Amino Acids (2010) 38:199–212
DOI 10.1007/s00726-008-0228-1

ORIGINAL ARTICLE

Gaussian process: an alternative approach for QSAM modeling of peptides

Peng Zhou · Xiang Chen · Yuqian Wu · Zhicai Shang

Received: 1 August 2008 / Accepted: 18 December 2008 / Published online: 4 January 2009

Abstract Different statistical modeling methods (SMMs) are used for nonlinear system classification and regression. On the basis of Bayesian probabilistic inference, Gaussian process (GP) is preliminarily used in the field of quantitative structure-activity relationship (QSAR) but has not yet been applied to quantitative sequence-activity model (QSAM) of biosystems. This paper proposes the application of GP as an alternative tool for the QSAM modeling of peptides. To investigate the modeling performance of GP, three classical peptide panels were used: Angiotensin-I converting enzyme inhibitory dipeptides, bradykinin-potentiating pentapeptides and cationic antimicrobial pentadecapeptides. On this basis, we made a comprehensive comparison between the GP and some widely used SMMs such as PLS, artificial neural network (ANN) and support vector machine (SVM), and gave the conclusions as follow: (1) for those of structurally complicated peptides, particularly the polypeptides, linear PLS was incapable of capturing all dependences hidden in the peptide systems, (2) even in assistance with the monitoring technique, ANN was inclined to be overtrained in the cases of insufficient number of peptide samples, (3) SVM and GP performed best for the three peptide panels. Moreover, since GP was able to correlate the linear and nonlinear-hybrid relationship, it was slightly superior to SVM at most peptide sets.

Keywords Gaussian process · Statistical modeling method · Quantitative sequence-activity model · Peptide · Amino acid descriptor

Introduction

Initially proposed by Jonsson et al. (1993), quantitative sequence-activity model (QSAM) is the subject that employs quantitative structure-activity relationship (QSAR) strategies to quantify biosequence-activity/function relationship for the peptides, proteins and nucleic acids (Zhou et al. 2008a; Dea-Ayuela et al. 2008; Gonzalez-Diaz et al. 2007, 2008). Early in 1966, pioneering work was made by Sneath (1966) who derived amino acid descriptors from qualitative (interval) data for the 20 coded amino acids. After that, Kidera et al. (1985) coded the amino acids using ten orthogonal factors derived from factor analysis (FA) of 188 properties. Hellberg et al. (1986) employed principal component analysis (PCA) to extract important information of amino acids, deriving famous amino acid principal property z scales which later were widely applied in peptide activity prediction (Jenssen et al. 2005; Wu et al. 2006), protein design (Genst et al. 2002; Freyhult et al. 2003) and protein-peptide binding affinity analysis (Guan et al. 2005). Sandberg et al. (1998) further extended z scales to 87 amino acids (20 coded ones plus 67 non-coded ones), this extended z scales were thus possible to structurally characterize peptides and proteins containing non-coded amino acids. Other frequently used amino acid descriptors of isotropic surface area and electronic charge index (ISA-ECI) were obtained by theoretical analysis of amino acid size and

Electronic supplementary material The online version of this article (doi:10.1007/s00726-008-0228-1) contains supplementary material, which is available to authorized users.

P. Zhou · X. Chen · Z. Shang (✉)
Department of Chemistry, Zhejiang University,
310027 Hangzhou, China
e-mail: shangzc@zju.edu.cn

Y. Wu
Institute of Agricultural and Life Sciences,
Chongqing University, 400044 Chongqing, China

electron distribution (Collantes and Dunn 1995). ISA-ECI was successfully applied into peptide library design (Cho et al. 1998), epitope identification (Lin et al. 2004) and molecular descriptor construction (Armas et al. 2005). Recently, we have proposed *T* scale (Tian et al. 2007a), 3D-HoVAIF (Tian et al. 2007b) and nonbonding interaction analysis (Zhou et al. 2007), using these methods we performed a series of QSAM studies on various peptide systems such as CTL epitopes, bitter-tasting dipeptides and thromboplastin inhibitors.

The previous studies of QSAM mainly concentrated on structural characterization of biomolecules and development of novel biosequence descriptors, laying less emphasis on the application of new statistical modeling methods (SMMs) to ascertain the complex relationship between the bio-structures/sequences and their functions/activities. At present, standard QSAM modeling tool is the partial least square (PLS) (Wold et al. 1984). Despite this method is able to handle the small-sample, high-dimensional and strong collinear data, it only derives the linear models and thus cannot be used to modeling complex biosystems. Several machine learning algorithms have already been successfully applied in the QSAM field. In the early stage, artificial neural network (ANN) was the main tool to perform structure-activity studies and sequence analysis for biomolecules, and some active oligopeptides were designed by ANN approach (Schneider et al. 1998; Patel et al. 1998). Hereafter, with the rapid development of support vector machine (SVM), SVM-based QSAM protocols were carried out on various biological sets of *Escherichia coli* promoters (Kiryu et al. 2005), MHC-restricted peptides (Liu et al. 2006) and proteins (Ladiwala et al. 2006), suggesting SVM is a powerful tool in both classification and regression for the complex problems. Presently, online SVM identification and prediction of MHC-binding peptides are available (Tung and Ho 2007). In addition, some other successful cases such as conformation searching (Wilson and Cui 2004), combinatorial peptide library design (Zhou et al. 2006) and bioactivity prediction (Udaka et al. 2002) were fulfilled by simulated annealing (SA), genetic algorithm (GA) and hidden Markov model (HMM). These works significantly improved the development of QSAM and bioinformatics.

In the current study, we introduced a new machine learning method that was preliminarily used in QSAR field, called the Gaussian process (GP) (Rasmussen and Williams 2006). Pioneering works were made by Burden (2001) who demonstrated GP applications in QSAR modeling of three sample panels of benzodiazepine, substituted benzene and muscarinic datasets. After that Enot et al. (2001), Tino et al. (2004), Schwaighofer et al. (2007) and Schroeter et al. (2007) used GP to successfully perform statistical predictions for a series of pharmacokinetic properties such as lipophilicity, solubility and lipophilicity, etc. Recently, Obrezanova and co-workers (2007) adopted GP to implement automatic QSAR modeling of ADME properties. Based on GP, Ažman and Kocijan (2007) addressed simulation of the nitrification process in a wastewater treatment plant and biomass growth in the Lagoon of Venice. In chemometrics, Chen et al. (2007) fulfilled multivariate spectroscopic calibration using GP regression approach. All these works confirmed that GP is a promising machine learning tool that can be used to information mining for complex chemical and biological systems. However, published works on GP applications into QSAM are yet in absence. In view of that, this study is dedicated to introducing GP in QSAM modeling of bioactive peptides. We made a comprehensive comparison of GP with several frequently used SMMs of PLS, ANN and SVM, and some empirical rules about GP applications in QSAM modeling of peptides were also suggested.

Methods

GP for regression

The regression problem is addressed as follows. We have a training set $\mathbf{D}$ of n observations, $\mathbf{D} = \{\mathbf{X}, \mathbf{y}\}$, where vector $\mathbf{y} = \{y^{(i)}\}_{i=1}^{n}$ is the set of observed activities and matrix $\mathbf{X} = \{\mathbf{x}^{(i)}\}_{i=1}^{n}$ is the set of the peptide descriptors, then we want to find a function $f(\mathbf{x})$, which is associated with each training sample $\mathbf{x}^{(i)}$, and that the function values $f(\mathbf{x}^{(i)})$ preserve the preference relations observed $y^{(i)}$ in the dataset $\mathbf{D}$. Conventionally, one function form (namely basis function) is selected with free parameters, for example by only considering linear functions of the input, and then function parameters are determined on the basis of observed data. This approach has an obvious problem in that if the underlying relation between descriptors and activities is not well consistent with the considered function form, then the predictions will be poor (Rasmussen and Williams 2006). Here, we introduce a different modeling pathway as Gaussian process (GP). The GP method is introduced by taking a Bayesian nonparametric perspective on the formulation of the basis function regression model, which means that the actual number of "hyperparameters" (distinguishing with "parameters" of basis function in conventional methods) required scale linearly with the number of inputs being processed. In spite of this inconvenient feature, the flexibility and transparency of the modeling makes it an attractive method being intensively studied (Zhou et al. 2008b).

Initially proposed by O'Hagan, GP is based on casting the problem of building a model for some data in the form of a Bayesian inference (O'Hagan 1978). Bayes' theorem

updates probabilities given new evidence (observed data) in the following way:

$$P(f(\mathbf{x})|\mathbf{D}) \propto P(\mathbf{y}|f(\mathbf{x}),\mathbf{X})P(f(\mathbf{x})) \quad (1)$$

where $P(f(\mathbf{x})|\mathbf{D})$ is the posterior, $P(\mathbf{y}|f(\mathbf{x}),\ \mathbf{X})$ is the likelihood, and $P(f(\mathbf{x}))$ is the prior. In GP, it is assumed that $f(\mathbf{x})$ is a random function, where functional values $f(\mathbf{x}^{(1)}), f(\mathbf{x}^{(2)}), \ldots, f(\mathbf{x}^{(n)})$ for any finite set of n points form a Gaussian distribution. A GP is completely specified by its mean zero and $n \times n$ covariance matrix $\mathbf{P}$ for noise-free function values $f(\mathbf{x})$:

$$f(\mathbf{x}) \sim N(0, \mathbf{P}) \quad (2)$$

Based upon that, we have further assumed that the observed activities $\mathbf{y}$ differ from the function values $f(\mathbf{x})$ by additive noise which follows an independent and identical normal distribution with zero mean and variance σ_v^2. In reference to Eq. 2, the prior distribution for the observed activities y is described as following:

$$\mathbf{y} \sim N(0, \mathbf{C}) \quad (3)$$

where $\mathbf{C}$ is the covariance matrix for the noisy observed activities $\mathbf{y}$, $\mathbf{C} = \mathbf{P} + \sigma_v^2\mathrm{I}$, I is an identity matrix. $\mathbf{C}$ defines the similarity between different peptides in the input space (see below), with the matrix element $\mathrm{Cov}(y^{(i)}, y^{(j)})$ $(i, j \in 1, 2, \ldots, n)$ denoting the covariance of peptide pair i–j. Here, based on a set of n training points in dataset $\mathbf{D} = \{\mathbf{X}, \mathbf{y}\}$, we wish to find the predictive distribution of y^* corresponding to a new given input $\mathbf{x}^*$. For the collection of random variables $(y_1, y_2, y_3, \ldots, y_n, y^*)$ we can write joint distribution:

$$\begin{pmatrix} \mathrm{y} \\ y^* \end{pmatrix} \sim N(0, \mathrm{C}^*) \quad (4)$$

where $\mathbf{C}^*$ is a $(n + 1) \times (n + 1)$ covariance matrix:

$$\mathrm{C}^* = \begin{bmatrix} [\mathrm{C}], & [\mathrm{k}] \\ [\mathrm{k}^{\mathrm{T}}], & [\kappa] \end{bmatrix} \quad (5)$$

where $\mathbf{k}$ denote the vector of covariance between the new point $\mathbf{x}^*$ and the n training points $\mathbf{X}$, and κ is the autocovariance of $\mathbf{x}^*$. Thus, we can obtain a prediction of the GP model at the input $\mathbf{x}^*$. Dividing the joint distribution (Eq. 4) by the training point distribution (Eq. 3), we obtain the predictive distribution of y^* (Ažman and Kocijan 2007):

$$\mathrm{P}(y^*) = \frac{\mathrm{P}\begin{pmatrix} \mathrm{y} \\ y^* \end{pmatrix}}{\mathrm{P}(\mathrm{y})} \quad (6)$$

It can be shown that this is a Gaussian distribution:

$$y^* \sim N(\mathrm{E}(y^*),\ \mathrm{V}(y^*)) \quad (7)$$

where $\mathrm{E}(y^*) = \mathbf{k}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{y}$, $\mathrm{V}(y^*) = \kappa - \mathbf{k}^{\mathrm{T}}\mathbf{C}^{-1}\ \mathbf{k}$, indicating the expectation and variance of the predictive distribution of y^*, respectively. Usually, the most probable $\mathrm{E}(y^*)$ is used as the prediction result of GP regression. Different from other modeling methods, GP also provides the variance $\mathrm{V}(y^*)$ of the predictive distribution, indicating the distance from new peptide $\mathbf{x}^*$ to the training set samples. If the new input $\mathbf{x}^*$ is far away from the training set $\mathbf{D}$, the term $\mathbf{k}^{\mathrm{T}}\mathbf{C}^{-1}\ \mathbf{k}$ in predictive variance will be small, so that the predictive variance $\mathrm{V}(y^*)$ will be large, and then the predictive reliability of GP regression is low.

Now, GP regression problem is transformed to solving the covariance matrix $\mathbf{C}$. In terms of Mercer's theorem (Schlkopf et al. 1999), matrix element $\mathrm{Cov}(y^{(i)}, y^{(j)})$ in $\mathbf{C}$ can be calculated by the covariance function (or kernel) $K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$ in input space. The only constraint stated by Mackay is that the function must always generate a non-negative definite covariance matrix for any set of data points (MacKay 1998). Following is a common covariance function which is a linear combination of constant term, linear term, squared exponential term and noise term:

$$\begin{aligned} \mathrm{Cov}(y^{(i)}, y^{(j)}) &= K(\mathrm{x}^{(i)},\ \mathrm{x}^{(j)}) \\ &= \theta_0 + \theta_1 \sum_{m=1}^{M} x_m^{(i)} x_m^{(j)} \\ &\quad + \theta_2 \exp\left[-\frac{1}{2} \sum_{m=1}^{M} \left(\frac{x_m^{(i)} - x_m^{(j)}}{r_m} \right)^2 \right] + \sigma_v^2 \delta_{ij} \end{aligned} \quad (8)$$

where $x_m^{(i)}$ is the mth component of $\mathbf{x}^{(i)}$, M is the number of peptide descriptors (M dimensional input space), and θ_0, θ_1, θ_2, $\{r_m\}_{m=1}^M$, $\sigma_v^2 \in \boldsymbol{\Theta}$, are hyperparameters. θ_0, θ_1 and θ_2 are overall scales of constant, linear and squared exponential term, respectively; r_m are the length scales associated with each input and characterizes the distance in the mth direction over which y is expected to vary significantly. The last hyperparameter, σ_v^2 is noise variance which controls the tradeoff between smoothness and quality of fitting. Conventionally, hyperparameter set $\boldsymbol{\Theta}$ is determined by maximizing the (logarithm) marginal likelihood:

$$\ln P(\mathrm{y}|\mathrm{X},\ \Theta) = -\frac{1}{2}\mathrm{y}^{\mathrm{T}}\mathrm{C}^{-1}\mathrm{y} - \frac{1}{2}\ln|\mathrm{C}| - \frac{n}{2}\ln 2\pi \quad (9)$$

where $\mathrm{P}(\mathbf{y}|\mathbf{X},\ \boldsymbol{\Theta})$, the marginal likelihood, is also a Gaussian distribution. The three terms in Eq. 9 have readily interpretable roles: data-fit term, complexity penalty term and normalization constant term, respectively. Many approaches can be used to optimize the hyperparameter set $\boldsymbol{\Theta}$, including conjugate gradient method (Rasmussen 1996), Markov chain Monte Carlo sampling (Neal 1997), nested sampling (Skilling 2006), etc. In which the conjugate gradient method is the most common, well dealing with the tradeoff between computational accuracy and time. In this study, the conjugate gradient Polak-Ribiere method (Polyak 1969) was used to compute search

directions, and a line search using quadratic and cubic polynomial approximations and the Wolfe-Powell stopping criteria (Wolfe 1969) was used together with the slope ratio method for guessing initial step sizes. Setting of initialization $\mathbf{\Theta}$ was suggested by Obrezanova et al. (2007).

PLS, ANN and SVM for regression

PLS (Geladi and Kowalski 1986). PLS is a widely used modeling method to construct the linear relationship between peptide structures and activities. It has many advantages such as overcoming collinearity issue and is particularly suitable for the problems of which the sample size is smaller than variable numbers. The PLS principle is as follows: original descriptor matrix $\mathbf{X}$ is subject to a bi-linear decomposition, $\mathbf{X} = \mathbf{TP}^{\mathrm{T}} + \mathbf{F}$, where matrix $\mathbf{T}$ contains mutually orthogonal latent variable or score which is a linear combination of the variables in matrix $\mathbf{X}$. While PLS also implements bi-linear decomposition on activity $\mathbf{Y} = \mathbf{UQ}^{\mathrm{T}} + \mathbf{E}$, where $\mathbf{U}$ comprises latent variable of $\mathbf{Y}$ (if $\mathbf{Y}$ is a vector $\mathbf{y}$, then $\mathbf{U} = \mathbf{u} = \mathbf{y} = \mathbf{Y}$). The latent variable $\mathbf{T}$ is extracted by decomposing $\mathbf{X}$ in the consideration of maximally overlapping with latent variable $\mathbf{U}$ derived from $\mathbf{Y}$ decomposition. Therefore $\mathbf{U} = \mathbf{cT} + \mathbf{e}$, where $\mathbf{e}$ is error vector, and the coefficient $\mathbf{c}$ is determined by least square approach. In this study, cross-validation was used to assess the significance of each extracted PLS component, if correlation q^2 corresponding to a newly constructed latent variable is smaller than 0.097, then the latent variable is not significant and removed from the model.

Artificial neural network (Haykin 1999)

Feed-forward fully connected neural network was adopted by using the back-propagation algorithm for training (Rumelhart et al. 1986). The hidden layer and output layer are activated by Sigmoid and linear functions, respectively. Network is trained by the methods of gradient degression with momentum and self-adaptive learning velocity, and the neuron number of hidden layer is determined using minimum error with fixed training times (Heravi and Parastar 2000). In addition, a monitoring set was randomly selected from the training samples to reduce the possibility of overtraining risk.

Support vector machine (Cortes and Vapnik 1995)

SVM is a machine learning algorithm on the basis of statistical learning theory (SLT). In SVM, structural risk minimization (SRM) is instead of traditional empirical risk minimization (ERM), and it is particularly suitable for small-sample, high-dimensional and strong collinear problems. The central strategy of SVM regression is as this: under a given accuracy ε, a regression hyperplane $f(\mathbf{x}) = (\mathbf{w}\cdot\mathbf{x}) + \mathrm{b}$ is used to best fit sample points in the data space (i.e. data set D). Following that, slack variable $\xi \geq 0$, $\xi^* \geq 0$ and penalty parameter $\mathrm{C} > 0$ are further introduced, and by the Lagrange method, quadratic convex programming is transformed to the dual problem. Similar to GP, SVM employs kernel function $K(\mathbf{x}\cdot\mathbf{x}')$ to implement the inner product operation of high dimensional Hilbert space in the input space. Its decision function can thus be obtained as $f(\mathrm{x}) = \sum_{i=1}^{n} (\alpha_i^* - \alpha_i) y_i K(\mathrm{x}\cdot\mathrm{x}_i) + b^*$, in which only few Lagrange multiplier α_i, α_i^* are not zero, so-called the support vectors. In SVM, the parameters required to be optimized include insensitive ε, penalty parameter C and kernel parameter γ. In this study, radial basis function (RBF) was served as the SVM kernel, and parameters ε, C and γ were optimized by grid-searching.

Peptide characterization

Peptide characterization methods can be classified into global and local descriptors (Zhou et al. 2008a), the latter is also called as the amino acid descriptor. Doytchinova et al. (2005) demonstrated local descriptor was superior to global descriptor in statistical quality and interpretability. Therefore, peptide sequence characterization in this study was fulfilled by four groups of local descriptors, i.e., z scales, extended z scales, ISA-ECI and DPPS (Table 1). z scales, the famous principal properties of amino acids, were derived from 29 measured/calculated properties of 20 coded amino acids (Hellberg et al. 1987). PCA was applied to the standardized 20×29 property matrix and yielded three significant components z_1, z_2 and z_3 which represent the amino acid properties as hydrophilicity, size and polarity, respectively. Sandberg et al. (1998) further extended z scales to 87 amino acids (20 coded ones plus 67 non-coded ones). The number of components of the extended z scales was thus increased from the original 3 to 5. ISA-ECI was one of the most widely used local descriptors (Collantes and Dunn 1995). ECI was calculated as the sum of the absolute values of the charges for each atom presented in the amino acid side-chains, ISA was derived by summing the surfaces of the side-chain atoms accessible to nonspecific solvent interactions. DPPS, divided physicochemical property scores of amino acids, was a recently proposed descriptor in our laboratory (Tian et al. 2008). Four DPPS components as D_1, D_2, D_3 and D_4 were separately generated by the PCA processing of 23 electronic properties, 37 steric properties, 54 hydrophobic properties and 5 hydrogen bond properties, thus this amino acid descriptor possesses definite physicochemical meaning.

Table 1 ISA-ECI, z scales, extended z scales and DPPS descriptors for 20 coded amino acids

AAs	ISA-ECI		z scales			Extended z scales					DPPS			
	ISA	ECI	z_1	z_2	z_3	z_1	z_2	z_3	z_4	z_5	D_1	D_2	D_3	D_4
Ala, A	62.90	0.05	0.07	−1.73	0.09	0.24	−2.32	0.60	−0.14	1.30	−1.02	−6.15	0.04	−1.94
Arg, R	52.98	1.69	2.88	2.52	−3.44	3.52	2.50	−3.50	1.99	−0.17	1.99	4.78	−9.06	4.41
Asn, N	17.87	1.31	3.22	1.45	0.84	3.05	1.62	1.04	−1.15	1.61	−2.19	−2.30	−5.71	1.73
Asp, D	18.46	1.25	3.64	1.13	2.36	3.98	0.93	1.93	−2.46	0.75	−6.60	−3.25	−7.36	1.24
Cys, C	78.51	0.15	0.71	−0.97	4.13	0.84	−1.67	3.71	0.18	−2.65	0.21	−2.27	3.11	−1.70
Gln, Q	19.53	1.36	2.18	0.53	−1.14	1.75	0.50	−1.44	−1.34	0.66	−0.47	0.39	−5.46	1.93
Glu, E	30.19	1.31	3.08	0.39	−0.07	3.11	0.26	−0.11	−3.04	−0.25	−5.39	−0.23	−6.84	1.41
Gly, G	19.93	0.02	2.23	−5.36	0.30	2.05	−4.06	0.36	−0.82	−0.38	−2.86	−11.45	−2.11	−2.16
His, H	87.38	0.56	2.41	1.74	1.11	2.47	1.95	0.26	3.90	0.09	0.73	1.60	−1.94	0.44
Ile, I	149.77	0.09	−4.44	−1.68	−1.03	−3.89	−1.73	−1.71	−0.84	0.26	1.91	2.70	8.93	−1.10
Leu, L	154.35	0.10	−4.19	−1.03	−0.98	−4.28	−1.30	−1.49	−0.72	0.84	1.64	2.62	7.72	−1.03
Lys, K	102.78	0.53	2.84	1.41	−3.14	2.29	0.89	−2.49	1.49	0.31	2.47	2.77	−6.18	2.19
Met, M	132.22	0.34	−2.49	−0.27	−0.41	−2.85	−0.22	0.47	1.94	−0.98	1.93	2.79	5.33	−0.99
Phe, F	189.42	0.14	−4.92	1.30	0.45	−4.22	1.94	1.06	0.54	−0.62	2.68	5.02	8.60	−1.40
Pro, P	122.35	0.16	−1.22	0.88	2.23	−1.66	0.27	1.84	0.70	2.00	0.45	−3.79	0.70	−1.67
Ser, S	19.75	0.56	1.96	−1.63	0.57	2.39	−1.07	1.15	−1.39	0.67	−1.76	−5.72	−4.14	−0.13
Thr, T	59.44	0.65	0.92	−2.09	−1.40	0.75	−2.18	−1.12	−1.46	−0.40	−0.55	−2.76	−2.46	0.17
Trp, W	179.16	1.08	−4.75	3.65	0.85	−4.36	3.94	0.59	3.44	−1.59	3.88	9.31	7.53	−0.23
Tyr, Y	132.16	0.72	−1.39	2.32	0.01	−2.54	2.44	0.43	0.04	−1.47	2.10	5.90	3.71	0.25
Val, V	120.91	0.07	−2.69	−2.53	−1.29	−2.59	−2.64	−1.54	−0.85	−0.02	0.83	0.05	5.61	−1.44

Two ISA-ECI components have distinct orders of magnitude, so autoscaling was performed to eliminate the difference before modeling.

Model validation

Splitting data set into a training set and a test set. Usually, training set space is expected to sufficiently cover the test set samples, i.e., training and test set should have a desired similarity, while sufficient internal diversity should also be ensured for both the training set and the test set. Based upon these considerations, here a simple method for splitting peptide set is presented. Similarity between two peptides A and B with length L is expressed as:

$$S_{AB} = \sum_{l=1}^{L} \delta(P_l^A, P_l^B) \tag{10}$$

where P_l^A and P_l^B denote the lth residue position at peptide A and B, respectively, $\delta(P_l^A, P_l^B)$ is the discriminant function: in case $P_l^A = P_l^B$, then $\delta(P_l^A, P_l^B) = 1$, otherwise, $\delta(P_l^A, P_l^B) = 0$. Dissimilarity (or diversity) between peptide A and B is thus described as $L - S_{AB}$. Obviously, S_{AB} and $L - S_{AB} \in 0, 1, 2, \ldots, L$. For a splitting scheme G, a peptide set consisting of N samples is divided as training/test set as the proportion of n/m ($N = n + m$) with the following formula served as its scoring function:

$$SpScore(\mathrm{G}) = \frac{1}{19}\left[\frac{1}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(L - S_{ij}\right) + \frac{1}{m(m-1)} \times \sum_{i=1}^{m-1}\sum_{j=i+1}^{m}\left(L - S_{ij}\right)\right] + \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} S_{ij} \tag{11}$$

The two terms in the square bracket denote the average internal diversity of training set and test set respectively, and the last term indicates the average similarity between training and test set. For a random splitting, the ratio of diversity to similarity is expected as $\mathrm{E}\left(\frac{L-S}{S}\right) = 19$ which is served as background and deducted in Eq. 11 (i.e., the average internal diversity in the square bracket is divided by 19). When the number of training and test samples is given, we attempt to find an optimal splitting scheme $\mathrm{G}_{\mathrm{optimal}}$ of which the *SpScore* achieves maximum. This is a combinational optimization problem. In this study, splitting scheme G was optimized by Monte Carlo sampling, with *SpScore*(G) as the scoring function.

Internal validation

Training set was validated by cross-validation: leave-one-out and leave −1/3-out cross-validation were performed for PLS. In consideration of computational efficiency, only

leave-1/3-out cross-validation for SVM and GP were implemented. No cross-validation for ANN due to monitoring set was used to prevent overtraining.

External validation

For the test set, besides the traditional coefficient of determination r^2_{pred} and root mean square error of prediction (RMSP) (Gedeck et al. 2006), the statistics proposed by Golbraikh and Tropsha (2002) were used in this study. Corresponding criteria for a QSAM model to perform high predictive power are described as follows (Tropsha et al. 2003):

$$q^2_{\text{ext}} = 1 - \frac{\sum_{i=1}^{n_{\text{ext}}} \left(y_i^{\text{obsd}} - y_i^{\text{pred}}\right)^2}{\sum_{i=1}^{n_{\text{ext}}} \left(y_i^{\text{obsd}} - \bar{y}_{\text{tr}}\right)^2} \tag{12}$$

$$\frac{r^2_{\text{pred}} - r^2_{0,\text{ext}}}{r^2_{\text{pred}}} < 0.1 \text{ or } \frac{r^2_{\text{pred}} - r'^2_{0,\text{ext}}}{r^2_{\text{pred}}} < 0.1 \tag{13}$$

$$0.85 \le k \le 1.15 \text{ or } 0.85 \le k' \le 1.15 \tag{14}$$

where q^2_{ext} (external q^2) is external correlation coefficient indicating unbiased predictability on the test set, $r^2_{0,\text{ext}}$ and $r'^2_{0,\text{ext}}$are the coefficients of determination for the regression through origin (predicted vs. observed activities $r^2_{0,\text{ext}}$, and observed vs. predicted activities $r'^2_{0,\text{ext}}$), and k together with k' are the slopes of the origin-passed regression line.

Software used

SIMCA-P 10.0 (Umetrics AB, 2002) was used to perform PLS analysis. Matlab toolboxes of NNET, SVM (Gunn 1998) and GPML (Rasmussen and Williams 2006) were used for implementations of ANN, SVM and GP, respectively. In this study, we made some modifications for these Matlab programs, adding the functions as of ANN training monitoring, SVM grid-searching and GP cross-validation, and so on. Splitting dataset was carried out using in-home program SPEP written in C++.

Results and discussion

Angiotensin-I converting enzyme (ACE) inhibitory dipeptides

The set of 58 ACE inhibitory dipeptides synthesized by Cushman et al. (1980) is a benchmark of QSAM studies. By *SpScore* approach, this dataset was divided into a training set and a test set with the proportion of 40/18. While for the ANN, 8 samples were randomly selected from training set to enter into the monitoring set.

Table 2 lists the statistics of QSAM models constructed by different methods. The results obtained by linear PLS are relatively poor in both of fitting ability r^2 on the training set and predictive power r^2_{pred} on the test set. But, the z scales-based PLS model achieves a good predictive power on the test set ($r^2_{\text{pred}} = 0.802$), suggesting favorable linear relationship between the z scales and bioactivities of ACE inhibitory dipeptides, this was previously confirmed (Hellberg et al. 1991). The ANN model is indicated to be slightly overfitted; the fitting ability r^2 on the internal training set is all above 0.9, while the predictive power on external test set is below 0.8. However, the ANN model is acceptable due to a monitoring set is used here and its overfitting is not very significant. Statistical qualities of SVM and GP models are approximate to each other, and the both performed favorably. In comparison, GP is slightly superior to SVM. In the previous studies, it was demonstrated that there existed some linear relationship between the dipeptide structures and their bioactivities (Hellberg et al. 1991; Cocchi and Johansson 1993; Zaliani and Gancia 1999). In SVM only the nonlinear RBF kernel was employed, while in GP the both linear and nonlinear terms were included in its covariance function and thus better modeling sequence-activity relationship for this panel of peptides. In Table 2, statistics of most models met the Tropsha's criteria, indicating these models constructed by different methods were reliable. For the three groups of descriptors, z scales performed best, ISA-ECI and DPPS were secondary. In addition, for such small dataset the computational time of GP is less than that of ANN and SVM.

Figure 1 shows the linear relationship between the observed and calculated activities for the three GP models based upon z scales, ISA-ECI and DPPS, respectively. Noise deviation σ_v of the three optimized GP models are about 0.4, indicating this group of sample set includes about 1/10 of noise in the observed activity. In the three GP models, z scales perform well on both training and test sets in contrast with the other two descriptors. Besides the GP, other modeling methods, if constructed based on the z scales, also gave good results, suggesting for this dataset z scales sufficiently described the sequence-activity relationship. Table 3 lists optimal values of the z scales-based GP hyperparameters. By comparing the overall scales θ, the model was revealed to include a considerable nonlinear component (θ_2) and also some linear component (θ_0 and θ_1). Analyses of the length scales r corresponding to the six z scales (for a dipeptide, z_{11}, z_{12} and z_{13} correspond to three z scales of the first residue, and z_{21}, z_{22} and z_{23} are for the second residue). z_{12} and z_{22}, representing bulk properties of the two dipeptide residues, are the most

Table 2 Modeling statistics of the ACE inhibitory dipeptide panel by using PLS, ANN, SVM and GP approaches

Method	Descriptor	Training set (40 samples)			Test set (18 samples)						
		r^2	q^2	RMSE	r^2_{pred}	RMSP	Tropsha's statistics				
							q^2_{extd}	$r^2_{0,ext}$	$r'^2_{0,ext}$	k	k'
PLS	ISA-ECI[a]	0.610	0.532[e] 0.514[f]	0.613	0.635	0.642	0.642	0.585	0.617	0.966	0.998
	z scale[b]	0.738	0.709[e] 0.696[f]	0.503	0.802	0.472	0.806	0.789	0.744	0.980	1.000
	Extended z scale[c]	0.752	0.714[e] 0.701[f]	0.494	0.782	0.495	0.793	0.774	0.733	0.974	1.002
	DPPS[d]	0.704	0.626[e] 0.609[f]	0.541	0.583	0.685	0.592	0.521	0.558	0.955	1.003
ANN	ISA-ECI	0.934[g]	0.846[h]	0.243	0.728	0.520	0.734	0.714	0.681	0.961	1.012
	z scale	0.970[g]	0.863[h]	0.147	0.761	0.501	0.773	0.654	0.688	0.971	0.999
	Extended z scale	0.965[g]	0.849[h]	0.152	0.734	0.515	0.746	0.689	0.701	0.967	1.007
	DPPS	0.942[g]	0.891[h]	0.225	0.789	0.493	0.796	0.729	0.772	0.982	1.005
SVM	ISA-ECI	0.871	0.810[f]	0.396	0.802	0.453	0.811	0.784	0.763	0.972	1.011
	z scale	0.923	0.857[f]	0.264	0.847	0.419	0.855	0.830	0.792	0.958	1.023
	Extended z scale	0.926	0.844[f]	0.260	0.839	0.426	0.848	0.822	0.789	0.964	1.013
	DPPS	0.918	0.846[f]	0.291	0.835	0.441	0.847	0.794	0.821	0.965	1.018
GP	ISA-ECI	0.895	0.826[f]	0.310	0.841	0.423	0.844	0.815	0.837	0.975	1.010
	z scale	0.969	0.918[f]	0.167	0.848	0.414	0.851	0.802	0.825	0.956	1.031
	Extended z scale	0.954	0.916[f]	0.169	0.846	0.416	0.852	0.801	0.827	0.962	1.025
	DPPS	0.946	0.813[f]	0.222	0.668	0.611	0.675	0.651	0.632	0.944	1.015

[a] Number of significant latent variables is 1

[b] Number of significant latent variables is 1

[c] Number of significant latent variables is 2

[d] Number of significant latent variables is 2

[e] Leave-one-out cross-validation q^2

[f] Leave-1/3-out cross-validation q^2

[g] Coefficient of determination derived from 32 training samples

[h] Coefficient of determination derived from 40 (training + monitoring) samples

important to the model (having the smallest r values). It can be concluded that the ACE inhibitory activity is directly related with dipeptide size, which agrees with experimentally measured values, i.e., dipeptides with bulk residues usually have high activities, such as VW, IY, etc. In addition, hydrophilicity (indicated by z_{11} and z_{21}) also partially relates to the activity, while electronic properties (indicated by z_{13} and z_{23}) have relatively few effects on activity. In the z scale-based GP model, the predictive RMSP on test set is just 0.414, it can be considered as a good prediction if deducting the noise deviation σ_v of 0.385 from the experimentally measured activity.

Bradykinin-potentiating pentapeptides

The second peptide panel used in this study was 31 bradykinin-potentiating pentapeptides (BPPs) that were reported by Ufkes et al. (1978, 1982). The activities of first 15 BPPs were determined in 1978 and those of the last 16, including one inactive, were measured in 1982. The bioactivities were expressed as the logarithm of the relative activity index compared to the first peptide VESSK. The total 31 BPPs were divided into training/test set as of 25/6. For the ANN, five samples were randomly selected from the training set to enter into the monitoring set. To decrease complexity of the ANN, PCA was employed to reduce dimension of the input variables since few samples included in the training set.

As can be seen from Table 4 that the results obtained from three nonlinear modeling methods of ANN, SVM and GP were notably batter than that from the linear PLS, It is revealed that significant nonlinear relationship is existed between the structural characteristics and bioactivities of BPPs. For the PLS models, the modeling performance on

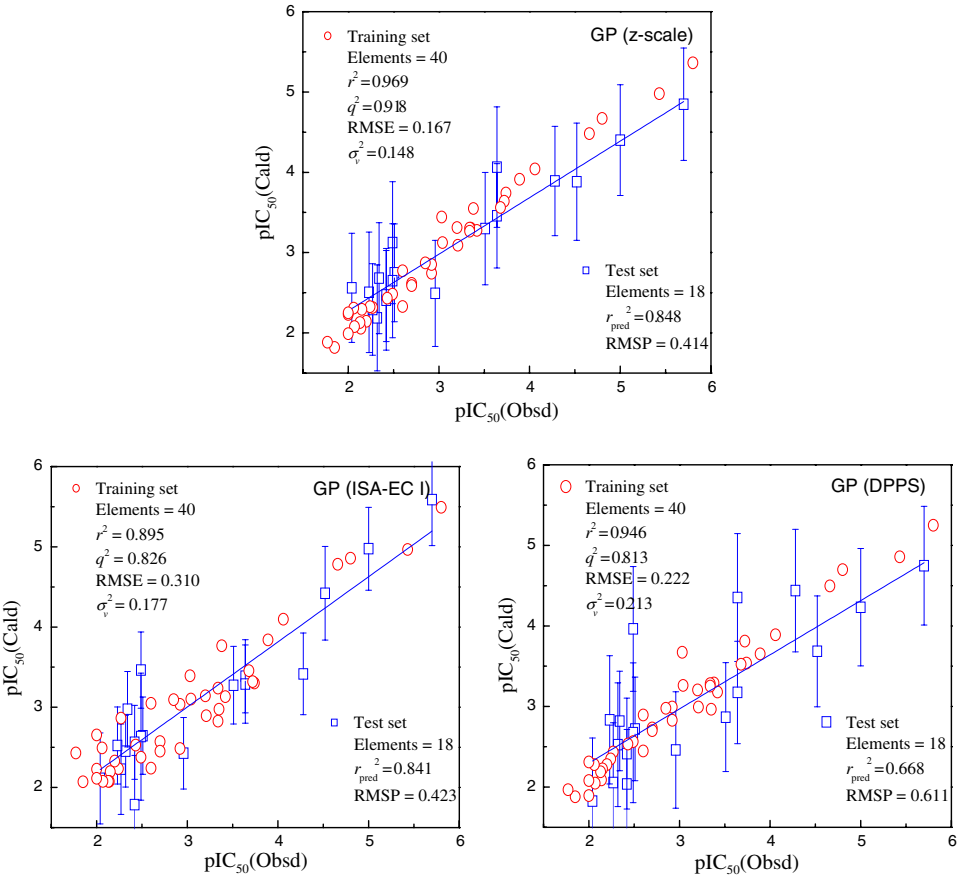

Fig. 1 Calculated versus observed pIC_{50} values for the ACE inhibitory dipeptide panel using GP approach. Training samples are denoted by *circles*, test samples are represented by *squares* and shown with *error bars* (Cald ± predictive deviation, including noise)

Table 3 Optimal values for the hyperparameters of z scale-based GP model

Hyperparameter	θ_0	θ_1	θ_2	σ_v^2	Position 1			Position 2		
					r_1 (z_{11})	r_2 (z_{12})	r_3 (z_{13})	r_4 (z_{21})	r_5 (z_{22})	r_6 (z_{23})
Optimal value	0.0227	0.0109	0.3108	0.1484	3.2900	0.5963	65.6279	1.6609	0.8462	14.6265

both training and test sets are poor, and the optimal ISA-ECI-based PLS model has a predictive r_{pred}^2 on the test set only of 0.518. For the three ANN models, although coefficient of determinations r^2 on the training set are all above 0.95, they are poor predictions on the test set, only slightly superior to the PLS models. It can be considered as the ANN models were insufficiently trained by the small-sample dataset. SVM performance approximately equals to the GP, they achieved a favorable result on both the training and test sets. The best prediction models were obtained by SVM-DPPS and GP-ISA-ECI, with the predictive r_{pred}^2 on the test set of 0.711 and 0.708, respectively. For the BPP set, most Tropsha's statistics of the three nonlinear methods were satisfied the criteria specified in Eqs. 12, 13, 14, but the slope k of SVM and GP models was slightly less than 0.85, it can be explained as the underestimation of the sample VKWAP which possesses the highest activity in the test set and only one peptide in the training set has higher activity than VKWAP, so the models may be insufficiently trained with respect to high-active samples. In addition, the performances of the three amino acid descriptors are approximately equivalent.

In Fig. 2, the three GP models are considered to be satisfactory for the BPP set. In which the z scales-based GP model involves slight systematic deviation for the test set ($k = 0.667$), ISA-ECI-based GP model is the best one ($r_{pred}^2 = 0.708$), while the DPP-based GP model possesses a smallest noise variance σ_v^2 of 0.067. Table 5 lists optimal values of hyperparameters for the ISA-ECI-based GP model. By analyses of the overall scales, nonlinear components are indicated to be far more than the linear ones in the covariance function (i.e., $\theta_2 \gg \theta_1$ and θ_0), this further confirmed that the BPP structures are nonlinearly related with their activities significantly. By comparing length

Table 4 Modeling statistics of the BPP panel by using PLS, ANN, SVM and GP approaches

Method	Descriptor	Training set (25 samples)			Test set (6 samples)						
		r^2	q^2	RMSE	r^2_{pred}	RMSP	Tropsha's statistics				
							q^2_{extd}	$r^2_{0,ext}$	$r'^2_{0,ext}$	k	k'
PLS	ISA-ECI[a]	0.767	0.626[e] 0.538[f]	0.425	0.518	0.764	0.521	0.487	0.452	0.658	1.103
	z scale[b]	0.664	0.591[e] 0.506[f]	0.499	0.469	0.814	0.473	0.398	0.447	0.697	1.094
	Extended z scale[c]	0.745	0.621[e] 0.546[f]	0.436	0.505	0.768	0.517	0.436	0.475	0.704	1.006
	DPPS[d]	0.690	0.606[e] 0.569[f]	0.479	0.474	0.783	0.480	0.424	0.454	0.745	1.002
ANN	ISA-ECI	0.954[g]	0.892[h]	0.162	0.524	0.865	0.533	0.508	0.465	0.896	1.026
	z scale	0.966[g]	0.887[h]	0.156	0.612	0.654	0.620	0.552	0.601	0.865	0.998
	Extended z scale	0.971[g]	0.894[h]	0.152	0.644	0.601	0.657	0.564	0.612	0.897	1.003
	DPPS	0.987[g]	0.901[h]	0.132	0.579	0.722	0.584	0.518	0.553	0.804	1.005
SVM	ISA-ECI	0.956	0.844[f]	0.168	0.689	0.489	0.691	0.632	0.669	0.766	1.006
	z scale	0.938	0.805[f]	0.207	0.597	0.813	0.604	0.588	0.549	0.715	1.019
	Extended z scale	0.929	0.810[f]	0.267	0.614	0.667	0.625	0.592	0.558	0.734	1.012
	DPPS	0.966	0.839[f]	0.156	0.711	0.467	0.723	0.652	0.697	0.812	1.013
GP	ISA-ECI	0.967	0.846[f]	0.150	0.708	0.494	0.715	0.695	0.651	0.732	1.080
	z scale	0.964	0.855[f]	0.157	0.693	0.520	0.712	0.681	0.646	0.667	1.088
	Extended z scale	0.965	0.851[f]	0.156	0.702	0.500	0.710	0.688	0.645	0.702	1.022
	DPPS	0.932	0.817[e]	0.215	0.701	0.501	0.707	0.685	0.643	0.797	0.978

[a] Number of significant latent variables is 1

[b] Number of significant latent variables is 2

[c] Number of significant latent variables is 3

[d] Number of significant latent variables is 2

[e] Leave-one-out cross-validation q^2

[f] Leave-1/3-out cross-validation q^2

[g] Coefficient of determination derived from 20 training samples

[h] Coefficient of determination derived from 25 (training + monitoring) samples

scales, electronic property of position 1 and steric property of positions 3 and 4 were revealed to exert a relatively important effect on peptide activities (r_2, r_5 and $r_7 < 1$), conversely, the steric property of position 2 nearly has no contributions to activity ($r_3 = 61.29$). The properties of other positions have some effects on the activity. In investigations of the noise variance σ_v^2 in the three GP models, it was indicated that the experimental error was relatively large, so the predictive accuracies were decreased by the noise distribution.

Cationic antimicrobial pentadecapeptides

The third peptide panel used here consists of 101 cationic antimicrobial pentadecapeptides (CAMPs) which were collected from the SAPD database (Wade and Englund 2002). Peptide antibacterial activity was expressed as the logarithm bactericidal potency which is the averages of potency values for 24 test bacteria such as *E. coli*, *Bacteroides fragilis*, *Staphylococcus aureus*, etc. We divided this data set into training/test set as of 70/31. For the ANN, 15 out of 70 samples in the training set were randomly selected as the monitoring set. Considering CAMP are all pentadecapeptides, high-dimensional variable space would be generated when using amino acid descriptors to characterize peptide sequences, so PCA was employed to reduce variable dimension prior to ANN training.

Table 6 lists the statistics of PLS, ANN, SVM and GP models. Nonlinear methods as ANN, SVM and GP are significantly superior to linear PLS, it is suggested antibacterial activity of CAMP is in strongly nonlinear relationship with structural characteristics. The statistical qualities of ANN, SVM and GP models are very close, and also note that the overtraining is not obvious in the ANN

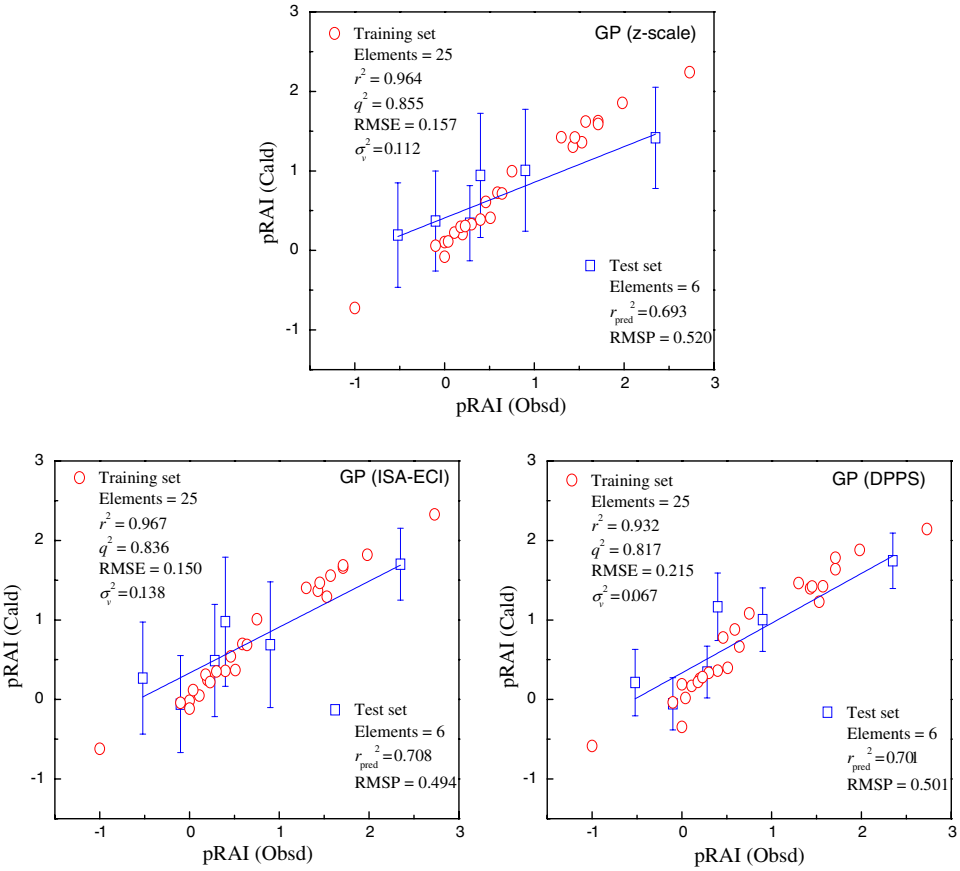

Fig. 2 Calculated versus observed pRAI values for the BPP panel using GP approach. Training samples are denoted by *circles*, test samples are represented by *squares* and shown with *error bars* (Cald ± predictive deviation, including noise)

Table 5 Optimal values of hyperparameters for the ISA-ECI-based GP model

Hyperparamter	θ_0	θ_1	θ_2	σ_v^2	Position 1		Position 2		Position 3		Position 4		Position 5	
					r_1 (ISA)	r_2 (ECI)	r_3 (ISA)	r_4 (ECI)	r_5 (ISA)	r_6 (ECI)	r_7 (ISA)	r_8 (ECI)	r_9 (ISA)	r_{10} (ECI)
Optimal value	0.0007	0.0545	0.2680	0.1378	1.1503	0.5211	61.2885	2.0574	0.2028	4.6783	0.0632	1.1343	5.1288	1.6293

model, this can be considered as that the CAMP set contains enough number of samples to sufficiently train the ANN model. For these nonlinear models, the fitting r^2 on training set are in the range of 0.85–0.95, and most of the predictive r^2_{pred} are above 0.65. The best predictive result was obtained by the DPPS-based GP model, with statistics r^2, q^2, RMSE, r^2_{pred}, q^2_{ext} and RMSP of 0.918, 0.823, 0.108, 0.752, 0.760 and 0.184, respectively. In further analyses of Tropsha's statistics listed in Table 6, the slope k of several models is less than 0.85, indicating slight systematic errors are existed in the predicted results of these models. By comparing the three kinds of amino acid descriptors, z scales and DPPS performed better than ISA-ECI for the CAMP set.

Figure 3 shows the calculation results for the CAMP set using different GP models. As can be seen, z scales- and DPPS-based GP models possessed a good performance, while the ISA-ECI-based GP model was relatively inferior, in this model the calculated values for many high-active peptides in the training set were nearly identical. By investigating these samples, we found that their sequences are very similar, with difference caused by only one or two distinct residues, and in addition, these different residues are almost conservative substitutions. So the ISA-ECI-based GP model was indicated to be in low resolution on CAMP structures, and thus incapable of describing the activity differences caused by small structural changes. Noise deviation σ_v of the best DPPS-based GP model is 0.167, which is about 1/5 of the average observed activities, indicating large error were presented in the antibacterial assay for the CAMP set. For the test set, the predictive RMSP by DPPS-based GP model is 0.184 (i.e., mean squared error of prediction MSEP was 0.034), note that is a good prediction due to the added noise with

Table 6 Modeling statistics of the CAMP panel by using PLS, ANN, SVM and GP approaches

Method	Descriptor	Training set (70 samples)			Test set (31 samples)						
		r^2	q^2	RMSE	r^2_{pred}	RMSP	Tropsha's statistics				
							q^2_{extd}	$r^2_{0,ext}$	$r'^2_{0,ext}$	k	k'
PLS	ISA-ECI[a]	0.627	0.412[e] 0.369[f]	0.301	0.218	0.413	0.239	0.186	0.156	0.684	1.180
	z scale[b]	0.712	0.495[e] 0.458[f]	0.234	0.429	0.338	0.450	0.362	0.396	0.701	1.122
	Extended z scale[c]	0.753	0.521[e] 0.510[f]	0.152	0.511	0.287	0.524	0.456	0.482	0.768	1.102
	DPPS[d]	0.733	0.484[e] 0.476[f]	0.201	0.412	0.345	0.431	0.385	0.354	0.721	1.119
ANN	ISA-ECI	0.896[g]	0.842[h]	0.124	0.648	0.233	0.654	0.620	0.589	0.755	1.109
	z scale	0.926[g]	0.871[h]	0.094	0.715	0.194	0.724	0.665	0.691	0.875	1.002
	Extended z scale	0.924[g]	0.864[h]	0.095	0.733	0.181	0.741	0.678	0.702	0.886	1.001
	DPPS	0.935[g]	0.877[h]	0.087	0.722	0.187	0.736	0.704	0.663	0.812	1.019
SVM	ISA-ECI	0.868	0.724[f]	0.138	0.645	0.236	0.649	0.594	0.622	0.734	1.112
	z scale	0.892	0.768[f]	0.127	0.732	0.179	0.742	0.719	0.681	0.798	1.095
	Extended z scale	0.891	0.774	0.128	0.742	0.173	0.753	0.722	0.701	0.802	1.036
	DPPS	0.889	0.754[f]	0.129	0.724	0.183	0.733	0.678	0.704	0.871	1.007
GP	ISA-ECI	0.865	0.712[f]	0.140	0.665	0.214	0.676	0.642	0.624	0.789	1.083
	z scale	0.881	0.774[f]	0.131	0.708	0.200	0.717	0.687	0.656	0.812	1.080
	Extended z scale	0.898	0.786[f]	0.126	0.714	0.197	0.723	0.695	0.664	0.834	1.061
	DPPS	0.918	0.823[f]	0.108	0.752	0.184	0.760	0.732	0.698	0.865	1.012

[a] Number of significant latent variables is 2

[b] Number of significant latent variables is 2

[c] Number of significant latent variables is 3

[d] Number of significant latent variables is 3

[e] Leave-one-out cross-validation q^2

[f] Leave-1/3-out cross-validation q^2

[g] Coefficient of determination derived from 55 training samples

[h] Coefficient of determination derived from 70 (training + monitoring) samples

variance 0.028. By analyses of the optimal hyperparameters (here is not given due to a much number) of DPPS-based GP model, a strongly nonlinear component was found ($\theta_2/\theta_1 \approx 64$). The electronic property (D_1) and hydrophobicity (D_3) are the most contributors to the peptide activity, second is hydrogen bond (D_4), and steric property (D_2) exert a insignificant effect on activity. This conclusion agrees well with CAMP antibacterial mechanism; since the sterilization function of CAMP is fulfilled by penetrating cell membrane, hydrophobicity and electronic property play the critical roles in inducing CAMP transmembrane (Jenssen et al. 2006). From Fig. 3, we can seen that the low-active peptides are overestimated by the GP models, it can be explained as the three amino acid descriptors are possibly unable to reproduce their structural information, due to the local descriptor is incapable of describing the interactive effect between peptide residues, particularly for the polypeptides.

Conclusions

Being a novel SMM, GP is preliminarily applied in QSAR field but has not yet been used to QSAM modeling. In current study, three peptide panels spanning 2–15 residues were modeled by GP approach, a systematic comparison between GP and PLS, ANN and SVM were also made. The results show that the GP modeling of peptides has the advantages as follows.

1. Since the covariance function consists of linear and nonlinear terms, the GP is able to model the linear and nonlinear-hybrid relationship between peptide

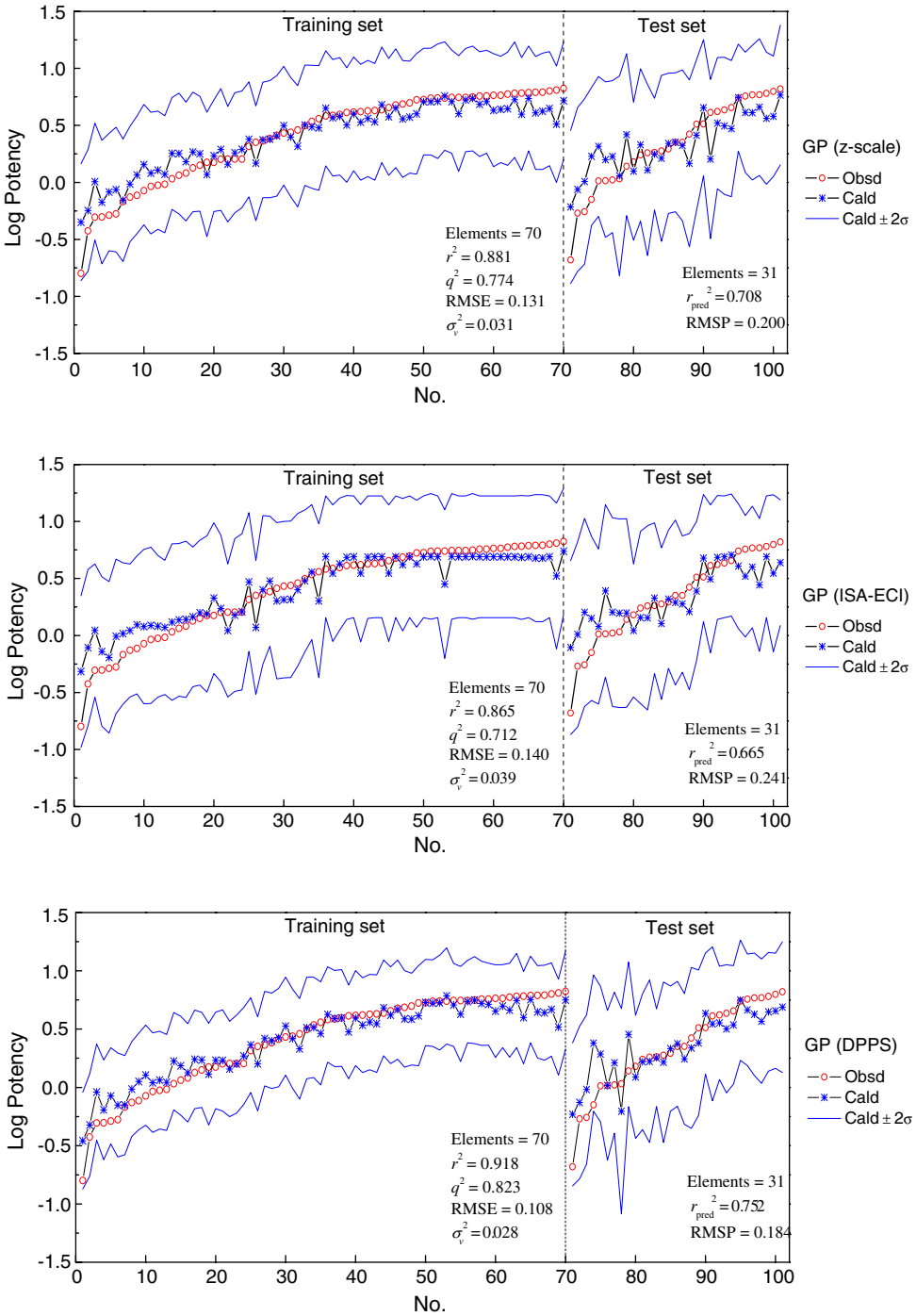

Fig. 3 GP modeling results for the CAMP panel using three kinds of amino acid descriptors. Samples in the training and test sets are numbered in term of their activities. The observed and calculated activities are represented by *circle* and *asterisk*, respectively. A 95% of confidence region (including noise) is also presented, i.e., Cald $\pm$ 2σ (σ is the standard deviation of calculation values)

structures and their activities. Furthermore, the ratio of linear to nonlinear components in the GP models can be automatically determined by adjusting the overall scales. By analyzing the constructed GP models, the sequence-activity relationship for oligopeptides (e.g., ACE inhibitory dipeptides) is shown to be a mix of linear and nonlinear, while for structurally complex polypeptides (e.g., BPPs and CAMPs) the sequence-activity relationship is obviously presented as a nonlinear form.

2. GP length scales are automatically determined by ARD algorithm and give a straightforward insight into the importance of amino acid descriptors, thus it can be used to evaluate the contribution of different properties and different positions in peptide sequence to the activity.

3. GP noise variance is served an indicator of the degree of experimental error involved in the data sets and can be used to estimate the predictive limit. For example, the CAMP bactericidal potency is the average of 24 antibacterial activities, so it contains a large uncertainty. This was well reflected by the resulted GP noise variance.

References

Armas RR, Gonzalez-Diaz H, Molina R, Uriarte E (2005) Stochastic-based descriptors studying biopolymers biological properties: extended MARCH-INSIDE methodology describing antibacterial activity of lactoferricin derivatives. Biopolymers 77:247–256. doi:10.1002/bip.20202

Ažman K, Kocijan J (2007) Application of Gaussian processes for black-box modeling of biosystems. ISA Trans 46:443–457. doi:10.1016/j.isatra.2007.04.001

Burden FR (2001) Quantitative structure-activity relationship studies using Gaussian processes. J Chem Inf Comput Sci 41:830–835. doi:10.1021/ci000459c

Chen T, Morris J, Martin E (2007) Gaussian process regression for multivariate spectroscopic calibration. Chemom Intell Lab Syst 87:59–71. doi:10.1016/j.chemolab.2006.09.004

Cho SJ, Zheng W, Tropsha A (1998) Rational design of targeted combinatorial peptide libraries using chemical similarity probe and the inverse QSAR approaches. J Chem Inf Comput Sci 38:259–268. doi:10.1021/ci9700945

Cocchi M, Johansson E (1993) Amino acids characterization by GRID and multivariate data analysis. Quant Struct Act Relat 12:1–8. doi:10.1002/qsar.19930120102

Collantes ER, Dunn WJ (1995) Amino acid side chain descriptors for quantitative structure-activity relationship studies of peptide analogues. J Med Chem 38:2705–2713. doi:10.1021/jm00014a022

Cortes C, Vapnik V (1995) Support vector networks. Mach Learn 20:273–293

Cushman DW, Ondetti MA, Cheung HS, Antonaccio MJ, Murthy VS, Rubin B (1980) Inhibitors of angiotensin converting enzymes. Adv Exp Med Biol 130:199–225

Dea-Ayuela MA, Perez-Castillo Y, Meneses-Marcel A, Ubeira FM, Bolas-Fernandez F, Chou KC, Gonzalez-Diaz H (2008) HP-Lattice QSAR for dynein proteins: experimental proteomics (2D-electrophoresis, mass spectrometry) and theoretic study of a *Leishmania infantum* sequence. Bioorg Med Chem 16:7770–7776. doi:10.1016/j.bmc.2008.07.023

Doytchinova IA, Walshe V, Borrow P, Flower DR (2005) Towards the chemometric dissection of peptide-HLA-A*0201 binding affinity: comparison of local and global QSAR models. J Comput Aided Mol Des 19:203–212. doi:10.1007/s10822-005-3993-x

Enot D, Gautier R, Le Marouille J (2001) Gaussian process: an efficient technique to solve quantitative structure-property relationship problems. SAR QSAR Environ Res 12:461–469. doi:10.1080/10629360108035385

Freyhult EK, Andersson K, Gustafsson MG (2003) Structural modeling extends QSAR analysis of antibody-lysozyme interactions to 3D-QSAR. Biophys J 84:2264–2272

Gedeck P, Rohde B, Bartels C (2006) QSAR—how good is it in practice? Comparison of descriptor sets on an unbiased cross section of corporate data sets. J Chem Inf Model 46:1924–1936. doi:10.1021/ci050413p

Geladi P, Kowalski B (1986) Partial least squares regression: a tutorial. Anal Chim Acta 185:1–17. doi:10.1016/0003-2670(86)80028-9

Genst ED, Areskoug D, Decanniere K, Muyldermans S, Andersson K (2002) Kinetic and affinity predictions of a protein-protein interaction using multivariate experimental design. J Biol Chem 277:29897–29907. doi:10.1074/jbc.M202359200

Golbraikh A, Tropsha A (2002) Beware of q2!. J Mol Graph Model 20:269–276. doi:10.1016/S1093-3263(01)00123-1

Gonzalez-Diaz H, Vilar S, Santana L, Uriarte E (2007) Medicinal chemistry and bioinformatics—current trends in drugs discovery with networks topological indices. Curr Top Med Chem 7:1015–1029. doi:10.2174/156802607780906771

Gonzalez-Diaz H, Gonzalez-Diaz Y, Santana L, Ubeira FM, Uriarte E (2008) Proteomics, networks and connectivity indices. Proteomics 8:750–778. doi:10.1002/pmic.200700638

Guan P, Doytchinova IA, Walshe VA, Borrow P, Flower DR (2005) Analysis of peptide-protein binding using amino acid descriptors: prediction and experimental verification for human histocompatibility complex HLA-A*0201. J Med Chem 48:7418–7425. doi:10.1021/jm0505258

Gunn S (1998) Support vector machines for classification and regression. Technical report. University of Southampton, Southampton

Haykin S (1999) Neural networks, a comprehensive foundation. Prentice Hall, Upper Saddle River, NJ

Hellberg S, Sjöström M, Wold S (1986) The prediction of bradykinin potentiating potency of pentapeptides. An example of a peptide quantitative structure-activity relationship. Acta Chem Scand B 40:135–140. doi:10.3891/acta.chem.scand.40b-0135

Hellberg S, Sjöström M, Skagerberg B, Wold S (1987) Peptide quantitative structure-activity relationships, a multivariate approach. J Med Chem 30:1126–1135. doi:10.1021/jm00390a003

Hellberg S, Eriksson L, Jonsson J, Lindgren F, Sjöström M, Skagerberg B, Wold S, Andrews P (1991) Minimum analogue peptide sets (MAPS) for quantitative structure-activity relationships. Int J Pept Protein Res 37:414–424

Heravi MJ, Parastar F (2000) Use of artificial neural networks in a QSAR study of anti-HIV activity for a large group of HEPT derivatives. J Chem Inf Comput Sci 40:147–154. doi:10.1021/ci990314+

Jenssen H, Gutteberg TJ, Lejon T (2005) Modeling of anti-HSV activity of lactoferricin analogues using amino acid descriptors. J Pept Sci 11:97–103. doi:10.1002/psc.604

Jenssen H, Hamill P, Hancock REW (2006) Peptide antimicrobial agents. Clin Microbiol Rev 19:491–511. doi:10.1128/CMR.00056-05

Jonsson J, Norberg T, Carlsson L, Gustafsson C, Wold S (1993) Quantitative sequence-activity models (QSAM) tools for sequence design. Nucleic Acids Res 21:733–739. doi:10.1093/nar/21.3.733

Kidera A, Konishi Y, Oka M (1985) Statistical analysis of the physical properties of the 20 naturally occurring amino acids. J Protein Chem 4:23–55. doi:10.1007/BF01025492

Kiryu H, Oshima T, Asai K (2005) Extracting relations between promoter sequences and their strengths from microarray data. Bioinformatics 21:1062–1068. doi:10.1093/bioinformatics/bti094

Ladiwala A, Xia F, Luo Q, Breneman CM, Cramer SM (2006) Investigation of protein retention and selectivity in HIC systems using quantitative structure retention relationship models. Biotechnol Bioeng 93:836–850. doi:10.1002/bit.20771

Lin Z, Wu Y, Zhu B, Ni B, Wang L (2004) Toward the quantitative prediction of T-cell epitopes: QSAR studies on peptides having affinity with the class I MHC molecular HLA-A*0201. J Comput Biol 11:683–694. doi:10.1089/cmb.2004.11.683

Liu W, Meng X, Xu Q, Flower DR, Li T (2006) Quantitative prediction of mouse class I MHC peptide binding affinity using support vector machine regression (SVR) models. BMC Bioinformatics 7:182. doi:10.1186/1471-2105-7-182

MacKay DJC (1998) Introduction to Gaussian processes. In: Bishop CM (ed) Neural networks and machine learning. Springer, Heidelberg

Neal RM (1997) Monte Carlo implementation of Gaussian process models for Bayesian regression and classification. Technical report, Department of Statistics, University of Toronto

O'Hagan A (1978) Curve fitting and optimal design for prediction. J R Stat Soc B 40:1–42

Obrezanova O, Csányi G, Gola JMR, Segall MD (2007) Gaussian processes: a method for automatic QSAR modeling of ADME properties. J Chem Inf Model 47:1847–1857. doi:10.1021/ci7000633

Patel S, Stott IP, Bhakoo M, Elliott P (1998) Patenting computer-designed peptides. J Comput Aided Mol Des 12:543–556. doi:10.1023/A:1008095802767

Polyak BT (1969) The conjugate gradient method in extreme problems. USSR Comput Math Math Phys 9:94–112. doi:10.1016/0041-5553(69)90035-4

Rasmussen CE (1996) Evaluation of Gaussian processes and other methods for non-linear regression. PhD thesis, University of Toronto, Canada

Rasmussen CE, Williams CKI (2006) Gaussian processes for machine learning. MIT Press, MA

Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back propagating errors. Nature 323:533–536. doi:10.1038/323533a0

Sandberg M, Eriksson L, Jonsson J, Sjöström M, Wold S (1998) New chemical descriptors for the design of biologically active peptides. A multivariate characterization of 87 amino acids. J Med Chem 41:2481–2491. doi:10.1021/jm9700575

Schlkopf B, Mika S, Burges C (1999) Input space versus feature space in kernel-based methods. IEEE Trans Neural Netw 10:1000–1017. doi:10.1109/72.788641

Schneider G, Schrödl W, Wallukat G, Müller J, Nissen E, Rönspeck W, Wrede P, Kunze R (1998) Peptide design by artificial neural networks and computer-based evolutionary search. Proc Natl Acad Sci USA 95:12179–12184. doi:10.1073/pnas.95.21.12179

Schroeter TS, Schwaighofer A, Mika S, Laak AT, Suelzle D, Ganzer U, Heinrich N, Müller K-R (2007) Predicting lipophilicity of drug-discovery molecules using Gaussian process models. Chem Med Chem 2:1265–1267. doi:10.1002/cmdc.200700041

Schwaighofer A, Schroeter T, Mika S, Laub J, Laak AT, Sulzle D, Ganzer U, Heinrich N, Muller KR (2007) Accurate solubility prediction with error bars for electrolytes: a machine learning approach. J Chem Inf Model 47:407–424. doi:10.1021/ci600205g

Skilling J (2006) Nested sampling for general Bayesiam computations. Bayesian Anal 1:833–860. doi:10.1214/06-BA127

Sneath PH (1966) Relations between chemical structure and biological activity in peptides. J Theor Biol 12:157–195. doi:10.1016/0022-5193(66)90112-3

Tian F, Zhou P, Li Z (2007a) T-scale as a novel vector of topological descriptors for amino acids and its application in QSARs of peptides. J Mol Struct 830:106–115. doi:10.1016/j.molstruc.2006.07.004

Tian F, Zhou P, Lv F, Song R, Li Z (2007b) Three-dimensional holograph vector of atomic interaction field (3D-HoVAIF): a novel rotation-translation invariant 3D structure descriptor and its applications to peptides. J Pept Sci 13:549–566. doi:10.1002/psc.892

Tian F, Li Y, Lv F, Yang Q, Zhou P (2008) In silico quantitative prediction of peptides binding affinity to human MHC molecule: an intuitive quantitative structure-activity relationship approach. Amino Acids (in press). doi:10.1007/s00726-008-0116-8

Tino P, Nabney IT, Williams BS, Losel J, Sun Y (2004) Nonlinear prediction of quantitative structure-activity relationships. J Chem Inf Comput Sci 44:1647–1653. doi:10.1021/ci034255i

Tropsha A, Gramatica P, Gombar VK (2003) The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. QSAR Comb Sci 22:69–77. doi:10.1002/qsar.200390007

Tung C-W, Ho S-Y (2007) POPI: predicting immunogenicity of MHC class I binding peptides by mining informative physicochemical properties. Bioinformatics 23:942–949. doi:10.1093/bioinformatics/btm061

Udaka K, Mamitsuka H, Nakaseko Y, Abe N (2002) Empirical evaluation of a dynamic experiment design method for prediction of MHC class I-binding peptides. J Immunol 169:5744–5753

Ufkes JGR, Visser RJ, Heuver G, van der Meer C (1978) Structure-activity relationships of bradykinin potentiating peptides. Eur J Pharmacol 50:119–122. doi:10.1016/0014-2999(78)90006-7

Ufkes JGR, Visser RJ, Heuver G, Wynne HJ, van der Meer C (1982) Further studies on the structure-activity relationships of bradykinin potentiating peptides. Eur J Pharmacol 79:155–158. doi:10.1016/0014-2999(82)90590-8

Wade D, Englund J (2002) Synthetic antibiotic peptides database. Protein Pept Lett 9:53–57. doi:10.2174/0929866023408986

Wilson SR, Cui W (2004) Applications of simulated annealing to peptides. Biopolymers 29:225–235. doi:10.1002/bip.360290127

Wold S, Ruhe A, Wold H, Dunn WJIII (1984) The collinearity problem in linear regression—the partial least squares (PLS) approach to generalized inverses. Siam J Sci Stat Comput 5:735–743. doi:10.1137/0905052

Wolfe P (1969) Convergence conditions for ascent methods. SIAM Rev 11:226–235. doi:10.1137/1011036

Wu J, Aluko RE, Nakai S (2006) Structural requirements of angiotensin I-converting enzyme inhibitory peptides: quantitative structure-activity relationship modeling of peptides containing 4–10 amino acid residues. QSAR Comb Sci 25:873–880. doi:10.1002/qsar.200630005

Zaliani A, Gancia E (1999) MS-WHIM scores for amino acids: a new 3D-description for peptide QSAR and QSPR studies. J Chem Inf Comput Sci 39:525–533. doi:10.1021/ci980211b

Zhou P, Li Z, Tian F, Zhang M (2006) QSAM-based computer-aided virtual vaccine library design. Acta Chim Sin 64:2065–2070

Zhou P, Tian F, Li Z (2007) A structure-based, quantitative structure-activity relationship approach for predicting HLA-A*0201-restricted cytotoxic T lymphocyte epitopes. Chem Biol Drug Des 69:56–67. doi:10.1111/j.1747-0285.2007.00472.x

Zhou P, Tian F, Wu Y, Li Z, Shang Z (2008a) Quantitative sequence–activity model (QSAM): applying QSAR strategy to model and predict bioactivity and function of peptides, proteins and nucleic acids. Curr Comput Aided Drug Des 4:311–321. doi:10.2174/157340908786785994

Zhou P, Tian F, Chen X, Shang Z (2008b) Modeling and prediction of binding affinities between the human amphiphysin SH3 domain and its peptide ligands using genetic algorithm-Gaussian processes. Biopolymers (Pept Sci) 90:792–802. doi:10.1002/bip.21091