The Pharma Innovation Journal 2021; 10(12): 3014-3021

www.ThePharmaJournal.com

The Pharma Innovation

ISSN (E): 2277- 7695
ISSN (P): 2349-8242
NAAS Rating: 5.23
TPI 2021; 10(12): 3014-3021

www.thepharmajournal.com
Received: 10-09-2021
Accepted: 23-10-2021

K Vignesh
Emlink Technologies Private Limited, Hyderabad, Telangana, India

K Kanagarajadurai
Veterinary University Training and Diagnostic Centre, Tamil Nadu Veterinary and Animal Sciences University, Madurai, Tamil Nadu, India

Corresponding Author:
K Kanagarajadurai
Veterinary University Training and Diagnostic Centre, Tamil Nadu Veterinary and Animal Sciences University, Madurai, Tamil Nadu, India

Classification and functional characterization of chemokines through machine learning approach

K Vignesh and K Kanagarajadurai

Abstract
Chemokines are inflammatory responsible proteins, which mediates varied immune functions like pulling leukocytes towards the inflammatory site, angiogenesis, T-cell differentiation & *etc*. At present, chemokines are classified in to CC, CXC, CX3C and XC (or just C) based on the patterns of their first two cysteine residues at their N-terminal region. Chemokines bind to G-protein coupled receptors (GPCRs). These GPCRs are classified based on the chemokines that binds to that receptor. More than one ligand binds to a single receptor and similarly, a single ligand binds to more than one receptor. This promiscuous nature of the chemokines and their receptors also extends across different classes (for eg. CCL1 ligand binds to CX3CR1 receptor despite their higher affinity to CCR1). These discrepancies maybe attributed to the classification of Chemokine receptors, which are not based on the receptor properties. So, the current study has been designed to classify the chemokines receptors using Support Vector Machine (SVM) based on the receptor properties exclusively.
There were 19 SVM (Support Vector Machine) models of chemokine receptors were generated to predict any protein sequence to be a chemokines or non-chemokine receptor sequence. Despite that it can also identify its receptor classification. The *Relief* and *mRMR* algorithms plays a major role in determining the sensitivity and efficiency of the SVM models. In order to get a better understanding of the SVM output, a phylogenetic tree was constructed using these SVM values. The cluster or a group of receptors based on evolutionary relationship is supported by the work published by other group. The accuracy of the receptor SVM models varies from 83.87% to 100%.
This prediction method of classifying protein sequences by using SVM models, treating each receptor independent of the other and extending it for inferring phylogenetic relationship between them is a novel approach. The achieved accuracy is more, since refinement in accuray was done using Relief and *mRMR* algorithms. Similar approach may be employed to understand the relationship between protein sequences of interest.

Keywords: Chemokine receptors, GPCRs, support vector machine (SVM), machine learning algorithm, evolutionary relationship and phylogenetic tree

1. Introduction
Chemokines are a family of small chemo-attractive proteins that are involved in host defence as regulators of leukocyte trafficking, organogenesis, hematopoiesis and neuronal communication [1]. Sequence identity is usually low among chemokines. All chemokines have a characteristic cysteine motif [2, 3].
Chemokines are classified into 4 classes such as CC, CXC, C (or XC) and CX3C based on the patterns of their first 2 cysteine residues that are located near the N-terminal end [4]. Chemokines are involved in variety of physiological mechanisms other than the endothelial migration and trafficking [5, 6].
Chemokines bind to their receptors, which are G-protein coupled receptors. Unlike chemokines, chemokine receptors share a higher degree of sequence identity both, within the species as well as between species. Chemokine receptors are classified into 4 classes as their ligands depending on the type of ligand that binds to them. More than one ligand binds to a single receptor [7] and similarly, a single ligand binds to more than one receptor. This promiscuous nature of the chemokines and their receptors, also extends across different classes (For example, CCL2 from herpes virus, is capable of binding to CCR4, CCR8, CXCR3, CXCR4, XCR1, and CX3CR1) [8, 9]. These discrepancies maybe attributed to the classification of chemokine receptors which do not follow a proper classification scheme as their ligands.
SVM is a machine learning technique that can perform statistical analyses such as classification or regression on a given dataset [10].

Support Vector Machines can detect hidden patterns in complex datasets and they are independent of the sequence similarity. SVM is robust to outliers, i.e. its sensitivity is not affected by the presence of outliers (data which originally belongs to one class but are classified into the other) [11]. These features of SVM make it an ideal choice for its use in this classification process.

The objective of our work is to provide a new classification scheme for the chemokine receptors using SVM, based on the receptor properties exclusively. The existing classification is based, only on the type of ligand binding to the receptor and not on the receptor properties.

2. Materials and Methods

2.1 Software / Database / Tools Used

SVM^{light}, developed by Joachims T [12] is used to build the SVM models. Samples are classified based on hyperplanes. In case, if there is more than one hyperplane that can classify the positive and negative examples, the hyperplane with maximum width (maximum margin) is chosen as the efficient one [13]. *Blast 2.2.25+* software tool package [14] available from NCBI is an offline version of NCBI BLAST. The query sequences can be searched, by employing the tools in the blast package against a non-redundant database (*nr database)* which is downloadable. *HMMER 2.0* software package [15] was used for building the HMM profiles of each and every chemokine receptor sequence. The HMM profile is built and calibrated using the modules available with this package. *SWISS-PROT* database is used for retrieving non-chemokine GPCPR sequences which are used for constructing the negative dataset. *AA index* database available at "http://www.genome.jp/aaindex/" is used for retrieving values for converting the protein sequences into their vector format. This database has numeric values for each and every amino acid for 544 properties [16]. These numeric values are used to convert each protein to a 544 dimensional vector. The *mRMR* (minimal Redundance Maximal Relevance) algorithm available at "penglab.janelia.org/proj/mRMR/" is used for eliminating redundant features. It requires more time and memory if used on a large dataset. So, *Relief* algorithm is applied first, to reduce the dimensionality of the dataset and then mRMR algorithm is employed [17]. *PHYLIP* (PHYlogeny Inference Package) is a free computational phylogenetics package of programs for inferring evolutionary relationships [18].

2.2 Dataset

2.2.1 Test set

19 Human chemokine receptor sequences are taken as the test set (i.e. these 19 human chemokine receptor sequences are classified in this process). These receptor sequences are retrieved from SWISS-PROT database.

2.2.2 Positive Dataset

2 different types of datasets are used for training and validating the SVM models. Positive dataset consists of putative as well as annotated chemokine receptor sequences from various species that are related to human chemokine receptors sequences.

2.2.3 Negative Dataset

Negative dataset is created using non-chemokine sequences that are retrieved from SWISS-PROT. An Equal number of sequences are taken in both positive and negative datasets for simplicity in dividing these sequences into training and validation sets (ST 2).

2.3 Methodology

2.3.1 Creating the Positive and negative dataset

The positive dataset is created by combining true as well as hypothetical proteins from the PSI-BLAST output of all 19 receptor sequences. An *hmmsearch* is done using the HMM profile against this dataset and sequences having E-value better than (less in magnitude) E^{-100} are taken for the creation of positive dataset. Mammalian protein sequences, not belonging to the chemokine family and ranging from 350-370 amino acids in length are retrieved from SWISS-PROT. It includes both non-chemokine GPCR and non-GPCR sequences. The *hmmsearch* is employed using the HMM profile and an equal number of sequences are chosen for the creation of negative dataset. After generating 19 datasets corresponding to 19 human chemokine receptor sequences, SVM models are created for each receptor.

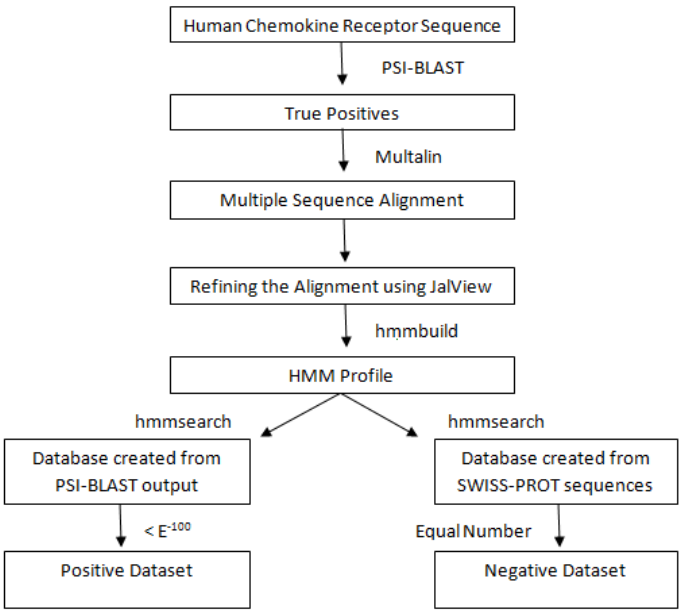

Fig 1: Methodology flowchart for creating positive and negative dataset.

2.3.2 Training and Validation Set

SVM has been trained before it can be used for classifying the sequences. The datasets created by the above said process are further refined in such a way that the positive dataset for each receptor has minimum 70% of true positives (For instance, the positive dataset for CCR1 should have a minimum of 70% CCR1 sequences and the rest can be other chemokine receptor sequences). This rule is imposed to avoid biasing. CCR5 and CXCR4 sequences have extensively been studied and hence, they are available in a large amount compared to other sequences. So, the dataset of other sequences are prone to pick up these 2 sequences in a greater amount because of their availability. This biasing results in a decrease in the sensitivity and accuracy of the SVM model. To overcome the biasness, the above said criteria have been imposed.

The positive dataset was then divided in a 3:1 ratio, i.e. the training set was 3 times the number of sequences as the validation set. The 70% condition was maintained in the training and validation set also.

2.3.3 Conversion of protein sequence into vector Format

AA index is a database that has numeric values for each of the twenty amino acids for 544 physicochemical properties. Protein sequences are thus converted into 544 dimensional vectors based on these 544 properties. The conversion is done according to the below mentioned formula.

$$\text{Value} = \frac{\text{Sum of values of each amino acid in the sequence for a particular property}}{\text{Sequence length}}$$

2.3.5 Building the SVM Model

The sequences that are now represented as 544 dimensional vectors are given as input for the SVM model. The accuracy of each model after training them with all 544 features is tabulated. The accuracy of the models is examined and some of the properties are eliminated to improve the accuracy employing the feature selection algorithms Relief and mRMR.

3. Results and Discussion

The total of 19 chemokikne receptors sequences were considered for this study and the sequences were extracted from Swiss-Prot database (Supplementary Table 1). The total number of sequences from PSI-BLAST of 19 chemokine receptors is 74000 among which number of non-redundant sequences is 1924. Thus, 1924 non-redundant sequences were taken to create the positive dataset. A dataset for each of the receptor was constructed with the equal number of positive and negative sequences (Supplementary Table 2). The number of training and validation set for each receptor from the said database were grouped as discussed in methodology (Supplementary Table 3).

The SVM models were built and accuracy of the model was examined as discussed in the methodology (Table 1). Some of the models were showing poor accuracy due to the fact that not all features are necessary for the classification process (Supplementary Table 4). As shown in ST4, the values of the properties *Chemical shift* and *Average flexibility indices* do not vary much across various proteins as compared to the properties *Hydrophobicity* and *Membrane Buried Preference*. Hence, the former was not as effective in classifying the receptor as compared to the latter. These properties can be eliminated to improve the accuracy. Feature selection algorithms Relief and mRMR were employed to eliminate these nonessential features and obtain the final SVM models (Table 1), which were then used to classify the test set. The test set consists of the 19 human chemokine receptors that were retrieved from SWISS-PROT. They were converted into vectors based on the same features which were used for building that particular SVM model and then classified using SVM model.

Table 1: Accuracy SVM model: Accuracy of each SVM model after training with 544 properties. Improved SVM model after refinement with feature selection algorithms Relief and mRMR.

SVM Model	Accuracy (%) (Training with 544 properties)	Accuracy (%) (After refinement with feature algorithms)
CCR1	70	88
CCR2	51.61	90.32
CCR3	48.15	100
CCR4	47.62	85.71
CCR5	64.62	92.45
CCR6	47.62	90.48
CCR7	57.69	100
CCR8	47.83	91.30
CCR9	62.50	91.67
CCR10	46.67	100
CXCR1	48.15	85.19
CXCR2	48	88.00
CXCR3	50	93.75
CXCR4	72.18	96.37
CXCR5	61.79	100
CXCR6	59.09	100
CXCR7	54.55	95.45
XCR1	47.06	100
CX3CR1	51.61	83.87

Supplementary Table 1: Chemokine Receptors: Human chemokine receptors and their SWISS-PROT accession IDs

Type	Receptor	Swiss-prot id
CC	CCR1	P32246
	CCR2	P41597
	CCR3	P51677
	CCR4	P51679
	CCR5	P51681
	CCR6	P51684
	CCR7	P32248
	CCR8	P51685
	CCR9	P51686
	CCR10	P46092
CXC	CXCR1	P25024
	CXCR2	P25025
	CXCR3	P49682
	CXCR4	P61073
	CXCR5	P32302
	CXCR6	O00574
	CXCR7	P25106
CX3C	CX3CR1	P49238
XC	XCR1	P46094

Supplementary Table 2: Chemokine Receptors –Dataset: Number of sequences in the positive and final (including positive and negative datasets) dataset for each receptor

Receptor	No. of sequences in the positive dataset	Total number of sequences (including positive and negative dataset)	Receptor	No. of sequences in the positive dataset	Total number of sequences (including positive and negative dataset)
CCR1	27	54	CXCR1	54	108
CCR2	61	122	CXCR2	50	100
CCR3	54	108	CXCR3	32	64
CCR4	41	82	CXCR4	496	992
CCR5	423	846	CXCR5	34	68
CCR6	41	82	CXCR6	41	82
CCR7	51	102	CXCR7	43	86
CCR8	52	104	XCR1	21	42
CCR9	48	96	CX3CR1	30	60
CCR10	29	58			

Supplementary Table 3: SVM - Training and Validation Set: Number of sequences in the training and validation set for each receptor

Receptor	Training Set	Validation Set	Receptor	Training Set	Validation Set
CCR1	40	14	CXCR1	81	27
CCR2	91	31	CXCR2	75	25
CCR3	81	27	CXCR3	48	16
CCR4	61	21	CXCR4	744	248
CCR5	634	212	CXCR5	51	17
CCR6	61	21	CXCR6	61	21
CCR7	76	26	CXCR7	64	22
CCR8	78	26	XCR1	31	11
CCR9	72	24	CX3CR1	45	15
CCR10	43	15			

Supplementary Table 4: Feature Vectors: Values of a few receptor sequences for a few properties

Protein	Chemical Shift	Hydrophobicity	Membrane Buried preference	Average flexibility indices
CCR1	4.34	1.11	1.27	0.42
CCR2	4.33	1.05	1.22	0.43
CXCR1	4.35	1.04	1.32	0.42
CXCR2	4.34	1.06	1.33	0.42
CX3CR1	4.36	1.06	1.27	0.42
XCR1	4.36	1.09	1.31	0.42

3.1 Observations

The values obtained by classifying the receptors by various models are given in Supplymentary Figure 1. Each row represents the SVM model that has been used and each column represents the receptor that is classified. The values for each receptor denote the class to which it is classified. The sign denotes the class to which it is classified and the magnitude denotes the distance between the hyperplane and the vector that has been classified. A threshold value of 0.50 was set and thus, receptors with values greater than 0.50 are considered to be classified by each SVM model as belonging to its own class. Supplementary table 5 represents a list of receptor models and the receptors that are classified into their own class.

Supplementary Table 5: Receptor Classification based on SVM: A list of receptor models and the receptors that are classified into its own class

Model	Receptors classified as belonging to its own class (arranged in the order of decreasing SVM values)
CCR1	CCR1, CCR3,
CCR2	CCR2, CCR3, CCR5
CCR3	CCR3, CCR1, CX3CR1
CCR4	CCR4
CCR5	CCR5, CCR3, CCR1, CCR8
CCR6	CCR6
CCR7	CXCR6, CCR6, CCR7, CCR9, CCR8, CX3CR1
CCR8	CCR8, CCR1, CX3CR1, CCR7, CCR3, CCR4
CCR9	CCR9, CXCR6
CCR10	CCR10
CX3CR1	CX3CR1, CCR5
CXCR1	CXCR1, CXCR2
CXCR2	CXCR2, CXCR1
CXCR3	CXCR3, CCR10
CXCR4	CXCR4, CCR8
CXCR5	CXCR5, CXCR2

CXCR6	CXCR6, CCR7, CCR8, CCR5
CXCR7	CXCR7, CCR7, CCR4
XCR1	XCR1, CCR1, CCR3, CCR8, CCR5

3.2 Analysis

The accuracy of the receptor SVM models vary from 83.87% to 100%. This classification may have included few false positives as well as few false negatives. For example, the SVM models for CCR3, CCR7, CCR8 classified CX3CR1 as belonging to their own class, however the SVM model for CX3CR1 classified only CCR5 and CX3CR1 as belonging to its own class. Also, the SVM model for CCR1 classified CCR3 with a better value than CCR1. These discrepancies may be due to two reasons: 1) differences in the number of sequences in the training set, 2) variation in the selection of features and the differences in the number of features for each class.

3.3 Inferring Evolutionary Relationship

In order to get a better understanding of the SVM output, a phylogenetic tree was constructed using these SVM values. This is a novel methodology, which convert the SVM values i.e. the sequence properties in to normalised values and compare the closeness between any two given receptors and considered as a distance between them. Hence, the matrix built using this SVM values was considered as distance matrix. The matrix displayed above was normalised in such a way that each diagonal element had a value of 1 (*i.e.* the SVM model for a particular receptor classified its own receptor with a value of 1). Any diagonal value greater than 1 was reduced to 1 and any value less than 1 was incremented to 1. This normalisation process was done to each row separately based on the diagonal element values.

The values obtained after normalisation are shown in Supplementary Figure 2. A distance matrix was then constructed using these normalised values. The distance between any 2 receptors is taken as the average of the values between them. Thus, the distance between CCR1 and CCR2 would be (-0.96 + 0.77)/2. This process was repeated for all receptors and the final value was subtracted from 1 (Figure 2). Thus, for example

The final distance between CCR1 and CCR2 = 1 – ((-0.96 + 0.77)/2) = 1.1

	CCR1	CCR2	CCR3	CCR4	CCR5	CCR6	CCR7	CCR8	CCR9	CCR10	CX3CR1	CXCR1	CXCR2	CXCR3	CXCR4	CXCR5	CXCR6	CXCR7	CR1
CCR1	0																		
CCR2	1.1	0																	
CCR3	0	0.95	0																
CCR4	1.65	0.85	1.18	0															
CCR5	1.09	0.89	0.83	1.07	0														
CCR6	2.16	1.77	1.87	1.55	1.93	0													
CCR7	1.45	1.66	0.76	0.85	1.3	0.42	0												
CCR8	0.99	1.02	0.75	0.6	1.14	1.68	0.2	0											
CCR9	1.76	1.69	1.32	1.1	1.38	1.21	0.41	1.08	0										
CCR10	1.69	1.73	1.77	1.8	2.04	1.5	1.67	1.74	1.64	0									
CX3CR1	1.28	1.61	0.84	1.02	0.89	1.43	0.86	0.7	1.18	1.71	0								
CXCR1	1.73	1.57	1.41	1.49	1.56	1.73	1.18	1.76	1.28	1.79	1.06	0							
CXCR2	1.05	1.48	1.4	1.22	1.44	1.53	1.33	1.13	1.23	1.59	1.17	0.32	0						
CXCR3	1.71	1.73	1.69	1.69	2.23	1.85	1.35	1.8	1.75	0.61	1.31	1.71	1.2	0					
CXCR4	1.18	1.16	1.46	1.07	1.76	1.39	1.14	0.91	1.48	1.57	1.51	1.81	1.59	1.5	0				
CXCR5	1.01	1.21	1.33	1.32	1.83	1.4	1	1.76	1.42	1.52	1.23	0.95	0.63	1.15	1.79	0			
CXCR6	1.35	1.3	0.83	0.76	0.63	1.6	0.1	0.36	0.65	1.77	0.98	1.23	1.44	1.97	1.51	1.51	0		
CXCR7	1.96	1.45	1.1	0.69	1.52	1.92	0.72	1.08	0.99	1.98	0.94	1.2	1.68	2.3	1.72	1.63	1.17	0	
CR1	0.85	0.86	0.61	0.96	0.63	1.71	1.53	0.64	1.68	1.68	1.69	1.53	1.64	1.75	1.33	1.71	0.6	1.59	0

Fig 2: Distance Matrix: The values are colour coded. Blue: < 0.5; Green: 0.5 – 1; Red: 1 – 1.5; Black: >1.5;

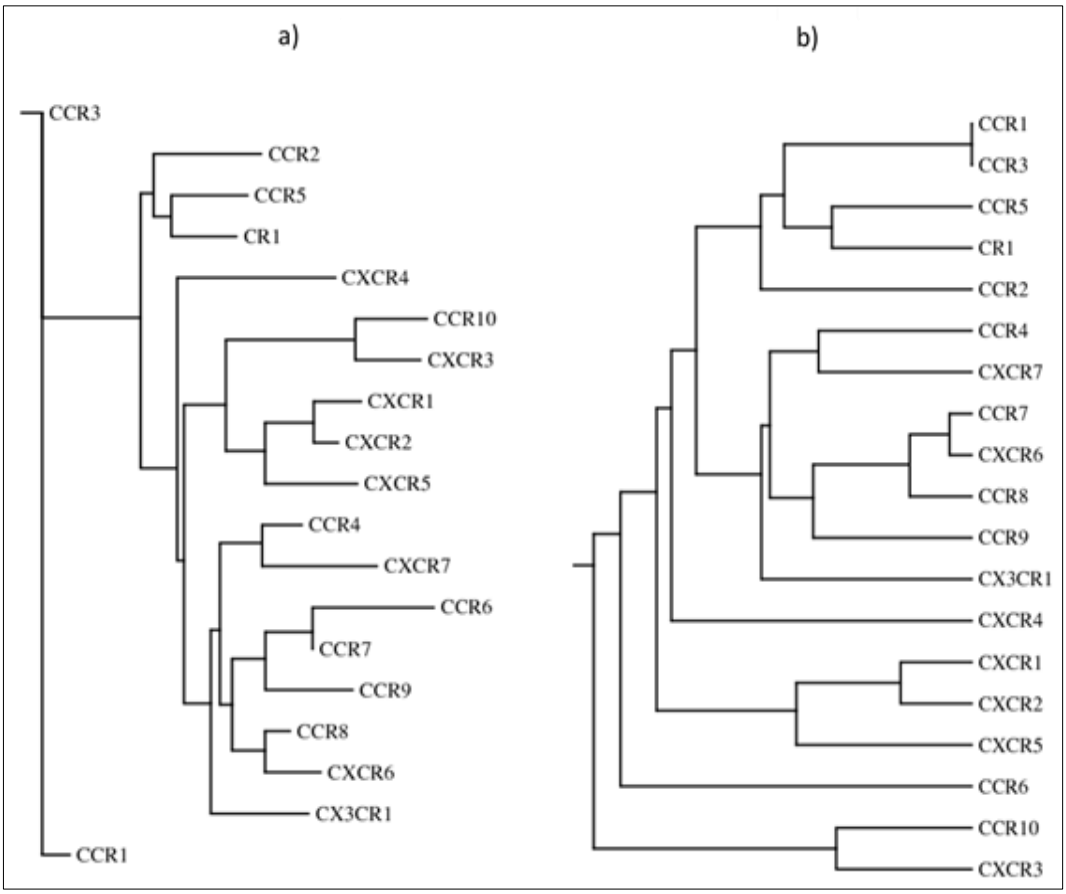

Fig 3: Phylogenetic tree. a) Tree constructed using Neighbour-Joining method; b) Tree constructed using UPGMA method

	CCR1	CCR2	CCR3	CCR4	CCR5	CCR6	CCR7	CCR8	CCR9	CCR10	CX3CR1	CXCR1	CXCR2	CXCR3	CXCR4	CXCR5	CXCR6	CXCR7	CR1
CCR1	0.98	-0.98	1.07	-1.08	-0.95	-1.37	-1.03	-0.96	-1.3	-0.22	-0.23	-0.87	0.07	-0.71	0.18	0.32	-1.1	-1.26	-0.72
CCR2	0.47	0.7	0.65	0.24	0.58	-0.92	-0.28	0.36	-0.63	-1.02	-0.7	-0.85	-0.98	-1.25	0.47	0.19	0.17	-0.26	-0.09
CCR3	1.01	-0.84	1.01	-0.1	-0.62	-0.82	0.4	0	-0.32	-0.21	0.64	-0.2	-0.6	-0.36	-0.43	-0.17	-0.07	0.48	-0.21
CCR4	-0.23	-0.24	-0.24	1	-0.28	-0.31	0.5	0.29	-0.24	-0.24	0.22	-0.38	-0.25	-0.31	-0.25	-0.22	0.09	0.12	-0.24
CCR5	0.98	-0.44	1.2	0.37	1.23	-0.88	-0.71	0.94	-0.73	-0.64	-0.32	-0.29	-0.49	-0.88	-0.37	-0.7	0.27	-0.2	0.32
CCR6	-0.96	-0.92	-0.9	-0.79	-0.75	1	0.15	-0.74	0.49	-0.56	-0.23	-0.9	-0.87	-0.96	0.07	-0.12	-0.7	-0.79	-0.86
CCR7	0.14	-1.31	0.11	-0.19	0.36	1.13	1.02	0.81	1.02	-0.3	0.59	-0.11	-0.45	-0.33	0.03	0.18	1.25	-0.17	-0.04
CCR8	0.96	-0.7	0.51	0.51	-0.98	-0.62	0.81	1	-0.12	-0.12	0.89	-0.92	-0.07	-0.05	-0.35	-1.03	0.49	0.16	-0.07
CCR9	-0.18	-0.99	-0.24	0.11	0.27	-0.84	0.24	0.03	1.06	-0.24	-0.16	-0.16	-0.21	-0.39	-0.62	-0.26	0.67	-0.18	-0.24
CCR10	-1.1	-0.66	-1.24	-1.28	-1.14	-0.37	-0.95	-1.28	-0.9	1.07	-1.03	-0.91	-0.92	-0.16	-0.48	-0.46	-1.17	-1.21	-1.42
CX3CR1	-0.35	-0.81	-0.31	-0.26	0.76	-0.62	-0.29	-0.29	-0.13	-0.31	1	-0.12	-0.26	-0.31	-0.33	-0.38	0.06	-0.01	-0.29
CXCR1	-0.41	-0.4	-0.41	-0.4	-0.41	-0.37	-0.04	-0.41	-0.14	-0.41	0.19	1.19	0.87	-0.53	-0.39	-0	-0.33	0.11	-0.41
CXCR2	-0.12	-0.22	-0.12	-0.12	-0.09	-0.12	-0.12	-0.12	-0.12	-0.12	-0.01	0.74	1.06	-0.26	-0.1	-0.09	-0.11	-0.08	-0.12
CXCR3	-0.73	-0.52	-1.01	-1.07	-1.36	-0.74	-0.36	-1.55	-1.06	1.32	-0.31	-0.7	-0.08	0.99	-0.19	-0.24	-1.31	-1.57	-1.64
CXCR4	-0.52	-1.05	-0.45	0.15	-0.88	-0.82	-0.26	0.56	-0.25	-0.55	-0.66	-1	-0.98	-0.78	1.03	-0.93	-0.19	-0.45	-0.71
CXCR5	-0.37	-0.92	-0.48	-0.43	-0.74	-0.68	-0.17	-0.5	-0.52	-0.51	-0.08	0.28	0.87	-0.08	-0.62	0.99	-0.51	-0.18	-0.51
CXCR6	0.44	-1.01	0.47	0.44	0.75	-0.45	0.84	0.84	0.14	-0.24	0.02	0.12	-0.65	-0.59	-0.75	-0.47	1.05	-0.12	0.39
CXCR7	-0.49	-0.75	-0.49	0.67	-0.43	-0.87	0.92	-0.14	0.44	-0.49	0.3	-0.13	-1.04	-0.86	-0.78	-0.9	0.02	1.18	-0.49
CR1	1.01	0.08	1.01	0.32	0.65	-0.54	-0.99	0.79	-1.05	0.15	-1.07	-0.44	-1.08	0.15	0.09	-0.91	0.46	-0.5	1.01

Supplementary Fig 1: SVM Values: Values obtained by using the SVM models for classifying the receptor sequences. The values are colour coded. Blue (>0.5), Green (0 to 0.5), Black (-0.5 to 0) and Red (<-0.5)

	CCR1	CCR2	CCR3	CCR4	CCR5	CCR6	CCR7	CCR8	CCR9	CCR10	CX3CR1	CXCR1	CXCR2	CXCR3	CXCR4	CXCR5	CXCR6	CXCR7	CR1
CCR1	1	-0.96	1	-1.06	-0.93	-1.35	-1.01	-0.94	-1.28	-0.2	-0.21	-0.85	0.09	-0.69	0.2	0.34	-1.08	-1.24	-0.7
CCR2	0.77	1	0.95	0.54	0.88	-0.62	0.02	0.66	-0.33	-0.72	-0.4	-0.55	-0.68	-0.95	0.77	0.49	0.47	0.04	0.21
CCR3	1	-0.85	1	-0.11	-0.63	-0.83	0.39	-0.01	-0.33	-0.22	0.63	-0.21	-0.61	-0.37	-0.44	-0.18	-0.08	0.47	-0.22
CCR4	-0.23	-0.24	-0.24	1	-0.28	-0.31	0.5	0.29	-0.24	-0.24	0.22	-0.38	-0.25	-0.31	-0.25	-0.22	0.09	0.12	-0.24
CCR5	0.75	-0.67	0.97	0.14	1	-1.11	-0.94	0.71	-0.96	-0.87	-0.55	-0.52	-0.72	-1.11	-0.6	-0.93	0.04	-0.43	0.09
CCR6	-0.96	-0.92	-0.9	-0.79	-0.75	1	0.15	-0.74	0.49	-0.56	-0.23	-0.9	-0.87	-0.96	0.07	-0.12	-0.7	-0.79	-0.86
CCR7	0.12	-1.33	0.09	-0.21	0.34	1	1	0.79	1	-0.32	0.57	-0.13	-0.47	-0.35	0.01	0.16	1	-0.19	-0.06
CCR8	0.96	-0.7	0.51	0.51	-0.98	-0.62	0.81	1	-0.12	-0.12	0.89	-0.92	-0.07	-0.05	-0.35	-1.03	0.49	0.16	-0.07
CCR9	-0.24	-1.05	-0.3	0.05	0.21	-0.9	0.18	-0.03	1	-0.3	-0.22	-0.22	-0.27	-0.45	-0.68	-0.32	0.61	-0.24	-0.3
CCR10	-1.17	-0.73	-1.31	-1.35	-1.21	-0.44	-1.02	-1.35	-0.97	1	-1.1	-0.98	-0.99	-0.23	-0.55	-0.53	-1.24	-1.28	-1.49
CX3CR1	-0.35	-0.81	-0.31	-0.26	0.76	-0.62	-0.29	-0.29	-0.13	-0.31	1	-0.12	-0.26	-0.31	-0.33	-0.38	0.06	-0.01	-0.29
CXCR1	-0.6	-0.59	-0.6	-0.59	-0.6	-0.56	-0.23	-0.6	-0.33	-0.6	0	1	0.68	-0.72	-0.58	-0.19	-0.52	-0.08	-0.6
CXCR2	-0.18	-0.28	-0.18	-0.18	-0.15	-0.18	-0.18	-0.18	-0.18	-0.18	-0.07	0.68	1	-0.32	-0.16	-0.15	-0.17	-0.14	-0.18
CXCR3	-0.72	-0.51	-1	-1.06	-1.35	-0.73	-0.35	-1.54	-1.05	1	-0.3	-0.69	-0.07	1	-0.18	-0.23	-1.3	-1.56	-1.63
CXCR4	-0.55	-1.08	-0.48	0.12	-0.91	-0.85	-0.29	0.53	-0.28	-0.58	-0.69	-1.03	-1.01	-0.81	1	-0.96	-0.22	-0.48	-0.74
CXCR5	-0.36	-0.91	-0.47	-0.42	-0.73	-0.67	-0.16	-0.49	-0.51	-0.5	-0.07	0.29	0.88	-0.07	-0.61	1	-0.5	-0.17	-0.5
CXCR6	0.39	-1.06	0.42	0.39	0.7	-0.5	0.79	0.79	0.09	-0.29	-0.03	0.07	-0.7	-0.64	-0.8	-0.52	1	-0.17	0.34
CXCR7	-0.67	-0.93	-0.67	0.49	-0.61	-1.05	0.74	-0.32	0.26	-0.67	0.12	-0.31	-1.22	-1.04	-0.96	-1.08	-0.16	1	-0.67
CR1	1	0.07	1	0.31	0.64	-0.55	-1	0.78	-1.06	0.14	-1.08	-0.45	-1.09	0.14	0.08	-0.92	0.45	-0.51	1

Supplementary Fig 2: Normalised Matrix: Values after normalisation. The values are colour coded. Blue (>0.5), Green (0 to 0.5), Black (-0.5 to 0) and Red (<-0.5).

Phylogenetic trees were constructed from this distance matrix by both UPGMA and Neighbour Joining (NJ) methods. Neighbour module of the *PHYLIP* software was used for constructing the tree via UPGMA and Neighbour Joining method. Figure 3 represent the phylogenetic trees constructed by NJ and UPGMA methods, respectively.

There were 5 major clusters found both in NJ and UPGMA methods (Table 2). However, CXCR4 was not clustered with any other receptor in both the trees. But, CCR6 and CX3CR1 were clustered differently when using these methods. CCR6 was clustered with CCR7, CCR8, CCR9 and CXCR6 when using NJ method and CX3CR1 was far away from this cluster. But when using UPGMA method, CCR6 was far away from this cluster but CX3CR1 was nearer to this cluster. The distance matrix shows that the shortest distance for CX3CR1 is with CCR8 (0.7) and for CCR6 is with CCR7 (0.42). Thus, the lack of consistency between both these trees with respect to these 2 receptors might be due to either 1) the difference in the algorithm of NJ and UPGMA methods or 2) the differences in the number of sequences and the features considered for each set may vary greatly for training the dataset.

Table 2: Evolutionary Tree clusters: The group members in clusters obtained between Neighbour-Joining and UPGMA methods

Cluster/ Method	Neighbour Joining method	UPGMA method
Cluster 1	CXCR7, CCR4	CXCR7, CCR4
Cluster 2	CXCR6, CCR7, CCR6, CCR9, CCR8	CCR7, CXCR6, CCR8, CCR9
Cluster 3	XCR1, CCR5, CCR2, CCR3, CCR1	CCR1, CCR3, CCR5, XCR1, CCR2
Cluster 4	CXCR2, CXCR1, CXCR5	CXCR1, CXCR2, CXCR5
Cluster 5	CCR10, CXCR3	CCR10, CXCR3

The difference in the method of tree construction between these two methods can account for the difference in topology of the tree with respect to the receptors CX3CR1 and CCR6 as the sum of the distances of CCR6 with other receptors is 28.65 and that of CX3CR1 is 21.41. But according to UPGMA method, the distance between CCR6 and CCR7 is 0.42 which is less than the distance between CX3CR1 and CCR8, which is 0.7. Hence we include these receptors into cluster 2 based on their distance with the members of this cluster in the distance matrix.

4. Conclusion

The human chemokine receptors classified by this process shows that members from different classes are clustered together (for eg. CCR4 and CXCR7). This may be contrary to the existing classification due to the fact that the current study considered the receptor properties exclusively and doesn't take into the account of their interactions with their cognate ligand(s). The cluster or group of receptors based on evolutionary relationship are supported by the work published by *Wang J., et al.*, (2005), despite few changes noticed as differences. Wang *et al*., proposed the relationship between the chemokine receptors, which was based only on distance between the sequences. In their work, they have shown CCR1, CCR3, CCR5, CCR2, CCR4, CCR8, CX3CR1, and XCR1 as one cluster where as CCR8 and CX3CR1 are linked

to this cluster through CCR4. Since our work involves CXCR7 which was not considered in Wang *et al*., work and CCR4 is clustered with CXCR7, the receptors CCR8 and CX3CR1 have got clustered with CXCR6 etc. It is interesting to notice that they are all present in the same Chromosome (Chromosome 3). Cluster 1 involving CCR4 and CXCR7. Cluster 2 involving CCR7, CXCR6, CCR8, CCR9, CX3CR1 and CCR6.Cluster 3 involving CCR1, CCR3, CCR5, XCR1 and CCR2.Cluster 4 involving CXCR1, CXCR2 and CXCR5.Cluster 5 involving CCR10 and CXCR3. The receptor CXCR4 is not clustered with any other receptor.

This method of classifying protein sequences by using SVM models, treating each receptor independent of the other and extending it for inferring phylogenetic relationship between them is a novel approach. Although the SVM models are statistically validated based on their accuracy and hence their specificity and sensitivity, the phylogenetic trees still need to be statistically validated. Apart from statistical validation it also needs to be validated biologically, which require lot of information.

The clustering of CCR4 and CXCR7 needs to be verified by experimental methods. Thus, in the absence of structural information, this method of classification using Support Vector Machines can be used for classification process. Better methods for normalising the output values from the SVM models for constructing the distance matrix can lead to better and reliable phylogenetic trees.

5. Acknowledgement

Both the authors thank the Centre of Excellence in Bioinformatics at School of Biotechnology, Madurai Kamaraj University, Madurai, and Tamil Nadu for the Computational facilities to carry out the work.

6. References

1. Olson TS, *et al*. Chemokines and chemokine receptors in leukocyte trafficking. Am J Physiol Regul Integr Comp Physiol, 2002;283:R7-R28.
2. Onuffer JJ, *et al*. Chemokines, chemokine receptors and small-molecule antagonists: recent developments. Trends Pharmacol Sci, 2002;23:459-467.
3. Zlotnik A, *et al*. Chemokines: a new classification system and their role in immunity. Immunity, 2000;12:121-127.
4. Murphy PM. Chemokines: In Fundamental Immunology 5th edition. Edited by: William EP. New York: Lippincott Williams & Wilkins, 2003, 801-840.
5. Nomiyama H, *et al*. Organization of the chemokine genes in the human and mouse major clusters of CC and CXC chemokines: diversification between the two species. Genes and Immunity. 2001;2:110-113.
6. Devora Rossi, *et al*. The Biology of Chemokines and their receptors. Annual Reviews Immunology. 2000;18:217-242.
7. Murdoch C, *et al*. Chemokine receptors and their role in inflammation and infectious diseases. Blood. 2000;10:3032-43.
8. Karuppiah Kanagarajadurai and Ramanathan Sowdhamini: Sequence and structural analyses of interleukin-8-like chemokine superfamily. In Silico Biology. 2008;8:0025.
9. Chien EY, *et al*. Structures of the CXCR4 chemokine GPCR with small-molecule and cyclic peptide antagonists. Science, 2010;330:1066-1071.
10. Joachims T. 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. 1999.
11. Christopher JC. Burges: A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery, 1998;2:121–167.
12. Joachims T. Learning to classify text using support vector machines. Boston: Kluwer Academic Publishers. 2002.
13. Hwanjo Yu, Sungchul Kim. SVM Tutorial: Classification, Regression, and Ranking. www.people.sabanciuniv.edu/berrin/cs512/lectures/11-svmtutorial2.pdf
14. Camacho C, *et al*. BLAST+: architecture and applications. BMC Bioinformatics, 2009;10:421.
15. Eddy S. HMMER User's Guide. www.selab.janelia.org. 1992.
16. Shuichi Kawashima, *et al*. AAindex: Amino Acid Index Database. Nucleic Acids Research, 1999;27(1).
17. Yi Zhang, *et al*. Gene selection algorithm by combining reliefF and mRMR. BMC Genomics. 2008;9(Suppl 2):S27.
18. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of Molecular Evolution. 1981;17(6):368-376.
19. Dereeper A, *et al*. BLAST-EXPLORER helps you building datasets for phylogenetic analysis. BMC Evolutionary Biology. 2010;12;10:8.