IDENTIFICATION OF CONSERVED FUNCTIONAL MOTIFS IN LACTOFERRIN USING MEME

*** Shashank Rana**[1]
Shashank.bioinfo@gmail.com
Shrikant Sharma[1]
shribioinfo@gmail.com

1. Research Scholar, Bioinformatics Facility, Department of Immunology, College of Biotechnology, Sardar VallabhBhai Patel University of Agriculture & Technology. Meerut (U.P.)

Raghvendar Singh[2]
Raghvendar@gmail.com

2. Head of Department, Bioinformatics Facility, Department of Immunology, College of Biotechnology, SardarVallabhBhai Patel University of Agriculture & Technology. Meerut (U.P.)

Corresponding Author-*Shashank Rana (Shashank.bioinfo@gmail.com)

Abstract

Lactoferrin is an iron binding globular protein with antimicrobial activity was firstly isolated in bovine milk. Lactoferrin (LF) is structurally similar to the transferrins. So it is also known as lactotransferrin (LTF) is a globular multifunctional protein. Our work was on motif discovery by using OOPS modal of MEME (Multiple EM for Motif Elicitation) tool. The aim of motif discovery is to detect short, highly conserved patterns in a collection of unaligned DNA or protein sequences. We have taken fifteen LTF AA sequencesfrom different resources.By analysis of these sequences, three motifs were retrieved. It is to be noted that all fifteen sequences contains all three motifs but start points of those are different. Each motif has 15 sites and 50 widths. On the bases of motif analysis, it is distinct that LTF retrieved from any milk recourse, have common conserved patterns.

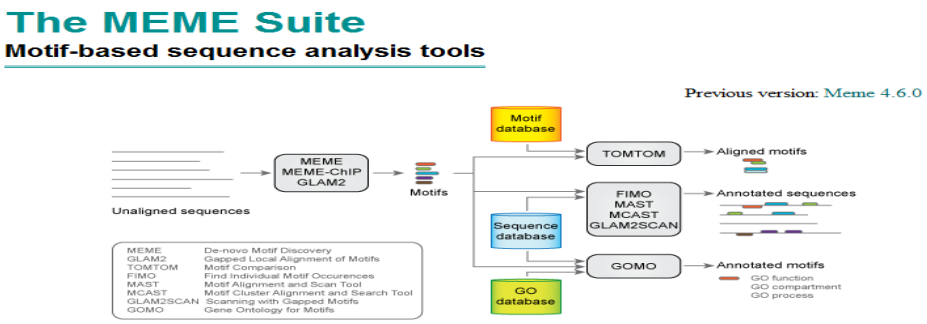
Keywords- ***MEME, LTF, OOPS, Conserved pattern, MAST***

Introduction

Lactoferrin form bovine milk was first isolated by Sorenson and Sorenson firstly in 1939 [1]. It is well known fact that LF have an iron binding properties and it also have similarity with transferrin, therefore also called lactotransferrin. Lactoferrin is considered a multifunctional or multi-tasking protein. LF has antibacterial, antiviral, antifungal, anti-inflammatory, antioxidant and Immunomodulatory activities [2]. LF is reported to have metal transfer, antiviral, anti-ageing agents. Javed et. al. (2001) reported lactoferrin from camel has two lobs N (Iron binding) and C (discharge of Iron) [7]. We have taken fifteen protein sequences of LTF from different milk producing animals. The length of shortest sequence was 692 residues and the longest sequence 711 residues and the modal for motif generation was OOPS. The purpose of MEME (Multiple EM for Motif Elicitation) (rhymes with 'team') [5] is to allow users to discover signals (called 'motifs') in DNA or protein sequences. For motif study of lactoferrin, we use by default value as described by MEME suite. The MEME Suite is well known software package with a

unified web server interface that enables us to perform all four types of motif analysis wiz 1) motif discovery, 2) motif–motif database searching, 3) motif-sequence database searching and 4) assignment of function [6].

Figure-1: MEME overview



Materials and Methods

In present study, we have selected fifteen lactoferrin sequences (amino acid) in FASTA format retrieved from NCBI [8] (Table-1). Motif analysis in lactoferrin sequences was conducted by using OOPS model of MEME.. The output of this modal of MEME shows color graphical alignment as well as common regular expression of motifs. On the hand, the block represents start and end point of the amino acid sequences with motif length. This is well known fact that E-value describes the statistical significance of the motif. MEME [9] (ver.4.6.1) usually finds the most statistically significant (low E-value) motifs first. The E-value is an assessment of the probable number of motifs with the given log probability ratio (or higher) along with the same width and site count present, that one would find in a similarly sized set of random sequences. On the other hand, motif width defines that each motif describes a pattern of a fixed with as no gaps are allowed in MEME motifs. In MEME package, sites define the conserved regions present in the particular motifs. Site numbers are the important contributing factor to the construction of the motifs. The information content of the motif (In bits), is equal to the sum of the uncorrected information content, R (), in the columns of the LOGO as described in user manual of MEME suite and MEME suites follows position specific probability matrices that specify the probability of each possible letter appearing at each possible position in an occurrence of the motif and are displayed as "sequence LOGOS", containing stacks of letters at each position in the motif. It is to be noted that the total height of the stack is the "information content" of that position in the motif in bits. For identification of motifs in proteins, the categories are based on the biochemical properties of the various amino acids. Further we also analyse motif by using Motif Alignment and Search Tool [10] (MAST ver.4.6.1). .

Results and Discussion

According to Bailey et al. (2006), by default, MEME looks for up to three motifs, each of which may be present in some or all of the input sequences. MEME chooses the width and number of occurrences of each motif automatically in order to minimize the ‘E-value’ of the motif—the probability of finding an equally well-conserved pattern in random sequences. By default, only motif widths between 6 and 50 are considered, The MEME output is HTML and shows the motifs as local multiple alignments of (subsets of) the input sequences, as well as in several other formats. ‘Block diagrams’ show the relative positions of the motifs in each of the input sequences. After the submission of sequences in query box of MEME, results display in the form of graph and seq will displays in the form of sequence logo or regular expression(Bailey et al., 2006) (Table-2). Motif overview in figure-2 has shown 6.5e-590 E-value of motif one, 5.4e-541 E-value of motif two and 2.4e-515 E-value of motif three.In Results were analyzed on the bases

of e-value and p-value. Where second one defines about the conserved pattern of motifs and first one describe about the width of the same match. Higher p-value described the best match whereas lower the e-value better the results. E-value is defined as an estimate of the expected number of motifs with the given log likelihood ratio (or higher), and with the same width and site count, that one would find in a similarly sized set of random sequences. After the submission of multiple amino acid sequences to MEME, we find that all the sequences have all three motifs but the starts points of all these motifs are vary sequence to sequence. Figure-3, 5 and 7 represent the number of site which contributing to the construction of the motif. In each protein sequence define the site in colour format. These are shown aligned with each other. Every site is recognized by the name of the sequence where it occurs, the strand, and the Start position in the sequence where the site begins. The sites are listed in order of increasing statistical significance (p-value).In figure-4, 6 and 8 shown the block diagram of each site and represent the motif location. The occurrence of the motif in the protein sequences are shown as coloured blocks on a line. The block one , two and three represent all of fifteen sequences. One figure is printed for each sequence to show all the sites contributing to that motif in that particular sequence and the sequences were listed below in same order as in the input. The motif occurrences shown combined block diagram (Figure-9) might not be closely the similar as those reported in each motif segment because only motifs with a position p-value of 0.0001 that don't overlap other, more significant motif occurrences are shown in combined block.Following on, we were submit our MEME result in MAST (Figure-10).The combined block diagram represent on the bases of E-value. The sequence 1,2,3,5,6,8,9,10,11,12,13 and 14 were shown zero E-value. Other than Sequence 7, 4 and 15 shown 4e-200, 8.8e-194 and 1.5e-164 E- value (Figure-11). The sequence 1 is show higher similarity and sequence 15 show lower similarity in lactoferrin protein sequence. of E-value. The sequence 1,2,3,5,6,8,9,10,11,12,13 and 14 were shown zero E-value. Other than Sequence 7, 4 and 15 shown 4e-200, 8.8e-194 and 1.5e-164 E-value (Figure-11). The sequence 1 is show higher similarity and sequence 15 show lower similarity in lactoferrin protein sequence.

Table-1: Scientific name of different species

Sequence No.	Scientific Name
1	Homo sapiens
2	Pan troglodytes
3	Macaca cyclopis
4	Oryctolagus cuniculus
5	Camelus dromedarius
6	Equus caballus
7	Mus musculus
8	Sus scrofa
9	Ovis aries
10	Capra hircus
11	Bos taurus
12	Bos indicus x Bos taurus
13	Bos indicus
14	Bos grunniens
15	Rattus norvegicus

Table-2: motif information with sequence logo and regular expression

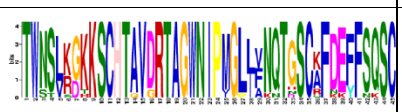
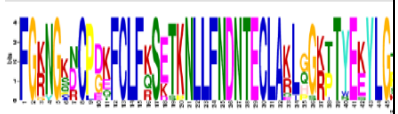
Sr. No.	Motif no.	Width	E Value	Sites	Sequence logo	Regular expression
1	1	50	6.5e-590	15		TWNS[LV][KR][GD]KKSCHTAVDRTAGWNIPMGL[LI][FVA]NQTGSC[AK]FDE[FY]FSQSCAPG[AS]D
2	2	50	5.4e-541	15		[QR]THYYAVAVVKKG[SG][NS]FQL[ND][DE]LQG[LR]KSCHTGLGR [ST]AGW[NI][IV]P[IM]G[IT]LRP[FY]L[NS]W
3	3	50	2.4e-515	15		FG[KR]NG][KS][DNR]CP[DG][KE][FCLF][KQR]S[EK]TKNLLFNDNTECLA[KR]L[GQH]G[KR][TP]TYE[KE]YLG[TP][EQ]YV[TA]

Figure-2: Conserved pattern of Lactoferrin

Figure-3: Site of Block one

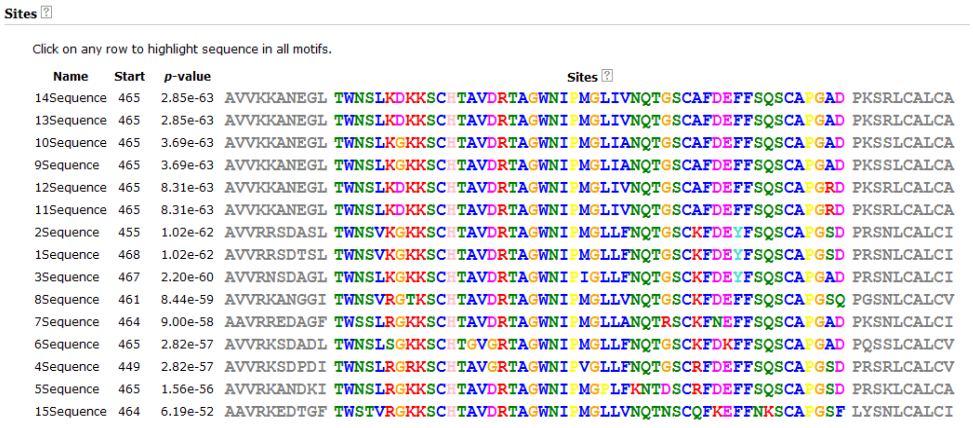


Figure-4: Block One Show the Motif Location in each Lactoferrin sequences

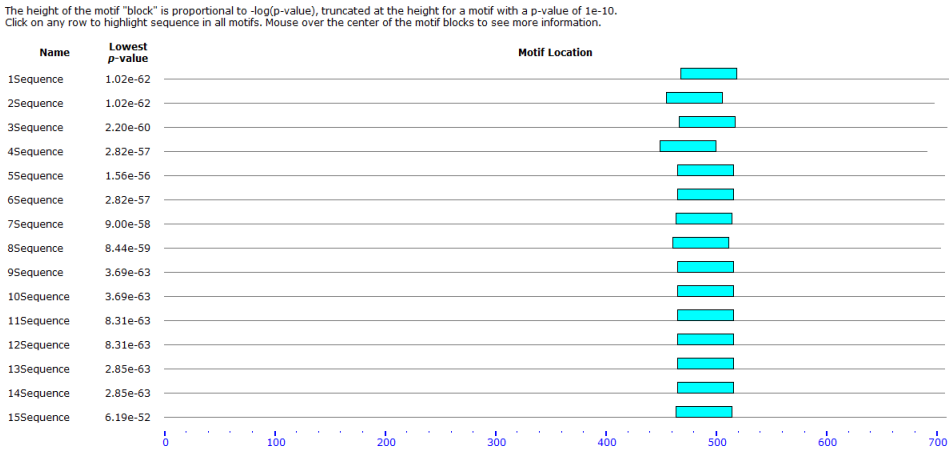


Figure-5: Site of Block Two

Sites

Click on any row to highlight sequence in all motifs.

Name	Start	*p*-value	Sites
14Sequence	107	1.52e-64	AEIYGTKESP QTHYYAVAVVKKGSNFQLDQLQGRKSCHTGLGRSAGWIIPMGILRPYLSW TESLEPLQG
13Sequence	107	1.52e-64	AEIYGTKESP QTHYYAVAVVKKGSNFQLDQLQGRKSCHTGLGRSAGWIIPMGILRPYLSW TESLEPLQG
11Sequence	107	1.57e-62	AEIYGTKESP QTHYYAVAVVKKGSNFQLDQLQGRKSCHTGLGRSAGWVIPMGILRPYLSW TESLEPLQG
9Sequence	107	2.47e-62	AEIYGTEKSP QTHYYAVAVVKKGSNFQLDQLQGQKSCHMGLGRSAGWNIPMGILRPFLSW TESAEPLQG
12Sequence	107	9.48e-62	AEIYGTKEAP QTHYYAVAVVKKGSNFQLDQLQGRKSCHTGLARSAGWIIPMGILRPYLSW TESLEPLQG
10Sequence	107	4.55e-59	AEIYGTEKSP QTHYYAVAVVKKGSNFQLDQLQGQKSCHAGLGRSAGWNIPVGILRPFLSW TESAEPLQG
5Sequence	107	2.42e-58	AEVYGTENNP QTHYYAVAIAKKGTNFQLNQLQGLKSCHTGLGRSAGWNIPMGLLRPFLDW TGPPEPLQK
6Sequence	107	1.22e-57	AEVYQTRGKP QTRYYAVAVVKKGSGFQLNQLQGVKSCHTGLGRSAGWNIPIGTLRPYLNW TGPPEPLQK
8Sequence	103	2.70e-56	AEIYGTEENP QTYYYAVAVVKKGFNFQLNQLQGRKSCHTGLGRSAGWNIPIGLLRRFLDW AGPPEPLQK
4Sequence	107	1.49e-55	VEVYGTEAKP QTHYYAVAVARKGSSFQLNQLRGRKSCHTGFGRTAGWNIPIGTLRPFLNW TGPPKPLSA
1Sequence	108	2.04e-55	AEVYGTERQP RTHYYAVAVVKKGGSFQLNELQGLKSCHTGLRRTAGWNVPIGTLRPFLNW TGPPEPIEA
3Sequence	107	3.80e-55	AEVYGTEEKP RTHYYAVAVVKKGGRFQLNELQGLKSCHTGLNRTAGWIVPIGMLRPFLNW TGPPEAIEA
7Sequence	106	7.07e-55	AEVYGTKEQP RTHYYAVAVVKNSSNFHLNQLQGLRSCHTGIGRSAGWKIPIGTLRPYLNW NGPPASLEE
2Sequence	95	1.45e-54	AEVYGTERQP RTHYYAVAVVKKGGSFQLNELQGLKSCHTGLRRTAGWTVPIGTLRPFLNW TGPPEPIEA
15Sequence	106	2.47e-45	AEVYGTKEKP RIHYYAVAVVKNSSDIRLNQLQGLKSCHAGFDTSAGWIAPLGALRPYLNW DEKSVSLEE

Figure-6: Block Two Show the Motif Location in each Lactoferrin sequences

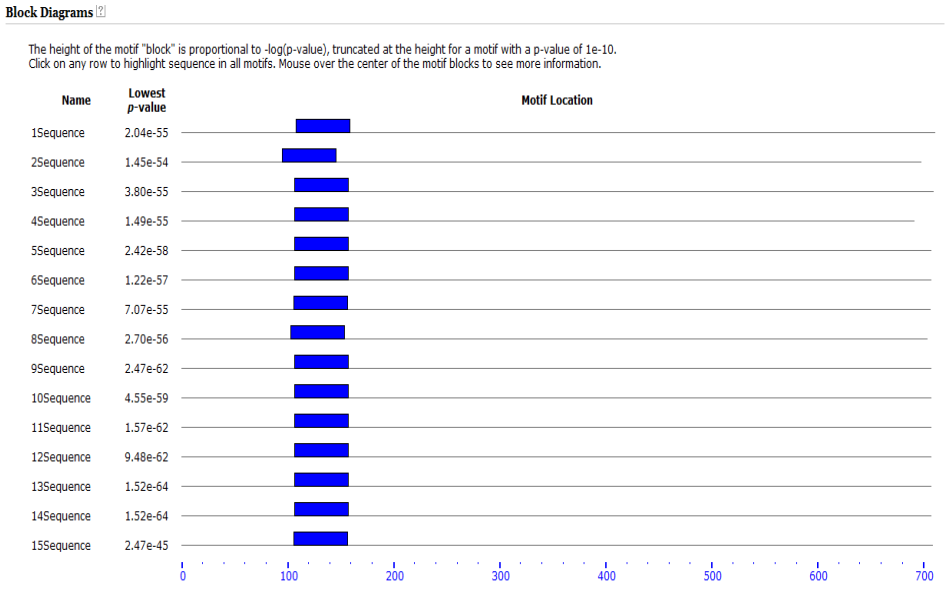


Figure-7: Site of Block Three

Figure-8: Block Three Show the Motif Location in each Lactoferrin sequences

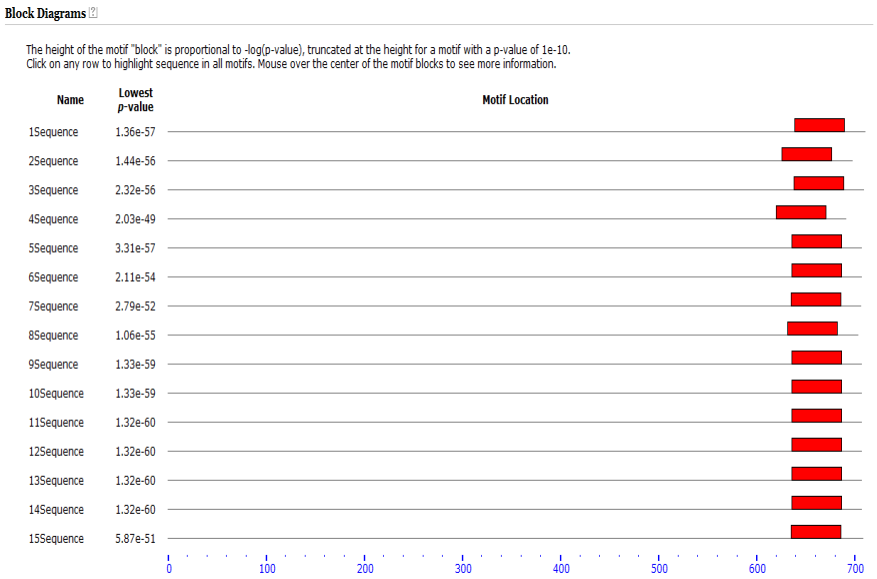


Figure- 9: Combined block diagram show the Motif location of each block

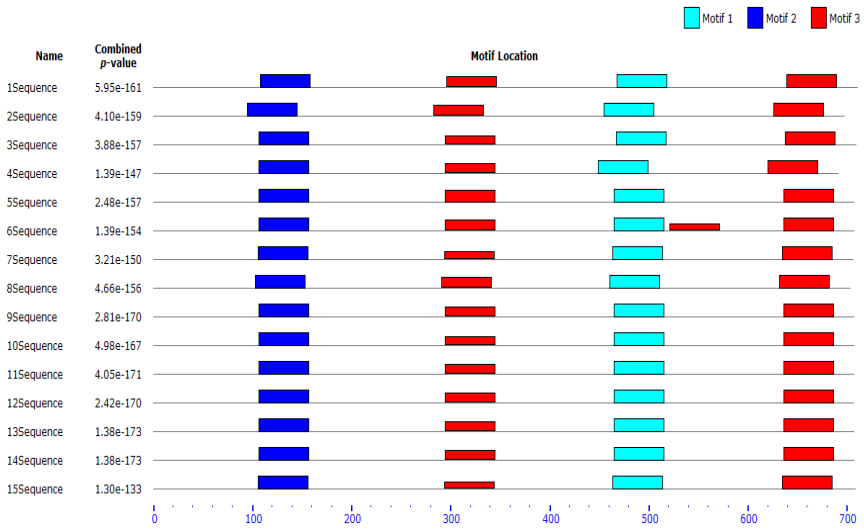
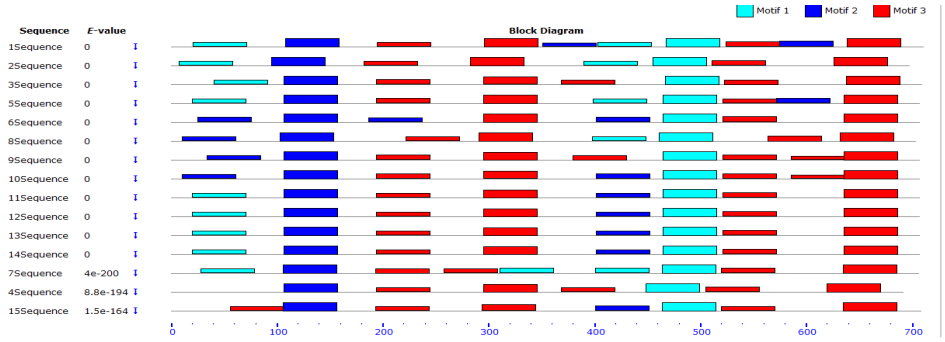


Figure-10: Submit motifs in MAST

Motif	Width	Best possible match	Similarity 1	2	3
1	50	TWNSLKGKKSCXTAVDRTAGWNIXMGLIFNQTGSCKFDEFFSQSCAPGSD	-	0.42	0.20
2	50	QTHYYAVAVVKKGSNFQLNQLQGRKSCXTGLGRSAGWIIXMGILRPYLNW	0.42	-	0.15
3	50	FGKNGKNCXDKFCLFKSETKNLLFNDNTECLAKLHGRTTYEEYLGTEYVT	0.20	0.15	-

Figure-11:Block diagram show top scoring sequence



Conclusion

At last,Our research explain that using multiple motifs gives much better database search results than using single motifs. Multiple motifs contain more information quality of the protein family than do single motifs.MEME easily accessible tobiologists who want to analyze their own sequences ofnucleic acids and proteins.This study suggest that lactoferrin sequence of different species show same conserver region. When these species were originate from different ancestor and different region. On the base of our result we assumed that lactoferrin illustrate similar function in all species. The MAST result also prove that lactoferrin have play same function in these species but their percentage is differ species to species. *Homo sapiens* contain higher concentration of lactoferrinwith 0.0 E-value and sequence15 *Rattusnorvegicus*with E-value 1.5e-164 have low concentration.As stated above MEMEdescribes that although lactoferrin present in different source of origin,they contains common patterns of amino acids. It is again noted that motifs may overlap withone another due to the reason of commonconsensus patterns.

Reference

1. Sorensen, M., Sorensen, S.P.L., (1939)**.** The proteins in whey. C. R. Lab. Carlsberg, 23, 55–99.
2. Adlerova, L., Bartoskova, A. and Faldyna, M. (2008), Lactoferrin: a review. Veterinarni Medicina, 53: 457-468.
3. Baker, E.N., Baker, H.M., Kidd, R.D.,(2002). Lactoferrin and transferrin: Functional variations on a common structural framework. Biochem.Cell Biol.,80, 27-34.PMID: 11908640.
4. Baker, E.N., Baker, H.M., (2005). Molecular structure, binding properties and dynamics of lactoferrin. Cellular and Molecular Life Sciences, 62, 2531–2539.PMID: 16261257.
5. Bailey, T.L., Elkan, C.,(1995). The value of prior knowledge in discovering motifs with MEME.Molecular biology. 3, 21–29. PMID: 7584439.
6. Bailey TL, Williams N, Misleh C, Li WW.(2006) Meme: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.,***34, W369**–W373. PMID: 16845028.
7. Javed A. Khan, Pravindra Kumar, M. Paramasivam, Raghvendra S. Yadav, Mohan S. Sahani, Sujata Sharma, Srinivasan A. , Singh, Tej P., (2001). Camel lactoferrin, a transferrin-cum-lactoferrin: crystal structure of camel apolactoferrin at 2.6 Å resolution and structural basis of its dual role. Journal of Moecluar Biology, Vol.309, issue 3,751-761.
8. http://www.ncbi.nlm.nih.gov/
9. http://meme.nbcr.net/meme4_6_1/cgi-bin/meme.cgi
10. http://meme.nbcr.net/meme4_6_1/cgi-bin/mast.cgi