PROTEINS: Structure, Function, and Bioinformatics 67:834–852 (2007)

Amino Acid Sequence Autocorrelation Vectors and Bayesian-Regularized Genetic Neural Networks for Modeling Protein Conformational Stability: Gene V Protein Mutants

Leyden Fernández,[1] Julio Caballero,[1] José Ignacio Abreu,[1,2] and Michael Fernández[1]*
[1]*Molecular Modeling Group, Center for Biotechnological Studies, Faculty of Agronomy, University of Matanzas, 44740 Matanzas, Cuba*
[2]*Artificial Intelligence Lab, Faculty of Informatics, University of Matanzas, 44740 Matanzas, Cuba*

ABSTRACT **Development of novel computational approaches for modeling protein properties from their primary structure is the 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 data base. A total of 720 *AASA* descriptors were tested for building predictive models of the change of thermal unfolding Gibbs free energy change ($\Delta\Delta G$) of gene V protein upon mutation. 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 66% variance of the data in training and test sets respectively. Furthermore, the optimum *AASA* vector subset not only helped to successfully model unfolding stability but also well distributed wild-type and gene V protein mutants on a stability self-organized map (SOM), when used for unsupervised training of competitive neurons. Proteins 2007;67:834–852.** © 2007 Wiley-Liss, Inc.

Key words: protein stability prediction; point mutations; Bayesian regularization; artificial neural networks; genetic algorithm

INTRODUCTION

Evidence is accumulating that many disease-causing mutations exert their effects by altering protein folding. 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 accurate 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 in 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 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, EEEF approach combines a physical description of the interactions with some data obtained from experiments previously ran on proteins. Examples of such algorithms are the helix/coil transition algorithm AGADIR[11,12] or FOLDEF, a fast and accurate EEEF approach based on AGADIR algorithm that uses a

*Correspondence to: Michael Fernández, Molecular Modeling Group, Center for Biotechnological Studies, Faculty of Agronomy, University of Matanzas, 44740 Matanzas, Cuba.
E-mail: michael.fernandez@umcc.cu

Received 17 March 2006; Revised 28 September 2006; Accepted 8 November 2006

Published online 21 March 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21349

full atomic description of the structure of the proteins, which was reported by Guerois et al.[13] for predicting conformational stability of more than 1000 mutants.

Furthermore, other stability prediction studies not based on protein force-field calculations have been focused on correlations of free energy change with structural, sequence information and amino acid properties such as hydrophobicity, accessible surface area, etc. In this sense, Gromiha et al. had reported some of the seminal works in this topic.[14–16] 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.[14] 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 properties but at 3D structure neighboring residues at specific radius from the mutation points. In such studies, it was reported that the role of structural and sequence information in the prediction of protein stability changes by comparing buried and partially buried mutations.[15] 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 highest segment length effects for helical, strand, and coil mutations are, respectively, 0, 9, and 4 residues on both sides of the mutant residues.[16]

On the other hand, empirical equations involving physical properties calculated from mutant structures have been reported. Several studies concerning mutations on human lysozymes[17–28] 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[29] 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.[1–13] Despite the size of protein crystallographic data base that 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.[30] In this context, Levin and Satir[31] successfully evaluated the functional significance of mutations on hemoglobin using amino acid similarity matrixes. Recently, Frenz[30] 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.[32–34] 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 the qualitative and quantitative predicting of 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.[35,36] 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 pseudo-graph of Cα atoms. In these reports, protein stability studies, on specifically how alanine substitution mutation on 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 like MRA or partial least square.[37–42] 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, stability of gene V protein mutants was successfully modeled from their amino acid sequences. The structure of the gene V protein is interesting because its small size and the large number of mutants available have made it a useful model for determining the effects of amino acid substitutions on protein stability and function. We attempted to predict gene V protein conformational stability by extending the concept of structural autocorrelation vectors[43–48] in molecules to protein primary structure. Protein structure information was encoded by means of amino acid sequence autocorrelation (*AASA*) vectors weighted by 48 physicochemical, energetic, and conformational amino acid/residues prop-

erties extracted from the AAindex amino acid database.[49–51] In this way, a large set of descriptors was computed and by employing a nonlinear modeling technique recently employed by our group, Bayesian-regularized genetic neural networks (BRGNNs),[39–43,49] optimum ANN-based predictive models of conformational stability were built with reduced subset of variables. To provide robust models, we employed data-diverse ensembles of BRGNN for calculating the conformational stability. In addition to the regression model, we built a self-organizing map (SOM) of gene V protein conformational stabilities using the inputs of the optimum BRGNN predictor for unsupervised training of competitive neurons.

EXPERIMENTAL

Amino Acid Sequence Autocorrelation vector approach

Conformational stability of a protein depends on a variety of intramolecular interactions such as hydrophobic, electrostatic, van der Waals, and hydrogen-bond that are ruled by the amino acid sequence. Therefore, in structure-property/activity studies the strategy for encoding structural information must, 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, and so the structural encoding approaches must allow comparing such structures.[48]

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, and so they are canonical. And third, the length of the correlation vector is independent of the size of the molecule.[48]

For the autocorrelation vectors in molecules, H-depleted molecular structure is represented as a graph and physico-chemical properties of atoms as real values are 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[43–45] has been successfully used in the last decades for modeling biological activities[45,46] and pharmaceutical research.[47,48] In recent works, our group has obtained outstanding results when such chemical code was used in combination with ANN approach in biological QSAR studies.[38,41] 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[45] is defined as follows:

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

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

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

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 the protein primary structure as a linear graph with nodes formed by amino acid residues. Autocorrelation approach in protein stability studies mainly differs from the method of Gromiha et al.,[16] 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 whole protein. Particularly, amino acid sequence autocorrelation (*AASA*) vectors of lag l are calculated as follows:

$$AASAlp_k = \frac{1}{L} \sum_i \delta_{ij} p_{ki} p_{kj} \quad (3)$$

where $AASAlp_k$ 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:

$$AASA1p_k = \frac{1}{9}(p_{kA}\,p_{kS} + p_{kS}\,p_{kT} + p_{kT}\,p_{kC} + p_{kC}\,p_{kG} + p_{kG}\,p_{kF} + p_{kF}\,p_{kH} + p_{kH}\,p_{kC} + p_{kC}\,p_{kS} + p_{kS}\,p_{kD}) \quad (4)$$

$$AASA5p_k = \frac{1}{5}(p_{kA}\,p_{kF} + p_{kS}\,p_{kH} + p_{kT}\,p_{kC} + p_{kC}\,p_{kS} + p_{kG}\,p_{kD}) \quad (5)$$

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.[43,44] 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, they used 48 physicochemical, energetic, and conformational amino acid/residues properties (Table I) selected by Gromiha et al.[14]

TABLE I. Numerical Values of 48 Selected Physicochemical, Energetic, and Conformational Properties of the 20 amino acids/residues[10]

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	μ	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	ΔGh	−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		
Property[a,b]		L	M	N	P	Q	R	S	T	V	W	Y
1	K_0	−31.78	−31.18	−30.9	−23.25	−32.6	−26.62	−29.88	−31.23	−30.62	−30.24	−35.01
2	H_t	2.17	1.67	0.09	2.77	0	0.85	0.07	0.07	1.87	3.77	2.67
3	H_P	14.19	13.62	11.72	11.06	11.78	12.4	11.68	12.12	14.73	13.96	13.57
4	P	0.13	1.43	3.38	1.58	3.53	52	1.67	1.66	0.13	2.1	1.61
5	pH_i	5.98	5.74	5.41	6.3	5.65	10.76	5.68	5.66	5.96	5.89	5.66
6	pK'	2.36	2.28	2.02	1.99	2.17	1.81	2.21	2.1	2.32	2.38	2.2
7	M_w	131	149	132	115	146	174	105	119	117	204	181
8	P_1	21.4	16.25	12.82	17.43	14.45	14.28	9.47	15.77	21.57	21.61	18.03
9	R_f	17.6	14.7	5.4	14.8	9	4.6	6.9	9.5	14.3	17	15
10	μ	18.78	21.64	13.28	10.93	17.56	26.66	6.35	11.01	13.92	42.53	31.55
11	H_{nc}	1.06	0.64	−0.78	0.12	−0.85	−2.53	−0.18	−0.05	1.08	0.81	0.26

TABLE I. (Continued)

Property[a,b]		L	M	N	P	Q	R	S	T	V	W	Y
12	E_{sm}	1.32	1.3	1.18	1.24	1.12	0.92	1.3	1.25	1.25	1.03	1.03
13	E_1	0.65	0.65	0.38	0.46	0.4	0.55	0.45	0.52	0.73	0.83	0.65
14	E_t	1.97	1.96	1.56	1.7	1.52	1.48	1.75	1.77	1.98	1.87	1.69
15	P_α	1.21	1.45	0.67	0.57	1.11	0.98	0.77	0.83	1.06	1.08	0.69
16	P_β	1.3	1.05	0.89	0.55	1.1	0.93	0.75	1.19	1.7	1.37	1.47
17	P_t	0.59	0.6	1.56	1.52	0.98	0.95	1.43	0.96	0.5	0.96	1.14
18	P_C	0.69	0.59	1.37	1.61	0.87	1.07	1.34	1.08	0.63	0.76	1.07
19	C_a	35	43	28	22	36	55	20	28	33	61	46
20	F	0.79	0.78	1.04	1.16	1.07	1.05	1.13	0.96	0.79	0.77	1.01
21	P_r	0.5	0.42	0.15	0.18	0.11	0.07	0.23	0.23	0.48	0.4	0.26
22	R_a	5.88	5.21	2.12	2.12	2.7	2.53	2.43	2.6	7.14	6.25	3.03
23	N_s	7.37	6.39	5.04	5.65	5.45	5.7	5.53	5.81	7.62	6.98	6.73
24	α_n	1.91	1.25	0.53	0	0.98	0.67	0.7	0.75	1.42	1.33	0.58
25	α_c	0.58	0.73	0.93	2.19	1.2	0.39	0.81	1.25	0.63	1.4	0.72
26	α_m	1.05	1.47	0.93	0	1.63	1.59	0.87	0.46	1.2	0.46	0.52
27	V^0	107.75	105.35	78.01	82.83	93.9	127.34	60.62	76.83	90.78	143.91	123.6
28	N_m	2.19	2.27	1.84	1.32	2.03	1.94	1.57	1.57	1.63	1.9	1.67
29	N_1	4.59	4.14	3.64	3.57	3.06	3.78	3.75	4.09	5.43	4.83	4.93
30	H_{gm}	14.13	13.86	13.02	12.35	12.61	13.1	13.39	12.7	14.56	15.48	13.88
31	ASA_D	161.4	189.8	134.9	135.1	164.9	210.2	111.4	130.4	143.9	208.8	196.4
32	ASA_N	31.1	41.3	60.5	60.7	71.5	94.5	48.7	52	28.1	39.5	50.4
33	ΔASA	129.8	147.9	74	73.5	93.3	116	62.8	78	115.6	167.8	145.9
34	ΔGh	0.27	−0.6	−3.55	0.32	−3.92	−5.96	−3.82	−1.97	0.13	−3.8	−5.64
35	G_{hD}	0.35	−0.71	−6.63	0.56	−7.12	−12.78	−6.18	−3.66	0.18	−4.71	−8.45
36	G_{hN}	0.07	−0.1	−3.03	0.23	−3.15	−6.85	−2.36	−1.69	0.04	−0.88	−2.82
37	ΔH_h	−3.52	−4.16	−5.68	−1.95	−6.23	−10.43	−5.94	−4.39	−3.15	−8.99	−10.67
38	$-T\Delta S_h$	3.79	3.56	2.13	2.27	2.31	4.47	2.12	2.42	3.28	5.19	5.03
39	ΔC_{ph}	38.26	31.67	3.91	23.69	3.74	16.66	6.14	16.11	32.58	37.69	30.54
40	ΔG_c	−0.35	1.13	3.26	−0.39	3.69	5.25	3.42	1.74	−0.19	5.59	6.56
41	ΔH_c	3.69	7.06	3.64	1.97	4.47	6.03	5.8	4.42	3.45	13.46	14.41
42	$-T\Delta S_c$	−4.04	−5.93	−0.39	−2.36	−0.78	−0.78	−2.38	−2.68	−3.64	−7.87	−7.95
43	ΔG	−0.08	0.53	−0.3	−0.06	−0.23	−0.71	−0.4	−0.24	−0.06	1.78	0.91
44	ΔH	0.17	2.89	−2.03	0.02	−1.76	−4.4	−0.16	0.04	0.3	4.47	3.73
45	$-T\Delta S$	−0.24	−2.36	1.74	−0.08	1.53	3.69	−0.24	−0.28	−0.36	−2.69	−2.82
46	V	4	4	4	3	5	7	2	3	3	10	8
47	s	2	0	2	0	3	5	0	1	1	2	2
48	f	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; M_w, 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; B_r, buriedness; R_a, solvent-accessible reduction ratio; N_s, 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 nonhydrogen 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.

from the AAindex data base[50–52] 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, ranged from 1 to 15, to access long range interactions in the sequence due to tertiary structure arrangements. Computational code for *AASA* vector calculation was written in Matlab environment.[53] A data matrix of 720 *AASA* vectors, 48 properties × 15 different lags, were generated with the autocorrelation vectors calculated for each gene V protein 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 260 descriptors were obtained. Afterwards, optimum predictive models were built with reduced subsets of variables by means of BRGNN algorithm.

Bayesian-Regularized Genetic Neural Networks 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.[54–56]

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[54] incorporating Bayesian regularization that has been successfully introduced by our group for modeling the inhibitory activity of several therapeutic target enzymes.[39–42] BRGNN was programmed within Matlab environment[53] using genetic algorithm and neural networks toolboxes. BRGNN technique leads to neural networks trained with optimum inputs selected from the whole *AASA* vector data matrix (see Fig. 1).

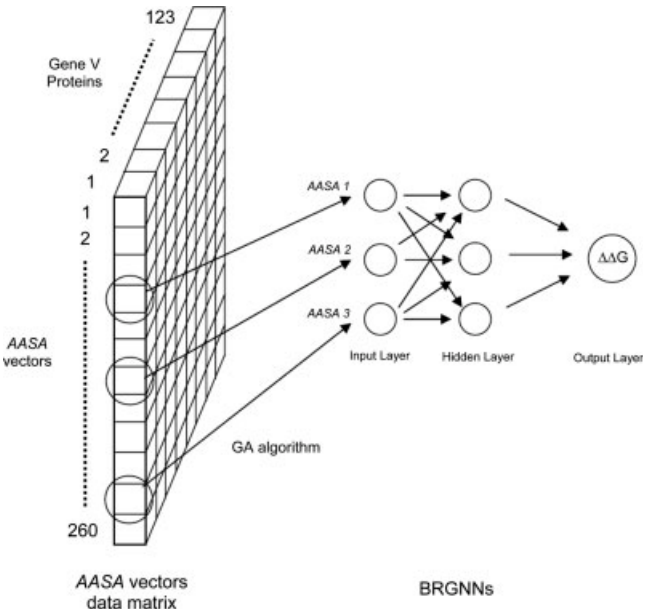


Fig. 1. Schematic representation of Bayesian-Regularized Genetic Neural Network (BRGNN) technique with a prototype back-propagation neural network with 3-3-1 architecture. *AASA* vectors chosen by genetic algorithm constitute inputs and network is trained against change of unfolding Gibbs free energy change ($\Delta\Delta G$) of gene V protein mutants.

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 "layers."[57,58] 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.[57] The employed transfer function, commonly hyperbolic tangent function, allows to establish 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, network loses its ability to generalize. Error on the training set is driven to a very small value, but when new data is 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 over-fitting.[59,60] To accomplish this purpose, BRANNs include an error term that regularizes the weights by penalizing overly large magnitudes.

Assuming a set of pairs $D = \{x_i, t_i\}$, where $i = 1, \ldots, N$ 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:

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

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

For completing the interpolation model, a prior probability distribution must be defined which embodies our prior knowledge on the sort of mappings that are "reasonable."[61] 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, a prior is defined as that which expresses the sort of smoothness the interpolant is expected to have. The model has a prior of the form:

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

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:

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

where $P(w \mid D,\alpha,\beta,\mathrm{M})$ is the posterior probability, that is the plausibility of a weight distribution considering the information of the data set in the model used, $P(w \mid \alpha,M)$ is the prior density, which represents our knowledge of the weights before any data is collected, $P(D \mid w,\beta,M)$ is the likelihood function, which is the probability of the data occurring, given the weights and $P(D \mid \alpha,\beta,\mathrm{M})$ 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:

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

where Z_F depends on objective function parameters. So under this framework, minimization of F is identical to find the (locally) most probable parameters.[59,60]

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.[59,60]

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.[38–42,55,56]

Fully connected, three-layer BRANNs with back-propagation training were implemented in MATLAB environment.[46] 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.[62] 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}.

Genetic algorithm

GAs are governed by biological evolution rules.[63] 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.[64] 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 (260 rows $\times$ 123 columns) that will be tested as BRANN inputs. So and Karplus,[54] used a variety of fitness functions which are proportional to the residual error of the training set, the test set, or even the crossvalidation 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 the 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.

Similar to the study of So and Karplus,[54] our study 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.[65]

Different to the 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 than a threshold MSE value, which the Bayesian regularization guarantees to posses good generalization abilities (see Fig. 1). This is because we used MSE of data training fitting instead of 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, an optimum generalizable network with the highest predictive power can be derived. This process also avoids chance correlations. This approach has 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.

Artificial neural network ensembles

Artificial neural network ensemble (NNE) is a learning paradigm where many ANNs are jointly used to solve a problem. On the basis of this judgement, 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.[66]

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[67] 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 (over-fitting, under-fitting, training with different topologies and/or training parameters, etc.), manipulating the response value (adding noise), or manipulating the training set.[68] Since BRANN predictors have demonstrated to be highly stable to network topology variations,[55,56] the latter method was used for introducing diversity in BRGNN ensembles.

Here we used a perturbation technique called subagging, but results are not expected to be different for traditional bagging.[69] A bootstrapped generated training set is obtained after 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.[69]

Model's validation

In this work, we validated our regression model using a reasonable method recently employed by our group that consists of a robust validation process by means of NNE.[39] Recently Baumann[69] 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 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 (see Artificial Neural Network Ensembles). 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.

Self-Organizing Maps

Although back-propagated neural networks have been extensively preferred for nonlinear QSAR modeling, SOMs have also been reported as useful ANNs, accounting for important merits and widespread applications.[70,71]

SOMs[72] 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 an 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 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)].

$$\mathrm{out}_{cs} \leftarrow \min\left[\sum_{i=1}^{m}(AASA_{si} - wij)^2\right] \qquad (14)$$

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 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. Eq. (15) gives the correction formula for the weights.

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

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 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 gene V protein 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 MATLAB environment.[53] 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.

Gene V Protein Mutant Data Set

Gene V protein was used in our study as a benchmark data to test the *AASA* vector approach. Gene V protein (87 residues) is a good model for protein stability studies because a wide thermodynamic data of mutants is available in very homogeneous conditions. Gene V protein data (wild-type and 122 mutants) was collected from the Protherm database.[73] Table II shows the change in Gibbs free energy upon unfolding at 25°C and pH = 7.0 in the presence of Gdn·HCl for wild-type and mutants in comparison to wild-type enzyme. The gene V protein is reversibly denatured by Gdn·HCl, and the denaturation can be monitored by the disappearance of a tyrosine CD signal at 229 nm.[74,75] Stabilities are expressed as free energy changes upon unfolding in kJ/mol of dimeric protein. Stabilities of mutant are compared to that of the wild-type in the presence of 2.0 M Gdn·HCl to yield the difference ($\Delta\Delta G$).

RESULTS AND DISCUSSION

Optimum Amino Acid Sequence Autocorrelation Vectors and Bayesian-Regularized Genetic Neural Networks Simulations

By using the amino acid sequences of the 123 gene V proteins under study (wild-type and mutants), *AASA*

TABLE II. Experimental and Calculated Change of Unfolding Gibbs Free Energy Change ($\Delta\Delta G$)[a] at 25°C, pH = 7.0 and 2.0 M Gdn·HCl for gene V protein wild-type and mutants according to a 50-members neural network ensemble of optimum model BRGNN 3

	$\Delta\Delta G$ (kJ/mol)				$\Delta\Delta G$ (kJ/mol)		
Mutant	Exp.	$Cal._{train}$[b]	$Cal._{test}$[c]	Mutant	Exp.	$Cal._{train}$[b]	$Cal._{test}$[c]
Wild	0	−5.2	−4.9	V35M/I47 L	−7.1	−10.1	−10.2
V35A/I47A	−45.2	−43.2	−38.7	S67T	−6.7	−3.7	−3.5
L37A	−32.2	−29.2	−25.1	L32R	−6.7	−8.5	−10.3
I47T	−31	−29.0	−27.4	E40C	−6.7	−7.0	−8.4
V35C/I47C	−30.1	−28.7	−24.3	V43T	−6.7	−4.5	−0.8
I47A	−29.7	−27.9	−27.0	D50H	−6.7	−4.4	3.7
I78T	−27.6	−30.1	−35.0	C33V/V35C	−6.6	−5.7	−5.2
L49A	−25.5	−23.9	−23.9	L65P	−6.3	−6.7	−6.3
C33M/I47C	−24	−18.0	−14.0	R82C	−6.3	−4.7	−3.4
L 49 T	−23.8	−23.2	−21.7	V35C	−5.9	−8.1	−8.4
V35L/I47M	−22.6	−17.9	−14.6	L37I	−5.9	0.6	2.7
I 47 C	−22.2	−25.0	−26.4	I78V	−5.4	−11.5	−14.0
V 35 T	−22.2	−16.2	−14.6	K69H	−5.4	−5.7	−4.9
L 37 T	−21.8	−22.7	−23.8	M77L	−5	−2.2	−2.0
L 81 T	−21.3	−18.9	−17.3	V35I/I47L	−4.9	−10.5	−12.1
V35L/I47V	−21.3	−12.1	−10.7	V35M	−4.6	−8.1	−9.7
V 70 P	−21.3	−19.0	−17.2	E30N	−4.6	−7.7	−8.1
V 63 T	−20.9	−20.3	−20.9	V45C/R82C	−4.4	−6.3	−13.7
F 68 V	−20.9	−17.3	−16.2	D36N	−4.2	−6.2	−6.5
C 33 T	−19.2	−16.2	−9.9	C33I	−3.8	−3.9	−3.9
L 37 C	−19.2	−18.2	−18.0	L32H	−3.8	−2.9	1.4
V35A/I47 V	−18.9	−20.9	−21.3	T48C	−3.3	−3.8	−7.5
V35A/I47M	−18.9	−22.6	−26.1	M77T	−3.3	−5.9	−10.4
I78C	−18.4	−17.6	−16.5	I6V	−2.9	−6.0	−6.2
F68L	−18	−16.0	−14.7	F13T	−2.9	−4.1	−9.9
C33S	−18	−18.5	−19.7	V35I	−2.9	−2.2	−2.3
L65P/F68 L	−17.8	−20.1	−29.2	I47L	−2.9	−10.4	−10.7
V35F/I47L	−17.7	−13.7	−9.8	A86T	−2.9	−6.3	−6.9
L49C	−17.2	−15.8	−15.5	T62C	−2.9	−5.3	−9.5
V63C	−17.2	−14.4	−13.4	Y41F/F73W	−2.8	−0.9	0.2
H64C/F68L	−17	−16.0	−14.3	Y41F	−2.5	−2.8	−5.4
V35L/I47F	−16.7	−13.3	−12.8	V19T	−2.5	−3.9	−3.9
L28V/F68 L	−16.2	−14.9	−14.9	C33A	−2.1	−7.9	−11.4
S67C	−15.5	−8.0	−3.7	Y26R	−1.7	−2.0	−4.5
L81C	−15.5	−14.4	−12.5	E40T	−1.7	−4.4	−7.0
V35A/I47 F	−15.4	−16.1	−17.4	Y41A	−1.7	−3.6	−22.9
V35M/I47M	−15.1	−12.8	−9.0	V19C	−1.3	−8.5	−12.3
V35L/I47L	−15	−13.3	−12.9	C33V	−0.8	−1.5	−0.3
C33M	−14.6	−12.3	−7.2	L81V	−0.8	−4.3	−4.1
L37V	−14.6	−5.6	−5.2	M77F	−0.8	−1.1	−1.0
V45T	−14.6	−16.1	−18.8	V45C	−0.4	−5.4	−10.8
V70T	−14.6	−15.0	−15.3	I6V/M77V	−0.2	−4.5	−5.5
V70C	−13.8	−9.6	−8.3	K69M	0.4	−0.5	−1.1
V35F	−13.4	−9.4	−6.2	L32Y/R82 C	0.6	2.9	9.6
V35I/I47V	−13.1	−8.5	−8.1	I6V/M77I	1.4	−0.6	0.0
V45L	−12.6	−8.3	−7.6	H64C	2.1	0.8	−1.0
V35A/I47L	−12.4	−10.6	−10.8	A86V	2.1	−2.5	−3.5
L49V	−12.1	−4.4	−3.1	E30M	2.5	−0.8	−3.2
V35I/I47M	−11.9	−17.1	−18.7	I6V/E30F	3	1.9	1.3
V35L	−11.3	−7.6	−7.1	F73W	3.3	2.7	1.9
I47V	−10.9	−10.1	−9.7	K24V	3.3	6.8	9.6
C33L	−10.9	−7.2	−5.8	L32Y	4.2	0.8	0.4
V35M/I47F	−9.8	−11.5	−14.6	L28V	4.6	−3.0	−3.6
V35A	−9.2	−12.0	−12.1	E30F/A86T	4.9	4.4	5.9
I47M	−9.2	−13.6	−14.8	M77V	5	−1.8	−2.9
V45A	−8.8	−9.5	−10.8	T62V	5.4	2.3	−4.3
D36C	−8.8	−3.4	0.9	F13T/E30F	6.7	8.2	14.6
V43C	−8.8	−9.7	−9.8	M77I	6.7	3.8	1.9

TABLE II. (Continued)

	ΔΔ*G* (kJ/mol)				ΔΔ*G* (kJ/mol)		
Mutant	Exp.	$Cal._{train}$[b]	$Cal._{test}$[c]	Mutant	Exp.	$Cal._{train}$[b]	$Cal._{test}$[c]
M77A	−8.8	−9.0	−9.0	E30F	8.4	4.6	2.5
V35I/I47F	−8.7	−10.0	−9.8	E30F/A86V	8.6	7.2	3.0
I47F	−8.4	−10.1	−11.2	L32W	11.7	9.6	3.9
L49I	−7.9	0.6	3.9				

[a]ΔΔ*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 using a 50-members ensemble.

TABLE III. Statistics of the Optimum BRGNN Predictors for the Conformational Stability of Wild-Type and Mutant Gene V Proteins

AASA vectors	BRGNN model	Hidd. nod.	Num. par.	Opt. par.	R^2	S	R^2_{cv}	S_{cv}
*AASA5*H_t	1	2	23	20	0.778	4.703	0.521	5.632
*AASA15*H_t	2	3	34	28	0.81	4.356	0.603	6.473
*AASA2*R_α	**3**	**4**	**45**	**36**	**0.853**	**3.837**	**0.653**	**6.058**
AASA7V	4	5	56	41	0.874	3.556	0.639	6.284
AASA13f	5	6	67	49	0.897	3.205	0.548	7.226
*AASA13*Δ*G*	6	7	78	53	0.905	3.092	0.582	6.925
*AASA7*μ	7	8	89	53	0.905	3.092	0.561	7.293
AASA8α_m	8	9	100	51	0.897	3.205	0.571	7.021
AASA5α_m	9	10	111	51	0.897	3.205	0.531	7.778

Optimum neural network predictor appears in bold letter.
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.

vectors were computed weighted by a variety of physicochemical, energetic, and conformational properties that appear in Table I. In this way, we gathered in a pool of variables, the structural information that can be relevant for modeling the conformational stability of gene V 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 (Bayesian-Regularized Artificial Neural Networks). The nonlinear subspaces in the data set were explored varying the number of network inputs from 6 to 12. 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.

Table III shows the statistical parameters for the optimum BRGNN predictors with nine inputs but varying the number of hidden nodes. By inspection of Table III, it can be observed that Bayesian regularization yielded quite stable and reliable networks. The behavior of the networks was asymptotic with respect to the number of hidden nodes with maximum number of optimum parameters about 50. However, considering the crossvalidation statistics among those neural networks the optimum predictor was BRGNN 3 with four hidden nodes and 36 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 0.85 and 0.65, respectively. The good behavior of the nonlinear models describing the conformational stability of the studied proteins suggests that the *AASA* vectors built a nonlinear vectorial space that well resembles gene V protein stability pattern.

Table IV shows the optimum subset of nine *AASA* vectors and the correlation matrix of such descriptors. Variables in the model mean that *AASA5*H_t and *AASA15*H_t are the amino acid sequence autocorrelations of lag 5 and 15 weighted by thermodynamic transfer hydrophobicity; *AASA2*R_a is the amino acid sequence autocorrelation of lag 2 weighted by solvent-accessible reduction

TABLE IV. Correlation Matrix of the Inputs of the Optimum Predictor BRGNN 3

	*AASA5*H_t	*AASA15*H_t	*AASA2*R_α	*AASA7V*	*AASA13f*	*AASA13*ΔG	*AASA7*μ	*AASA8*α_m	*AASA5*α_m
*AASA5*H_t	1.000	0.510	0.386	0.148	0.005	0.032	0.046	0.025	0.002
*AASA15*H_t		1.000	0.327	0.120	0.000	0.001	0.064	0.013	0.000
*AASA2*R_α			1.000	0.040	0.002	0.034	0.049	0.021	0.001
AASA7V				1.000	0.241	0.028	0.033	0.006	0.014
AASA13f					1.000	0.012	0.005	0.029	0.015
AASA13ΔG						1.000	0.006	0.050	0.005
AASA7μ							1.000	0.109	0.121
AASA8α_m								1.000	0.601
AASA5α_m									1.000

ratio; *AASA7V* is the amino acid sequence autocorrelation of lag 7 weighted by volume (number of nonhydrogen side-chain atoms); *AASA13f* is the amino acid sequence autocorrelation of lag 13 weighted by flexibility (number of side-chain dihedral angles); *AASA13*ΔG is the amino acid sequence autocorrelation of lag 13 weighted by unfolding Gibbs free energy change; *AASA7*μ is the amino acid sequence autocorrelation of lag 7 weighted by refractive index; *AASA8*α_m and *AASA5*α_m are the amino acid sequence autocorrelations of lag 8 and 5 weighted by power to be at the middle of an α-helix. As can be observed in Table IV, there is not significant intercorrelation among selected descriptors, and 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: one thermodynamical (ΔG), five structural (H_t, R_a, V, f, μ,) and one secondary structure-related (α_m) properties. Distribution on the amino acid sequence of unfolding Gibbs free energy change at lag 13 reflects the significance of an adequate amino acid substitution pattern at large ranges in the sequence, resembling a certain thermodynamic pattern in gene V protein. Shape-related amino acid properties (volume, flexibility, and refractive index) appear relevant at autocorrelations of middle and large ranges (lags 7 and 13) on the sequence. This fact suggests that an adequate packing of protein side-chains at protein segments of such lengths but tridimensional close due to folding arrangements contributes to a stable folded state. Likewise, autocorrelations of hydrophobicity-related properties (thermodynamic transfer hydrophobicity and solvent-accessible reduction ratio) at lags 2, 5, and 15 on the sequence could be related with the importance of having an adequate hydrophobicity-polarity distribution at short, middle, and large ranges on the protein structure. Furthermore, distributions at lags 5 and 8 of the power to be at the middle of a α-helix should contribute to an optimum secondary structure pattern that is essential for conformational stability of gene V protein.

Interestingly, solvent-accessible surface area for denatured protein was reported by Gromiha et al.[14] among the properties most linearly correlated with the changes of unfolding Gibbs free energy change for a diverse set of protein mutants. Although secondary structure propensity-related properties did not linearly correlate with the changes of unfolding Gibbs free energy change for the large set of proteins in Ref. 14, this property highly correlated but in a nonlinear way with the conformational stability variations upon mutations in the particular case of gene V protein.

Data-Diverse Ensembles of Bayesian-Regularized Genetic Neural Networks

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 gene V proteins. This approach recently applied by us[39] consists in training several BRGNNs with different randomly partitioned training sets of 99 proteins (80% of the data) and predicting the activity of the rest 24 proteins (20% 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 II two calculated $\Delta\Delta G$ values for each protein: one averaged 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 RMSE of training and test sets, respectively, vs. the number of networks in the ensemble. Concerning this, Figure 2 shows plots of RMSE values for NNEs with number of members varying from 2 to 100. As can be observed, such statistical quantity 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 88% and 66% 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 or residues properties. The conformational stability pattern of gene V proteins that the optimum nine vectors resembled was successfully learned by the ensemble of BRGNNs during supervised training.

Regression Model's Interpretation

To gain a deeper understanding on the relative effects of each autocorrelation vector in the model BRGNN 3, a

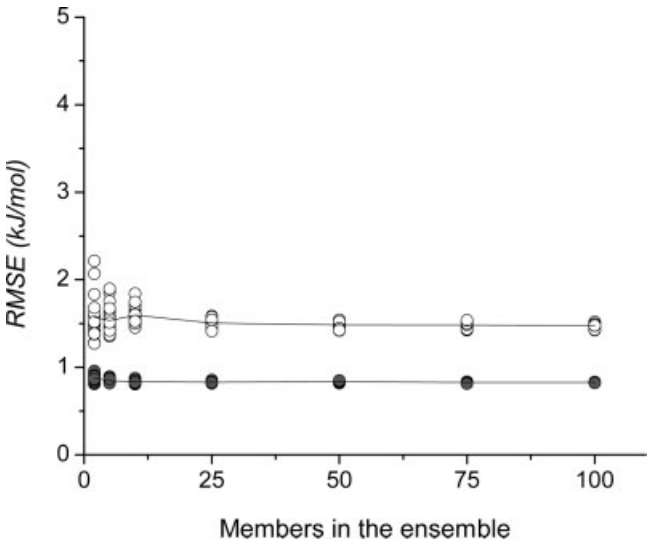


Fig. 2. Plots of *RMSE* of training (●) and test (○) sets for change of unfolding Gibbs free energy change ($\Delta\Delta G$) average values for 20 ensembles vs. number of neural networks in each ensemble.

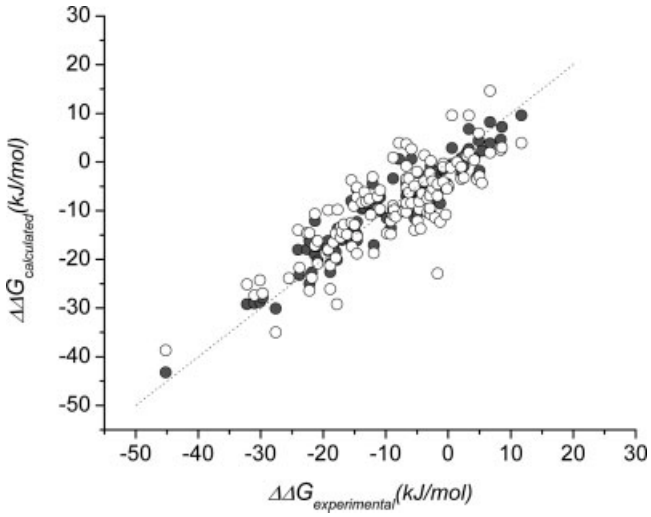


Fig. 3. Plots of average calculated vs. experimental change of unfolding Gibbs free energy change ($\Delta\Delta G$) of gene V protein mutants in training (●) and test (○) sets according to 50-member ensemble of the optimum network BRGNN 3.

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.[76] Their method allows the understanding of 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. The ranking scheme is carried out by determining the square contribution values (*SCV*) for each hidden neuron (see Ref. 76 for details). This method for ANN model interpretation

TABLE V. Effective Weight Matrix for the 9-4-1 BRGNN 3 Model Developed for Conformational Stability of Gene V Protein Mutants[a]

	Hidden nodes			
	2	3	1	4
***AASA5*H_t**	**−1.436**	**1.448**	**0.293**	**0.493**
***AASA15*H_t**	**0.612**	**0.718**	**−0.213**	**−0.097**
*AASA2*R_α	0.259	−0.689	0.038	0.072
AASA7V	0.158	1.093	0.036	−0.445
AASA13f	−0.022	0.084	0.796	−0.702
AASA13ΔG	0.249	−0.769	0.454	−0.249
AASA7μ,	0.201	0.033	−0.528	0.125
***AASA8*α_m**	**−1.725**	**0.409**	**−0.275**	**0.615**
***AASA5*α_m**	**−1.021**	**0.947**	**−0.482**	**0.640**
SCV	0.742	0.217	0.021	0.019

AASA vectors with the highest impacts appear in bold letter.
[a]The columns are ordered by the *SCV*s for the hidden neurons, shown in the last row.

is similar in manner to the partial least squares interpretation method for linear models described by Stanton.[77]

The results of the model interpretation analysis appear in Table V. As can be observed, among the four hidden nodes in the predictor BRGNN 3 the most ranked node is Node 2 having an SCV value about 0.75, which is 3.5-fold higher than hidden Node 3 and about 38-fold higher than the hidden Nodes 1 and 4. According to Guha's analysis,[76] 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 V, descriptors having weights $>|0.5|$ on the most ranked node are the most relevant descriptors. As can be observed, such descriptors are *AASA5*H_t, *AASA15*H_t, *AASA8*α_m, and *AASA5*α_m which represent autocorrelations of thermodynamic transfer hydrophobicity and the power to be at the middle of a α-helix on the gene V proteins structures.

The high relevance of a hydrophobicity related property are in concordance with unfolding denaturation mechanism hypothesis. For the denaturation process of globular proteins, Privalov and Gill[78] stated that hydratation equilibrium, polar interactions between solvent and polar residues in the protein, is the main cause of unfolding meanwhile hydrophobic interactions in the protein core contribute to keep the folded state. On the other hand, the high relevance of the power to be at the middle of a α-helix strongly suggested that optimum secondary structure pattern is another key factor for a stable tertiary conformation. Point mutations studies have highlighted the role of secondary structure propensities in protein stability. Manipulating favorable and unfavorable secondary structure propensities at certain positions in a protein can produce significant variations in protein stability.[13]

Taking into account that conformational stability is a more complex protein property in comparison to other physical stability measurements such as protein melting

TABLE VI. Comparison between a previous report on gene V protein conformational stability prediction and the result obtained here by the *AASA* approach combined with BRGNN according to a LOO crossvalidation test. Results of conformational stability predictions for single mutants of gene V protein using the D-FIRE method[80] and the Capriotti predictor[34] are also shown

Method	Number of mutants	R^2
AASA and BRGNN	123	0.653
Change of free energy transfer of amino acid residues[a]	66	0.168
D-FIRE[b]	90	0.403
Capriotti[c]	90	0.382

[a] R^2 value of jack-knife crossvalidation taken from Table 1 in Ref. 29.
[b] R^2 value from the calculation of ΔΔG for the 90 single mutants using in our study using the web implementation of the D-FIRE method in Ref. 80. (http://sparks.informatics.iupui.edu/hzhou/mutation.html)
[c] R^2 value from the calculation of ΔΔG for the 90 single mutants using in our study using the web implementation of the Capriotti method in Ref. 34. (http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant2.0/I-Mutant2.0.cgi)

point, the accuracy over 85% of our approach for modeling the stability of gene V protein 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 stability of Staphylococcal nuclease mutants using similarity score vectors.[30] In addition, the statistical quality of our ensemble model is in concordance with the report of Marrero–Ponce et al.[36] in which they extended topological indexes to the study of biological macromolecules. In this 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 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 not more than 60% of validation data variance was described; although models were developed for larger and more varied data sets (>1000).[13,29,32–34,79,80] Our neural network model for the conformation stability of gene V protein mutants largely overcomes the predictions of gene V protein mutant ΔΔ*G* values derived from a previous report in Ref. 29 from Zhou and Zhou. Table VI shows a comparative analysis of the results obtained here for the gene V protein and the predictions reported by Zhou and Zhou[29] for this protein. In this report, Zhou and Zhou used a varied data set of 1023 protein mutants in order to derive a new stability scale for amino acid residues that was nonbiased to a specific protein. However, the general scale they derived completely fails to predict gene V protein conformational stability variations upon mutations, yielding very poor results for the gene V protein, as they described only 20% of the validation variance of the 66 single mutants tested, in comparison to 60% in our approach. Furthermore, the overall performance of this method was also very poor, describing about 40% of crossvalidation variance averaged over all the tested mutants.[29]

In addition, prediction results for gene V protein single mutants calculated from the web implementations of the Zhou and Zhou method D-FIRE[80] and the Capriotti et al. predictor[34] are also presented in Table VI. These methods are developed for prediction of protein mutant stability changes upon single mutations only. As can be observed, the D-FIRE and Capriotti methods were unable to describe more than 40% of the conformational stability variance yielded for the 90 single gene V protein mutants in this study, in comparison to more than 60% of our predictor for all the data set including also double mutants. Indeed, a major advantage of our approach is that it can handle any mutant despite the amount of mutated residues, because *AASA* vectors are calculated all over the protein sequence without considering any information depending on the characteristic of a particular mutation. General predictors use a large dataset of proteins as they are designed for giving an approximation of the ΔΔ*G* for any protein, while our approach is useful for making predictions only over gene V protein mutants. However, the bad results of such general methods over gene V protein made our protein specific-method suitable for proteins having some thermodynamic data reported for its mutants.

We tried to establish the number of experimental data cases that should be included in the network training set for a successful application of our approach in the particular case of the gene V protein. The average test set R^2 values for 50 randomly constructed training/test set partitions reached values lower than 0.50 for training set having less than a 30% of the dataset (37 mutants). So, a minimum of 30% of the dataset included in the training set is here sufficient for having an acceptable 50% of the validation variance explained according to the BRGNN model developed.

AASA vectors were capable of resembling an amino acid interaction pattern in human gene V protein 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 information derived from protein primary sequence. 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 mentioned above but with lower generalization abilities. It is noteworthy that our predictor, different to the most of the reported approaches, successfully encompasses both single and double point mutants. The aim of our work was just to present a reliable predictor for the conformational stability of a sole protein using its sequence and a wide thermodynamic data of their mutants. Requiring some previous thermodynamic experimental data for generating a training set is a disad-

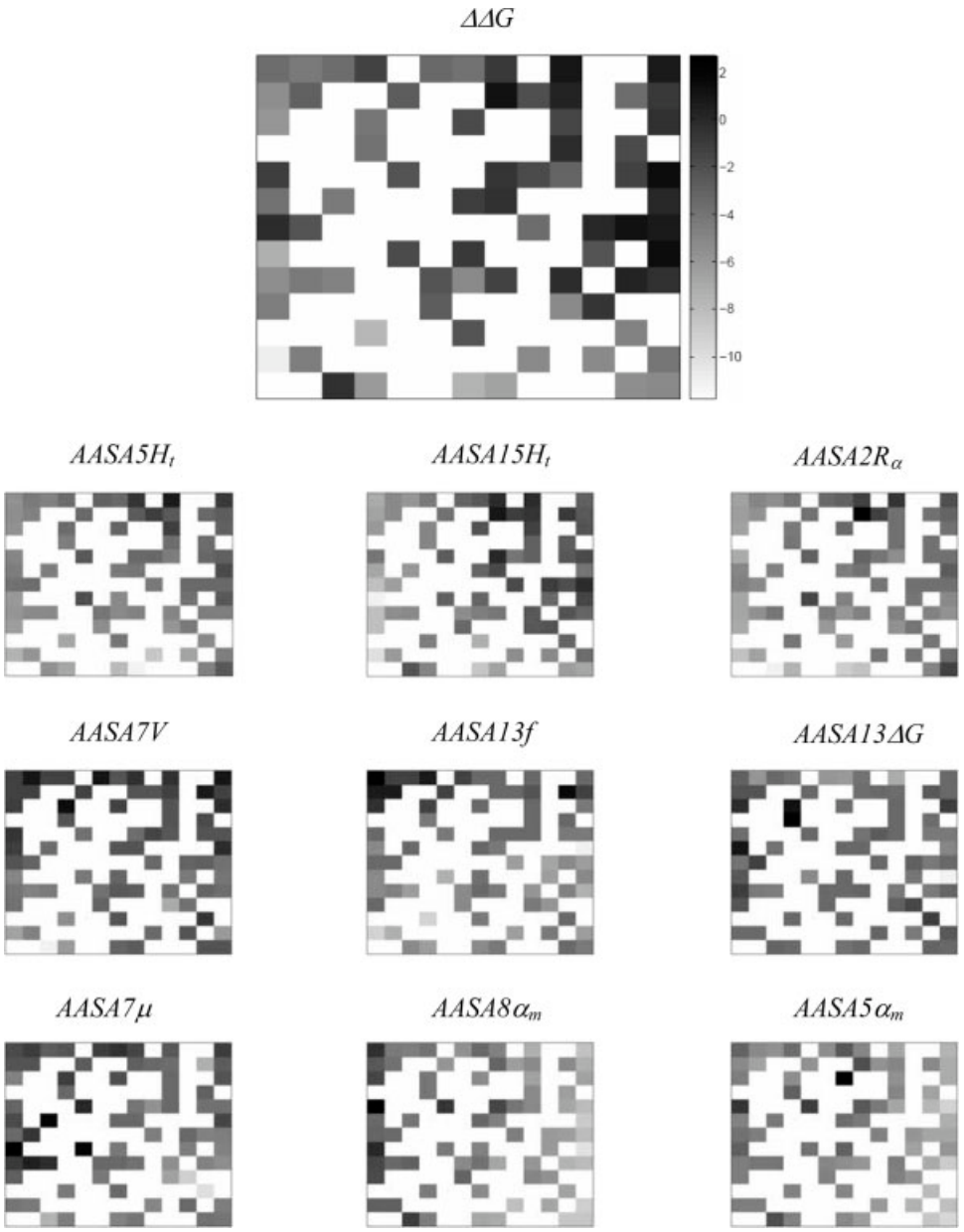


Fig. 4. Kohonen self-organizing maps (SOMs) of the change of unfolding Gibbs free energy change ($\Delta\Delta G$) of gene V protein mutants and the normalized values of the nine optimum *AASA* vectors in predictor BRGNN 3.

vantage, whereas our modeling technique is a viable alternative for the stability prediction of proteins for which X-ray structural information is lacking but some thermodynamic data exists.

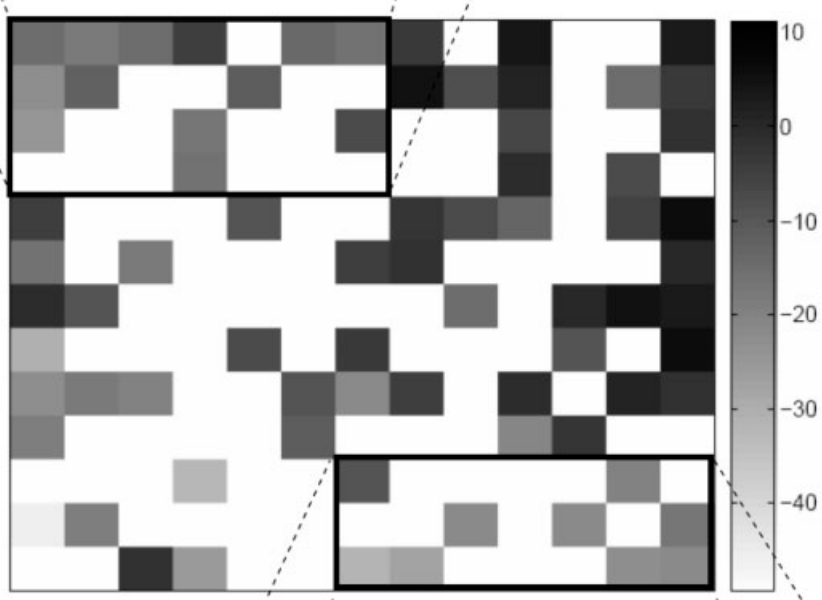
Self-Organizing Maps (SOMs) of Gene V Protein Mutants Stability

Finally, we aimed to settle some similarity among gene V protein mutants by building an SOM of the conformational stability using the optimum subset of *AASA* vectors. Figure 4 depicts 13 × 13 SOM of the $\Delta\Delta G$ values for the studied proteins and SOM of the *AASA* vectors which are inputs of the optimum predictor BRGNN 3. Seventy-four neurons were occupied out of a total of 169 neurons, yielding about 44% of occupancy in the map. As can be observed, proteins with similar stability range were located at neighboring neurons in the map. Distributions of autocorrelation vectors on the map

Fig. 5. Kohonen self-organizing map (SOM) of the change of unfolding Gibbs free energy change ($\Delta\Delta G$) of gene V protein mutants. Conformational stability legend is placed at the right-hand of the map. Underlined mutants mean wrong located mutants.

V35M/I47M V35M/I47L Y26R	V35F/I47L	V35L/I47L V35A/I47L V35I/I47L	V35M		V35F	S67C T62C
V35L/I47M L32R	V35I/I47M			V35L		
C33M/I47C V35M/I47F I47M			V35L/I47F V35I/I47F I47F			D50H
			L81C			

Zone 1



Zone 2

V45A					C33T C33S	
		L37T		L81T		H64C/F68L M77T
C33M/I47C V35M/I47F I47M	I78T L49T V63T V45T V70T V43T F13T V19T		V35L/I47F V35I/I47F I47F		V35T	V70P L65P/F68L L65P

Figure 5.

depict complex patterns, although the distributions of the most relevant AASA vectors (*AASA5*H_t, *AASA15*H_t, *AASA8*α_m and *AASA5*α_m) tend to well match to $\Delta\Delta G$ map. Less stable gene V protein mutants were placed at the left-middle, left-bottom, and right-bottom regions of the map. Otherwise, most stable mutants were distributed at right-top and right-middle zones of the map.

By analyzing the $\Delta\Delta G$ topological map in Figure 5, some structural similarities among less stable mutants can be addressed taking into to account their allocation at adjacent neurons with similar level of conformational stability. In this sense, destabilizing mutations of hydrophobic residues Valine 35 and Isoleucine 47 in the interior of the protein by other hydrophobic residues were located at the right-top region, Zone 1 in the map. Although positions 35 and 47 respond little differently to substitution, nongain in stability in comparison to wild-type protein is achieved according to mutation studies.[74,75] The volume available for a side-chain at position 35 is quite small and the protein interior at this site has a higher energetic penalty for conformational alterations, whereas groups around the site at Position 47 are more easily rearranged. That is why hydrophobic mutations on Valine 35 may have 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, the surroundings of position 47 may be readily deformable or there may be compensating effects that yield no net packing energy change. However, replacement of Isoleucine in wild-type gene V protein decreases hydrophobicity in the protein core leading to a fall in protein stability.

Another interesting region is denoted as Zone 2 on the map. Mutants in this region correspond to destabilizing mutations, and most of them are substitutions of small and/or flexible residues such as valine, leucine, isoleucine, and methionine by bulkier and/or more rigid amino acids: phenylalanine, tyrosine, and proline. Such mutations could lead to energetic unfavored arrangements of groups inside the protein due to the introduction of bulkier and/or more rigid side-chains, decreasing the so-called protein interior packing energy.[74,75] Those mutations decrease side-chain entropy and could contribute to destabilize the folded state if the entropy cost is bigger than the favorable interaction energy brought by the hydrogen bond and/or van der Waals interactions added by residue substitution.

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 the most of them require X-ray structural analysis and have limited prediction accuracy. This fact is 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. Because of 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 of proteins as a tool for encoding protein structural information for supervised training of ANNs. In this sense, novel *AASA* vectors were obtained by calculating autocorrelations on the protein primary structure of 48 amino acid/residue properties selected from the AAindex data base. BRGNNs proved 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 gene V protein mutants that describes about 86% and 66% 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 a 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. Although one model per protein is required according to the approach presented here, a general model encompassing a large and varied mutant data (>2000) as well as protein-specific models for other proteins are under development by our group at the present time.

REFERENCES

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 DN, Marcus JS, Ross SA, Mayo SL. Prudent modeling of core polar residues in computational protein design. J Mol Biol 2003;329:611–622.
4. Looger LL, Dwyer MA, Smith JJ, Helling HW. Computational design of receptor and sensor proteins with novel functions. Nature 2003;423:185–190.
5. Dang LX, Merz KM, Kollman PA. 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 CM, Srinivasan N, Blundell TL. Prediction of the stability of protein mutants based on structural environment-dependent 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 AR, 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.
12. 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.
13. Guerois R, Nielsen JE, 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.
14. Gromiha MM, Oobatake M, Kono H, Uedaira H, Sarai A. Relationship between amino acid properties and protein stability: buried mutations. J Prot Chem 1999;18:565–578.
15. Gromiha MM, 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.
16. Gromiha MM, 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.
17. 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.
18. 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.
19. 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.
20. 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.
21. 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.
22. 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.
23. 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.
24. 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.
25. 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.
26. 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.
27. 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.
28. 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.
29. Zhou H, Zhou Y. Stability scale and atomic solvation parameters extracted from 1023 mutation experiments. Proteins 2002;49: 483–492.
30. Frenz CM. Neural network-based prediction of mutation-induced protein stability changes in Staphylococcal nuclease at 20 residue positions. Proteins 2005;59:147–151.
31. Levin S, Satir BH. POLINA: Detection and evaluation of single amino acid substitutions in protein superfamilies. Bioinformatics 1998;14:374–375.
32. Capriotti E, Fariselli P, Casadio R. A neural-network-based method for predicting protein stability changes upon single mutations. Bioinformatics 2004;20:63–68.
33. Capriotti E, Fariselli P, Calabrese R, Casadio R. Prediction of protein stability changes from sequences using support vector machines. Bioinformatics 2005;21:54–58.
34. Capriotti E, Fariselli P, Casadio R. I-Mutant20: predicting stability changes upon mutation from the protein sequence or structure. Nucl Acids Res 2005;33:306–310.
35. 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.
36. Marrero-Ponce Y, Medina-Marrero R, Castillo-Garit JA, Romero-Zaldivar V, Torrens F, Castro EA. 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.
37. Fernández M, Caballero J, Helguera AM, Castro EA, González MP. Quantitative structure-activity relationship to predict differential inhibition of aldose reductase by flavonoid compounds. Bioorg Med Chem 2005;13:3269–3277.
38. Fernández M, Tundidor-Camba A, Caballero J. 2D Autocorrelation modeling of the activity of trihalobenzocycloheptapyridine analogues as farnesyl protein transferase inhibitors. Mol Simulat 2005;31:575–584.
39. Fernández M, Tundidor-Camba A, Caballero J. Modeling of cyclin-dependent kinase inhibition by 1H-pyrazolo [3,4-d] pyrimidine derivatives using artificial neural networks ensembles. J Chem Inf Comput Sci 2005;45:1884–1895.
40. González MP, Caballero J, Tundidor-Camba A, Helguera AM, 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.
41. 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.
42. 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.
43. Moran PAP. Notes on continuous stochastic processes. Biometrika 1950;37:17–23.
44. Geary RF. The contiguity ratio and statistical mapping. Incorporated Statistician 1954;5:115–145.
45. Moreau G, Broto P. Autocorrelation of a topological structure: A new molecular descriptor. Nouv J Chim 1980;4:359–360.
46. Moreau G, Broto P. Autocorrelation of molecular structures: application to SAR studies. Nouv J Chim 1980;4:757–764.
47. 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.
48. 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.
49. Guha R, Jurs PC. 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.
50. 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.
51. 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.
52. Kawashima S, Kanehisa M. AAindex: amino acid index database. Nucleic Acids Res 2000;28:374.

53. MATLAB 7.0. Mathworks, Natick, MA. Available at http://www.mathworks.com.
54. So S, Karplus M. Evolutionary optimization in quantitative structure-activity relationship: an application of genetic neural networks. J Med Chem 1996;39:1521–1530.
55. Burden FR, Winkler DA. Robust QSAR models using bayesian regularized neural networks. J Med Chem 1999;42:3183–3187.
56. Winkler DA, Burden FR. Bayesian neural nets for modeling in drug discovery. Biosilico 2004;2:104–111.
57. Zupan J, Gasteiger J. Neural networks: a new method for solving chemical problems or just a passing phase? Anal Chim Acta 1991;248:1–30.
58. Aoyama T, Suzuki Y, Ichikawa H. Neural networks applied to structure-activity relationships. J Med Chem 1990;33:905–908.
59. Mackay DJC. Bayesian interpolation. Neural Comput 1992; 4:415–447.
60. Mackay DJC. A practical bayesian framework for backprop networks. Neural Comput 1992;4:448–472.
61. Lampinen J, Vehtari A. Bayesian approach for neural networks—review and case studies. Neural Networks 2001;14:7–24.
62. Foresee FD, Hagan MT. Gauss-Newton approximation to Bayesian learning. In Proceedings of the 1997 International Joint Conference on Neural Networks. Houston: IEEE; 1997. pp 1930–1935.
63. Holland H. Adaption in natural and artificial systems. Ann Arbor, MI: The University of Michigan Press; 1975.
64. Cartwright HM. Applications of artificial intelligence in chemistry. Oxford: Oxford University Press; 1993.
65. Hemmateenejad B, Safarpour MA, 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.
66. Hansen LK, Salamon P. Neural network ensembles. IEEE Trans Pattern Anal Machine Intell 1990;12:993–1001.
67. Krogh A, Vedelsby J. Neural network ensembles, cross-validation and active learning. In: Tesauro G, Touretzky D, Lean T, editors. Advances in neural information processing systems 7. MA: MIT Press; 1995. pp 231–238.
68. Agrafiotis DK, Cedeño W, Lobanov VS. On the use of neural network ensembles in QSAR and QSPR. J Chem Inf Comput Sci 2002;42:903–911.
69. 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.
70. 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.
71. de-Sousa JA, 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.
72. Kohonen T. Self-organized formation of topologically correct feature maps. Biol Cybern 1982;43:59–69.
73. Bava KA, Gromiha MM, 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.
74. Sandberg WS, Terwilliger TC. Energetics of repacking a protein interior. Proc Natl Acad Sci USA 1991;88:1706–1710.
75. Sandberg WS, Terwilliger TC. Engineering multiple properties of a protein by combinatorial mutagenesis. Proc Natl Acad Sci USA 1993;90:8367–8371.
76. Guha R, Stanton DT, Jurs PC. Interpreting computational neural network QSAR models: a detailed interpretation of the weights and biases. J Chem Inf Model 2005;45:1109–1121.
77. Stanton DT. On the physical interpretation of QSAR models. J Chem Inf Comput Sci 2003;43:1423–1433.
78. Privalov PL, Gill SJ. Stability of protein structure and hydrophobic interaction. Adv Prot Chem 1988;39:191–234.
79. Bordner AJ, Abagyan RA. Large-scale prediction of protein geometry and stability changes for arbitrary single point mutations. Proteins 2004;57:400–413.
80. Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 2002;11:2714–2726.