

COMMUNICATION

CTKPred: an SVM-based method for the prediction and classification of the cytokine superfamily

Ni Huang, Hu Chen and Zhirong Sun¹

Institute of Bioinformatics and System Biology, MOE Key Laboratory of Bioinformatics, State Key Laboratory of Biomembrane and Membrane Biotechnology, Department of Biological Science and Biotechnology, Tsinghua University, Beijing 100084, China

¹To whom correspondence should be addressed.
E-mail: sunzhr@mail.tsinghua.edu.cn

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 with 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/>.

Keywords: classification/dipeptide composition/cytokine/prediction/support vector machine/SVM

Introduction

Cytokines, a diverse group of polypeptides that are generally associated with inflammation, immune activation and cell differentiation or death, include interleukins (IL), interferons (IFNs), tumor necrosis factors (TNFs) and various growth factors, including transforming growth factor β (TGF- β), fibroblast growth factor (FGF), heparin binding growth factor (HBGF) and neuron growth factor (NGF) (Benveniste, 1998). Recent studies have revealed that this superfamily of proteins participate in various new biological processes (Kleemann *et al.*, 2000; Allan and Rothwell, 2001; Derouet *et al.*, 2004; Dranoff, 2004; Ueki *et al.*, 2004). For example, the mixture of cytokines that is produced in the tumor microenvironment has an important role in cancer pathogenesis: cytokines that are released in response to infection, inflammation and immunity can function to inhibit tumor development and progression (Dranoff, 2004). Cytokines also respond to brain injury and have diverse actions that can cause, exacerbate, mediate and/or inhibit cellular injury and repair (Allan and Rothwell, 2001).

Besides its many novel functions, an increasing number of newly discovered molecules have been identified as members of the cytokine superfamily. Although the sequences of these molecules are quickly accumulating, for a large proportion of them their precise function remains unclear. Indeed, laboratory work is essential and irreplaceable in the procedure to confirm a protein's structure and function, but might appear too expensive and lengthy when applied on a large scale. Computational methods, however, provide the possibility of a quicker and less expensive solution. Although several methods, such as BLAST, HMM and ANN, have been exploited for protein family prediction, less effort has been devoted to the prediction of cytokines from sequence data (Altschul *et al.*, 1990; Papasaikas *et al.*, 2003; Bhasin and Raghava, 2004).

This paper describes a support vector machine (SVM)-based method developed for the recognition of cytokines on the basis of dipeptide composition. The method uses a three-step strategy. First, a protein sequence is examined to determine whether it belongs to the cytokine superfamily. If it is recognized as a cytokine, the method then predicts to which family of cytokine it belongs. Finally, it classifies the protein to subfamily level if it belongs to the TGF- β family of cytokines. The performance of this method was evaluated in each step on independent and non-redundant datasets created in this study. An online web server was also developed on the basis of the above method and is freely accessible at <http://bioinfo.tsinghua.edu.cn/~huangni/CTKPred/>.

Method and procedure

We adopted a three-step strategy for recognizing cytokines from protein sequences and further classifying cytokines to subfamily level. The method was trained using fixed-length vectors obtained on the basis of the dipeptide composition of proteins. The accuracy of each step was evaluated by cross-validation.

Recognition of cytokine superfamily

First, we developed an SVM module for identifying cytokines from protein sequence data uncovered by various genome-sequencing projects. The original dataset, obtained from <http://cytokine.medic.kumamoto-u.ac.jp/>, consisted of 1173 cytokines belonging to the eight major classes. Next we excluded highly homologous sequences within the dataset 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 (TGFB) 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 $\text{dep } (i)$ 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:

$$\begin{aligned} \text{sensitivity} &= \text{TP}/(\text{TP} + \text{FN}) \\ \text{specificity} &= \text{TN}/(\text{TN} + \text{FP}) \\ \text{accuracy} &= (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}) \\ \text{MCC} &= \frac{(\text{TP} \times \text{TN} - \text{FP} \times \text{FN})}{\{[(\text{TP} + \text{FN})(\text{TN} + \text{FP})]^{0.5} (\text{TP} + \text{FP})(\text{TN} + \text{FN})\}} \end{aligned}$$

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

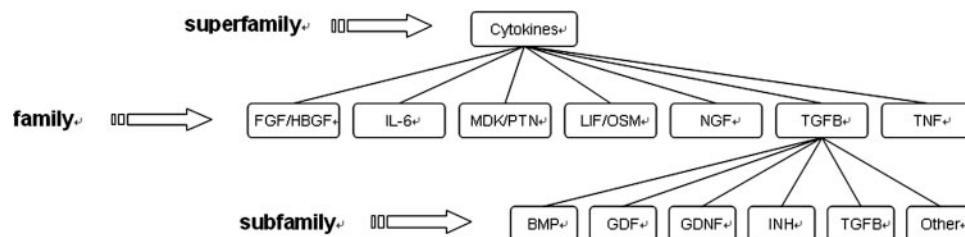


Fig. 1. The hierarchical structure of the cytokine superfamily. The cytokine superfamily consists of seven major families of proteins; each can be further divided into subfamilies, e.g. the largest family, TGF-β, is comprised of six major subfamilies.

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.

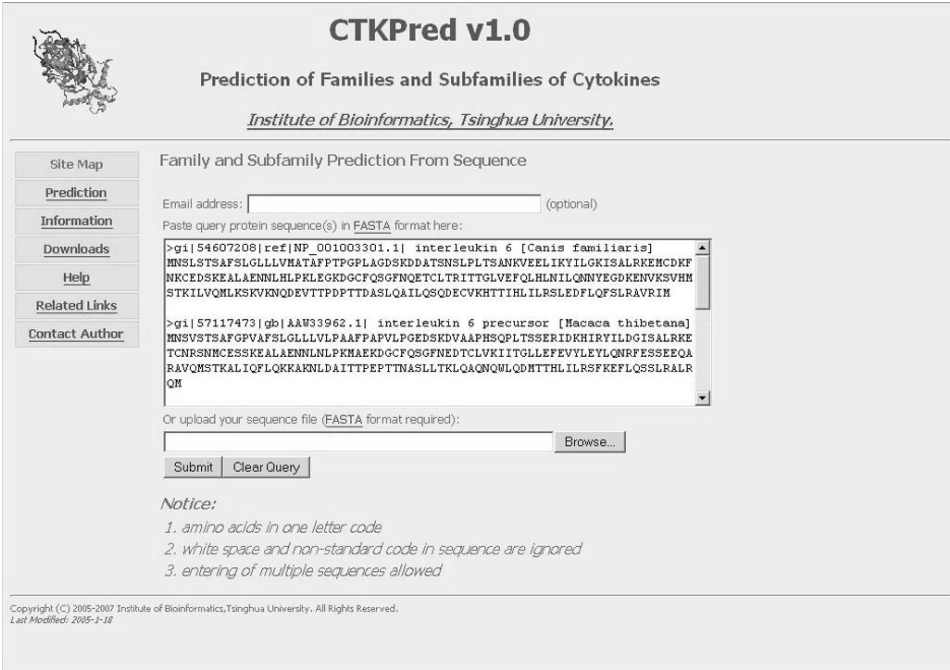
To predict further the subfamilies of the recognized cytokines in order to assign its function, we again constructed a multi-class SVM for classifying the TGF- β family. The performance was evaluated through a two-fold cross-validation owing to the smaller number of sequences and the results are shown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

Table I. Performance of cytokine superfamily recognition

Methods	Sensitivity (%)	Specificity (%)	Accuracy (%)	MCC	Time span
SVM	92.5	97.2	95.3	0.90	<2 s ^a
Pfam	92.9	94.7	94.0	0.87	~20 h ^a

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

(a)



CTKPred v1.0
Prediction of Families and Subfamilies of Cytokines
Institute of Bioinformatics, Tsinghua University.

Family and Subfamily Prediction From Sequence

Site Map | **Prediction** | Information | Downloads | Help | Related Links | Contact Author

Email address: (optional)

Paste query protein sequence(s) in FASTA format here:

```
>gi|54607208|ref|NP_001003301.1| interleukin 6 [Canis familiaris]
MNSLSTSAFSLGLLVHATAFPPLAGDSKDDATSNLPLTSANKVEELIKYILGKISALRKEHCKDF
NKEDSKEALAENNLHLPKLEKGGCFQSGFNQETCLTRITTLGVEFQLHLNLIKQNYEGDKENKSVHM
STKILVQLKSKVKNQDEVTTDPDPTDASLQAILQSQDQECVKHTTIHLILRSLEDLQFSLRAVRIM

>gi|57117473|gb|AAW33962.1| interleukin 6 precursor [Macaca thibetana]
MNSVSTSAFGPVAFSLGLLVLPAAFPAPVLPGEDSKDVAAPHSQPLTSSERIDKHIRYILDGISAALRKE
TCNRSNMCESSEKALAENNLNLPKMAEKDGCFCQSGFNEDTCLVKIITGLLEFEVYLEYLQNRFESEEA
RAVQMSKRALIQFLQKARKNLDAITTPPTNASLLTKLQAGNQWLQDHTTHLILRSFKEFLQSSLRALR
QM
```

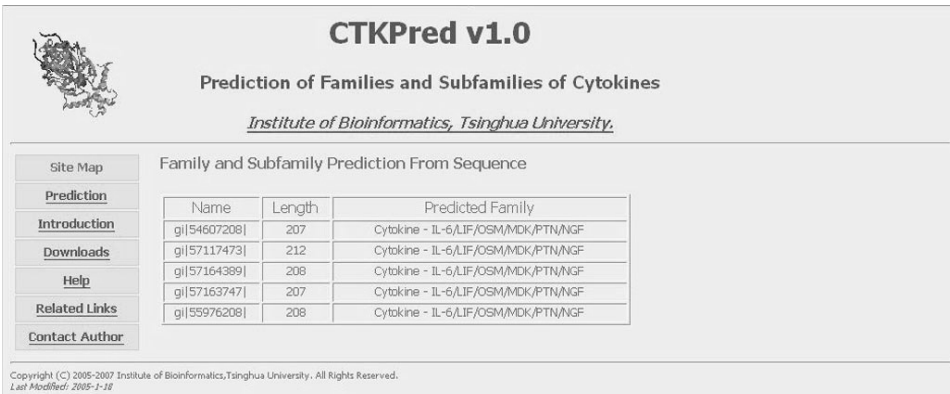
Or upload your sequence file (FASTA format required):

Notice:

1. amino acids in one letter code
2. white space and non-standard code in sequence are ignored
3. entering of multiple sequences allowed

Copyright (C) 2005-2007 Institute of Bioinformatics, Tsinghua University. All Rights Reserved.
Last Modified: 2005-3-18

(b)



CTKPred v1.0
Prediction of Families and Subfamilies of Cytokines
Institute of Bioinformatics, Tsinghua University.

Family and Subfamily Prediction From Sequence

Site Map | **Prediction** | Introduction | Downloads | Help | Related Links | Contact Author

Name	Length	Predicted Family
gi 54607208	207	Cytokine - IL-6/LIF/OSM/MDK/PTN/NGF
gi 57117473	212	Cytokine - IL-6/LIF/OSM/MDK/PTN/NGF
gi 57164389	208	Cytokine - IL-6/LIF/OSM/MDK/PTN/NGF
gi 57163747	207	Cytokine - IL-6/LIF/OSM/MDK/PTN/NGF
gi 55976208	208	Cytokine - IL-6/LIF/OSM/MDK/PTN/NGF

Copyright (C) 2005-2007 Institute of Bioinformatics, Tsinghua University. All Rights Reserved.
Last Modified: 2005-3-18

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.

Table II. Performance of cytokine family recognition

Cytokine family	Accuracy (%)	MCC	Sensitivity (%)	Specificity (%)
FGF/HBGF	97.5	0.92	92.7	98.6
Joint class ^a	98.4	0.94	91.0	99.7
TGF- β	95.8	0.92	97.4	94.7
TNF	97.7	0.94	94.0	98.8

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

Table III. Performance of cytokine subfamily recognition

TGF- β subfamily	Accuracy (%)	MCC	Sensitivity (%)	Specificity (%)
BMP	86	0.67	87.5	85.5
GDF	93	0.76	82.4	95.2
GDNF	98	0.86	75	100
INH	92	0.65	46.7	100
TGFB	99	0.96	100	98.9
Other	84	0.56	66.7	89.5

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

- Allan,S.M. and Rothwell,N.J. (2001) *Nat. Rev. Neurosci.*, **2**, 734–744.
 Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.
 Benveniste,E.N. (1998) *Cytokine Growth Factor Rev.*, **9**, 259–275.
 Hasin,M. and Raghava,G.P. (2004) *Nucleic Acids Res.*, **32**(Web Server issue), W383–W399.
 Chang,C.-C. and Lin,C.-J. (2001) <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
 Cristianini,N. and Shawe-Taylor,J. (2000) *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge.
 Derouet,D., Rousseau,F., Alfonsi,F., Froger,J., Hermann,J., Barbier,F., Perret,D., Diveu,C., Guillet,C., Preisser,L., et al. (2004) *Proc. Natl Acad. Sci. USA*, **101**, 4827–4832.

- Dranoff,G. (2004) *Nat Rev. Cancer*, **4**, 11–22.
 Guo,J., Chen,H., Sun,Z. and Lin,Y. (2004) *Proteins*, **54**, 738–743.
 Hua,S. and Sun,Z. (2001a) *J. Mol. Biol.*, **308**, 397–407.
 Hua,S. and Sun,Z. (2001b) *Bioinformatics*, **17**, 721–728.
 Kleemann,R., Hausser,A., Geiger,G., Mischke,R., Burger-Kentischer,A., Flieger,O., Johannes,F.J., Roger,T., Calandra,T., Kapurniotu,A. et al. (2000) *Nature* **408**, 211–216.
 Li,W., Jaroszewski,L. and Godzik,A. (2001) *Bioinformatics*, **17**, 282–283.
 Li,W., Jaroszewski,L. and Godzik,A. (2002) *Bioinformatics*, **18**, 77–82.
 Papasaikas,P.K., Bagos,P.G., Litou,Z.I. and Hamodrakas,S.J. (2003) *SAR QSAR Environ. Res.*, **14**, 413–420.
 Ueki,K., Kondo,T., Tseng,Y.H. and Kahn,C.R. (2004) *Proc. Natl Acad. Sci. USA*, **101**, 10422–10427.
 Vapnik,V.N. (1979) *Estimation of Dependencies Based on Empirical Data*. Springer-Verlag, Berlin.

Received March 5, 2005; accepted May 18, 2005

Edited by Paul Carter