Mol. Cells *27*, 547-556, May 31, 2009
DOI/10.1007/s10059-009-0082-3

Molecules and Cells

©2009 KSMCB

Prediction of Parathyroid Hormone Signalling Potency Using SVMs

Ahrim Yoo, Sunggeon Ko[1], Sung-Kil Lim[2], Weontae Lee[1,*], and Dae Ryook Yang*

Parathyroid hormone is the most important endocrine regulator of calcium concentration. Its N-terminal fragment (1-34) has sufficient activity for biological function. Recently, site-directed mutagenesis studies demonstrated that substitutions at several positions within shorter analogues (1-14) can enhance the bioactivity to greater than that of PTH (1-34). However, designing the optimal sequence combination is not simple due to complex combinatorial problems. In this study, support vector machines were introduced to predict the biological activity of modified PTH (1-14) analogues using mono-substituted experimental data and to analyze the key physicochemical properties at each position that correlated with bioactivity. This systematic approach can reduce the time and effort needed to obtain desirable molecules by bench experiments and provide useful information in the design of simpler activating molecules.

INTRODUCTION

Bioinformatics is a field of managing and interpreting information from biological sequences and structures. An enormous amount of data from gene and protein sequence analysis, as well as protein structure prediction, is published in this field. All of these analyses are designed to better understand biological systems. Because protein is one essential component of biological systems, a crucial first step in understanding such a system is to know the functions of a protein. In this paper, we sought to predict the function of a target peptide from published experimental data.

Parathyroid hormone (PTH), the target peptide, is the most important endocrine regulator of calcium concentration in extracellular fluid (Chorev, 2002). Injected intermittently, this hormone stimulates a net increase in bone mass. Therefore, PTH has considerable therapeutic potential for the treatment of osteoporosis. PTH consists of 84 amino acids, but the N-terminal fragment, PTH (1-34), is sufficient for receptor binding and activation (Tsomaia et al., 2004). In fact, PTH in the form of PTH (1-34) was approved by the FDA in 2002 for treating osteoporosis (Potts, 2005). Experiments with truncated PTH fragments has provided useful information in assigning the roles of the N- and C-terminal regions (Shimizu et al., 2000; 2001; Tsomaia et al., 2004). The results showed that PTH analogues with N-terminal truncations, such as PTH (3-34) and PTH (7-34), efficiently bound the PTH receptor, but lost their cAMP-stimulating potency. Therefore, the N-terminal residues are involved in receptor activation while the C-terminal residues participate in receptor binding.

Shimizu et al. found that the shorter fragment PTH (1-14) exhibited cyclic AMP (cAMP)-stimulating potency albeit weaker than that of PTH (1-34) (Luke et al., 1999; Shimizu et al., 2000). Moreover, their results indicated that the potency of PTH (1-14) analogues depends on their sequence combination. Therefore, site-directed mutagenesis at several positions of PTH (1-14) can enhance agonist potency to a level greater than that of PTH (1-34) and also reduce the size of the agonist peptide. The first step in designing a fragment with an optimal sequence combination involves determining which residues to substitute with different amino acids. Subsequently, the designed fragment is synthesized, and cAMP expression level induced by the peptide is measured. However, the first step is not simple due to complex combinatorial problems. In addition, these experiments require considerable time and effort. Therefore, a systematic approach is needed to predict the signaling potency of mutated fragments effectively.

In this study, we applied support vector machines (SVMs) to the single-substitution experimental data. SVMs, introduced by Vapnik (1995), have been used broadly to solve classification and regression problems. The basic idea behind applying SVMs to classification problems is to map training data into a higher-dimensional feature space via functions, $\phi(x_i)$, and to find a separating hyperplane with maximum margin in this higher dimensional space. SVMs have been successfully applied in a wide range of fields, such as text categorization, analysis of microarray gene expression data, and identification of critical positions in a protein (Dubey et al., 2004; Sarda et al., 2005). Bhasin and Raghava used the regression method to predict the affinity of Transporter-associated with antigen processing subunit 1 and 2 (TAP) binding peptides with the training feature set extracted from the sequence and 33 physicochemical properties of amino acids (Bhasin and Raghava, 2004). Using SVMs, a good correlation was achieved between the experimentally determined and predicted binding affinities of TAP peptides. Supper also used SVMs to classify major histocompatibility

Department of Chemical and Biological Engineering, Korea University, Seoul 136-713, Korea, [1]Department of Biochemistry, Yonsei University, Seoul 120-749, Korea, [2]Department of Internal Medicine, School of Medicine, Yonsei University, Seoul 120-749, Korea
*Correspondence: wlee@spin.yonsei.ac.kr (WL); drying@korea.ac.kr (DRY)

Received November 26, 2008; revised April 11, 2009; accepted April 14, 2009; published online May 15, 2009

Keywords: function, parathyroid hormone, peptide analogue, SVM

complex (MHC) class I binding peptides (Supper, 2005) by building the training set from sequence and physicochemical properties. These studies demonstrated that SVMs are a useful method for the prediction of peptide signaling potency.

The objective of this study was to build a model to classify cAMP expression levels and to identify the physicochemical properties closely related to the stimulating potency. First, we built a model using SVMs. The training set consisted of a target class and input feature sets. The target class was divided into eight levels according to the degree of cAMP expression, and the input sets were built based on features extracted from sequence and physicochemical properties. The physicochemical properties of amino acids were derived from the Amino Acid Index Database (AAindex) (Kawashima et al., 1999). This flat-file database provides 434 amino acid indices in the form of a real value for each amino acid with respect to a certain property. For the feature sets, 295 properties were selected from the AAindex as a result of correlation analysis. After obtaining the model, we reduced the number of properties to remove redundancy and make it easier to analyze the correlation between stimulating potency and physicochemical properties. Last, we investigated which properties at each position influenced the signaling potency of PTH (1-14) analogues using the model and test sets.

MATERIALS AND METHODS

Support vector machines

In this study, we applied a method using support vector machines (SVMs) to predict the expression level of cAMP by PTH (1-14) analogues. The SVMs produce a model that can predict the class of input set in a high dimensional feature space, as shown in Fig. 1A (Scholkopf and Smola, 2002; Vapnik, 1995). Each training data set consists of two major elements: a set of feature values (input) and a target value (output).

As shown in Fig. 1A, the input data sets (x_i) are mapped into the feature space by the mapping function $\phi(x_i)$ and then separated by the hyperplane expressed as follows:

$$F(x_i) = w\phi(x_i) + b \tag{1}$$

where w is the weighting vector and b is the bias.

The geometric margin of the hyperplane (w, b) is represented as $M = 1/\|w\|$ and the following equation (2) is the objective function to find the maximum-margin hyperplane:

$$\begin{aligned} &\min_{w,b,\xi} \frac{1}{2} w^T w + C\sum_{i=1}^{M} \xi_i \\ &s.t.\ y_i (w^T \phi(x_i) + b) \geq 1-\xi_i \qquad \forall\, i = 1,\ldots, M \\ &\quad \xi_i \geq 0 \qquad\qquad\qquad\qquad \forall\, i = 1,\ldots, M \end{aligned} \tag{2}$$

In this objective function, the first part maximizes the margin, while the second part relaxes the maximum-margin constraints. The slack variable (ξ_i) is an upper bound on the training classification error. The parameter (C) controls the relative weighting between maximizing the width of the margin and minimizing the error of misclassification (Fig. 1B).

This problem was solved by transforming the equation into the equivalent Lagrangian problem (Cristianini and Shawe-Taylor, 2000). The primal Lagrangian for this problem is defined as:

$$L(w,\xi,b,\alpha) = \frac{1}{2} w^T w + C\sum_{i=1}^{M} \xi_i - \sum_{i=1}^{M} \alpha_i [y_i (w^T \phi(x_i) + b) - 1 + \xi_i] \tag{3}$$

where α is the Lagrange multiplier.

The Lagrangian dual form (4) of primal problem simplifies the optimization problem.

$$\begin{aligned} &\max_{\alpha} \sum_{i=1}^{M} \alpha_i - \frac{1}{2} \sum_{i=1}^{M} \sum_{j=1}^{M} y_i y_j \alpha_i \alpha_j K(x_i, x_j) \\ &s.\ t.\ \sum_{i=1}^{M} y_i \alpha_i = 0 \\ &C \geq \alpha_i \geq 0 \end{aligned} \tag{4}$$

In this case, the decision function is written as

$$f(x) = \sum \alpha_i y_i K(x_i, x_j) + b \tag{5}$$

where $K(x_i, x_j)$ is the kernel defined as the inner product of two vectors in the feature space:

$$K(x_i, x_j) = \phi(x_i)^T \phi(x_j) \tag{6}$$

Most kernels used in this type of problem are polynomial, Gaussian, or radial basis (RBF). We selected the RBF kernel for this study.

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2), \quad \gamma > 0 \tag{7}$$

The use of kernels makes it possible to implicitly map the training data into a feature space. In particular, the decision function can be constructed by the kernel $K(x_i, x_j)$ as in Equation 5 without an explicit mapping function. This kernel reduces the computation time inherent in evaluating the feature map.

In binary classification, the class of input data is determined by the sign of the decision function. However, cAMP expression levels were divided into eight categories in this study; hence, the binary problem had to be extended into a multiple classification problem. Multiple classification was achieved by using the SVM^{multiclass} package (Joachims, 1999; Toschantaridis et al., 2004). This package enables the user to select the built-in kernel types and tune the parameters. The most common approach for solving multiple classification problems using SVMs is based on reducing a single multiclass problem into multiple binary problems. The optimization algorithm of Crammer and Singer (2001) was implemented in the SVM^{multiclass} package. When k classes were given, they suggested another approach to solve a single optimization problem with the following objective function (8):

$$\begin{aligned} &\min_{w_m,\xi_i} \frac{1}{2} \sum_{i=1}^{k} w_m^T w_m + C\sum_{i=1}^{l} \xi_i w_{y_i}^T \phi(x_i) \\ &s.\ t. \quad -w_m^T \phi(x_i) \geq e_i^m - \xi_i \qquad \forall i = 1,\ldots,l \\ &\text{where} \quad e_i^m \equiv 1 - \delta_{y_i,m} \text{ and } \delta_{y_i,m} \equiv \begin{cases} 1 & \text{if } y_i = m \\ 0 & \text{if } y_i \neq m \end{cases} \end{aligned} \tag{8}$$

Then the decision function is defined as:

$$\text{Arg} \quad \max_{m=1,\ldots,k} w_m^T \phi(x) \tag{9}$$

This function uses only l slack variables $\xi_i (i = 1,\ldots,l)$. In other methods, the number of classifiers is $k(k\text{-}1)/2$ because the gap must be defined between each pair of decision planes. In the single optimization problem suggested by Crammer and Singer (2001), the number of classifiers was less than k because slack variables were expressed as (Crammer and Singer, 2001):

$$\xi_i = \max\{(\max(w_m^T \phi(x_i) + e_i^m) - w_{y_i}^T \phi(x_i)), 0\} \tag{10}$$

This makes the multi-classification problem simpler. In this

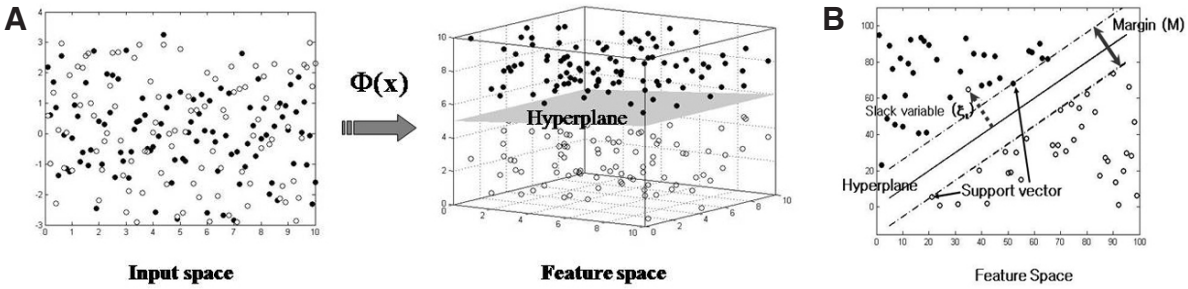

Fig. 1. Principle of support vector machines. (A) Transformation of data from input space into the feature space. (B) Principle of binary data classification by maximizing the margin.

study, the signaling potency of PTH (1-14) was predicted using the SVM^{multiclass} package in which the optimization algorithm of Crammer and Singer was implemented.

Data encoding

As the first step in building the classification model, we grouped the target value into eight classes according to the level of cAMP expression. The signaling potency data used in this study for PTH analogues was gathered from published papers (Luke et al., 1999; Shimizu et al., 2000). Shimizu et al. (2000) tested the control peptide, native PTH (1-14), and its analogues for their ability to stimulate cAMP formation at a single dose of 100 μM. We identified 132 single-substituted analogues to analyze which properties at each position affect the activity. The results demonstrated that the cAMP-stimula-ting potency depended upon the amino acid composition. The experimental data of 125 mono-substituted PTH analogues, excluding mutants with D-amino acids, were used to train the SVMs. For the single-substituted analogues, the level of expression ranged from 0-200 pmol/well. We divided this range into eight equally spaced levels and defined the target value as an integer between one and eight. The control peptide, PTH (1-14), belonged to the fifth class. When the C-terminus of native PTH (1-14) was truncated, its potency became weak. If an analogue was categorized in a class lower than third, it did not display any signaling potency. If an analogue was placed higher than the seventh class, it possessed a strong signaling potency.

The next step involved determining the feature vector. As shown in Fig. 2, three kinds of input sets were used: set_1 = [binary encoded sequence], set_2 = [binary encoded sequence + physicochemical properties], and set_3 = [physicochemical properties]. The first set was encoded based only on amino acid composition since cAMP expression levels depend on the sequence of amino acids. Although various PTH (1-14) mutants were tested by Shimizu et al. (2003), the residues at each position were not substituted with each different amino acid. Therefore, physicochemical properties were used to supplement the insufficiency in the experimental data since the twenty amino acids can be categorized into groups based on their properties. Therefore, the second set was based on information extracted from sequence and physicochemical properties. The third set was constructed to examine if it is possible to classify test sets with only physicochemical properties. The amino acid of each position was encoded with a 20-dimensional vector containing 19 zeros, and the physicochemical properties were represented as normalized real values. Therefore, the input vector has (20 + *n* properties) × 14 dimensions in set_2.

The AAindex database was used for the set of physicochemical properties (Kawashima et al., 1999). The AAindex1 is a flat-file database that provides various physicochemical properties of amino acids. To train the SVMs, 295 different properties were extracted from 484 indices with p-values of less than 0.05 in the correlation analysis. Pearson's correlation coefficient between the cAMP expression level and the physicochemical properties at each position of the analogs was tested using a Student's *t*-test (Martinez and Martinez, 2002). A low p-value for the test (e.g., less than 0.05) indicated a statistically significant relationship between the two variables, namely the physicochemical properties and the signaling potency of the mutants.

Analysis of physicochemical properties

The models developed by the SVMs had two uses. One was to design PTH (1-14) analogs with maximum potency, since the model can predict the stimulus potency level of arbitrary mutants. The other was to identify the physical properties closely related to the potency. The only criterion for whether certain physicochemical properties were related to the signaling potency was the correlation analysis before the model was constructed. At first, 295 properties were extracted using the correlation analysis function of the MATLAB 7.2 statistics toolbox. However, after the optimal hyperplane was determined, it was possible to reduce the size of the property set by reversing the SVM training process. If removal of a certain property in the training set led to deterioration of the classification performance, then that property was essential for the potency of PTH (1-14). We could therefore find these critical properties by repeating this procedure.

RESULTS AND DISCUSSION

Prediction with SVMs

A model to predict the activation level of the PTH (1-14) analogues was generated using the SVM^{multiclass} package. This package enables the user to select a kernel type and define parameters (Joachims, 1999). In this study, we selected the RBF kernel and the user-defined parameters, γ and C, ranging from 0.01-1.2 and 10-1,000, respectively. For the SVM training, we divided the target class into eight levels based on the degree of cAMP expression, and three feature sets were prepared as shown in Fig. 2, namely set_1, based on binary encoding of the sequence, set_2, based on the binary encoding of sequence plus physicochemical properties, and set_3, based on the physicochemical properties.

Validation with a test set was conducted to develop the model and evaluate performance. The test set consisted of 20 mutants (Table 2) that were not included in the training set but were derived from experimental data by Shimizu et al. (2000; 2001). Shimizu et al. (2000) sought to enhance the activity of the analogues by modifying two or more residues based on single mutation data. Their results showed that the effects of

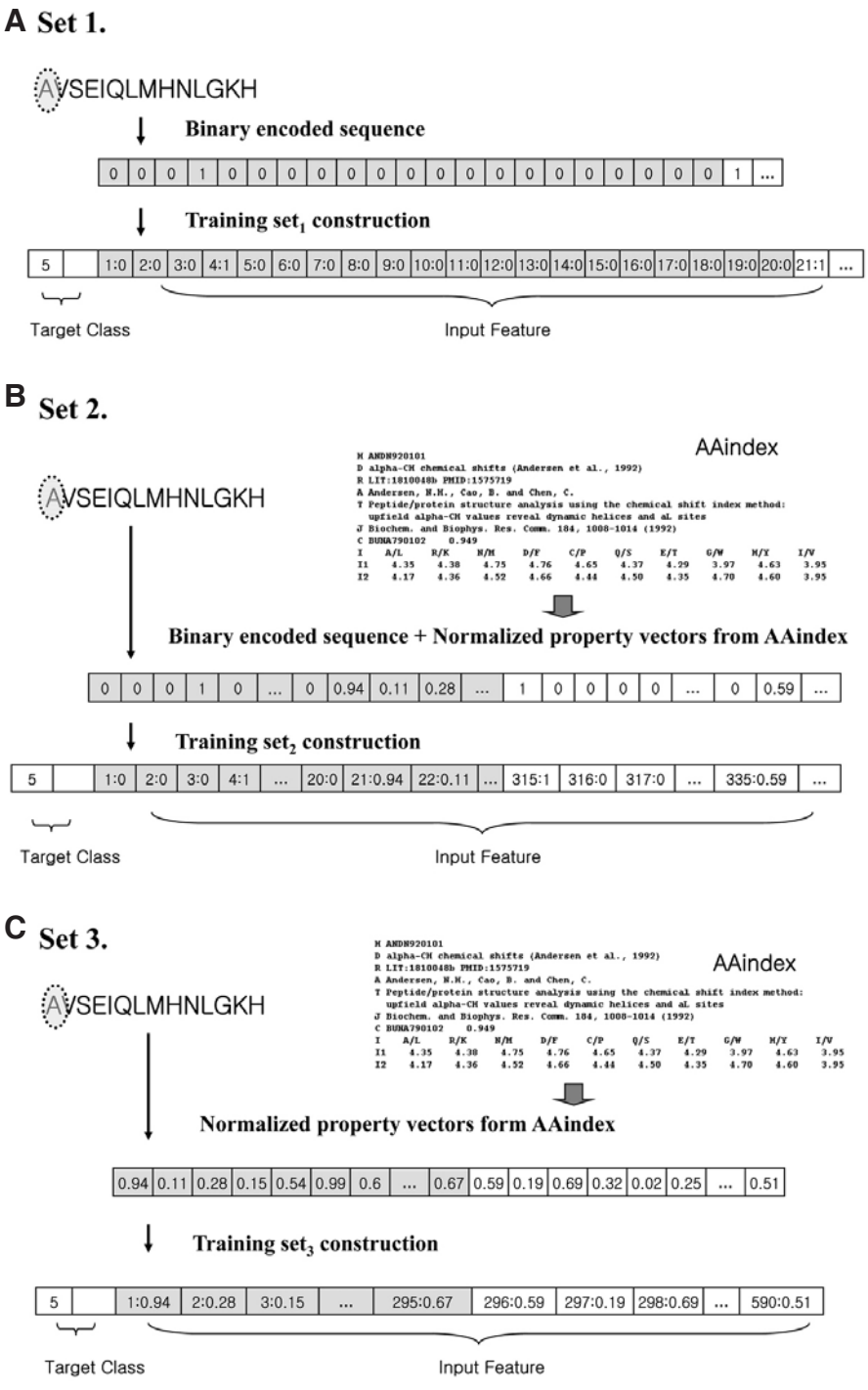

Fig. 2. Feature vector construction. (A) Set_1, based on binary encoding of sequence, (B) Set_2, based on the binary encoding of sequence plus physicochemical properties, and (C) Set_3, based on the physicochemical properties.

Table 1. Classification results of SVMs and the optimal parameters (γ/C) of the RBF kernel. The total of 20 analogues substituted at two or more positions were used as the test set. Three kinds of feature vector were used: Set_1, based on binary encoding of sequence, Set_2, based on the binary encoding of sequence plus physicochemical properties, and Set_3, based on the physicochemical properties.

	Set_1	Set_2	Set_3
Accuracy (true/total)	65%	90%	90%
SVMs Parameter (γ/C)	0.5/700	0.9/1,000	0.9/1,000

these substitutions on the potency of stimulation were additive. The aim of this paper was to predict the signaling potency of the PTH (1-14) analogues using the mono-substitution experimental data published by Shimizu et al. (2000). The classification results and the optimal parameters of the RBF kernel are summarized in Table 1.

The data in Table 1 reveal that SVMs can classify the cAMP expression level. The prediction based on sequence plus physicochemical properties was better than that based on sequence alone (90% vs. 65%). The physicochemical properties in set_2

Table 2. Training and test set to evaluate the performance of the SVM model. The test set consisted of 20 mutants (126-145) that were not included in the training set; most were from the experimental data of Shimizu et al.

No.	Sequence	Signaling activity (cAMP picomol/well)	Status
1	AVSEIQLMHNLGKH	105	Training
2	SVSEIQLMHNLGKH	90	Training
3	VVSEIQLMHNLGKH	90	Training
4	LVSEIQLMHNLGKH	20	Training
5	NVSEIQLMHNLGKH	5	Training
6	EVSEIQLMHNLGKH	5	Training
7	KVSEIQLMHNLGKH	5	Training
8	FVSEIQLMHNLGKH	10	Training
9	WVSEIQLMHNLGKH	10	Training
10	PVSEIQLMHNLGKH	80	Training
11	AASEIQLMHNLGKH	5	Training
12	ALSEIQLMHNLGKH	5	Training
13	AISEIQLMHNLGKH	35	Training
14	AMSEIQLMHNLGKH	5	Training
15	AESEIQLMHNLGKH	5	Training
16	AKSEIQLMHNLGKH	5	Training
17	AFSEIQLMHNLGKH	5	Training
18	AWSEIQLMHNLGKH	5	Training
19	APSEIQLMHNLGKH	5	Training
20	AVGEIQLMHNLGKH	75	Training
21	AVAEIQLMHNLGKH	150	Training
22	AVTEIQLMHNLGKH	30	Training
23	AVTEIQLMHNLGKH	10	Training
24	AVEEIQLMHNLGKH	5	Training
25	AVQEIQLMHNLGKH	10	Training
26	AVKEIQLMHNLGKH	5	Training
27	AVFEIQLMHNLGKH	5	Training
28	AVWEIQLMHNLGKH	5	Training
29	AVPEIQLMHNLGKH	5	Training
30	AVSAIQLMHNLGKH	5	Training
31	AVSIIQLMHNLGKH	5	Training
32	AVSDIQLMHNLGKH	5	Training
33	AVSQIQLMHNLGKH	10	Training
34	AVSKIQLMHNLGKH	10	Training
35	AVSHIQLMHNLGKH	10	Training
36	AVSFIQLMHNLGKH	5	Training
37	AVSWIQLMHNLGKH	5	Training
38	AVSPIQLMHNLGKH	5	Training
39	AVSEAQLMHNLGKH	5	Training
40	AVSEVQLMHNLGKH	10	Training
41	AVSELQLMHNLGKH	7	Training
42	AVSEMQLMHNLGKH	5	Training
43	AVSEEQLMHNLGKH	4	Training
44	AVSEKQLMHNLGKH	4	Training
45	AVSEFQLMHNLGKH	5	Training
46	AVSEWQLMHNLGKH	5	Training
47	AVSEPQLMHNLGKH	5	Training
48	AVSEIALMHNLGKH	10	Training
49	AVSEIELMHNLGKH	11	Training
50	AVSEIKLMHNLGKH	6	Training
51	AVSEIFLMHNLGKH	6	Training
52	AVSEIWLMHNLGKH	5	Training
53	AVSEIPLMHNLGKH	5	Training
54	AVSEIQAMHNLGKH	5	Training
55	AVSEIQEMHNLGKH	6	Training
56	AVSEIQKMHNLGKH	5	Training
57	AVSEIQFMHNLGKH	120	Training
58	AVSEIQWMHNLGKH	25	Training
59	AVSEIQPMHNLGKH	5	Training
60	AVSEIQLAHNLGKH	5	Training
61	AVSEIQLEHNLGKH	5	Training
62	AVSEIQLFHNLGKH	6	Training
63	AVSEIQLWHNLGKH	5	Training
64	AVSEIQLPHNLGKH	5	Training
65	AVSEIQLMANLGKH	7	Training
66	AVSEIQLMKNLGKH	7	Training
67	AVSEIQLMRNLGKH	5	Training
68	AVSEIQLMFNLGKH	5	Training
69	AVSEIQLMWNLGKH	5	Training
70	AVSEIQLMPNLGKH	6	Training
71	AVSEIQLMHGLGKH	24	Training
72	AVSEIQLMHALGKH	160	Training
73	AVSEIQLMHSLGKH	70	Training
74	AVSEIQLMHVLGKH	11	Training
75	AVSEIQLMHLLGKH	22	Training
76	AVSEIQLMHDLGKH	150	Training
77	AVSEIQLMHELGKH	135	Training
78	AVSEIQLMHQLGKH	185	Training
79	AVSEIQLMHRLGKH	20	Training
80	AVSEIQLMHHLGKH	40	Training
81	AVSEIQLMHFLGKH	21	Training
82	AVSEIQLMHWLGKH	38	Training
83	AVSEIQLMHPLGKH	5	Training
84	AVSEIQLMHNAGKH	125	Training
85	AVSEIQLMHNSGKH	8	Training
86	AVSEIQLMHNVGKH	100	Training
87	AVSEIQLMHNIGKH	140	Training
88	AVSEIQLMHNMGKH	150	Training
89	AVSEIQLMHNEGKH	5	Training
90	AVSEIQLMHNQGKH	45	Training
91	AVSEIQLMHNKGKH	170	Training
92	AVSEIQLMHNRGKH	200	Training
93	AVSEIQLMHNHGKH	20	Training
94	AVSEIQLMHNFGKH	70	Training
95	AVSEIQLMHNWGKH	130	Training
96	AVSEIQLMHNPGKH	5	Training
97	AVSEIQLMHNLAKH	160	Training

No.	Sequence	Signaling activity (cAMP picomol/well)	Status
98	AVSEIQLMHNLLKH	5	Training
99	AVSEIQLMHNLNKH	30	Training
100	AVSEIQLMHNLEKH	5	Training
101	AVSEIQLMHNLKKH	40	Training
102	AVSEIQLMHNLRKH	110	Training
103	AVSEIQLMHNLHKH	110	Training
104	AVSEIQLMHNLFKH	25	Training
105	AVSEIQLMHNLWKH	60	Training
106	AVSEIQLMHNLPKH	10	Training
107	AVSEIQLMHNLGAH	140	Training
108	AVSEIQLMHNLGLH	110	Training
109	AVSEIQLMHNLGEH	70	Training
110	AVSEIQLMHNLGQH	95	Training
111	AVSEIQLMHNLGRH	150	Training
112	AVSEIQLMHNLGHH	120	Training
113	AVSEIQLMHNLGFH	60	Training
114	AVSEIQLMHNLGWH	130	Training
115	AVSEIQLMHNLGPH	40	Training
116	AVSEIQLMHNLGKA	80	Training
117	AVSEIQLMHNLGKS	30	Training
118	AVSEIQLMHNLGK	130	Training
119	AVSEIQLMHNLGKE	10	Training
120	AVSEIQLMHNLGKQ	30	Training
121	AVSEIQLMHNLGKK	60	Training
122	AVSEIQLMHNLGKR	120	Training
123	AVSEIQLMHNLGKF	160	Training
124	AVSEIQLMHNLGKW	165	Training
125	AVSEIQLMHNLGKP	25	Training
126	AVSEIQLMHNLGKH	105	Test
127	AVAEIQLMHNLGKH	150	Test
128	AVSEIQLMHALGKH	160	Test
129	AVSEIQLMHNLAKH	170	Test
130	AVSEIQLMHNRGKH	190	Test
131	AVAEIQLMHARAKW	273	Test
132	AVAEIQLMHQRAKH	223	Test
133	AVAEIQLMHARAKH	269	Test
134	AVSEIQLMHARAKH	254	Test
135	AVAEIQLMHNRAKH	259	Test
136	AVAEIQLMHARGKH	255	Test
137	AVSEIQLMHNRAKH	364	Test
138	AVSEIQLMHARGKH	337	Test
139	AVAEIQLMHNRGKH	316	Test
140	AVAEIQLMHALAKH	217	Test
141	AVSEIQLMHALAKH	344	Test
142	AVAEIQLMHNLAKH	294	Test
143	AVAEIQLMHALGKH	230	Test
144	AVAEIQLMHNLGKG	0	Test
145	AVSEIQLMVNLGKH	0	Test

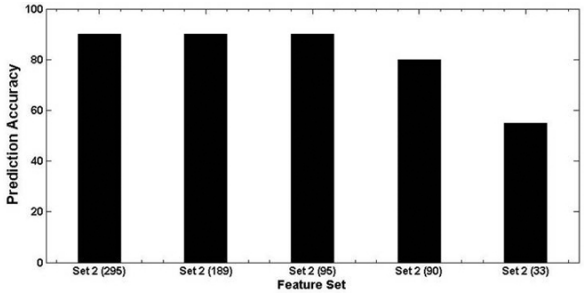

Fig. 3. Prediction results with reduced numbers of physicochemical property sets.

made up for the insufficiency of the data in set_1 as mentioned in "Materials and Methods", allowing for a better predictive value. These results are consistent with those published by Bhasin et al. (2004) and Sarda et al. (2005). The former improved the prediction results for the affinity of TAP binding peptides by adding property features into the sequence feature sets, and the latter developed a protein localization prediction algorithm with physicochemical properties from the AAindex. Furthermore, the results obtained from training with set_3 were the same as that with set_2. Since set_3 was composed of only property features, the results suggest that signaling potency was more affected by the physicochemical properties at each position than by the amino acid. These results demonstrate that the property features correlate better with the stimulus potency of PTH (1-14) and that the model generated on the basis of sequence plus properties is more reasonable.

Analysis of positional properties

The next step was to determine which property at each position correlates with potency. At first, 295 properties were selected from the AAindex based on the correlation analysis (Kawashima et al., 1999). The optimal model described above enabled us to reduce the size of the property sets by investigating the effects of eliminating a certain property. If the elimination did not affect the performance of the SVM model, that property is not involved with potency. Figure 3 illustrates the prediction results with decreasing numbers of feature sets based on their p-value. Performance was maintained up to a p-value of 0.015. Through this process, we chose 95 properties as shown in Table 3. The positional correlation between these 95 properties and the signaling potency was analyzed using Pearson's correlation coefficient based on the approach used in Bhasin et al. (2004).

Figure 4 shows the positional correlation of 20 properties among the feature sets (95 total). These 20 properties relate to two or more positions of the PTH (1-14) analogues. The x and y axes represent peptide position (1-14) and correlation coefficient, respectively. The title of each graph represents the AAindex ID. The analysis indicated that a certain residue and specific properties were preferred at each position. The characteristics of AURR980103, RICJ880106, AURR980106, TANS770107, SUEM840101, and RICJ880109 relate to helix frequencies, and they have a positive correlation with R9, R4/R10, R4, R6, R11/R14, and R11/13, respectively (Aurora and Rose, 1998; Richardson and Richardson, 1988; Tanaka and Scheraga, 1977). The secondary structure was also an important positional property, consistent with the experimental results published by Shimizu et al. (2003). They showed that modified PTH (1-14) analogues, such as [$Aib^{1,3}$, Gln^{10}, Har^{11}, Ala^{12}, Trp^{14}] PTH (1-14), exhibit enhanced signaling potency beyond that of PTH (1-34). The α-aminoisobutyric acid (Aib) and homoarginine

Table 3. Physicochemical properties of feature sets selected by property analysis using SVMs. The names of the properties in each column represent the AAindex ID.

AAindex				
NISK860101	RICJ880111	BIOV880102	LIFS790103	NADH010106
PARS000101	CIDH920104	CHAM830106	NADH010105	PARS000102
VINM940102	WIMW960101	NAKH900108	PONP800108	RICJ880106
AURR980103	LEVM760102	WILM950102	TAKK010101	BHAR880101
VINM940103	BROC820102	CHAM830104	FUKS010111	RICJ880109
CIDH920105	CHOP780210	MUNV940103	PONP800102	OOBM850103
KARP850101	FAUJ880105	QIAN880137	PONP800106	CHOC760101
RACS770101	LEVM760105	AURR980106	VENT840101	OOBM770102
ROSG850101	ROBB790101	KARP850102	WOLS870103	AURR980104
VINM940101	NOZY710101	PONP800101	GUYH850101	MEEJ810101
WERD780101	QIAN880121	CHAM830102	FODM020101	SUEM840101
MEIH800101	SNEP660103	CHOP780206	MANP780101	MEIH800102
OOBM770103	MAXF760103	PONP800107	WILM950101	WILM950104
PONP930101	MIYS850101	PRAM820103	RACS770102	KRIW790103
RICJ880107	ROSM880103	ROBB760105	DAYM780201	FASG760101
CIDH920101	YANJ020101	DAWD720101	NADH010104	CHARGE0101
LEVM760104	ZASB820101	MEIH800103	AURR980105	KLEP840101
NISK800101	BIOV880101	ROSG850102	FINA910101	GRAR740103
PARJ860101	SWER830101	FINA910104	TANS770106	FUKS010103

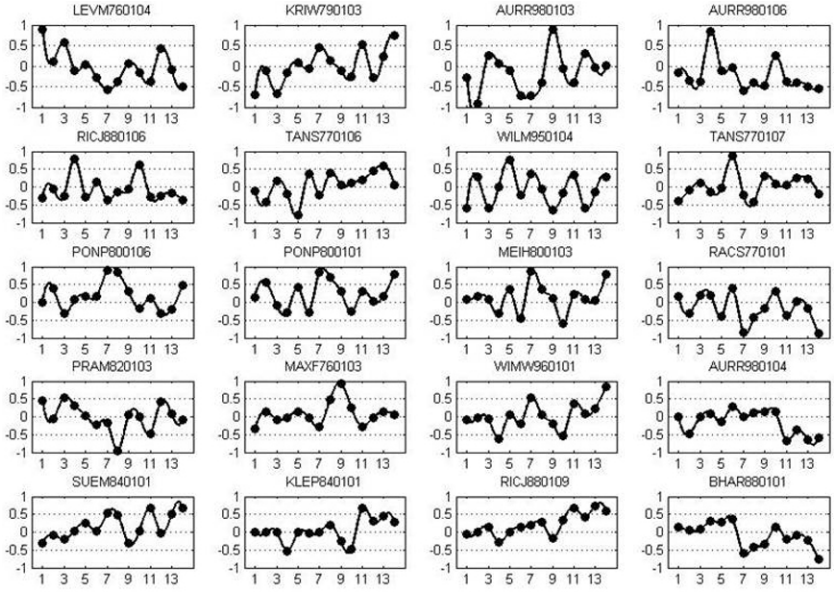

Fig. 4. Positional correlation between physicochemical properties of amino acids and signaling potency of PTH (1-14) analogues. The x and y represent peptide position (1-14) and correlation coefficient, respectively. The title of each graph represents the AAindex ID.

residue (Har) are modified amino acids that promote α-helix formation (Karle and Balaram, 1990; Mita et al., 2004). This indicated that the helix-promoting amino acids improve the capacity of PTH to activate its receptor. To test whether an α-helix is required for receptor activation, Tsomaia et al. (2004) designed lactam-bridged analogues. The analogues exhibited stronger potency than the linear peptides due to the helicity induced by the lactam bridge (Tsomaia et al., 2004). The shorter analogue [Ala^{10}, Gln^{10}, Har^{11}] PTH (1-11) also displayed signaling potency when the helicity was induced in the N-terminal region (Shimizu et al., 2001), supporting the hypothesis that helix formation is important for receptor activation. However, the results reported by Barazza et al. (2005) indicated that the presence of a stable N-terminal helical sequence was important, but not sufficient, for biological activity. They designed PTH (1-11) analogues substituted at positions 1, 3, or both by Aib, Ac_5c (1-aminocyclopentane-1-carboxylic acid) and Ac_6c (1-aminocyclohexane-1-carboxylic acid) to investigate the role of the α helix. Like Aib, Ac_5c and Ac_6c can promote α helix conformation because they reduce the flexibility of the peptides by imposing steric constrictions into the peptide chain that limit conformational freedom (Kowalczyk et al., 2005). Barazza et al.

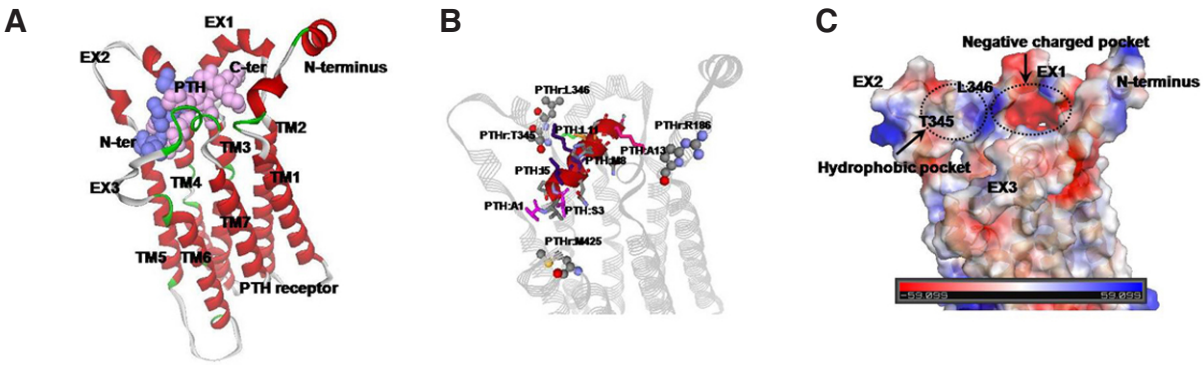

Fig. 5. Comparative modelling structures of PTH (1-14)/PTHr. The structures were constructed using Modeller 7v7. (A) The complex structure between ligand and receptor. The hydrophobic residues of the ligand PTH (1-14) are denoted in blue and the hydrophilic residues are denoted in pink. (B) The binding pocket and the interactions between ligand and receptor: the 1st residue of PTH (R1, pink) interacted with Met425 of PTHr, the 13th residue of PTH (R13, magenta) interacted with Arg186 of PTHr, and the hydrophobic residues of PTH (R5, R7, and R8, dark blue) interacted with the hydrophobic residues of PTHr (Thr345, Leu346). The 11th residue, which prefers a positive charge, is shown in green. (C) Molecular surface of receptor, colour coded according to the electrostatic potential. This model was built using PYMOL (www.pymol.org). The hydrophobic and negatively charged pockets were observed at the entrance of the receptor.

observed that substitutions at residues 1 and 3 with Ac_5c and Aib or with Aib and Ac_5c enhanced the extent of helix formation. Ligand-stimulated cAMP accumulation experiments indicated that the analogues were active. The first analogue, which had the most helical content, exhibited biological activity 3500-fold higher than that of native PTH (1-11) and only 15-fold weaker than that of native PTH (1-34) (Barazza et al., 2005). However, the helix content alone was not sufficient to explain the activity difference among the analogues. While Ac_6c also enhanced the helix content, Ac_5c analogue had much higher activity. Compared to Ac_6c, the side chain ring of Ac_5c is reduced by one CH_2 group; thus, this side chain is less bulky and less flexible. These results showed that both N-terminal helical stability and side chain properties are important. This is consistent with our results from the property analysis using SVMs.

LEVM760104 exhibited strong positive correlation with the first (R1) and third residues (R3) (Levitt, 1976). This property is the side-chain torsion angle taken from 13 well-refined protein conformations. The standard geometry of R1/R3 is important for potency. The side-chain volume, KRIW790103, correlated negatively with R1/R3, yet positively with R11/R14 (Krigbaum and Komriya, 1979). RACS770101, which was the average reduced distance of C_α, also showed negative correlation with R14 (Rackovsky and Scheraga, 1977). In other words, R1 and R3 preferred a smaller volume, while R11 and R14 preferred a higher volume. The strong preference for hydrophobicity was observed in R7, R8, and R14. PONP800106, PONP800101, WILM950104, WIMW960101, and MEIH800103 were scales of amino acid hydrophobicity in specific environments (Meirovitch et al., 1980; Ponnuswamy et al., 1980; Wilce et al., 1995). R8 may favor a specific backbone topology and prefer to interact with the hydrophobic residues of the receptor rather than with solvent because it correlated positively with a frequency of zeta R (MAXF760103) and negatively with the solvent accessible property (PRAM820103) (Maxifield and Scherga, 1976; Prabhakaran and Ponnuswamy, 1982). R14 also showed strong negative correlation with the average flexibility parameter, BHAR880101 (Bhaskaran and Ponnuswamy, 1988). Therefore, high-volume, non-flexible, and hydrophobic properties are required for receptor binding. R11 demonstrated strong positive correlation with net charge, KLEP840101, suggesting a positive charge was preferred at the 11th position (Klein et al., 1984).

According to published mutagenesis studies, N-terminal residues are involved in receptor activation while C-terminal residues participate in receptor binding (Potts, 2005). Compared to published data, our findings showed that signaling potency depends on the size and conformation of the N-terminal residues and binding depends on electrostatic properties, size, and hydrophobicity of the C-terminal residues.

Model evaluation via comparative studies

The sequence activity relationship (SAR) was analyzed using the model we developed. This result was compared to that of structure modeling to verify the proposed method. Protein structure can be determined either by X-ray crystallography or by NMR spectroscopy. However, due to experimental difficulties in determining the structure of G protein-coupled receptors like the PTH receptor, Jin et al. (2000) built a three-dimensional structure of the receptor using molecular modeling (PDB ID:1et3). Rölz et al. built a complex model based on the crystal structure of the hormone PTH (1-34) and the structure of the receptor. We used this complex model to analyze the correlation between the experimental data and molecular interactions (Jin et al., 2000; Rölz et al., 1999). Because only the xyz coordinates of C_α were published in the theoretical model deposited in PDB, side chains were built using the Modeller 7v7 comparative modeling package in this study (Sail and Blundell, 1993). Figure 5A shows the complex structure model of PTH (1-14)/PTHr. This model was used to support three characteristics found in the properties analysis by SVMs. First, R1 was located on or near the top of transmembrane helix TM6 and extracellular loop EC3 as shown in Fig. 5B. This finding was consistent with the published result that R1 crosslinks to Met425 of the receptor in photoaffinity crosslinking studies (Bisello et al., 1998). Replacement of Ala by Ser, as in PTHrP, should have only minor effects on the binding affinity as both residues were of similar size and no specific interactions of the serine hydroxyl group were identified (Rölz et al., 1999). Therefore, size was important for the residue to fit into the binding pocket, as mentioned in the previous section. The small side chain volume of R3 was also important for the stability of the α helix structure. Second, the hydrophobic pockets of Thr345 and Leu346 were observed around R7 and R8. Residues of both R5 and R8 were directed toward the top of the receptor molecule where

the hydrophobic pockets were composed of the extracellular ends of hPTHr TM7. These findings support the analysis results that these three residues (R5, R7, R8) prefer hydrophobic properties. Third, the electrostatic potential model (Fig. 5C) supports the positive charge preference of R11. The activity of the analogues increased when this residue was mutated to positively charged residues such as Lys and Arg in the property analysis with SVMs. Figure 5C shows that R11 was located within the negatively charged pocket formed by EC1. Therefore, its position can be stabilized through electrostatic interactions with the receptor. The predicted structure was used to evaluate the SAR analysis, and the results suggest that the model developed using SVMs is useful in the property analysis of the target peptide.

CONCLUSIONS

The objectives of this study were to predict the signaling potency of modified PTH (1-14) analogues using the mono-substitution experimental data, and to analyze the key physicochemical properties at each position that correlate with bioactivity. The model was based on single mutants performed well. Inclusion of the physicochemical properties enhanced the classification performance. The important properties found using the proposed procedure were consistent with published data. This systematic approach can reduce the time and effort needed to obtain bioactive molecules and provide useful information in the design of more potent drug candidates.

ACKNOWLEDGMENTS

This work was supported by the Korea Science and Engineering Foundation grant (to W.L) funded by the Korea government (Ministry of Science and Technology) (R01-2007-000-10161-0). This work was also supported in part by the Brain Korea 21 program.

REFERENCES

Aurora, R., and Rose, G.D. (1998). Helix capping. Protein Sci. *7*, 21-38.

Barazza, A., Wittelsberger, A., Fiori, N., Schievano, E., Mammi, S., Toniolo, C., Alexander, J.M., Rosenblatt, M., Peggion, E., and Chorev, M. (2005). Bioactive N-terminal undecapeptides derived from parathyroid hormone: the role of alpha-helicity. J. Pept. Res. *65*, 23-35.

Bhasin, M., and Raghava, G.P. (2004). Analysis and prediction of affinity of TAP binding peptides using cascade SVM. Protein Sci. *13*, 596-607.

Bhaskaran, R., and Ponnuswamy, P.K. (1988). Positional flexibilities of amino acid residues in globular proteins. Int. J. Peptide Protein Res. *32*, 241-255.

Bisello, A., Adams, A.E., Mierke, D.F., Pellegrini, M., Rosenblatt, M., Suva, L.J., and Chorev, M. (1998). Parathyroid hormone-receptor interactions identified directly by photocross-linking and molecular modeling studies. J. Biol. Chem. *273*, 22498-22505.

Chorev, M. (2002). Parathyroid hormone 1 receptor: insights into structure and function. Receptors Channels *8*, 219-242.

Crammer, K., and Singer, Y. (2001). On the algorithm implementation of multi-class SVM. J.M.L.R. *2*, 265-292.

Cristianini, N., and Shawe-Taylor, J. (2000). An introduction to support vector machines and other kernel-based learning methods (New York: Cambridge University Press).

Dubey, A., Realff, M.J., Lee, J.H., and Bommarium, A.S. (2004). Support vector machines for learning to identify the critical position of a protein. J. Theor. Biol. *234*, 351-361.

Jin, L., Briggs, S.L., Chandrasekhar, S., Chirgadze, N.Y., Clawson, D.K., Schevitz, R.W., Smiley, D.L., Tashjian, A.H., and Zhang, F. (2000). Crystal structure of human parathyroid hormone 1-34 at 0.9-A resolution. J. Biol. Chem. *275*, 27238-27244.

Joachims, T. (1999). Making large-scale SVM learning practical. In B. Scholkopf, ed., Advances in kernel methods - support vector learning, (Cambridge, Massachusetts: The MIT Press).

Karle, I.L., and Balaram, P. (1990). Structural characteristics of alpha-helical peptide molecules containing Aib residues. Biochemistry *29*, 6747-6756.

Kawashima, S., Ogata, H., and Kanehisa, M. (1999). AAindex: amino acid index database. Nucleic Acids Res. *27*, 368-369.

Klein, P., Kanehisa, M., and DeLisi, C. (1984). Prediction of protein function from sequence properties. Discriminant analysis of a data base. Biochim. Biophys. Acta *787*, 221-226.

Kowalczyk, W., Prahl, A., Dawidowska, O., Derdowska, I., Sobolewski, D., Hartrodt, B., Neubert, K., Slaninova, J., and Lammek, B. (2005). The influence of 1-aminocyclopentane-1-carboxylic acid at position 2 or 3 of AVP and its analogues on their pharmacological properties. J. Pept. Sci. *11*, 584-588.

Krigbaum, W.R., and Komoriya, A. (1979). Local interactions as a structure determinant for protein molecules: II. Biochim. Biophys. Acta *576*, 204-248.

Levitt, M. (1976). A simplified representation of protein conformations for rapid simulation of protein folding. J. Mol. Biol. *104*, 59-107.

Luck, M.D., Carter, P.H., and Gardella, T.J. (1999). The (1-14) fragment of parathyroid hormone (PTH) activates intact and amino-terminally truncated PTH-1 receptors. Mol. Endocrinol. *13*, 670-680.

Martinez, W.L., and Martinez, A.R. (2002). Computational statistics handbook with MATLAB. (Boca Raton, London, New York, Washington D.C.: Chapman and Hall/CRC).

Maxfield, F.R., and Scheraga, H.A. (1976). Status of empirical methods for the prediction of protein backbone topography. Biochemistry *15*, 5138-5153.

Meirovitch, H., Rackovsky, S., and Scheraga, H.A. (1980). Empirical studies of hydrophobicity: 1. Effect of protein size on the hydrophobic behavior of amino acids. Macromolecules *13*, 1398-1405.

Mita, K., Ichimura, S., and Zama, M. (2004). Conformation of poly (L-homoarginine). Biopolylmers *19*, 1123-1135.

Ponnuswamy, P.K., Prabhakaran, M., and Manavalan, P. (1980). Hydrophobic packing and spatial arrangement of amino acid residues in globular proteins. Biochim. Biophys. Acta *623*, 301-316.

Potts, J.T. (2005). Parathyroid hormone: past and present. J. Endocrinol. *187*, 311-325.

Prabhakaran, M., and Ponnuswamy, P.K. (1982). Shape and surface features of globular proteins. Macromolecules *15*, 314-320.

Rackovsky, S., and Scheraga, H.A. (1977). Hydrophobicity, hydrophilicity, and the radial and orientational distributions of residues in native proteins. Proc. Natl. Acad. Sci. USA *74*, 5248-5251.

Richardson, J.S., and Richardson, D.C. (1988). Amino acid preferences for specific locations at the ends of alpha helices. Science *240*, 1648-1652.

Rölz, C., Pellegrini, M., and Mierke, D.F. (1999). Molecular characterization of the receptor-ligand complex for parathyroid hormone. Biochemistry *38*, 6397-6405.

Sali, A., and Blundell, T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. *234*, 779-815.

Sarda, D., Chua, G.H., Li, K.B., and Krishnan, A. (2005). pSLIP: SVM based protein subcellular localization prediction using multiple physicochemical properties. BMC Bioinformatics *6*, 152.

Scholkopf, B., and Smola, A.J. (2002). Learning with Kernels (Cambridge: The MIT Press).

Shimizu, M., Potts, J.T., Jr., and Gardella, T.J. (2000). Minimiz. of parathyroid hormone. Novel amino-terminal parathyroid hormone fragments with enhanced potency in activating the type-1 parathyroid hormone receptor. J. Biol. Chem. *275*, 21836-21843.

Shimizu, M., Carter, P.H., Khatri, A., Potts, J.T., Jr., and Gardella, T.J. (2001). Enhanced activity in parathyroid hormone-(1-14) and -(1-11): novel peptides for probing ligand-receptor interactions. Endocrinology *142*, 3068-3074.

Supper, J. (2005). Predicting MHC class I binding peptides based on amino acid properties using decision trees and support vector machines (Tübingen: Department for Simulation of Biological Systems, University of Tübingen).

Tanaka, S., and Scheraga, H.A. (1977). Statistical mechanical treatment of protein conformation. 5. A multistate model for specific-sequence copolymers of amino acids. Macromolecules *10*, 9-20.

Toschantaridis, I., Hofmann, T., Joachims, T., and Altun, Y. (2004).

Support vector machine learning for interdependent and structured output spaces. (Banff, Canada: 21st International Conference on Machine Learning).

Tsomaia, N., Pellegrini, M., Hyde, K., Gardella, T.J., and Mierke, D.F. (2004). Toward parathyroid hormone minimization: conformational studies of cyclic PTH(1-14) analogues. Biochemistry *43*, 690-699.

Vapnik, V.N. (1995). The Natural of Statistical Learning Theory (New York: Springer Verlag).

Wilce, M.C., Aguilar, M.I., and Hearn, M.T. (1995). Physicochemical basis of amino acid hydrophobicity scales: evaluation of four new scales of amino acid hydrophobicity coefficient derived from RP-HPLC of peptides. Anal. Chem. *67*, 1210-1219.