the descriptors instead of local sequence segments over the mutated point. In this way, the calculated autocorrelation vectors encode structural information concerning the whole protein. Particularly, Amino Acid Sequence Autocorrelation (AASA) vectors of lag l are calculated as follows

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

where $\mathrm{AASA}lp_k$ is the AASA at spatial 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 $\delta(l, d_{ij})$ is a Dirac-delta function.

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

$$\mathrm{AASA1}p_k = \frac{1}{9}(p_{kA}\cdot p_{kS} + p_{kS}\cdot p_{kT} + p_{kT}\cdot p_{kC} + p_{kC}\cdot p_{kG} + p_{kG}\cdot p_{kF} + p_{kF}\cdot p_{kH} + p_{kH}\cdot p_{kC} + p_{kC}\cdot p_{kS} + p_{kS}\cdot p_{kD}) \tag{4}$$

$$\mathrm{AASA5}p_k = \frac{1}{5}(p_{kA}\cdot p_{kF} + p_{kS}\cdot p_{kH} + p_{kT}\cdot p_{kC} + p_{kC}\cdot p_{kS} + p_{kG}\cdot p_{kD}) \tag{5}$$

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.[30,31] In a protein, autocorrelation analysis tests whether the value of a property at one residue is independent of the values of the property at neighboring 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 were used 48 physicochemical, energetic, and conformational amino acid/residues properties (Table 1) selected by Gromiha et al.[13] from the AAindex database[36] in a previous study concerning relationships between amino acid/residues properties and protein stability for a large set of proteins. In our work, spatial lag, l, was ranging from 1 to 15 with the aim of accessing to long-range interactions in the sequence due to tertiary structure arrangements. Computational code for AASA vector calculation was written in Matlab environment.[37] A data matrix of 720 AASA vectors, 48 properties × 15 different lags, was generated with the autocorrelation vectors calculated for each lysozyme 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 255 descriptors were obtained. Afterward, optimum predictive models were built with reduced subsets of variables by means of the BRGNN algorithm.

2.2. Bayesian-Regularized Genetic Neural Networks (BRGNN) Approach. In the context of ANN-based modeling of biological interactions we introduced Bayesian-Regularized Genetic Neural Networks (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.[38,39]

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. The BRGNN approach is a version of the So and Karplus report[38] incorporating Bayesian regularization that has been successfully introduced by our group for modeling the inhibitory activity of several therapeutic target enzymes.[26−29] BRGNN was programmed within the Matlab environment[33] using Genetic Algorithm and Neural Networks Toolboxes. The BRGNN technique leads to neural networks trained with optimum inputs selected from the whole AASA vector data matrix (Figure 1).

2.2.1. Bayesian-Regularized Artificial Neural Networks. ANNs are computer-based models in which a number of processing elements, also called neurons, units, or nodes, are interconnected by links in a netlike structure forming

Table 1. Numerical Values of 48 Selected Physicochemical, Energetic, and Conformational Properties of the 20 Amino Acids/Residues[13]

property[a,b]		A	C	D	E	F	G	H	I	K
1	K_0	−25.5	−32.82	−33.12	−36.17	−34.54	−27	−31.84	−31.78	−32.4
2	H_t	0.87	1.52	0.66	0.67	2.87	0.1	0.87	3.15	1.64
3	H_P	13.05	14.3	11.1	11.41	13.89	12.2	12.42	15.34	11.01
4	P	0	1.48	49.7	49.9	0.35	0	51.6	0.1	49.5
5	pH_i	6	5.05	2.77	5.22	5.48	5.97	7.59	6.02	9.74
6	pK'	2.34	1.65	2.01	2.19	1.89	2.34	1.82	1.36	2.18
7	M_w	89	121	133	147	165	75	155	131	146
8	P_1	11.5	13.46	11.68	13.57	19.8	3.4	13.67	21.4	15.71
9	R_f	9.9	2.8	2.8	3.2	18.8	5.6	8.2	17.1	3.5
10	m	14.34	35.77	12	17.26	29.4	0	21.81	19.06	21.29
11	H_{nc}	0.62	0.29	0.9	−0.74	1.19	0.48	−0.4	1.38	−1.5
12	E_{sm}	1.4	1.37	1.16	1.16	1.14	1.36	1.22	1.19	1.07
13	E_1	0.49	0.67	0.35	0.37	0.72	0.53	0.54	0.76	0.3
14	E_t	1.9	2.04	1.52	1.54	1.86	1.9	1.76	1.95	1.37
15	P_α	1.42	0.7	1.01	1.51	1.13	0.57	1	1.08	1.16
16	P_β	0.83	1.19	0.54	0.37	1.38	0.75	0.87	1.6	0.74
17	P_t	0.66	1.19	1.46	0.74	0.6	1.56	0.95	0.47	1.01
18	P_C	0.71	1.19	1.21	0.84	0.71	1.52	1.07	0.66	0.99
19	C_a	20	25	26	33	46	13	37	39	46
20	F	0.96	0.87	1.14	1.07	0.69	1.16	0.8	0.76	1.14
21	P_r	0.38	0.57	0.14	0.09	0.51	0.38	0.31	0.56	0.04
22	R_a	3.7	3.03	2.6	3.3	6.6	3.13	3.57	7.69	1.79
23	N_s	6.05	7.86	4.95	5.1	6.62	6.16	5.8	7.51	4.88
24	α_n	1.59	0.33	0.53	1.45	1.14	0.53	0.89	1.22	1.13
25	α_c	1.44	0.76	2.13	2.01	1.01	0.62	0.56	0.68	0.59
26	α_m	1.22	1.53	0.56	1.28	1.13	0.4	2.23	0.77	1.65
27	V^0	60.46	67.7	73.83	85.88	121.48	43.25	98.79	107.72	108.5
28	N_m	2.11	1.88	1.8	2.09	1.98	1.53	1.98	1.77	1.96
29	N_1	3.92	5.55	2.85	2.72	4.53	4.31	3.77	5.58	2.79
30	H_{gm}	13.85	15.37	11.61	11.38	13.93	13.34	13.82	15.28	11.58
31	ASA_D	104	132.5	132.2	161.9	182	73.4	165.8	171.5	195.2
32	ASA_N	33.2	17.9	62.4	81	33.1	29.2	57.7	28.3	107.5
33	ΔASA	70.9	114.3	69.6	80.5	148.4	44	107.9	142.7	87.5
34	ΔG_h	−0.54	−1.64	−2.97	−3.71	−1.06	−0.59	−3.38	0.32	−2.19
35	G_{hD}	−0.58	−1.91	−6.1	7.37	−1.35	−0.82	−5.57	0.4	−5.97
36	G_{hN}	−0.06	−0.27	−3.11	−3.62	−0.28	−0.23	−2.18	0.07	−1.7
37	ΔH_h	−2.24	−3.43	−4.54	−5.63	−5.11	−1.46	−6.83	−3.84	−5.02
38	$-T\Delta S_h$	1.7	1.79	1.57	1.92	4.05	0.87	3.45	4.16	2.83
39	ΔC_{ph}	14.22	9.41	2.73	3.17	39.06	4.88	20.05	41.98	17.68
40	ΔG_c	0.51	2.71	2.89	3.58	3.22	0.68	3.95	−0.4	1.87
41	ΔH_c	2.77	8.64	4.72	5.69	11.93	1.23	7.64	4.03	3.57
42	$-T\Delta S_c$	−2.25	−5.92	−1.83	−2.11	−8.71	−0.55	−3.69	−4.42	−1.7
43	ΔG	−0.02	1.08	−0.08	−0.13	2.16	0.09	0.56	−0.08	−0.32
44	ΔH	0.51	5.21	0.18	0.05	6.82	−0.23	0.79	0.19	−1.45
45	$-T\Delta S$	−0.54	−4.14	−0.26	−0.19	−4.66	0.31	−0.23	−0.27	1.13
46	V	1	2	4	5	7	0	6	4	5
47	s	0	0	2	3	2	0	2	1	0
48	f	0	1	2	3	2	0	2	2	4

L	M	N	P	Q	R	S	T	V	W	Y
−31.78	−31.18	−30.9	−23.25	−32.6	−26.62	−29.88	−31.23	−30.62	−30.24	−35.01
2.17	1.67	0.09	2.77	0	0.85	0.07	0.07	1.87	3.77	2.67
14.19	13.62	11.72	11.06	11.78	12.4	11.68	12.12	14.73	13.96	13.57
0.13	1.43	3.38	1.58	3.53	52	1.67	1.66	0.13	2.1	1.61
5.98	5.74	5.41	6.3	5.65	10.76	5.68	5.66	5.96	5.89	5.66
2.36	2.28	2.02	1.99	2.17	1.81	2.21	2.1	2.32	2.38	2.2
131	149	132	115	146	174	105	119	117	204	181
21.4	16.25	12.82	17.43	14.45	14.28	9.47	15.77	21.57	21.61	18.03
17.6	14.7	5.4	14.8	9	4.6	6.9	9.5	14.3	17	15
18.78	21.64	13.28	10.93	17.56	26.66	6.35	11.01	13.92	42.53	31.55
1.06	0.64	−0.78	0.12	−0.85	−2.53	−0.18	−0.05	1.08	0.81	0.26
1.32	1.3	1.18	1.24	1.12	0.92	1.3	1.25	1.25	1.03	1.03
0.65	0.65	0.38	0.46	0.4	0.55	0.45	0.52	0.73	0.83	0.65
1.97	1.96	1.56	1.7	1.52	1.48	1.75	1.77	1.98	1.87	1.69
1.21	1.45	0.67	0.57	1.11	0.98	0.77	0.83	1.06	1.08	0.69
1.3	1.05	0.89	0.55	1.1	0.93	0.75	1.19	1.7	1.37	1.47
0.59	0.6	1.56	1.52	0.98	0.95	1.43	0.96	0.5	0.96	1.14
0.69	0.59	1.37	1.61	0.87	1.07	1.34	1.08	0.63	0.76	1.07
35	43	28	22	36	55	20	28	33	61	46
0.79	0.78	1.04	1.16	1.07	1.05	1.13	0.96	0.79	0.77	1.01
0.5	0.42	0.15	0.18	0.11	0.07	0.23	0.23	0.48	0.4	0.26
5.88	5.21	2.12	2.12	2.7	2.53	2.43	2.6	7.14	6.25	3.03
7.37	6.39	5.04	5.65	5.45	5.7	5.53	5.81	7.62	6.98	6.73

Table 1 (Continued)

L	M	N	P	Q	R	S	T	V	W	Y
1.91	1.25	0.53	0	0.98	0.67	0.7	0.75	1.42	1.33	0.58
0.58	0.73	0.93	2.19	1.2	0.39	0.81	1.25	0.63	1.4	0.72
1.05	1.47	0.93	0	1.63	1.59	0.87	0.46	1.2	0.46	0.52
107.75	105.35	78.01	82.83	93.9	127.34	60.62	76.83	90.78	143.91	123.6
2.19	2.27	1.84	1.32	2.03	1.94	1.57	1.57	1.63	1.9	1.67
4.59	4.14	3.64	3.57	3.06	3.78	3.75	4.09	5.43	4.83	4.93
14.13	13.86	13.02	12.35	12.61	13.1	13.39	12.7	14.56	15.48	13.88
161.4	189.8	134.9	135.1	164.9	210.2	111.4	130.4	143.9	208.8	196.4
31.1	41.3	60.5	60.7	71.5	94.5	48.7	52	28.1	39.5	50.4
129.8	147.9	74	73.5	93.3	116	62.8	78	115.6	167.8	145.9
0.27	−0.6	−3.55	0.32	−3.92	−5.96	−3.82	−1.97	0.13	−3.8	−5.64
0.35	−0.71	−6.63	0.56	−7.12	−12.78	−6.18	−3.66	0.18	−4.71	−8.45
0.07	−0.1	−3.03	0.23	−3.15	−6.85	−2.36	−1.69	0.04	−0.88	−2.82
−3.52	−4.16	−5.68	−1.95	−6.23	−10.43	−5.94	−4.39	−3.15	−8.99	−10.67
3.79	3.56	2.13	2.27	2.31	4.47	2.12	2.42	3.28	5.19	5.03
38.26	31.67	3.91	23.69	3.74	16.66	6.14	16.11	32.58	37.69	30.54
−0.35	1.13	3.26	−0.39	3.69	5.25	3.42	1.74	−0.19	5.59	6.56
3.69	7.06	3.64	1.97	4.47	6.03	5.8	4.42	3.45	13.46	14.41
−4.04	−5.93	−0.39	−2.36	−0.78	−0.78	−2.38	−2.68	−3.64	−7.87	−7.95
−0.08	0.53	−0.3	−0.06	−0.23	−0.71	−0.4	−0.24	−0.06	1.78	0.91
0.17	2.89	−2.03	0.02	−1.76	−4.4	−0.16	0.04	0.3	4.47	3.73
−0.24	−2.36	1.74	−0.08	1.53	3.69	−0.24	−0.28	−0.36	−2.69	−2.82
4	4	4	3	5	7	2	3	3	10	8
2	0	2	0	3	5	0	1	1	2	2
2	3	2	0	3	5	1	1	1	2	2

[a] K^0, compressibility; H_t, thermodynamic transfer hydrophobicity; H_p, surrounding hydrophobicity; P, polarity; pH_i, isoelectric point; pK', equilibrium constant with reference to the ionization property of COOH group; Mw, molecular weight; B_1, bulkiness; R_f, chromatographic index; μ, refractive index; H_{nc}, normalized consensus hydrophobicity; E_{sm}, short- and medium-range nonbonded energy; E_1 long-range nonbonded energy; E_t, total nonbonded energy ($E_{sm} + E_1$); P_α, P_β, P_t, and P_c are, respectively, α-helical, β-structure, turn, and coil tendencies; C_a, helical contact area; F, mean rms fluctuational displacement; Br, buriedness; R_a, solvent-accessible reduction ratio; Ns, average number of surrounding residues; α_n, α_c, and α_m are, respectively, power to be at the N-terminal, C-terminal, and middle of α-helix; $V°$, partial specific volume; N_m and N_1 are, respectively, average medium- and long-range contacts; H_{gm}, combined surrounding hydrophobicity (globular and membrane); ASA_D, ASA_N, and ΔASA are, respectively, solvent-accessible surface area for denatured, native, and unfolding; ΔG_h,\ G_{hD}, and G_{hN} are, respectively, Gibbs free energy change of hydration for unfolding, denatured, and native protein; ΔH_h, unfolding enthalpy change of hydration; $-T\Delta S_h$, unfolding entropy change of hydration; ΔC_{ph}, unfolding hydration heat capacity change; ΔG_C, ΔH_C, and $-T\Delta S_C$ are, respectively, unfolding Gibbs free energy, unfolding enthalpy, and unfolding entropy changes of side-chain; ΔG, ΔH, and $-T\Delta S$ are, respectively, unfolding Gibbs free energy change, unfolding enthalpy change, and unfolding entropy change of protein; V, volume (number of non-hydrogen side-chain atoms); s, shape (position of branch point in a side chain); f, flexibility (number of side-chain dihedral angles). [b] K^0 in m^3/mol/Pa ($\times 10^{-15}$); H_t, H_p, H_{nc}, H_{gm}, ΔG_h, G_{hD}, G_{hN}, ΔH_h, $-T\Delta S_h$, ΔG_C, ΔH_c, $-T\Delta S_c$, ΔG, ΔH, and $-T\Delta S$ in kcal/mol; P in Debye; P_{hi} and pK' in pH units; E_{sm}, E_1, and E_t, in kcal/mol/atom; B_1, C_α, ASA_D, ASA_N, and ΔASA in $Å^2$; F in Å; $V°$ in m^3/mol ($\times 10^{-6}$); ΔC_{ph} in cal/mol/K; and the rest are dimensionless quantities.

"layers".[40,41] Every connection between two neurons is associated with a weight, a positive or negative real number that multiplies the signal from the preceding neuron. Neurons are commonly distributed among the input, hidden, and output layers. Neurons in the input layer receive their values from independent variables; in turn, the hidden neurons collect values from precedent neurons, giving a result that is passed to a successor one. Finally, neurons in the output layer take values from other units and correspond to different dependent variables.

Commonly, ANNs are adjusted, or trained, so that a particular input leads to a specific target output. According to this, the output j is obtained from the input j, by application of eq 6

$$\text{out}_j = f(\text{inp}_j) \tag{6}$$

where the function f is called transfer function. When the ANN is training, the weights are updated in order to minimize network error. In contrast to common statistical methods, ANNs are not restricted to linear correlations or linear subspaces.[40] The employed transfer function, commonly hyperbolic tangent function, allows for establishing nonlinear relations. Thus, ANNs can take into account nonlinear structures and structures of arbitrarily shaped clusters or curved manifolds.

While more connections take effect, the ANN adjusts better the relation input−output. However, when parameters increase, the network loses its ability to generalize. The error on the training set is driven to a very small value, but when new data are presented to the network the error is large. In this process, the predictor has memorized the training examples, but it has not learned to generalize to new situations; it means network overfits the data.

Typically, training aims to reduce the sum of squared errors:

$$F = \text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - t_i)^2 \tag{7}$$

In this equation F is the network performance function, MSE is the mean of the sum of squares of the network errors, N is the number of mutants, y_i is the predicted stability of the mutant i, and t_i is the experimental stability of the mutant i.

MacKay's Bayesian-regularized ANNs (BRANNs) have been designed to resist overfitting.[42] To accomplish this purpose, BRANNs include an error term that regularizes the weights by penalizing overly large magnitudes.

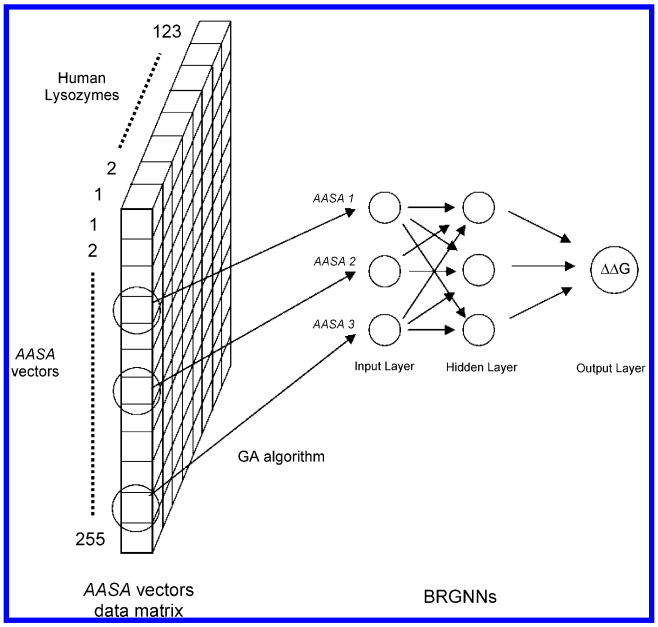

Figure 1. Schematic representation of Bayesian-Regularized Genetic Neural Network (BRGNN) technique with a prototype back-propagation neural network with 3−3−1 architecture. AASA vectors chosen by genetic algorithm constitute inputs and network is trained against the change of thermal unfolding Gibbs free energy change ($\Delta\Delta G$) of human lysozymes.

Assuming a set of pairs $D = \{x_i, t_i\}$, where $i = 1...N$ is a label running over the pairs, the data set can be modeled as deviating from this mapping under some additive noise process (v_i):

$$t_i = y_i + v_i \tag{8}$$

If v is modeled as zero-mean Gaussian noise with standard deviation σ_v, then the probability of the data given the parameters w is

$$P(D|w,\beta,M) = \frac{1}{Z_D(\beta)} \exp(-\beta \times \mathrm{MSE}) \tag{9}$$

where M is the particular neural network model used, and $\beta = 1/\sigma_v^2$ and the normalization constant are given by $Z_D(\beta) = (\pi/\beta)^{N/2}$. $P(D|w,\beta,M)$ is called the likelihood. The maximum likelihood parameters w_{ML} (the w that minimizes MSE) depend sensitively on the details of the noise in the data.

For completing the interpolation model, it must be defined a prior probability distribution which embodies our prior knowledge on the sort of mappings that are "reasonable".[43] Typically this is quite a broad distribution, reflecting the fact that we only have a vague belief in a range of possible parameter values. Once we have observed the data, Bayes' theorem can be used to update our beliefs, and we obtain the posterior probability density. As a result, the posterior distribution is concentrated on a smaller range of values than the prior distribution. Since a neural network with large weights will usually give rise to a mapping with large curvature, we favor small values for the network weights. At this point, it is defined a prior that expresses the sort of smoothness it is expected for the interpolant to have. The model has a prior of the form

$$P(w|\alpha,M) = \frac{1}{Z_W(\alpha)} \exp(-\alpha \times \mathrm{MSW}) \tag{10}$$

where α represents the inverse variance of the distribution, and the normalization constant is given by $Z_W(\alpha) = (\pi/\alpha)^{N/2}$. MSW is the mean of the sum of the squares of the network weights and is commonly referred to as a regularizing function.

Considering the first level of inference, if α and β are known, then the posterior probability of the parameters w is

$$P(w|D,\alpha,\beta,M) = \frac{P(D|w,\beta,M) \times P(w|\alpha,M)}{P(D|\alpha,\beta,M)} \tag{11}$$

where $P(w|D,\alpha,\beta,M)$ is the posterior probability, that is the plausibility of a weight distribution considering the information of the data set in the model used, $P(w|\alpha,M)$ is the prior density, which represents our knowledge of the weights before any data is collected, $P(D|w,\beta,M)$ is the likelihood function, which is the probability of the data occurring, given the weights, and $P(D|\alpha,\beta,M)$ is a normalization factor, which guarantees that the total probability is 1.

Considering that the noise in the training set data is Gaussian and that the prior distribution for the weights is Gaussian, the posterior probability fulfills the relation

$$P(w|D,\alpha,\beta,M) = \frac{1}{Z_F} \exp(-F) \tag{12}$$

where Z_F depends on objective function parameters. So under this framework, the minimization of F is identical to finding the (locally) most probable parameters.[42]

In short, Bayesian regularization involves modifying the performance function (F) defined in eq 7, which is possible improving generalization by adding an additional term.

$$F = \beta \times \mathrm{MSE} + \alpha \times \mathrm{MSW} \tag{13}$$

The relative size of the objective function parameters α and β dictates the emphasis for getting a smoother network response. MacKay's Bayesian framework automatically adapts the regularization parameters to maximize the evidence of the training data.[42]

Bayesian regularization overcomes the remaining deficiencies of neural networks and produces predictors that are robust and well matched to the data; in this sense, BRANNs have been successfully applied in structure−property/activity analysis.[22−25,35]

Fully connected, three-layer BRANNs with back-propagation training were implemented in MATLAB environment.[33] 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.[44] 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]

2.2.2. Genetic Algorithm. GAs are governed by biological evolution rules.[45] They are stochastic optimization methods that have been inspired by evolutionary principles. The

distinctive aspect of a GA is that it investigates many possible solutions simultaneously, each of which explores different regions in parameter space.[46] The first step is to create a population of *N* individuals. Each individual encodes the same number of randomly chosen descriptors. The fitness of each individual in this generation is determined. In the second step, a fraction of children of the next generation is produced by crossover (crossover children) and the rest by mutation (mutation children) from the parents on the basis of their scaled fitness scores. The new offspring contains characteristics from two or one of its parents.

In the BRGNN approach, individuals in the populations are BRANN predictors with a fixed architecture, and the MSE of data fitting was tried as the individual fitness function. An individual is represented by a string of integers which means the numbering of the rows in the all-descriptors matrix (255 rows × 123 columns) will be tested as BRANN inputs. So and Karplus[38] used a variety of fitness functions which are proportional to the residual error of the training set, the test set, or even the cross-validation set from the neural network simulations. However, since we implemented regularized networks, we tried the MSE of data fitting as the individual fitness function. The first step is to create a gene pool (population of neural network predictors) of *N* individuals. Each individual encodes the same number of descriptors; the descriptors are randomly chosen from a common data matrix and in a way such that (1) no two individuals can have exactly the same set of descriptors and (2) all descriptors in a given individual must be different. The fitness of each individual in this generation is determined by the MSE of the model and scaled using and scaling function. A top scaling fitness function scaled a top fraction of the individuals in a population equally; these individuals have the same probability to be reproduced, while the rest are assigned the value 0.

The next step, a fraction of children of the next generation, is produced by crossover (crossover children) and the rest by mutation (mutation children) from the parents. Sexual and asexual reproductions take place so that the new offspring contains characteristics from two or one of its parents. In a sexual reproduction two individuals are selected probabilistically on the basis of their scaled fitness scores and serve as parents. Next, in a crossover each parent contributes a random selection of half of its descriptor set, and a child is constructed by combining these two halves of "genetic code". Finally, the rest of the individuals in the new generation are obtained by asexual reproduction when parents selected randomly are subjected to a random mutation in one of its genes; i.e., one descriptor is replaced by another.

Similarly to So and Karplus,[38] we also included elitism which protects the fittest individual in any given generation from crossover or mutation during reproduction. The genetic content of this individual simply moves on to the next generation intact. This selection, crossover, and mutation process is repeated until all of the *N* parents in the population are replaced by their children. The fitness score of each member of this new generation is again evaluated, and the reproductive cycle is continued until 90% of the generations showed the same target fitness score.[47]

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 guarantees to possess good generalization abilities (Figure 1). This is due to the fact that we used MSE of data training fitting instead of cross-validation or test set MSE values as cost function, and therefore the optimum model cannot be directly derived from the best fitted model yielded by the genetic search. However, from cross-validation experiments over the subpopulation of well fitted models it can derive an optimum generalizable network with the highest predictive power. This process also ensures avoiding chance correlations. This approach has shown to be highly efficient in comparison with the cross-validation-based GA approach since only optimum models, according to the Bayesian regularization, are cross-validated at the end of the routine and not all the models generated throughout all the search process.

2.2.3. 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.[48]

An effective NNE should consist of a set of ANNs that are not highly correct and 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[49] later proved that the ensemble error can be divided into a term measuring the average generalization error of each individual network and a term called diversity that measures the disagreement among the networks. In this way, the MSE of the ensemble estimator is guaranteed to be less than or equal to the averaged MSE of the component estimators.

Model diversity can be introduced by manipulating the input features (feature selection), randomizing the training procedure (overfitting, underfitting, training with different topologies, and/or training parameters, etc.), manipulating the response value (adding noise), or manipulating the training set.[50] Since BRANN predictors have demonstrated to be highly stable to network topology variations,[39] the later method was used for introducing diversity in BRGNN ensembles.

Here we used a perturbation technique called subagging, but the results are not expected to be different for traditional bagging.[51] A bootstrapped generated training set is obtained; afterward the repetitions in the bootstrap sample are removed (i.e. remove objects that were drawn twice, thrice, etc.). The resulting set encompasses the training set, while the remaining objects, which are not part of the training set, represent the test set (set difference between all objects and the training set). Note that removal of the repetitions after the bootstrap sampling is the only difference between subagging and bagging.[51]

2.2.4. Model's Validation. In this work, we validated our regression model using a reasonable method recently employed by our group which consists of a robust validation process by means of NNE.[26] Recently Baumann[51] demonstrated that ensemble averaging significantly improves prediction accuracy by averaging the predictions of several models that are obtained in parallel with bootstrapped training

sets and provides 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 (see section 2.2.3). 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.

2.3. Self-Organizing Maps (SOM). Despite the fact that back-propagated neural networks have been extensively preferred for nonlinear QSAR modeling, SOMs have been also reported as useful ANNs accounting important merits and widespread applications.[52,53]

SOMs[54] are a class of unsupervised neural networks whose characteristic feature is their ability to map nonlinear relations in multidimensional data sets into easily visualizable two-dimensional grids of neurons. SOMs are also referred to as self-organized topological feature maps since the basic function of a SOM is to display the topology of a data set, that is, the relationships between members of the set. These relationships are gathered in several clusters; each local group has the result that topologically close neurons react similarly when they receive similar input. Essentially, SOMs permit to perceive similarities in objects.

In SOMs the input units are fully connected to the 2D Kohonen layer. Each neuron within the Kohonen layer has a well-defined topology, which means a defined number of neurons in its neighborhood. The SOMs are trained through an unsupervised competitive learning process using a "winner takes it all" policy. Under this process, mutant *s*, characterized by *m* autocorrelation vectors $AASA_{si}$, will be projected into neuron c_s, that has weights w_{ji}, most similar to the input variables (eq 14).

$$\text{out}_{c_s} \leftarrow [\sum_{i=1}^{m}(\text{AASA}_{si} - w_{ij})^2] \qquad (14)$$

Albeit all neurons in the active layer obtain the same multidimensional input pattern at the same time, only one is selected to represent this pattern. That neuron is avowed as the winner because it has the smallest Euclidian distance between the presented *m*-dimensional input pattern vector $AASA_s$ ($AASA_{s1}$, $AASA_{s2}$, ..., $AASA_{si}$, ..., $AASA_{sm}$) and the *m*-dimensional weight vector w_i (w_{i1}, w_{i2}, ..., w_{ij}, ..., w_{im}) of the *i* neurons.

Learning within a Kohonen layer consists of the adjustment of the weights, w_{ij}, in such a way that the weights of the winning neuron c_s are shifted closer to the values of the input data. However, not only the weights of the winning neuron are adjusted but also those of the neighboring neurons. Equation 15 gives the correction formula for the weights.

$$\Delta w_{ij} = w_{ij} + f \times (\text{AASA}_{si} - w_{ij}) \qquad (15)$$

The correction factor *f* has the largest value for the weights in the winning neuron c_s and decreases with increasing distance between winning and neighboring neurons. Therefore, when a training case is presented to the network, and the winning neuron is found, the winner updates its weights using the current learning rate, while the neighbors scale down their weights proportional to the distance to the winner.

To settle conformational similarities among human lysozyme wild-type and mutants, a Kohonen SOM was built. The optimum AASA vectors selected by BRGNN were used for unsupervised training of 13×13 neuron maps. SOMs were implemented in a MATLAB environment.[37] Neurons were initially located at a grid topology. The ordering phase was developed in 1000 steps with 0.9 learning rate until tuning neighborhood distance (1.0) was achieved. The tuning phase learning rate was 0.02. Training was performed for a period of 2000 epochs in an unsupervised manner.

2.4. Human Lysozyme Data Set. Human lysozyme (130 residues) is a good model for protein stability studies because it is possible to obtain qualitative thermodynamic parameters from differential scanning calorimetry (DSC) measurements of heat-denaturation process. Human lysozyme mutant data (wild-type and 122 mutants) was collected from Protherm database.[55] Table 2 shows differences of Gibbs free energy change for the thermal denaturation process at 64.9 °C and pH = 2.7 for wild-type and mutants in comparison to wild-type enzyme.

3. RESULTS AND DISCUSSION

By using the amino acid sequences of the 123 lysozymes under study (wild-type and mutants) AASA vectors were computed weighted by a variety of physicochemical, energetic, and conformational properties. In this way, we gathered in a pool of descriptors, the structural information that can be relevant for modeling the conformational stability of lysozyme mutants. Inside the BRGNN framework, GA searches for the best fitted BRANN, in such a way that from one generation to another the algorithm tried to minimize the MSE of the networks (fitness function). By employing this approach instead of a more complicated and time-consuming cross-validation based fitness function, we gain in CPU time and simplicity of the routine. 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 2.2.1).

Nonlinear models were searched by the BRGNN technique varying the network inputs from 6 to 12 AASA vectors. As result of the algorithm a small population of well fitted models was obtained. Afterward those models were tested in cross-validation experiments, and the model with the best cross-validation statistics was selected as optimum. In Table 3 appear statistical parameters for the optimum BRGNN predictors with nine inputs but varying the number of hidden nodes. As can be observed, an optimum predictor was found having three hidden nodes. This optimum nonlinear model describes about an 86% variance of data fitting and a 68%

Table 2. Experimental and Calculated Change of Thermal Unfolding Gibbs Free Energy Change ($\Delta\Delta G$[a]) at 64.9 °C for Wild-Type and Mutant Human Lysozymes According to a 50 Members Neural Network Ensemble of Optimum Model BRGNN 2

	$\Delta\Delta G$ (kJ/mol)				$\Delta\Delta G$ (kJ/mol)		
mutant	exp	cal_{train}[b]	cal_{test}[c]	mutant	exp	cal_{train}[b]	cal_{test}[c]
wild-type	0.0	−0.7	−0.8	S80A	2.0	−1.4	−2.1
A32L	−0.4	1.9	4.0	S82A	1.6	−1.5	−1.8
A32S	−1.4	−2.0	−4.0	T11A	1.6	0.7	0.3
A92S	3.4	1.3	0.9	T11V	1.3	0.0	0.0
A96M	0.1	1.8	2.4	T40A	−6.3	−6.0	−6.6
A96S	−4.2	−1.9	−1.5	T40V	−5.6	−3.1	−2.1
A9S	−0.1	−0.9	−1.2	T43A	−1.5	−2.1	−1.9
D102N	0.7	−0.6	−0.7	T43V	4.0	−0.4	−1.2
D120N	0.4	−0.7	−0.8	T52A	−3.8	−3.4	−3.6
D18N	−2.2	−0.7	−0.7	T52V	−3.6	−3.1	−2.7
D49N	0.0	−0.6	−0.6	T70A	−6.2	−1.7	−1.1
D67N	−1.0	−0.7	−0.7	T70V	−2.9	0.2	0.6
E35L	−2.2	−0.3	0.4	V100A	−1.1	−1.2	−0.4
E7Q	0.4	−0.3	−0.2	V100F	−6.9	−7.2	−7.9
G105A	−2.6	−1.9	−2.0	V100T	−1.2	−3.5	−4.4
G127A	−2.3	−0.6	−0.6	V110A	2.2	−0.8	−3.0
G129A	0.6	−0.4	−0.3	V110D	0.7	−1.2	−3.9
G16A	−5.8	−2.3	−1.5	V110F	−0.2	−0.4	−2.1
G19A	−7.4	−8.4	−10.7	V110G	2.0	2.0	4.3
G22A	−7.5	−5.6	−3.3	V110I	3.6	1.1	0.8
G37A	−1.2	0.1	0.4	V110L	0.3	0.4	0.4
G37Q	−1.1	1.4	2.3	V110M	2.2	−0.4	−2.0
G48A	1.9	−0.3	−0.7	V110N	0.3	0.2	0.1
G68A	−0.5	−0.5	−0.5	V110R	3.7	1.2	−2.7
G72A	−1.5	−2.1	−1.7	V110Y	−0.6	−1.9	−4.6
H78A	−0.6	−3.3	−3.5	V121A	−6.0	−3.7	−3.3
H78G	−0.5	−1.0	−1.9	V125A	−5.5	−2.6	−2.2
I106A	−3.9	−4.4	−5.0	V130A	−3.5	−3.1	−3.6
I106V	−3.0	−3.3	−3.2	V2A	−6.3	−5.5	−5.8
*I*23A	−10.6	−11.0	−12.4	V2F	−3.6	−2.6	−2.9
*I*23V	−1.5	−3.0	−3.0	V2G	−9.6	−7.1	−5.3
I56A	−15.5	−14.7	−10.8	V2I	4.6	0.4	−0.2
I56F	−17.1	−16.8	−16.7	V2L	−0.2	−0.9	−0.9
I56L	−0.4	−1.8	−1.9	V2M	−1.3	−0.1	−0.1
I56M	−7.4	−6.4	−5.1	V2N	−5.6	−5.4	−5.1
I56T	−15.2	−15.1	−15.1	V2R	−1.6	−1.1	−0.7
I56V	−5.0	−4.3	−4.3	V2S	−5.9	−4.9	−4.7
I59A	−7.2	−8.2	−9.5	V2Y	−1.5	−2.5	−2.7
I59F	−3.4	−4.5	−5.0	V74A	−1.5	−2.2	−2.0
I59G	−16.0	−13.9	−12.1	V74D	−1.8	−1.0	−0.6
I59L	0.0	−1.7	−1.8	V74F	−1.2	0.3	1.0
I59M	−5.4	−2.7	−1.9	V74G	−0.9	−1.5	−2.3
I59S	−15.0	−15.6	−15.6	V74I	1.9	0.3	0.2
I59T	−9.3	−9.8	−9.8	V74L	0.8	0.1	−0.2
I59V	−4.6	−2.6	−2.5	V74M	2.7	0.1	−0.6
I59Y	−15.8	−14.7	−12.3	V74N	−1.4	−1.5	−1.0
I89A	−11.3	−10.2	−9.3	V74R	−0.3	0.6	1.4
I89V	−2.0	−2.9	−2.9	V74S	−1.6	−1.7	−1.5
K1A	−2.5	−3.9	−4.3	V74Y	−1.0	0.8	1.2
K1M	−0.5	−0.2	−0.3	V93A	−3.1	−4.9	−5.2
L8T	−15.6	−13.9	−5.0	V93T	−2.8	−2.5	−2.6
N118A	0.8	1.2	7.9	V99A	−4.1	−3.4	−2.6
N118G	0.2	0.7	0.2	V99T	−2.1	−2.7	−2.4
Q58G	7.8	6.9	6.5	Y124F	−1.5	0.0	0.1
R21A	5.5	4.8	3.9	Y20F	−2.1	−1.0	−1.0
R21G	4.8	4.6	2.9	Y38A	−10.4	−9.2	−1.5
R50A	1.8	2.0	5.0	Y38F	−0.8	−1.8	−2.3
R50G	1.1	2.0	6.6	Y38G	−9.7	−9.7	−11.7
S24A	−2.2	−3.7	−5.7	Y45F	0.3	−1.4	−1.5
S36A	−4.7	−4.8	−4.2	Y54F	−4.0	−1.9	−1.9
S51A	−1.0	−1.5	−1.2	Y63F	−1.0	−0.2	−0.7
S61A	−5.7	−1.1	−0.8				

[a] $\Delta\Delta G$ negative and positive values mean destabilizing and stabilizing mutations, respectively. [b] Calculated as average over training sets using a 50 members ensemble. [c] Calculated as average over test sets in the using a 50 members ensemble.

Table 3. Statistics of the Optimum BRGNN Predictors for the Conformational Stability of Wild-Type and Mutant Human Lysozymes[a]

descriptors	BRGNN model	hidd. nod.[b]	R^2	Q^2
AASA6ASA_D AASA2ΔG_c AASA4*s* AASA8H_t	1	2	0.623	0.289
AASA7-$T\Delta S_h$ AASA1ΔC_{ph} AASA1α_N AASA7α_N	**2**	**3**	**0.861**	**0.682**
AASA9-$T\Delta S$	3	4	0.859	0.517
	4	5	0.912	0.522
	5	6	0.931	0.544

[a] Optimum neural network predictor appears in bold. [b] hidd. nod. represents the number of hidden nodes; R^2 and Q^2 are the square correlation coefficients of data fitting and leave-one-out (LOO) cross-validation, respectively.

variance of the leave-one-out (LOO) cross-validation process. The good behavior of the nonlinear models describing the conformational stability of the study proteins suggests that the AASA vectors built a nonlinear vectorial space that well resembles human lysozyme stability pattern.

Concerning the possibility of change correlations, following the method used by So and Karplus in ref 38, we performed a randomization test. Randomized values were given to the dependent variable ($\Delta\Delta G$), and networks were trained using this randomized target and the real set of independent variables (optimum AASA vectors). By repeating this processes 500 times, no correlation was found between R^2 values for training and test sets, similar to the results of So and Karplus.[38]

Table 4 shows an optimum subset of nine AASA vectors and the correlation matrix of such descriptors. Variables in the model mean are as follows: AASA6ASA_D is the amino acid sequence autocorrelation of lag 6 weighted by solvent-accessible surface area for denatured protein; AASA2ΔG_c is the amino acid sequence autocorrelation of lag 2 weighted by unfolding Gibbs free energy changes of chain; AASA4*s* is the amino acid sequence autocorrelation of lag 4 weighted by shape (position of branch point in a side chain); AASA8H_t is the amino acid sequence autocorrelation of lag 8 weighted by thermodynamic transfer hydrophobicity; AASA7-$T\Delta S_h$ is the amino acid sequence autocorrelation of lag 7 weighted by unfolding entropy change of hydratation; AASA1ΔC_{ph} is the amino acid sequence autocorrelation of lag 1 weighted by unfolding hydration heat capacity change; AASA1α_N is the amino acid sequence autocorrelation of lag 1 weighted by power to be at the N-terminal of an α-helix; AASA7α_N is the amino acid sequence autocorrelation of lag 7 weighted by power to be at the N-terminal of an α-helix; and AASA9-$T\Delta S$ is the amino acid sequence autocorrelation of lag 9 weighted by unfolding entropy change. As can be observed in Table 4, only two pair of correlations appear significant ($R^2 > 0.7$): AASA6ASA_D vs AASA7-$T\Delta S_h$ and AASA1α_N vs AASA7α_N. Despite some observed intercorrelation, the adequate fitting of the data set and the LOO cross-validation obtained by such a descriptor subset reflect that relevant structural information is brought to the model by each AASA descriptor.

Interestingly, relevant amino acid/residue properties appear weighting the selected optimum AASA vectors: four thermodynamical (ΔG_c, -$T\Delta S_h$, ΔC_{ph}, and -$T\Delta S$), three structural

Table 4. Correlation Matrix of the Inputs of the Optimum Predictor BRGNN 2

	AASA6ASA$_D$	AASA2ΔG_c	AASA4*s*	AASA8H_t	AASA7-$T\Delta S_h$	AASA1ΔC_{ph}	AASA1α_N	AASA7α_N	AASA9-$T\Delta S$
AASA6ASA$_D$	1	0.092	0.519	0.149	0.724	0.115	0	0	0
AASA2ΔG_c		1	0.117	0.067	0.051	0.142	0.338	0.318	0.002
AASA4*s*			1	0.001	0.23	0.001	0.007	0.01	0.003
AASA8H_t				1	0.153	0.259	0.082	0.106	0
AASA7-$T\Delta S_h$					1	0.187	0.002	0.001	0
AASA1ΔC_{ph}						1	0.324	0.255	0.009
AASA1α_N							1	0.785	0
AASA7α_N								1	0.001
AASA9-$T\Delta S$									1

(ASA$_D$, *s*, and H_t) and one secondary structure-related (α_N) properties. Distributions on the amino acid sequence of unfolding Gibbs free energy changes of chain at lag 2; unfolding entropy changes of hydration at lag 7; unfolding hydration heat capacity change at lag 1; and unfolding entropy change at lag 9 reflect the significance of an adequate amino acid distribution at short and large ranges in the sequence, resembling certain thermodynamic pattern of human lysozymes. Shape-related amino acid property appears relevant at an autocorrelation of middle range (lag 4) on the sequence. This fact suggests that an adequate packing of protein segments of such length in the protein structure contributes to a stable folded state. Likewise, autocorrelations of solvent-accessible surface area for denatured protein and thermodynamic transfer hydrophobicity at lags 6 and 8 on the sequence could be related with the importance of having an adequate hydrophobicity-polarity distribution at the middle range on the protein tertiary structure. Furthermore, distributions at lags 1 and 7 of the power to be at the N-terminal of an α-helix should contribute to an optimum secondary structure patron that is essential for the conformational stability of human lysozyme. It is noteworthy that, among the properties here appearing as relevant for nonlinear modeling of the conformational stability of human lysozyme, the solvent-accessible surface area for denatured protein, shape (position of branch point in a side chain), unfolding entropy change of hydration, and unfolding hydration heat capacity change were reported by Gromiha et al.[13] among most linearly correlated properties with the change of unfolding Gibbs free energies for a diverse set of protein mutants.

To build a robust model we used ensembles of BRGNNs instead of a single network to calculate the $\Delta\Delta G$ values for wild-type and mutant lysozymes. This approach recently applied by us[26] consists of training several BRGNNs with different randomly partitioned training sets of 99 proteins (80% of the data) and predicting the stability of the rest of the 24 proteins (20% of the data) in the test sets. In this regard, the outputs of the trained networks were combined to form one unified prediction. As a result, the generalization ability of the neural network system is significantly improved since changing the elements that constitute test and training sets is a way to introduce diversity to the ensemble.[50] In this sense, we reported in Table 2 two calculated $\Delta\Delta G$ values for each protein: one average over training sets and another over the test sets. The optimum number of elements in the ensemble predictor was selected by studying the behavior of R^2 and RMSE of the training and test sets, respectively, vs the number of networks in the ensemble. Concerning this, Figure 2 shows plots of R^2 (A) and RMSE values for NNEs

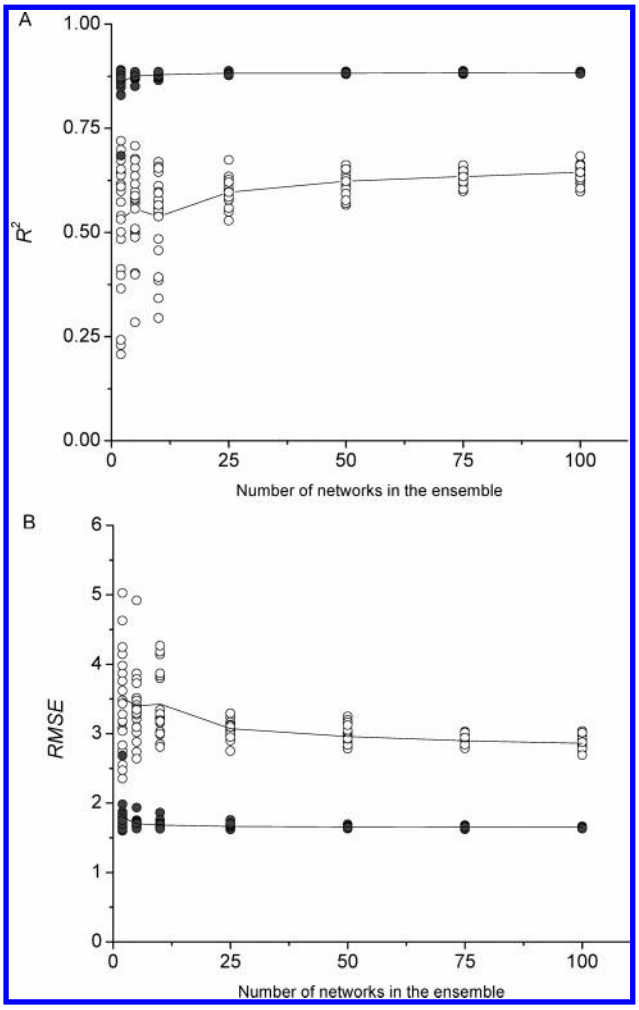

Figure 2. Plots of R^2 (A) and RMSE (B) of training (●) and test (○) sets for $\Delta\Delta G$ average values for 20 ensembles vs number of neural networks in each ensemble.

with the number of members varying from 2 to 100. As can be observed such statistical quantities remained stable for ensembles having 50 and more members. Considering this, we selected the optimum ensemble having 50 networks.

Figure 3 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 ensemble predictor. The accuracy for data fitting was about 86% and 68% for proteins in training and test sets, respectively. AASA vectors approach well fit in a nonlinear way the $\Delta\Delta G$ by means of a combination of sequence information and amino acid/residues properties. The conformational stability pattern of human lysozyme that the optimum nine vectors resembled was successfully learned by the ensemble of BRGNNs during supervised training.

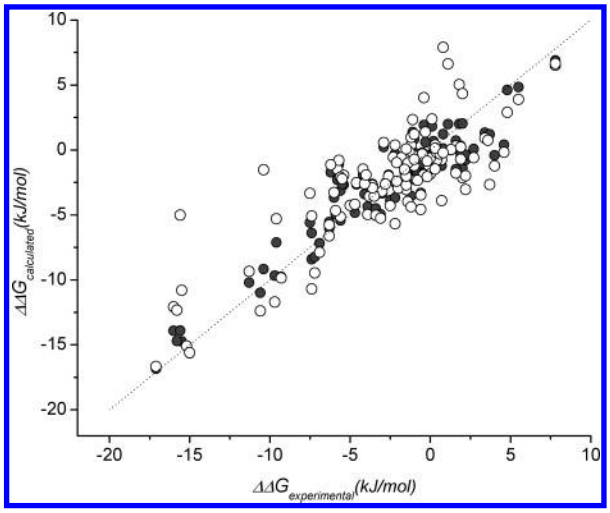

Figure 3. Plots of average calculated vs experimental change of thermal unfolding Gibbs free energy change ($\Delta\Delta G$) for human lysozymes in training (●) and test (○) sets according to a 50 member ensemble of the optimum network BRGNN 2.

To gain a deeper insight into the impact of each input in the predictor BRGNN 2 we performed a sensibility analysis following the method recently proposed by Guha and Jurs.[56] The descriptor under study was scrambled in the optimum predictor, and the Root Mean Square Errors ($RMSE_i$) between the observed and estimated stability value for all mutants were calculated. Finally, the contribution factor *Ci* of descriptor *i* (i=1−9) is given by

$$Ci = \frac{100 \times \mathrm{RMSE}_i}{\sum \mathrm{RMSE}_i} \qquad (16)$$

The results of the sensitivity analysis appear in Figure 4. As can be observed, the order of relevance for the neural network inputs is AASA2ΔG_c > AASA1ΔC_{ph} > AASA8H_t > AASA4*s* ≅ AASA7-$T\Delta S_h$ ≅ AASA7α_N > AASA6ASA_D ≅ AASA9-$T\Delta S$ ≅ AASA1α_N. However, values of *Ci* for each descriptor (Figure 4A) are rather similar suggesting a similar contribution of each descriptor to the model. Besides the analysis of the impact of each variable, we performed an analysis of the influence of each selected property to the modeling of conformational stability of human lysozyme (Figure 4B). In this connection, we accessed to the impact of the properties in the model by adding up the impacts of descriptors weighted by similar properties. Figure 4B depicts that entropy changes and the power to be at the N-terminal of a α-helix have a strong contribution to the stability pattern of human lysozyme here established. This result agrees well with reports regarding point mutations in proteins that state that reduction of entropy of the 3D structure of proteins positively affects the conformational stability.[16,57] On the other hand, the high relevance of the power to be at the N-terminal of an α-helix strongly suggested that the optimum secondary structure patron is another key factor for a stable tertiary conformation. Regarding other properties, occurrence in the model of hydrophobicity-related properties (ASA_D and H_t) is in concordance with the thermal denaturation mechanism hypothesis. For the thermal denaturation process of globular proteins, Privalov and Gill[58] stated that hydration equilibrium at high temperatures, polar interactions between solvent and polar residues in the protein, is the main cause

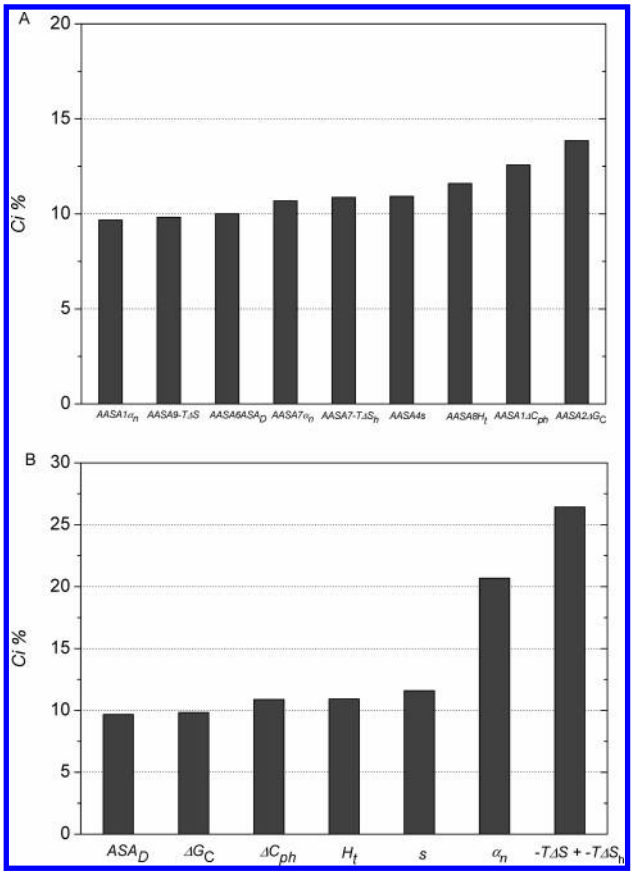

Figure 4. Relative impacts of the inputs in the optimum predictor BRGNN 2. Relative impacts of each AASA vector (A) and overall relative impacts of the amino acid/residue properties (B).

of unfolding, meanwhile hydrophobic interactions contribute to keep the folded state.

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 86% of our approach for modeling the stability of lysozyme mutants is remarkably good. In this sense, our result is about the same range of about 90% obtained by Frenz for the ANN-prediction of the relative physical stability, melting point, of a smaller set of 41 staphylococcal nuclease mutants using similarity score vectors.[18] In addition, the statistical quality of our ensemble model is in concordance with the report of Marrero-Ponce et al.[22] in which they extended topological indexes to the study of biological macromolecules. In such a report, protein linear indices of the "macromolecular pseudograph Cα-atom adjacency matrix" were applied to the prediction of melting points of Arc repressor mutants, and a nonlinear model was obtained using a piecewise multilinear equation that described about 93% of data variance.

Concerning the prediction of Gibbs free energy change of proteins, our approach overcomes previous reports in which no more than 60% of data variance was described, although models were developed for larger and more varied data sets (>1000).[12,17,20,57] AASA vectors were able to resemble amino acid interaction patterns in a human lysozyme that was learned by BRGNNs 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 only reduced

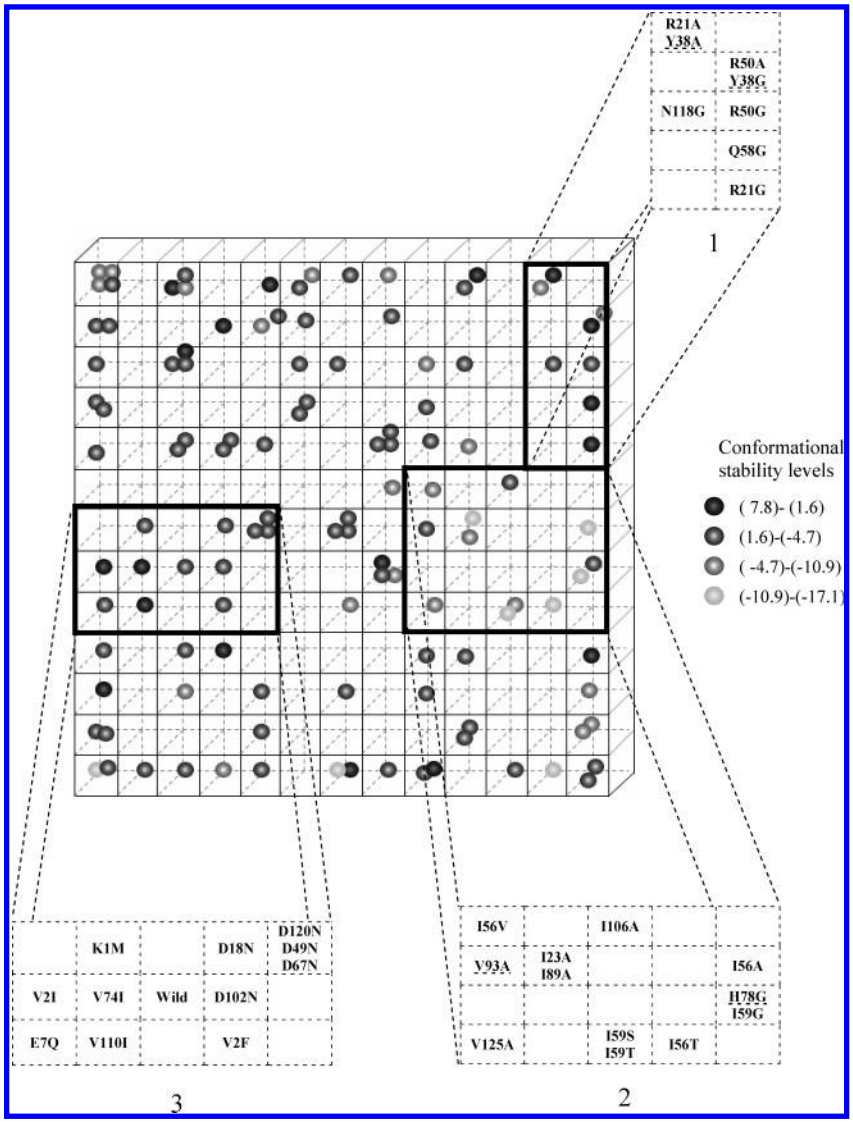

Figure 5. Kohonen Self-Organizing Map (SOM) of the change of thermal unfolding Gibbs free energy change ($\Delta\Delta G$) of human lysozyme proteins. Conformational stability legend is placed at the right-hand of the map. Underlined mutants mean wrong located mutant.

information derived from protein primary sequence. At the moment, the prediction approach presented here is protein-specific, and then one model for each protein of interest should be obtained. We gain in quality of predictions in comparison to more comprehensive models mentioned above but with lower generalization abilities. The aim of our work was just to present a reliable predictor for the conformational stability of a sole protein using its sequence and wide thermodynamic data of their mutants. Despite the disadvantage of requiring some thermodynamic experimental data for generating a training set, our modeling technique is a viable alternative for protein stability prediction when X-ray structural information is lacking but some thermodynamic data exist.

Finally, we aimed to settle some similarity among human lysozyme mutants by building a SOM of the conformational stability using the optimum subset of AASA vectors. Figure 5 depicts a 13×13 SOM of the $\Delta\Delta G$ values for the studied proteins. Eighty-four neurons were occupied for a total of 169 neurons yielding about 50% of occupancy in the map. One, 6, and 20 neurons were shared by 3, 4, and 2 proteins at the same time, respectively. As can be observed, proteins with a similar stability range were located at neighboring neurons in the map. Less stable lysozyme mutants were placed at the left-middle and left-bottom regions of the map. Otherwise, most stable mutants were distributed at right-top, left-top, and right-middle zones at the map, surrounded by middle stable mutants. By analyzing the map, some structural similarities among mutants can be addressed taking into to account their allocation at neurons with a similar level of conformational stability. In this sense, stabilizing mutations of bulky residues by small residues (alanine and glycine) were located at the right-top region, zone 1 in the map. Those mutants can accommodate the mutated residue in the main-chain conformation without an unfavorable energetic cost. On the other hand, another interesting region is denoted as zone 2 on the map. Most of the mutants in this region correspond to low stable mutations, specifically change of interior isoleucines that can reduce hydrophobic interactions in the protein core decreasing protein rigidity. Finally, we highlighted the region where wild-type lysozyme was allocated, zone 3. At neighboring neurons appeared two kinds of mutants, in which one residue was changed for a similar amino acid such as valine (V) by isoluecine (I), glutamic acid (E) by glutamine (Q), aspartic acid (D) by asparagine (N), and another in which mutations were carried

out at the initial part of the sequence (K1M, V2F, V2I). Allocation of those mutants in the same neighborhood that wild-type lysozyme is expected since such mutants are quite similar to wild-type protein.

CONCLUSIONS

Protein structures are stabilized by numerous intramolecular interactions such as hydrophobic, electrostatic, van der Waals, and hydrogen bond. Stability changes induced by mutations have been analyzed by various computational methods, but most of them require X-ray structural analysis and have limited prediction accuracy. This fact makes it useful to have simpler methods for predicting the mutation-induced stability changes.

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 a structure−properties relationship approach for protein modeling. We extended the concept of autocorrelation vectors in molecules to the amino acid sequence of proteins as a tool for encoding protein structural information for supervised training of ANNs. In this sense, novel Amino Acid Sequence Autocorrelation (AASA) vectors were obtained by calculating autocorrelations on the protein primary structure of 48 amino acid/residue properties selected from the AAindex database. BRGNNs showed again to be a powerful technique for feature selection and mathematical modeling. This approach yielded a reliable and robust nine-input ensemble model for the conformational stability of human lysozyme mutants that describes about 86% and 68% of training and test set variances. Furthermore, conformational similarities among mutants were addressed analyzing a SOM built with the subset of AASA vectors in the optimum BRGNN predictor.

The present work demonstrates the successful application of the AASA vectors to the modeling of protein conformational stability in combination with the BRGNN approach. Encoding amino acid properties and protein primary structure information on the same pool of descriptors are more appropriate than other approaches considering only amino acid substitution information. This approach leads to a powerful method for the scientific community interested in protein prediction studies. Despite the fact that one model per protein is required according to the approach present here, a general model encompassing large and varied mutant data (>2000) as well as protein-specific models for other proteins will be developed by our group shortly.

REFERENCES AND NOTES

(1) Saven, J. Combinatorial Protein Design. *Curr. Opin. Struct. Biol.* **2002**, *12*, 453−458.

(2) Mendes, J.; Guerois, R.; Serrano, L. Energy Estimation in Protein Design. *Curr. Opin. Struct. Biol.* **2002**, *12*, 441−446.

(3) Bolon, D. N.; Marcus, J. S.; Ross S. A.; Mayo, S. L. Prudent Modeling of Core Polar Residues in Computational Protein Design. *J. Mol. Biol.* **2003**, *329*, 611−622.

(4) Looger, L. L.; Dwyer, M. A.; Smith, J. J.; Helling, H. W. Computational Design of Receptor and Sensor Proteins with Novel Functions. *Nature* **2003**, *423*, 185−190.

(5) Dang, L. X.; Merz; K. M.; Kollman, P. A. Free-energy Calculations on Protein Stability: Thr-1573Val-157 Mutation of T4 Lysozyme. *J. Am. Chem. Soc.* **1989**, *111*, 8505−8508.

(6) Lazaridis, T., Karplus, M. Effective energy functions for protein structure prediction. *Curr. Opin. Struct. Biol*. **2000**, *10*, 139−145.

(7) Lee, C.; Levitt, M. Accurate Prediction of the Stability and Activity Effects of Site- Directed Mutagenesis on a Protein Core. *Nature* **1991**, *352*, 448−451.

(8) Lee, C. Testing Homology Modeling on Mutant Proteins: Predicting Structural and Thermodynamic Effects in the Ala98-Val Mutants of T4 Lysozyme. *Fold. Des.* **1995**, *1*, 1−12.

(9) Topham, C. M.; Srinivasan, N.; Blundell, T. L. Prediction of the Stability of Protein Mutants Based on Structural Environmentdependent Amino Acid Substitution and Propensity Tables. *Protein Eng.* **1997**, *10*, 7−21.

(10) Gilis, D.; Rooman, M. Prediction of Stability Changes upon Single site Mutations Using Database-Derived Potentials. *Theor. Chem. Acc.* **1999**, *101*, 46−50.

(11) Lacroix, E.; Viguera; A. R.; Serrano, L. Elucidating the folding problem of alpha-helices: local motifs, long-range electrostatics, ionic-strength dependence and prediction of NMR parameters. *J. Mol. Biol.* **1998**, *284*, 173−191. (b) Munoz, V.; Serrano, L. Development of the multiple sequence approximation within the AGADIR model of alpha-helix formation: comparison with Zimm-Bragg and Lifson-Roig formalisms. *Biopolymers* **1997**, *41*, 495−509.

(12) Guerois, R.; Nielsen, J. E.; Serrano, L. Predicting Changes in the Stability of Proteins and Protein Complexes: A Study of More Than 1000 Mutations. *J. Mol. Biol.* **2002**, *320*, 369−387.

(13) Gromiha, M. M.; Oobatake, M.; Kono, H.; Uedaira, H.; Sarai, A. Relationship Between Amino Acid Properties and Protein Stability: Buried Mutations. *J. Protein Chem.* **1999**, *18*, 565−578.

(14) Gromiha, M. M.; Oobatake, M.; Kono, H.; Uedaira, H.; Sarai, A. Role of Structural and Sequence Information in the Prediction of Protein Stability Changes: Comparison between Buried and Partially Buried Mutations. *Protein. Eng.* **1999**, *12*, 549−555.

(15) Gromiha, M. M.; Oobatake, M.; Kono, H.; Uedaira, H.; Sarai, A. Importance of Surrounding Residues for Protein Stability of Partially Buried Mutations. *J. Biomol. Struct. Dyn.* **2000**, *18*, 1−16.

(16) (a) Takano, K.; Ogasahara, K.; Kaneda, H.; Yamagata, Y.; Fujii, S.; Kanaya, E.; Kikuchi, M., Oobatake, M.; Yutani, K. Contribution of Hydrophobic Residues to the Stability of Human Lysozyme: Calorimetric Studies and X-ray Structural Analysis of the Five Isoleucine to Valine Mutants. *J. Mol. Biol*. **1995**, *254*, 62−76. (c) Takano, K.; Yamagata, Y.; Fujii, S.; Yutani, K. Contribution of the Hydrophobic Effect to the Stability of Human Lysozyme: Calorimetric Studies and X-ray Structural Analyses of the Nine Valine to Alanine Mutants. *Biochemistry* **1997**, *36*, 688−698. (d) Takano, K.; Funahashi, J.; Yamagata, Y.; Fujii, S.; Yutani, K. Contribution of Water Molecules in the Interior of a Protein to the Conformational Stability. *J. Mol. Biol.* **1997**, *274*, 132−142. (e) Takano, K.; Yamagata, Y.; Yutani, K. A General Rule for the Relationship Between Hydrophobic Effect and Conformational Stability of a Protein: Stability and Structure of a Series of Hydrophobic Mutants of Human Lysozyme. *J. Mol. Biol.* **1998**, *280*, 749−761. (f) Yamagata, Y.; Kubota, M.; Sumikawa, Y.; Funahashi, J.; Takano, K.; Fujii, S.; Yutani, K. Contribution of Hydrogen Bonds to the Conformational Stability of Human Lysozyme: Calorimetry and X-ray Analysis of Six Tyrosine Phenylalanine Mutants. *Biochemistry* **1998**, *37*, 9355−9362. (g) Takano, K.; Yamagata, Y.; Kubota, M.; Funahashi, J.; Fujii, S.; Yutani, K. Contribution of Hydrogen Bonds to the Conformational Stability of Human Lysozyme: Calorimetry and X-ray Analysis of Six Ser → Ala Mutants. *Biochemistry* **1999**, *38*, 6623−6629. (h) Takano, K.; Yamagata, Y.; Funahashi, J.; Hioki, Y.; Kuramitsu, S.; Yutani, K. Contribution of Intra- and Intermolecular Hydrogen Bonds to the Conformational Stability of Human Lysozyme. *Biochemistry* **1999**, *38*, 12698−12708. (i) Funahashi, J.; Takano, K.; Yamagata, Y.; Yutani, K. Contribution of Amino Acid Substitutions at Two Different Interior Positions to the Conformational Stability of Human Lysozyme. *Protein Eng.* **1999**, *12*, 841−850. (j) Takano, K.; Ota, M.; Ogasahara, K.; Yamagata, Y.; Nishikawa, K.; Yutani, K. Experimental verification of the "stability profile of mutant protein" (SPMP) data using mutant human lysozymes. *Protein Eng.* **1999**, *12*, 663−672. (k) Takano, K.; Tsuchimori, K.; Yamagata, Y.; Yutani, K. Contribution of Salt Bridges near the Surface of a Protein to the Conformational Stability. *Biochemistry* **2000**, *39*, 12375−12381. (l) Funahashi, J.; Takano, K.; Yamagata, Y.; Yutani, K. Role of Surface Hydrophobic Residues in the Conformational Stability of Human Lysozyme at Three Different Positions. *Biochemistry* **2000**, *39*, 14448−14456. (m) Takano, K.; Yamagata, Y.; Yutani, K. Contribution of Polar Groups in the Interior of a Protein to the Conformational Stability. *Biochemistry* **2001**, *40*, 4853−4858.

(17) Zhou, H.; Zhou, Y. Stability Scale and Atomic Solvation Parameters Extracted From 1023 Mutation Experiment. *Proteins* **2002**, *49*, 483−492.

(18) Frenz, C. M. Neural Network-Based Prediction of Mutation-Induced Protein Stability Changes in Staphylococcal Nuclease at 20 Residue Positions. *Proteins* **2005**, *59*, 147−151.

(19) Levin, S.; Satir, B. H. POLINA: Detection and Evaluation of Single Amino Acid Substitutions in Protein Superfamilies. *Bioinformatics* **1998**, *14*, 374−375.
(20) (a) Capriotti, E.; Fariselli, P.; Casadio, R. A neural-network-based method for predicting protein stability changes upon single mutations. *Bioinformatics* **2004**, *20*, 63−68. (b) Capriotti, E.; Fariselli, P.; Calabrese, R.; Casadio, R. Prediction of protein stability changes from sequences using support vector machines. *Bioinformatics* **2005**, *21*, 54−58.
(21) Ramos de Armas, R.; González-Díaz, H.; Molina, R.; Uriarte E.; Markovian Backbone Negentropies: Molecular Descriptors for Protein Research. I. Predicting Protein Stability in Arc Repressor Mutants. *Proteins* **2004**, 56, 715−723.
(22) Marrero-Ponce, Y.; Medina-Marrero, R.; Castillo-Garit, J. A.; Romero-Zaldivar, V.; F. Torrens; Castro, E. A. 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.* **2005**, *13*, 3003−3015.
(23) Guha, R.; Jurs, P. C. Development of Linear, Ensemble, and Nonlinear Models for the Prediction and Interpretation of the Biological Activity of a Set of PDGFR Inhibitors. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2179−2189.
(24) Fernández, M.; Caballero, J.; Helguera, A. M.; Castro, E. A.; González, M. P. Quantitative Structure−Activity Relationship to Predict Differential Inhibition of Aldose Reductase by Flavonoid Compounds. *Bioorg. Med. Chem.* **2005**, *13*, 3269−3277.
(25) Fernández, M.; Tundidor-Camba, A.; Caballero, J. 2D Autocorrelation Modeling of the Activity of Trihalobenzocycloheptapyridine Analogues as Farnesyl Protein Transferase Inhibitors, *Mol. Simul.* **2005**, *31*, 575−584.
(26) Fernández, M.; Tundidor-Camba, A.; Caballero, J. Modeling of Cyclin-Dependent Kinase Inhibition by 1*H*-pyrazolo [3,4-*d*] pyrimidine Derivatives using Artificial Neural Networks Ensembles. *J. Chem. Inf. Comput. Sci*. **2005**, *45*, 1884−1895.
(27) González, M. P.; Caballero, J.; Tundidor-Camba, A.; Helguera, A. M.; Fernández, M. Modeling of Farnesyltransferase Inhibition by some Thiol and Non-Thiol Peptidomimetic Inhibitors using Genetic Neural Networks and RDF Approaches. *Bioorg. Med. Chem.* **2006**, *14*, 200−213.
(28) Fernández, M.; Caballero, J. Modeling of Activity of Cyclic Urea HIV-1 Protease Inhibitors using Regularized-Artificial Neural Networks. *Bioorg. Med. Chem.* **2006**, *14*, 280−294.
(29) Caballero, J.; Fernández, M. Linear and Nonlinear Modeling of Antifungal Activity of Some Heterocyclic Ring Derivatives using Multiple Linear Regression and Bayesian-Regularized Neural Networks. *J. Mol. Model.* **2006**, *12*, 168−181.
(30) Moran, P. A. P. Notes on Continuous Stochastic Processes. *Biometrika* **1950**, *37*, 17−23.
(31) Geary, R. F. The contiguity ratio and statistical mapping. *The Incorporated Statistician* **1954**, *5*, 115−145.
(32) Moreau, G.; Broto, P. Autocorrelation of a topological structure: A new molecular descriptor. *Nouv. J. Chim*. **1980**, *4*, 359−360.
(33) Moreau, G.; Broto, P. Autocorrelation of Molecular Structures: Application to SAR Studies. *Nouv. J. Chim*. **1980**, *4*, 757−764.
(34) Wagener, M.; Sadowski, J.; Gasteiger, J. Autocorrelation of Molecular Properties for Modelling Corticosteroid Binding Globulin and Cytosolic Ah Receptor Activity by Neural Networks. *J. Am. Chem. Soc*. **1995**, *117*, 7769−7775.
(35) Bauknecht, H.; Zell, A.; Bayer, H.; Levi, P.; Wagener, M.; Sadowski, J.; Gasteiger, J. Locating Biologically Active Compounds in Medium-Sized Heterogeneous Datasets by Topological Autocorrelation Vectors: Dopamine and Benzodiazepine Agonists. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1205−1213.
(36) (a) Nakai, K.; Kidera, A.; Kanehisa, M. Cluster analysis of amino acid indices for prediction of protein structure and function. *Protein Eng.* **1988**, *2*, 93−100. (b) Tomii, K.; Kanehisa, M. Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Eng.* **1996**, *9*, 27−36. (c) Kawashima, S.; Kanehisa, M. AAindex: amino acid index database. *Nucleic Acids Res.* **2000**, *28*, 374−374.
(37) MATLAB 7.0. program, available from The Mathworks Inc., Natick, MA. http://www.mathworks.com.
(38) So, S.; Karplus, M. Evolutionary Optimization in Quantitative Structure−Activity Relationship: An Application of Genetic Neural Networks. *J. Med. Chem.* **1996**, *39*, 1521−1530.
(39) (a) Burden, F. R.; Winkler, D. A. Robust QSAR Models Using Bayesian Regularized Neural Networks. *J. Med. Chem.* **1999**, *42*, 3183−3187. (b) Winkler, D. A.; Burden, F. R. Bayesian neural nets for modeling in drug discovery. *Biosilico* **2004**, *2*, 104−111.
(40) Zupan, J.; Gasteiger, J. Neural networks: a new method for solving chemical problems or just a passing fase? *Anal. Chim. Acta* **1991**, *248*, 1−30.
(41) Aoyama, T.; Suzuki, Y.; Ichikawa, H. Neural Networks applied to Structure−Activity Relationships. *J. Med. Chem.* **1990**, *33*, 905−908.
(42) (a) Mackay, D. J. C. Bayesian Interpolation. *Neural Comput.* **1992**, *4*, 415−447. (b) Mackay, D. J. C. A practical Bayesian Framework for Backprop Networks. *Neural Comput.* **1992**, *4*, 448−472.
(43) Lampinen, J.; Vehtari, A. Bayesian Approach for Neural Networks − Review and Case Studies. *Neural Networks* **2001**, *14*, 7−24.
(44) Foresee, F. D.; Hagan M. T. *Gauss−Newton approximation to Bayesian learning.* Proceedings of the 1997 International Joint Conference on Neural Networks, 1997; IEEE: Houston, pp 1930−1935.
(45) Holland, H. *Adaption in natural and artificial systems*; The University of Michigan Press: Ann Arbor, MI, 1975.
(46) Cartwright, H. M. *Applications of artificial intelligence in chemistry*; Oxford University Press: Oxford, 1993.
(47) Hemmateenejad, B.; Safarpour, M. A.; Miri, R., Nesari, N. Toward an Optimal Procedure for PC-ANN Model Building: Prediction of the Carcinogenic Activity of a Large Set of Drugs. *J. Chem. Inf. Model.* **2005**, *45*, 190−199.
(48) Hansen, L. K.; Salamon, P. Neural Network Ensembles. *IEEE Trans. Pattern Anal. Machine Intell.* **1990**, *12*, 993−1001.
(49) Krogh, A.; Vedelsby, J. Neural network ensembles, cross-validation and active learning. In *Advances in Neural Information Processing Systems 7*; Tesauro, G., Touretzky, D., Lean, T., Eds.; MIT Press: 1995; pp 231−238.
(50) Agrafiotis, D. K.; Cedeño, W.; Lobanov, V. S. On the Use of Neural Network Ensembles in QSAR and QSPR. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 903−911.
(51) Baumann, K. Chance Correlation in Variable Subset Regression: Influence of the Objective Function, the Selection Mechanism, and Ensemble Averaging. *QSAR Comb. Sci.* **2005**, *24*, 1033−1046.
(52) Yan, A.; Gasteiger, J.; Krug, M.; Anzali. S. Linear and Nonlinear Functions on Modeling of Aqueous Solubility of Organic Compounds by Two Structure Representation Methods. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 75−87.
(53) de-Sousa, J. A.; Gasteiger, J. New Description of Molecular Chirality and Its Application to the Prediction of the Preferred Enantiomer in Stereoselective Reactions. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 369−375.
(54) Kohonen, T. Self-organized Formation of Topologically Correct Feature Maps. *Biol. Cybern.* **1982**, *43*, 59−69.
(55) Bava, K. A.; Gromiha, M. M.; Uedaira, H.; Kitajima, K.; Sarai, A. ProTherm, version 4.0: Thermodynamic Database for Proteins and Mutants. *Nucleic Acids Res.* **2004**, *32*, 120−121. http://gibk26.bse.kyutech.ac.jp/jouhou/protherm/protherm.html.
(56) Guha, R.; Jurs, P. C. Interpreting Computational Neural Network QSAR Models: A Measure of Descriptor Importance. *J. Chem. Inf. Model.* **2005**, *45*, 800−806.
(57) Bordner, A. J.; Abagyan, R. A. Large-Scale Prediction of Protein Geometry and Stability Changes for Arbitrary Single Point Mutations. *Proteins* **2004**, *57*, 400−413.
(58) Privalov, P. L.; Gill, S. J. Stability of Protein Structure and Hydrophobic Interaction. *Adv. Prot. Chem.* **1988**, *39*, 191−234.

CI050507Z