Prediction of Parathyroid Hormone Signalling Potency Using SVMs

Ahrim Yoo, Sunggeon Ko, Sung-Kil Lim, Weontae Lee, 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, \( \phi(x) \), 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...
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) are mapped into the feature space by the mapping function φ(x) and then separated by the hyperplane expressed as follows:

\[ F(x) = w^T \phi(x) + b \]  

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

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

\[
\begin{align*}
\min_{w,b} & \quad \frac{1}{2} w^T w + C \sum_{i=1}^{M} \zeta_i \\
\text{s.t.} & \quad y_i (w^T \phi(x_i) + b) \geq 1 - \zeta_i, \quad \forall j = 1, \ldots, M \\
& \quad \zeta_i \geq 0
\end{align*}
\]  

(2)

In this objective function, the first part maximizes the margin, while the second part relaxes the maximum-margin constraints. The slack variable (\( \zeta_i \)) 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:

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

where \( \alpha \) is the Lagrange multiplier.

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

\[
\begin{align*}
\max & \quad \sum_{i=1}^{M} \alpha_i - \frac{1}{2} \sum_{i=1}^{M} \sum_{j=1}^{M} y_i y_j \alpha_i \alpha_j K(x_i, x_j) \\
\text{s.t.} & \quad \sum_{i=1}^{M} y_i \alpha_i = 0 \\
& \quad C \geq \alpha_i \geq 0
\end{align*}
\]  

(4)

In this case, the decision function is written as

\[
f(x) = \sum_{i=1}^{M} \alpha_i y_i K(x_i, x) + b
\]  

(5)

where \( K(x, x) \) is the kernel defined as the inner product of two vectors in the feature space:

\[
K(x_i, x_j) = \phi(x_i)^T \phi(x_j)
\]  

(6)

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

\[
K(x_i, x_j) = \exp(-|x_i - x_j|^2), \quad \gamma > 0
\]  

(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 \( K(x, x) \) 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 SVMmulticlass 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 SVMmulticlass package. When k classes were given, they suggested another approach to solve a single optimization problem with the following objective function (8):

\[
\begin{align*}
\min_{w_m} & \quad \frac{1}{2} \sum_{i=1}^{M} w_m^T w_m + C \sum_{i=1}^{M} \xi_i \\
\text{s.t.} & \quad w_m^T \phi(x_i) \geq \delta_m - \xi_i, \quad \forall i = 1, \ldots, M \\
& \quad \xi_i \geq 0
\end{align*}
\]  

(8)

where \( \xi_i = 1 - \delta_m, \xi_m = 0 \) if \( y_i = m \) and \( \delta_m = 1 \).

Then the decision function is defined as:

\[
\arg \max_{m=1,\ldots,M} w_m^T \phi(x)
\]  

(9)

This function uses only \( l \) slack variables \( \xi_i (i = 1, \ldots, l) \). In other methods, the number of classifiers is \( k(k-1)/2 \) 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):

\[
\xi_i = \max(0, \max(w_m^T \phi(x_i) + \delta_m) - w_m^T \phi(x_i), 0)
\]  

(10)

This makes the multi-classification problem simpler. In this