Protein Engineering, Design & Selection vol. 22 no. 7 pp. 441–444, 2009
Published online June 2, 2009 doi:10.1093/protein/gzp016

SHORT COMMUNICATION

Prediction and classification of chemokines and their receptors

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

Bioinformatics Center, Institute of Microbial Technology, Sector 39A, Chandigarh, India

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

Chemokines are low molecular mass cytokine-like proteins that orchestrate myriads of immune functions like leukocyte trafficking, T cell differentiation, angiogenesis, hematopoeosis and mast cell degranulation. Chemokines also play a role as HIV-1 inhibitor and act as potent natural adjuvant in antitumor immunotherapy. Receptors for these molecules are all seven-pass transmembrane G-protein-coupled receptors that are intimately involved with chemokines in a wide array of physiological and pathological conditions. These receptors also have a major role as co-receptors for HIV-1 entry into target cells. Therefore, chemokine receptors have proven to be excellent targets for small molecule in pharmaceutical industry. The immense importance of chemokines and their receptors motivated us to develop a support vector machine-based method ChemoPred to predict this important class of proteins and further classify them into subfamilies. ChemoPred is capable of predicting chemokines and chemokine receptors with an accuracy of 95.08% and 92.19%, respectively. The overall accuracy of classification of chemokines into three subfamilies was 96.00% and that of chemokine receptors into three families was 92.87%. The server ChemoPred is freely available at www.imtech.res.in/raghava/chemopred.

Keywords: chemokine/chemokine receptor/prediction/SVM

Introduction

Chemokines are small protein sequences that are important components of innate immune system and are believed to regulate leukocyte trafficking during normal physiological processes and pathological conditions. Chemokines are large peptides, 70–125 amino acids in length, and are rich in basic amino acid residues. On the basis of the number and spacing of conserved cysteine residues that they contain, chemokines are further divided into four subfamilies (Rossi and Zlotnik, 2000) viz CXC or alpha chemokines, CC or beta chemokines, C or gamma chemokine and CX3C or delta chemokines. Chemokines were named so due to their chemoattractant properties that were first demonstrated in a chemotaxis assay for neutrophils (Yoshimura *et al.*, 1987). It has been now shown that chemokine has a role to play in leukocyte trafficking (Yoshimura *et al.*, 1987; Cyster, 1999; Peveri *et al.*, 1999), hematopoeisis (Cook, 1996; Quackenbush *et al.*, 1997; Kim and Broxmeyer, 1999), organogenesis (Ma *et al.*, 1998), neuronal communication (Harrison *et al.*, 1998) and modulate angiogenesis (Moore *et al.*, 1998). They are also suggested to act as inflammatory mediators (Gura, 1996; Luster, 1998). Some chemokines can function as inhibitors of HIV-1 (Cocchi *et al.*, 1995). Chemokines are also proposed to act as potent natural adjuvants for antitumor immunotherapy (Narvaiza, 2000; Ruehlmann *et al.*, 2001).

Chemokines bind to a set of receptors called chemokine receptors to perform all these vital functions. Engagement of chemokine receptors by appropriate chemokines leads to a cascade of downstream signaling event that ultimately results in the functions performed by chemokines as mentioned above. These receptors are a seven-pass transmembrane receptor proteins belonging to a subfamily of G-protein-coupled receptor (Rojo *et al.*, 1999; Sallusto *et al.*, 2000; Thelen, 2001). Like their ligands (i.e. the chemokines), chemokine receptors are also further subdivided into four subfamilies, i.e. CXCR, CCR, CR and CX3CR, based upon their primary amino acid sequence. Chemokines and chemokine receptors are intimately involved in a wide array of diseases (Sallusto *et al.*, 1998; Zlotnik *et al.*, 1999). Potent chemokine receptor agonists may, therefore, be useful to treat varied conditions such as inflammation and neoplasia. With the discovery of the fact that some chemokine receptors act as co-receptors for HIV entry into target cells (Alkhatib *et al.*, 1996; Choe *et al.*, 1996; Deng *et al.*, 1996; Doranz *et al.*, 1996; Dragic *et al.*, 1996), research has really been accelerated to find small chemicals that can block these receptors and thereby HIV entry into the target cells. Therefore, these receptors are considered to be of prime importance as potential drug target by the pharmaceutical industry.

Advances over the past few years have included the discovery of new chemokines, receptors and antagonists, and a greater appreciation for the diverse biological functions displayed by this cytokine family. Keeping in mind such diverse roles played by the chemokines and their receptors, identification of these 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 chemokines and chemokine receptors 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. In this paper, an attempt has been made to achieve to predict and classify chemokines and chemokine receptors. A support vector machine (SVM)-based approach was adopted in order to predict and further classify further them into subfamilies.

Methods

Data sets

Positive data set The positive data sets for both chemokines and chemokine receptors were downloaded from http://

cytokine.medic.kumamoto-u.ac.jp/. Chemokine data set had 431 protein sequences, whereas the data set for chemokine receptor had 314 protein sequences. CD-HIT software (Li and Godzik, 2006) was used to remove sequences that are highly homologous from these data sets, the cut-off being 90%. Thus, no two sequences in the data sets were >90% similar, and hence, the data sets used are non-redundant. After using CD-HIT, 193 chemokine and 96 chemokine receptor sequences were left in the data sets. These data sets (as appear in the respective order) were used as positive data sets for developing the method.

Negative data set A systematic approach was adopted to select the negative data sets to be used for training the method. Protein sequences were retrieved from UniProt (Apweiler *et al.*, 2004) using the mammalian proteins 'BUTNOT chemokines' criteria. Now a BLAST (Altschul *et al.*, 1990) search was performed against the database of these sequences for each and every protein in the positive data set. The result obtained was sequences that were homologous to chemokines but were performing different function. One hundred and ninety-three proteins were selected from the resulting sequences. These were used as negative data set against chemokines. Similar strategy was used to get 96 protein sequences to be used as negative data set against the positive chemokine receptors.

Prediction

The data set for chemokine prediction contained a total of 386 sequences (193 positive and 193 negative examples), whereas that for the chemokine receptor prediction contained a total of 192 sequences (96 positive and 96 negative). Amino acid compositions of the sequences were given as the input pattern to train the SVM. For evaluating and developing the method, a cut-off value was chosen where the sensitivity and specificity were nearly equal or the difference between them is the least.

Evaluation of the models was done using 5-fold cross-validation technique. The data were randomly divided into five sets, each set containing almost equal number of examples. The method was trained on four sets and tested on the remaining one set. This was repeated five times, so that each set was used once as test set.

We adopted another strategy for evaluating the performance of the method. We randomly picked 20% of the data set as the testing set and the remaining data set was used as the training set to train the model. This procedure was repeated 100 times and the results obtained in each cycle were averaged to get the final result.

Classification

Chemokine classification The data set contains sequences from four major families of chemokines containing about five C family sequences, 109 CC family sequences, 76 CXC family sequences and four CX3C family sequences. Therefore, a prediction method to classify the chemokines into families was also developed. As the number of sequences in C and CX3C families was not enough, these were clubbed together to form a single class called 'joint class'. Thus, the method was developed to predict three families of chemokines instead of four. For family classification, the data set consisted of 193 chemokine 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 chemokine family classification, the number of classes was equal to 3. 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 I.

Chemokine receptor classification The data set for chemokine receptors contained 96 sequences belonging to four families. These include five examples from CR family, 58 examples from CCR family, 29 examples from CXCR family and four examples from CX3CR family. As the number of sequences in CR and CX3CR families was very few, they were combined into a single family called joint class. Then, N SVMs were constructed as described in chemokine classification case. For chemokine receptor family classification, the number of classes was equal to 3. 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 I.

Support vector machine

In this study, all SVM models have been developed using a freely available program SVM_LIGHT (Joachims, 1999). This program allows users to run SVM using various kernels and parameters. In this study, the accuracy was computed at a cut-off score where sensitivity and specificity are nearly equal.

Evaluation parameters

The evaluation of performance of the method was done by calculating the sensitivity, specificity, accuracy and the Methew's correlation coefficient (MCC) of the prediction. Sensitivity is the percentage of positive examples (chemokines or chemokine receptors in this case), which are correctly predicted as positive. Specificity is the percentage of negative examples (non-chemokines or non-chemokine receptors in this case), which are correctly predicted to be negative. Accuracy is the percentage of correctly predicted positive and negative examples. It is a good measure to assess the performance of any method when the data set is balanced (equal number of positive and negative examples). It is also known as percent coverage. The formulae for calculating these parameters are as follows:

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

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

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

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

Table I. Performance of method ChemoPred for chemokine and its receptors subfamily classification

	Chemokine				Chemokine receptors			
	Sensitivity	Specificity	Accuracy	MCC	Sensitivity	Specificity	Accuracy	MCC
Joint class	100.00%	100.00%	100.00%	1.00	100.00%	100.00%	100.00%	1.00
CC	97.25%	96.43%	96.89%	0.94	94.83%	94.74%	94.79%	0.89
CXC	96.05%	97.44%	96.89%	0.93	93.10%	95.52%	94.79%	0.88

where, TP is true positive; TN, true negative; FP, false positive and FN, false negative.

Results and discussions

Chemokines and chemokine receptors are an important class of innate immunity molecules. To the best of author's knowledge, this is the first method to predict and classify chemokine and their receptors. Although methods have been developed earlier to predict and classify cytokines (Huang *et al.*, 2005; Lata and Raghava, 2008) and some receptors such as nuclear receptor (Bhasin and Raghava, 2004) and GPCR (Bhasin and Raghava, 2004), no such method is present for chemokines and their receptors. In this method, we adopted a systematic strategy to select the negative data set rather than selecting some random sequences as negative examples as selection of negative data set is always a challenge for any study. In order to minimize error in our negative examples, we selected negative examples/proteins only from Swiss-Prot (a database of annotated proteins). In other words, function of these proteins is known and it is not chemokines as per Swiss-Prot. Despite our careful selection of negative examples, it is still possible that some of negative examples may have chemokines. During creation of negative data set using BLAST, we cautiously removed the examples that by any chance happened to be the chemokines or chemokine receptors. This was done in order to lessen the bias in the method had the random sequences (which may be distantly remote and easily distinguishable) been used.

ChemoPred achieved an accuracy of 95.08% for chemokine prediction and 92.16% for chemokine receptor prediction using 5-fold cross-validation technique. The sensitivity and specificity achieved for chemokine prediction were 94.82% and 95.34%, respectively. The sensitivity and specificity for chemokines receptor prediction were 90.62% and 93.75%, respectively. We also plotted the sensitivity versus 1-specificity chart, i.e. receiver operator curve (ROC), from the prediction methods. The area under curve for chemokine prediction was 0.987 (Fig. 1) and that for their receptor prediction was 0.906 (Fig. 2).

On using the random samplings evaluation technique, the sensitivity, specificity and accuracy achieved for chemokines were 95.82, 95.84 and 95.44, respectively. For chemokine receptors, the sensitivity, specificity and accuracy achieved using this technique were 88.63, 92.98 and 90.55, respectively.

An attempt was made to further classify the chemokines and chemokine receptors into subfamilies. The amino acid composition-based chemokine subfamily classification

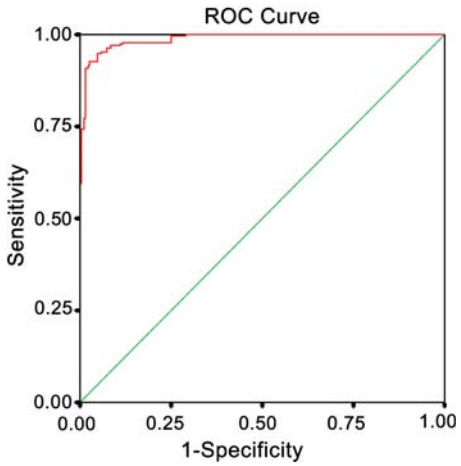

Fig. 1. ROC curve for chemokine prediction.

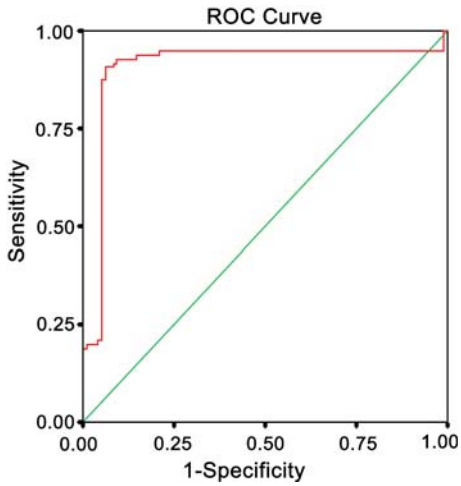

Fig. 2. ROC curve for chemokine receptor prediction.

(using 5-fold cross-validation technique) achieved an overall accuracy of 97.02% and an average MCC of 0.95. The overall accuracy and average MCC obtained in chemokine receptor subfamily classification were 90.17% and 0.92 (Table I).

ChemoPred can predict as well as classify a chemokine protein with high accuracy as well as with high sensitivity and specificity. 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.

Webserver

All the modules constructed in this study have been implemented on the World Wide Web as a dynamic webserver 'ChemoPred' which is available freely at www.imtech.res.in/raghava/chemopred. All the scripts of the method were written in CGI-PERL and the interface was designed using HTML. It is a user-friendly webserver that allows the users to submit their query sequence by typing or pasting it in the box or by using the file upload facility. The user can choose if they want to go for chemokine prediction or chemokine receptor prediction in the submission form. The result of predicted superfamily and subfamily of the query protein is displayed in a user-friendly tabular format.

Funding

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

References

Alkhatib,G., Combadiere,C., Broder,C.C., Feng,Y., Kennedy,P.E., Murphy,P.M. and Berger,E.A. (1996) *Science*, **272**, 1955–1958.
Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.
Apweiler,R., *et al*. (2004) *Nucleic Acids Res.*, **32**, D115–D119.
Bhasin,M. and Raghava,G.P.S. (2004) *Nucleic Acids Res.*, **32**, W383–W389.
Bhasin,M. and Raghava,G.P.S. (2004) *J. Biol. Chem.*, **279**, 23262–23266.
Choe,H., *et al*. (1996) *Cell*, **85**, 1135–1148.
Cocchi,F., DeVico,A.L., Garzino-Demo,A., Arya,S.K., Gallo,R.C. and Lusso,P. (1995) *Science*, **270**, 1811–1815.
Cook,D.N. (1996) *J. Leukoc. Biol.*, **59**, 61–66.
Cyster,J.G. (1999) *Science*, **286**, 2098–2102.
Deng,H., *et al*. (1996) *Nature*, **381**, 661–666.
Doranz,B.J., Rucker,J., Yi,Y., Smyth,R.J., Samson,M., Peiper,S.C., Parmentier,M., Collman,R.G. and Doms,R.W. (1996) *Cell*, **85**, 1149–1158.
Dragic,T., *et al*. (1996) *Nature*, **381**, 667–673.
Gura,T. (1996) *Science*, **272**, 954–956.
Harrison,J.K., *et al*. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 10896–10901.
Huang,N., Chen,H. and Sun,Z. (2005) *Protein Eng. Des. Sel.*, **18**, 365–368.
Joachims,T. (1999) Making large-Scale SVM learning practical. In Scholkopf,B., Burges,C. and Smola,A. (eds), *Advances in Kernel Methods—Support Vector Learning.* MIIT Press, Cambridge, MA, London, England.
Kim,C.H. and Broxmeyer,H.E. (1999) *J. Leukoc. Biol.*, **65**, 6–15.
Lata,S. and Raghava,G.P. (2008) *Protein Eng. Des. Sel.*, **21**, 279–282.
Li,W. and Godzik,A. (2006) *Bioinformatics*, **22**, 1658–1659.
Luster,A.D. (1998) *N. Engl. J. Med.*, **338**, 436–445.
Ma,Q., Jones,D., Borghesani,P.R., Segal,R.A., Nagasawa,T., Kishimoto,T., Bronson,R.T. and Springer,T.A. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 9448–9453.
Moore,B.B., Keane,M.P., Addison,C.L., Arenberg,D.A. and Strieter,R.M. (1998) *J. Invest. Med.*, **46**, 113–120.
Narvaiza,I. (2000) *J. Immunol.*, **164**, 3112–3122.
Peveri,P., Walz,A., Dewald,B. and Baggiolini,M. (1998) *J. Exp. Med.*, **167**, 1547–1559.
Quackenbush,E.J., Aguirre,V., Wershil,B.K. and Gutierrez-Ramos,J.C. (1997) *J. Leukoc. Biol.*, **62**, 661–666.
Rojo,D., Suetomi,K. and Navarro,J. (1999) *Biol. Res.*, **32**, 263–272.
Rossi,D. and Zlotnik,A. (2000) *Annu. Rev. Immunol.*, **18**, 217–242.
Ruehlmann,J.M., *et al*. (2001) *Cancer Res.*, **61**, 8498–8503.
Sallusto,F., Lanzavecchia,A. and Mackay,C.R. (1998) *Immunol. Today*, **19**, 568–574.
Sallusto,F., Mackay,C.R. and Lanzavecchia,A. (2000) *Annu. Rev. Immunol.*, **18**, 593–620.
Thelen,M. (2001) *Nat. Immunol.*, **2**, 129–134.
Yoshimura,T., Matsushima,K., Tanaka,S., Robinson,E.A., Appella,E., Oppenheim,J.J. and Leonard,E.J. (1987) *Proc. Natl Acad. Sci. USA*, **84**, 9233–9237.
Zlotnik,A., Morales,J. and Hedrick,J.A. (1999) *Crit. Rev. Immunol.*, **19**, 1–47.

Received September 21, 2008; revised February 23, 2009; accepted May 3, 2009

Edited by Feng Ni