The Pharma Innovation Journal 2021; SP-10(12): 1904-1912

www.ThePharmaJournal.com

The Pharma Innovation

ISSN (E): 2277- 7695
ISSN (P): 2349-8242
NAAS Rating: 5.23
TPI 2021; SP-10(12): 1904-1912

www.thepharmajournal.com
Received: 18-10-2021
Accepted: 20-11-2021

Saravana Prasanna
Centre of Excellence in Bioinformatics, School of Biotechnology, Madurai Kamaraj University, Madurai, Tamil Nadu, 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

Computational approaches to probe amino acid preference at secondary structure of chemokines

Saravana Prasanna and K Kanagarajadurai

Abstract
Chemokines are small and approximately ~8-12 kDa in weight. They are classified into 4 main types with respect to the spacing of the first two cysteine residues at N-terminal region: a) CXC or alpha, b) CC or β, c) C or gamma, and d) CX3C or delta chemokines. The major function of chemokines is considered to be the ability to mediate the migration of leukocytes. However, it is interesting to note that each chemokine tends to mediate various cell types. A known fact is that similar functional proteins tend to have similar structures and intern similar sequences. Chemokines are varying in their sequence identity between 20 and 99 despite that they have similar structure and function. The variation in sequence identity gives a clue about their sub functional variation under the major function, i.e. chemokines mediate various leukocytes under. However, how chemokines manage to maintain the 2D and 3D structural similarity in proteins. It brings an interest to study the sequence-structure-function relationship in terms of their amino acid preference. Hence, the objective of this study is to analyze the structurally similar chemokine proteins to understand the amino acid preference at their secondary structures.
In this study, the chemokines which are having 3D structures are aligned and the structure guided sequence alignment is extracted. Structural properties such as, solvent accessibility, hydrophobicity, retention coefficient, multiple contact index and dihedral angle for each residue in the secondary structures are calculated and their patterns are plotted.
Results of this study reveal that it is possible to explain how different residues tend to form similar secondary structures. The residues that are involved in forming the β-strand at the core region are different from the residues that are forming β-strand at the surface. These results indicate that the preference of amino acid is varying even in the same secondary structural elements depends on the location of the secondary structure present, i.e. whether they present in the core or surface of the protein. This is a novel algorithm, which may be applied to predict the existing or novel protein fold.

Keywords: chemokines, amino acid preference, amino acid alphabet (AAA) propensity, physico-chemical properties, protein structure analysis, solvent accessibility

Introduction
Chemokines are small group of proteins weighing approximately ~8-12 KDa. They have been classified in to four groups based on the presence of four cysteines at N-terminal region of the protein, CC, CXC, XC and CX_3C. C-type (XC) chemokines lack first and third cysteins, but maintain the fold and function. Chemokines are reported to perform immune functions like leukocyte trafficking [1], T cell differentiation, angiogenesis, hematopeosis [2] and mast cell degranulation [3]. Chemokines are also seen to play a role as HIV-1 inhibitor [4] and act as potential adjuvant in the immunotherapy [5]. However, chemokines are known for its leukocyte trafficking, which is considered as the major function. Chemokine receptors belong to the largest subfamily of class-A GPCRs [6].
The structural integrity of chemokines is maintained at their three dimensional space by the conserved cysteine residues. The first chemokine to be characterized was Interleukin 8 (CXCL8), which was more than two decades ago. There are few chemokines including CXCL8 that possess Glu, Leu and Arg residues at their N-terminal region, hence; they are called ELR+ chemokines, which are reported to have angiogenic property. Other chemokines, which lakes this ELR motif is called ELR- chemokines and are reported to be angiostatic in nature. The N-terminal region is reported to be involved in binding with its corresponding G-protein coupled receptor. The biological form of the chemokines may vary, i.e. they may either monomer or dimer. The dimerization pattern of chemokines is reported to be different for each chemokine to perform its biological function. The monomeric and dimeric forms of CXCL8 are reported to induce different responses by interacting with the same receptor[7]. The CXCL8 is an example for α+β protein with three anti-parallel β strands connected by loops and a terminal alpha helix.

CXCL8 forms a globular dimer when seen in solution consisting of six anti parallel β strands (3 from each) and two alpha helices seen across the β sheet. The dimer formation with in each subclass may vary. For example, CXC type subfamily forms a dimer by interaction between β_1 strands, whereas CC type forms a dimer by interaction between short N-terminal β strands.

Due to the high similarity at secondary and tertiary structure of chemokines[8] despite their poor sequence identity, the main objective of the work is to study the molecular properties of amino acid using computational approaches to analyze the preference of amino acids at their secondary structure. Since, chemokines are having one alpha helix and one 3_{10} helix, this study has been focused to find out the amino acid preference at the β strands.

Materials and Methods

Software/Database/Tools

The Structural Classification of Protein (SCOP) database[9] has been used to find the information on protein domains, which share the same structural conformations at their family and superfamily level using evolutionary structure-function relationship. In SCOP, domains are hierarchically classified into species, proteins, families, superfamilies, folds, and classes.

Dali server [10] compares protein in its 3D form for a given query protein. Dali database has been used to find the pre-computed structural neighbourhood. A set of superimposed structures is viewed through Jmol [11]. scratch server[12] has been used to predict secondary structure, solvent accessibility, disordered regions or domains, disulphide bridges, single mutation stability, residue contacts versus average, individual residue contacts and tertiary structure information. ACCpro has been used to predict the relative solvent accessibility of protein residues [12, 13]. either as buried or exposed with respect to a specified threshold between 0% and 95% at steps of 5%. A Multiple Structural Alignment Algorithm (MUSTANG) [13] has been considered as an effective tool to align distantly related proteins, which has been used to align chemokines as they are highly distantly related in terms of their sequence identity.

JOY [14] has been used to annotate the protein sequence alignments with 3D structural features, which helps to understand the conservation of amino acids in their specific local environments. JOY has been used to get the output files of alignment file, hydrogen bond data, secondary structure data, solvent accessibility data and template representation of data. SCWRL has been used for adding missing side chain of a residue in a protein structure. SCWRL [15] uses a backbone-dependent rotamer library based on the kernel density estimates, average over a number of sample conformations about the positions in the library. ProtScale [16] has been used to compute and display the values by the amino acid scale on a query protein. The amino acid scale is the numerical value which is assigned for every amino acid in the table. The most common scale used is the hydrophobicity scale which is derived from experiments of polar and nonpolar solvents with aims to predict the membrane spanning regions which are more hydrophobic and the secondary structure conformation scales.

PROFcon [17] is a server, which predicts contact of a residue with other residues in the proteins which are set a minimum of 6 residues apart. It classifies them as either a long range contact or a short range contact with respect to the distance between them. It undergoes iterative steps to confirm the multiple contacts seen in the protein.

Anglor [18] is a machine learning algorithm that predicts the backbone torsional angles from amino acid sequence using the *ab initio* method. The prediction is perfected by neural networks and a support vector machine which takes a training set of existing data and predicts the dihedral angles. The dihedral angles are predicted along with the secondary structure.

Protein Structure analysis

The CXCL8 like protein structures were retrieved from the SCOP database. Two criteria were set for choosing the structures as follows, i) only X-ray structures (no NMR structures) and ii) the crystal with a high resolution (≤2.5 A). The structure, which shows high sequence similarity (>90%) was removed for making a non-redundant dataset.

The collected set of CXCL8 like proteins is aligned structurally using multiple structural alignment tool, Mustang to obtain a structure-based sequence alignment. The secondary structural regions were mapped on to the multiple sequence alignment for better visualization of various residues that are contributed to form the same secondary structures.

The solvent accessible surface area (PSA) of the collected proteins was calculated using the JOY script. An output of ".psa" file carries the percentage of accessible surface area for every residue in the given protein. The PSA is listed from a scale of 0 to 9, where 0 is least accessible and 9 is most accessible. Residues were classified to their amino acid alphabets [19] for residues according to the PSA value calculated. The residues are labelled as Buried (PSA<30%), Neutral (30% to 70%), Surfaced (PSA>70%).

The hydropathy index is used to check its role in the structure using the ExPASy PROT scale server. The window size is set to minimum to cover most of the residues in the input sequence and the value is normalized for our calculation. The hydrophobicity is higher when the hydropathy value is high. The residues are classified into the buried (Hval>7), neutral (4<Hval< 7) or exposed (Hval <3) using the hydropathy values.

The retention coefficient was computed using the ExPASy PROTscale server. The window size was set to be smaller and the values were normalized. The residues were classified into the buried (RC>7), neutral (4<RC< 7) or exposed (RC <3) using the hydropathy values.

The MCI was calculated using the PROFcon server and residues were classified as long range or short-range ordered. The residues are given alphabets L for Long-range order protein or S for short-range ordered protein. The distribution of the long-range and short-range contacts in the secondarry structure is analyzed.

The amino acids have a preference to form a secondary structure of their interest at certain conditions. The 20 amino acids are sorted according to their tendencies to form β-strands [20]. The thermal denaturation transition energy and free energy scale of the 20 amino acids are calculated. A few amino acids, such as, Tyr, Thr, Ile, Phe, Trp, Val, Cys, Met and Ser are considered to be the most favorable residues in forming the β-strand and a few other amino acids, such as, Glu, Ala, Asp, Gly and Pro are listed as the least favorable β-strand forming residues. These residues were analysed for their tendency to form β-strand with the variation in their degree of preference.

Results and Discussions

The CXCL8 chemokine monomer (PDB id: 3IL8) has a flexible N-terminal region and a small 3_{10}-helix, which is followed by three-stranded anti-parallel β-sheet followed by a large α-helix (Figure 1a). As discussed in methodology, there were total of 172 CXCL8 like protein 3D structures were obtained while searching for similar structures in SCOP database. After careful considerations as mentioned in the methodology only 23 high resolution (<2.5A) crystal structures were selected, which were non-mutated as well as non-redundant (Supplementary Table 1). The final set of 23 chemokines that includes 3IL8 was aligned structurally using MUSTANG tool (Figure 2b) and a structure-based sequence alignment was extracted expressed using JMOL (Figure 1c).

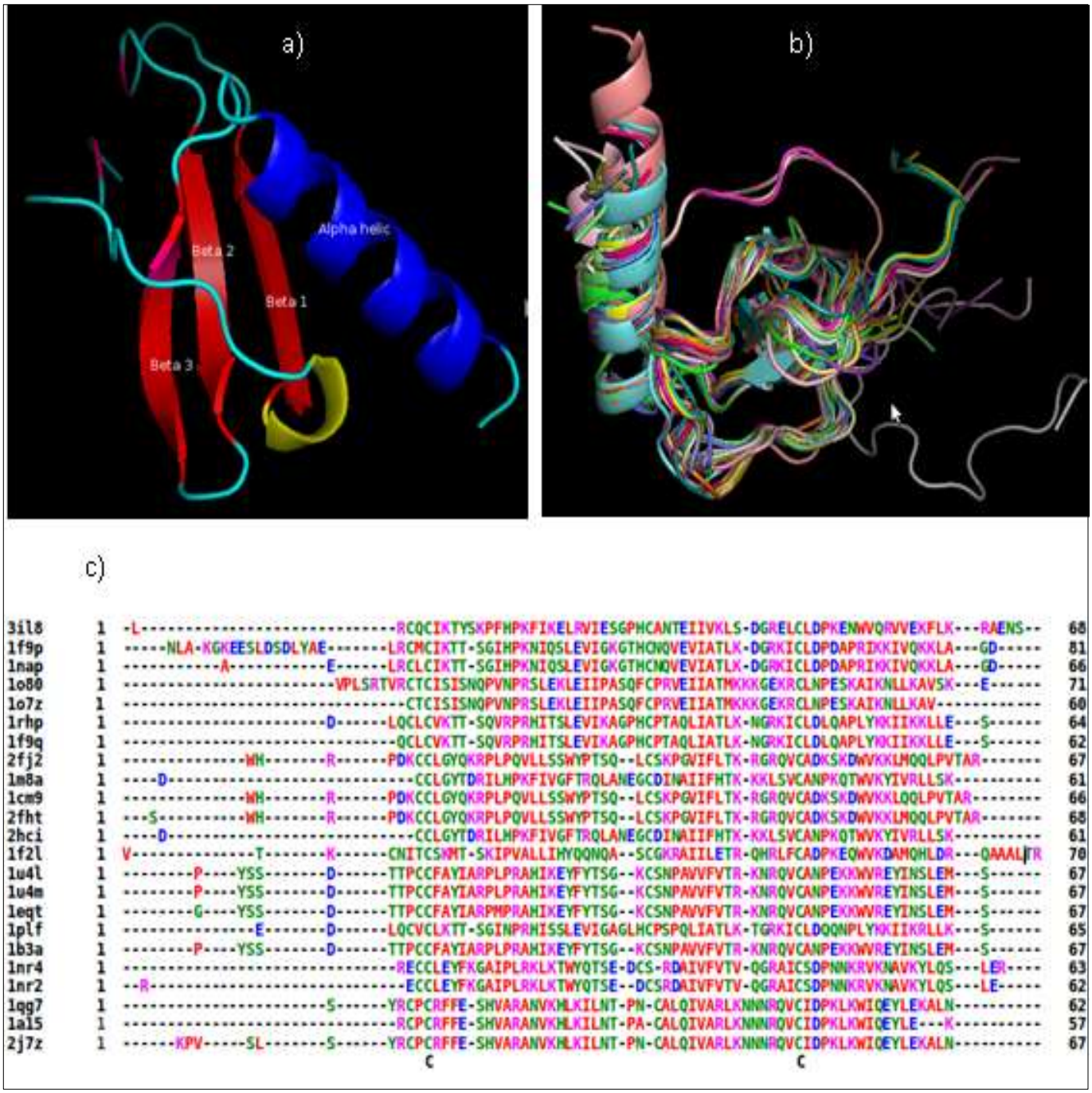


Fig 1: Chemokine structures and their alignment. a) CXCL8 structure (pdb id: 3IL8); b) Multiple structural alignment. The alignment of 23 chemokine protein structures; c) A structure-based sequence alignment.

Supplementary Table 1: The final list of CXCL8 like proteins. They are selected based on the criteria discussed in methodology

Sl no	Protein	Resolution (Å)	Chain with the domain	Sequence range	Sequence length	RMSD value	Species
1	3IL8	2	A	5 – 72	68	1.6	Homo sapiens
2	1PLF	2.2	B	21 – 85	65	1.9	Bos taurus
3	1RHP	2.4	B	7 – 70	64	1.8	Homo sapiens
4	1O7Z	1.92	A	9 – 69	61	1.4	Homo sapiens
5	1O80	2	B	1 – 75	75	2	Homo sapiens
6	1M8A	1.7	B	5 – 65	61	1.6	Homo sapiens ccl20/mip-3a
7	2HCI	1.81	B	5 – 65	61	1.8	Homo sapiens ccl20/mip-3a
8	2FHT	2.5	A	4 – 71	68	3	Kaposi's sarcoma-associated herpesvirus, VMIP-II
9	1CM9	2.1	B	8 – 74	67	3.1	Kaposi's sarcoma-associated herpesvirus, VMIP-II
10	2FJ2	2.3	B	5 – 71	67	3	Kaposi's sarcoma-associated herpesvirus, VMIP-II
11	1B3A	1.6	A	2 – 68	67	3.3	Homo sapiens
12	1EQT	1.6	A	2 – 68	67	3.3	Homo sapiens
13	1U4L	2	A	2 – 68	67	3.3	Homo sapiens
14	1U4M	2	A	2 – 68	67	3.3	Homo sapiens
15	1F2L	2	D	5 – 74	70	1.9	Homo sapiens
16	1F9P	1.93	A	6 – 86	81	1.4	Homo sapiens
17	1NAP	1.9	A	21 – 86	66	1.8	Homo sapiens
18	2J7Z	1.95	B	1 – 67	67	3.2	Homo sapiens
19	1QG7	2	A	6 – 67	62	2.7	Homo sapiens
20	1A15	2.2	A	1 – 67	67	2.8	Homo sapiens
21	1NR4	1.72	H	8 – 70	73	1.9	Homo sapiens
22	1NR2	2.18	B	8 – 69	62	1.9	Homo sapiens
23	1F9Q	2	C	209 – 270	62	1.6	Homo sapiens

The solvent accessibility test was carried out using the JOY tool. The percentage of solvent accessibility was calculated for every residue in the protein structure in scale from 0 to 9. A few residues were not assigned by the program to provide the PSA values because of the missing atoms in those residues. To avoid this technical error, a software tool, SCWRL-4 was used to add the missing atoms to those residues. The alphabets N (Neutral), B (Buried), and S (Surface) was assigned to every residue in the protein. The residues were replaced by the amino acid alphabet in the structural alignment and the result was observed. The results revealed that the β-1 strand and β-2 strands of all the 23 proteins showed similar kind of PSA pattern but the third strand was showing a varying pattern when compared to the first two β-strands. The second β strand showed very low PSA value throughout the 23 structures and is seen to be buried in nature (Supplementary Figure 1a).

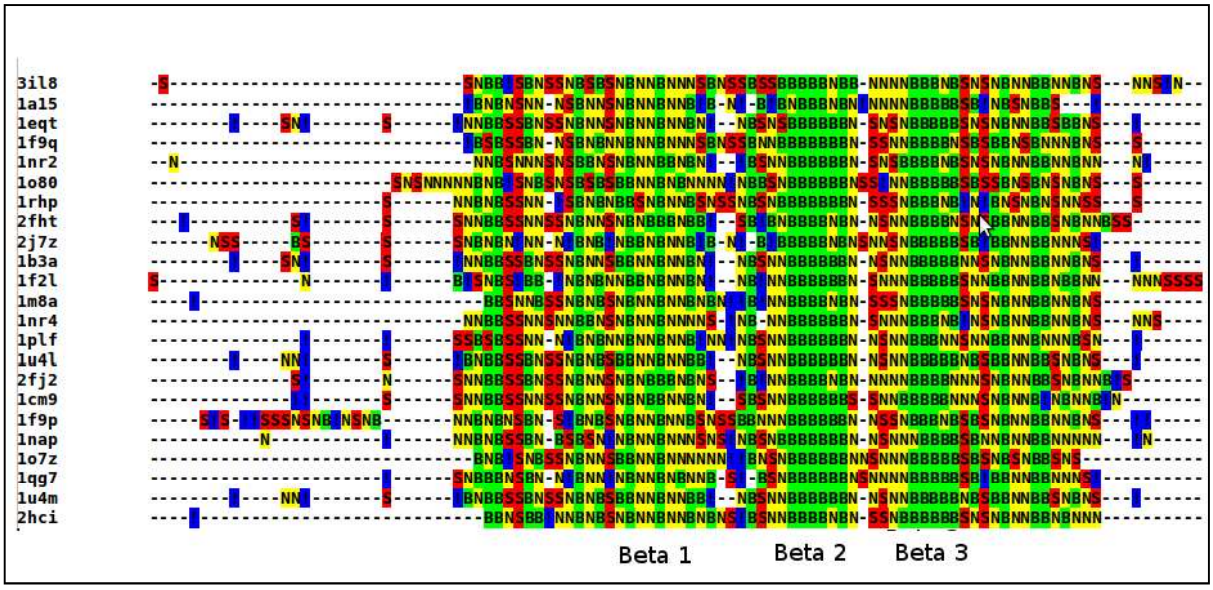
The amino acid alphabets show patterns, which were different in each β-strand. The β_1-strand shows an alternative neutral and buried alphabet as a pattern and β_2-strand was almost completely occupied by buried residues and the β_3-strand was showing buried alphabet (Supplementary Figure 1a). These results reveal that the β_2-strand was completely inside the core of the three dimentional structure, whereas β_3-strand was buried more than 75% in the core region and the β_1-strand was buried less than 50% in the core region. However, the alpha helix at the end was showing varying pattern of all three AAA, which shows that half of the helix was buried and another half was on the surface.

The hydropathy index values obtained were more favorable towards the hydrophilic nature, which may be due to the small window size. When the window size was increased, not all the residues were taken for calculation. The minimum window size was 2, which still leads to miss the first residue in the sequence. Due to the hydrophilic nature of the result most of the residues seems to fall on the neutral classification. When the hydropathy values and PSA was combined using their relation to hydrophobicity the AAA pattern seen in the residues still show a more prominent conservation in β_1 and β_2 strands than the β-3 strand.

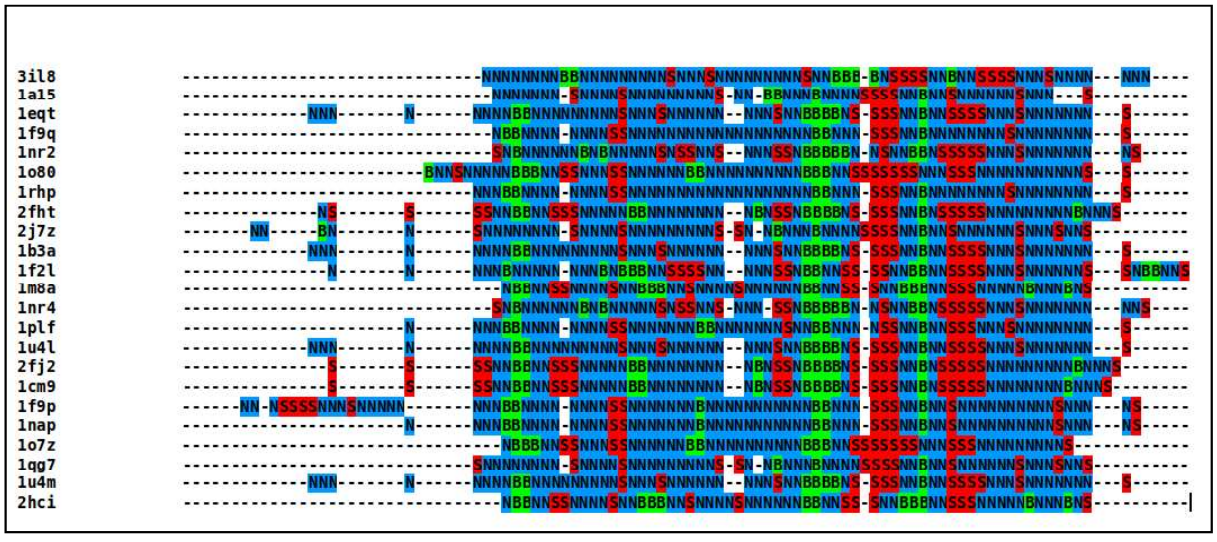
The PROTscale server was used to find the hydropathy values for the residues in the collected set of CXCL8 like proteins. The computed values were normalized and an amino acid alphabet was given to them with respect to their hydropathy values. The minimum window size was used to make sure that all the residues were included in the calculations but when the window size reduced the result was more favorable towards the hydrophilic result. During the computation not all the residues were included, the first residue missed to take part in the calculation. With the collected values the residues were labeled N (Neutral), B (Buried), and S (Surface). The small window size is affecting the actual hydropathy values of the residues (Supplementary Figure 1b). Since the Hydropathy values and the percentage of solvent accessibility are related towards the hydrophobic nature of the protein, both the features were combined and normalized values were provided. Each residue names were changed to their alphabets with respect to the combined values of the PSA and Hydropathy values (Supplementary Figure 1c). The amino acid alphabets were labeled as N (Neutral), B (Buried), and S (Surface). The retention coefficient of the protein was calculated by PROTscale server. The higher the retention coefficient value of the protein then lower the hydrophobicity was. The values obtained for PSA, hydropathy and retention coefficient were normalized to understand any relationship between them. The average values of the three features were normalized to a scale between 0 and 9 (Figure 2).

The normalized average values of the PSA, hydropathy index, and retention coefficient shows a range for the secondary structure elements in the protein. However, the range is not very specific although a small variation in range was seen in the β-strand compared to the helices. The helices have a higher range compared to the β-strands. The mean values were 3-6 for the β_1-strand, 2-5 for the β_2-strand, 4-6 for β_3-strand and 4-7 for the helices.

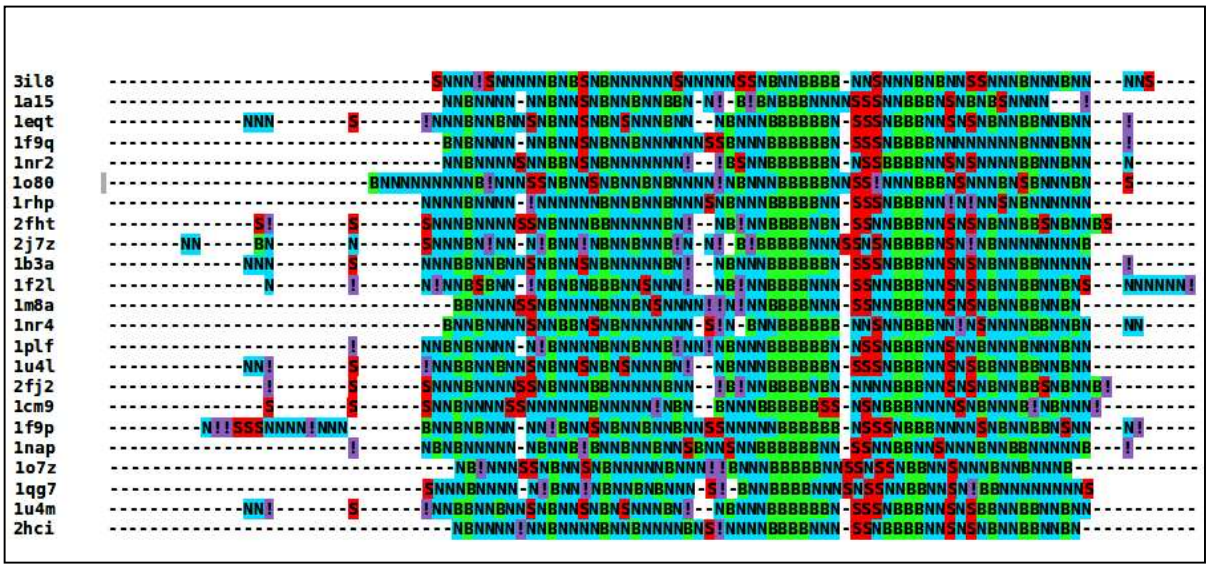
Fig 2: A relationship between structural properties. The PSA, hydropathy index and retention coefficient values of the secondary structure elements in chemokines were normalized and mapped on the alignment.



a)



b)



d)

Supplementary Fig 1: Solvent accessibility and hydrophathy properties of chemokines. a) The multiple structural alignment with Amino Acid Alphabet using the percentage of solvent accessibility; b) The Amino acid alphabet using the Hydropathy values of the residues; c) The percentage of combined and normalized values of solvent accessibility and hydropathy. N-Neutral B-Buried S-Surface !--Residues with missing atoms

The value of multiple contact index (MCI) of each residue was collected using the PROFcon server. The output has value for each residue against every residue in the protein. These values are indicative of either long range order or short range order contacts. Residues with contacts distance more than 6.5 A^{o} is considered as a long range contact. We considered only a 7 of the 23 proteins from the dataset for simplification and chose to label the residues as L for long range order residue or S for short range order residue (Figure 3).

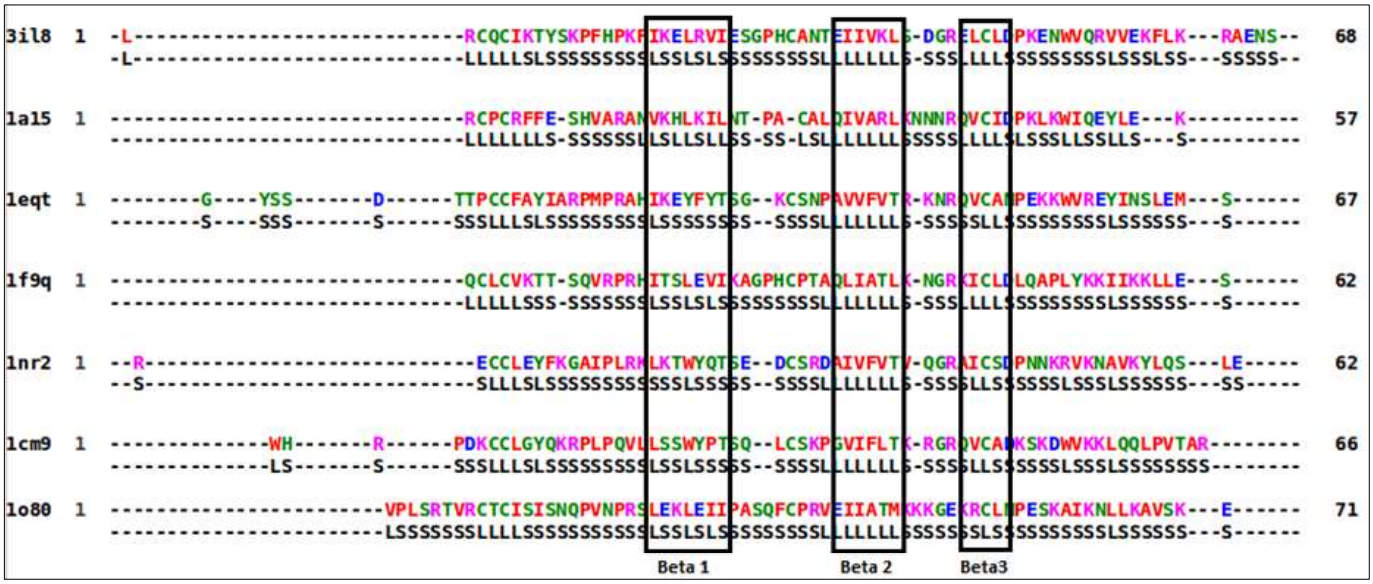


Fig 3: Multiple contact index (MCI). The secondary structure elements of 7 selected chemokines were annotated for their long and short range contact values obtained through PROFcon server.

The MCI shows that the β_2-strand residues were completely in the long range contacting order, which might be important for the key determinant of protein folding [21]. It is comparable with the results of PSA, where β_1-strand shows an alternative neutral and buried alphabet pattern and β_2-strand was almost completely occupied by buried residues and the β_3-strand was showing buried alphabet. It reveals that the β_2-strand, which is in the core of the structure have high tendency to form long-range interaction compared to β_1-strand that do not have high tendency to form a long-range interaction.

Every individual residue seen in the β-strands was taken and checked for its favorability to form β-strands. The dihedral angle of the β-strands forming residues were taken from three sample chemokines (PDB ID: 1EQT, 1U4L and 3IL8) and plotted on a Ramachandran map (Supplementary Figure 2). The same residues were studied for their favorability to form β-strands (Supplementary Figure3).

A correlation coefficient analysis was made between the PSA scale (X-axis) and the favorability of the β-strand forming residues (Y-axis) and was plotted on a scatter plot for two of the proteins, 1EQT and 1NR2 (Figure 3).

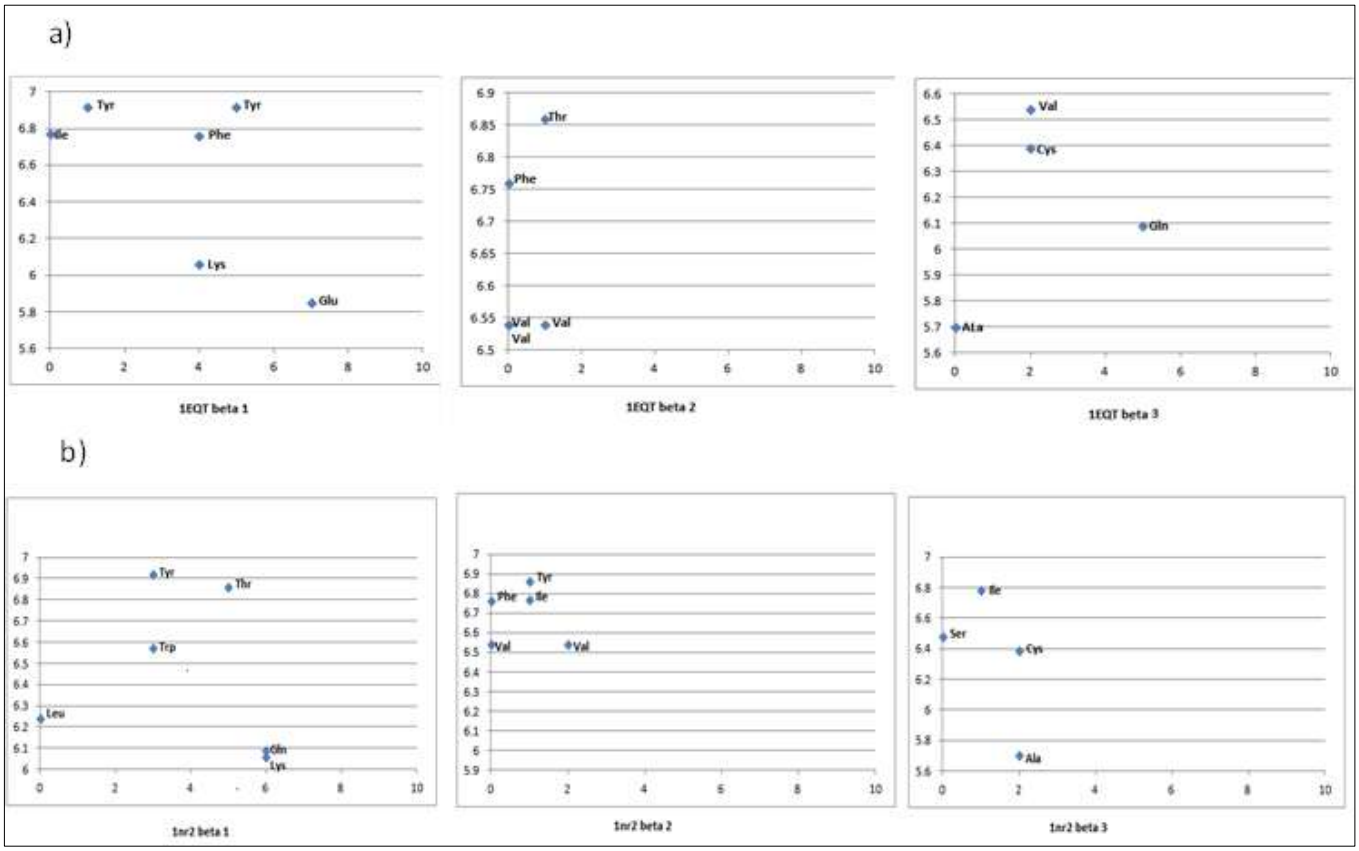
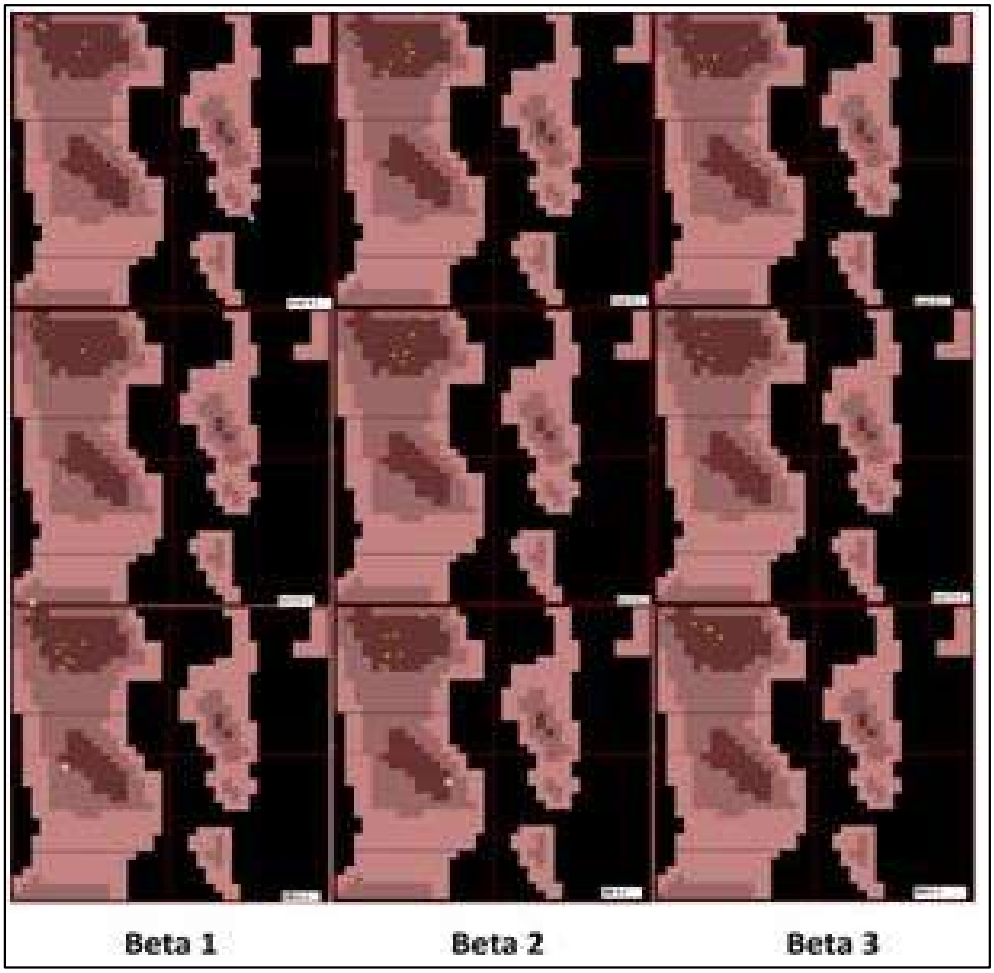
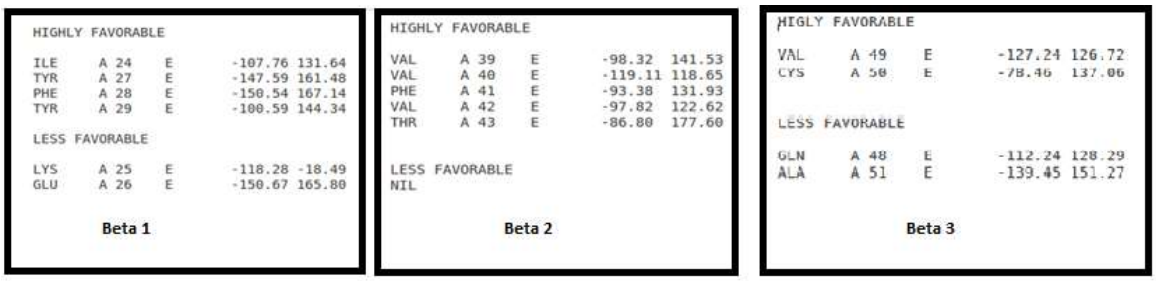


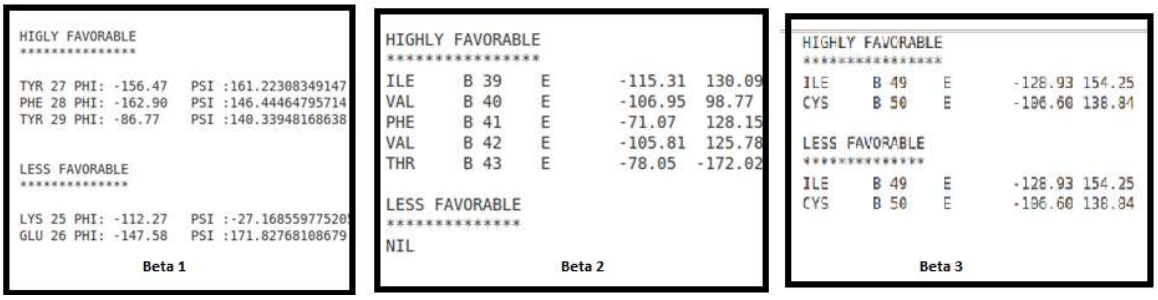
Fig 3: Correlation coefficient between PSA scale (X-axis) and the favorability of the β-strand forming residue (Y-axis). The scatter plots of the residues in all three β-strands were plotted against their respective PSA.



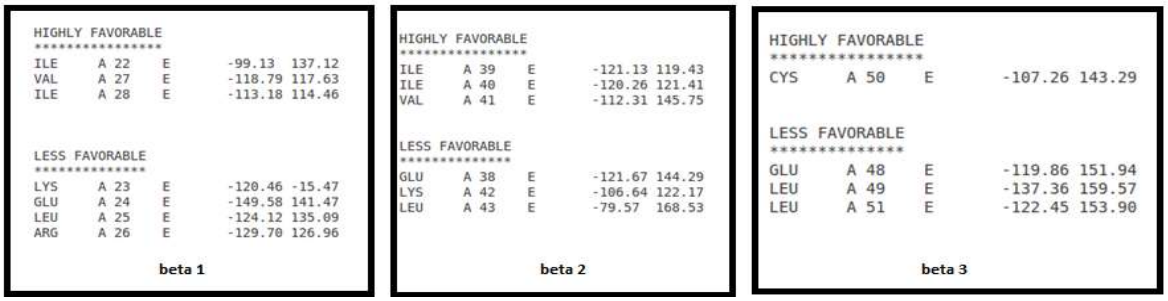
Supplementary Fig 2: The Ramachandran plot for the dihedral angles. The dihedral angles of amino acids from three β strands of 1EQT, 1U4L and 3IL8.



a)



b)



c)

Supplementary Fig 3: β-strand forming favourable residues from three different proteins. a) 1EQT; b) IU4L and c) 3IL8

The plot shows most of the residues correlate with low PSA values and high favorability values. The analysis clearly indicates the favorability of β-strand forming tendency of the amino acids was decreased when PSA increases. The result reveals that high favourable β-strand forming residues are tend to form β-strand in the core of the structure whereas less favourable β-strand forming residues are tend to form β-strand which are not in the core of the structure. The correlation coefficient between the percentage of solvent accessibility and retention coefficient seems to be higher when compared with other structural properties of proteins. The correlation coefficient was 0.67 for one set of data.

Conclusion

By comparing the physico-chemical properties of residues that tend to form secondary structures, it is possible to explain how different residues tend to form similar secondary structure. Although different residues tend to form similar β - strands in proteins, the residues that are involved in forming the β-strand at the core region are different from the residues that are forming β-strand other than core region. The core β-strand forming residues are highly favorable residues and their phi-psi values are tend to fall within a short-range, whereas other than core region or surface β-strand forming residues are less favorable residues and their phi-psi values are tend to fall in a broad range that are extended sheet in nature. However, to understand the preference of helix forming residues the data might not be sufficient, since chemokines have a single α-helix and a small 3_{10}-helix. But, sufficient informations are given for β-strand forming residues. These results indicate residues that tend to form similar secondary structures with differences in physico-chemical property in it might be responsible for different protein folds and new folds due to evolution in protein. This is a novel approach, which throw a light on applications that deal with prediction of either existing or novel fold of unknown protein.

Acknowledgement

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

References

1. Gura T. Chemokines take center stage in inflammatory ills. Science. 1996;272:954-956.
2. Cocchi F, DeVico AL, Garzino-Demo A, Arya SK, Gallo RC, Lusso P. Identification of RANTES, MIP-1α, and MIP-1β as the major HIV-suppressive factors produced by CD8+ T cells. Science. 1995;270(5243):1811-1815.
3. Narvaiza I. Intratumoral Coinjection of Two Adenoviruses, One Encoding the Chemokine IFN-γ-Inducible Protein-10 and Another Encoding IL-12, Results in Marked Antitumoral Synergy J Immunol. 2000;164:3112-3122.

4. Baggiolini M, Dewald B, Moser B. Human chemokines: an update. Annu. Rev. Immunol. 1997;15:675-705.
5. Bazan JF, Bacon KB, Hardiman G, Wang W, Soo K, Rossi D, *et al*. A new class of membrane-bound chemokine with a CX3C motif. Nature. 1997;385:640-644.
6. Onuffer JJ, Horuk R. Chemokines, chemokine receptors and small antagonists: recent developments. Trends Pharmacol Sci. 2002;23:459-467.
7. Drury LJ, Ziarek JJ, Gravel S, Veldkamp CT, Takekoshi T, Hwang ST, *et al*. Dwinell, Proc. Natl. Acad. Sci. USA. 2011;108:17655.
8. Chothia C, Lesk AM. How different amino acid sequences determine similar protein structures: The structure and evolutionary dynamics of the globins. JMB. 1980;136(3):225-230.
9. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: A structural classification of proteins database for the investigation of sequences and structures. J Mol. Biol. 1995;247:536-540.
10. Holm L, Rosenstrom P. Dali server: Conservation mapping in 3D. Nucleic acid Res. 2010;38:W545-W549.
11. Jmol: an open-source Java viewer for chemical structures in 3D. http://www.jmol.org/k
12. Pollastri G, Baldi P, Fariselli P, Casadio R. Prediction of Coordination Number and Relative Solvent Accessibility in Proteins. Proteins. 2002;47:142-153,
13. Arun S. Konagurthu, James C. Whisstock, Peter J. Stuckey Arthur M. Lesk. MUSTANG: A Multiple Structural Alignment Algorithm. Proteins. 2006;64(3):559-574.
14. Cook DN. The role of MIP-1α in inflammation and hematopoiesis. J. Leukoc. Biol. 1996;59:61-66.
15. Qiang Wang, Adrian Canutescu A, Roland Dunbrack L Jr. SCWRL and MolIDE: Computer programs for side-chain conformation prediction and homology modeling. Nat Protoc. 2008;3(12):1832-1847.
16. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, *et al*. Protein Identification and Analysis Tools on the ExPASy Server, 2005, 571-607.
17. Marco Punta, Burkhard Rost1. PROFcon: novel prediction of long-range contacts. Structural bioinformatics. 2005;21(13):2960-2968.
18. Wu S, Zhang Y. ANGLOR: A Composite Machine-Learning Algorithm for Protein Backbone Torsion Angle Prediction. Plos One. 2008;3(10):e3400. doi:10.1371/journal.pone.0003400.
19. Kanagarajadurai K, Sowdhamini R. Sequence and structural analyses of interleukin-8-like chemokine superfamily. In Silico Biol. 2008;8:307-330.
20. Catherine K. Smith$ Jane M, Withka O, Lynne Regan J. A Thermodynamic Scale for the & Sheet Forming Tendencies of the Amino Acids. Biochemistry. 1994;33:5510-5517.
21. Michael Gromiha M. Multiple Contact Network Is a Key Determinant to Protein Folding Rates. J Chem. Inf. Model. 2009;49:1130-1135.