BIT2009

PROTEIN ENGINEERING AND DESIGN

PROJECT REPORT

PHYLOGENETIC RELATIONSHIP OF CONSERVED AND NON-CONSERVED PROTEINS ACROSS MODEL ORGANISMS

PROJECT GUIDE: Dr. G. Jayaraman

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

November,2019

Place: Vellore Date: 08/11/19

CONTENT

Abstract

Conserved and non-conserved sequences in protein signify the relation of different species at a molecular level. Analysis of these sequences is used not only in phylogenetic studies but also in epitope mapping and vaccine design. The most common technique used for studies of sequences across a broad range of species is multiple sequence analysis (MSA). When proteins are used in analytical studies along with sequences, structures are also considered as primary evidence. This study compares both structures and sequences of one conserved and one non-conserved protein from different species across a broad range of phylogeny. Histone, protein responsible for DNA packaging, is the subject of the study. Histone H4 and Histone H1 are considered as most and least conserved among all histone proteins respectively. This study is to provide the idea on how the sequence of protein and their structures have changed over time and adapted to the environment.

Key words: Phylogenetic, Multiple sequence alignment, Conserved protein, Non-conserved protein, Histone

Objective

1) Brief idea on conserved and non-conserved proteins; their contribution to evolutionary biology.

2) Structural and functional comparison of a conserved protein and a non-conserved protein from selected species.

3) Similarities between protein characteristics of closely related species and distant species.

Introduction

Conserved and non-conserved proteins:

Phylogenesis is the evolutionary development of particular features of an organism as well as the diversification of organisms, which sets them apart from one another.

Conserved proteins: These proteins are present in all genomes sequenced so far, from archaea and bacteria to man. So, one can say, that there is no life on earth without this small set of proteins. The genes encoding these proteins are usually essential.

Highly conserved proteins are useful in constructing phylogenesis because proteins are made by genes. These genes contain the genetic material of the organism. If the genetic material is in good condition, it is easier to study and the relationship of the particular organism with others is easily studied.

Histones as conserved proteins:

Histone proteins are among the most highly conserved proteins in eukaryotes, emphasizing their important role in the biology of the nucleus. In contrast mature sperm cells largely use protamines to package their genomic DNA, most likely because this allows them to achieve an even higher packaging ratio.

Classes and histone variants:

Five major families of histones exist: H1/H5, H2A, H2B, H3, and H4.Histones H2A, H2B, H3 and H4 are known as the core histones, while histones H1/H5 are known as the linker histones.

The core histones all exist as dimers, which are similar in that they all possess the histone fold domain: three alpha helices linked by two loops. It is this helical structure that allows for interaction between distinct dimers, particularly in a head-tail fashion (also called the handshake motif).The resulting four distinct dimers then come together to form one octameric nucleosome core, approximately 63 Angstroms in diameter (a solenoid (DNA)-like particle). Around 146 base pairs (bp) of DNA wrap around this core particle 1.65 times in a left-handed super-helical turn to give a particle of around 100 Angstroms across. The linker histone H1 binds the nucleosome at the entry and exit sites of the DNA, thus locking the DNA into place and allowing the formation of higher order structure. The most basic such formation is the 10 nm fibre or beads on a string conformation. This involves the wrapping of DNA around nucleosomes with approximately 50 base pairs of DNA separating each pair of nucleosomes (also referred to as

linker DNA). Higher-order structures include the 30 nm fibre (forming an irregular zigzag) and 100 nm fibre, these being the structures found in normal cells. During mitosis and meiosis, the condensed chromosomes are assembled through interactions between nucleosomes and other regulatory proteins.

Histones are subdivided into canonical replication-dependent histones that are expressed during the S-phase of cell cycle and replication-independent histone variants, expressed during the whole cell cycle. In animals, genes encoding canonical histones are typically clustered along the chromosome, lack introns and use a stem loop structure at the 3' end instead of a polyA tail. Genes encoding histone variants are usually not clustered, have introns and their mRNAs are regulated with polyA tails. Complex multicellular organisms typically have a higher number of histone variants providing a variety of different functions. Recent data are accumulating about the roles of diverse histone variants highlighting the functional links between variants and the delicate regulation of organism development. Histone variants from different organisms, their classification and variant specific features can be found in "HistoneDB 2.0 - Variants" database.

Multiple sequence alignment

Sequence alignment is a way of arranging sequences of DNA, RNA, or proteins in order to distinguish regions of similarity. A multiple sequence alignment (MSA) is a sequence alignment of three or more biological sequences such as protein, DNA, or RNA. Typically it is implied that the set of sequences share an evolutionary relationship, which means they are all descendents from a common ancestor. These regions may correspond to functional, structural, or evolutionary relationships between the sequences. Alignments can reflect a degree of evolutionary change between sequences that are descendants from a common ancestor.

Phylogenetics is the study of evolutionary relatedness amongst organisms. The genetic relationships between species can be represented using phylogenetic trees. Phylogenetics is the study of evolutionary relatedness amongst organisms. The genetic relationships between species can be represented using phylogenetic trees. A way of visually representing these relationships is with a phylogenetic tree.

These trees show the evolutionary relationships amongst various species by way of common ancestors. Each node in the tree with descendants represents the most recent common ancestor of the descendents.

SSIM

The *structural similarity (SSIM)* index is a method for predicting the perceived quality of digital television and cinematic pictures, as well as other kinds of digital images and videos.

SSIM is used for measuring the similarity between two images. The SSIM index is a full reference metric; in other words, the measurement or prediction of image quality is based on an initial uncompressed or distortion-free image as reference. SSIM is designed to improve on traditional methods such as peak signal-to-noise ratio (PSNR) and mean squared error (MSE).

In bioinformatics, a sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. Aligned sequences of nucleotide or amino acid residues are typically represented as rows within a matrix. Gaps are inserted between the residues so that identical or similar characters are aligned in successive columns. Sequence alignments are also used for non-biological sequences, such as calculating the distance cost between strings in a natural language or in financial data.

In this project we are giving a brief idea on conserved and non-conserved proteins and their contribution to evolutionary biology, their structural and functional comparison of a conserved and a non-conserved protein from selected specie and similarities in protein characteristics of closely related species and distant species.

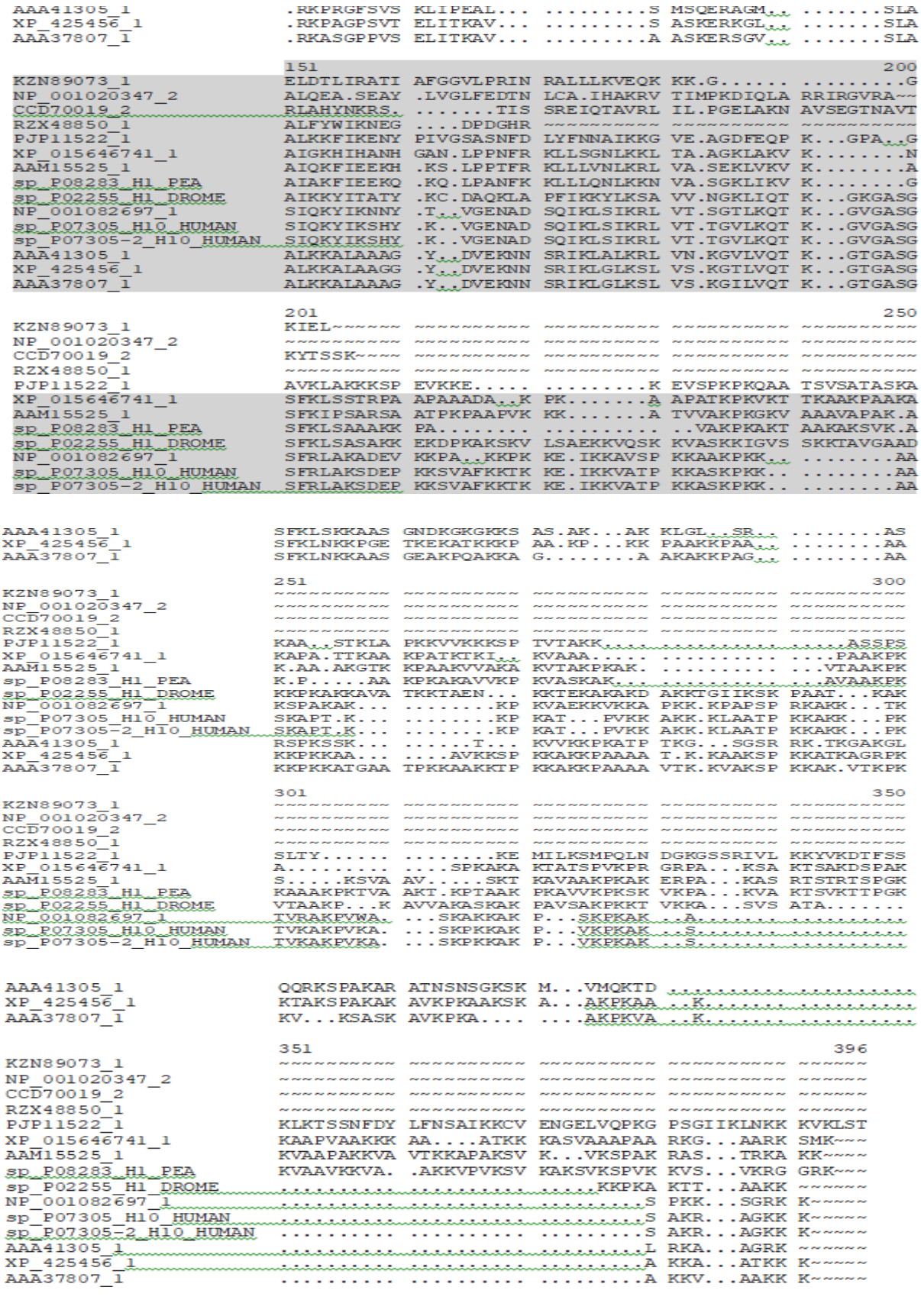
Result

Non-conserved Protein H1 Histone:

After retrieving protein sequences from NCBI, UniProt and PDB, they were subjected to MSA using ClustalW and MSF. The identity matrix and phylogenetic tree obtained are also included here.

MSA using MSF

```
!!AA_MULTIPLE_ALIGNMENT 1.0
 squid.msf  MSF: 396  Type: P  September 06, 2019 18:01  Check: 5691  ..

 Name: KZN89073_1              Len:    396  Check:  574  Weight: -1.00
 Name: NP_001020347_2          Len:    396  Check: 7871  Weight: -1.00
 Name: CCD70019_2              Len:    396  Check: 7855  Weight: -1.00
 Name: RZX48850_1              Len:    396  Check: 8422  Weight: -1.00
 Name: PJP11522_1              Len:    396  Check: 8045  Weight: -1.00
 Name: XP_015646741_1          Len:    396  Check: 1168  Weight: -1.00
 Name: AAM15525_1              Len:    396  Check: 5759  Weight: -1.00
 Name: sp_P08283_H1_PEA        Len:    396  Check: 9459  Weight: -1.00
 Name: sp_P02255_H1_DROME      Len:    396  Check: 8863  Weight: -1.00
 Name: NP_001082697_1          Len:    396  Check: 3160  Weight: -1.00
 Name: sp_P07305_H10_HUMAN     Len:    396  Check: 1153  Weight: -1.00
 Name: sp_P07305-2_H10_HUMAN   Len:    396  Check:  935  Weight: -1.00
 Name: AAA41305_1              Len:    396  Check: 4093  Weight: -1.00
 Name: XP_425456_1             Len:    396  Check: 5829  Weight: -1.00
 Name: AAA37807_1              Len:    396  Check: 2505  Weight: -1.00

//

                              1                                                50
KZN89073_1                    ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~M PGGKGKSIGG
NP_001020347_2                MARTK..... .......... ...QTA.... ..........  ..RKS..TGG
CCD70019_2                    ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
RZX48850_1                    ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
PJP11522_1                    MAPKK..... .......... ......STTK T..T.....S KGKKPATSKG
XP_015646741_1                MATEEVVPE. ..VPVTEVEA AAAEEAVEET TAAEEKAAKP AKEKKKAGRP
AAM15525_1                    MSIEE..... ENVPTTVD.S GAADTTVKSP .......... EK.KPAAKGG
sp_P08283_H1_PEA              MATEEPIVAV ETVPEPIV.T E..PTTITEP E.......VP EKEEPKAEVE
sp_P02255_H1_DROME            MSDSA..... .......... ...VATSASP VAAPP...A. TVEKKVVQ..
NP_001082697_1                ~~MTE..... .......... ...NSA.......... PAAKPR....
sp_P07305_H10_HUMAN           MTENS..... .......... ...TSA.......... PAAKPK....
sp_P07305-2_H10_HUMAN         ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~

AAA41305_1                    MSETA..... .......... ...PAA..SS TLVP....A. PVEKPATK..
XP_425456_1                   MSETA..... .......... ...PVA..AP AVSA..... PGAKAAAK..
AAA37807_1                    MSEAA..... .......... ...PAA...P AAAP..... PAEKAPAK..

                              51                                               100
KZN89073_1                    KGGKGDAAGK A.QKSHSAKA GL....QFPC G.....RVKR FLKNNTQNKM
NP_001020347_2                KAPRKQLATK AARKSAPATG GVKKPHRYRP GTVALREIRR YQKST.....
CCD70019_2                    ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
RZX48850_1                    ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
PJP11522_1                    KEKS......T.TSKAAIKKT TAKK..... .......... ..........
XP_015646741_1                PKEKKEAKPA KEKKVKEAKA KKPR..... .......... ..........
AAM15525_1                    KSKKTTTAKA TKKPVKAAAP TKKK..... .......... ..........
sp_P08283_H1_PEA              KTKK...A.. ..KGSKPKKA SKPR..... .......... ..........
sp_P02255_H1_DROME            .........K KASGSAGTKA KKAS..... .......... ..........
NP_001082697_1                ......... ........RS KASK..... .......... ..........
sp_P07305_H10_HUMAN           ......... ........RA KASK..... .......... ..........
sp_P07305-2_H10_HUMAN         ~~~~~~~~~~ ~~~~~~~~~~ ~MTK..... .......... ..........
AAA41305_1                    .........R R....GKKPG MATA..... .......... ..........
XP_425456_1                   .........K P....KKAAG GAKA..... .......... ..........
AAA37807_1                    .........K K....AAKKP AGVR..... .......... ..........

                              101                                              150
KZN89073_1                    RVGAKAAVYV TAVLEYLTAE VLELAGNAAK DLKVKRITPR HLQLAIRGDE
NP_001020347_2                ELLIRKLPFQ RLVREIA... .........Q DFKTDLRFQS S......AVM
CCD70019_2                    ~~~~~~~~~~M SIMNSFV... .........N DVFERIAA.. .......EAS
RZX48850_1                    ~~~~MTIYIT EIITGAI... .........Y .......... .......TV
PJP11522_1                    .EEASSKSYR ELIIEGL... .........T ALKERKGS.. .......SRP
XP_015646741_1                .VAAAHPPYA EMIMEAI... .........V ALKERTGS.. .......SSQ
AAM15525_1                    .TTSSHPTYE EMIKDAI... .........V TLKERTGS.. .......SQY
sp_P08283_H1_PEA              .NPASHPTYE EMIKDAI... .........V SLKEKNGS.. .......SQY
sp_P02255_H1_DROME            .ATPSHPPTQ QMVDASI... .........K NLKERGGS.. .......SLL
NP_001082697_1                .KSTDHPKYS DMILDAV... .........Q AEKSRSGS.. .......SRQ
sp_P07305_H10_HUMAN           .KSTDHPKYS DMIVAAI... .........Q AEKNRAGS.. .......SRQ
sp_P07305-2_H10_HUMAN         .KSTDHPKYS DMIVAAI... .........Q AEKNRAGS.. .......SRQ
```

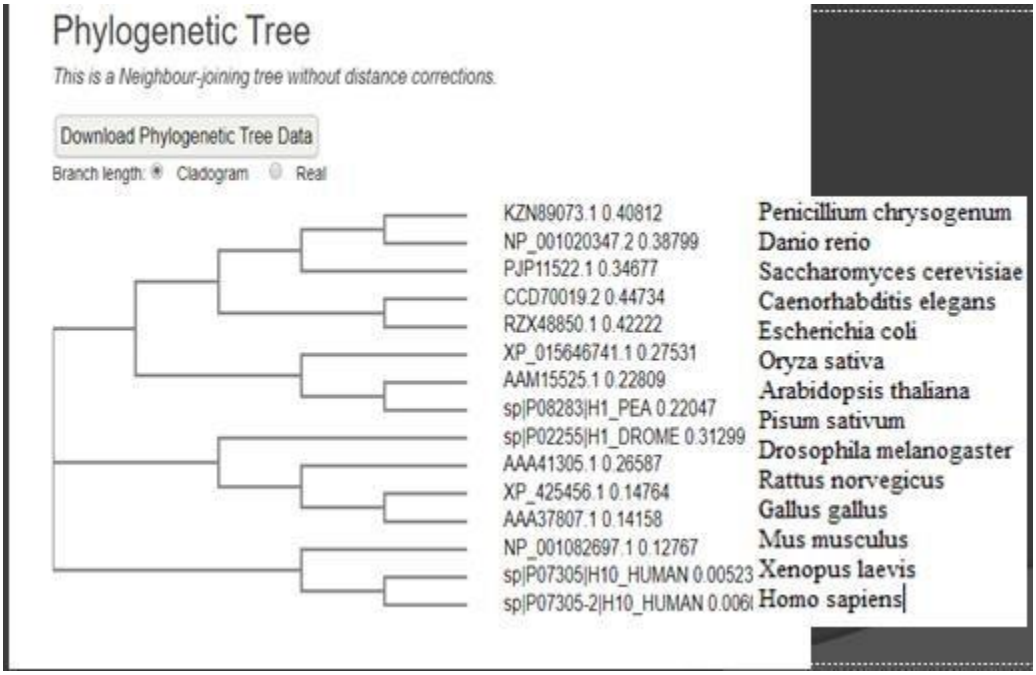

AAA41305_1 .RKPRGFSVS KLIPEAL... .....S MSQERAGM.. .....SLA
XP_425456_1 .RKPAGPSVT ELITKAV... .....S ASKERKGL.. .....SLA
AAA37807_1 .RKASGPPVS ELITKAV... .....A ASKERSGV.. .....SLA
151 200
KZN89073_1 ELDTLIRATI AFGGVLPRIN RALLLKVEQK KK.G..... .....G
NP_001020347_2 ALQEA.SEAY .LVGLFEDTN LCA.IHAKRV TIMPKDIQLA RRIRGVRA~~
CCD70019_2 RLAHYNKRS. .....TIS SREIQTAVRL IL.PGELAKN AVSEGTNAVT
RZX48850_1 ALFYWIKNEG ....DPDGHR ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
PJP11522_1 ALKKFIKENY PIVGSASNFD LYFNNAIKKG VE.AGDFEQP K...GPA..G
XP_015646741_1 AIGKHIHANH GAN.LPPNFR KLLSGNLKKL TA.AGKLAKV K.....N
AAM15525_1 AIQKFIEEKH .KS.LPPTFR KLLLVNLKRL VA.SEKLVKV K.....A
sp_P08283_H1_PEA AIAKFIEEKQ .KQ.LPANFK KLLLQNLKKN VA.SGKLIKV K.....G
sp_P02255_H1_DROME AIKKYITATY .KC.DAQKLA PFIKKYLKSA VV.NGKLIQT K...GKGASG
NP_001082697_1 SIQKYIKNNY .T..VGENAD SQIKLSIKRL VT.SGTLKQT K...GVGASG
sp_P07305_H10_HUMAN SIQKYIKSHY .K..VGENAD SQIKLSIKRL VT.TGVLKQT K...GVGASG
sp_P07305-2_H10_HUMAN SIQKYIKSHY .K..VGENAD SQIKLSIKRL VT.TGVLKQT K...GVGASG
AAA41305_1 ALKKALAAAG .Y..DVEKNN SRIKLALKRL VN.KGVLVQT K...GTGASG
XP_425456_1 ALKKALAAGG .Y..DVEKNN SRIKLGLKSL VS.KGTLVQT K...GTGASG
AAA37807_1 ALKKALAAAG .Y..DVEKNN SRIKLGLKSL VS.KGILVQT K...GTGASG
201 250
KZN89073_1 KIEL~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
NP_001020347_2 ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
CCD70019_2 KYTSSK~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
RZX48850_1 ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
PJP11522_1 AVKLAKKKSP EVKKE..... .....K EVSPKPKQAA TSVSATASKA
XP_015646741_1 SFKLSSTRPA APAAADA..K PK.....A APATKPKVKT TKAAKPAAKA
AAM15525_1 SFKIPSARSA ATPKPAAPVK KK.....A TVVAKPKGKV AAAVAPAK.A
sp_P08283_H1_PEA SFKLSAAAKK PA..... ....VAKPKAKT AAKAKSVK.A
sp_P02255_H1_DROME SFKLSASAKK EKDPKAKSKV LSAEKKVQSK KVASKKIGVS SKKTAVGAAD
NP_001082697_1 SFRLAKADEV KKPA..KKPK KE.IKKAVSP KKAAKPKK.. .....AA
sp_P07305_H10_HUMAN SFRLAKSDEP KKSVAFKKTK KE.IKKVATP KKASKPKK.. .....AA
sp_P07305-2_H10_HUMAN SFRLAKSDEP KKSVAFKKTK KE.IKKVATP KKASKPKK.. .....AA
AAA41305_1 SFKLSKKAAS GNDKGKGKKS AS.AK...AK KLGL..SR.. .....AS
XP_425456_1 SFKLNKKPGE TKEKATKKKP AA.KP...KK PAAKKPAA.. .....AA
AAA37807_1 SFKLNKKAAS GEAKPQAKKA G.....A AKAKKPAG.. .....AA
251 300
KZN89073_1 ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
NP_001020347_2 ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
CCD70019_2 ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
RZX48850_1 ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
PJP11522_1 KAA..STKLA PKKVVKKKSP TVTAKK..... .....ASSPS
XP_015646741_1 KAPA.TTKAA KPATKTKI.. KVAAA..... ....PAAKPK
AAM15525_1 K.AA.AKGTK KPAAKVVAKA KVTAKPKAK. ....VTAAKPK
sp_P08283_H1_PEA K.P.....AA KPKAKAVVKP KVASKAK.. ....AVAAKPK
sp_P02255_H1_DROME KKPKAKKAVA TKKTAEN... KKTEKAKAKD AKKTGIIKSK PAAT...KAK
NP_001082697_1 KSPAKAK... .....KP KVAEKKVKKA PKK.KPAPSP RKAKK...TK
sp_P07305_H10_HUMAN SKAPT.K... .....KP KAT...PVKK AKK.KLAATP KKAKK...PK
sp_P07305-2_H10_HUMAN SKAPT.K... .....KP KAT...PVKK AKK.KLAATP KKAKK...PK
AAA41305_1 RSPKSSK... .....T... KVVKKPKATP TKG...SGSR RK.TKGAKGL
XP_425456_1 KKPKKAA... ....AVKKSP KKAKKPAAAA T.K.KAAKSP KKATKAGRPK
AAA37807_1 KKPKKATGAA TPKKAAKKTP KKAKKPAAAA VTK.KVAKSP KKAK.VTKPK
301 350
KZN89073_1 ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
NP_001020347_2 ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
CCD70019_2 ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
RZX48850_1 ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
PJP11522_1 SLTY..... .....KE MILKSMPQLN DGKGSSRIVL KKYVKDTFSS
XP_015646741_1 A..... ....SPKAKA KTATSPVKPR GRPA...KSA KTSAKDSPAK
AAM15525_1 S.....KSVA AV.....SKT KAVAAKPKAK ERPA...KAS RTSTRTSPGK
sp_P08283_H1_PEA KAAAKPKTVA AKT.KPTAAK PKAVVKPKSK VKPA...KVA KTSVKTTPGK
sp_P02255_H1_DROME VTAAKP...K AVVAKASKAK PAVSAKPKKT VKKA...SVS ATA.....
NP_001082697_1 TVRAKPVWA. ...SKAKKAK P...SKPKAK ..A.....
sp_P07305_H10_HUMAN TVKAKPVKA. ...SKPKKAK P...VKPKAK ..S.....
sp_P07305-2_H10_HUMAN TVKAKPVKA. ...SKPKKAK P...VKPKAK ..S.....
AAA41305_1 QQRKSPAKAR ATNSNSGKSK M...VMQKTD .....
XP_425456_1 KTAKSPAKAK AVKPKAAKSK A...AKPKAA ..K.....
AAA37807_1 KV...KSASK AVKPKA.... ....AKPKVA ..K.....
351 396
KZN89073_1 ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~
NP_001020347_2 ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~
CCD70019_2 ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~
RZX48850_1 ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~
PJP11522_1 KLKTSSNFDY LFNSAIKKCV ENGELVQPKG PSGIIKLNKK KVKLST
XP_015646741_1 KAAPVAAKKK AA....ATKK KASVAAAPAA RKG...AARK SMK~~~
AAM15525_1 KVAAPAKKVA VTKKAPAKSV K...VKSPAK RAS...TRKA KK~~~~
sp_P08283_H1_PEA KVAAVKKVA. .AKKVPVKSV KAKSVKSPVK KVS...VKRG GRK~~~
sp_P02255_H1_DROME ....KKPKA KTT...AAKK ~~~~~~
NP_001082697_1 ....S PKK...SGRK K~~~~~
sp_P07305_H10_HUMAN ....S AKR...AGKK K~~~~~
sp_P07305-2_H10_HUMAN ....S AKR...AGKK K~~~~~
AAA41305_1 ....L RKA...AGRK ~~~~~~
XP_425456_1 ....A KKA...ATKK K~~~~~
AAA37807_1 ....A KKV...AAKK K~~~~~

MSA using ClustalW

```
CLUSTAL O(1.2.4) multiple sequence alignment


KZN89073.1----------------------------------------------------------MPGGKGKSIGGKGGKGDAAGK
NP_001020347.2             MARTK------------------QTA----------------RKS--TGGKAPRKQLATK
CCD70019.2                 ------------------------------------------------------------
RZX48850.1                 ------------------------------------------------------------
PJP11522.1                 MAPKK---------------------STTKT--T-----SKGKKPATSKGKEKS------
XP_015646741.1             MATEEVVPE---VPVTEVEAAAAEEAVEETTAAEEKAAKPAKEKKKAGRPPKEKKEAKPA
AAM15525.1                 MSIEE-----ENVPTTVD-SGAADTTVKSP----------EK-KPAAKGGKSKKTTTAKA
sp|P08283|H1_PEA           MATEEPIVAVETVPEPIV-TE--PTTITEPE-------VPEKEEPKAEVEKTKK---A--
sp|P02255|H1_DROME         MSDSA------------------VATSASPVAAPP---A-TVEKKVVQ-----------K
NP_001082697.1             --MTE------------------NSA---------------PAAKPR--------------
sp|P07305|H10_HUMAN        MTENS------------------TSA---------------PAAKPK--------------
sp|P07305-2|H10_HUMAN      ------------------------------------------------------------
AAA41305.1                 MSETA------------------PAA--SSTLVP----A-PVEKPATK-----------R
XP_425456.1                MSETA------------------PVA--APAVSA------PGAKAAAK-----------K
AAA37807.1                 MSEAA------------------PAA---PAAAP------PAEKAPAK-----------K


KZN89073.1                 A-QKSHSAKAGL----QFPCG-----RVKRFLKNNTQNKMRVGAKAAVYVTAVLEYLTAE
NP_001020347.2             AARKSAPATGGVKKPHRYRPGTVALREIRRYQKST-----ELLIRKLPFQRLVREIA---
CCD70019.2                 ----------------------------------------------MSIMNSFV---
RZX48850.1                 ---------------------------------------------MTIYITEIITGAI---
PJP11522.1                 -TSKAAIKKTTAKK---------------------------EEASSKSYRELIIEGL---
XP_015646741.1             KEKKVKEAKAKKPR---------------------------VAAAHPPYAEMIMEAI---
AAM15525.1                 TKKPVKAAAPTKKK---------------------------TTSSHPTYEEMIKDAI---
sp|P08283|H1_PEA           --KGSKPKKASKPR---------------------------NPASHPTYEEMIKDAI---
sp|P02255|H1_DROME         KASGSAGTKAKKAS---------------------------ATPSHPPTQQMVDASI---
NP_001082697.1             --------RSKASK---------------------------KSTDHPKYSDMILDAV---
sp|P07305|H10_HUMAN        --------RAKASK---------------------------KSTDHPKYSDMIVAAI---
sp|P07305-2|H10_HUMAN      -----------MTK---------------------------KSTDHPKYSDMIVAAI---
AAA41305.1                 R----GKKPGMATA---------------------------RKPRGFSVSKLIPEAL---
XP_425456.1                P----KKAAGGAKA---------------------------RKPAGPSVTELITKAV---
AAA37807.1                 K----AAKKPAGVR---------------------------RKASGPPVSELITKAV---
                                                                                 :

KZN89073.1                 VLELAGNAAKDLKVKRITPRHLQLAIRGDEELDTLIRATIAFGGVLPRINRALLLKVEQK
NP_001020347.2             ---------QDFKTDLRFQSS------AVMALQEA-SEAY-LVGLFEDTNLCA-IHAKRV
CCD70019.2                 ---------NDVFERIAA---------EASRLAHYNKRS--------TISSREIQTAVRL
RZX48850.1                 ---------Y------------------TVALFYWIKNEG----DPDGHR----------
PJP11522.1                 ---------TALKERKGS---------SRPALKKFIKENYPIVGSASNFDLYFNNAIKKG
XP_015646741.1             ---------VALKERTGS---------SSQAIGKHIHANHGAN-LPPNFRKLLSGNLKKL
AAM15525.1                 ---------VTLKERTGS---------SQYAIQKFIEEKH-KS-LPPTFRKLLLVNLKRL
sp|P08283|H1_PEA           ---------VSLKEKNGS---------SQYAIAKFIEEKQ-KQ-LPANFKKLLLQNLKKN
sp|P02255|H1_DROME         ---------KNLKERGGS---------SLLAIKKYITATY-KC-DAQKLAPFIKKYLKSA
NP_001082697.1             ---------QAEKSRSGS---------SRQSIQKYIKNNY-T--VGENADSQIKLSIKRL
sp|P07305|H10_HUMAN        ---------QAEKNRAGS---------SRQSIQKYIKSHY-K--VGENADSQIKLSIKRL
sp|P07305-2|H10_HUMAN      ---------QAEKNRAGS---------SRQSIQKYIKSHY-K--VGENADSQIKLSIKRL
AAA41305.1                 ---------SMSQERAGM---------SLAALKKALAAAG-Y--DVEKNNSRIKLALKRL
XP_425456.1                ---------SASKERKGL---------SLAALKKALAAGG-Y--DVEKNNSRIKLGLKSL
AAA37807.1                 ---------AASKERSGV---------SLAALKKALAAAG-Y--DVEKNNSRIKLGLKSL
                                                       :

KZN89073.1                 KK-G---------------GKIEL-------------------------------------
NP_001020347.2             TIMPKDIQLARRIRGVRA-----------------------------------------
CCD70019.2                 IL-PGELAKNAVSEGTNAVTKYTSSK---------------------------------
RZX48850.1                 ------------------------------------------------------------
PJP11522.1                 VE-AGDFEQPK---GPA--GAVKLAKKKSPEVKKE--------------KEVSPKPKQAA
XP_015646741.1             TA-AGKLAKVK--------NSFKLSSTRPAAPAAADA--KPK       AAPATKPKVKT
AAM15525.1                 VA-SEKLVKVK--------ASFKIPSARSAATPKPAAPVKKK       ATVVAKPKGKV
sp|P08283|H1_PEA           VA-SGKLIKVK--------GSFKLSAAAKKPA--------------------VAKPKAKT
sp|P02255|H1_DROME         VV-NGKLIQTK   GKGASGSFKLSASAKKEKDPKAKSKVLSAEKKVQSKKVASKKIGVS
NP_001082697.1             VT-SGTLKQTK---GVGASGSFRLAKADEVKKPA--KKPKKE-IKKAVSPKKAAKPKK--
sp|P07305|H10_HUMAN        VT-TGVLKQTK---GVGASGSFRLAKSDEPKKSVAFKKTKKE-IKKVATPKKASKPKK--
sp|P07305-2|H10_HUMAN      VT-TGVLKQTK---GVGASGSFRLAKSDEPKKSVAFKKTKKE-IKKVATPKKASKPKK--
AAA41305.1                 VN-KGVLVQTK---GTGASGSFKLSKKAASGNDKGKGKKSAS-AK---AKKLGL--SR--
XP_425456.1                VS-KGTLVQTK---GTGASGSFKLNKKPGETKEKATKKKPAA-KP---KKPAAKKPAA--
AAA37807.1                 VS-KGILVQTK---GTGASGSFKLNKKAASGEAKPQAKKAG--------AAKAKKPAG--
```

```
KZN89073.1                ------------------------------------------------------------
NP_001020347.2            ------------------------------------------------------------
CCD70019.2                ------------------------------------------------------------
RZX48850.1                ------------------------------------------------------------
PJP11522.1                TSVSATASKAKAA--STKLAPKKVVKKKSPTVTAKK-------------------ASSPS
XP_015646741.1            TKAAKPAAKAKAPA-TTKAAKPATKTKI--KVAAA-------------------PAAKPK
AAM15525.1                AAAVAPAK-AK-AA-AKGTKKPAAKVVAKAKVTAKPKAK---------------VTAAKPK
sp|P08283|H1_PEA          AAKAKSVK-AK-P-----AAKPKAKAVVKPKVASKAK-----------------AVAAKPK
sp|P02255|H1_DROME        SKKTAVGAADKKPKAKKAVATKKTAEN---KKTEKAKAKDAKKTGIIKSKPAAT---KAK
NP_001082697.1            --------AAKSPAKAK-----------KPKVAEKKVKKAPKK-KPAPSPRKAKK---TK
sp|P07305|H10_HUMAN       --------AASKAPT-K-----------KPKAT---PVKKAKK-KLAATPKKAKK---PK
sp|P07305-2|H10_HUMAN     --------AASKAPT-K-----------KPKAT---PVKKAKK-KLAATPKKAKK---PK
AAA41305.1                --------ASRSPKSSK---------T---KVVKKPKATPTKG---SGSRRK-TKGAKGL
XP_425456.1               --------AAKKPKKAA-------AVKKSPKKAKKPAAAAT-K-KAAKSPKKATKAGRPK
AAA37807.1----------------------AAKKPKKATGAATPKKAAKKTPKKAKKPAAAAVTK-KVAKSPKKAK-VTKPK


KZN89073.1                ------------------------------------------------------------
NP_001020347.2            ------------------------------------------------------------
CCD70019.2                ------------------------------------------------------------
RZX48850.1                ------------------------------------------------------------
PJP11522.1                SLTY--------------KEMILKSMPQLNDGKGSSRIVLKKYVKDTFSSKLKTSSNFDY
XP_015646741.1            A-------------SPKAKAKTATSPVKPRGRPA---KSAKTSAKDSPAKKAAPVAAKKK
AAM15525.1                S-----KSVAAV-----SKTKAVAAKPKAKERPA---KASRTSTRTSPGKKVAAPAKKVA
sp|P08283|H1_PEA          KAAAKPKTVAAKT-KPTAAKPKAVVKPKSKVKPA---KVAKTSVKTTPGKKVAAVKKVA-
sp|P02255|H1_DROME        VTAAKP---KAVVAKASKAKPAVSAKPKKTVKKA---SVSATA-----------------
NP_001082697.1            TVRAKPVWA----SKAKKAKP---SKPKAK--A---------------------------
sp|P07305|H10_HUMAN       TVKAKPVKA----SKPKKAKP---VKPKAK--S---------------------------
sp|P07305-2|H10_HUMAN     TVKAKPVKA----SKPKKAKP---VKPKAK--S---------------------------
AAA41305.1                QQRKSPAKARATNSNSGKSKM---VMQKTD------------------------------
XP_425456.1               KTAKSPAKAKAVKPKAAKSKA---AKPKAA--K---------------------------
AAA37807.1                KV---KSASKAVKPKA--------AKPKVA--K---------------------------


KZN89073.1                -------------------------------------
NP_001020347.2            -------------------------------------
CCD70019.2                -------------------------------------
RZX48850.1                -------------------------------------
PJP11522.1                LFNSAIKKCVENGELVQPKGPSGIIKLNKKKVKLST
XP_015646741.1            AA----ATKKKASVAAAPAARKG---AARKSMK---
AAM15525.1                VTKKAPAKSVK---VKSPAKRAS---TRKAKK----
sp|P08283|H1_PEA          -AKKVPVKSVKAKSVKSPVKKVS---VKRGGRK---
sp|P02255|H1_DROME        ---------------KKPKAKTT---AAKK------
NP_001082697.1            -------------------SPKK---SGRKK-----
sp|P07305|H10_HUMAN       -------------------SAKR---AGKKK-----
sp|P07305-2|H10_HUMAN     -------------------SAKR---AGKKK-----
AAA41305.1                -------------------LRKA---AGRK------
XP_425456.1               -------------------AKKA---ATKKK-----
AAA37807.1                -------------------AKKV---AAKKK-----
```

Phylogenetic Tree:



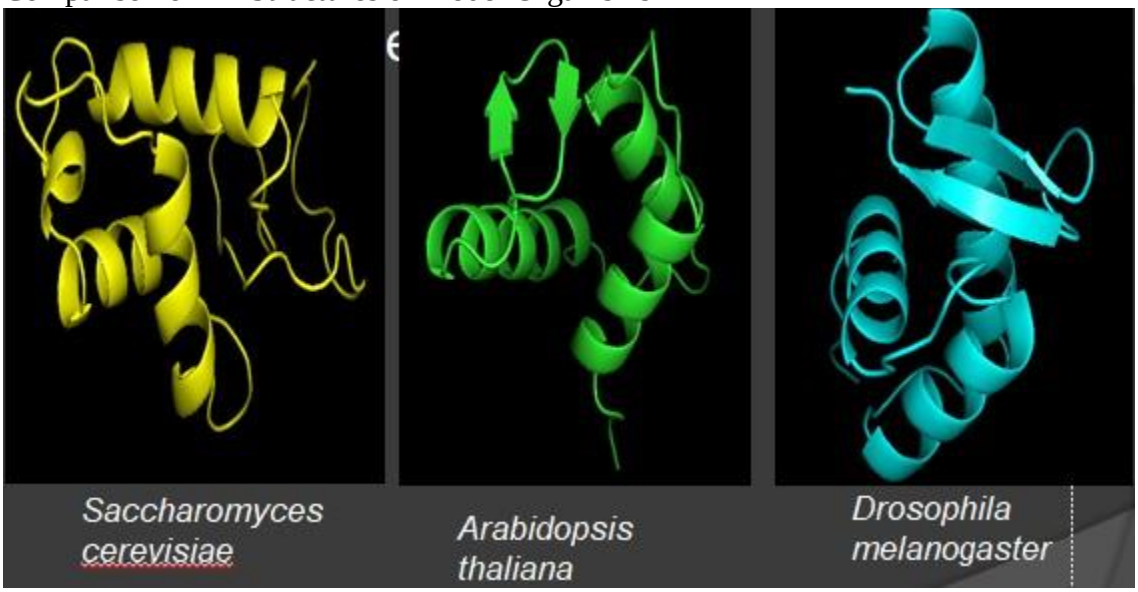
Identity Matrix:

```
#
#
#  Percent Identity  Matrix - created by Clustal2.1
#
#
```

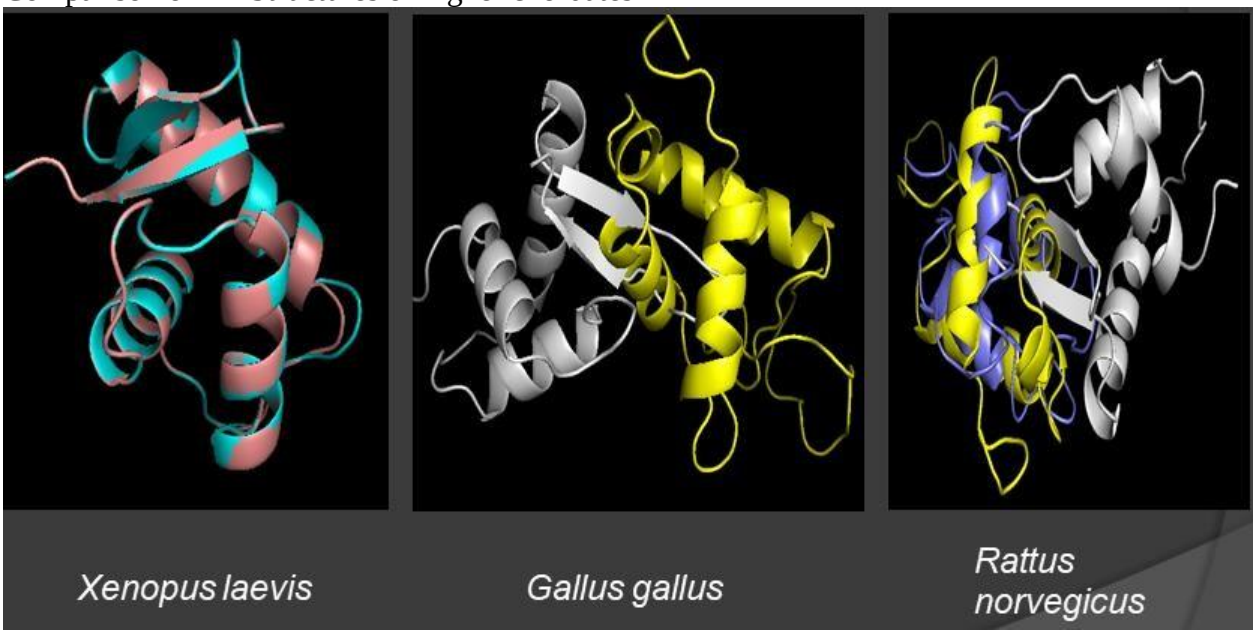
1: KZN89073.1	100.00	20.39	8.00	6.25	18.68	19.59	14.89	18.89	16.67	13.70	15.07	10.94	12.50	16.25	15.00
2: NP_001020347.2	20.39	100.00	12.28	6.45	23.96	13.73	16.83	12.63	15.79	17.07	16.67	15.71	20.00	21.11	18.89
3: CCD70019.2	8.00	12.28	100.00	13.04	11.29	13.56	11.86	11.86	12.50	18.75	17.19	17.19	21.88	20.31	20.31
4: RZX48850.1	6.25	6.45	13.04	100.00	18.75	25.00	25.00	21.88	12.50	15.62	15.62	15.62	18.75	28.12	25.00
5: PJP11522.1	18.68	23.96	11.29	18.75	100.00	25.11	25.11	24.89	23.50	30.61	30.14	30.30	26.11	30.12	29.34
6: XP_015646741.1	19.59	13.73	13.56	25.00	25.11	100.00	44.84	46.06	27.36	33.11	31.58	31.11	21.89	27.43	35.06
7: AAM15525.1	14.89	16.83	11.86	25.00	25.11	44.84	100.00	55.14	30.00	34.39	31.85	33.33	22.81	30.56	31.67
8: sp|P08283|H1_PEA	18.89	12.63	11.86	21.88	24.89	46.06	55.14	100.00	37.50	36.00	34.23	35.61	23.31	28.65	33.53
9: sp|P02255|H1_DROME	16.67	15.79	12.50	12.50	23.50	27.36	30.00	37.50	100.00	35.29	38.50	40.00	33.66	39.42	43.84
10: NP_001082697.1	13.70	17.07	18.75	15.62	30.61	33.11	34.39	36.00	35.29	100.00	73.68	73.71	34.07	41.05	40.45
11: sp|P07305|H10_HUMAN	15.07	16.67	17.19	15.62	30.14	31.58	31.85	34.23	38.50	73.68	100.00	98.87	32.97	39.47	38.20
12: sp|P07305-2|H10_HUMAN	10.94	15.71	17.19	15.62	30.30	31.11	33.33	35.61	40.00	73.71	98.87	100.00	32.73	39.31	38.51
13: AAA41305.1	12.50	20.00	21.88	18.75	26.11	21.89	22.81	23.31	33.66	34.07	32.97	32.73	100.00	52.91	49.74
14: XP_425456.1	16.25	21.11	20.31	28.12	30.12	27.43	30.56	28.65	39.42	41.05	39.47	39.31	52.91	100.00	71.08
15: AAA37807.1	15.00	18.89	20.31	25.00	29.34	35.06	31.67	33.53	43.84	40.45	38.20	38.51	49.74	71.08	100.00

Structural Analysis:

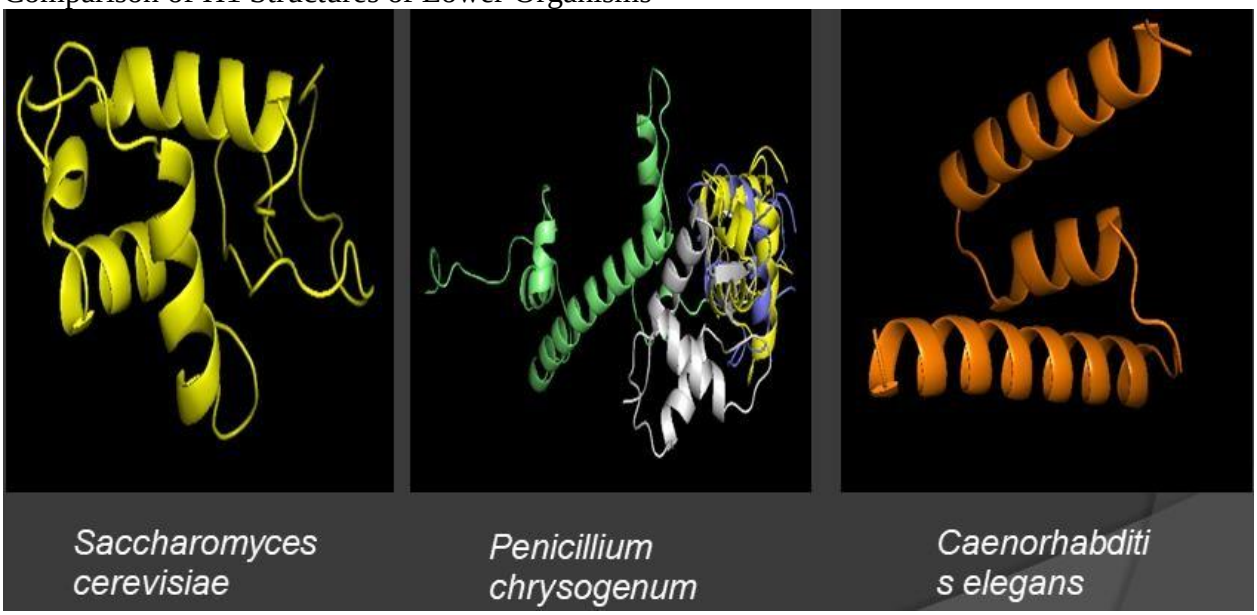
Comparison of H1 Structures of Model Organisms



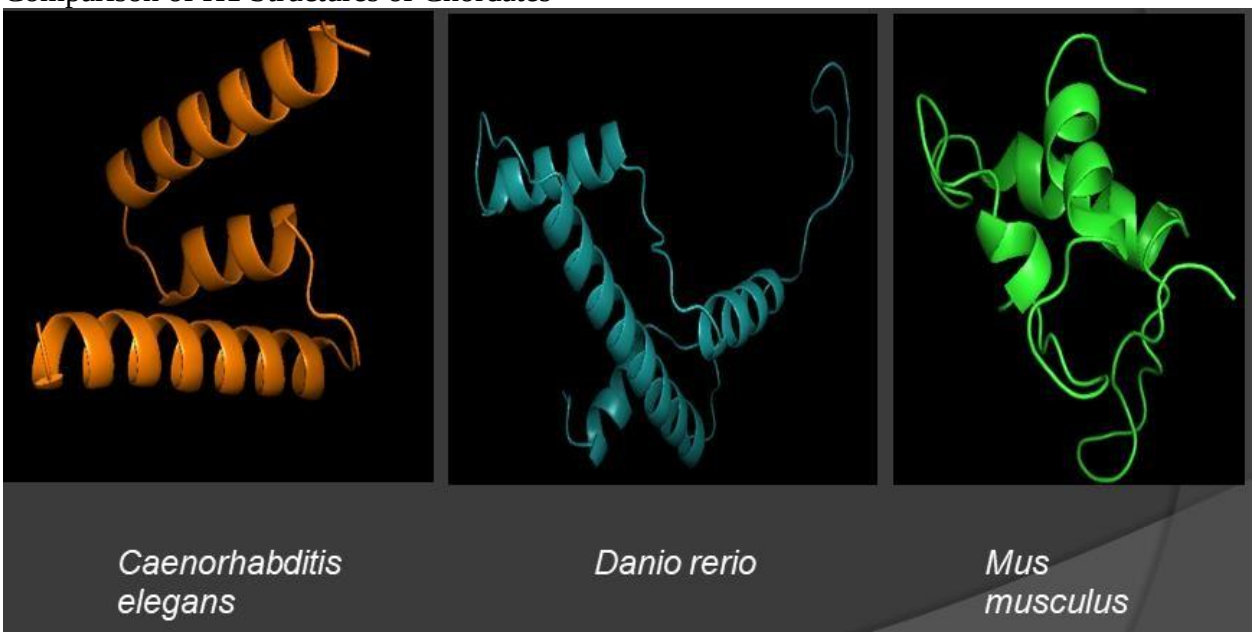
Comparison of H1 Structures of higher chordates



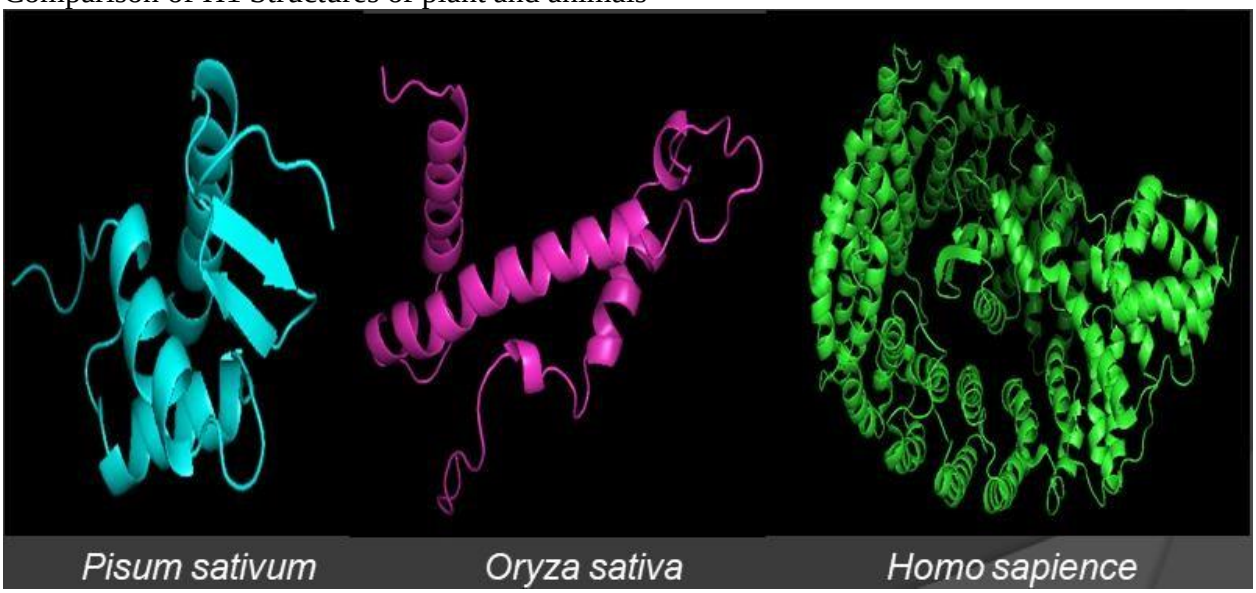
Comparison of H1 Structures of Lower Organisms



Comparison of H1 Structures of Chordates

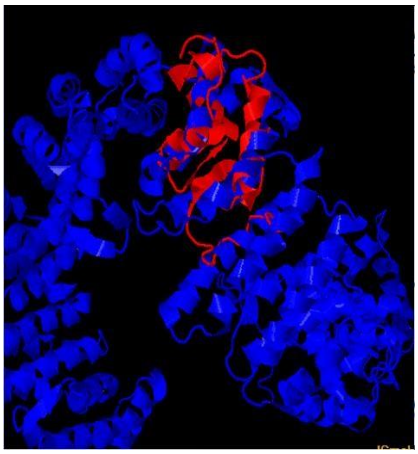


Comparison of H1 Structures of plant and animals



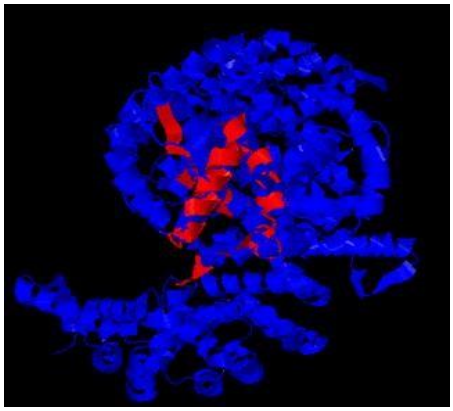
Superimposition of Human and Yeast

```
Name of Chain_1: A399181
Name of Chain_2: B399181
Length of Chain_1:  873 residues
Length of Chain_2:   92 residues

Aligned length=   65, RMSD=   4.88, Seq_ID=n_identical/n_aligned= 0.077
TM-score= 0.06194 (if normalized by length of Chain_1)
TM-score= 0.35128 (if normalized by length of Chain_2)
(You should use TM-score normalized by length of the reference protein)
```



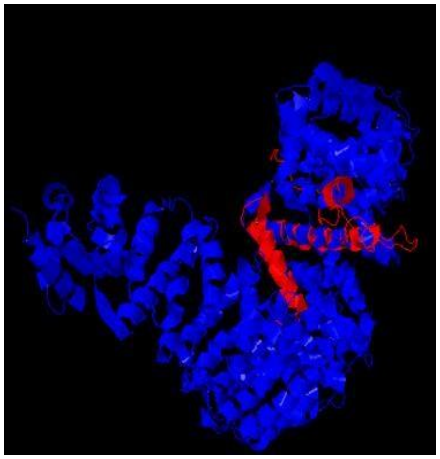
Superimposition of Human and *A. thaliana*

```
Name of Chain_1: A180298
Name of Chain_2: B180298
Length of Chain_1:  873 residues
Length of Chain_2:   70 residues

Aligned length=   51, RMSD=   3.71, Seq_ID=n_identical/n_aligned= 0.078
TM-score= 0.05216 (if normalized by length of Chain_1)
TM-score= 0.39895 (if normalized by length of Chain_2)
```



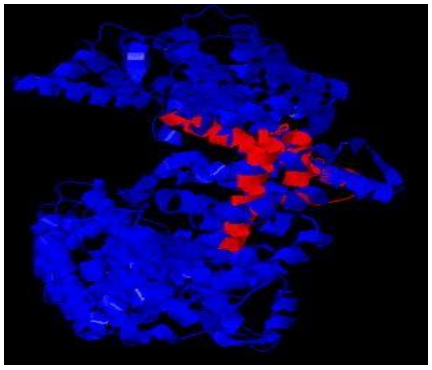
Superimposition of Human and Rice

```
Name of Chain_1: A934281
Name of Chain_2: B934281
Length of Chain_1:  873 residues
Length of Chain_2:  103 residues

Aligned length=   63, RMSD=   4.65, Seq_ID=n_identical/n_aligned= 0.079
TM-score= 0.06161 (if normalized by length of Chain_1)
TM-score= 0.35981 (if normalized by length of Chain_2)
```



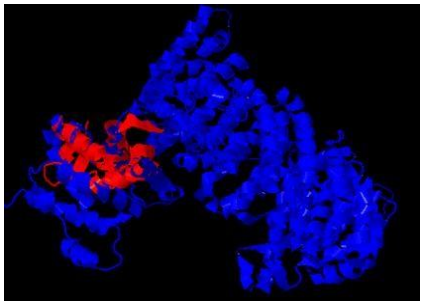
Superimposition of Human and *C. elegans*

```
Name of Chain_1: A416365
Name of Chain_2: B416365
Length of Chain_1:  873 residues
Length of Chain_2:   67 residues

Aligned length=   49, RMSD=   3.46, Seq_ID=n_identical/n_aligned= 0.082
TM-score= 0.05112 (if normalized by length of Chain_1)
TM-score= 0.46322 (if normalized by length of Chain_2)
```



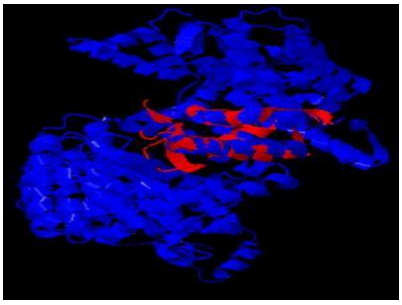
Superimposition of Human and Fruit fly

```
Name of Chain_1: A320088
Name of Chain_2: B320088
Length of Chain_1:  873 residues
Length of Chain_2:   76 residues

Aligned length=   55, RMSD=   4.10, Seq_ID=n_identical/n_aligned= 0.036
TM-score= 0.05505 (if normalized by length of Chain_1)
TM-score= 0.36224 (if normalized by length of Chain_2)
```



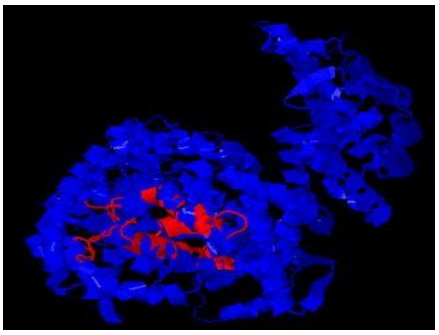
Superimposition of Human and Frog

```
Name of Chain_1: A282950
Name of Chain_2: B282950
Length of Chain_1:  873 residues
Length of Chain_2:   76 residues

Aligned length=   62, RMSD=   4.70, Seq_ID=n_identical/n_aligned= 0.065
TM-score= 0.05978 (if normalized by length of Chain_1)
TM-score= 0.36509 (if normalized by length of Chain_2)
```



Superimposition of Human and Rat

```
Name of Chain_1: A73000
Name of Chain_2: B73000
Length of Chain_1:  873 residues
Length of Chain_2:   75 residues

Aligned length=   53, RMSD=   3.72, Seq_ID=n_identical/n_aligned= 0.057
TM-score= 0.05443 (if normalized by length of Chain_1)
TM-score= 0.41870 (if normalized by length of Chain_2)
```



Conserved Protein H4 Histone:

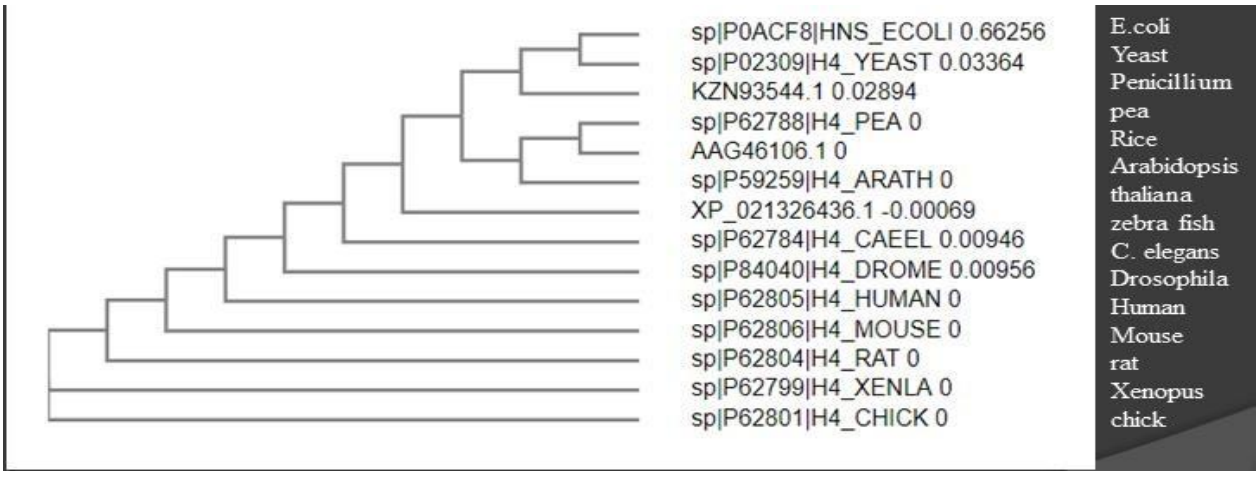
After retrieving protein sequences from NCBI, UniProt and PDB, they were subjected to MSA using ClustalW and MSF. The identity matrix and phylogenetic tree obtained are also included here.

MSA using MSF:

```
!!AA_MULTIPLE_ALIGNMENT 1.0
 squid.msf  MSF: 172  Type: P  October 19, 2019 15:01  Check: 7213  ..

 Name: sp_P0ACF8_HNS_ECOLI  Len:    172  Check:  492  Weight: -1.00
 Name: sp_P02309_H4_YEAST   Len:    172  Check: 5053  Weight: -1.00
 Name: KZN93544_1           Len:    172  Check: 3872  Weight: -1.00
 Name: sp_P62788_H4_PEA     Len:    172  Check: 3667  Weight: -1.00
 Name: AAG46106_1           Len:    172  Check: 3667  Weight: -1.00
 Name: sp_P59259_H4_ARATH   Len:    172  Check: 3667  Weight: -1.00
 Name: sp_P62784_H4_CAEEL   Len:    172  Check: 3224  Weight: -1.00
 Name: sp_P62805_H4_HUMAN   Len:    172  Check: 4091  Weight: -1.00
 Name: sp_P62806_H4_MOUSE   Len:    172  Check: 4091  Weight: -1.00
 Name: sp_P62804_H4_RAT     Len:    172  Check: 4091  Weight: -1.00
 Name: XP_021326436_1       Len:    172  Check: 9009  Weight: -1.00
 Name: sp_P62799_H4_XENLA   Len:    172  Check: 4091  Weight: -1.00
 Name: sp_P62801_H4_CHICK   Len:    172  Check: 4091  Weight: -1.00
 Name: sp_P84040_H4_DROME   Len:    172  Check: 4107  Weight: -1.00

//

                          1                                                50
sp_P0ACF8_HNS_ECOLI  MSE..... ..........  ..........  ..ALKIL.NN IRTL.....R
sp_P02309_H4_YEAST   ~~~~~~~~~~~~ ~~~~~MSGRGK GGKGLGKGGA KRHRKILRDN IQGITKPAIR
KZN93544_1           ~~~~~~~~~~~~ ~~~~~MSGRGK GGKGLGKGGA KRHRKILRDN IQGITKPAIR
sp_P62788_H4_PEA     ~~~~~~~~~~~~ ~~~~~MSGRGK GGKGLGKGGA KRHRKVLRDN IQGITKPAIR
AAG46106_1           ~~~~~~~~~~~~ ~~~~~MSGRGK GGKGLGKGGA KRHRKVLRDN IQGITKPAIR
sp_P59259_H4_ARATH   ~~~~~~~~~~~~ ~~~~~MSGRGK GGKGLGKGGA KRHRKVLRDN IQGITKPAIR
sp_P62784_H4_CAEEL   ~~~~~~~~~~~~ ~~~~~MSGRGK GGKGLGKGGA KRHRKVLRDN IQGITKPAIR
sp_P62805_H4_HUMAN   ~~~~~~~~~~~~ ~~~~~MSGRGK GGKGLGKGGA KRHRKVLRDN IQGITKPAIR
sp_P62806_H4_MOUSE   ~~~~~~~~~~~~ ~~~~~MSGRGK GGKGLGKGGA KRHRKVLRDN IQGITKPAIR
sp_P62804_H4_RAT     ~~~~~~~~~~~~ ~~~~~MSGRGK GGKGLGKGGA KRHRKVLRDN IQGITKPAIR
XP_021326436_1       MSRQALVESS SKLIMSGRGK GGKGLGKGGA KRHRKVLRDN IQGITKPAIR
sp_P62799_H4_XENLA   ~~~~~~~~~~~~ ~~~~~MSGRGK GGKGLGKGGA KRHRKVLRDN IQGITKPAIR
sp_P62801_H4_CHICK   ~~~~~~~~~~~~ ~~~~~MSGRGK GGKGLGKGGA KRHRKVLRDN IQGITKPAIR
sp_P84040_H4_DROME   ~~~~~~~~~~~~ ~~~~~MTGRGK GGKGLGKGGA KRHRKVLRDN IQGITKPAIR
```

```
                    51                                                  100
sp_P0ACF8_HNS_ECOLI AQARECTLET LEEMLEKLEV VVNERREEES AAAAEVEERT RKLQQYREML
sp_P02309_H4_YEAST  RLARRGGVKR ISG.....  .......... ...LIYEEVR AVLKSFLESV
KZN93544_1          RLARRGGVKR ISA.....  .......... ...MIYEETR GVLKTFLEGV
sp_P62788_H4_PEA    RLARRGGVKR ISG.....  .......... ...LIYEETR GVLKIFLENV
AAG46106_1          RLARRGGVKR ISG.....  .......... ...LIYEETR GVLKIFLENV
sp_P59259_H4_ARATH  RLARRGGVKR ISG.....  .......... ...LIYEETR GVLKIFLENV
sp_P62784_H4_CAEEL  RLARRGGVKR ISG.....  .......... ...LIYEETR GVLKVFLENV
sp_P62805_H4_HUMAN  RLARRGGVKR ISG.....  .......... ...LIYEETR GVLKVFLENV
sp_P62806_H4_MOUSE  RLARRGGVKR ISG.....  .......... ...LIYEETR GVLKVFLENV
sp_P62804_H4_RAT    RLARRGGVKR ISG.....  .......... ...LIYEETR GVLKVFLENV
XP_021326436_1      RLARRGGVKR ISG.....  .......... ...LIYEETR GVLKVFLENV
sp_P62799_H4_XENLA  RLARRGGVKR ISG.....  .......... ...LIYEETR GVLKVFLENV
sp_P62801_H4_CHICK  RLARRGGVKR ISG.....  .......... ...LIYEETR GVLKVFLENV
sp_P84040_H4_DROME  RLARRGGVKR ISG.....  .......... ...LIYEETR GVLKVFLENV

                    101                                                 150
sp_P0ACF8_HNS_ECOLI IADGIDPNEL LNSLAAVKSG TKAKRAQRPA KYSYVDENGE TKTWTGQGRT
sp_P02309_H4_YEAST  IRDSVTYTEH ..........  ..AKRKT...  ..........  ...V.....T
KZN93544_1          IRDAVTYTEH ..........  ..AKRKT...  ..........  ...V.....T
sp_P62788_H4_PEA    IRDAVTYTEH ..........  ..ARRKT...  ..........  ...V.....T
AAG46106_1          IRDAVTYTEH ..........  ..ARRKT...  ..........  ...V.....T
sp_P59259_H4_ARATH  IRDAVTYTEH ..........  ..ARRKT...  ..........  ...V.....T
sp_P62784_H4_CAEEL  IRDAVTYCEH ..........  ..AKRKT...  ..........  ...V.....T
sp_P62805_H4_HUMAN  IRDAVTYTEH ..........  ..AKRKT...  ..........  ...V.....T
sp_P62806_H4_MOUSE  IRDAVTYTEH ..........  ..AKRKT...  ..........  ...V.....T
sp_P62804_H4_RAT    IRDAVTYTEH ..........  ..AKRKT...  ..........  ...V.....T
XP_021326436_1      IRDAVTYTEH ..........  ..AKRKT...  ..........  ...V.....T
sp_P62799_H4_XENLA  IRDAVTYTEH ..........  ..AKRKT...  ..........  ...V.....T
sp_P62801_H4_CHICK  IRDAVTYTEH ..........  ..AKRKT...  ..........  ...V.....T
sp_P84040_H4_DROME  IRDAVTYTEH ..........  ..AKRKT...  ..........  ...V.....T

                    151                 172
sp_P0ACF8_HNS_ECOLI PAVIKKAMDE QGKSLDDFLI KQ
sp_P02309_H4_YEAST  SLDVVYALKR QGRTLYGFGG ~~
KZN93544_1          SLDVVYALKR QGRTLYGFGG ~~
sp_P62788_H4_PEA    AMDVVYALKR QGRTLYGFGG ~~
AAG46106_1          AMDVVYALKR QGRTLYGFGG ~~
sp_P59259_H4_ARATH  AMDVVYALKR QGRTLYGFGG ~~
sp_P62784_H4_CAEEL  AMDVVYALKR QGRTLYGFGG ~~
sp_P62805_H4_HUMAN  AMDVVYALKR QGRTLYGFGG ~~
sp_P62806_H4_MOUSE  AMDVVYALKR QGRTLYGFGG ~~
sp_P62804_H4_RAT    AMDVVYALKR QGRTLYGFGG ~~
XP_021326436_1      AMDVVYALKR QGRTLYGFGG ~~
sp_P62799_H4_XENLA  AMDVVYALKR QGRTLYGFGG ~~
sp_P62801_H4_CHICK  AMDVVYALKR QGRTLYGFGG ~~
sp_P84040_H4_DROME  AMDVVYALKR QGRTLYGFGG ~~
```

Phylogenetic Tree:

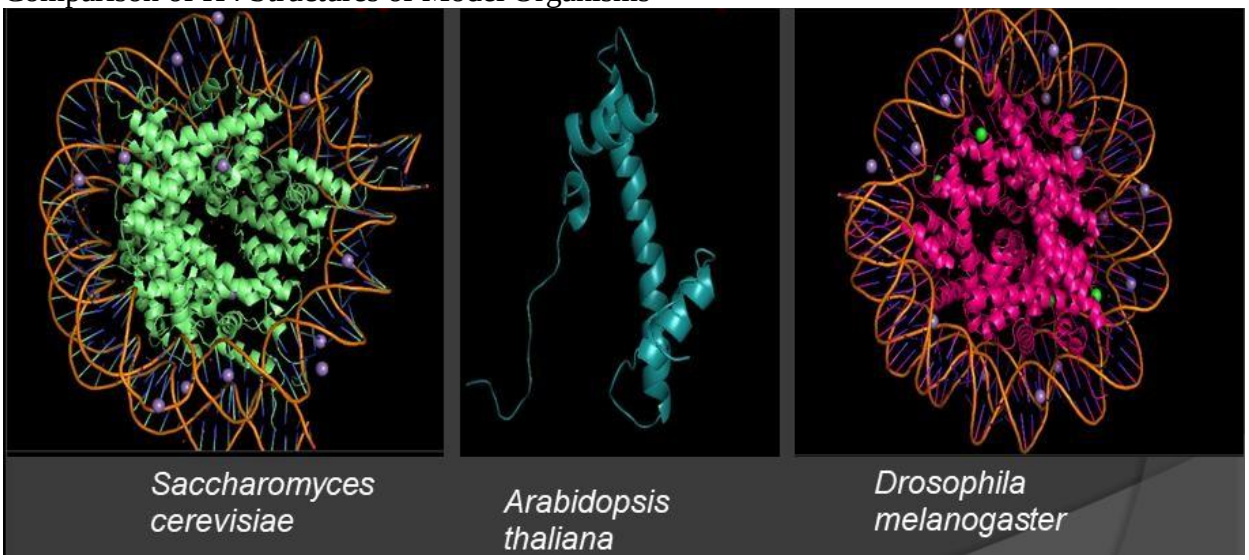


Identity Matrix

1: sp|P0ACF8|HNS_ECOLI	100.00	30.38	30.38	27.85	27.85	27.85	29.11	29.11	29.11	29.11	30.49	29.11	29.11	29.11
2: sp|P02309|H4_YEAST	30.38	100.00	93.20	91.26	91.26	91.26	91.26	92.23	92.23	92.23	92.23	92.23	92.23	91.26
3: KZN93544.1	30.38	93.20	100.00	92.23	92.23	92.23	92.23	93.20	93.20	93.20	93.20	93.20	93.20	92.23
4: sp|P62788|H4_PEA	27.85	91.26	92.23	100.00	100.00	100.00	97.09	98.06	98.06	98.06	98.06	98.06	98.06	97.09
5: AAG46106.1	27.85	91.26	92.23	100.00	100.00	100.00	97.09	98.06	98.06	98.06	98.06	98.06	98.06	97.09
6: sp|P59259|H4_ARATH	27.85	91.26	92.23	100.00	100.00	100.00	97.09	98.06	98.06	98.06	98.06	98.06	98.06	97.09
7: sp|P62784|H4_CAEEL	29.11	91.26	92.23	97.09	97.09	97.09	100.00	99.03	99.03	99.03	99.03	99.03	99.03	98.06
8: sp|P62805|H4_HUMAN	29.11	92.23	93.20	98.06	98.06	98.06	99.03	100.00	100.00	100.00	100.00	100.00	100.00	99.03
9: sp|P62806|H4_MOUSE	29.11	92.23	93.20	98.06	98.06	98.06	99.03	100.00	100.00	100.00	100.00	100.00	100.00	99.03
10: sp|P62804|H4_RAT	29.11	92.23	93.20	98.06	98.06	98.06	99.03	100.00	100.00	100.00	100.00	100.00	100.00	99.03
11: XP_021326436.1	30.49	92.23	93.20	98.06	98.06	98.06	99.03	100.00	100.00	100.00	100.00	100.00	100.00	99.03
12: sp|P62799|H4_XENLA	29.11	92.23	93.20	98.06	98.06	98.06	99.03	100.00	100.00	100.00	100.00	100.00	100.00	99.03
13: sp|P62801|H4_CHICK	29.11	92.23	93.20	98.06	98.06	98.06	99.03	100.00	100.00	100.00	100.00	100.00	100.00	99.03
14: sp|P84040|H4_DROME	29.11	91.26	92.23	97.09	97.09	97.09	98.06	99.03	99.03	99.03	99.03	99.03	99.03	100.00

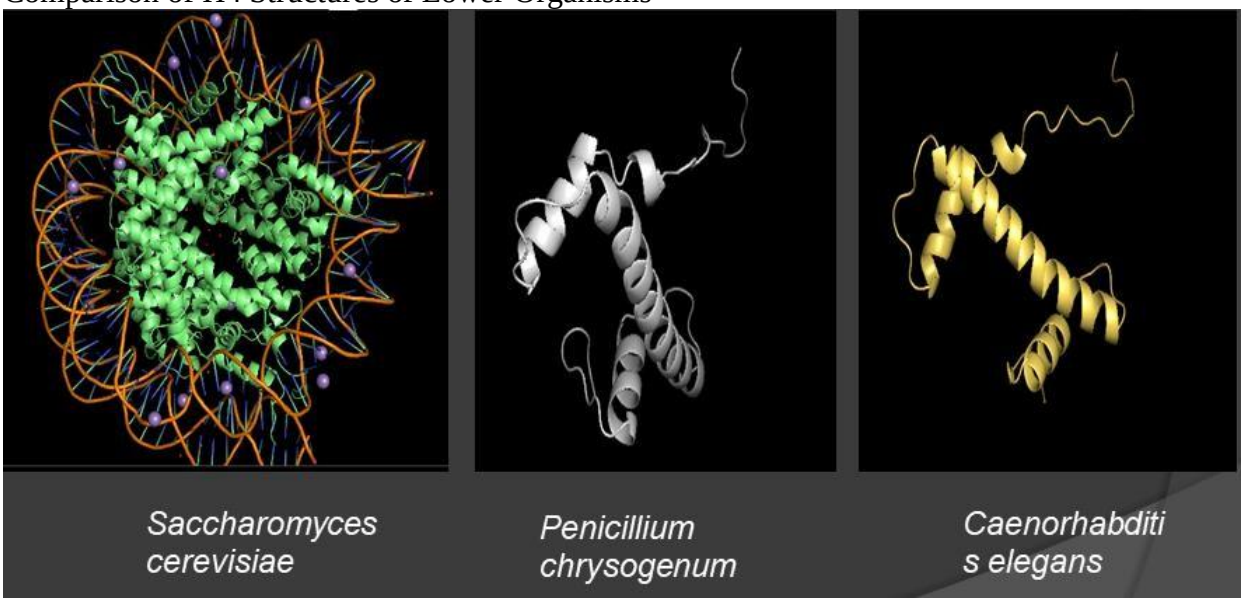
Structural Analysis:

Comparison of H4 Structures of Model Organisms

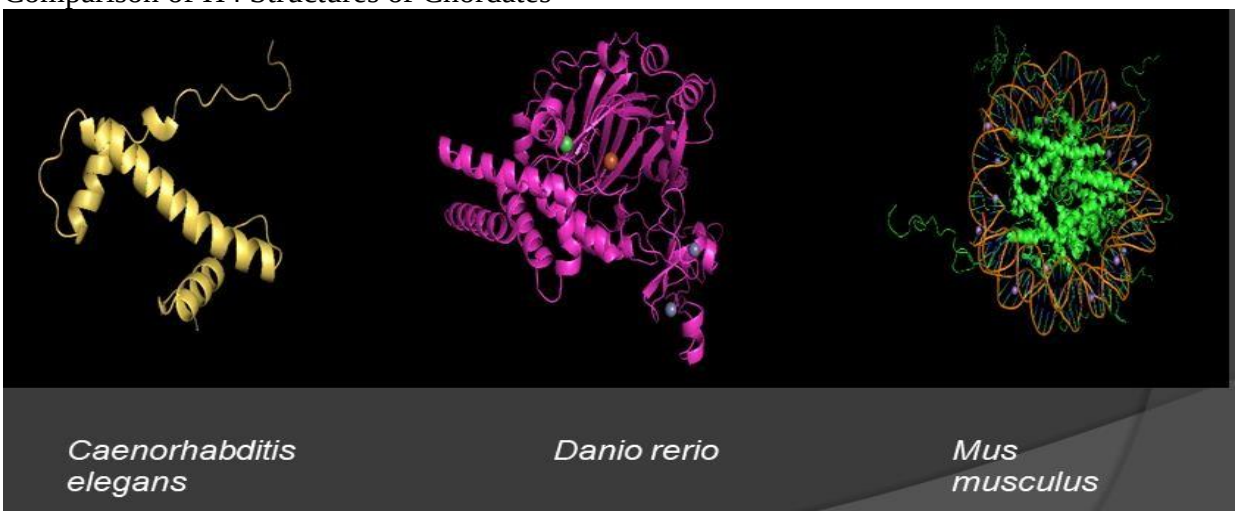


Comparison of H4 Structures of Higher Chordates

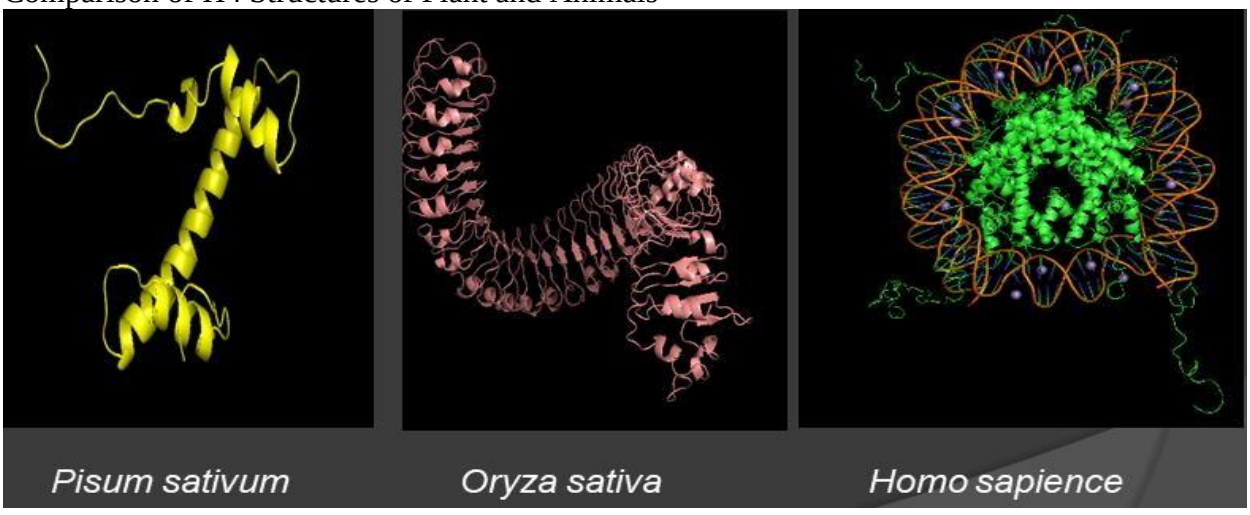
Comparison of H4 Structures of Lower Organisms



Comparison of H4 Structures of Chordates

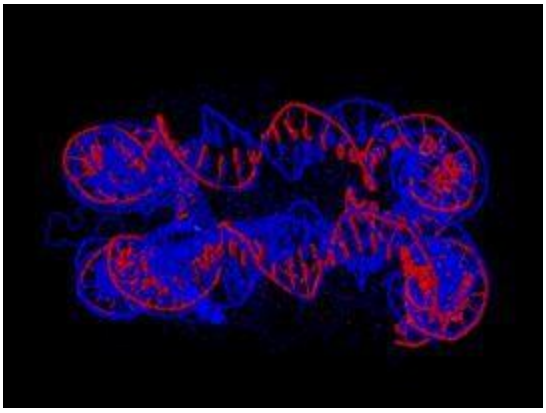


Comparison of H4 Structures of Plant and Animals



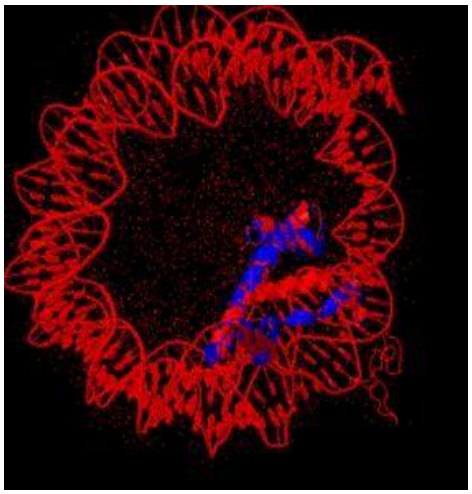
Superimposing H4 from Human and Yeast

```
Name of Chain_1: A522102
Name of Chain_2: B522102
Length of Chain_1:  135 residues
Length of Chain_2:   97 residues

Aligned length=   97, RMSD=   0.41, Seq_ID=n_identical/n_aligned= 0.866
TM-score= 0.71247 (if normalized by length of Chain_1)
TM-score= 0.98818 (if normalized by length of Chain_2)
(You should use TM-score normalized by length of the reference protein)
```



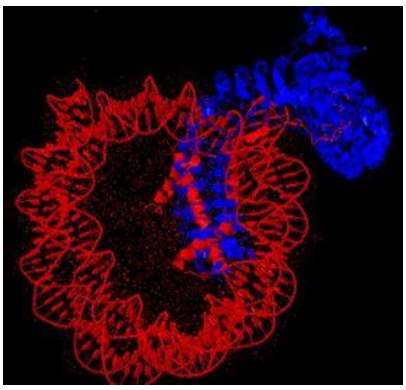
Superimposing H4 from Human and *A. thaliana*

```
Name of Chain_1: A600170
Name of Chain_2: B600170
Length of Chain_1:   91 residues
Length of Chain_2:  135 residues

Aligned length=   86, RMSD=   2.80, Seq_ID=n_identical/n_aligned= 0.163
TM-score= 0.72335 (if normalized by length of Chain_1)
TM-score= 0.51906 (if normalized by length of Chain_2)
```



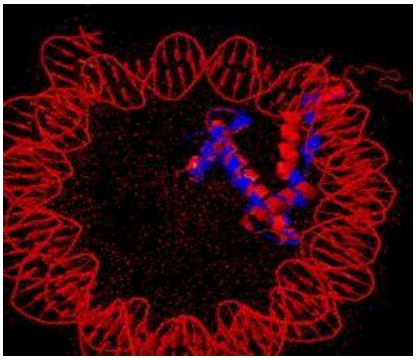
Superimposing H4 from Human and Rice

```
Name of Chain_1: A205474
Name of Chain_2: B205474
Length of Chain_1:  742 residues
Length of Chain_2:  135 residues

Aligned length=   80, RMSD=   4.62, Seq_ID=n_identical/n_aligned= 0.063
TM-score= 0.08923 (if normalized by length of Chain_1)
TM-score= 0.33629 (if normalized by length of Chain_2)
```



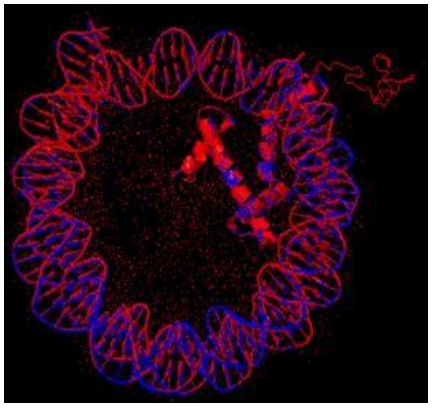
Superimposing H4 from Human and *C. elegans*

```
Name of Chain_1: A149083
Name of Chain_2: B149083
Length of Chain_1:   91 residues
Length of Chain_2:  135 residues

Aligned length=   86, RMSD=   2.80, Seq_ID=n_identical/n_aligned= 0.174
TM-score= 0.72339 (if normalized by length of Chain_1)
TM-score= 0.51908 (if normalized by length of Chain_2)
```



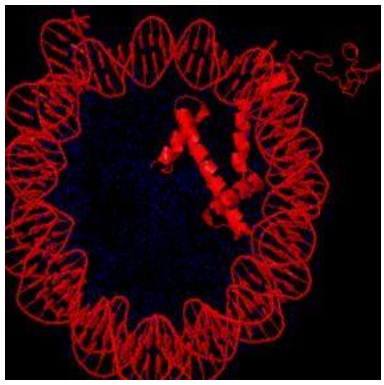
Superimposing H4 from Human and Fruit fly

```
Name of Chain_1: A594101
Name of Chain_2: B594101
Length of Chain_1:   99 residues
Length of Chain_2:  135 residues

Aligned length=   99, RMSD=   0.15, Seq_ID=n_identical/n_aligned= 0.990
TM-score= 0.99823 (if normalized by length of Chain_1)
TM-score= 0.73241 (if normalized by length of Chain_2)
```



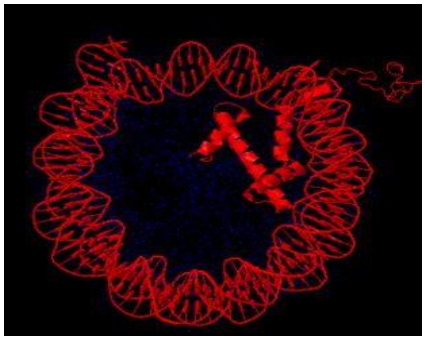
Superimposing H4 from Human and Frog

```
Name of Chain_1: A210060
Name of Chain_2: B210060
Length of Chain_1:  135 residues
Length of Chain_2:  135 residues

Aligned length=  135, RMSD=   0.00, Seq_ID=n_identical/n_aligned= 1.000
TM-score= 1.00000 (if normalized by length of Chain_1)
TM-score= 1.00000 (if normalized by length of Chain_2)
```



Superimposing H4 from Human and Rat

```
Name of Chain_1: A718509
Name of Chain_2: B718509
Length of Chain_1:  135 residues
Length of Chain_2:  135 residues

Aligned length=  135, RMSD=   0.00, Seq_ID=n_identical/n_aligned= 1.000
TM-score= 1.00000 (if normalized by length of Chain_1)
TM-score= 1.00000 (if normalized by length of Chain_2)
(You should use TM-score normalized by length of the reference protein)
```



Discussion

From the result it can be deduced that identity between the sequences H1 histone shows 70% sequence similarity in almost all organisms. In case of, H4 histone except for E. coli all others have above 90% identity amongst each other. This infers that the sequence of amino acid that has been conserved throughout the evolutionary line.

From H4 MSA analysis the phylogenetic tree constructed has the most similarity in accordance with the conventional data available so far. The evolution of H1 histone protein has shown anomalies as compared with the conventional ones. This phylogenetic relationship has been adapted in regards with the function of protein.

Structural visualization shows much similarity shows much similarity with the species. However RMSD value collected from the TM shows higher range as compared to H4. In case of, H4 histone proteins both structural visualization data and structure superimposed data has shown high similarities. The RMSD values for most of pairs of histones are considerably low. The flexibility of the two structures is calculated by RMSD. So, higher the flexibility, higher the RMSD and lower the similarity and vice-versa.

The H4 histone forms the central core of the nucleosome that is needed to wind the major length of the DNA. The H4 histone is mainly composed of lysine and cysteine residues that have highly cationic and need for binding of the negatively charged DNA. This is the major reason that the H4 histone is ought to be conserved throughout the evolutionary lines. On the other hand, H1 histone is needed as the linker protein to bind the larger loops of DNA together. So it is only required in the organisms having larger genome which are found at higher order in the evolutionary lines. These proteins are rich in Lysine and Arginine, which are responsible for the positive charge, and thus the identity percentages for non-conserved proteins are also in range of 15-20%.

Evolution has come across due to countless mutations in DNA, yet the reason for to choose Protein for this study was that every change in DNA ultimately results in a protein adapted to the environment. Hence, by analyzing the protein responsible for a specific function, one can postulate the phylogeny. The other reason is the presence of non-coding sequences in higher organism which, while comparing to lower organism, can act as obstacles.

E. coli does not contain any histone H4 or Histone H, hence variant or homologous of these sequences from the closest related species are available on Databanks. This explains the low identity percentage in H4 Histone identity matrix. Due to extremely small size and one of the primary organisms to be evolved, these do not have Histone but for DNA compacting NHPs (Non-Histone Proteins) are used.

Conclusion

From the study we can conclude that the phylogenetic relationship can be deduced from studying conserved and non-conserved region of protein. These regions can also give idea on which amino acids have been changed for certain specific adaptation that the protein had made, thus will be a great asset in protein engineering and design. Except for that, we can decipher the relation of protein structure and its sequence's relation. From the study we can proof the Anfinson's postulate by the evidence of high RMSD value associated with less identity percentage. Thus this study proves the importance of protein sequence in folding. The phylogenic relation obtained from MSA analysis has similarities to that of evolutionary tree provided by scientists. Importance and contribution of protein in phylogeny is thus been deduced.

Methodology

a) **Retrieval of protein sequence and structure**

Protein sequence FASTA files are retrieved from Protein Databanks like NCBI, UniProt and PDB depending on their availability. Structures are primarily collected in .pdb format from UniProt but for those whose structures are not available SwissModel software is used to create 3D structure through Homology Modeling.

b) **Multiple sequence analysis, Phylogenetic tree and Identity matrix generation**

Multiple sequence analysis (MSA) data has been collected using ClustalW and MSF which uses Needleman-wunsch algorithm for global alignment. From EMBL MSA portal the Phylogenetic Tree using neighborhood joining method without distance is obtained. Identity Matrix is also retrieved from the data.

c) **Structure observation and analysis**

Structure that are retrieved, are viewed using PyMol and thus divided into different sets of evolution to visualize the similarities and differences.

d) **Structure superimpose**

Pair-wise structure superimposing is done using TM-align developed by Zhang Lab in University of Michigan. Human H1 and H4 Histone have been kept constant in all the combination taken for the alignment.

e) **Comparison and relation**

RMSD values obtained from different pairs are compared with the phylogenic tree and the identity matrix to find the relation between structure and sequence of a protein conserved throughout the evolution and of another protein which is non-conserved.

References

- Conserved Proteins Are Fragile; Raquel Assis and Alexey S. Kondrashov
- Evolutionary Conserved Positions DefineProtein Conformational Diversity; Tadeo E. Saldaño, Alexander M. Monzon, Gustavo Parisi, Sebastian Fernandez-Alberti
- Using Evolutionary Rates to Investigate Protein Functional Divergence and Conservation: A Case Study of the Carbonic Anhydrases; Bjarne Knudsen, Michael M. Miyamoto, Philip J. Laipis and David N. Silverman
- BIS2Analyzer: a server for co-evolution analysis of conserved protein families; Francesco Oteri, Francesca Nadalin, Raphael Champeimont1 and Alessandra Carbone
- Structural and Evolutionary Studies on Sterol 14-Demethylase P450 (CYP51), the Most Conserved P450 Monooxygenase: II. Evolutionary Analysis of Protein and Gene Structures; Yuzo Yoshida, Mitsuhide Noshiro, Yuri Aoyama,Takeshi Kawamoto, Tadao Horiuchi, and Osamu Gotoh
- Evolutionarily Conserved Protein Sequences of Influenza A Viruses, Avian and Human, as Vaccine Targets A. T. Heiny1, Olivo Miotto1,2, Kellathur N. Srinivasan, Asif M. Khan, G. L. Zhang, Vladimir Brusic, Tin Wee Tan1, J. Thomas August
- Evolutionary-Conserved Allosteric Properties of Three Neuronal Calcium Sensor Proteins Valerio Marino and Daniele Dell'Orco
- Protein dispensability and rate of evolution Aaron E. Hirsh & Hunter B. Fraser
- Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families Steve W. bockess and Rams Rangenatban
- Evolutionary Conservation of Protein Backbone Flexibility Sandra Maguid, Sebastian Fernandez-Alberti, Gustavo Parisi, Julian Echave
- Protein Phylogenies and Signature Sequences: A Reappraisal of Evolutionary Relationships among Archaebacteria, Eubacteria, and Eukaryotes Radhey S Gupta.
- Identification and phylogenetic sorting of bacterial lineages with universally conserved genes and proteins by Scott R. Santos and Howard Ochman
- Phylogenetic analysis of the core histones H2A, H2B, H3, and H4 by Thomas H.Thatcher+ and Martin A.Gorovsky Department of Biology, University of Rochester, Rochester, NY 14627, USA