Available online at www.sciencedirect.com
ScienceDirect
Journal of Molecular Graphics and Modelling 26 (2007) 166–178
Journal of Molecular Graphics and Modelling
www.elsevier.com/locate/JMGM

Proteometric study of ghrelin receptor function variations upon mutations using amino acid sequence autocorrelation vectors and genetic algorithm-based least square support vector machines

Julio Caballero [a], Leyden Fernández [a], Miguel Garriga [a,b], José Ignacio Abreu [a,c], Simona Collina [d], Michael Fernández [a,*]

[a] *Molecular Modeling Group, Center for Biotechnological Studies, Faculty of Agronomy, University of Matanzas, 44740 Matanzas, Cuba*
[b] *Plant Biotechnology Group, Center for Biotechnological Studies, Faculty of Agronomy, University of Matanzas, Matanzas, C.P. 44740, Cuba*
[c] *Artificial Intelligence Lab, Faculty of Informatics, University of Matanzas, 44740 Matanzas, Cuba*
[d] *Department of Pharmaceutical Chemistry, University of Pavia, via Taramelli, 12, 27100 Pavia, Italy*

Received 20 June 2006; received in revised form 8 November 2006; accepted 8 November 2006
Available online 15 November 2006

Abstract

Functional variations on the human ghrelin receptor upon mutations have been associated with a syndrome of short stature and obesity, of which the obesity appears to develop around puberty. In this work, we reported a proteometrics analysis of the constitutive and ghrelin-induced activities of wild-type and mutant ghrelin receptors using amino acid sequence autocorrelation (AASA) approach for protein structural information encoding. 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. Genetic algorithm-based multilinear regression analysis (GA-MRA) and genetic algorithm-based least square support vector machines (GA-LSSVM) were used for building linear and non-linear models of the receptor activity. A genetic optimized radial basis function (RBF) kernel yielded the optimum GA-LSSVM models describing 88% and 95% of the cross-validation variance for the constitutive and ghrelin-induced activities, respectively. AASA vectors in the optimum models mainly appeared weighted by hydrophobicity-related properties. However, differently to the constitutive activity, the ghrelin-induced activity was also highly dependent of the steric features of the receptor.

Keywords: 7TM protein; Mutational studies; Kernel-based methods; QSAR; Ghrelin; Autocorrelation vectors; Constitutive activity

1. Introduction

Simple reflection of molecular activation mechanism provoke that seven-transmembrane segment (7TM) or G protein-coupled receptors can signal without any agonist present [1]. The receptor inactive and active conformations are in equilibrium and can relatively easily access the active conformation without the presence of an agonist [2]. Such ligand-independent or constitutive signalling is generally ignored even it varies among receptors because in most cases only represents a small fraction of the maximal signalling capacity. However, constitutive activity appears to have important functional consequences among receptors involved in the control of appetite and energy expenditure. The cannabinoid type 1 (CB1) receptor, which is the target for the novel antiobesity drug rimonabant and the ghrelin receptor both signal with around 50% activity in the absence of ligand [3]. But the lack of appropriate pharmacological tools makes hard to establish *in vivo* the physiological importance of such constitutive signalling. Besides, inverse agonists, that in most cases act as antagonists, also block the action of the endogenous agonist masking the in vivo setting to differentiate between an effect on constitutive receptor signalling and a blocking effect on receptor access to an endogenous ligand.

Ghrelin is a hormone and neuropeptide involved in growth hormone (GH) release and control of food intake and energy

* Corresponding author. Tel.: +53 45 26 1251; fax: +53 45 25 3101.
E-mail addresses: michael.fernandez@umcc.cu, michael_llamosa@yahoo.com (M. Fernández).

1093-3263/$ – see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2006.11.002

expenditure [4]. Besides being a transmitter in discrete neuronal networks, ghrelin functions as a hormonal ''hunger signal'' from enteroendocrine cells in the stomach to various target cells located in afferent vagal neurons, the brain stem, and the arcuate nucleus of the hypothalamus [5]. Ghrelin concentrations in plasma are opposite to that of gastrointestinal tract hormones in general: a surge before the first meal of the day is followed by a prolonged nadir caused by the inhibitory effect of food being present in the upper gastrointestinal tract. This dynamic pattern suggests that, between meals, constitutive GH secretagogue receptor activity could play an important role in modulating the orexigenic signals in the regulatory pathways that are integrating anorexigenic signals such as leptin, insulin, and peptide $YY_{3–36}$, etc. [5]. The amount of constitutive signalling is directly proportional to the expression level of the receptor, and the expression of the ghrelin receptor is highly regulated, for example, by fasting [5,6].

Recently, natural mutations in the ghrelin receptor which are associated with a selective loss of constitutive activity without affecting ghrelin affinity, potency, or efficacy, segregates in families with the development of short stature and obesity [7]. This fact suggests that obesity is part of the phenotype associated with a ghrelin receptor lacking constitutive activity.

Furthermore, through a mutational analysis of the opposing faces of TMs III, VI, and VII of the ghrelin receptor in particular, an aromatic cluster was identified that seems to be structurally important for the constitutive activity of this family of receptors [8]. This was verified through a series of corresponding substitutions in both the ghrelin receptor and GPR39 at position VI:16 with residues of variable aromaticity and size through which the constitutive activity could systematically be tuned up and down.

Current computational methods of assessing protein function are based to a large extend on prediction based on sequence similarity of proteins with other proteins having known functions. The accuracy of such predictions depends on the ability of the computational methods to extend sequence similarity to functional similarity [9]. Conventional approaches to molecular recognition have hitherto essentially required determining protein three-dimensional structures, which is resource-demanding and error-prone, and generally requires prior knowledge such as three-dimensional structure of a homologous protein. But the great gap between the amount of known protein sequences and the elucidated structures is a large drawback for the three-dimensional-based protein function modeling [9].

In chemistry and related fields, chemometrics has been developed for more than 30 years, consisting on the use of mathematical, statistical and symbolic methods to improve the understanding of chemical information [9]. Chemometrics has been most successfully applied in four areas, namely multivariate calibration, quantitative structure–activity relationship (QSAR) studies, pattern recognition, classification and discriminate analysis and multivariate modeling and monitoring process [9]. But recently the development of the bioinformatics has brought up chemometrics studies focused on proteins, so-called proteometrics studies.

QSAR studies of proteins have been developed by extending conventional graph–theoretical representation of chemical structures to protein sequence and three-dimensional structure in combination with statistical methods for regression and/or classification analysis [10,11]. This approach has been applied to protein stability studies and protein classification. Similarly, a novel bioinformatics approach named as proteochemometrics including the mapping of molecular recognition and not necessary needing knowledge of the three-dimensional structure of biomacromolecules has successfully applied to the study of protein–ligand interactions [9]. In this context, we recently introduced the amino acid sequence autocorrelation (AASA) vectors for modeling the conformational stability of human lysozyme mutants [11]. AASA vectors are weighted by 48 physicochemical, energetic, and conformational amino acid/residues properties extracted from the AAindex amino acid database [12]. Property autocorrelations are calculated over the protein sequence at different spatial lags. The generated X-ray structure-independent data that contained very rich information of the protein primary sequence also contains quasi-sequence order effects. In this way, a large dataset is computed and optimum models are obtained by employing linear and non-linear modeling techniques combined with genetic algorithm (GA) feature selection.

In this work, a proteometric analysis of the variation of the human ghrelin receptor constitutive and ghrelin-induced activities upon mutations was developed using AASA vectors for protein information encoding. Linear and non-linear regression models of the constitutive activity of the wild-type and mutant human ghrelin receptors were obtained using GA-based multilinear regression analysis (MRA) and non-linear GA-based least square support vector machines (LSSVM).

2. Methods and experimental procedure

2.1. Amino acid sequence autocorrelation vector (AASA) approach

Protein interactions depend 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 comparing such structures [13].

Autocorrelation vectors have several useful properties. Firstly, a substantial reduction in data can be achieved by limiting the topological distance, l. Secondly, the autocorrelation coefficients are independent of the original atom numberings, so they are canonical. And thirdly, the length of the correlation vector is independent of the size of the molecule [13].

For the autocorrelation vectors in molecules, H-depleted molecular structure is represented as a graph and physicochemical properties of atoms as real values assigned to the

graph vertices. These descriptors can be obtained by summing up the products of certain properties of two atoms, located at given topological distances or spatial lag in the graph. Two-dimensional spatial autocorrelations [14–16] has been successfully used in the last decades for modeling biological activities [16,17] and pharmaceutical research [13,18]. In recent works, our group has obtained outstanding results when such chemical code was used in combination with ANN approach in biological QSAR studies [19]. Such results 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 function/property prediction.

Broto–Moreau's autocorrelation coefficient [16] 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

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

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 mainly differs from the Gromiha et al. [20] method in considering the whole amino acid sequence of the protein for calculation of the descriptors instead local sequence segments over the mutated point. In this way, the calculated autocorrelation vectors encode structural information concerning whole protein. Particularly, amino acid sequence autocorrelation (AASA) vectors of lag l are calculated as follows:

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

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

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

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

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

Autocorrelation measures the level of interdependence between properties, and the nature and strength of that interdependence. It may be classified 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 [14,15]. 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 were used 48 physicochemical, energetic, and conformational amino acid/residues properties (Table 1) selected by Gromiha et al. [21] from the AAindex database [12] in a previous study concerning relationships between amino acid/residues properties and protein stability for a large set of proteins. These properties were recently used by us for generating human lysozymes AASA vectors for modeling conformational stability [11] and by Grommiha et al. [22] for predicting with protein folding rates. 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 [23]. A data matrix of 720 AASA vectors, 48 properties × 15 different lags, were generated with the autocorrelation vectors calculated for each ghrelin receptor mutant. Descriptors that stayed constant or almost constant were eliminated and pairs of variables with a square correlation coefficient (R^2) greater than 0.9 were classified as intercorrelated, and only one of these was included for building the model. Finally, 143 descriptors were obtained. Afterwards, optimum predictive models were built with reduced subsets of variables by means of MRA and LSSVM approaches combined with GA feature selection, so-called GA-MRA and GA-LSSVM. The dataset was normalized for LSSVM model generation.

2.2. Least squares support vector machine

In recent years, the support vector machine (SVM), based on statistical learning theory, as a powerful new tool for data classification and function estimation, has been developed [24]. SVM maps input data into a high-dimensional feature space where it may become linearly separable. Recently SVM has been applied to a wide variety of domains such as pattern recognition and object detection [25], function estimation [26], etc.

One reason that SVM often performs better than earlier methods is that SVM was designed to minimize structural risk whereas previous techniques were usually based on minimization of empirical risk. So SVM is usually less vulnerable to the overfitting problem. Especially, Suykens and Vandewalle [27] proposed a modified version of SVM called least squares SVM (LSSVM), which resulted in a set of linear equations instead of a quadratic programming problem, which can extend the application of the SVM. To understand LSSVM well, here LSSVM for classification and regression was introduced first.

Excellent introductions to SVM appear in Refs. [25,26]. The theory of LSSVM has also been described clearly by Suykens

Table 1
Numerical values of 48 selected physicochemical, energetic, and conformational properties of the 20 amino acids/residues [12,21]

Property[a,b]	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
1. K_0	−25.5	−32.82	−33.12	−36.17	−34.54	−27	−31.84	−31.78	−32.4	−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	0.87	1.52	0.66	0.67	2.87	0.1	0.87	3.15	1.64	2.17	1.67	0.09	2.77	0	0.85	0.07	0.07	1.87	3.77	2.67
3. H_P	13.05	14.3	11.1	11.41	13.89	12.2	12.42	15.34	11.01	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	1.48	49.7	49.9	0.35	0	51.6	0.1	49.5	0.13	1.43	3.38	1.58	3.53	52	1.67	1.66	0.13	2.1	1.61
5. pH_i	6	5.05	2.77	5.22	5.48	5.97	7.59	6.02	9.74	5.98	5.74	5.41	6.3	5.65	10.76	5.68	5.66	5.96	5.89	5.66
6. pK'	2.34	1.65	2.01	2.19	1.89	2.34	1.82	1.36	2.18	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	89	121	133	147	165	75	155	131	146	131	149	132	115	146	174	105	119	117	204	181
8. P_1	11.5	13.46	11.68	13.57	19.8	3.4	13.67	21.4	15.71	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	9.9	2.8	2.8	3.2	18.8	5.6	8.2	17.1	3.5	17.6	14.7	5.4	14.8	9	4.6	6.9	9.5	14.3	17	15
10. μ	14.34	35.77	12	17.26	29.4	0	21.81	19.06	21.29	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}	0.62	0.29	0.9	−0.74	1.19	0.48	−0.4	1.38	−1.5	1.06	0.64	−0.78	0.12	−0.85	−2.53	−0.18	−0.05	1.08	0.81	0.26
12. E_{sm}	1.4	1.37	1.16	1.16	1.14	1.36	1.22	1.19	1.07	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	0.49	0.67	0.35	0.37	0.72	0.53	0.54	0.76	0.3	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	1.9	2.04	1.52	1.54	1.86	1.9	1.76	1.95	1.37	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	1.42	0.7	1.01	1.51	1.13	0.57	1	1.08	1.16	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	0.83	1.19	0.54	0.37	1.38	0.75	0.87	1.6	0.74	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	0.66	1.19	1.46	0.74	0.6	1.56	0.95	0.47	1.01	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	0.71	1.19	1.21	0.84	0.71	1.52	1.07	0.66	0.99	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	20	25	26	33	46	13	37	39	46	35	43	28	22	36	55	20	28	33	61	46
20. F	0.96	0.87	1.14	1.07	0.69	1.16	0.8	0.76	1.14	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	0.38	0.57	0.14	0.09	0.51	0.38	0.31	0.56	0.04	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	3.7	3.03	2.6	3.3	6.6	3.13	3.57	7.69	1.79	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	6.05	7.86	4.95	5.1	6.62	6.16	5.8	7.51	4.88	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	1.59	0.33	0.53	1.45	1.14	0.53	0.89	1.22	1.13	1.91	1.25	0.53	0	0.98	0.67	0.7	0.75	1.42	1.33	0.58
25. α_c	1.44	0.76	2.13	2.01	1.01	0.62	0.56	0.68	0.59	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	1.22	1.53	0.56	1.28	1.13	0.4	2.23	0.77	1.65	1.05	1.47	0.93	0	1.63	1.59	0.87	0.46	1.2	0.46	0.52
27. V°	60.46	67.7	73.83	85.88	121.48	43.25	98.79	107.72	108.5	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	2.11	1.88	1.8	2.09	1.98	1.53	1.98	1.77	1.96	2.19	2.27	1.84	1.32	2.03	1.94	1.57	1.57	1.63	1.9	1.67
29. N_l	3.92	5.55	2.85	2.72	4.53	4.31	3.77	5.58	2.79	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}	13.85	15.37	11.61	11.38	13.93	13.34	13.82	15.28	11.58	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	104	132.5	132.2	161.9	182	73.4	165.8	171.5	195.2	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	33.2	17.9	62.4	81	33.1	29.2	57.7	28.3	107.5	31.1	41.3	60.5	60.7	71.5	94.5	48.7	52	28.1	39.5	50.4
33. ΔASA	70.9	114.3	69.6	80.5	148.4	44	107.9	142.7	87.5	129.8	147.9	74	73.5	93.3	116	62.8	78	115.6	167.8	145.9
34. ΔGh	−0.54	−1.64	−2.97	−3.71	−1.06	−0.59	−3.38	0.32	−2.19	0.27	−0.6	−3.55	0.32	−3.92	−5.96	−3.82	−1.97	0.13	−3.8	−5.64
35. ΔG_{hD}	−0.58	−1.91	−6.1	7.37	−1.35	−0.82	−5.57	0.4	−5.97	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}	−0.06	−0.27	−3.11	−3.62	−0.28	−0.23	−2.18	0.07	−1.7	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	−2.24	−3.43	−4.54	−5.63	−5.11	−1.46	−6.83	−3.84	−5.02	−3.52	−4.16	−5.68	−1.95	−6.23	−10.43	−5.94	−4.39	−3.15	−8.99	−10.67
38. $-T\Delta S_h$	1.7	1.79	1.57	1.92	4.05	0.87	3.45	4.16	2.83	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}	14.22	9.41	2.73	3.17	39.06	4.88	20.05	41.98	17.68	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	0.51	2.71	2.89	3.58	3.22	0.68	3.95	−0.4	1.87	−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	2.77	8.64	4.72	5.69	11.93	1.23	7.64	4.03	3.57	3.69	7.06	3.64	1.97	4.47	6.03	5.8	4.42	3.45	13.46	14.41
42. $-T\Delta S_c$	−2.25	−5.92	−1.83	−2.11	−8.71	−0.55	−3.69	−4.42	−1.7	−4.04	−5.93	−0.39	−2.36	−0.78	−0.78	−2.38	−2.68	−3.64	−7.87	−7.95
43. ΔG	−0.02	1.08	−0.08	−0.13	2.16	0.09	0.56	−0.08	−0.32	−0.08	0.53	−0.3	−0.06	−0.23	−0.71	−0.4	−0.24	−0.06	1.78	0.91
44. ΔH	0.51	5.21	0.18	0.05	6.82	−0.23	0.79	0.19	−1.45	0.17	2.89	−2.03	0.02	−1.76	−4.4	−0.16	0.04	0.3	4.47	3.73
45. $-T\Delta S$	−0.54	−4.14	−0.26	−0.19	−4.66	0.31	−0.23	−0.27	1.13	−0.24	−2.36	1.74	−0.08	1.53	3.69	−0.24	−0.28	−0.36	−2.69	−2.82
46. V	1	2	4	5	7	0	6	4	5	4	4	4	3	5	7	2	3	3	10	8
47. s	0	0	2	3	2	0	2	1	0	2	0	2	0	3	5	0	1	1	2	2
48. f	0	1	2	3	2	0	2	2	4	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 non-bonded energy; E_1 long-range non-bonded energy; E_t, total non-bonded 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 r.m.s. fluctuational displacement; Br, 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_l 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.

and Vandewalle [27]. For this reason, we will only briefly describe the main idea of LSSVM and the differences between SVM and LSSVM here.

2.2.1. LSSVM for classification [27,28]

Consider a binary classification training sample $\{(x_i, y_i)\}_{i=1,2,\ldots,l}$, where x_i is the vector of input pattern for the ith example and y_i is the corresponding target output. The pattern represented by the subset $y_i = +1$ belongs to class 1, and the pattern represented by the subset $y_i = -1$ belongs to class 2. The original SVM classifier satisfies the following conditions:

$$y_i \lfloor w^{\mathrm{T}}\varphi(x)_i + b \rfloor \geq 1, \quad i = 1, \ldots, l \tag{6}$$

where φ: $R^n \rightarrow R^m$ is the feature map mapping the input space to a usually high-dimensional feature space where a hyperplane defined by the pair $(w \in R^m, b \in R)$ linearly separates the data points. The classification function is then given by

$$y(x) = \mathrm{sign}\{w^{\mathrm{T}}\varphi(x) + b\} \tag{7}$$

It is usually unnecessary to compute with the feature map, and it is only needed to work instead with a kernel function in the original space given by

$$K(x_i, x_j) = \varphi(x_i)^{\mathrm{T}}\varphi(x_j) \tag{8}$$

In the case of noisy data, poor generalization is yielded by forcing zero training error. So, some data points may be misclassified by introducing a set of slack variables:

$$\varepsilon_i > 0, \quad i = 1, \ldots, l \tag{9}$$

The relaxed separation constraint is given as

$$y_i \lfloor w^{\mathrm{T}}\varphi(x)_i + b \rfloor \geq 1 - \varepsilon_i, \quad i = 1, \ldots, l \tag{10}$$

The optimal separating hyperplane can be found by the following minimization problem:

$$\min_{w,b,\varepsilon} J(w, b) = \frac{1}{2}w^{\mathrm{T}}w + C\sum_{i=1}^{l}\varepsilon_i \tag{11}$$

subject to the constraints

$$\left\{ \begin{array}{ll} y_i[w^{\mathrm{T}}\varphi(x) + b] \geq 1 - \varepsilon_i, & i = 1, \ldots, l \\ \varepsilon_i \geq 0, & i = 1, \ldots, l \end{array} \right\}$$

where C is a regularization parameter used to decide a trade off between the training error and the margin. The dual of system in Eq. (11) leads to a well-known convex quadratic programming problem via the Karush–Kuhn–Tucker condition.

Standard formulation for Vapnik's SVM classifier was modified by Suykens and Vandewalle into the following LSSVM formulation:

$$\min_{w,b,e} J(w, b) = \frac{1}{2}w^{\mathrm{T}}w + \gamma\frac{1}{2}\sum_{i=1}^{l}e_i^2 \tag{12}$$

subject to the equality constraint

$$y_i \lfloor w^{\mathrm{T}}\varphi(x_i) + b \rfloor = 1 - e_i, \quad i = 1, \ldots, l$$

Passing from Eq. (11) to Eq. (12) involves replacing the inequality constraints by equality constraints and a squared error term (hence least square) similar to ridge regression. The corresponding Lagrange for Eq. (12) is

$$L(w, b, e, \alpha) = J(w, e) - \sum_{i=1}^{l}\alpha_i\{y_i[w^{\mathrm{T}}\varphi(x_i) + b] - 1 + e_i\} \tag{13}$$

where the α_i are Lagrange multipliers. As was shown in Ref. [27], the optimality condition leads to the following $(N + 1) \times (N + 1)$ linear system:

$$\begin{bmatrix} 0 & Y^{\mathrm{T}} \\ Y & ZZ^{\mathrm{T}} + \gamma^{-1}I \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{14}$$

where $Z = \lfloor \varphi(x_1)^{\mathrm{T}}y_1, \ldots, \varphi(x_l)^{\mathrm{T}}y_l \rfloor$, $Y = [y_1, \ldots, y_l]$, $1 = [1, \ldots, 1]$.

Mercer's condition is applied within the matrix ZZ^{T}:

$$ZZ^{\mathrm{T}} = y_i y_j \varphi(x_i)^{\mathrm{T}}\varphi(x_j) = y_i y_j K(x_i, x_j)$$

Thus, we would only need to use kernel function K in the training algorithm, and would never need to explicitly, even know what φ is. The LSSVM classifier is then constructed as follows:

$$f(x) = \mathrm{sign}\left(\sum_{i=1}^{l} y_i\alpha_i K(x, x_i) + b\right) \tag{15}$$

2.2.2. LSSVM for function estimation [29]

For the function estimation problem, given a training data set of l points $\{(x_i, y_i)\}_{i=1,2,\ldots,l}$, with input data $x_i \in R^n$ and output data $y_i \in R$, one considers the following optimization problem in primal weight space:

$$\min_{w,b,e} J(w, b) = \frac{1}{2}w^{\mathrm{T}}w + \gamma\frac{1}{2}\sum_{i=1}^{l}e_i^2 \tag{16}$$

subject to the equality constraint

$$y_i = w^{\mathrm{T}}\varphi(x_i) + b + e_i, \quad i = 1, \ldots, l$$

The constraint is the only difference between the classification and function estimation problem according to Eqs. (12) and (16) that leads to a different Lagrange style for function estimation,

$$L(w, b, e, \alpha) = J(w, e) - \sum_{i=1}^{l}\alpha_i\{w^{\mathrm{T}}\varphi(x_i) + b + e_i + y_i\} \tag{17}$$

The resulting LSSVM model for function estimation becomes

$$f(x) = \sum_{i=1}^{l}\alpha_i K(x, x_i) + b \tag{18}$$

2.2.3. LSSVM implementation

Differently to the Lagrange multipliers, the kernel and its specific parameters together with regularization parameter,

named γ in the LSSVM (Eqs. (12) and (16)), cannot be set from the optimization problem but have to be tuned by the user. These can be optimized by the use of Vapnik–Chervonenkis bounds, cross-validation, an independent optimization set, or Bayesian learning.

In this paper, the following radial basis function (RBF) was used as kernel function:

$$K(x_i, x_j) = \exp\left(-\frac{||x_i - x_j||^2}{2\sigma^2} \right)$$

GA-based search was implemented for selection of the optimum input variable subset and for setting the optimized values of the two parameters: the regularization parameter γ and the width of the RBF kernel σ^2.

A five-fold-out (FFO) cross-validation was used for testing model predictive power. Five data subsets were created, in the cross-validation process four subsets are used for training the LSSVM and the rest subset is then predicted. This process is repeated until all the subsets have been predicted. A ''venetian-blind'' method was used for creating the data subsets. Firstly, dataset are sorted according to the dependent variable and secondly, cases are added to each subset in consecutive order, in such a way that all the subsets are representative samples of the whole dataset. FFO cross-validation mean square error (MSE_{FFO}) was used as cost function for driving the GA search.

$$MSE_{FFO} = \frac{\sum_{i=1}^{n} (t_i - o_i)^2}{n}$$

where t_i are the target outputs (experimental output); o_i the actual outputs; n is the number of samples. LSSVM simulations were implemented using the LSSVM toolbox for Matlab by Pelckmans et al. [30].

2.3. Genetic algorithm (GA) feature selection and hyperparameter optimization

The use of SVM approach for solving classification and function mapping problems in biological QSAR studies has been growing very rapidly in the last years [31]. However, choosing the adequate descriptors for predictor training in QSAR studies is difficult because there are no absolute rules that govern this choice. Recently, evolutionary algorithms and specifically genetic algorithms have been used for variable selection problems [19,32–34]. Deriving an optimal QSAR model through variable selection needs to be addressed, since 143 AASA vectors were available for QSAR analysis and only a subset of them is statistically significant in terms of correlation with the proteins activity. In this sense, linear and non-linear GA searches were carried out for building optimum models.

GAs are governed by biological evolution rules [35]. 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 [36]. In the case, where the number of features to be selected can be known beforehand, as in our study, a binary encoding in the prior way due to the much smaller search space of size would be inefficient. Therefore, in this case, it is reasonable to switch to a decimal encoding indicates the number of the feature, which is selected. Of course, one has to make sure that each is unique in the code [37]. 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 cross-over (cross-over 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.

MSE_{FFO} was used as individual fitness or cost function. The first step is to create a gene pool (population of MRA or SVM 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 cross-validation percent of correct classifications 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 cross-over (cross-over 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 cross-over, 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. We also included elitism which protects the fittest individual in any given generation from cross-over or mutation during reproduction. The genetic content of this individual simply moves on to the next generation intact. This selection, cross-over 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 a 90% of the generations showed the same target fitness score [33].

For non-linear LSSVM models, GA was also used for the optimization of kernel regularization parameter γ and width of an RBF kernel σ^2 as Fröhlich et al. suggested [37]. We can simply concatenate a representation of the parameter to our existing chromosome. That means, we are trying to select an optimal feature subset and an optimal γ at the same time. This is reasonable, because the choice of the parameter is influenced by

the feature subset taken into account and vice versa. Usually, it is not necessary to consider any arbitrary value but only certain discrete values with the form: $n \times 10^k$, where $n = 1, \ldots, 9$ and $k = -3, \ldots, 4$. So, these values can be calculated by randomly generating n and k values as integers between (1, ..., 9) and (−3, ..., 4), respectively. In a similar way, we used GA to optimize the width of an RBF kernel but in this case n and k values were integers between (1, ..., 9) and (−2, ..., 1). Then, our chromosome was concatenate with another gene with discrete values in the interval (0.001–90,000) for encoding the γ parameter and similarly the width of the RBF kernel was encoded in a gene containing discrete values ranging in the interval (0.01–90).

The GA implemented in this paper is a version of a previous reported of our group [34] but incorporating LSSVM hyperparameter optimization that was programmed within the Matlab environment [23] using Genetic Algorithm [38] and LSSVM Toolboxes [30].

2.4. Ghrelin receptor mutants dataset

In an attempt to model the function of ghrelin receptor, we tried to obtain regression models for the change in constitutive and ghrelin-induced activities upon mutations. Ghrelin receptor is a 7TM protein with 366 residues [8]. General serpentine model of the 7TM protein family and helical wheel diagram of the ghrelin receptor are depicted in Fig. 1A and B, respectively. Mutated residues are indicated in black on gray. In white on black are the three Phe residues that were subjected to more elaborate mutagenesis. Mutations are focused on the inner faces of the extracellular ends of TMs III, VI and VII. By means of a GA-based QSAR study we aimed to identify relevant structural features influencing the activity of the ghrelin receptor. In this connection, the novel reported AASA vectors were used for protein structural information encoding [11]. Amino acid sequence of human ghrelin receptor (primary accession number Q92847) was obtained from the Swiss-Prot/

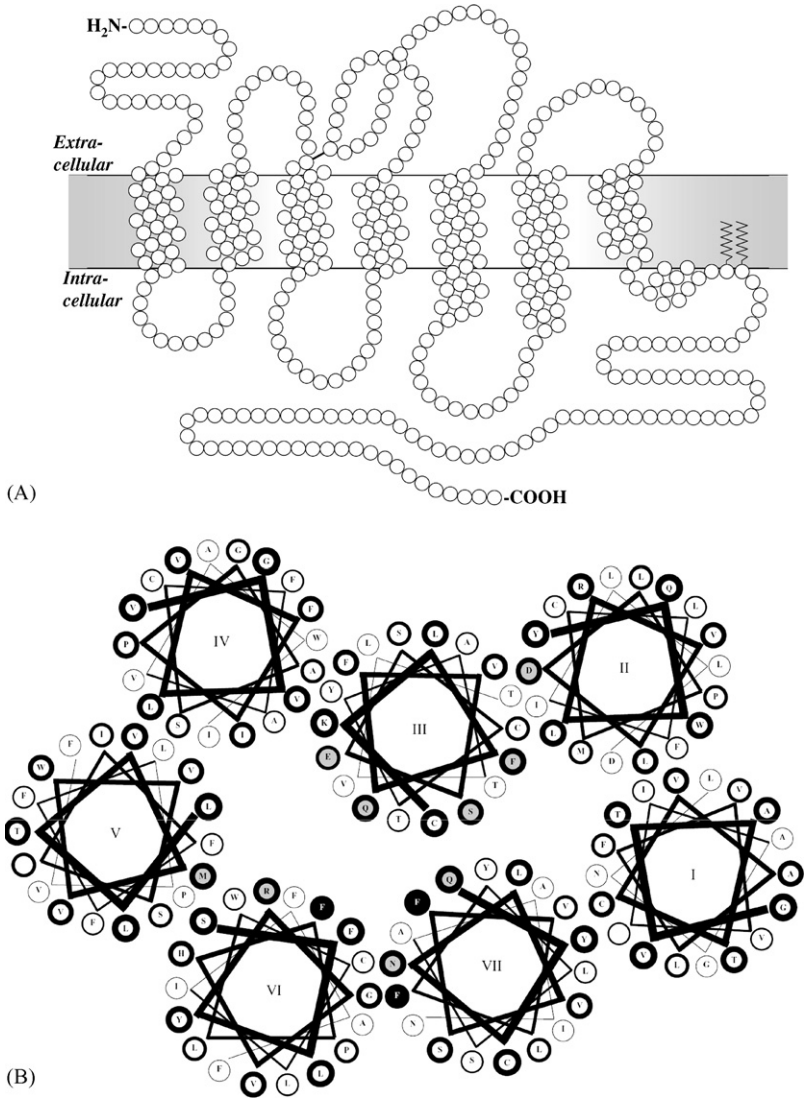

Fig. 1. General serpentine model of the 7TM protein family (A) and helical wheel diagram of the ghrelin receptor (B). Mutated residues are indicated in black on gray. In white on black are the three Phe residues that were subjected to more elaborate mutagenesis.

Table 2
Experimental and predicted constitutive- and ghrelin-induced activities of the ghrelin receptor wild-type and mutants according to LSSVM models

Ghrelin receptors	$Act_{Constitutive}$		$Act_{Ghrelin}$	
	Experimental[a]	GA-LSSVM 1	Experimental[a]	GA-LSSVM 2
Wild-type	2.000	2.040	2.301	2.336
D99N	2.093	2.104	2.344	2.334
FIII:04S	2.009	2.010	2.281	2.225
QIII:05L	1.114	1.120	1.041	1.045
SIII:08A	1.978	1.938	2.328	2.321
EIII:09Q	1.949	2.000	2.276	2.336
E196Q	2.149	2.073	2.380	2.336
R198L	1.763	1.766	2.276	2.261
MV:05A	1.820	1.811	2.297	2.286
FVI:16A	1.176	1.177	2.161	2.196
FVI:16N	1.519	1.511	2.238	2.292
FVI:16Y	1.914	1.945	2.258	2.327
RVI:20A	1.146	1.163	1.204	1.217
QVII:02A	1.898	1.882	2.270	2.238
NVII:02A	1.826	1.868	2.281	2.328
FVII:06A	1.230	1.257	1.204	1.217
FVII:06L	2.127	2.056	2.538	2.336
FVII:06H	1.924	1.956	2.267	2.306
FVII:06Y	1.919	1.921	2.322	2.330
FVII:09A	1.531	1.525	2.324	2.305
FVII:09L	1.568	1.585	2.228	2.336
FVII:09H	1.699	1.690	2.307	2.284
FVII:09Y	2.061	2.035	2.375	2.315

[a] From Ref. [8].

TrEMBL database [39]. Mutant sequences were generated by residue substitution from the wild-type sequence. Constitutive and ghrelin-induced activities of wild-type and 22 ghrelin receptor mutants were collected from the report of Holst et al. [8] (Table 2). Receptor activities are expressed as natural logarithm of the percentage of basal activity of the wild-type ghrelin receptor [8]. Correlation between constitutive and ghrelin-induced activities is depicted in Fig. 2, as can be observed both activities have low intercorrelation (R^2 = 0.624). This fact suggests that mutations on the receptor provoked different effects over each activity.

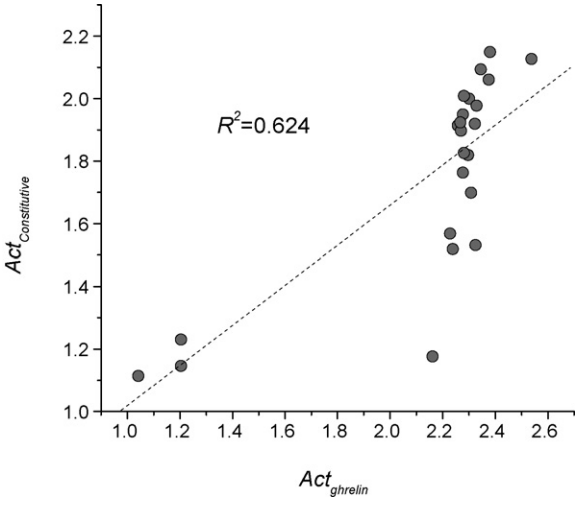

Fig. 2. Correlation plots between constitutive and ghrelin-induced activities of the wild-type and mutant ghrelin receptors.

3. Result and discussion

3.1. GA-based multilinear regression analysis (GA-MRA)

The implemented GA-based multilinear regression (GA-MRA) algorithm searches for the best multilinear regression model, in such a way that from one generation to another the algorithm tries to minimize the MSE_{FFO} (cost function). FFO cross-validation groups were selected according to a ''venetian-blind'' method that assures a very similar range and dispersion of the dependent variable (receptor activity) for each deletion group (see Section 2.2.3). The linear subspaces in the dataset were explored varying the number of variables in the models from 2 to 6.

Optimum model selection was subjected to the principle of parsimony; we chose functions with higher statistical signification but having as few parameters as possible. For that reason, despite to develop several models for the AASA descriptor changing the variables number in every step of the analysis, the best models that we found for the constitutive and ghrelin-induced activity of the ghrelin receptors were described with the following linear equations and with the statistical parameters of the regression presented next.

Constitutive activity of the ghrelin receptor wild-type and mutants:

$$\begin{aligned}\mathrm{Act_{Basal}} &= 17.446\mathrm{AASA}\text{-}T\Delta S_h 3 - 10.735\mathrm{AASA}\text{-}T\Delta S6 \\ &\quad + 7.955\mathrm{AASA}R_\alpha 1 + 25.124\mathrm{AASA}H_t 6 \\ &\quad - 0.334\mathrm{AASA}\Delta C_{ph}1 - 186.979; \qquad N \\ &= 23; \quad R^2 = 0.910; \quad S = 0.111; \quad F \\ &= 34.730; \quad R^2_{\mathrm{FFO}} = 0.834; \quad S_{\mathrm{FFO}} = 0.132 \end{aligned} \tag{19}$$

Ghrelin-induced activity of the ghrelin receptor wild-type and mutants:

$$\begin{aligned}\mathrm{Act_{Ghrelin}} &= 29.526\mathrm{AASA}\,f4 + 1.208\mathrm{AASA}R_f 6 \\ &\quad - 1.329\mathrm{AASA}R_f 5 + 0.407\mathrm{AASA}\Delta C_{ph}6 \\ &\quad - 0.341\mathrm{AASA}\Delta C_{ph}1 - 107.164; \qquad N \\ &= 23; \quad R^2 = 0.941; \quad S = 0.112; \quad F \\ &= 53.691; \quad R^2_{\mathrm{FFO}} = 0.839; \quad S_{\mathrm{FFO}} = 0.163 \end{aligned} \tag{20}$$

where Act_{Basal} and $Act_{Ghrelin}$ are the natural logarithms of the basal and ghrelin-induced activity of the ghrelin receptor wild-type and mutants, respectively, which are expressed as percentage of basal activity of the wild-type ghrelin receptor; N the number of proteins included in the model; R^2 the square of the correlation coefficient; S the standard deviation of the regression; F the Fisher ratio; R^2_{FFO} and S_{FFO} are the square regression coefficient and the standard deviation of the FFO cross-validation, respectively.

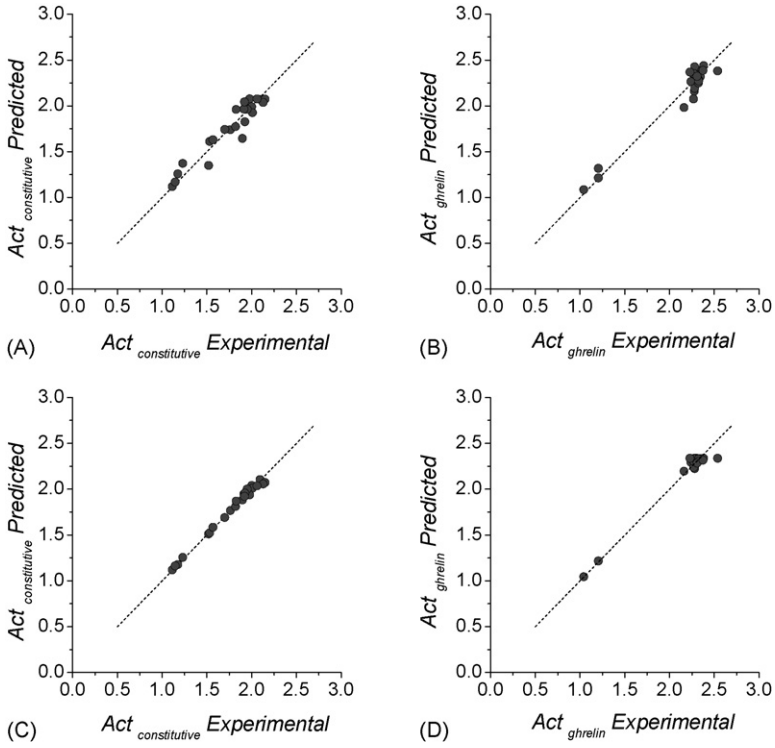

Fig. 3. Plots of predicted vs. experimental constitutive (A) and (C) and ghrelin-induced (B) and (D) activities of the ghrelin receptor according to linear models Eq. (11) (A), Eq. (12) (C) and non-linear models GA-LSSVM 1 (B) and GA-LSSVM 2 (D).

The optimum linear model for the constitutive activity of the ghrelin receptor included five AASA vectors and it is represented by Eq. (19), which has high statistic quality and, more important, good predictive power describing more than an 80% of the FOO cross-validation data variance. Fig. 3A depicts the plots of observed versus predicted values using Eq. (19). Variables in Eq. (19) mean: AASA-$T\Delta S_h$3 is the amino acid sequence autocorrelation at lag 3 of the unfolding entropy change of hydration; AASA-$T\Delta S$6 is the amino acid sequence autocorrelation at lag 6 of the unfolding entropy change; AASAR_a1 is the amino acid sequence autocorrelation at lag 1 of the solvent-accessible reduction ratio; AASAH_t6 is the amino acid sequence autocorrelation at lag 6 of the thermodynamic transfer hydrophobicity and AASAΔC_{ph}1 is the amino acid sequence autocorrelation at lag 1 of the unfolding hydration heat capacity change.

Similarly, the optimum model for the ghrelin-induced activity of the ghrelin receptor has five AASA vectors (Eq. (20)). This linear model also exhibits high statistical quality and predictive power, describing also more than an 80% of the data variance in the cross-validation analysis. Fig. 3 B depicts the plots of observed versus predicted values using Eq. (20). Variables in Eq. (20) mean: AASAf4 is the amino acid sequence autocorrelation at lag 4 of the flexibility (number of side-chain dihedral angles); AASAR_f6 is the amino acid sequence autocorrelation at lag 6 of the chromatographic index; AASAR_f5 is the amino acid sequence autocorrelation at lag 5 of the chromatographic index; AASAΔC_{ph}5 is the amino acid sequence autocorrelation at lag 5 of the unfolding hydration heat capacity change and AASAΔC_{ph}1 is the amino acid sequence autocorrelation at lag 1 of the unfolding hydration heat capacity change.

Despite different optimum subsets of AASA vectors appear in Eqs. (19) and (20) and only one descriptor (AASAΔC_{ph}1) coincides in both models, the occurrence in both equations of several amino acid/residue properties ($-T\Delta S_h$, R_a, H_t, ΔC_{ph}, R_f) accounting for the hydrophobic features of the residues suggested that differences in both constitutive and ghrelin-induced activities among ghrelin receptor wild-type and mutants are mainly due to hydrophobicity change after mutations. This fact was expectable since the most of the point mutations here studied correspond to substitution of aromatic residue phenylalanine especially at VI and VII TM regions [8] (Fig. 1, Table 2). Interesting, occurrence in both models of the amino acid sequence autocorrelation at lag 1 of the unfolding hydration heat capacity change (AASAΔC_{ph}1) suggests that an optimum distribution at lag 1 of residues resembling a specific hydration pattern that is essential for the receptor activity. However, it is noteworthy that a steric-related property f, flexibility (number of side-chain dihedral angles), appears in Eq. (20) weighting an autocorrelation vector of lag 4 (AASAf4). This is the main difference between both linear models and suggests that in the case of ghrelin-induced activity the receptor should interact with the ghrelin agonist adopting and optimum conformation. Consequently, the receptor activity after mutations is affected not only by hydrophobicity changes but also by steric changes upon mutations in the side-chains of the inner faces of the TM regions.

3.2. GA-based least-square support vector machine (GA-LSSVM) modeling

In this study, we aimed to search for reliable non-linear models of the constitutive and ghrelin-induced activity of the ghrelin receptor upon mutations. In this sense, besides the linear models obtained by GA-MRA we used a GA-based LSSVM approach for building optimum non-linear models. Non-linearity was accessed by using a RBF kernel inside the SVM framework. As was point out in Section 2.3, in the case of the LSSVM modeling, the GA algorithm was also used for optimizing the hyperparameters, the kernel regularization parameter γ and the width of an RBF kernel σ^2. Similarly to the linear models, non-linear subspace in the dataset was searched varying problem dimension from 2 to 6. FFO cross-validation subsets were selected according to ''venetian-blind'' method and the MSE_{FFO} was minimized throughout the GA search.

Fig. 4A depicts the behaviour of the FFO cross-validation versus the number of variables in the LSSVM models for the constitutive activity. As can be observed maximum explained variance of the FFO cross-validation was achieved for models having four AASA vectors. This maximum value was about 88% overcoming the linear model presented in Eq. (19) even having one descriptor less. Fig. 3C depicts the plots of observed versus predicted values using GA-LSSVM 1 model fitting the dataset with a value of $R^2 = 0.996$. The optimum non-linear predictor yielded better results suggesting that a non-linear relationship among the AASA vectors and the constitutive activity of the ghrelin receptor is more adequate than a linear one. AASA vectors in the GA-LSSVM 1 predictor for the constitutive activity of ghrelin receptor appear in Table 3 as well as the hyperparameters and statistical quantities of such model. In the optimum non-linear predictor two variables, AASAH_t6 and AASA-$T\Delta S_h$3, also appear in Eq. (19). The other two variables mean: AASAH_t1 is the amino acid sequence autocorrelation at lag 1 of the thermodynamic transfer hydrophobicity and AASAΔG_c3 is the amino acid sequence autocorrelation at lag 3 of the unfolding Gibbs free energy changes of side-chain. The optimum non-linear model GA-LSSVM 1 also reflects a great influence of hydrophobicity-related properties (H_t and $-T\Delta S_h$) in the receptor constitutive activity. The model exhibits influence of both short (lags 1 and 3) and medium (lag 6) range hydrophobic interactions at protein sequence.

In the case of non-linear modeling of the ghrelin-induced receptor activity, Fig. 4B depicts the behaviour of the model FFO cross-validation versus the number of variables in the LSSVM models. It is remarkable that the optimum LSSVM model has only two inputs and describes about 93% of the data variance in the cross-validation analysis, widely overcoming the linear model for the ghrelin-induced receptor activity in Eq. (20). Fig. 3D depicts the plots of observed versus predicted values using GA-LSSVM 2 model fitting the dataset with a

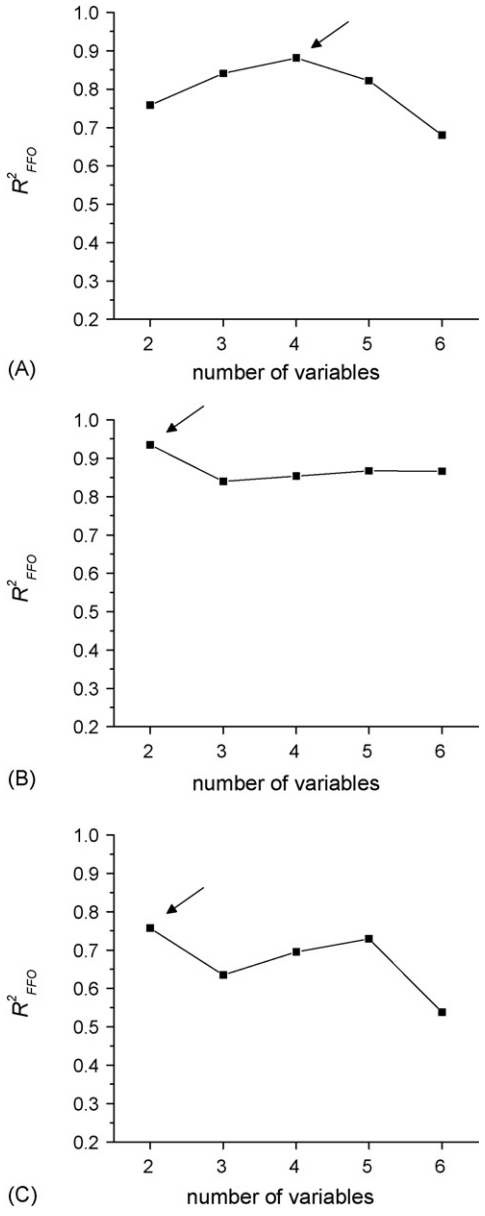

Fig. 4. Plots of the FFO variance described throughout de GA-LSSVM search vs. the number of variables in the models for the constitutive (A), ghrelin-induced (B) and coupled (C) ghrelin receptor activities. Arrows point out the optimum number of variables.

Table 3
Optimum GA-LSSVM models for the constitutive (GA-LSSVM 1), ghrelin-induced (GA-LSSVM 2) and coupled (GA-LSSVM 3) ghrelin receptor activities

Variables	Model	γ	σ^2	R^2	S	R^2_{FFO}	S_{FFO}
AASAH_t6; AASA-$T\Delta S_h$3; AASAH_t1; AASAΔG_c	GA-LSSVM 1	500	10	0.991	0.032	0.881	0.113
AASAΔC_{ph}1; AASAs4	GA-LSSVM 2	80,000	60	0.977	0.035	0.935	0.112
AASA-$T\Delta S_h$3; AASA-$T\Delta S_c$4	GA-LSSVM 3	300	2	0.996	0.027	0.757	0.305

γ is the LSSVM regularization parameter; σ^2 is the width of the RBF kernel using in the LSSVM; R^2 is the square of the correlation coefficient; S is the standard deviation of the regression; R^2_{FFO} and S_{FFO} are the square regression coefficient and the standard deviation of the FFO cross-validation, respectively.

value of $R^2 = 0.977$. It is remarkable the accuracy obtaining even when three mutants (QIII:05L, RVI:20A, FVII:06A) have low activities, very different from the rest of the data. The GA-LSSVM 2 model is able of accurately predicts the activity of each one of these cases by learning the activity pattern from the other two. Indeed, according to the data distribution in the five subsets yielded by the ''venetian-blind'' method for cross-validation, always two of the less active mutants were used for LSSVM training meanwhile another was predicted.

Input variables in GA-LSSVM 2 predictor appear in Table 3 as well as the hyperparameters and statistical quantities. Interestingly, a variable in Eqs. (19) and (20), AASA$\Delta C_{ph}1$, is also present in model GA-LSSVM 2. As we previously pointed out this AASA vector encoded the hydration pattern of the ghrelin receptor that directly impact in its activity. The other autocorrelation vector in the model is AASA*s*4, which means the autocorrelation at spatial lag 4 of the shape (position of branch point in a side-chain). So, ghrelin-induced activity was also found depending on a non-linear way of a descriptor accounting for steric variations upon mutations that it is weighted by *s*, shape (position of branch point in a side-chain). This result also support the previous finding in Eq. (20) that ghrelin-induced activity of the ghrelin receptor, beside of the hydrophobicity distribution, is also dependent of the conformations of the residue side-chains of the inner faces of the TM regions.

Both non-linear SVM models over-performed their correspondent linear equations; however the most outstanding model was obtained for the ghrelin-induced activity. By including only two AASA vectors this model describes near the 95% of FFO cross-validation variance. The graphical representation of this two-dimensional predictor is depicted in Fig. 5. This graph renders the ghrelin-induced activity surface of the wild-type and mutant ghrelin receptors. AASA vectors showed to encode relevant non-linear information about the ghrelin receptor activity. The activity pattern depicted in Fig. 4 resembled the variation of the receptor activity according to the selected optimum autocorrelation vectors, AASA$\Delta C_{ph}1$ and AASA*s*4. The first descriptor encodes the hydrophobic interactions between residues at the shortest range and the second one the steric interaction at medium range in the protein sequence.

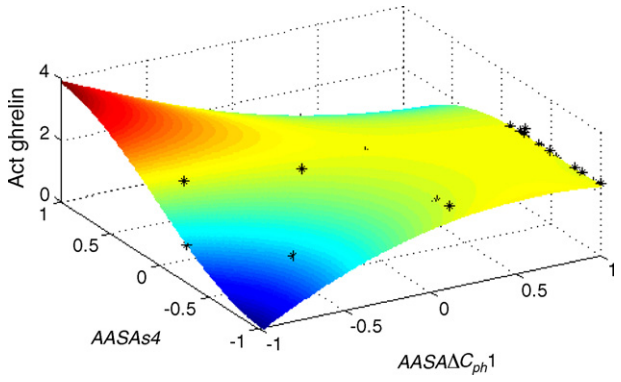

Fig. 5. Ghrelin-induced activity surface generated by the two-dimensional predictor GA-LSSVM 2. AASA vector values are normalized in the range (−1, 1).

Finally, GA-LSSVM algorithm was also used for building a general model of the ghrelin receptor activity by searching a sole descriptor subset that well describes both constitutive and ghrelin-induced activities. The behaviour of the FFO cross-validation for the coupled-activity model versus the number of inputs throughout the GA search is depicted in Fig. 4C. As can be observed, the optimum predictor was found with two input variables but the explained FFO cross-validation variance about 74% was very poor in comparison with the independent predictors, GA-LSSVM 1 and GA-LSSVM 2. Inputs variables, hyperparameters and statistical quantities of the optimum coupled-data predictor GA-LSSVM 3 appear in Table 3. The input variables in this predictor are weighted by $-T\Delta S_h$ and $-T\Delta S_c$, the unfolding entropy change of hydration and the unfolding entropy change of protein side-chain, respectively. This fact suggests that when we modelled both activities the main driving force of the model is the entropy change upon mutations on the ghrelin receptor. Despite the relevant contribution of the hydrophobicity, the coupled model shows that receptor interaction is also highly dependent of the protein overall entropy.

The ghrelin receptor mutation series here studied mainly represents variations in the TM regions of the protein. However, the autocorrelation formalism applied over all the protein sequence well described functional variations upon mutations. According to the models developed, the main driving force of the receptor activity was the hydrophobic interactions. This fact well agrees with the proposed models for the ghrelin receptor interactions [8]. But interesting, the ghrelin-induced activity showed according to both linear and non-linear models to be highly dependent also of the steric conformation of the receptor, this fact suggested that agonist-receptor interactions not only depends on the intensity of hydrophobic interactions but also on an specific conformational matching. Similarly, the relevance of the protein entropy also suggested that certain freedom degree of the side-chains in the TM region is essential for the active form of the receptor. In this sense, the mutational analysis in Ref. [8] showed that the constitutive activity of the ghrelin receptor could systematically be tuned up and down depending on the size and hydrophobicity of the side-chain in position VI:16 in the context of an aromatic residue at VII:09 and a large hydrophobic residue at VII:06 [8].

In systematic search among subjects with short stature, it has been found identical mutations in the ghrelin receptor that carried out lacks in constitutive activity. This phenomenon has been associated to obesity propensity in humans. Authors had suggested that selective loss of ghrelin receptor constitutive activity causes a syndrome of short stature and obesity, of which the obesity appears to develop around puberty [7,8]. The models here developed despite the fact that its range of application is limited for the reduced mutant data here used can be useful for estimation of functional variation of the ghrelin receptor upon mutations. The main advantages of the approach here presented is that non-three-dimensional structure of the receptor protein was needed so only sequence information was used for building high quality predictive models.

4. Conclusions

Functional variations induced by mutations are the main causes of several genetic pathologies and syndromes. Due to the availability of some functional variation data upon mutations of some proteins it is possible to use function–structure relationship approach for protein function 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 proteometrics studies. 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. The autocorrelation formalism that was previously applied to protein stability study was now successfully used for 7TM protein function prediction.

The GA-LSSVM algorithm present here successfully allowed simultaneously optimizing the input variable subset and the LSSVM hyperparameter. This approach showed to be a powerful technique for feature selection and mathematical modeling by yielding reliable and robust non-linear models for the constitutive and ghrelin-induced activities of the ghrelin receptor upon mutations that describe about 88% and 93% of FFO cross-validation, respectively.

The present work demonstrates the successful application of the AASA vectors to the modeling of protein function in combination with GA-LSSVM 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.

References

[1] T. Costa, S. Cotecchia, Historical review: negative efficacy and the constitutive activity of Gprotein-coupled receptors, Trends Pharmacol. Sci. 26 (2005) 618–624.

[2] T.W. Schwartz, T.M. Frimurer, B. Holst, M.M. Rosenkilde, C.E. Elling, Molecular mechanism of 7TM receptor activation: a global toggle switch model, Annu. Rev. Pharmacol. Toxicol. 46 (2006) 481–519.

[3] (a) R.A. Adan, M.J. Kas, Inverse agonism gains weight, Trends Pharmacol. Sci. 24 (2003) 315–321;
(b) B. Holst, A. Cygankiewicz, J.T. Halkjar, M. Ankersen, T.W. Schwartz, High constitutive signaling of the ghrelin receptor-identification of a potent inverse agonist, Mol. Endocrinol. 17 (2003) 2201–2210;
(c) C. Leterrier, D. Bonnard, D. Carrel, J. Rossier, Z. Lenkei, Constitutive endocytic cycle of the CB1 cannabinoid receptor, J. Biol. Chem. 279 (2004) 36013–36021.

[4] (a) M. Korbonits, A.P. Goldstone, M. Gueorguiev, A.B. Grossman, Ghrelin: a hormone with multiple functions, Front. Neuroendocrinol. 25 (2004) 27–68;
(b) J.A. van der Lely, M. Tschop, M.L. Heiman, E. Ghigo, Biological, physiological, pathophysiological, and pharmacological aspects of ghrelin, Endocrinol. Rev. 25 (2004) 426–457.

[5] (a) T.W. Holst, Schwartz, Constitutive ghrelin receptor activity as a signaling set-point in appetite regulation, Trends Pharmacol. Sci. 25 (2004) 113–117;
(b) K.L. Grove, M.A. Cowley, Is ghrelin a signal for the development of metabolic systems? J. Clin. Invest. 115 (2005) 3393–3397.

[6] G. Burdyga, S. Lal, A. Varro, R. Dimaline, D.G. Thompson, G.J. Dockray, Expression of cannabinoid CB1 receptors by vagal afferent neurons is inhibited by cholecystokinin, J. Neurosci. 24 (2004) 2708–2715.

[7] J. Pantel, M. Legendre, S. Cabrol, L. Hilal, Y. Hajaji, S. Morisset, S. Nivot, M.P. Vie-Luton, D. Grouselle, M. de Kerdanet, A. Kadiri, J. Epelbaum, Y. Le Bouc, S. Amselem, Loss of constitutive activity of the growth hormone secretagogue receptor in familial short stature, J. Clin. Invest. 116 (2006) 760–768.

[8] B. Holst, N.D. Holliday, A. Bach, C.E. Elling, H.M. Cox, T.W. Schwartz, Common structural basis for constitutive activity of the ghrelin receptor family, J. Biol. Chem. 279 (2004) 53806–53817.

[9] S.J.E. Wikberg, M. Lapinsh, P. Prusis, Proteochemometrics: a tool for modelling the molecular interaction space, in: H. Kubinyi, G. Müller (Eds.), Chemogenomics in Drug Discovery—A Medicinal Chemistry Perspective, Wiley–VCH, Weinheim, 2004, pp. 289–309.

[10] (a) R. Ramos de Armas, H. González-Díaz, R. Molina, E. Uriarte, Markovian backbone negentropies: molecular descriptors for protein research. I. Predicting protein stability in arc repressor mutants, Proteins 56 (2004) 715–723;
(b) H. González-Díaz, R. Molina, E. Uriarte, Recognition of stable protein mutants with 3D stochastic average electrostatic potentials, FEBS Lett. 579 (2005) 4297–4301;
(c) Y. Marrero-Ponce, R. Medina-Marrero, J.A. Castillo-Garit, V. Romero-Zaldivar, F. Torrens, E.A. Castro, Protein linear indices of the 'macromolecular pseudograph α-carbon atom adjacency matrix' in bioinformatics. Part 1. Prediction of protein stability effects of a complete set of alanine substitutions in arc repressor, Bioorg. Med. Chem. 13 (2005) 3003–3015;
(d) G. Agüero-Chapin, H. González-Díaz, R. Molina, J. Varona Santos, E. Uriarte, Y. González-Díaz, Novel 2D maps and coupling numbers for protein sequences. The first QSAR study of polygalacturonase; isolation and prediction of a novel sequence from *Psidium guajava* L., FEBS Lett. 580 (2006) 723–730.

[11] J. Caballero, L. Fernández, J.I. Abreu, M. Fernández, Amino acid sequence autocorrelation vectors and ensembles of Bayesian-regularized genetic neural networks for prediction of conformational stability of human lysozyme mutants, J. Chem. Inf. Model. 46 (2006) 1255–1268.

[12] (a) K. Nakai, A. Kidera, M. Kanehisa, Cluster analysis of amino acid indices for prediction of protein structure and function, Protein Eng. 2 (1988) 93–100;
(b) K. Tomii, M. Kanehisa, Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins, Protein Eng. 9 (1996) 27–36;
(c) S. Kawashima, M. Kanehisa, AAindex: amino acid index database, Nucleic Acids Res. 28 (2000) 374.

[13] H. Bauknecht, A. Zell, H. Bayer, P. Levi, M. Wagener, J. Sadowski, J. Gasteiger, Locating biologically active compounds in medium-sized heterogeneous datasets by topological autocorrelation vectors: dopamine and benzodiazepine agonists, J. Chem. Inf. Comput. Sci. 36 (1996) 1205–1213.

[14] P.A.P. Moran, Notes on continuous stochastic processes, Biometrika 37 (1950) 17–23.

[15] R.F. Geary, The contiguity ratio and statistical mapping, Incorp. Stat. 5 (1954) 115–145.

[16] G. Moreau, P. Broto, Autocorrelation of a topological structure: a new molecular descriptor, Nouv. J. Chim. 4 (1980) 359–360.

[17] G. Moreau, P. Broto, Autocorrelation of molecular structures: application to SAR studies, Nouv. J. Chim. 4 (1980) 757–764.

[18] M. Wagener, J. Sadowski, J. Gasteiger, Autocorrelation of molecular properties for modelling corticosteroid binding globulin and cytosolic Ah receptor activity by neural networks, J. Am. Chem. Soc. 117 (1995) 7769–7775.

[19] (a) M. Fernández, A. Tundidor-Camba, J. Caballero, 2D autocorrelation modeling of the activity of trihalobenzocycloheptapyridine analogues as farnesyl protein transferase inhibitors, Mol. Simulat. 31 (2005) 575–584;
(b) J. Caballero, M. Garriga, M. Fernández, 2D autocorrelation modeling of the negative inotropic activity of calcium entry blockers using Bayesian-regularized genetic neural networks, Bioorg. Med. Chem. 14 (2006) 3330–3340;

(c) M. Fernández, J. Caballero, Bayesian-regularized genetic neural networks applied to the modelling of non-peptide antagonists for the human luteinizing hormone-releasing hormone receptor, J. Mol. Graph. Model. 25 (2006) 410–422.

[20] M.M. Gromiha, M. Oobatake, H. Kono, H. Uedaira, A. Sarai, Importance of surrounding residues for protein stability of partially buried mutations, J. Biomol. Struct. Dyn. 18 (2000) 1–16.

[21] M.M. Gromiha, M. Oobatake, H. Kono, H. Uedaira, A. Sarai, Relationship between amino acid properties and protein stability: buried mutations, J. Prot. Chem. 18 (1999) 565–578.

[22] M.M. Gromiha, A.M. Thangakani, S. Selvaraj, FOLD-RATE: prediction of protein folding rates from amino acid sequence, Nucleic Acids Res. 34 (2006) 70–74.

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

[24] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297.

[25] C.J.C. Burges, A tutorial on support vector machines for pattern recognition, Data Min. Knowl. Discov. 2 (1998) 1–47.

[26] V. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.

[27] J.A.K. Suykens, J. Vandewalle, Least squares support vector machine classifiers, Neural Process. Lett. 9 (1999) 293–300.

[28] (a) S. Viaene, B. Baesens, T. Van Gestel, J.A.K. Suykens, D. Van den Poel, J. Vanthienen, B. De Moor, G. Dedene, Knowledge discovery in a direct marketing case using least squares support vector machines, Int. J. Intell. Syst. 16 (2001) 1023–1036;
(b) K.S. Chua, Efficient computations for large least square support vector machine classifiers, Pattern Recognit. Lett. 24 (2003) 75–80.

[29] (a) J.A.K. Suykens, J. Vandewalle, Chaos control using least-squares support vector machines, Int. J. Circ. Theor. Appl. 27 (1999) 605–615;
(b) J.A.K. Suykens, J. Vandewalle, B. De Moor, Optimal control by least squares support vector machines, Neural Netw. 14 (2001) 23–35.

[30] K. Pelckmans, J.A.K. Suykens, T. Van Gestel, D. De Brabanter, L. Lukas, B. Hamers, B. De Moor, J. Vandewalle, LS-SVMlab: a Matlab/C Toolbox for Least Squares Support Vector Machines, Internal Report 02-44, ESATSISTA, K.U. Leuven: Leuven, 2002.

[31] (a) W. Lua, N. Donga, G. Náray-Szabó, Predicting anti-HIV-1 activities of HEPT-analog compounds by using support vector classification, QSAR Combust. Sci. 24 (2005) 1021–1025;
(b) X. Yao, H. Liu, R. Zhang, M. Liu, Z. Hu, A. Panaye, J.P. Doucet, B. Fan, QSAR and classification study of 1,4-dihydropyridine calcium channel antagonists based on least squares support vector machines, Mol. Pharm. Vol. 2 (2005) 348–356;
(c) H. Fröhlich, J.K. Wegner, A. Zell, Towards optimal descriptor subset selection with support vector machines in classification and regression, QSAR Combust. Sci. 23 (2004) 311–318.

[32] S. So, M. Karplus, Evolutionary optimization in quantitative structure–activity relationship: an application of genetic neural networks, J. Med. Chem. 39 (1996) 1521–1530.

[33] B. Hemmateenejad, M.A. Safarpour, R. Miri, N. Nesari, Toward an optimal procedure for PC-ANN model building: prediction of the carcinogenic activity of a large set of drugs, J. Chem. Inf. Model. 45 (2005) 190–199.

[34] (a) J. Caballero, M. Garriga, M. Fernández, Genetic neural network modeling of the selective inhibition of the intermediate-conductance Ca^{2+}-activated K^+ channel by some triarylmethanes using topological charge indexes descriptors, J. Comput. Aid. Mol. Des. 19 (2005) 771–789;
(b) J. Caballero, M. Fernández, Linear and nonlinear modeling of antifungal activity of some heterocyclic ring derivatives using multiple linear regression and Bayesian-regularized neural networks, J. Mol. Model. 12 (2006) 168–181.

[35] H. Holland, Adaptation in Natural and Artificial Systems, The University of Michigan Press, Ann Arbor, MI, 1975.

[36] H.M. Cartwright, Applications of Artificial Intelligence in Chemistry, Oxford University Press, Oxford, 1993.

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

[38] The MathWorks Inc, Genetic Algorithm and Direct Search Toolbox User's Guide for Use with MATLAB, The Mathworks Inc., Massachusetts, 2004

[39] A. Bairoch, R. Apweiler, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M.J. Martin, D.A. Natale, C. O'Donovan, N. Redaschi, L.S. Yeh, The Universal Protein Resource (UniProt), Nucleic Acids Res. 33 (2005) 154–159.