B.Sc. (Hons.) Biotechnology
Core Course 13:
Basics of Bioinformatics and Biostatistics (BIOT 3013)

Unit 4:
Introduction to bioinformatics & biological databases

Dr. Satarudra Prakash Singh
Department of Biotechnology
Mahatma Gandhi Central University, Motihari

Bioinformatics

- Bioinformatics is the application of information technology to store, organize and analyze large amount of biological data available in the form of sequence and structure of biomolecules (e.g., protein (the building blocks of organism) and nucleic acid (the information carrier).
- Other types of important data include Gene/protein expression, metabolic pathway, molecular interaction, mutations, genetic map etc.,.

Function of biological database

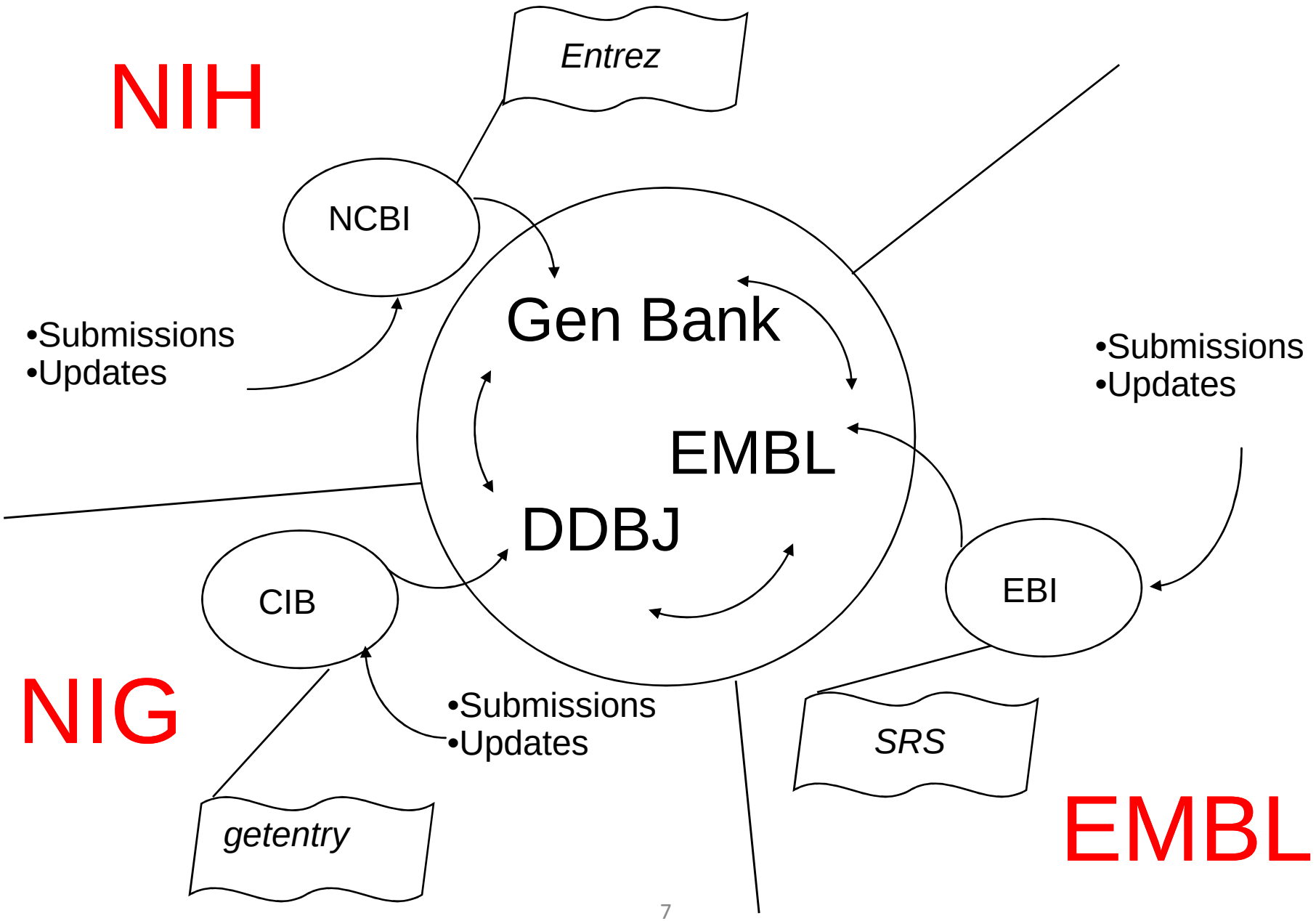
- To make biological data available in computer readable format.
- Easy access of biological information to every scientist.
- Databases in general classified into two categories: primary and secondary
- A primary database contain sequence and structure information alone generated from wet lab experiments.

Primary and secondary database

- A primary database contain sequence and structure information alone generated from wet lab experiments. Examples includes Gene bank, Swiss-Prot/ UniProt, DDBJ, PIR, PDB.
- Secondary database contain information derived from primary databases like conserved sequences, motif. Example includes SCOP, CATH, PROSITE, PRINT, BLOCK, Pfam, eMOTIF etc.

Primary nucleotide sequence database

- Gene bank (located in USA)
- DDBJ (Located in JAPAN)
- EMBL (Located in UK)
- All these databases have uniform data formats (but not identical) and exchange the information on daily basis.


NIH
Entrez
NCBI
•Submissions
•Updates
Gen Bank
EMBL
DDBJ
•Submissions
•Updates
CIB
EBI
NIG
•Submissions
•Updates
SRS
getentry
EMBL

GenBank Flat File (GBFF)

```
LOCUS       MUSNGH      1803 bp    mRNA            ROD       29-AUG-1997
DEFINITION  Mouse neuroblastoma and rat glioma hybridoma cell line NG108-15
            cell TA20 mRNA, complete cds.
ACCESSION   D25291
NID         g1850791
KEYWORDS    neurite extension activity; growth arrest; TA20.
SOURCE      Murinae gen. sp. mouse neuroblastma-rat glioma hybridoma
            cell_line:NG108-15 cDNA to mRNA.
  ORGANISM  Murinae gen. sp.
            Eukaryotae; mitochondrial eukaryotes; Metazoa; Chordata;
            Vertebrata; Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae;
            Murinae.
REFERENCE   1  (sites)
  AUTHORS   Tohda,C., Nagai,S., Tohda,M. and Nomura,Y.
  TITLE     A novel factor, TA20, involved in neuronal differentiation: cDNA
            cloning and expression
  JOURNAL   Neurosci. Res. 23 (1), 21-27 (1995)
  MEDLINE   96064354
REFERENCE   3  (bases 1 to 1803)
  AUTHORS   Tohda,C.
  TITLE     Direct Submission
  JOURNAL   Submitted (18-NOV-1993) to the DDBJ/EMBL/GenBank databases. Chihiro
            Tohda, Toyama Medical and Pharmaceutical University, Research
            Institute for Wakan-yaku, Analytical Research Center for
            Ethnomedicines; 2630 Sugitani, Toyama, Toyama 930-01, Japan
            (E-mail:CHIHIRO@ms.toyama-mpu.ac.jp, Tel:+81-764-34-2281(ex.2841),
            Fax:+81-764-34-5057)
COMMENT     On Feb 26, 1997 this sequence version replaced gi:793764.
FEATURES             Location/Qualifiers
     source          1..1803
                     /organism="Murinae gen. sp."
                     /note="source origin of sequence, either mouse or rat, has
                     not been identified"
                     /db_xref="taxon:39108"
                     /cell_line="NG108-15"
                     /cell_type="mouse neuroblastma-rat glioma hybridoma"
     misc_signal     156..163
                     /note="AP-2 binding site"
     GC_signal       647..655
                     /note="Sp1 binding site"
     TATA_signal     694..701
     gene            748..1311
                     /gene="TA20"
     CDS             748..1311
                     /gene="TA20"
                     /function="neurite extensiion activity and growth arrest
                     effect"
                     /codon_start=1
                     /db_xref="PID:d1005516"
                     /db_xref="PID:g793765"
                     /translation="MMKLWVPSRSLPNSPNHYRSFLSHTLHIRYNNSLFISNTHLSRR
                     KLRVTNPIYTRKRSLNIFYLLIPSCRTRLILWIIYIYRNLKHWSTSTVRSHSHSIYRL
                     RPSMRTNIILRCHSYYKPPISHPIYWNNPSRMNLRGLLSRQSHLDPILRFPLHLTIYY
                     RGPSNRSPPLPPRNRIKQPNRIKLRCR"
     polyA_site      1803
BASE COUNT      507 a    458 c    311 g    527 t
ORIGIN
        1 tcagtttttt ttttttttct ttttttttct ttttttttct ttttttttg ttgattcatg
       61 tccgtttaca tttggtaagt tcacaggcct cagtcaacac aattggactg ctcaggaaat
      121 cctccttggt gaccgcagta tacttggcct atgaacccaa gccacctatg gctaggtagg
      181 agaagctcaa ctgtagggct gactttggaa gagaatgcac atggctgtat cgacatttca
      241 catggtggac ctctggccag agtcagcagg ccgagggttc tcttccgggc tgctccctca
      301 ctgcttgact ctgcgtcagt gcgtccatac tgtgggcgga cgttattgct atttgccttc
      361 cattctgtac ggcattgcct ccatttagct ggagagggac agagcctggt tctctagggc
      421 gttttccattg gggcctggtg acaatccaaa agatgagggc tccaaacacc agaatcagaa
      481 ggcccagcgt atttgtaaaa acaccttctg gtgggaatga atggtacagg ggcgtttcag
      541 gacaaagaac agcttttctg tcactcccat gagaaccgtc gcaatcactg ttccgaagag
      601 gaggagtcca gaatacacgt gtatgggcat gacgattgcc cggagagagg cggagcccat
      661 ggaagcagaa agacgaaaaa cacacccatt atttaaaatt attaaccact cattcattga
      721 cctacctgcc ccatccaaca tttcatcatg atgaaacttt gggtcccttc taggagtctg
      781 cctaatagtc caaatcatta caggtctttt cttagccata cactacacat cagatacaat
      841 aacagccttt tcatcagtaa cacacatttg tcgagacgta aattacgggt gactaatccg
      901 atatatacac gcaaacggag cctcaatatt ttttatttgc ttattccttc atgtcggacg
      961 aggcttatat tatggatcat atacatttat agaaacctga aacattggag tacttctact
     1021 gttcgcagtc atagccacag catttatagg ctacgtcctt ccatgaggac aaatatcatt
     1081 ctgaggtgcc acagttatta caaacctcct atcagccatc ccatatattg gaacaaccct
     1141 agtcgaatga atttgagggg gcttctcagt agacaaagcc accttgaccc gattcttcgc
     1201 tttccacttc atcttaccat ttattatcgc ggccctagca atcgttcacc tcctcttcct
     1261 ccacgaaaca ggatcaaaca acccaacagg attaaactca gatgcagata aaattccatt
     1321 tcacccctac tatacatcaa agatatccta ggtatcctaa tcatattctt aattctcata
     1381 accctagtat tatttttccc agacatacta ggagacccag acaactacat accagctaat
     1441 ccactaaaca ccccacccca tattaaaccc gaatgatatt tcctatttgc atacgccatt
     1501 ctacgctcaa tccccaataa actaggaggt gtcctagcct taatcttatc tatcctaatt
     1561 ttagccctaa tacctttcct tcatacctca aagcaacgaa gcctaatatt ccgcccaatc
     1621 acacaaattt tgtactgaat cctagtagcc aacctactta tcttaacctg aattgggggc
     1681 caaccagtag acacccattt attatcattg gccaactagc ctccatctca tacttctcaa
     1741 tcatcttaat tcttatacca atctcaggaa ttatcgaaga caaaatacta aaattatatc
     1801 cat
//
```

Header

- Title
- Taxonomy
- Citation

Features (AA seq)

DNA Sequence

Primary protein database

- PIR (Protein information resource)-USA
- SWISS-PROT – Switzerland
- TrEMBL
- UniProt
- PDB

Swiss-Prot

```
ID   CYS3_YEAST     STANDARD;      PRT;   393 AA.
AC   P31373;
DT   01-JUL-1993 (REL. 26, CREATED)
DE   CYSTATHIONINE GAMMA-LYASE (EC 4.4.1.1) (GAMMA-CYSTATHIONASE).
GN   CYS3 OR CYI1 OR STR1 OR YAL012W OR FUN35.
OS   TAXONOMY
OC   SACCHAROMYCETACEAE; SACCHAROMYCES.

RX   CITATION
CC   -!- CATALYTIC ACTIVITY: L-CYSTATHIONINE + H(2)O = L-CYSTEINE +
CC       NH(3) + 2-OXOBUTANOATE.
CC   -!- COFACTOR: PYRIDOXAL PHOSPHATE.
CC   -!- PATHWAY: FINAL STEP IN THE TRANS-SULFURATION PATHWAY SYNTHESIZING
CC       L-CYSTEINE FROM L-METHIONINE.
CC   -!- SUBUNIT: HOMOTETRAMER.
CC   -!- SUBCELLULAR LOCATION: CYTOPLASMIC.
CC   -!- SIMILARITY: BELONGS TO THE TRANS-SULFURATION ENZYMES FAMILY.
CC   ---------------------------------------------------------------------------
CC   DISCLAMOR
CC   ---------------------------------------------------------------------------

DR   DATABASE cross-reference
KW   CYSTEINE BIOSYNTHESIS; LYASE; PYRIDOXAL PHOSPHATE.
FT   INIT_MET      0      0
FT   BINDING     203    203       PYRIDOXAL PHOSPHATE (BY SIMILARITY).
SQ   SEQUENCE   393 AA;  42411 MW;  55BA2771 CRC32;
     TLQESDKFAT KAIHAGEHVD VHGSVIEPIS LSTTFKQSSP ANPIGTYEYS RSQNPNRENL
     ERAVAALENA QYGLAFSSGS ATTATILQSL PQGSHAVSIG DVYGGTHRYF TKVANAHGVE
     TSFTNDLLND LPQLIKENTK LVWIETPTNP TLKVTDIQKV ADLIKKHAAG QDVILVVDNT
     FLSPYISNPL NFGADIVVHS ATKYINGHSD VVLGVLATNN KPLYERLQFL QNAIGAIPSP
     FDAWLTHRGL KTLHLRVRQA ALSANKIAEF LAADKENVVA VNYPGLKTHP NYDVVLKQHR
     DALGGGMISF RIKGGAEAAS KFASSTRLFT LAESLGGIES LLEVPAVMTH GGIPKEAREA
     SGVFDDLVRI SVGIEDTDDL LEDIKQALKQ ATN
//
```

```
ID   CYS3_YEAST     STANDARD;      PRT;   393 AA.
AC   P31373;
DT   01-JUL-1993 (REL. 26, CREATED)
DT   01-JUL-1993 (REL. 26, LAST SEQUENCE UPDATE)
DT   01-NOV-1995 (REL. 32, LAST ANNOTATION UPDATE)
DE   CYSTATHIONINE GAMMA-LYASE (EC 4.4.1.1) (GAMMA-CYSTATHIONASE).
GN   CYS3 OR CYI1 OR STR1 OR YAL012W OR FUN35.
OS   SACCHAROMYCES CEREVISIAE (BAKER'S YEAST).
OC   EUKARYOTA; FUNGI; ASCOMYCOTA; HEMIASCOMYCETES; SACCHAROMYCETALES;
OC   SACCHAROMYCETACEAE; SACCHAROMYCES.
RN   [1]
RP   SEQUENCE FROM N.A., AND PARTIAL SEQUENCE.
RX   MEDLINE; 92250430. [NCBI, ExPASy, Israel, Japan]
RA   ONO B.-I., TANAKA K., NAITO K., HEIKE C., SHINODA S., YAMAMOTO S.,
RA   OHMORI S., OSHIMA T., TOH-E A.;
RT   "Cloning and characterization of the CYS3 (CYI1) gene of
RT   Saccharomyces cerevisiae.";
RL   J. BACTERIOL. 174:3339-3347(1992).
RN   [2]
RP   SEQUENCE FROM N.A., AND CHARACTERIZATION.
RC   STRAIN=DBY939;
RX   MEDLINE; 93328685. [NCBI, ExPASy, Israel, Japan]
RA   YAMAGATA S., D'ANDREA R.J., FUJISAKI S., ISAJI M., NAKAMURA K.;
RT   "Cloning and bacterial expression of the CYS3 gene encoding
RT   cystathionine gamma-lyase of Saccharomyces cerevisiae and the
RT   physicochemical and enzymatic properties of the protein.";
RL   J. BACTERIOL. 175:4800-4808(1993).
RN   [3]
RP   SEQUENCE FROM N.A.
RC   STRAIN=S288C / AB972;
RX   MEDLINE; 93289814. [NCBI, ExPASy, Israel, Japan]
RA   BARTON A.B., KABACK D.B., CLARK M.W., KENG T., OUELLETTE B.F.F.,
RA   STORMS R.K., ZENG B., ZHONG W.W., FORTIN N., DELANEY S., BUSSEY H.;
RT   "Physical localization of yeast CYS3, a gene whose product resembles
RT   the rat gamma-cystathionase and Escherichia coli cystathionine gamma-
RT   synthase enzymes.";
RL   YEAST 9:363-369(1993).
RN   [4]
RP   SEQUENCE FROM N.A.
RC   STRAIN=S288C / AB972;
RX   MEDLINE; 93209532. [NCBI, ExPASy, Israel, Japan]
RA   OUELLETTE B.F.F., CLARK M.W., KENG T., STORMS R.K., ZHONG W.W.,
RA   ZENG B., FORTIN N., DELANEY S., BARTON A.B., KABACK D.B., BUSSEY H.;
RT   "Sequencing of chromosome I from Saccharomyces cerevisiae: analysis
RT   of a 32 kb region between the LTE1 and SPO7 genes.";
RL   GENOME 36:32-42(1993).
RN   [5]
RP   SEQUENCE OF 1-18, AND CHARACTERIZATION.
RX   MEDLINE; 93289817. [NCBI, ExPASy, Israel, Japan]
RA   ONO B.-I., ISHII N., NAITO K., MIYOSHI S.-I., SHINODA S., YAMAMOTO S.,
RA   OHMORI S.;
RT   "Cystathionine gamma-lyase of Saccharomyces cerevisiae: structural
RT   gene and cystathionine gamma-synthase activity.";
RL   YEAST 9:389-397(1993).
CC   -!- CATALYTIC ACTIVITY: L-CYSTATHIONINE + H(2)O = L-CYSTEINE +
CC       NH(3) + 2-OXOBUTANOATE.
CC   -!- COFACTOR: PYRIDOXAL PHOSPHATE.
CC   -!- PATHWAY: FINAL STEP IN THE TRANS-SULFURATION PATHWAY SYNTHESIZING
CC       L-CYSTEINE FROM L-METHIONINE.
CC   -!- SUBUNIT: HOMOTETRAMER.
CC   -!- SUBCELLULAR LOCATION: CYTOPLASMIC.
CC   -!- SIMILARITY: BELONGS TO THE TRANS-SULFURATION ENZYMES FAMILY.
CC   ---------------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between  the Swiss Institute of Bioinformatics  and the  EMBL outstation -
CC   the European Bioinformatics Institute.  There are no  restrictions on  its
CC   use  by  non-profit  institutions as long  as its content  is  in  no  way
CC   modified and this statement is not removed.  Usage  by  and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------------
DR   EMBL; L05146; AAC04945.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence]
DR   EMBL; L04459; AAA85217.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence]
DR   EMBL; D14135; BAA03190.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence]
DR   PIR; S31228; S31228.
DR   YEPD; 5280; -.
DR   SGD; L0000470; CYS3. [SGD / YPD]
DR   PFAM; PF01053; Cys_Met_Meta_PP; 1.
DR   PROSITE; PS00868; CYS_MET_METAB_PP; 1.
DR   DOMO; P31373.
DR   PRODOM [Domain structure / List of seq. sharing at least 1 domain]
DR   PROTOMAP; P31373.
DR   PRESAGE; P31373.
DR   SWISS-2DPAGE; GET REGION ON 2D PAGE.
KW   CYSTEINE BIOSYNTHESIS; LYASE; PYRIDOXAL PHOSPHATE.
FT   INIT_MET      0      0
FT   BINDING     203    203       PYRIDOXAL PHOSPHATE (BY SIMILARITY).
SQ   SEQUENCE   393 AA;  42411 MW;  55BA2771 CRC32;
     TLQESDKFAT KAIHAGEHVD VHGSVIEPIS LSTTFKQSSP ANPIGTYEYS RSQNPNRENL
     ERAVAALENA QYGLAFSSGS ATTATILQSL PQGSHAVSIG DVYGGTHRYF TKVANAHGVE
     TSFTNDLLND LPQLIKENTK LVWIETPTNP TLKVTDIQKV ADLIKKHAAG QDVILVVDNT
     FLSPYISNPL NFGADIVVHS ATKYINGHSD VVLGVLATNN KPLYERLQFL QNAIGAIPSP
     FDAWLTHRGL KTLHLRVRQA ALSANKIAEF LAADKENVVA VNYPGLKTHP NYDVVLKQHR
     DALGGGMISF RIKGGAEAAS KFASSTRLFT LAESLGGIES LLEVPAVMTH GGIPKEAREA
     SGVFDDLVRI SVGIEDTDDL LEDIKQALKQ ATN
//
```

TREMBL

- TrEMBL is a computer-annotated protein sequence database supplementing the SWISS-PROT Protein Sequence Data Bank.

- TrEMBL contains the translations of all coding sequences (CDS) present in the EMBL Nucleotide Sequence Database not yet integrated in SWISS-PROT.

- TrEMBL can be considered as a preliminary section of SWISS-PROT. For all TrEMBL entries which should finally be upgraded to the standard SWISS-PROT quality, SWISS-PROT accession numbers have been assigned.

UniProt

- New protein sequence database that is the result of a merge from SWISS-PROT and PIR. It will be the annotated curated protein sequence database.
- Data in UniProt is primarily derived from coding sequence annotations in EMBL (GenBank/DDBJ) nucleic acid sequence data.
- UniProt is a Flat-File database just like EMBL and GenBank
- Flat-File format is Swiss-Prot-like, or EMBL-like

PDB

- HEADER
- COMPND
- SOURCE
- AUTHOR
- DATE
- JRNL
- REMARK
- SECRES
- ATOM COORDINATES

```
HEADER    LEUCINE ZIPPER                          15-JUL-93   1DGC      1DGC   2
COMPND    GCN4 LEUCINE ZIPPER COMPLEXED WITH SPECIFIC                     1DGC   3
COMPND   2 ATF/CREB SITE DNA                                              1DGC   4
SOURCE    GCN4: YEAST (SACCHAROMYCES CEREVISIAE); DNA: SYNTHETIC          1DGC   5
AUTHOR    T.J.RICHMOND                                                    1DGC   6
REVDAT   1   22-JUN-94 1DGC    0                                          1DGC   7
JRNL        AUTH   P.KONIG,T.J.RICHMOND                                   1DGC   8
JRNL        TITL   THE X-RAY STRUCTURE OF THE GCN4-BZIP BOUND TO          1DGC   9
JRNL        TITL 2 ATF/CREB SITE DNA SHOWS THE COMPLEX DEPENDS ON DNA     1DGC  10
JRNL        TITL 3 FLEXIBILITY                                            1DGC  11
JRNL        REF    J.MOL.BIOL.                   V. 233   139 1993      1DGC  12
JRNL        REFN   ASTM JMOBAK  UK ISSN 0022-2836               0070     1DGC  13
REMARK   1                                                                1DGC  14
REMARK   2                                                                1DGC  15
REMARK   2 RESOLUTION. 3.0  ANGSTROMS.                                    1DGC  16
REMARK   3                                                                1DGC  17
REMARK   3 REFINEMENT.                                                    1DGC  18
REMARK   3   PROGRAM                  X-PLOR                              1DGC  19
REMARK   3   AUTHORS                  BRUNGER                             1DGC  20
REMARK   3   R VALUE                  0.216                               1DGC  21
REMARK   3   RMSD BOND DISTANCES       0.020  ANGSTROMS                   1DGC  22
REMARK   3   RMSD BOND ANGLES          3.86   DEGREES                     1DGC  23
REMARK   3                                                                1DGC  24
REMARK   3   NUMBER OF REFLECTIONS     3296                                1DGC  25
REMARK   3   RESOLUTION RANGE     10.0 - 3.0  ANGSTROMS                   1DGC  26
REMARK   3   DATA CUTOFF              3.0    SIGMA(F)                     1DGC  27
REMARK   3   PERCENT COMPLETION        98.2                               1DGC  28
REMARK   3                                                                1DGC  29
REMARK   3   NUMBER OF PROTEIN ATOMS                        456           1DGC  30
REMARK   3   NUMBER OF NUCLEIC ACID ATOMS                  386           1DGC  31
REMARK   4                                                                1DGC  32
REMARK   4 GCN4: TRANSCRIPTIONAL ACTIVATOR OF GENES ENCODING FOR AMINO    1DGC  33
REMARK   4 ACID BIOSYNTHETIC ENZYMES.                                     1DGC  34
REMARK   5                                                                1DGC  35
REMARK   5 AMINO ACIDS NUMBERING (RESIDUE NUMBER) CORRESPONDS TO THE      1DGC  36
REMARK   5 281 AMINO ACIDS OF INTACT GCN4.                                1DGC  37
REMARK   6                                                                1DGC  38
REMARK   6 BZIP SEQUENCE 220 - 281 USED FOR CRYSTALLIZATION.              1DGC  39
REMARK   7                                                                1DGC  40
REMARK   7 MODEL FROM AMINO ACIDS 227 - 281 SINCE AMINO ACIDS 220 -       1DGC  41
REMARK   7 226 ARE NOT WELL ORDERED.                                      1DGC  42
REMARK   8                                                                1DGC  43
REMARK   8 RESIDUE NUMBERING OF NUCLEOTIDES:                              1DGC  44
REMARK   8 5' T  G  G  A  G  A  T  G  A  C  G  T  C  A  T  C  T  C  C     1DGC  45
REMARK   8  -10 -9 -8 -7 -6 -5 -4 -3 -2 -1  1  2  3  4  5  6  7  8  9     1DGC  46
REMARK   9                                                                1DGC  47
REMARK   9 THE ASYMMETRIC UNIT CONTAINS ONE HALF OF PROTEIN/DNA           1DGC  48
REMARK   9 COMPLEX PER ASYMMETRIC UNIT.                                   1DGC  49
REMARK  10                                                                1DGC  50
REMARK  10 MOLECULAR DYAD AXIS OF PROTEIN DIMER AND PALINDROMIC HALF      1DGC  51
REMARK  10 SITES OF THE DNA COINCIDES WITH CRYSTALLOGRAPHIC TWO-FOLD      1DGC  52
REMARK  10 AXIS.  THE FULL PROTEIN/DNA COMPLEX CAN BE OBTAINED BY         1DGC  53
REMARK  10 APPLYING THE FOLLOWING TRANSFORMATION MATRIX AND               1DGC  54
REMARK  10 TRANSLATION VECTOR TO THE COORDINATES X Y Z:                   1DGC  55
REMARK  10                                                                1DGC  56
REMARK  10     0   -1   0   X       117.32       X SYMM                   1DGC  57
REMARK  10    -1    0   0   Y   +   117.32   =   Y SYMM                   1DGC  58
REMARK  10     0    0  -1   Z        43.33       Z SYMM                   1DGC  59
SEQRES   1 A   62  ILE VAL PRO GLU SER SER ASP PRO ALA ALA LEU LYS ARG    1DGC  60
SEQRES   2 A   62  ALA ARG ASN THR GLU ALA ALA ARG ARG SER ARG ALA ARG    1DGC  61
SEQRES   3 A   62  LYS LEU GLN ARG MET LYS GLN LEU GLU ASP LYS VAL GLU    1DGC  62
SEQRES   4 A   62  GLU LEU LEU SER LYS ASN TYR HIS LEU GLU ASN GLU VAL    1DGC  63
SEQRES   5 A   62  ALA ARG LEU LYS LYS LEU VAL GLY GLU ARG                1DGC  64
SEQRES   1 B   19    T   G   G   A   G   A   T   G   A   C   G   T   C   1DGC  65
SEQRES   2 B   19    A   T   C   T   C   C                               1DGC  66
HELIX    1 A   ALA A  228  LYS A  276  1                                  1DGC  67
CRYST1   58.660   58.660   86.660  90.00  90.00  90.00 P 41 21 2     8    1DGC  68
ORIGX1      1.000000  0.000000  0.000000        0.00000                   1DGC  69
ORIGX2      0.000000  1.000000  0.000000        0.00000                   1DGC  70
ORIGX3      0.000000  0.000000  1.000000        0.00000                   1DGC  71
SCALE1      0.017047  0.000000  0.000000        0.00000                   1DGC  72
SCALE2      0.000000  0.017047  0.000000        0.00000                   1DGC  73
SCALE3      0.000000  0.000000  0.011539        0.00000                   1DGC  74
ATOM      1  N   PRO A 227      35.313 108.011  15.140  1.00 38.94       1DGC  75
ATOM      2  CA  PRO A 227      34.172 107.658  15.972  1.00 39.82       1DGC  76


ATOM    842  C5    C B   9      57.692 100.286  22.744  1.00 29.82       1DGC 916
ATOM    843  C6    C B   9      58.128 100.193  21.465  1.00 30.63       1DGC 917
TER     844        C B   9                                               1DGC 918
MASTER      46    0    0    1    0    0    0    6  842    2    0    7    1DGC 919
END                                                                       1DGC 920
```

FASTA format

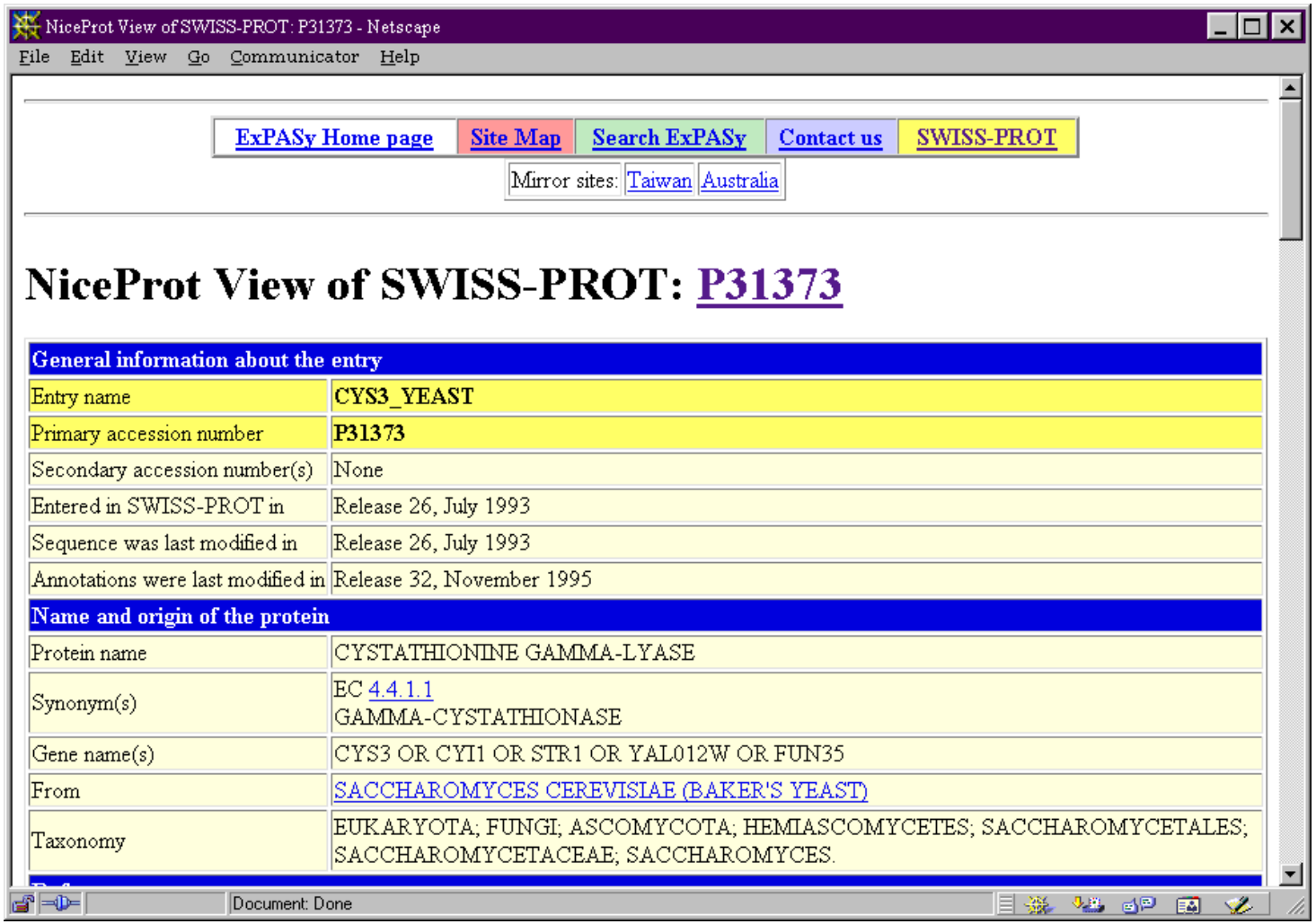
```
>
MSEYQPSLFALNPMGFSPLDGSKSTNENVSASTSTAKPMVGQLIFDKFIKTEEDPI
IKQDTPSNLDFDFALPQTATAPDAKTVLPIPELDDAVVESFFSSSTDSTPMFEYEN
LEDNSKEWTSLFDNDIPVTTDDVSLADKAIESTEEVSLVPSNLEVSTTSFLPTPVL
EDAKLTQTRKVKKPNSVVKKSHHVGKDDESRLDHLGVVAYNRKQRSIPLSPIVPES
SDPAALKRARNTEAARRSRARKLQRMKQLEDKVEELLSKNYHLENEVARLKKLVGE
R
```

“Ten Important Bioinformatics Databases”

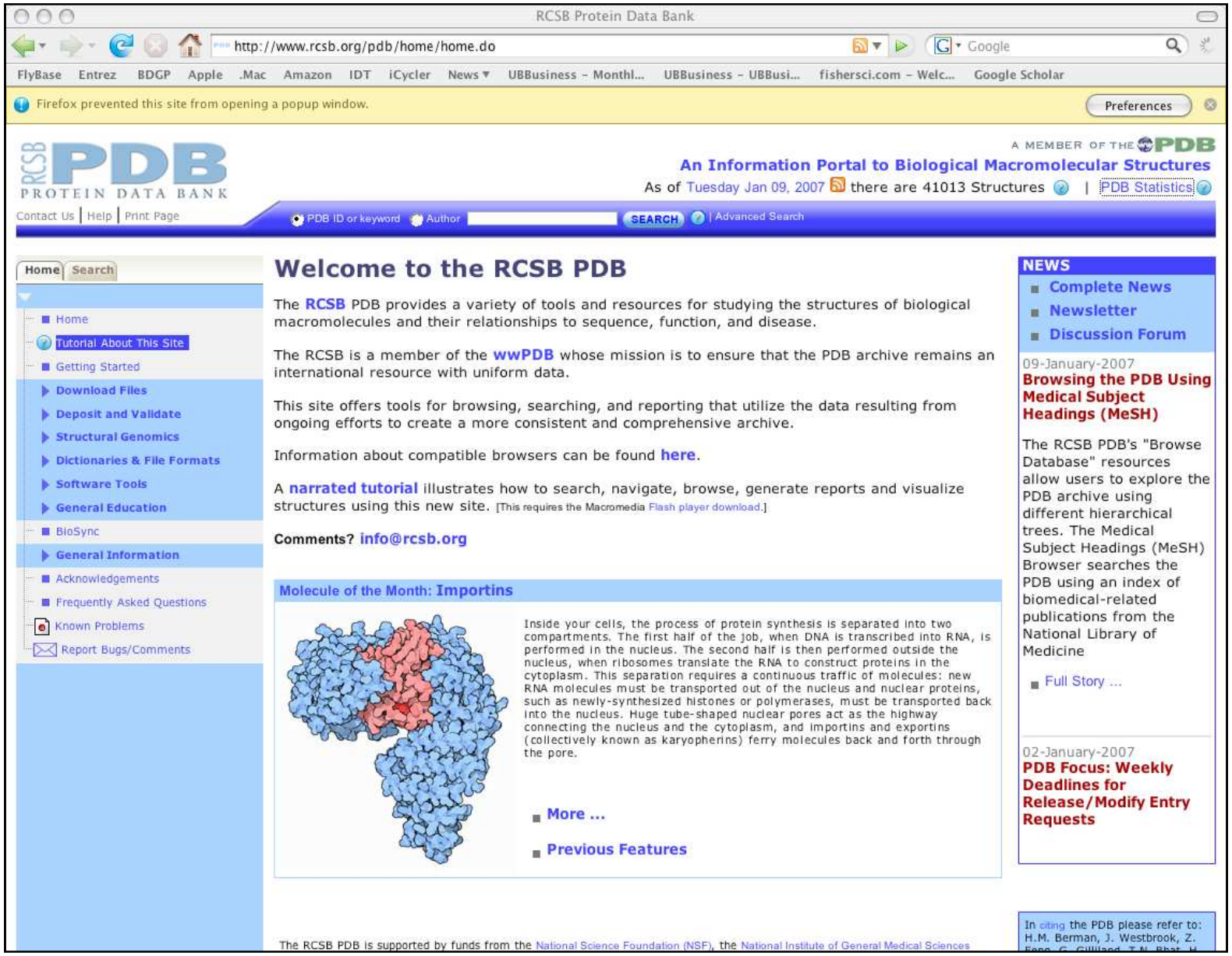
GenBank	www.ncbi.nlm.nih.gov	nucleotide sequences
Ensembl	www.ensembl.org	human/mouse genome
PubMed	www.ncbi.nlm.nih.gov	literature references
NR	www.ncbi.nlm.nih.gov	protein sequences
SWISS-PROT	www.expasy.ch	protein sequences
InterPro	www.ebi.ac.uk	protein domains
OMIM	www.ncbi.nlm.nih.gov	genetic diseases
Enzymes	www.chem.qmul.ac.uk	enzymes
PDB	www.rcsb.org/pdb/	protein structures
KEGG	www.genome.ad.jp	metabolic pathways

Source: *Bioinformatics for Dummies*

Swiss-Prot


NiceProt View of SWISS-PROT: P31373 - Netscape
File Edit View Go Communicator Help
ExPASy Home page
Site Map
Search ExPASy
Contact us
SWISS-PROT
Mirror sites: Taiwan Australia
NiceProt View of SWISS-PROT: P31373
General information about the entry
Entry name
CYS3_YEAST
Primary accession number
P31373
Secondary accession number(s)
None
Entered in SWISS-PROT in
Release 26, July 1993
Sequence was last modified in
Release 26, July 1993
Annotations were last modified in
Release 32, November 1995
Name and origin of the protein
Protein name
CYSTATHIONINE GAMMA-LYASE
Synonym(s)
EC 4.4.1.1
GAMMA-CYSTATHIONASE
Gene name(s)
CYS3 OR CYI1 OR STR1 OR YAL012W OR FUN35
From
SACCHAROMYCES CEREVISIAE (BAKER'S YEAST)
Taxonomy
EUKARYOTA; FUNGI; ASCOMYCOTA; HEMIASCOMYCETES; SACCHAROMYCETALES; SACCHAROMYCETACEAE; SACCHAROMYCES.
Document: Done

Protein Data Bank (PDB)



RCSB PDB : Structure Explorer

http://www.rcsb.org/pdb/explore.do?structureId=4HHB

FlyBase Entrez BDGP Apple .Mac Amazon IDT iCycler News ▾ UBBusiness - Monthl... UBBusiness - UBBusi... fishersci.com - Welc... Google Scholar

RCSB Protein Data Bank | RCSB PDB : Structure Explorer

RCSB PDB PROTEIN DATA BANK

A MEMBER OF THE PDB

An Information Portal to Biological Macromolecular Structures

As of Tuesday Jan 09, 2007 there are 41013 Structures | PDB Statistics

Contact Us | Help | Print Page

PDB ID or keyword | Author | SEARCH | Advanced Search

Home | Search | Structure

Structure Summary | Biology & Chemistry | Materials & Methods | Sequence Details | Geometry

- 4HHB
- Download Files
- FASTA Sequence
- Display Files
- Display Molecule
- Structural Reports
- External Links
- Structure Analysis
- Help

Learn more: [M] 4HHB (Replaced 1HHB)

Title	THE CRYSTAL STRUCTURE OF HUMAN DEOXYHAEMOGLOBIN AT 1.74 ANGSTROMS RESOLUTION
Authors	Fermi, G., Perutz, M.F.
Primary Citation	**Fermi, G., Perutz, M.F., Shaanan, B., Fourme, R.** The crystal structure of human deoxyhaemoglobin at 1.74 A resolution *J.Mol.Biol. v175 pp.159-174 , 1984* [Abstract]
History	Deposition 1984-03-07 Release 1984-07-17 Previous versions: 1HHB
Experimental Method	Type X-RAY DIFFRACTION Data N/A

Parameters	Resolution[A]	R-Value	R-Free	Space Group
	1.74	0.135 (work)	n/a	$P\,2_1\,(P\,1\,2_1\,1)$

Unit Cell						
Length [A]	a	63.15	b	83.59	c	53.80
Angles [°]	alpha	90.00	beta	99.34	gamma	90.00

Molecular Description Asymmetric Unit

Polymer: 1 Molecule: HEMOGLOBIN (DEOXY) (ALPHA CHAIN)
Chains: A,C
Polymer: 2 Molecule: HEMOGLOBIN (DEOXY) (BETA CHAIN)
Chains: B,D

Classification Oxygen Transport

Source Polymer: 1 Scientific Name: **Homo sapiens** Polymer: 2 Scientific Name: **Homo sapiens**

Chemical Component

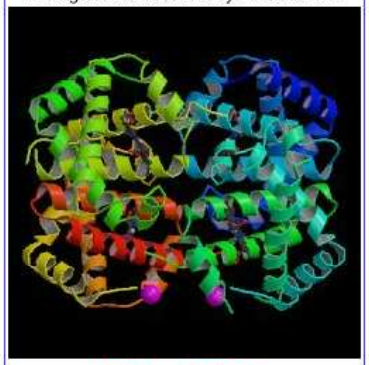
Identifier	Name	Formula	Drug Similarity	Ligand Structure	Ligand Interaction
PO4	PHOSPHATE ION	$O_4\,P^{3-}$	[View]	[View]	[View]
HEM	PROTOPORPHYRIN IX CONTAINING FE	$C_{34}\,H_{32}\,N_4\,O_4\,Fe$	[View]	[View]	[View]

SCOP Classification (version 1.69)

Domain Info	Class	Fold	Superfamily	Family	Domain	Species
d4hhba_	All alpha proteins	Globin-like	Globin-like	Globins	Hemoglobin, alpha-chain	Human (Homo sapiens)

Images and Visualization

Biological Molecule / Asymmetric Unit



Display Options

KiNG
Jmol
WebMol
Protein Workshop
QuickPDB
All Images

PubMed


Links
Conserved Domains
GEO Profiles
HomoloGene
Map Viewer
Nucleotide
OMIM
Full text in PMC
PubMed
SNP
SNP: Genotype
SNP: GeneView
Taxonomy
UniSTS
AceView
Ensembl
Evidence Viewe
GDB
HGNC
KEGG
MGC
ModelMaker
Reactome
UCSC
UniGene
LinkOut
NCBI
PubMed
National Library of Medicine
NLM
My NCBI
[Sign In] [Register]
All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books
Search PubMed for Go Clear
Limits Preview/Index History Clipboard Details
Display Abstract Show 20 Sort by Send to
All: 1 Review: 0
About Entrez
Text Version
Entrez PubMed
Overview
Help | FAQ
Tutorial
New/Noteworthy
E-Utilities
PubMed Services
Journals Database
MeSH Database
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
Special Queries
LinkOut
My NCBI (Cubby)
Related Resources
Order Documents
NLM Mobile
NLM Catalog
NLM Gateway
TOXNET
Consumer Health
Clinical Alerts
ClinicalTrials.gov
PubMed Central
1: Proc Natl Acad Sci U S A. 1986 Feb;83(3):634-8.
Related Articles, Links
FREE Full Text Article at www.pnas.org
Denison Library Full Text Online
Denison Library Print Collection
FREE full text article in PubMed Central
Three human alcohol dehydrogenase subunits: cDNA structure and molecular and evolutionary divergence.
Ikuta T, Szeto S, Yoshida A.
Class I human alcohol dehydrogenase (ADH; alcohol:NAD+ oxidoreductase, EC 1.1.1.1) consists of several homo- and heterodimers of alpha, beta, and gamma subunits that are governed by the ADH1, ADH2, and ADH3 loci. We previously cloned a full length of cDNA for the beta subunit, and the complete sequence of 374 amino acid residues was established. cDNAs for the alpha and gamma subunits were cloned and characterized. A human liver cDNA library, constructed in phage lambda gt11, was screened by using a synthetic oligonucleotide probe that was matched to the gamma but not to the beta sequence. Clone pUCADH gamma 21 and clone pUCADH alpha 15L differed from beta cDNA with respect to restriction sites and hybridization with the nucleotide probe. Clone pUCADH gamma 21 contained an insertion of 1.5 kilobase pairs (kbp) and encodes 374 amino acid residues compatible with the reported amino acid sequence of the gamma subunit. Clone pUCADH alpha 15L contained an insertion of 2.4 kbp and included nucleotide sequences that encode 374 amino acid residues for another subunit, the alpha subunit. In addition, this clone contained the sequences that encode the COOH-terminal part of the beta subunit at its extended 5' region. The amino acid sequences and coding regions of the cDNAs of the three subunits are very similar (approximately 93-95% identity). A high degree of resemblance is observed also in their 3' noncoding regions. However, distinctive differences exist in the vicinity of the Zn-binding cysteine residue at position 46--i.e., Cys-Gly-Thr in the alpha, Cys-Arg-Thr in the wild-type beta 1, Cys-His-Thr in the Oriental-type beta 2, and Cys-Arg-Ser in the gamma, reflecting the differences in their kinetic properties. Based on the cDNA sequences and the deduced amino acid sequences of the three subunits, their structural and evolutionary relationships are discussed.
PMID: 2935875 [PubMed - indexed for MEDLINE]
Display Abstract Show 20 Sort by Send to

Secondary Database

- PROSITE (https://prosite.expasy.org/)
- PRINTS (http://130.88.97.239/PRINTS/index.php)
- ProDOM (http://prodom.prabi.fr/prodom/current/html/home.php)
- Pfam (https://pfam.xfam.org/)
- SCOP (http://scop.mrc-lmb.cam.ac.uk/)
- CATH (https://www.cathdb.info/)

prosite.expasy.org

cup.edu.in | My Favorite Machin... | iMod Server home... | Other bookmarks

ASy
natics Resource Portal

PROSITE

Home | Contact

Home | ScanProsite | ProRule | Documents | Downloads | Links | Funding

Database of protein domains, families and functional sites

New SARS-CoV-2 relevant PROSITE motifs

PROSITE consists of documentation entries describing protein domains, families and functional sites as well as associated patterns and profiles to identify them [More... / References / Commercial users].
PROSITE is complemented by ProRule , a collection of rules based on profiles and patterns, which increases the discriminatory power of profiles and patterns by providing additional information about functionally and/or structurally critical amino acids [More...].

Release 2020_01 of 26-Feb-2020 contains 1846 documentation entries, 1311 patterns, 1265 profiles and 1289 ProRule.

Search

e.g. PDOC00022, PS50089, SH3, zinc finger

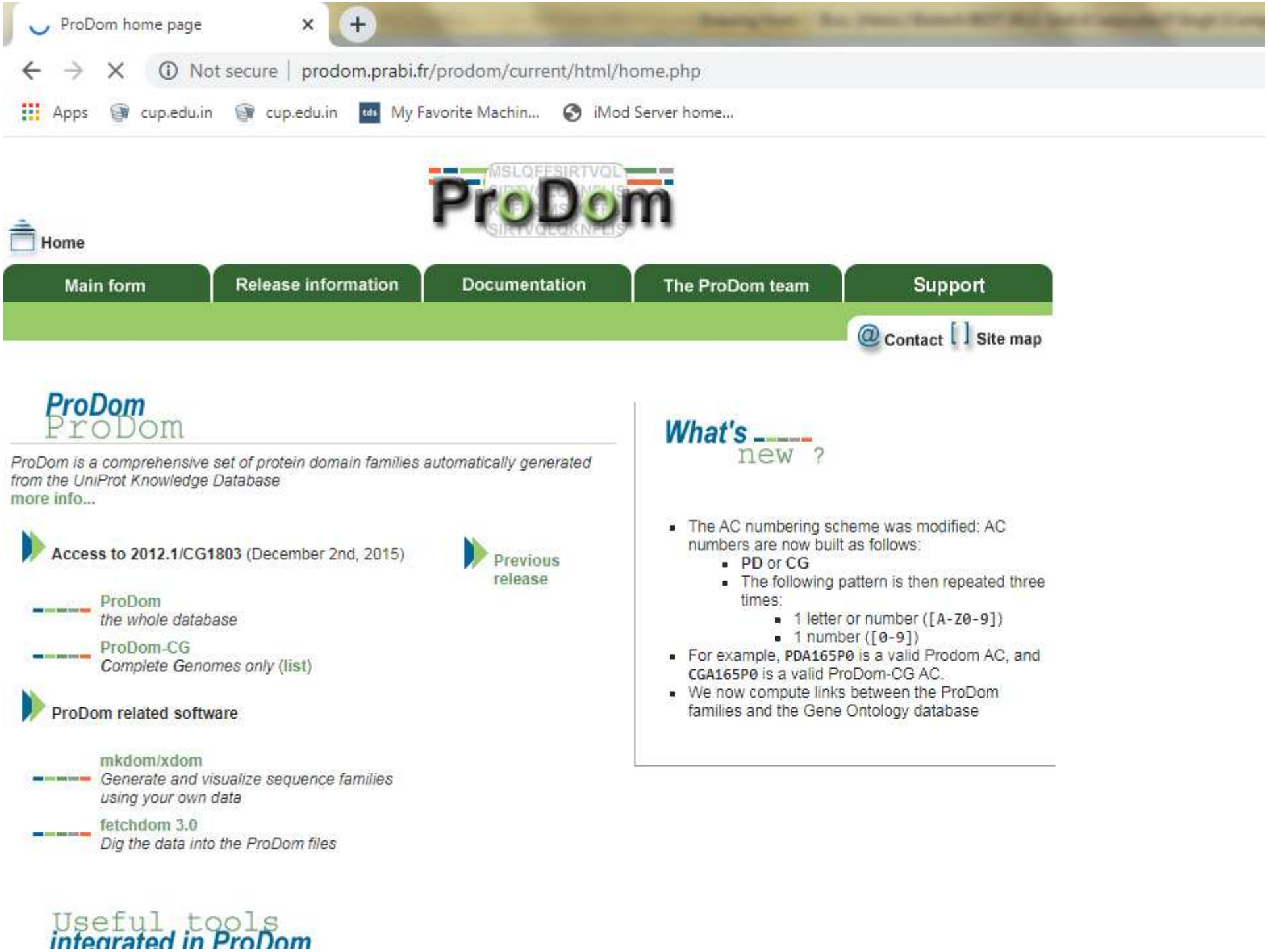
Search

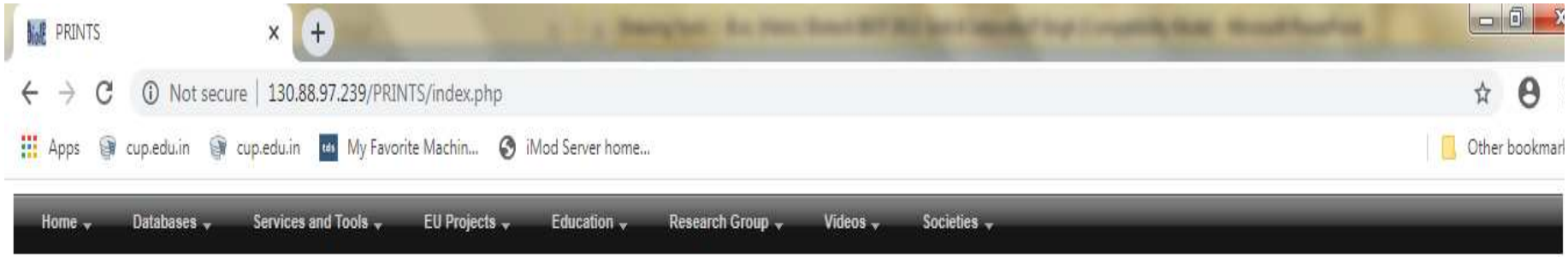
Browse

- by documentation entry
- by ProRule description
- by taxonomic scope
- by number of positive hits

Quick Scan mode of ScanProsite

Other tools


ProDom home page
Not secure | prodom.prabi.fr/prodom/current/html/home.php
Apps
cup.edu.in
cup.edu.in
My Favorite Machin...
iMod Server home...
ProDom
Home
Main form
Release information
Documentation
The ProDom team
Support
Contact
Site map
ProDom
ProDom is a comprehensive set of protein domain families automatically generated from the UniProt Knowledge Database
more info...
Access to 2012.1/CG1803 (December 2nd, 2015)
Previous release
ProDom
the whole database
ProDom-CG
Complete Genomes only (list)
ProDom related software
mkdom/xdom
Generate and visualize sequence families using your own data
fetchdom 3.0
Dig the data into the ProDom files
What's new ?
The AC numbering scheme was modified: AC numbers are now built as follows:
PD or CG
The following pattern is then repeated three times:
1 letter or number ([A-Z0-9])
1 number ([0-9])
For example, PDA165P0 is a valid Prodom AC, and CGA165P0 is a valid ProDom-CG AC.
We now compute links between the ProDom families and the Gene Ontology database
Useful tools
integrated in ProDom



PRINTS

PRINTS is a compendium of protein **fingerprints**. A fingerprint is a group of conserved motifs used to characterise a protein family; its diagnostic power is refined by iterative scanning of a *SWISS-PROT/TrEMBL* composite. Usually the motifs do not overlap, but are separated along a sequence, though they may be contiguous in 3D-space. Fingerprints can encode protein folds and functionalities more flexibly and powerfully than can single motifs, full diagnostic potency deriving from the mutual context provided by motif neighbours. References

Direct PRINTS access:

- By accession number
- By PRINTS code
- By database code
- By text
- By sequence
- By title
- By number of motifs
- By author
- By query language

PRINTS search:

- FPScan - search PRINTS with a query sequence/ID
- GRAPHScan - search a sequence with a named fingerprint
- FingerPRINTScan source is available: contact attwood@bioinf.man.ac.uk

m.xfam.org

cup.edu.in | My Favorite Machin... | iMod Server home...

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

keyword

Pfam 32.0 (September 2018, 17929 entries)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. **More...**

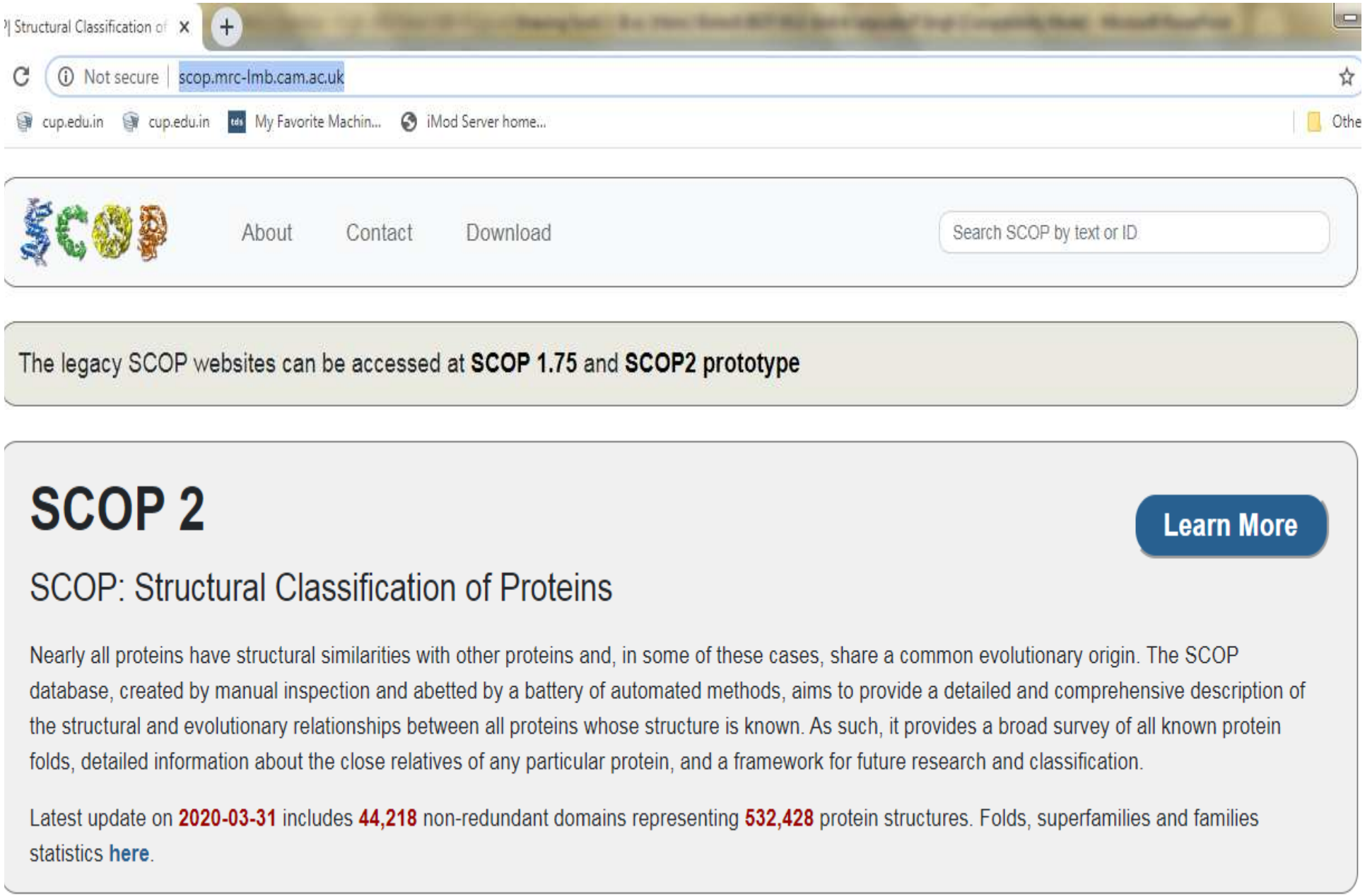
QUICK LINKS	YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...
SEQUENCE SEARCH	Analyze your protein sequence for Pfam matches
VIEW A PFAM ENTRY	View Pfam annotation and alignments
VIEW A CLAN	See groups of related entries
VIEW A SEQUENCE	Look at the domain organisation of a protein sequence
VIEW A STRUCTURE	Find the domains on a PDB structure
KEYWORD SEARCH	Query Pfam by keywords
JUMP TO	enter any accession or ID Go Example

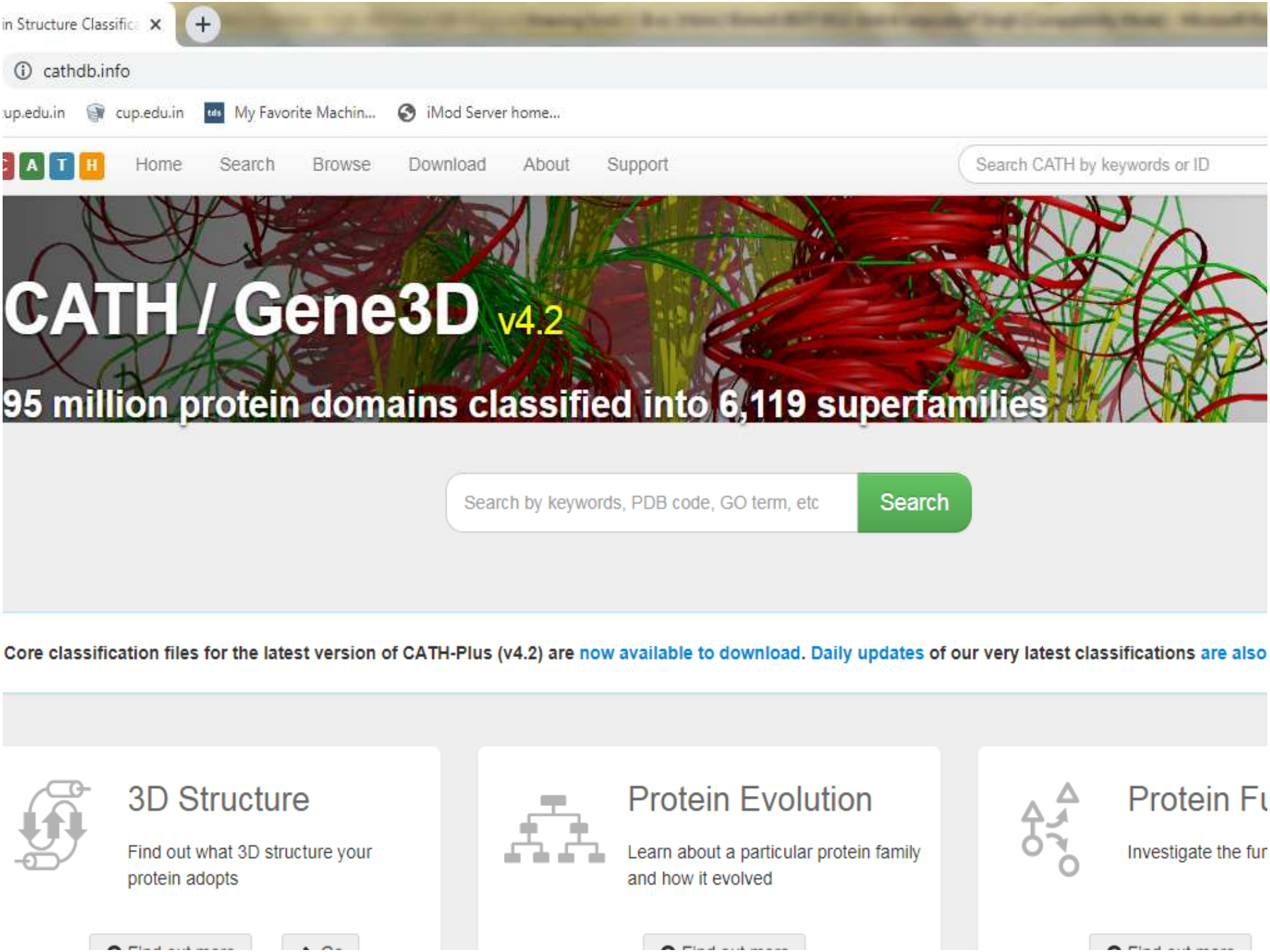
Enter any type of accession or ID to jump to the page for a Pfam entry or clan, UniProt sequence, PDB structure, etc.

Or view the help pages for more information

Recent Pfam blog posts

☒Hide this


Structural Classification of
Not secure
scop.mrc-lmb.cam.ac.uk
cup.edu.in
cup.edu.in
My Favorite Machin...
iMod Server home...
About
Contact
Download
Search SCOP by text or ID
The legacy SCOP websites can be accessed at SCOP 1.75 and SCOP2 prototype
SCOP 2
Learn More
SCOP: Structural Classification of Proteins
Nearly all proteins have structural similarities with other proteins and, in some of these cases, share a common evolutionary origin. The SCOP database, created by manual inspection and abetted by a battery of automated methods, aims to provide a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known. As such, it provides a broad survey of all known protein folds, detailed information about the close relatives of any particular protein, and a framework for future research and classification.
Latest update on 2020-03-31 includes 44,218 non-redundant domains representing 532,428 protein structures. Folds, superfamilies and families statistics here.


in Structure Classific
cathdb.info
up.edu.in
cup.edu.in
My Favorite Machin...
iMod Server home...
CATH
Home
Search
Browse
Download
About
Support
Search CATH by keywords or ID
CATH / Gene3D v4.2
95 million protein domains classified into 6,119 superfamilies
Search by keywords, PDB code, GO term, etc
Search
Core classification files for the latest version of CATH-Plus (v4.2) are now available to download. Daily updates of our very latest classifications are also
3D Structure
Find out what 3D structure your protein adopts
Protein Evolution
Learn about a particular protein family and how it evolved
Protein Fu
Investigate the fur

References

- http://www.nsscnilamel.org/images/Download/4198570bed66b665de0fb42e478bfe3c.pdf

Home Assignment

1. Discuss the scope of bioinformatics.
2. Differentiate between primary and secondary databases with examples.
3. Discuss the importance of biological databases.

Last data of submission 15.04.2020

Thank you.

Email: sprakashsingh@mgcub.ac.in