**Protein Engineering, Design & Selection** vol. 21 no. 4 pp. 279–282, 2008
Published online February 20, 2008 doi:10.1093/protein/gzn006

**SHORT COMMUNICATION**

# CytoPred: a server for prediction and classification of cytokines

**Sneh Lata and G.P.S. Raghava**[1]

Bioinformatics Centre, Institute of Microbial Technology, Chandigarh, India

[1]To whom correspondence should be addressed. E-mail: raghava@imtech.res.in

**Cytokines are messengers of immune system. They are small secreted proteins that mediate and regulate the immune system, inflammation and hematopoiesis. Recent studies have revealed important roles played by the cytokines in adjuvants as therapeutic targets and in cancer therapy. In this paper, an attempt has been made to predict this important class of proteins and classify further them into families and subfamilies. A PSI-BLAST+Support Vector Machine-based hybrid approach is adopted to develop the prediction methods. CytoPred is capable of predicting cytokines with an accuracy of 98.29%. The overall accuracy of classification of cytokines into four families and further classification into seven subfamilies is 99.77 and 97.24%, respectively. It has been shown by comparison that CytoPred performs better than the already existing CTKPred. A user-friendly server CytoPred has been developed and available at http://www.imtech.res.in/raghava/cytopred.**

*Keywords*: cytokine/prediction/PSI-BLAST/SVM

## Introduction

Cytokines are hormone-like proteins that enable immune cells to communicate, and play an integral role in the initiation, perpetuation and subsequent down-regulation of the immune response. They are small secreted proteins that possess pleiotropic functions and mediate systemic and local biological actions. Studies have revealed that cytokines have important role in the pathogenesis and progression of diseases like rheumatoid arthritis (Feldmann *et al*., 1996), Chron's disease (Pizarro and Cominelli, 2007), and inflammation, and hold promise to act as therapeutic targets. Various cytokines and growth factors are believed to orchestrate cellular behavior in a healing cornea (Saika, 2007) and also play a role in injury-induced neural damage and repair (Allan and Rothwell, 2001; Schroeter and Jander, 2005). Cytokines are used in cytokine therapies to help immune system to recognize and destroy those cells that are cancerous (Tagawa, 2000; Kalaaji, 2007; Weiss *et al*., 2007) and in therapies that are routinely used by people living with HIV (Kim *et al*., 1999). Therefore, they have applications in the treatment of hematologic malignancies and immunogenic tumors. Since cytokines were also found to be the effector molecules for many adjuvant effects (Afonso *et al*., 1994; Staats and Enni, 1999; Lori *et al*., 2006; Eaton, 2007; Pardoll, 1995), there has been an effort to build the optimal vaccine adjuvant effect one cytokine at a time.

Keeping in mind such diverse roles played by the cytokines, identification of these cytokine would enable us to dissect the complex reactions and to advance our knowledge on how an immune system is operated. A large amount of sequence data are piling up with the completion of ongoing genome-sequencing projects, but the functional class of several proteins still remains unclear. Thus, computer-aided prediction of cytokines from a large amount of sequence data whose function is still largely unknown would be very fruitful for biologists as the experimental determination of the functions would be a laborious and time-consuming job. Earlier an attempt has been made where a Support Vector Machine (SVM)-based method has been developed for the prediction and classification of cytokine superfamily (Huang *et al*., 2005). In this paper, an attempt has been made to achieve higher prediction accuracy for the cytokine prediction and classification. A hybrid approach of PSI-BLAST and SVM was adopted in order to predict the cytokines and classify further them into families and subfamilies.

## Methods

### Cytokine prediction

For cytokine prediction, the dataset contained a total of 1110 sequences, having 437 positive and 673 negative examples (randomly selected from the SCOP version 1.37 PDB90 domain data). This dataset was the same as used by the method CTKPred and is downloaded from http://cytokine.medic.kumamoto-u.ac.jp/. No two examples in the dataset are >90% similar, i.e. the dataset is non-redundant. The prediction for any query protein was done at first by similarity search-based module, the PSI-BLAST, and if in case it failed to generate any hits, the prediction was done based on SVM. For SVM-based predictions, dipeptide composition of the sequences was taken as an input. In this study, a cut-off value was chosen where the sensitivity and specificity were nearly equal or the difference between them is the least, for evaluating and developing SVM-based methods.

Evaluation of the cytokine prediction was done using 7-fold cross validation technique. The data were randomly divided into seven sets, each set containing almost equal number of examples. The method was trained on six sets and tested on the remaining one set. This was repeated seven times, so that each set was used once as test set. The method achieved an accuracy of 98.29% at cut-off where the sensitivity and specificity were nearly equal. The performance of the hybrid method for cytokine prediction is given in Table I.

### Cytokine classification

*Family classification* The dataset contains sequences form seven major families of cytokines containing about 83 FGF/

Downloaded from peds.oxfordjournals.org by guest on April 28, 2011

© The Author 2008. Published by Oxford University Press. All rights reserved.
For Permissions, please e-mail: journals.permissions@oxfordjournals.org

**Table I.** Comparison of performance of various methods for predicting cytokines

| Method | Sensitivity (%) | Specificity (%) | Accuracy (%) | MCC |
|---|---|---|---|---|
| SVM (aa comp.) | 91.99 | 96.14 | 94.50 | 0.88 |
| SVM (dip. comp.) | 92.91 | 97.47 | 95.68 | 0.91 |
| Hybrid (PSI-BLAST+SVM) | 98.40 | 98.22 | 98.29 | 0.96 |
| CTKPred | 92.5 | 97.2 | 95.3 | 0.90 |

aa comp., amino acid composition; Dip. Comp., dipeptide composition.

HBGF family sequences, 22 IL-6 family sequences, 12 LIF/OSM family sequences, 10 MDK/PTN family sequences, 24 NGF family sequences, 190 TGF-B family sequences and 96 TNF family sequences. Therefore, a prediction method to classify the cytokines into families was also developed. As the number of sequences in families LIF/OSM, DK/PTN and NGF were not enough, these were clubbed together to form a single class called ‘joint class’. Thus, the method was developed to predict four families of cytokines instead of seven. For family classification, the dataset consisted of 437 cytokine sequences only. As it is multi-class prediction problem, we developed a series of classifiers to handle the problem. N SVMs were constructed for N-class classification. For cytokine family classification, the number of classes was equal to 4. The *i*th SVM was trained with all the samples of *i*th class labeled positive and all other samples labeled negative. An unknown example was classified into the class that corresponds to the SVM with the highest output score. The results for the family prediction are given in Table II.

*Subfamily classification* Among all the families, TGF-$\beta$ was the only family having sufficient number of sequences that could be used for training and testing in order to develop a prediction method. So, we chose to develop a method that could also predict the subfamily of cytokines if they belonged to the TGF-$\beta$ family. As the TGF-$\beta$ family can be further divided into six subfamilies including bone morphogenetic protein (BMP), growth differentiation factor (GDF), glial-derived neurotrophic factor (GDNF), inhibin (INHA/INHB), transforming growth factor-$\beta$ (TGF$\beta$) and others, again a multi-class classification was done as described for cytokine family prediction. The evaluation was done by 2-fold cross validation. The results for subfamily prediction are given in Table III.

### Evaluation parameters

The evaluation of performance of the method was done by calculating the sensitivity, specificity, accuracy and the MCC of the prediction. The formulae for calculating these parameters are as follows:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \times 100$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \times 100$$

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \times 100$$

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

## Results and discussions

The performance of the method for predicting the cytokines, based on amino acid composition, dipeptide composition, CytoPred (based on Hybrid approach), is compiled and presented in Table I. Among these, the performance of the amino acid composition-based method was the poorest. Composition-based method gave an accuracy of 94.50% and an MCC of 0.88. The performance of dipeptide composition-based method performed better than simple composition-based method (with accuracy and MCC), but was further outperformed by the Hybrid approach-based method. The hybrid approach-based method performed best and achieved accuracy 98.29% and MCC 0.91. The performance is also compared with that of an already existing method, CTKPred (Table I). The hybrid approach-based method outshines the performance of CTKPred too.

An attempt was made to further classify the cytokines into families and subfamilies. The classification of cytokine family was done at first by using the dipeptide composition alone (Table II). The dipeptide-based cytokine family classification achieved an overall accuracy of 96.34% and an average MCC of 0.95. The classification was done by using the hybrid approach. The performance of the hybrid approach is compared one-to-one with that of the method CTKPred (Table II). The overall accuracy achieved by hybrid module was 99.77% and the average MCC was 0.99, which is better than that achieved by CTKPred.

Similar trend was followed in further classification of TGF-$\beta$ family into subfamilies. The detailed results of the

**Table II.** Performance of various methods (SVM alone, CytoPred and CTKPred for Cytokine family classification)

| Family | Sensitivity (%) | | | Specificity (%) | | | Accuracy (%) | | | MCC | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SVM (Dip.) | CytoPred (Hybrid) | CTKPred | SVM (Dip.) | CytoPred (Hybrid) | CTKPred | SVM (Dip.) | CytoPred (Hybrid) | CTKPred | SVM (Dip.) | CytoPred (Hybrid) | CTKPred |
| FGF/HBGF | 93.98 | 100 | 92.7 | 99.44 | 100 | 98.6 | 98.40 | 100 | 97.5 | 0.95 | 1.00 | 0.92 |
| TGF-$\beta$ | 97.37 | 100 | 91.0 | 97.17 | 99.73 | 99.7 | 97.25 | 99.77 | 98.4 | 0.94 | 0.99 | 0.94 |
| TNF | 96.68 | 100 | 97.4 | 100.00 | 99.60 | 94.7 | 99.31 | 99.77 | 95.8 | 0.98 | 1.00 | 0.92 |
| Joint class | 95.59 | 97.92 | 94.0 | 98.10 | 100 | 98.8 | 97.71 | 99.54 | 97.7 | 0.92 | 0.99 | 0.94 |

Dip., dipeptide composition.

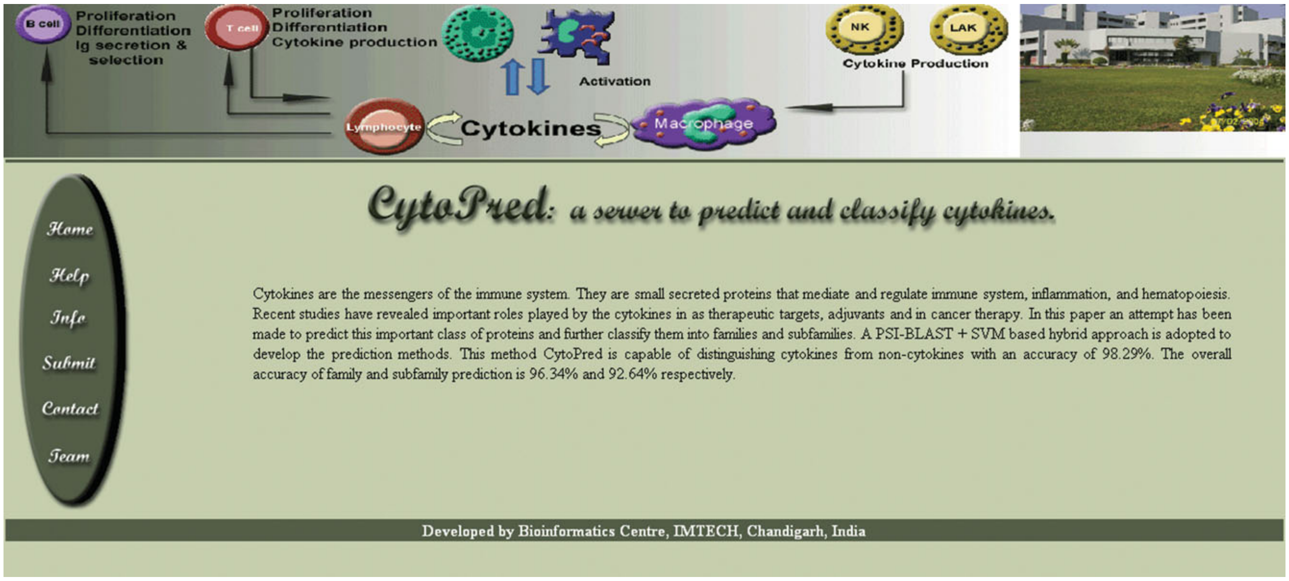
Downloaded from peds.oxfordjournals.org by guest on April 28, 2011

**Table III.** Performances of various methods (SVM alone, CytoPred and CTKPred for Cytokine sub-family classification)

| Subfamily | Sensitivity (%) | | | Specificity (%) | | | Accuracy (%) | | | MCC | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SVM (Dip.) | CytoPred (Hybrid) | CTKPred | SVM (Dip.) | CytoPred (Hybrid) | CTKPred | SVM (Dip.) | CytoPred (Hybrid) | CTKPred | SVM (Dip.) | CytoPred (Hybrid) | CTKPred |
| BMP | 80.43 | 97.83 | 87.5 | 92.36 | 95.83 | 85.5 | 89.47 | 96.32 | 86 | 0.72 | 0.91 | 0.67 |
| GDF | 84.38 | 96.88 | 82.4 | 96.84 | 98.73 | 95.2 | 94.74 | 98.42 | 93 | 0.81 | 0.94 | 0.76 |
| GDNF | 86.67 | 93.33 | 75 | 98.86 | 100.00 | 100 | 97.89 | 99.47 | 98 | 0.86 | 0.94 | 0.86 |
| INH | 92.86 | 100.00 | 46.7 | 98.77 | 98.77 | 100 | 97.89 | 98.95 | 92 | 0.92 | 0.96 | 0.65 |
| TGF-$\beta$ | 91.30 | 91.30 | 100.00 | 100.00 | 98.80 | 98.9 | 98.95 | 97.89 | 99 | 0.95 | 0.90 | 0.96 |
| Others | 71.74 | 82.61 | 66.7 | 90.97 | 99.31 | 89.5 | 86.32 | 95.26 | 84 | 0.63 | 0.87 | 0.56 |

Dip., dipeptide composition.

performance of the dipeptide composition-based module alone and the one-to-one comparison of hybrid model against CTKPred are given in Table III. The overall accuracy achieved by CytoPred in subfamily classification was 97.24% and the average MCC achieved was 0.92, which were again better than that achieved by CTKPred.

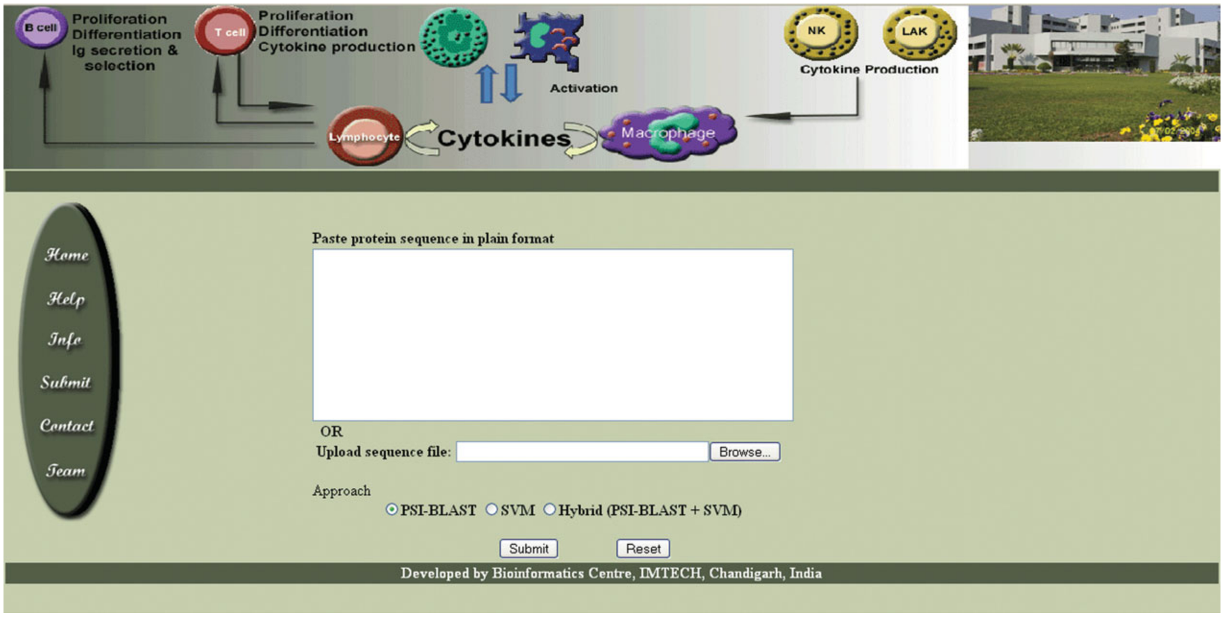


CytoPred: a server to predict and classify cytokines.

Cytokines are the messengers of the immune system. They are small secreted proteins that mediate and regulate immune system, inflammation, and hematopoiesis. Recent studies have revealed important roles played by the cytokines in as therapeutic targets, adjuvants and in cancer therapy. In this paper an attempt has been made to predict this important class of proteins and further classify them into families and subfamilies. A PSI-BLAST + SVM based hybrid approach is adopted to develop the prediction methods. This method CytoPred is capable of distinguishing cytokines from non-cytokines with an accuracy of 98.29%. The overall accuracy of family and subfamily prediction is 96.34% and 92.64% respectively.

Developed by Bioinformatics Centre, IMTECH, Chandigarh, India

**Fig. 1.** A snapshot of CytoPred server home page.



Paste protein sequence in plain format

OR

Upload sequence file:

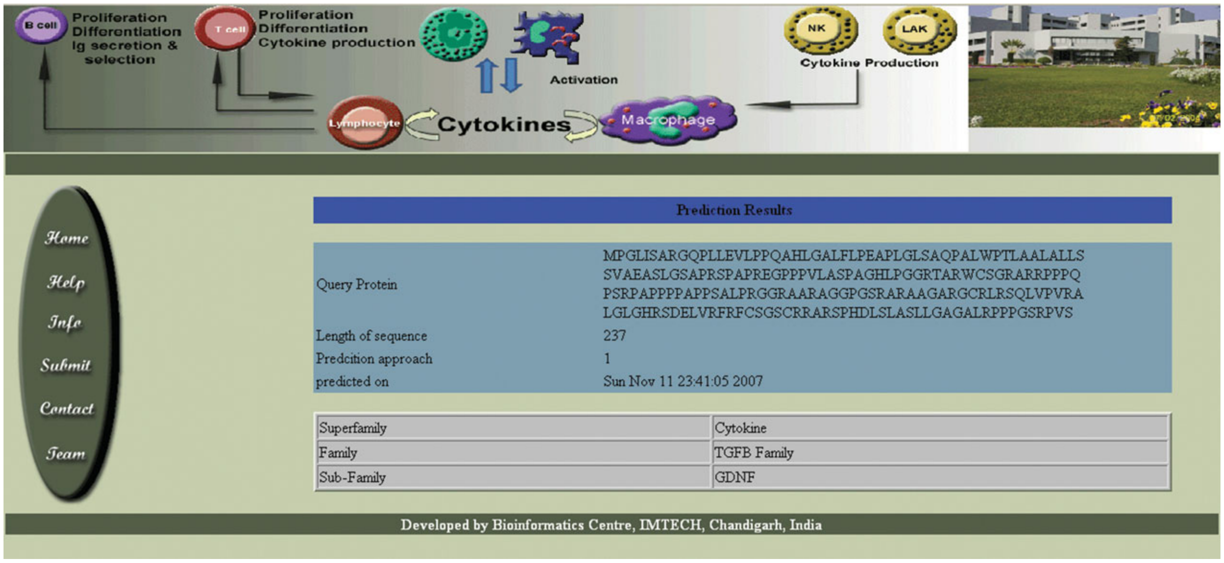
Approach

● PSI-BLAST ○ SVM ○ Hybrid (PSI-BLAST + SVM)

Developed by Bioinformatics Centre, IMTECH, Chandigarh, India

**Fig. 2.** Submission page of CytoPred.

Downloaded from peds.oxfordjournals.org by guest on April 28, 2011



Prediction Results

| | |
|---|---|
| Query Protein | MPGLISARGQPLLEVLPPQAHLGALFLPEAPLGLSAQPALWPTLAALALLS<br>SVAEASLGSAPRSPAPREGPPPVLASPAGHLPGGRTARWCSGRARRPPPQ<br>PSRPAPPPPAPPSALPRGGRAARAGGPGSRARAAGARGCRLRSQLVPVRA<br>LGLGHRSDELVRFRFCSGSCRRARSPHDLSLASLLGAGALRPPPGSRPVS |
| Length of sequence | 237 |
| Predcition approach | 1 |
| predicted on | Sun Nov 11 23:41:05 2007 |

| | |
|---|---|
| Superfamily | Cytokine |
| Family | TGFB Family |
| Sub-Family | GDNF |

Developed by Bioinformatics Centre, IMTECH, Chandigarh, India

**Fig. 3.** The result generated by server CytoPred.

CytoPred can predict as well as classify a cytokine protein with high accuracy as well as with high sensitivity and specificity. Thus, the PSI-BLAST+SVM-based method is better than SVM alone in predicting and classifying the query sequences to cytokines. We hope that our method would be of great help in order to annotate the proteins and would aid the experimental validation, in turn, saving time and labor.

## Web server

All the modules constructed in this study have been implemented on the World Wide Web as a dynamic web server 'CytoPred' (Fig. 1), which is available at http://www.imtech.res.in/raghava/cytopred. All the CGI scripts of the method were written in PERL5.0 and the interface was designed using HTML. The SVM and PSI-BLAST were implemented by obtaining SVM_light from http://www.cs.cornell.edu/People/tj/svm_light/ and PSI-BLAST from http://www.ncbi.nlm.nih.gov/blast/. It is a user-friendly web server which allows users to submit their protein sequence by typing or pasting in box or by using the file upload facility (Fig. 2). The server provides an option to select the prediction approach. In the case of default prediction, the server uses the hybrid module for prediction. However, if the user wants to make predictions for multiple sequences at a time, he is suggested to use the SVM-based model. The server presents the results of comprehensive analysis in user-friendly format (Fig. 3).

## Funding

Authors are thankful to Council of Scientific and Industrial Research (CSIR) and Department of Biotechnology (DBT), Government of India for financial support.

## References

Feldmann,M., Brennan,F.M. and Maini,R.N. (1996) *Annu. Rev. Immunol.*, **14**, 397–440.

Pizarro,T.T. and Cominelli,F. (2007) *Annu. Rev. Med.*, **58**, 433–444.

Saika,S. (2007) *Cornea*, **26**(Suppl. 9 1), S70–S74.

Allan,S.M. and Rothwell,N.J. (2001) *Nat. Rev. Neurosci.*, **2**, 734–744.

Schroeter,M. and Jander,S. (2005) *NeuroMol. Med.*, **7**, 183–195.

Tagawa,M. (2000) *Curr. Pharm. Des.*, **6**, 681–699.

Weiss,J.M., Subleski,J.J., Wigginton,J.M. and Wiltrout,R.H. (2007) *Expert Opin. Biol. Ther.*, **7**, 1705–1721.

Kalaaji,A.N. (2007) J. Drugs Dermatol, **6**, 374–378.

Kim,J.J., *et al.* (1999) *J. Interferon Cytokine Res.*, **19**, 77–84.

Afonso,L.C., *et al.* (1994) *Science*, **263**, 235–237.

Pardoll,D.M. (1995) *Annu. Rev. Immunol.*, **13**, 399–415.

Staats,H.F. and Enni,F.A. (1999) *J. Immunol.*, **162**, 6141–6147.

Lori,F., Weiner,D.B., Calarota,S.A., Kelly,L.M. and Lisziewicz,J. (2006) *Springer Semi. Immunopathol.*, **28**, 231.

Eaton,S.M., Maue,A.C., Blumerman,S.L. and Haynes,L. (2007) *J. Immunol.*, **178**, 85.22.

Huang,N., Chen,H. and Sun,Z. (2005) *Protein Eng. Des. Sel.*, **18**, 365–368.

**Received December 6, 2007; revised December 6, 2007; accepted January 18, 2008**

**Edited by Valerie Daggett**

Downloaded from peds.oxfordjournals.org by guest on April 28, 2011