Cell proliferation, differentiation and death are controlled by a multitude of cell–cell signals and loss of this control has devastating consequences. Prominent among these regulatory signals is the cytokine superfamily, which has crucial functions in the development, differentiation and regulation of immune cells. In this study, a support vector machine (SVM)-based method was developed for predicting families and subfamilies of cytokines using dipeptide composition. The taxonomy of the cytokine superfamily in which our method complies was described in the Cytokine Family cDNA Database (dbCFC) and the dataset used in this study for training and testing was obtained from the dbCFC and Structural Classification of Proteins (SCOP). The method classified cytokines and non-cytokines with an accuracy of 92.5% by 7-fold cross-validation. The method is further able to predict seven major classes of cytokine with an overall accuracy of 94.7%. A server for recognition and classification of cytokines based on multi-class SVMs has been set up at http://bioinfo.tsinghua.edu.cn/~huangni/CTKPred/. The performance of the module was evaluated using a 7-fold cross-validation test. The SVM was trained with a randomly selected from the SCOP version 1.37 PDB90 domain set of id90 and thus resulted in 437 sequences. Then the dataset was extended by adding 673 additional negative examples using CD-HIT software (Li et al, 2001, 2002) by a threshold of id90 and thus resulted in 437 sequences. Then the dataset was extended by adding 673 additional negative examples randomly selected from the SCOP version 1.37 PDB90 domain data. The performance of the module was evaluated using a 7-fold cross-validation test. The SVM was trained with a...
fixed-dimensions (400) vector obtained on the basis of the dipeptide composition of protein sequences.

**Recognition of cytokine family**

Cytokines can be divided into seven major classes: FGF/HBGF, IL-6, LIF/OSM, MDK/PTN, NGF, TGF-β and TNF. The dataset consisted of 83 sequences from FGF/HBGF, 22 sequences from IL-6, 12 sequences from LIF/OSM, 10 sequences from MDK/PTN, 24 sequences from NGF, 190 sequences from TGF-β and 96 sequences from TNF. Because of a lack of adequate sequences, we put IL-6, LIF/OSM, MDK/PTN and NGF into a single class (thus containing 68 sequences) through the rest of process (hence there were then four major classes). Classification of cytokines into one of these four classes is a multi-class classification problem. Therefore, a multi-class SVM was employed to classify sequences from all possible classes. The vectors were extracted from the dipeptide composition of proteins. The performance of SVM classification was evaluated using 7-fold cross-validation.

**Recognition of subfamilies**

Classifying a cytokine to the subfamily level is of greater significance to further specific studies. Therefore, we chose to classify the TGF-β family which possesses most known sequences to a lower level, since other families lack enough sequences for SVM training and cross-validation. As described in Figure 1, TGF-β can be divided into six major subfamilies: bone morphogenetic protein (BMP), growth differentiation factor (GDF), glial-derived neurotrophic factor (GDNF), inhibin (INHA/INHB), transforming growth factor β (TGFβ) and others. Again, a multi-class SVM was constructed for this multi-class classification problem and the performance was evaluated using 2-fold cross-validation because of the smaller number of sequences.

**Support vector machines**

SVMs are a class of statistical learning algorithms whose theoretical basis was first presented by Vapnik (1982). After the 1990s, they became extremely popular in the machine-learning community (Cristianini and Shawe-Taylor, 2000; Hua and Sun, 2001a,b; Bhasin and Raghava, 2004; Guo et al., 2004). In this study, the SVM was implemented using the freely downloadable software package libsvm written by Chang and Lin (2001). The software, which features an efficient multi-class classification, enables the user to define a number of parameters and to select from a choice of inbuilt kernel functions, including a radial basis function (RBF) and a polynomial kernel (of given degree). The experimentation was conducted using an RBF kernel. The SVM was provided with fixed-length vector input. The fixed-length feature vector was obtained from proteins of variable length using dipeptide composition.

**Dipeptide composition**

The dipeptide composition used as input provides global information on protein features in the form of a fixed-length vector. Dipeptide composition encapsulates information about the fraction of amino acids and their local order. The dipeptide composition of each protein was calculated using the following equation:

\[
\text{fraction of dep (i)} = \frac{\text{total number of dep (i)}}{\text{total number all possible dipeptides}}
\]

where dep (i) is a dipeptide i out of 400 dipeptides. In this study, three SVMs were constructed: one for discriminating cytokine proteins from other proteins such as globular proteins, the second for predicting the family of cytokines and the third for predicting the subfamily of certain cytokine families.

**Performance evaluation**

The performance of SVMs in distinguishing cytokines from non-cytokines was evaluated using 7-fold cross-validation. In this approach, the dataset was partitioned randomly into seven equal-sized sets. The training and testing of each classifier was carried out seven times using one distinct set for testing and the other sets for training. Four threshold-dependent parameters, sensitivity, specificity, accuracy and Matthews’s correlation coefficient (MCC) (Hua and Sun, 2001b), were used to measure the performance of this module. The performance of SVM modules constructed for recognizing cytokine family and subfamily were evaluated using 7- and 2-fold cross-validation, respectively, also measured by sensitivity, specificity, accuracy and MCC. Calculations of sensitivity, specificity, accuracy and MCC were carried out as follows:

\[
\text{sensitivity} = \frac{TP}{(TP + FN)}
\]

\[
\text{specificity} = \frac{TN}{(TN + FP)}
\]

\[
\text{accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)}
\]

\[
\text{MCC} = \frac{(TP \times TN - FP \times FN)}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}}^{0.5}
\]

where TP, TN, FP and FN represent true positive, true negative, false positive and false negative, respectively.

![Fig. 1](https://academic.oup.com/peds/article-abstract/18/8/365/1457986/146786)

Fig. 1. The hierarchical structure of the cytokine superfamily. The cytokine superfamily consists of seven major families of proteins; each can be further divided into subfamilies, e.g. the largest family, TGF-β, is comprised of six major subfamilies.
Results and discussion

The performance of the module developed for discriminating between cytokines and non-cytokines is summarized in Table I. The results show that the module can distinguish cytokines from other protein sequences with an accuracy of 95.3% and an MCC of 0.90, when evaluated through 7-fold cross-validation. The results were obtained using the RBF kernel with $\gamma = 100$ and parameter $C = 1000$.

This dipeptide composition-based method was compared with Pfam server prediction which based on HMM on the same dataset. The performance of Pfam is shown in Table I. The Pfam method discriminated between cytokines and non-cytokines with an accuracy of 94.1% and an MCC of 0.88, both of which are lower than with the dipeptide composition-based method. This confirms that the dipeptide composition is a better feature for recognizing cytokines from non-cytokine proteins. Further, the SVM method was much less time consuming than the HMM method.

To predict the family of cytokines, a multi-class SVM was constructed. The SVM was trained and tested using dipeptide composition and evaluated by 7-fold cross-validation. The performance in recognizing different classes of cytokines is summarized in Table II. As shown, our method discriminated the four families of cytokines with an accuracy of 96.9% and an MCC of 0.93 on average.

To predict further the subfamilies of the recognized cytokines in order to assign its function, we again constructed a multi-class SVM for classifying the TGF-β family. The performance was evaluated through a two-fold cross-validation owing to the smaller number of sequences and the results are shown in Table III. This method discriminated the six major classes of TGF-β subfamilies with an accuracy of 96.6% and an MCC of 0.92.

Table I. Performance of cytokine superfamily recognition

<table>
<thead>
<tr>
<th>Methods</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>Accuracy (%)</th>
<th>MCC</th>
<th>Time span</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>92.5</td>
<td>97.2</td>
<td>95.3</td>
<td>0.90</td>
<td>&lt;2 s$^a$</td>
</tr>
<tr>
<td>Pfam</td>
<td>92.9</td>
<td>94.7</td>
<td>94.0</td>
<td>0.87</td>
<td>~20 h$^a$</td>
</tr>
</tbody>
</table>

$^a$These time spans were obtained under identical conditions on an Anthlon 64 3000+, 1 G memory machine.

Fig. 2. Snapshot of the CTKPred web server interface. (a) Users can either input protein sequence in the text area or upload the sequence file in FASTA format. (b) The result of the prediction is displayed on-screen in a user-friendly format with basic sequence information. Users also have the option to receive the result by e-mail.
subfamilies of TGF-β with an accuracy of 90.1% and an MCC of 0.74. Less accurate results were obtained owing to the smaller number of sequences. It is well established that machine learning methods require large number of examples for reliable prediction.

Nevertheless, this dipeptide composition-based SVM approach provides a highly accurate and time-saving method that is able to recognize unknown sequences to cytokine subfamily level, and it is hoped that it will have broad applications ranging from assisting further experimental study to facilitating drug screening.

Cytokinepred server

Based on our study, we constructed a freely accessible web server at http://bioinfo.tsinghua.edu.cn/~huangni/CTKPred/ that allows users to recognize and classify cytokines from protein sequence. The common gateway interface (CGI) script is written in PERL version 5.8.4. Users can enter one or more protein sequences at a time in FASTA format by copy and paste or file upload. The result of the prediction will be displayed in a user-friendly format on the screen or e-mailed to the users if provided with a valid e-mail address. The interface of our web server is shown in Figure 2.

Acknowledgements

We are grateful to Dr Chih-Chung Chang and Dr Chih-Jen Lin for providing the LIBSVM software. This work was supported by a Foundational Science Research Grant from Tsinghua University (No. JC2001043), a National Nature Science Grant (No. 19947006) and the 863 Projects (2002AA234041).

References


Received March 5, 2005; accepted May 18, 2005

Edited by Paul Carter

Table II. Performance of cytokine family recognition

<table>
<thead>
<tr>
<th>Cytokine family</th>
<th>Accuracy (%)</th>
<th>MCC</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FGF/HBGF</td>
<td>97.5</td>
<td>0.92</td>
<td>92.7</td>
<td>98.6</td>
</tr>
<tr>
<td>Joint classa</td>
<td>98.4</td>
<td>0.94</td>
<td>91.0</td>
<td>99.7</td>
</tr>
<tr>
<td>TGF-β</td>
<td>95.8</td>
<td>0.92</td>
<td>97.4</td>
<td>94.7</td>
</tr>
<tr>
<td>TNF</td>
<td>97.7</td>
<td>0.94</td>
<td>94.0</td>
<td>98.8</td>
</tr>
</tbody>
</table>

Table III. Performance of cytokine subfamily recognition

<table>
<thead>
<tr>
<th>TGF-β subfamily</th>
<th>Accuracy (%)</th>
<th>MCC</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMP</td>
<td>86</td>
<td>0.67</td>
<td>87.5</td>
<td>85.5</td>
</tr>
<tr>
<td>GDF</td>
<td>93</td>
<td>0.76</td>
<td>82.4</td>
<td>95.2</td>
</tr>
<tr>
<td>GDNF</td>
<td>98</td>
<td>0.86</td>
<td>75</td>
<td>100</td>
</tr>
<tr>
<td>INH</td>
<td>92</td>
<td>0.65</td>
<td>46.7</td>
<td>100</td>
</tr>
<tr>
<td>TGFβ</td>
<td>99</td>
<td>0.96</td>
<td>100</td>
<td>98.9</td>
</tr>
<tr>
<td>Other</td>
<td>84</td>
<td>0.56</td>
<td>66.7</td>
<td>89.5</td>
</tr>
</tbody>
</table>

aJoint class includes IL-6, LIF/OSM, MDK/PTN and NGF.