

Multiple Sequences Alignment and Structure Prediction of T-cell Surface Glycoprotein CD4 Protein in Various Species

Sichen Zhou

Wuhan Britain-China School, Wuhan, Hubei, China.

Abstract

CD4 protein is present on the surface of 65% of human T cells, it plays an important functional role in the course of human immunodeficiency virus 1(HIV-1) infection. The difference in various CD4 proteins raise the great interest and necessary to investigate. In this project, the sequence diversity and structure difference among 15 species will be focused, and many free tools for biological data analysis, such as MEGAX and PyMOL. The result is that the more similar species evolve, the more similar or even the same length of CD4 will be, and the coverage of CD4 will also become the same.

Keywords

CD4 Protein; HIV; Sequence; Similar Species; Length.

1. Introduction

CD4 protein is present on the surface of 65% of human T cells, which is essential for the immune system of human. For instance, it plays an important functional role in the course of human immunodeficiency virus 1 (HIV-1) infection [1]. When the HIV virus infects cells with CD4 surface proteins, the gp120 protein on HIV virus will bind to CD4 protein, then to attack the T cells [2-3]. Further mutation studies reported that the region spanning amino acids 40–60 of CD4 contacts with the viral gp120. Meanwhile, sequence variations of CD4 protein will affect the binding affinity of gp120-CD4 complex, which is reasonable that site diversity results in the structural difference. In abundant researches, the sequences of CD4 proteins of many species have been identified (e.g. Homo sapiens, Macaca mulatta, Cercopithecus atys etc.) [4], and a small number of 3D structures have been determined (e.g. 5u1f) [5]. Those sequences and structures are very helpful to understand the molecular mechanisms of HIV infection. The differences in various CD4 proteins raise great interest and are necessary to investigate.

With the development of database and bioinformatics, it is not difficult to obtain the information for CD4 proteins. As introduced, UniProt [6] and PDB [7] database have abundant information about protein function, structure, gene annotation, etc. Meanwhile, there are many free tools for biological data analysis, like MEGAX [8] for sequence analysis, PyMOL [9] for structural analysis. Both data source and available tools can be used to perform the project about exploring the difference in sequences and structures of CD4 proteins.

In this project, the sequence diversity and structure difference among the 15 species will be focused. Firstly, the multiple sequence alignment is adopted to obtain the sequence features, both the conserved and variable regions are analyzed. Then, to learn the similarity of CD4 among those species, phylogenetic tree is constructed. As part of structure for those species is available, Swiss-model is used for protein structure prediction with template-based method, then structure alignment provides more details on CD4 difference among various species.

2. Methods

2.1. Data set preparation

To construct the data set for this project, the amino acid sequences of CD4 protein are downloaded from Uniprot (www.uniprot.org), and then can be organized into tables (as shown

in table 1). In total, 15 species are selected for the following sequence and structural analysis. The length of those protein sequences range from 397 to 463 in various species.

Table 1. CD4 protein sequences used in this project.

Uniprot ID	Gene name	Species	Common name	Length
P16003	CD4_MACMU	Macaca mulatta	Rhesus macaque	458
Q08336	CD4_CERAT	Cercocebus atys	Sooty mangabey	397
P05540	CD4_RAT	Rattus norvegicus	Rat	457
P16004	CD4_PANTR	Pan troglodytes	Chimpanzee	458
Q9XS78	CD4_DELLE	Delphinapterus leucas	Beluga whale	455
P01730	CD4_HUMAN	Homo sapiens	Human	458
Q29037	CD4_SAISC	Saimiri sciureus	Common squirrel monkey	457
P33705	CD4_CANLF	Canis lupus familiaris	Dog	463
Q08338	CD4_CHLAE	Chlorocebus aethiops	Green monkey	458
P06332	CD4_MOUSE	Mus musculus	Mouse	457
P79184	CD4_MACFU	Macaca fuscata fuscata	Japanese macaque	458
Q08339	CD4_ERYPA	Erythrocebus patas	Red guenon	397
P46630	CD4_RABIT	Oryctolagus cuniculus	Rabbit	459
Q08340	CD4_MACNE	Macaca nemestrina	Pig-tailed macaque	458
P79185	CD4_MACFA	Macaca fascicularis	Crab-eating macaque	458

2.2. Multiple sequence alignment

For multiple sequence alignment, ClustalX software in MEGAX was applied to analyze 15 CD4 sequences across diverse species. When running ClustalX, BLOSUM matrix was adopted and other parameters for the alignment were fixed with their default values.

2.3. Phylogenetic trees construction

Then, with the multiple sequence alignment results, MEGAX was used for phylogenetic tree building with Maximum Like-hood method, which is one of the criteria for choosing the best phylogenetic tree. To increase the reliability of the phylogenetic tree, the resampling method was adopted and the bootstrap value set as 1000. The rough results showed that sequences similarity is highly related to the amino acid lengths.

Table 2. Sequence identity and coverage between target CD4 protein and template used to predict the protein structure.

CD4 protein	Sequence identity	coverage
P05540	53.63	0.78
P06332	52.37	0.79
P16003	90.91	0.79
P16004	98.90	0.79
P29037	78.73	0.79
P33705	59.83	0.77
P46630	57.98	0.78
P79184	90.91	0.79
P79185	90.91	0.79
Q9XS78	60.34	0.79
Q08336	90.03	0.91
Q08338	90.08	0.79
Q08339	90.30	0.91
Q29037	78.73	0.79
P01730	100.00	0.79

2.4. Protein structure prediction

Next, Swiss-model [10] (swissmodel.expasy.org) is used to predict the three dimensional protein structure based on homology-modelling. To select templates with the highest quality, the template's quality is predicted from features of the target-template alignment in Swiss-model. The model quality is highly depended on the sequence identity and coverage between target sequence and template. Higher sequence identity and coverage will predict the protein

structures with higher accuracy. In this project, chain I of structure 5u1f Conserved coordinates in both target and the template is copied from the template to the predicted structure. Insertions and deletions in structures are built by using a fragment library. Side chains and loops with low sequence identity are then optimized by using a force field and energy function.

2.5. Structure alignment and analysis

Finally, to analyze the conformation difference in various species, PyMol is used to calculate RMSD and perform structural alignment and represent the differences in details. The coordinates of C α atom in residues are used to calculate the distances between atom pairs, best RMSD score is optimized after 5 cycles calculation and except for the atoms with lower alignment. The RMSD distance matrix will help to understand the structural differences of CD4 proteins in different species.

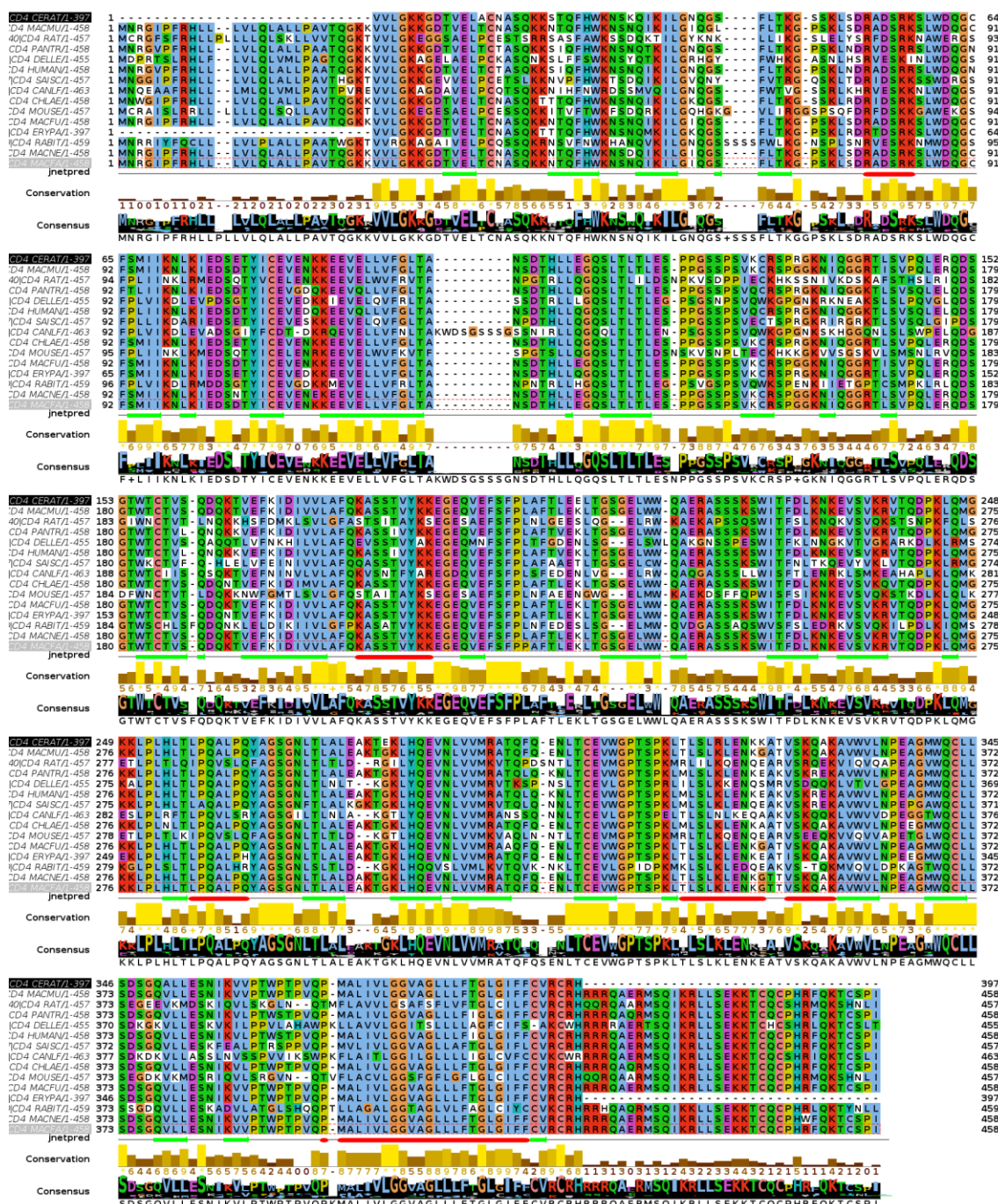


Figure 1. Multiple sequence alignment for CD4 protein in 15 species and conserved domain identification.

3. Results and Discussion

3.1. Multiple sequence alignment results and feature analysis

The CD4 proteins from different species have various amino acids length, then the conservation and variance in CD4 proteins are analyzed by using multiple sequence alignment. As shown in Figure 1, we found that CD4 protein in dogs (P33705) is different in particular, as two regions in.

This sequence is relatively longer than others, including -GSSSSFW- and -TAKWDSGSSSGSS- region, which are serine rich regions. It is well known protein has four-level structure, amino acid sequence is the primary structure, which is the basis of the higher structure, so the difference in primary sequence will result in the conformational change in three dimensional structure. Thus, from the results of multiple sequence alignment, we could suppose there are differences in 3D structure. In addition, before 3D structure prediction, the secondary structure for each CD4 protein prediction was firstly presented. The secondary structure contains α -helix, β -pleated sheet, β -turn, random coil, which will fold into the functional 3D structure. As shown in Figure 1, the predicted secondary structure corresponding to the sequence presented below the sequence alignment (green curves indicate β -pleated sheets), indicating that the CD4 protein is rich in β -pleated sheets.

Based on multiple sequence alignment, the phylogenetic tree for CD4 protein in 15 species is constructed. As shown in Figure 2, the confidence value from bootstrap (set to 500) is represented as purple cycle, the larger the mean is, the higher the confidence for the branch is. Here, bootstrap means that the phylogenetic tree has been reconstructed for 500 times, every time the tree is acquired, the confidence value will be obtained from statistical analysis. Thus, it can be learned that various primates are more similar than dog, mouse, and rabbit. Furthermore, it is reasonable that HIV will attack primates more easily than others as their CD4 protein similarity.

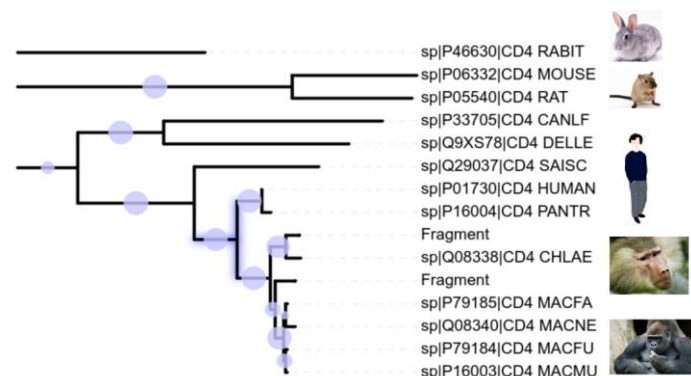


Figure 2. Phylogenetic tree for CD4 protein in 15 species.

3.2. Structural alignment and differences analysis

To learn the higher-level structural difference, the 3D structures are predicted by using homology modelling method. Then, the structural alignment for those predicted structures is performed, C_{α} atom in all residues was used for the vector calculation. To assess the structural difference, we used root-mean-square deviation (RMSD) as criteria, which defined as $RMSD = \sqrt{\frac{1}{N} \sum_{i=1}^N d_i^2}$, where N is the residue number in CD4 protein, i is the i th residue, d is the vector for

i th residue in two structures. As shown in Table 3, we built the RMSD matrix for all predicted structures. The largest RMSD value is 2.051, the difference between P33705 and P06332. While, RMSD value indicated an overall alignment as the total difference value is divided by the residue number.

Table 3. RMSD matrix among the 15 CD4 protein structures.

	P05540	P06332	P16003	P16004	P29037	P33705	P46630	P79184	P79185	Q08336	Q08338	Q08339	Q29037	Q9XS78
P01730	0.752	1.624	0.036	0.013	0.424	1.204	1.55	0.038	0.046	0.034	0.041	0.034	0.424	1.566
P05540		1.51	0.752	0.752	0.824	0.911	1.486	0.752	0.752	0.873	0.752	0.754	0.824	1.508
P06332			1.506	1.625	1.509	2.051	0.827	1.625	1.625	1.599	1.498	1.628	1.509	0.579
P16003				0.036	0.424	1.204	1.537	0.01	0.025	0.035	0.03	0.028	0.424	1.567
P16004					0.424	1.204	1.55	0.037	0.046	0.034	0.039	0.033	0.424	1.566
P29037						1.075	1.686	0.424	0.424	0.426	0.423	0.425	0	1.548
P33705							1.632	1.205	1.205	1.279	1.205	1.207	1.075	1.627
P46630								1.537	1.538	1.585	1.537	1.537	1.686	0.656
P79184									0.024	0.037	0.032	0.03	0.424	1.567
P79185										0.045	0.033	0.035	0.424	1.568
Q08336											0.039	0.033	0.426	1.615
Q08338												0.018	0.423	1.567
Q08339													0.425	1.569
Q29037														1.548
Q9XS78														

Then, as shown in Figure 3, the left panel is the predicted CD4 protein structures colored with RMSD hot-map, blue regions represent the conserved domain in structural alignment, red means flexible region with higher RMS value. In the right panel of Figure 3, the structural alignment between P01730 and P33705 was listed, and the three most different region in two structures were highlighted, which are EVEDQ, LEAKTGKL, and FTVEKLTGSG in P33705, and DTDK-, L-AK-GTL, and FEDENLVG-- in P01730. Protein P33705 have longer loops compare to the others.

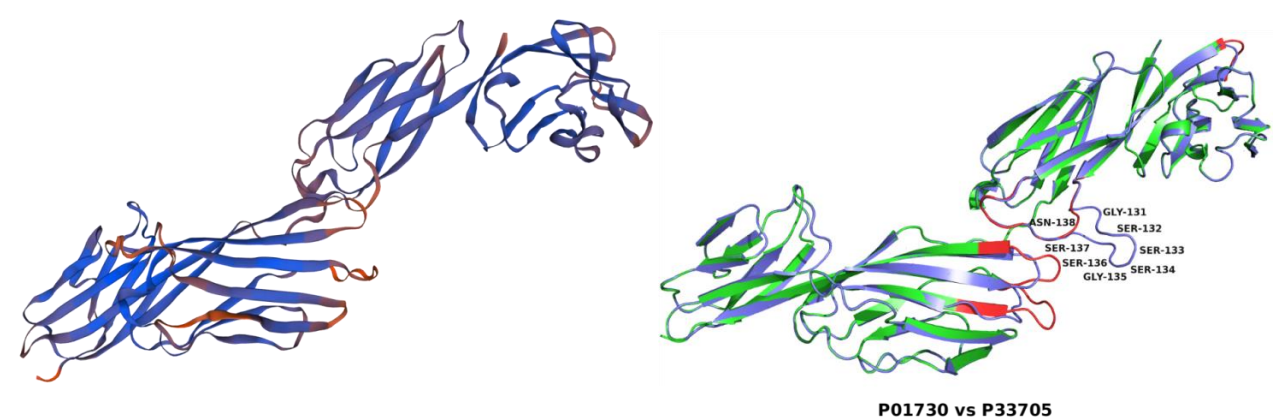


Figure 3. Structural difference in details of CD4 protein.

4. Conclusion

In the project, the paired sequence alignment tool the multi-sequence alignment tool and the sequence similarity search tool were used to conduct sequence alignment. Then, the sequence alignment results, the evolutionary tree construction results, the structure prediction results and the structure alignment results through the tools such as SWISS-MODEL and PyMOL were obtained.

According to the CD4 sequence comparison results of fifteen types including *Macaca mulatta*, *Cerocebus atys* and *Rattus norvegicus*, it can be known that the more similar species evolve, the more similar or even the same length of CD4 will be, such as *Homo sapiens*, *Chlorocebus aethiops* and *Macaca nemestrina*. Taking P1730 as a reference, it can also be known that the more similar species evolve, the higher the sequence identity of CD4 will be. Meanwhile, the coverage of CD4 will also become the same, such as P79184, P16003 and P16004.

References

- [1] Mascola JR, Haynes BF. HIV-1 neutralizing antibodies: understanding nature's pathways. *Immunol. Rev.* 2013; 254:225–244.
- [2] Kwong PD, et al. Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. *Nature.* 1998; 393:648–659.
- [3] Fouts T R, et al. Expression and characterization of a single-chain polypeptide analogue of the human immunodeficiency virus type 1 gp120-CD4 receptor complex. *Journal of virology*, 2000, 74(24): 11427-11436.
- [4] Meyerson N R, et al. Identification of owl monkey CD4 receptors broadly compatible with early-stage HIV-1 isolates[J]. *Journal of Virology*, 2015, 89(16): 8611-8622.
- [5] Liu Q, et al. Quaternary contact in the initial interaction of CD4 with the HIV-1 envelope trimer. *Nat Struct Mol Biol.* 2017 Apr;24(4):370-378. doi: 10.1038/nsmb.3382.
- [6] UniProt Consortium. UniProt: a hub for protein information. *Nucleic acids research* 43. D1 (2015): D204-D212.
- [7] Sussman, Joel L., et al. Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. *Acta Crystallographica Section D: Biological Crystallography* 54.6 (1998): 1078-1084.
- [8] Kumar, Sudhir, et al. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Molecular biology and evolution* 35.6 (2018): 1547-1549.
- [9] DeLano, Warren Lyford. PyMOL. (2002): 700.
- [10] Guex, Nicolas, and Manuel C. Peitsch. SWISS-MODEL and the Swiss-Pdb Viewer: an environment for comparative protein modeling. *Electrophoresis* 18.15 (1997): 2714-2723.