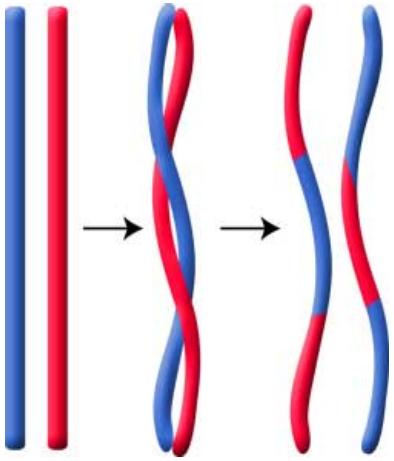
http://petang.cgu.edu.tw/bioinformatics/index.htm



Bioinformatics

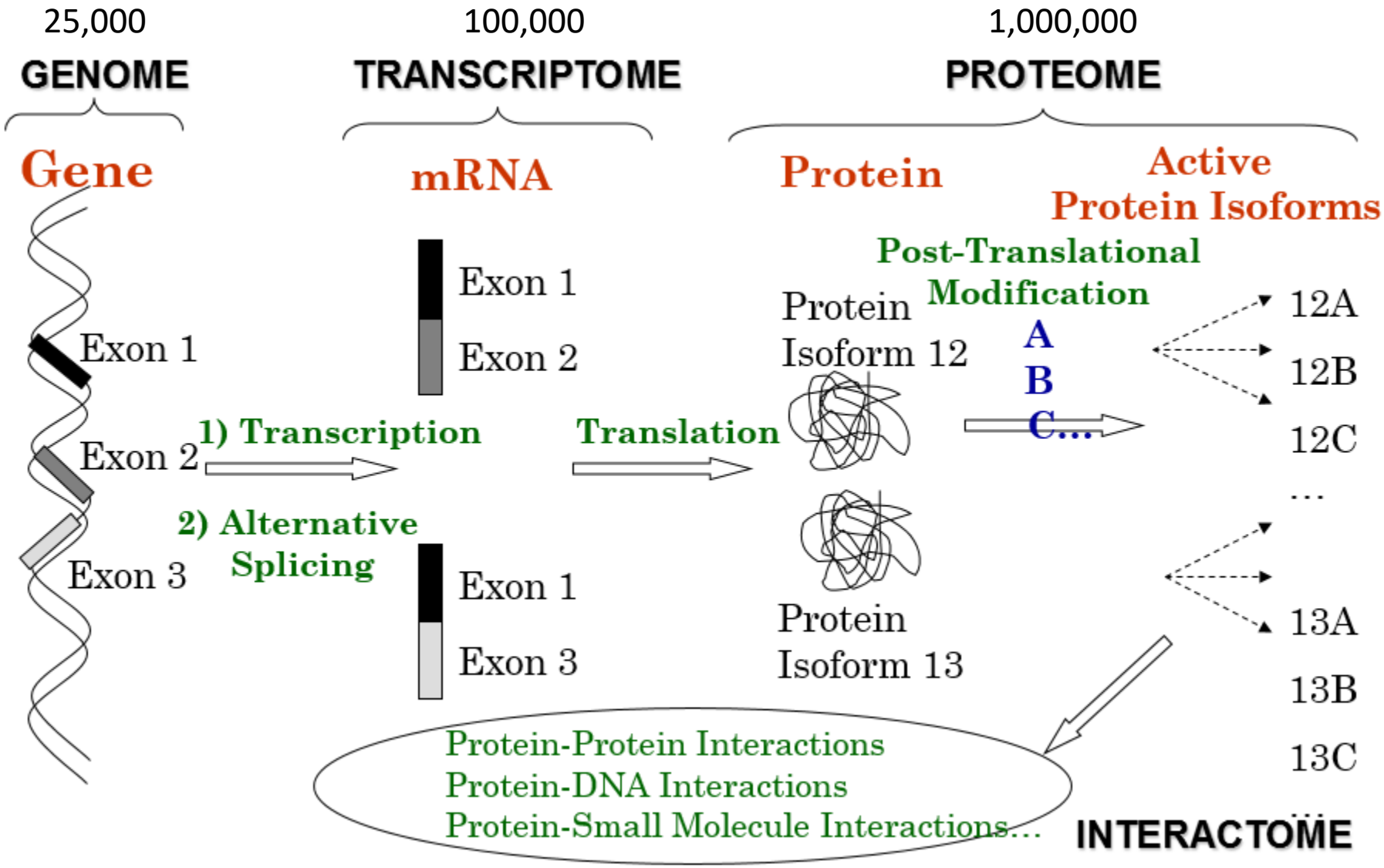
Lecture 8 – Protein Sequence Analysis

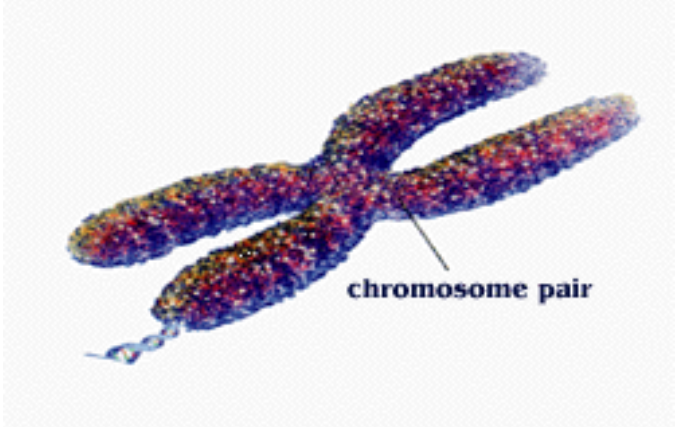
AGTCCGCGAATACAGGCTCGGT AGTCCGCGAATACAGGCTCGGT

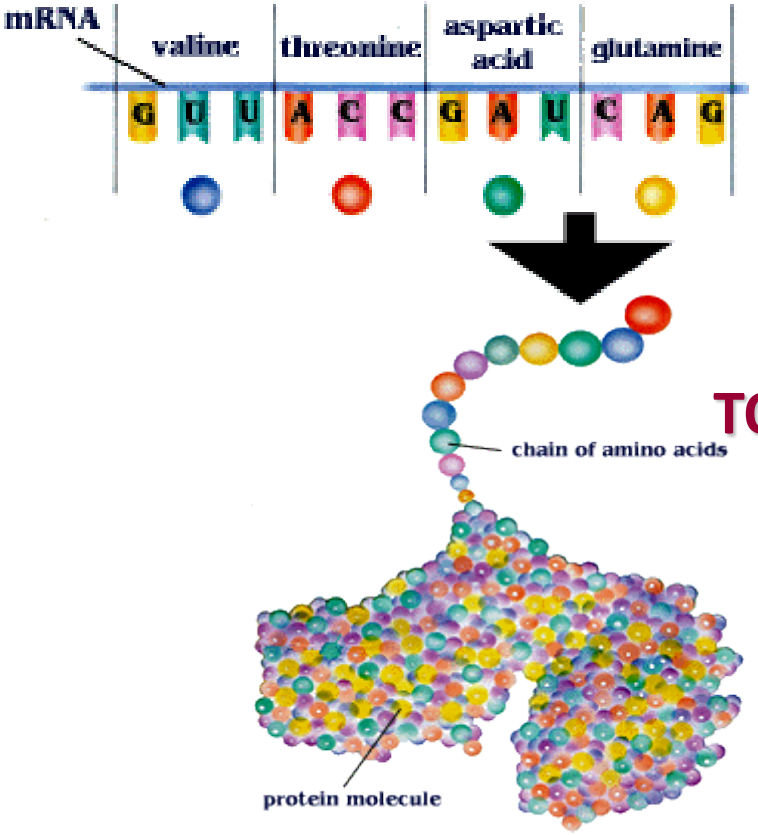
Petrus Tang, Ph.D. (鄧致剛**)**
Graduate Institute of Basic Medical Sciences
and
Bioinformatics Center, Chang Gung University.
petang@mail.cgu.edu.tw
EXT: 5136

助教：
蔡智宇(分機5690)

Why Proteomics: One-Gene-one-Protein?







PROTEIN DATABASES

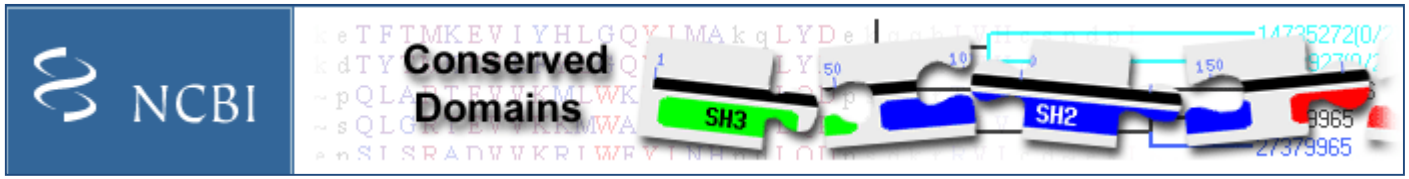
TOOLS

- **PROTEIN SEQUENCE**
- **PROPERTIES**
- **MOTIF/DOMAIN**
- **FOLDING**

Protein Sequence/Motif/Domain Databases



http://www.uniprot.org/



http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi



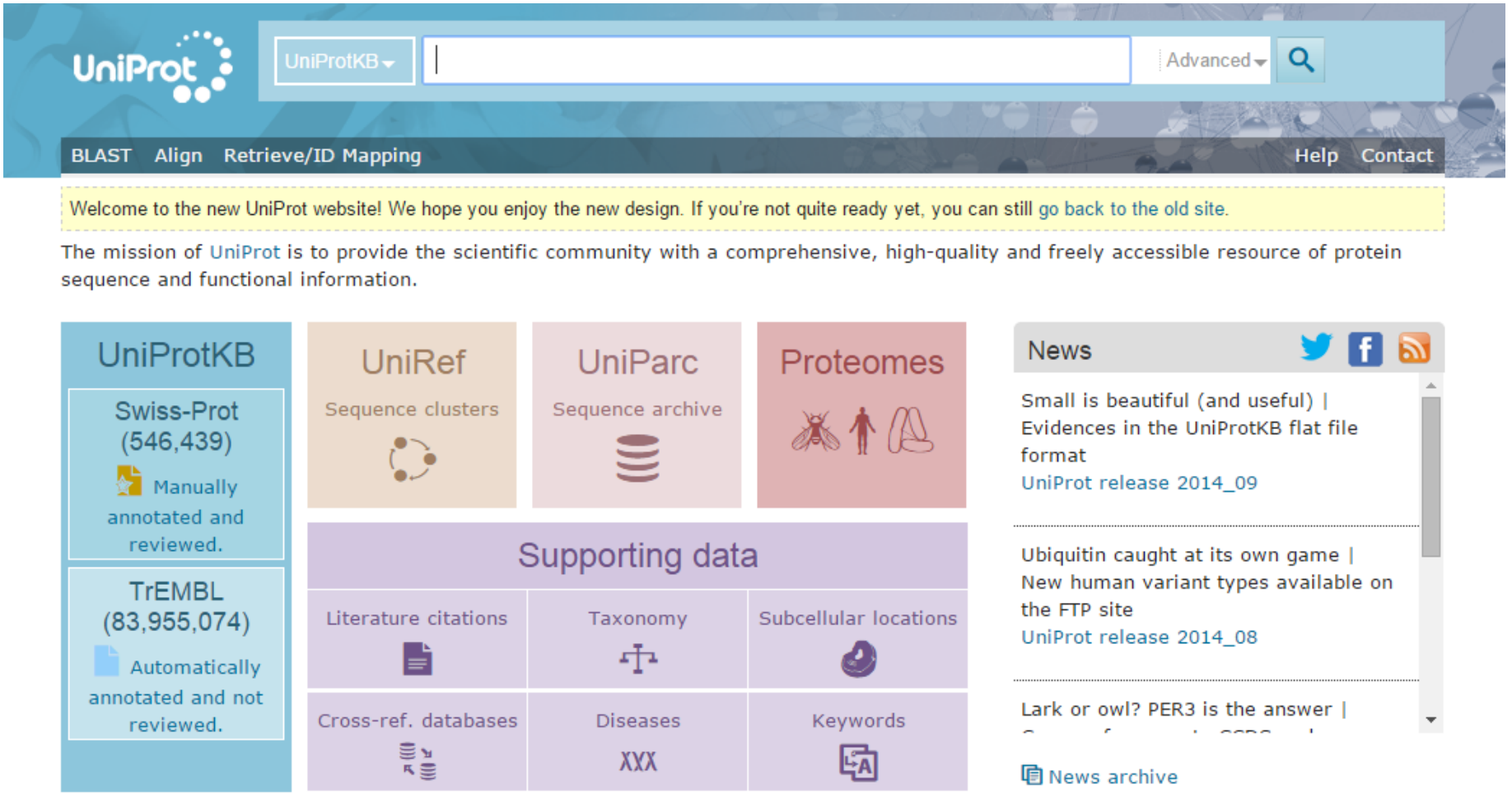
http://www.ebi.ac.uk/Tools/pfa/iprscan/

Protein Analysis Tools



http://tw.expasy.org/

http://www.uniprot.org/



Results

http://www.uniprot.org/uniprot/?query=p53&sort=score

Filter by

Reviewed (1,839) Swiss-Prot

Unreviewed (14,987) TrEMBL

Popular organisms

Human (1,056)
Mouse (803)
Rat (312)
Bovine (270)
Zebrafish (231)

Other organisms

Go

Search terms

Filter "p53" as:
gene name (119)
gene ontology (3,273)
organism (3,646)
protein family (630)
protein name (828)
strain (3,679)
taxonomy (3,680)

View by

Taxonomy
Keywords
Gene Ontology
Enzyme class
Pathway

UniRef

Your results in sequence clusters with identity of:
100%, 90% or 50%

Show help for UniProtKB

Basket

Columns | BLAST | Align | Download | Add to basket

1 to 25 of 16,826

Show 25

Entry	Entry name	Protein names	Gene names	Organism	Length
P04637	P53_HUMAN	**Cellular tumor antigen p53**	**TP53**, P53	Homo sapiens (Human)	393
P02340	P53_MOUSE	**Cellular tumor antigen p53**	**Tp53**, P53, Trp53	Mus musculus (Mouse)	387
P10361	P53_RAT	**Cellular tumor antigen p53**	**Tp53**, P53	Rattus norvegicus (Rat)	391
Q42578	PER53_ARATH	**Peroxidase 53**	**PER53**, P53, At5g06720, MPH15.8	Arabidopsis thaliana (Mouse-ear cress)	335
P25035	P53_ONCMY	**Cellular tumor antigen p53**	**tp53**, p53	Oncorhynchus mykiss (Rainbow trout) (Salmo gairdneri)	396
P79820	P53_ORYLA	**Cellular tumor antigen p53**	**tp53**, p53	Oryzias latipes (Medaka fish) (Japanese ricefish)	352
Q92143	P53_XIPMA	**Cellular tumor antigen p53**	**tp53**, p53	Xiphophorus maculatus (Southern platyfish) (Platypoecilus maculatus)	342
Q29537	P53_CANFA	**Cellular tumor antigen p53**	**TP53**, P53	Canis familiaris (Dog) (Canis lupus familiaris)	381
O09185	P53_CRIGR	**Cellular tumor antigen p53**	**TP53**, P53	Cricetulus griseus (Chinese hamster) (Cricetulus barabensis griseus)	393
Q8SPZ3	P53_DELLE	**Cellular tumor antigen p53**	**TP53**, P53	Delphinapterus leucas (Beluga whale)	387
P61260	P53_MACFU	**Cellular tumor antigen p53**	**TP53**, P53	Macaca fuscata fuscata (Japanese macaque)	393

Customize results table

Columns to be displayed

Reset to *default* | Save | Cancel

Drag and drop to re-order.

Entry | Entry name | Protein names | Gene names | Organism | Length

Add more columns

▶ Expand all

Search: *e.g. gene, ontology,...*

- ▶ Names & Taxonomy
- ▶ Sequences
- ▶ Function
- ▶ Miscellaneous
- ▶ Interaction
- ▶ Expression
- ▶ Gene Ontology (GO)
- ▶ Pathology & Biotech
- ▶ Subcellular location
- ▶ PTM / Processsing
- ▶ Structure
- ▶ Publications
- ▶ Date of
- ▶ Family & Domains
- ▶ Taxonomic lineage
- ▶ Taxonomy Id

Databases

- ▶ Sequence
- ▶ 3D structure
- ▶ Protein-protein
- ▶ Chemistry
- ▶ Protein family/group
- ▶ PTM
- ▶ Polymorphism
- ▶ 2D gel
- ▶ Proteomic
- ▶ Protocols and
- ▶ Genome annotation
- ▶ Organism-specific
- ▶ Phylogenomic
- ▶ Enzyme and
- ▶ Other
- ▶ Gene expression
- ▶ Family and domain

▼ Names & Taxonomy

- [x] *Entry name*
- [x] *Protein names*
- [x] *Gene names*
- [ ] Gene names (primary)
- [ ] Gene names (synonym)
- [ ] Gene names (ordered locus)
- [ ] Gene names (ORF)
- [x] *Organism*
- [ ] Organism ID
- [ ] Proteomes
- [ ] Taxonomic lineage
- [ ] Virus hosts

▼ Sequences

- [x] *Length*
- [ ] Fragment
- [ ] Gene encoded by
- [ ] Alternative products
- [ ] Erroneous gene model prediction
- [ ] Erroneous initiation
- [ ] Erroneous termination
- [ ] Erroneous translation
- [ ] Frameshift
- [ ] Mass spectrometry
- [ ] Polymorphism
- [ ] RNA editing
- [ ] Sequence caution
- [ ] Mass
- [ ] Sequence
- [ ] Alternative sequence
- [ ] Natural variant
- [ ] Non-adjacent residues
- [ ] Non-standard residue
- [ ] Non-terminal residue
- [ ] Sequence conflict
- [ ] Sequence uncertainty
- [ ] Sequence version

▼ Function

- [ ] EC number
- [ ] Absorption
- [ ] Catalytic activity
- [ ] Cofactor
- [ ] Enzyme regulation
- [ ] Function[CC][1]
- [ ] Kinetics
- [ ] Pathway
- [ ] Redox potential
- [ ] Temperature dependence
- [ ] pH dependence
- [ ] Active site
- [ ] Binding site
- [ ] DNA binding
- [ ] Metal binding
- [ ] Nucleotide binding
- [ ] Site

▼ Miscellaneous

- [x] 
- [ ] Annotation score
- [ ] Features
- [ ] Caution
- [ ] Miscellaneous[CC][1]
- [ ] Keywords
- [ ] Matched text
- [ ] Protein existence
- [ ] Tools

▶ Interaction

▶ Expression

▶ Gene Ontology (GO)

▶ Pathology & Biotech

▶ Subcellular location

▶ PTM / Processsing

▶ Structure

▶ Publications

▶ Date of

▶ Family & Domains

▶ Taxonomic lineage

▶ Taxonomy Id

Databases

▶ Sequence

▶ 3D structure

▶ Protein-protein interaction

▶ Chemistry

▶ Protein family/group

▶ PTM

▶ Polymorphism

▶ 2D gel

▶ Proteomic

▶ Protocols and materials

▶ Genome annotation

▶ Organism-specific

▶ Phylogenomic

▶ Enzyme and pathway

▶ Other

▶ Gene expression

▶ Family and domain

P04637 - P53_HUMAN

Protein: **Cellular tumor antigen p53**

Gene: **TP53**

Organism: *Homo sapiens (Human)*

Status: Reviewed - ●●●●●- Experimental evidence at protein level[i]

Display None

- FUNCTION
- NAMES & TAXONOMY
- SUBCELLULAR LOCATION
- PATHOLOGY & BIOTECH
- PTM / PROCESSING
- EXPRESSION
- INTERACTION
- STRUCTURE
- FAMILY & DOMAINS
- SEQUENCES (9)
- CROSS-REFERENCES
- PUBLICATIONS
- ENTRY INFORMATION
- MISCELLANEOUS

Top

BLAST | Align | Format | Add to basket | History | Comment (0) | Feedback | Help video

Function[i]

Acts as a tumor suppressor in many tumor types; induces growth arrest or apoptosis depending on the physiological circumstances and cell type. Involved in cell cycle regulation as a trans-activator that acts to negatively regulate cell division by controlling a set of genes required for this process. One of the activated genes is an inhibitor of cyclin-dependent kinases. Apoptosis induction seems to be mediated either by stimulation of BAX and FAS antigen expression, or by repression of Bcl-2 expression. In cooperation with mitochondrial PPIF is involved in activating oxidative stress-induced necrosis; the function is largely independent of transcription. Induces the transcription of long intergenic non-coding RNA p21 (lincRNA-p21) and lincRNA-Mkln1. LincRNA-p21 participates in TP53-dependent transcriptional repression leading to apoptosis and seem to have to effect on cell-cycle regulation. Implicated in Notch signaling cross-over. Prevents CDK7 kinase activity when associated to CAK complex in response to DNA damage, thus stopping cell cycle progression. Isoform 2 enhances the transactivation activity of isoform 1 from some but not all TP53-inducible promoters. Isoform 4 suppresses transactivation activity and impairs growth suppression mediated by isoform 1. Isoform 7 inhibits isoform 1-mediated apoptosis. 11 Publications

Cofactor[i]
Binds 1 zinc ion per subunit.

Sites

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Site[i]	120 – 120	1	Interaction with DNA			
Metal binding[i]	176 – 176	1	Zinc			
Metal binding[i]	179 – 179	1	Zinc			
Metal binding[i]	238 – 238	1	Zinc			
Metal binding[i]	242 – 242	1	Zinc			

PROSITE documentation PDOC00301 [for PROSITE entry PS00348]

p53 family signature

Description Technical section References Copyright Miscellaneous

Description

The p53 tumor antigen [1,2,3,4,5] is a protein found in increased amounts in a wide variety of transformed cells. It is also detectable in many proliferating nontransformed cells, but it is undetectable or present at low levels in resting cells. It is frequently mutated or inactivated in many types of cancer. p53 seems to act as a tumor suppressor in some, but probably not all, tumor types. p53 is probably involved in cell cycle regulation, and may be a trans-activator that acts to negatively regulate cellular division by controlling a set of genes required for this process.

p53 is a phosphoprotein of about 390 amino acids which can be subdivided into four domains: a highly charged acidic region of about 75 to 80 residues, a hydrophobic proline-rich domain (position 80 to 150), a central region (from 150 to about 300), and a highly basic C-terminal region. The sequence of p53 is well conserved in vertebrate species; attempts to identify p53 in other eukaryotic phylum has so far been unsuccessful.

The p53 protein belongs to a family [6] that also includes:

- p51 (p63 or Ket), a transcriptional activator.
- p73, a transcriptional activator that participates in the apoptotic response to DNA damage.

As a signature pattern for this family we selected a conserved stretch of 13 residues located in the central region of the protein. This region, known as domain IV in [3], is involved (along with an adjacent region) in the binding of the large T antigen of SV40. In man this region is the focus of a variety of point mutations in cancerous tumors.

Last update:

December 2001 / Pattern and text revised.

Technical section

PROSITE method (with tools and information) covered by this documentation:

P53, PS00348; p53 family signature (PATTERN)

- Consensus pattern:
  M-C-N-S-S-C-[MV]-G-G-M-N-R-R
- Sequences in UniProtKB/Swiss-Prot known to belong to this class: 40
  - detected by PS00348: 40 (true positives)
  - undetected by PS00348: 0 (false negative or 'partial')
- Other sequence(s) in UniProtKB/Swiss-Prot detected by PS00348:
  NONE.
- Retrieve an alignment of UniProtKB/Swiss-Prot true positive hits:
  Clustal format, color, condensed view / Clustal format, color / Clustal format, plain text / Fasta format
- Retrieve the sequence logo from the alignment
- Taxonomic tree view of all UniProtKB (Swiss-Prot + TrEMBL) entries matching PS00348
- Retrieve a list of all UniProtKB (Swiss-Prot + TrEMBL) entries matching PS00348
- Scan UniProtKB (Swiss-Prot and/or TrEMBL) entries against PS00348
- View ligand binding statistics of PS00348
- Matching PDB structures: 1GZH 1HU8 1KZY 1TSR ... [ALL]

Prosite search results

Search in PROSITE for: casein kinase

(Release 20.76, of 17-Oct-2011)

Enter search terms:

casein kinase

☐ Prefix and append wildcard '*' to words.

new search clear

By default, this search engine searches for complete words only. If you did not find
would try to do a substring match, you should perform a new search and select 'pre
to words'.

Number of documents in PROSITE containing the search term:2

- PDOC00006 Casein kinase II phosphorylation site
- PDOC00845 Casein kinase II regulatory subunit signature

Casein kinase II regulatory subunit signature

Description:

Casein kinase II (CK-2) [1] is an ubiquitous eukaryotic serine/threonine protein kinase which is found both in the cytoplasm and the nucleus and whose substrates are numerous. It generally phosphorylates Ser or Thr at the N-terminal of stretch of acidic residues (see <PDOC00006>). CK-2 exists as an heterotetramer composed of two catalytic subunits (α) and two regulatory subunits (β). In most species there are two closely related isoforms of the catalytic subunit: α and α'. Some species, such as fungi and plants, express two forms of regulatory subunits: β and β'. The β subunit plays a complex role in regulating the basal catalytic activity of the α subunit. The β subunit is a highly conserved protein of about 25 Kd that contains, in its central section, a cysteine-rich motif that could be involved in binding a metal such as zinc [2]. We have used this region as a signature pattern.

Last update:

December 2004 / Pattern and text revised.

Technical section:

PROSITE method (with tools and information) covered by this documentation:

CK2_BETA, PS01101; **Casein kinase II regulatory subunit signature** (PATTERN)

Consensus pattern:	**C-P-x-[LIVMYAT]-x-C-x(5)-[LI]-P-[LIVMCA]-G-x(9)-V-[KRM]-x(2)-C-[PA]-x-C**
Sequences known to belong to this class detected by the pattern:	ALL
Other sequence(s) detected in Swiss-Prot:	NONE

- Retrieve an alignment of Swiss-Prot true positive hits:
  Clustal format, color, condensed view / Clustal format, color / Clustal format, plain text / Fasta format
- Retrieve the sequence logo from the alignment
- Taxonomic tree view of all Swiss-Prot/TrEMBL entries matching PS01101
- Retrieve a list of all Swiss-Prot/TrEMBL entries matching PS01101
- Scan Swiss-Prot/TrEMBL entries against PS01101
- view ligand binding statistics

Casein kinase II phosphorylation site

Description:

Casein kinase II (CK-2) is a protein serine/threonine kinase whose activity is independent of cyclic nucleotides and calcium. CK-2 many different proteins. The substrate specificity [1] of this enzyme can be summarized as follows:

```
(1) Under comparable conditions Ser is favored over Thr.
(2) An acidic residue (either Asp or Glu) must be present three residues from
    the C-terminal of the phosphate acceptor site.
(3) Additional acidic  residues in  positions +1, +2, +4, and +5 increase the
    phosphorylation rate.  Most  physiological  substrates  have at least one
    acidic residue in these positions.
(4) Asp is preferred to Glu as the provider of acidic determinants.
(5) A basic residue at the N-terminal  of the  acceptor  site  decreases  the
    phosphorylation rate, while an acidic one will increase it.
```

Note:

This pattern is found in most of the known physiological substrates.

Last update:

May 1991 / Text revised.

Technical section:

PROSITE method (with tools and information) covered by this documentation:

CK2_PHOSPHO_SITE, PS00006; **Casein kinase II phosphorylation site** (PATTERN with a high probability of occurrence!)

Consensus pattern: **[ST]-x(2)-[DE][SorTisthephosphorylationsite]**

- Scan Swiss-Prot/TrEMBL entries against PS00006
- view ligand binding statistics

http://prosite.expasy.org/

Database of protein domains, families and functional sites

PROSITE consists of documentation entries describing protein domains, families and functional sites as well as associated patterns and profiles to identify them [More details / 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 details].

Release 20.76, of 17-Oct-2011 (1627 documentation entries, 1308 patterns, 946 profiles and 944 ProRule)

PROSITE access

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

Search

add wildcard '*'

Browse:

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

PROSITE tools

Scan a sequence against PROSITE patterns and profiles - quick scan

(Output includes graphical view and feature detection)



Enter your sequence or a UniProtKB (Swiss-Prot or TrEMBL) ID or AC [help]:

Scan Clear

exclude patterns with a high probability of occurrence

- **ScanProsite** - advanced scan
- **PRATT** - allows to interactively generate conserved patterns from a series of unaligned proteins.
- **MyDomains - Image Creator** new - allows to generate custom domain figures.



Sequences (9)

Sequence status: Complete.

This entry describes **9** isoforms produced by **alternative promoter usage** and **alternative splicing**. Align

Isoform 1 (identifier: **P04637-1**) [UniParc] FASTA Add to Basket

Also known as: p53, p53alpha

This isoform has been chosen as the 'canonical' sequence. All positional information in this entry refers to it. This is also the sequence that appears in the downloadable versions of the entry.

« Hide

```
>sp|P04637|P53_HUMAN Cellular tumor antigen p53 OS=Homo sapiens GN=TP53 PE=1 SV=4
MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGP
DEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAK
SVTCTYSPALNKMFCQLAKTCPVQLWVDSTPPPGTRVRAMAIYKQSQHMTEVVRRCPHHE
RCSDSDGLAPPQHLIRVEGNLRVEYLDDRNTFRHSVVVPYEPPEVGSDCTTIHYNYMCNS
SCMGGMNRRPILTIITLEDSSGNLLGRNSFEVRVCACPGRDRRTEEENLRKKGEPHHELP
PGSTKRALPNNTSSSPQPKKKPLDGEYFTLQIRGRERFEMFRELNEALELKDAQAGKEPG
GSRAHSSHLKSKKGQSTSRHKKLMFKTEGPDSD
```

Copy this sequence to your computer

```
FS DLWKLLPENN VLSPLPSQAM DDLMLSPDDI    50
AA PPVAPAPAAP TPAAPAPAPS WPLSSSVPSQ   100
AK SVTCTYSPAL NKMFCQLAKT CPVQLWVDST   150
MT EVVRRCPHHE RCSDSDGLAP PQHLIRVEGN   200
LRVEYLDDRN TFRHSVVVPY EPPEVGSDCT TIHYNYMCNS SCMGGMNRRP   250
ILTIITLEDS SGNLLGRNSF EVRVCACPGR DRRTEEENLR KKGEPHHELP   300
PGSTKRALPN NTSSSPQPKK KPLDGEYFTL QIRGRERFEM FRELNEALEL   350
KDAQAGKEPG GSRAHSSHLK SKKGQSTSRH KKLMFKTEGP DSD          393
```

Length: 393
Mass (Da): 43,653
Last modified: November 24, 2009 - v4
Checksum: AD5C149FD8106131

BLAST GO

Isoform 2 (identifier: **P04637-2**) [UniParc] FASTA Add to Basket

Also known as: I9RET, p53beta

The sequence of this isoform differs from the canonical sequence as follows:

332-341: IRGRERFEMF → DQTSFQKENC

342-393: Missing.

Note: Expressed in quiescent lymphocytes. Seems to be non-functional. May be produced at very low levels due to a premature stop codon in the mRNA, leading to nonsense-mediated mRNA decay.

Show »

Length: 341
Mass (Da): 37,826
Checksum: 2C5FE7A14A575E43

BLAST GO

Isoform 3 (identifier: **P04637-3**) [UniParc] FASTA Add to Basket

Also known as: p53gamma

The sequence of this isoform differs from the canonical sequence as follows:

332-346: IRGRERFEMFRELNE → MLLDLRWCYFLINSS

347-393: Missing.

Note: Expressed in quiescent lymphocytes. Seems to be non-functional. May be produced at very low levels due to a premature stop codon in the mRNA, leading to nonsense-mediated mRNA decay.

Show »

Length: 346
Mass (Da): 38,501
Checksum: 6F18E09F8CD9129F

BLAST GO

Example

cellular tumor antigen p53 isoform a [Homo sapiens]

>sp|P04637|P53_HUMAN Cellular tumor antigen p53 OS=Homo sapiens GN=TP53 PE=1 SV=4
MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGP
DEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAK
SVTCTYSPALNKMFCQLAKTCPVQLWVDSTPPPGTRVRAMAIYKQSQHMTEVVRRCPHHE
RCSDSDGLAPPQHLIRVEGNLRVEYLDDRNTFRHSVVVPYEPPEVGSDCTTIHYNYMCNS
SCMGGMNRRPILTIITLEDSSGNLLGRNSFEVRVCACPGRDRRTEEENLRKKGEPHHELP
PGSTKRALPNNTSSSPQPKKKPLDGEYFTLQIRGRERFEMFRELNEALELKDAQAGKEPG
GSRAHSSHLKSKKGQSTSRHKKLMFKTEGPDSD

UniProt

UniProtKB | Advanced

BLAST | Align | Retrieve/ID Mapping | Help | Contact

How to use this tool

The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences, which can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

1. Enter either a protein or nucleotide sequence or a UniProt identifier (e.g.P00750 or A4_HUMAN or UPI0000000001) into the form field.
2. Optionally, change the program parameters with the dropdown menus under the form.
3. Click the *Run BLAST* button.

Help | Tutorials and Videos | Downloads

BLAST

MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGP
DEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAK
SVTCTYSPALNKMFCQLAKTCPVQLWVDSTPPPGTRVRAMAIYKQSQHMTEVVRRCPHHE
RCSDSDGLAPPQHLIRVEGNLRVEYLDDRNTFRHSVVVPYEPPEVGSDCTTIHYNYMCNS
SCMGGMNRRPILTIIITLEDSSGNLLGRNSFEVRVCACPGRDRRTEEENLRKKGEPHHELP
PGSTKRALPNNTSSSPQPKKKPLDGEYFTLQIRGRERFEMFRELNEALELKDAQAGKEPG
GSRAHSSHLKSKKGQSTSRHKKLMFKTEGPDSD

Target database[i] | E-Threshold[i] | Matrix[i] | Filtering[i] | Gapped[i] | Hits[i]

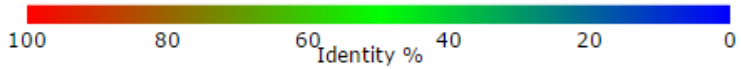
UniProtKB | 10 | Auto | None | yes | 250

Run Blast in a separate window.

Run BLAST | Clear



BLAST



Basket

Filter by

- Reviewed (59) Swiss-Prot
- Unreviewed (180) TrEMBL
- With 3D structure (4)
- Proteomes (100)

Organisms

- Human (51)
- Mouse (18)
- Zebrafish (7)
- Rat (4)
- Bovine (3)
- Other organisms

Go

Map To

- UniProtKB
- UniRef
- UniParc

View by

- Taxonomy
- Text version
- XML version

Demo

- Help video

Edit and resubmit

Order by: Score | Taxonomy: All

Overview

Show all 239

Entry	Protein names	Match hit (200 400 600)	Identity
H2QC53	Cellular tumor antigen p53 (Pan troglodytes)		100.0%
K7PPA8	Cellular tumor antigen p53 (Homo sapiens)		100.0%
P04637	Cellular tumor antigen p53 (Homo sapiens)		100.0%
G3R2U9	Cellular tumor antigen p53 (Gorilla gorilla gorilla)		99.0%

Alignments

Columns | BLAST | Align | Download | Add to basket

1 to 25 of 239 | Show 25

Entry	Alignment overview	Info	Status
Query: B20141016F000TRFJRY			
H2QC53	H2QC53_PANTR - Cellular tumor antigen p53 - Pan troglodytes ... - View alignment	E-value: 0.0 Score: 2,121 Ident.: 100.0%	
K7PPA8	K7PPA8_HUMAN - Cellular tumor antigen p53Homo sapiens (Human) - View alignment	E-value: 0.0 Score: 2,121 Ident.: 100.0%	
P04637	P53_HUMAN - Cellular tumor antigen p53Homo sapiens (Human) - View alignment	E-value: 0.0 Score: 2,121 Ident.: 100.0%	

http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml



NCBI

Conserved Domains

Conserved Domains and Protein Classification

RESOURCES | SEARCH | HOW TO | HELP | NEWS | FTP | PUBLICATIONS | DISCOVER

Resources

Conserved Domain Database (CDD)

CDD is a protein annotation resource that consists of a collection of well-annotated multiple sequence alignment models for ancient domains and full-length proteins. These are available as position-specific score matrices (PSSMs) for fast identification of conserved domains in protein sequences via RPS-BLAST. CDD content includes NCBI-curated domains, which use 3D-structure information to explicitly define domain boundaries and provide insights into **sequence/structure/function relationships**, as well as domain models imported from a number of external source databases (Pfam, SMART, COG, PRK, TIGRFAM).

Search | How To | Help | News | FTP | Publications

CD-Search & Batch CD-Search

CD-Search is NCBI's interface to searching the Conserved Domain Database with protein or nucleotide query sequences. It uses RPS-BLAST, a variant of PSI-BLAST, to quickly scan a set of pre-calculated position-specific scoring matrices (PSSMs) with a protein query. The results of CD-Search are presented as an annotation of protein domains on the user query sequence (illustrated example), and can be visualized as domain multiple sequence alignments with embedded user queries. High confidence associations between a query sequence and conserved domains are shown as **specific hits**. The CD-Search Help provides additional details, including information about running CD-Search locally.

Batch CD-Search serves as both a web application and a script interface for a conserved domain search on multiple protein sequences, accepting up to 100,000 proteins in a single job. It enables you to view a graphical display of the concise or full search result for any individual protein from your input list, or to download the results for the complete set of proteins. The Batch CD-Search Help provides additional details.

CD-Search (Help & FTP) | Batch CD-Search (Help) | Publications

CDART: Domain Architectures

Conserved Domain Architecture Retrieval Tool (CDART) performs similarity searches of the Entrez Protein database based on domain architecture, defined as the sequential order of conserved domains in protein queries. CDART finds protein similarities across significant evolutionary distances using sensitive domain profiles rather than direct sequence similarity. Proteins similar to the query are grouped and scored by architecture. You can search CDART directly with a query protein sequence, or, if a sequence of interest is already in the Entrez Protein database, simply retrieve the record, open its "*Links*" menu, and select "**Domain Relatives**" to see the precalculated CDART results (illustrated example). Relying on domain profiles allows CDART to be fast and, because it relies on annotated functional domains, informative.

About | Search | Help | FTP | Publications

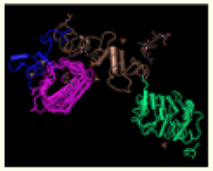
CDTree

CDTree is a helper application for your web browser that allows you to interactively view and examine conserved domain hierarchies curated at NCBI. CDTree works with Cn3D as its alignment viewer/editor, it is used in the CDD curation process and is a both **classification and research tool** for functional annotation and the study of protein and protein domain families.

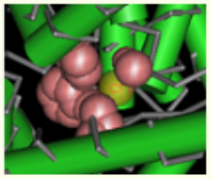
About | Install | Publications

Highlights

What is a conserved domain?



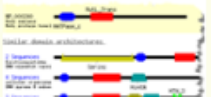
3-D structures and conserved core motifs:



Conserved features (binding and catalytic sites)



Proteins with Similar Domain Architectures



Conserved Domains and Protein Classification

RESOURCES

Search Methods: Quick Start Guide

Text Term Search

Retrieve conserved domain records that contain a term(s) of interest (e.g., chloride channel). Enter the terms in the query box at the top of this page or use the **Entrez** Conserved Domains Database (CDD) home page. See the help document for search tips, including a list of available search fields and examples of their use.

Protein or Nucleotide Query Sequence

Enter a protein or nucleotide query as an accession or GI number (e.g., AAC50285 or 463989), or as a sequence in FASTA format, to identify the protein's conserved domains and therefore its putative function:

送出查詢 Search Database CDD v3.12 - 46675 PSSMs

```
PGSTKRALPNNTSSSPQPKKKPLDGEYFTLQIRGRERFEMFRELN
EALELKDAQAGKEPG
GSRAHSSHLKSKKGQSTSRHKKLMFKTEGPDSD
```

The text box above provides a short cut to the **CD-Search** tool, using its default parameters. The help document provides details about database selection and search results. To view/change the default parameters, or to use advanced search options, enter your query directly on the **CD-Search** home page. *(Note: If a sequence of interest is already in the Entrez Protein database, you can simply follow the "**Conserved Domains**" link for that sequence record to view its pre-computed CD-Search results.)*

Batch of Protein Sequences

The **Batch CD-Search** tool allows the computation and download of conserved domain annotation for large sets of protein queries. Input up to 100,000 protein query sequences as a list of sequence identifiers and/or raw sequence data, then download output in a variety of formats (including tab-delimited text files) or view the search results graphically. See the help document for additional details, including information on using Batch CD-Search for scripted data downloads.

Direct fetch via UID

Retrieve a conserved domain record directly from the backend database by entering its unique identifier (UID), in the form of an accession (e.g., cd00400) or PSSM ID (e.g., 79359), in the text box below:

Fetch CDs via UID

(Note: the "text term search" function also allows you to enter either of those unique identifiers (UIDs), but it first searches the Entrez indices for the UID, then retrieves the record. The "direct fetch via UID" option bypasses the Entrez indices and simply retrieves the specified record.)

Find proteins with similar domain architectures

Enter a **protein query** as an accession or GI number (e.g., AAC50285 or 463989), or as a sequence in FASTA format, on the Conserved Domain Architecture Retrieval Tool (CDART) page to find other proteins with similar domain architectures. *(Note: If a sequence of interest is already in the Entrez Protein database, you can simply select "**Domain Relatives**" in that sequence record's "**Links**" menu to find proteins with similar architecture.)*

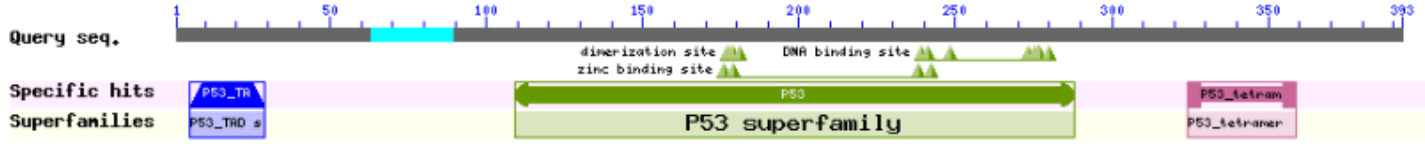
Conserved domains on [lcl|local_MEEPQSDPSV]

View: Concise Results

Local query sequence

Graphical summary

show options »



Search for similar domain architectures | Refine search

List of domain hits

Name	Accession	Description	Interval	E-value
[−] P53	cd08367	P53 DNA-binding domain; P53 is a tumor suppressor gene product; mutations in p53 or lack of ...	109-288	6.36e-96

P53 DNA-binding domain; P53 is a tumor suppressor gene product; mutations in p53 or lack of expression are found associated with a large fraction of all human cancers. P53 is activated by DNA damage and acts as a regulator of gene expression that ultimatively blocks progression through the cell cycle. P53 binds to DNA as a tetrameric transcription factor. In its inactive form, p53 is bound to the ring finger protein Mdm2, which promotes its ubiquitinylation and subsequent proteosomal degradation. Phosphorylation of p53 disrupts the Mdm2-p53 complex, while the stable and active p53 binds to regulatory regions of its target genes, such as the cyclin-kinase inhibitor p21, which complexes and inactivates cdk2 and other cyclin complexes.

:

Pssm-ID: 176262 **Cd Length:** 179 **Bit Score:** 286.86 **E-value:** 6.36e-96

```
                                  10        20        30        40        50        60        70        80
                                  ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
lcl|local_MEEPQSDPSV 109 FRLGFLHSGTAKSVTCTYSPALNKMFCQLAKTCPVQLWVDSTPPPGTRVRAMAIYKQSQHMTEVVRRCPHHERCsDSDGL 188
Cdd:cd08367           1 FEVTLDESGVAKSSTWTYSPKLNKLFVKMAKTCPIQFKVNPSPPPGLYVRAMLVYKDPEHVKEPVERCPNHRQG-DDGHT 79

                                  90       100       110       120       130       140       150       160
                                  ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
lcl|local_MEEPQSDPSV 189 APPQHLIRVEgNLRVEYLDDRNTFRHSVVVPYEPPEVGSDCTTIHYNYMCNSSCMGGMNRRPILTIITLEDSSGNLLGRN 268
Cdd:cd08367          80 APNSHVIRCE-NPQAEYVGDAFTGRLSVVVPLEPPQVGSEYVTVLLQFMCQNSCPGGINRRPIQLVFTLEDENGNVLGRR 158

                                 170       180
                                 ....*....|....*....|
lcl|local_MEEPQSDPSV 269 SFEVRVCACPGRDRRTEEEN 288
Cdd:cd08367          159 VIEVRVCACPGRDRKNEEKA 178
```

[+] P53_tetramer	pfam07710	P53 tetramerization motif;	324-359	1.34e-15
[+] P53_TAD	pfam08563	P53 transactivation motif; The binding of the p53 transactivation domain by regulatory ...	5-29	5.36e-07

cd08367: **P53** ?



P53 DNA-binding domain

P53 is a tumor suppressor gene product; mutations in p53 or lack of expression are found associated with a large fraction of all human cancers. P53 is activated by DNA damage and acts as a regulator of gene expression that ultimatively blocks progression through the cell cycle. P53 binds to DNA as a tetrameric transcription factor. In its inactive form, p53 is bound to the ring finger protein Mdm2, which promotes its ubiquitinylation and subsequent proteosomal degradation. Phosphorylation of p53 disrupts the Mdm2-p53 complex, while the stable and active p53 binds to regulatory regions of its target genes, such as the cyclin-kinase inhibitor p21, which complexes and inactivates cdk2 and other cyclin complexes.

Links ?

Source: pfam
Taxonomy: Fungi/Metazoa group
PubMed: 10 links
Book: 20 links
Protein: Representatives
Specific Protein
Related Protein
Related Structure
Architectures
Superfamily: cl14608
BioSystems: 1056 links

Conserved Features/Sites ? | PubMed References ? | Book References ?

DNA binding | zinc binding | dimerization

Feature 1: DNA binding site [nucleic acid binding site]



Evidence:

- **Structure:** 2AC0: Human p53 tetramer binds DNA, contacts at 4 A - View structure with Cn3D
- **Citation:** PMID 16793544
- **Structure:** 2GEQ: A mouse P53 core dimer binds DNA, contacts at 4 A - View structure with Cn3D
- **Citation:** PMID 16717092

Download Cn3D for Viewing 3D Structure

Scroll to Sequence Alignment Display

BLAST® Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

My NCBI [Sign In] [Register]

NCBI/ BLAST Home

BLAST finds regions of similarity between biological sequences. more...

New DELTA-BLAST, a more sensitive protein-protein search Go

BLAST Assembled Genomes

Find Genomic BLAST pages:

Enter organism name or id--completions will be suggested GO

- Human
- Mouse
- Rat
- Cow
- Pig
- Dog
- Rabbit
- Chimp
- Guinea pig
- Fruit fly
- Honey bee
- Chicken
- Zebrafish
- Clawed frog
- *Arabidopsis*
- Rice
- Yeast
- Microbes

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast, delta-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

Your Recent Results New!

All Recent results...

News

Find Genomic BLAST pages

You can now find Genomic BLAST pages using the search box from the BLAST homepage.

Thu, 02 Oct 2014 11:00:00 EST

More BLAST news...

Tip of the Day

More tips...

http://blast.ncbi.nlm.nih.gov/Blast.cgi



BLAST® Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

My NCBI [Sign In] [Register]

NCBI/ BLAST/ blastp suite

Standard Protein BLAST

blastn | blastp | blastx | tblastn | tblastx

BLASTP programs search protein databases using a protein query. more... Reset page Bookmark

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

Query subrange

From

To

```
SVTCTYSPALNKMFCQLAKTCPVQLWVDSTPPPGTRVRAMAIYKQSQHMTEVVRRCPHHE
RCSDSDGLAPPQHLIRVEGNLRVEYLDDRNTFRHSVVVPYEPPEVGSDCTTIHYNYMCNS
SCMGGMNRRPILTIITLEDSSGNLLGRNSFEVRVCACPGRDRRTEEENLRKKGEPHHELP
PGSTKRALPNNTSSSPQPKKKPLDGEYFTLQIRGRERFEMFRELNEALELKDAQAGKEPG
GSRAHSSHLKSKKGQSTSRHKKLMFKTEGPDSD
```

Or, upload file 選擇檔案 未選擇任何檔案

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism Optional: Enter organism name or id--completions will be suggested Exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude Optional: Models (XM/XP) Uncultured/environmental sample sequences

Entrez Query Optional: YouTube Create custom database

Enter an Entrez query to limit search

Program Selection

Algorithm:

- blastp (protein-protein BLAST)
- PSI-BLAST (Position-Specific Iterated BLAST)
- PHI-BLAST (Pattern Hit Initiated BLAST)
- DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

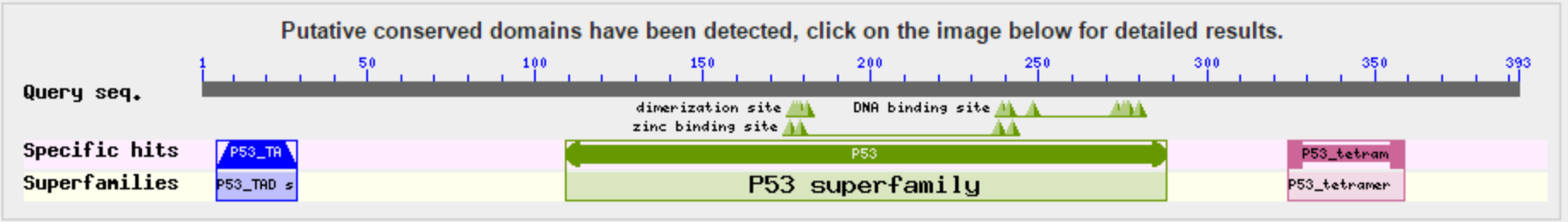
BLAST

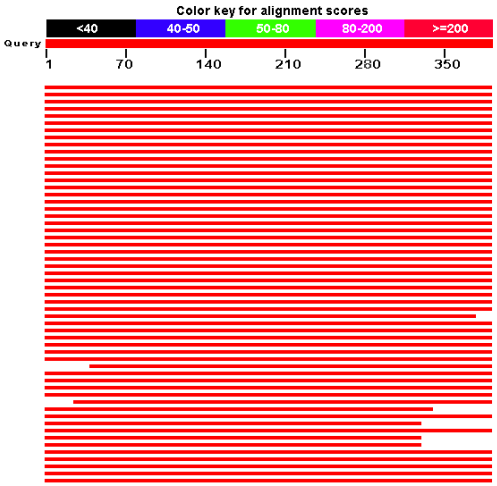
Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)

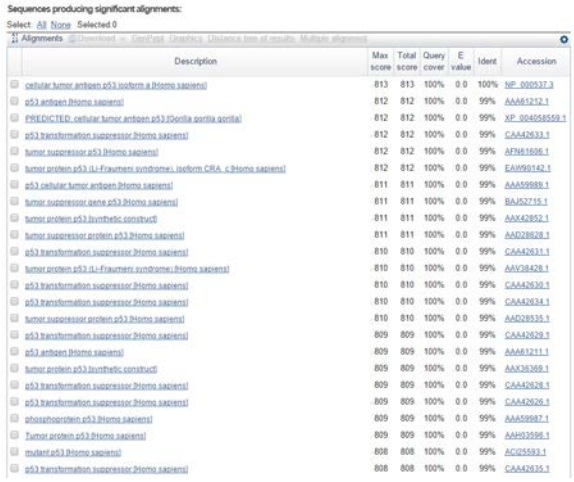
Show results in a new window

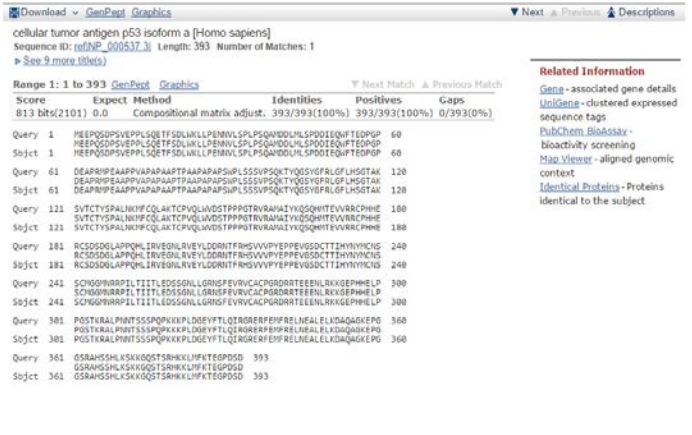
Algorithm parameters

Show Conserved Domains









InterPro

Protein sequence analysis & classification

http://www.ebi.ac.uk/interpro/about.html

PROSITE is a database of protein families and domains. It consists of biologically significant sites, patterns and profiles that help to reliably identify to which known protein family a new sequence belongs. PROSITE is base at the Swiss Institute of Bioinformatics (SIB), Geneva, Switzerland.

HAMAP stands for High-quality Automated and Manual Annotation of Proteins. HAMAP profiles are manually created by expert curators. They identify proteins that are part of well-conserved proteins families or subfamilies. HAMAP is based at the SIB Swiss Institute of Bioinformatics, Geneva, Switzerland.

Pfam is a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains. Pfam is based at the Wellcome Trust Sanger Institute, Hinxton, UK.

PRINTS is a compendium of protein fingerprints. A fingerprint is a group of conserved motifs used to characterise a protein family or domain. PRINTS is based at the University of Manchester, UK.

ProDom protein domain database consists of an automatic compilation of homologous domains. Current versions of ProDom are built using a novel procedure based on recursive PSI-BLAST searches. ProDom is based at PRABI Villeurbanne, France.

SMART (a Simple Modular Architecture Research Tool) allows the identification and annotation of genetically mobile domains and the analysis of domain architectures. SMART is based at at EMBL, Heidelberg, Germany.

TIGRFAMs is a collection of protein families, featuring curated multiple sequence alignments, hidden Markov models (HMMs) and annotation, which provides a tool for identifying functionally related proteins based on sequence homology. TIGRFAMs is based at the J. Craig Venter Institute, Rockville, MD, US.

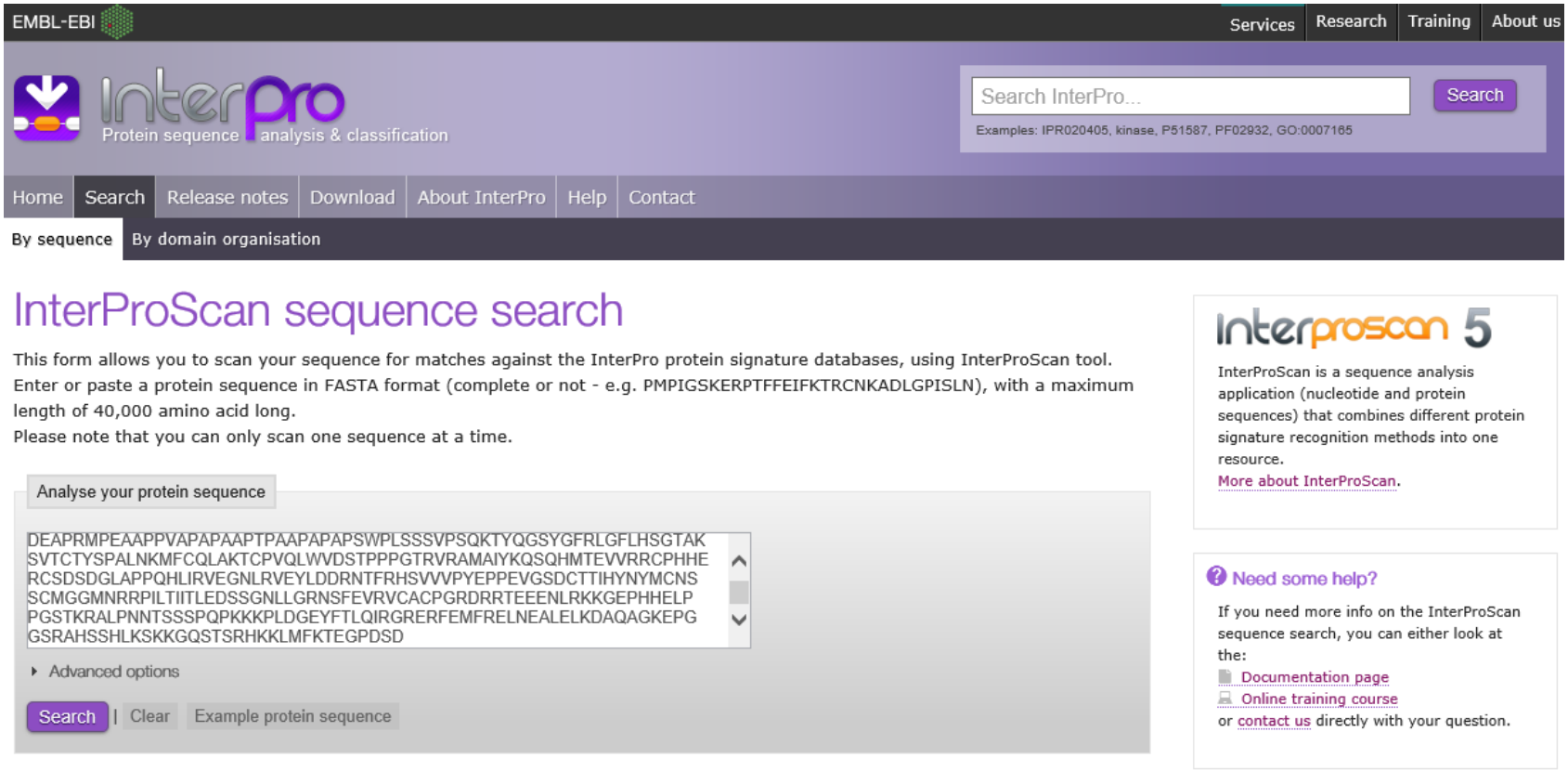
PIRSF protein classification system is a network with multiple levels of sequence diversity from superfamilies to subfamilies that reflects the evolutionary relationship of full-length proteins and domains. PIRSF is based at the Protein Information Resource, Georgetown University Medical Centre, Washington DC, US.

SUPERFAMILY is a library of profile hidden Markov models that represent all proteins of known structure. The library is based on the SCOP classification of proteins: each model corresponds to a SCOP domain and aims to represent the entire SCOP superfamily that the domain belongs to. SUPERFAMILY is based at the University of Bristol, UK.

CATH-Gene3D database describes protein families and domain architectures in complete genomes. Protein families are formed using a Markov clustering algorithm, followed by multi-linkage clustering according to sequence identity. Mapping of predicted structure and sequence domains is undertaken using hidden Markov models libraries representing CATH and Pfam domains. CATH-Gene3D is based at University College, London, UK.

PANTHER is a large collection of protein families that have been subdivided into functionally related subfamilies, using human expertise. These subfamilies model the divergence of specific functions within protein families, allowing more accurate association with function, as well as inference of amino acids important for functional specificity. Hidden Markov models (HMMs) are built for each family and subfamily for classifying additional protein sequences. PANTHER is based at at University of Southern California, CA, US.

http://www.ebi.ac.uk/Tools/pfa/iprscan5/



Protein

Export Select format

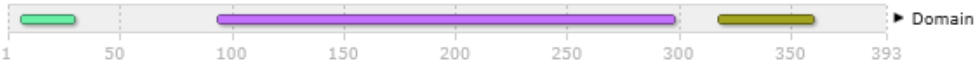
Submitted

Length 393 amino acids

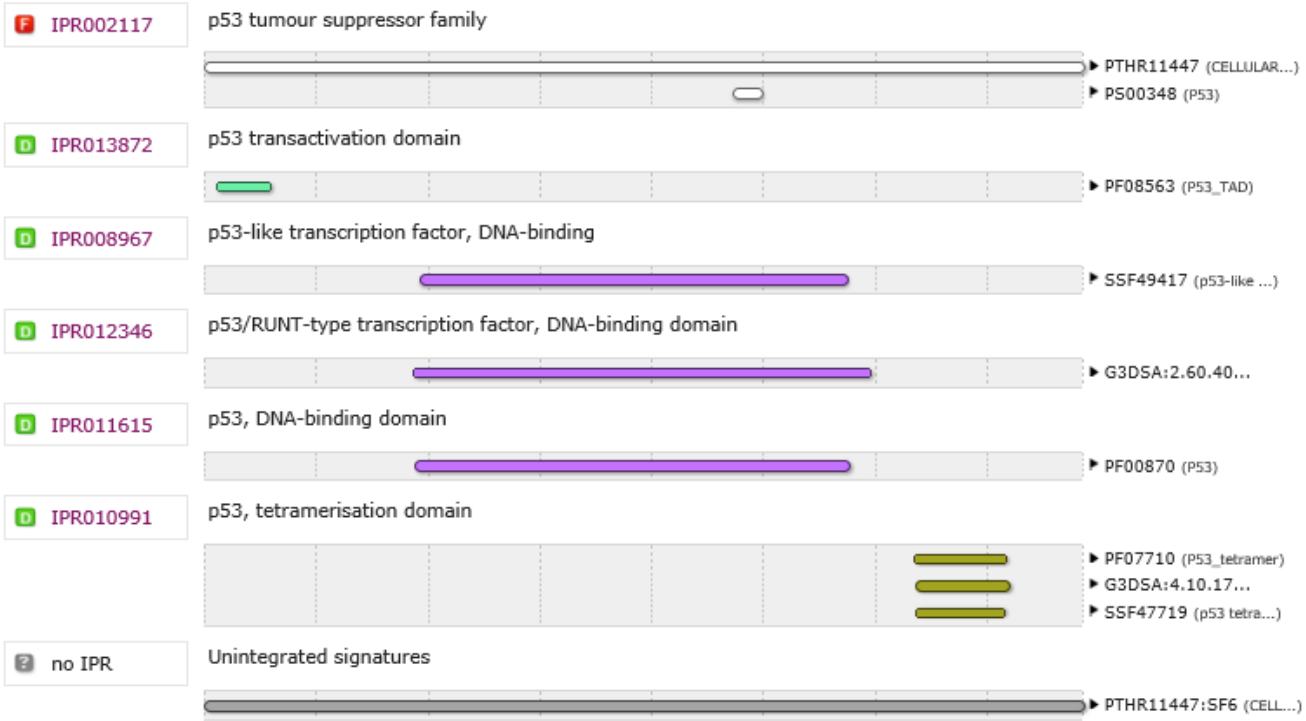
Protein family membership

p53 tumour suppressor family (IPR002117)

Domains and repeats



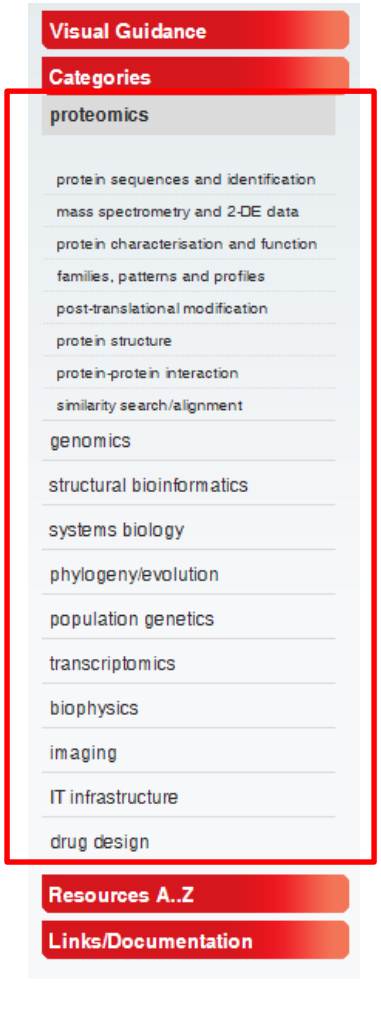
Detailed signature matches



Protein Sequence Analysis Tools

http://www.expasy.org/





ExPASy Molecular Biology Server

Expert **P**rotein **A**nalysis **Sy**stem is the new SIB Bioinformatics Resource Portal which provides access to scientific databases and software tools in different areas of life sciences including proteomics, genomics, phylogeny, systems biology, population genetics, transcriptomics etc

Tools

- SWISS-MODEL Workspace • structure homology-modeling • [more]
- SwissDock • protein ligand docking server • [more]

- 2ZIP • Prediction of leucine zipper domains • [more]
- 3of5 • find user-defined patterns in protein sequences • [more]
- AACompIdent • protein identification by aa composition • [more]
- AACompSim • amino acid composition comparison • [more]
- Agadir • Prediction of the helical content of peptides • [more]
- ALF • simulation of genome evolution • [more]
- Alignment tools • Four tools for multiple alignments • [more]
- AllAll • protein sequences comparisons • [more]
- APSSP • Advanced Protein Secondary Structure Prediction • [more]
- Ascalaph • Molecular modeling software • [more]
- big-PI • predict GPI modification sites • [more]
- Biochemical Pathways • Biochemical Pathways • [more]
- BLAST • sequence similarity search • [more]
- BLAST (UniProt) • BLAST search on the UniProt web site • [more]
- BLAST - NCBI • Biological sequence similarity search • [more]
- BLAST - PBIL • BLAST search on protein sequence databases • [more]
- Blast2Fasta • Blast to Fasta conversion • [more]
- boxshade • MSA pretty printer • [more]
- CFSSP • Protein secondary structure prediction • [more]
- ChloroP • chloroplast transit peptides & cleavage sites • [more]
- Click2Drug • Directory of computational drug design tools • [more]
- ClustalO (UniProt) • Align two or more protein sequences • [more]
- ClustalW • Multiple sequence alignment • [more]
- ClustalW - PBIL • Multiple sequence alignment program • [more]
- ClustalW2 • Multiple sequence alignment program • [more]
- Coiled-Coils prediction • Prediction of coiled coils regions • [more]
- COILS • Prediction of Coiled Coil Regions in Proteins • [more]
- ColorSeq • Color Protein Sequence • [more]
- Compute pI/MW • theoretical pI and Mw computation • [more]
- CPHmodels • Protein homology modeling • [more]
- CSS-Palm • Prediction of palmitoylation sites in proteins • [more]
- DAS-TMfilter • Prediction of transmembrane regions • [more]
- Decrease redundancy • Sequence redundancy reduction • [more]
- DIALIGN • Local multiple sequence aligment • [more]
- DictyOGlyc • GlcNAc O-glycosylation sites in D.discoideum • [more]
- DisEMBL • Prediction of disordered protein regions • [more]
- DLP-SVM • Domain linker predictor • [more]
- Dotlet • sequence similarity plots • [more]
- EasyProt • graphical platform for proteomics analysis • [more]
- ELM • Eukaryotic Linear Motifs • [more]
- EMBnet services • bioinformatics tools, databases and courses • [more]
- EMBOSS translation tools • sequence translation tools • [more]
- epestfind • Identification of PEST motifs • [more]
- FASTA/SSEARCH/GGSEARCH/GLSEARCH • Sequence similarity searching of protein db • [more]
- FindMod • protein post-translational modification prediction • [more]
- FindPept • peptide identification from unspecific cleavage • [more]
- FingerPRINTScan • scan sequences against PRINTS • [more]
- FUGUE • Sequence-structure homology recognition • [more]
- GENIO/logo • RNA/DNA & Amino Acid Sequence Logos • [more]
- Geno3D • Protein molecular modelling • [more]
- GlobPlot • Protein disorder/globularity/domain predictor • [more]
- GlycanMass • oligosaccharide structure mass calculation • [more]
- GlycoMod • oligosaccharide structure prediction • [more]
- Glycoviewer • visualize a set of glycan structures • [more]
- GOR • Protein secondary structure prediction • [more]
- GPI-SOM • identify GPI-anchor signals • [more]
- GPMAW lite • protein physical and chemical parameters • [more]
- Graphical Codon Usage Analyser • Codon bias • [more]
- HAMAP • UniProtKB family classification and annotation • [more]
- HAMAP-Scan • Scan protein sequences against the HAMAP families • [more]
- HCA (hydrophobic cluster analysis) • hydrophobic cluster analysis • [more]
- HCD/CID spectra merger • combine HCD and CID MS/MS spectra • [more]
- HeliQuest • alpha-helical properties screening • [more]
- HHpred • Homology detection & structure prediction • [more]
- HMMTOP • Prediction of transmembranes helices and topology • [more]
- HTMSRAP • Helical TransMembrane Prediction • [more]
- ImageMaster / Melanie • software for 2-D PAGE analysis • [more]
- InsPecT • MS/MS tool to identify modified peptides • [more]
- InterProScan • Family and domain datatabase search • [more]
- ISMARA • genome-wide expression data modeling • [more]
- IsotopIdent • theoretical isotopic distribution • [more]
- Jmol • 3D molecule visualization tool • [more]
- Jpred • Secondary Structure Prediction Server • [more]
- Kalign - EBI • Fast and accurate multiple sequence alignment • [more]
- Kalign - SBC • Fast and accurate multiple sequence alignment • [more]
- LALIGN • Pairwise alignment • [more]
- LipoP • predict lipoproteins and signal peptides • [more]
- LiveBench • Continuous Benchmarking of Structure Prediction Se • [more]
- MAFFT - CBRC • Multiple sequence aligment • [more]
- MAFFT - EBI • Multiple sequence aligment • [more]
- Make2D-DB II • package to build web-based proteomics database • [more]
- MakeMultimer.py • Reconstruction of multimeric molecules in crystals • [more]
- MALDIPepQuant • quantify MALDI peptides • [more]
- MARCOIL • coiled-coils prediction • [more]
- MarvinSpace • Visualization, modeling and interactive rendering • [more]
- Mascot • protein identification from mass spectrometry data • [more]
- MassSearch • sequence search by mass spectrometry weights • [more]

Practical: *Gene*; *RNA*; *Protein*

U62639 (Gene)

```
aaaaatgtat gtctgatttt gaaatgctca tttcctttga ggtttccatt tttgagttgc 61   ccgtaatttg tatttttctg aagatgagca attcaatttt taaattgccc gcacctctac 121
cgtttccatc gtgtattttg ttaaaatatt cacagattaa cccatttacc gtttcatcca 181  cctgtttttc ctcgaaaaga ttccaatgtt ctataattct acaaaacttc ccacgcgaga 241
aacaactgta ataaactgaa tatattatct atcgcatcgt tttcaaccag aattaagcaa 301  gaggttccac aactttaaac accaacaacg caatcctaaa tcatttgcaa gattttattt 361
cagatgctac actttctgcc tgaaaaaaat tctgaaaagc cgaacaataa ttcatggtaa 421  caatgaatgg cagatacatc aaagttttag atgaacaatt tttatgtatt aaatgtacat 481
ttaaaaacaa attgcacaac gattctacta ctgtcgcact aattttacgt atgtctgtac 541  ttgaagattt cgaattaatt tgttcaatat tgtgttaaaa tgtttgattt atacactcaa 601
atctttaaaa gatttattgg aaaagataaa tggttaattt aaaccaaaaa tttccatcaa 661  gccttttctg aaaacactaa aattattttc gtggtgggac caggcgcgcg cgtcccatga 721
tgttccttta atcaaaatgc atttctgtcc cggcgggaga aattgaattt tgattttaag 781  gcgcgaattt ttgcctaaaa acgatgccat tctttcattc ttttcataat ctcactcacc 841
atgagaacca tgcgccttgc ttggttgctc ccacttttta ttcacatact aatcaaggta 901  atttccccgt ttttctagtt ttttcaatgt attttcatgt ttcagaacac agctcaagct 961
ccggctgtca acaactcgac atgcgatcaa gcaaaggaat ttgattgcgg gaacgggaga 1021 ctccgatgca ttcccgcgga gtggcaatgc gacaacgtag cggactgcga caaaggaaga 1081
gacgaatcgg gctgctcata tgcgcatcat tgttcgacaa gcttcatgtt atgcaagaat 1141 ggactgtgtg tcgcaaatga gttcaaatgc gacggcgaag acgactgccg cgatggaagc 1201
gatgagcagc attgcgagta caatatcctg aagtctcgct tcgatggttc caatccttcg 1261 gctcctacca ctttcgttgg tcacaatggc ccagaatgcc atcctcctcg tttacgatgc 1321
cgatcaggac aatgtattca accagatctc gtttgtgatg gacatcagga ttgttctgga 1381 ggagatgatg aggtcaactg caccagaagg ggacatgaaa atatgcagtc ctcgactgat 1441
tttcacgatg atgttcatct tgtcgatcca acctttttcg ctaatgaaga caataaggta 1501 attgtttaat gtttattaat ccgttttaac ttttattttt cagtgtcgga gtggatacac 1561
aatgtgccat agcggagacg tctgcatacc tgacagtttt ctttgtgacg gcgatctaga 1621 ttgtgatgat gcttcggacg agaaaaactg ccaaactaat gctccaagcg aagaagaata 1681
tctttctggg caagccgatc acatgcattc gtgctcagca gcaggaatgt attcttgtgg 1741 aacaaaagga tccgaaattg gcgtttgtat tccgatgaat gccacgtgta atgggatcaa 1801
ggagtgtcca ctaggagatg acgagtcaaa acattgctcc gaatgtgcca gaaagcgatg 1861 tgaccacaca tgtatgaaca ctccacacgg ggctcgctgc atttgtcaag aaggatataa 1921
gcttgccgat gacggactca cttgcgagga tgaagatgag tgtgcaactc atgggcactt 1981 gtgccagcat ttctgtgaag atcgtttggg ttcctttgca tgcaaatgtg ccaacggtta 2041
tgagcttgaa acggatgggc attcttgtaa atacgaggca accactacgc cagaaggata 2101 tttgttcatc agtcttggtg gagaagttcg acagatgcca ttggcagatt tcaccgatgg 2161
ttcaaattac tcggcgattc aaaagtttgc tggccacgga accatcagat cgatcgactt 2221 catgcatcgc aacaacaaaa tgttcatgtc aatttctgat gagcacggtg atccaactgg 2281
cgaattgtca gtgtccgaca atggattgat gagagttctt cgagaaaatg tcattggagt 2341 gagcaacgtg gcagtcgact ggattggtgg aaacgttttc ttcacacaaa aatgtatgtt 2401
tatctaatgt ttaaattttt catttgtgat tcttacagct ccatctccaa gcgctgggat 2461 ttccatctgc acaatgagcg gaatgttctg tcgccgagtt atcgaaggca aagaacaagg 2521
acaatcctat cgtggtcttg ttgttcaccc gatgcgcggt ctcatcatct ggatcgattc 2581 ttatcagaaa tatcatcgca tcatgatggc taatatggat gggtctcagg tgagtcgatc 2641
gagtcgatct gatttagttc atttctaaat aaatttcagg tcagaatcct tctcgacaac 2701 aagttggaag ttccatcagc tcttgccatc gactacatcc gccacgatgt ctattttgga 2761
gatgttgaac gtcagttgat cgaaagagtc aatatcgaca cgaaagagcg ccgcgtagtg 2821 atttcgaacg gagttcatca tccgtatgac atggcttact tcaatggttt cctatactgg 2881
gcagattggt aagacatctt atctaattta tattttcaaa tttatttttc aggggaagcg 2941 agtcattaaa ggttcaagag atgacccatc atcattcgag tcctcaagtc atccatactt 3001
tcaatcgtta tccatatggt attgctgtca atcactcact ctaccagact ggtcctccat 3061 caaacccatg ccttgaactc gagtgcccat ggctctgcgt tattgtgcca aagagcgatt 3121
tcattatgac tgccaagtgt gtctgcccag acggatacac tcattccgtc actgaaaact 3181 cttgcatccc gcctgtgacg attgaggacg aggagaacct tgagaagctt tcccacattg 3241
gatctgcttt gatggccgaa tactgcgaag ctggtgtcgc gtgtatgaat ggaggagcct 3301 gccgtgaact acaaaatgag cacggaagag ctcatcgcat cgtttgtgat tgtgagggtc 3361
catatgacgg gcaatactgc gaacggctca atccagagaa gttctccgca atggaagagg 3421 aagattcgtc cttatggctt atcgttctgc ttctcatttt tctcatcatc gttgcggtag 3481
tcggaattat tgccttcctt tggttttctc aacaagagca tatgaaagat gtgatttcca 3541 ctgcccgtgt ccgtgttgat aacatggcta gaaaagcgga agatgctgca gctccaattg 3601
tcgagaagtt ccgcaaggtc actgataagc agaggagcac gcctcctaga gaaggttgtc 3661 aaacggcaac aaacgttgac ttcgtttcct acgagacaaa tgctgagaaa agaattcgga 3721
tggactcttc gccgacgtca tacggaaacc ccatgtacga tgaagttcct gaatcgtcaa 3781 ctggtttcgt cagatcggct tccgcaccat tcgctggagt cattcgattt gagaacgaca 3841
gcttgttgtg aattctacta caaaattact aaatcagatg tctgtaaagt atatctattt 3901 ttgcctattt attgcatgaa agttgataat gtcta
```

Practical: *Gene; RNA; Protein*

U62639 (mRNA)

```
atgagaacca tgcgccttgc ttggttgctc ccacttttta ttcacatact aatcaagaac 61   acagctcaag ctccggctgt caacaactcg acatgcgatc aagcaaagga atttgattgc 121
gggaacggga gactccgatg cattcccgcg gagtggcaat gcgacaacgt agcggactgc 181  gacaaaggaa gagacgaatc gggctgctca tatgcgcatc attgttcgac aagcttcatg 241
ttatgcaaga atggactgtg tgtcgcaaat gagttcaaat gcgacggcga agacgactgc 301  cgcgatggaa gcgatgagca gcattgcgag tacaatatcc tgaagtctcg cttcgatggt 361
tccaatcctt cggctcctac cactttcgtt ggtcacaatg gcccagaatg ccatcctcct 421  cgtttacgat gccgatcagg acaatgtatt caaccagatc tcgtttgtga tggacatcag 481
gattgttctg gaggagatga tgaggtcaac tgcaccagaa ggggacatga aaatatgcag 541  tcctcgactg attttcacga tgatgttcat cttgtcgatc caacctttt cgctaatgaa 601
gacaataagt gtcggagtgg atacacaatg tgccatagcg gagacgtctg catacctgac 661  agttttcttt gtgacggcga tctagattgt gatgatgctt cggacgagaa aaactgccaa 721
actaatgctc caagcgaaga agaatatctt tctgggcaag ccgatcacat gcattcgtgc 781  tcagcagcag gaatgtattc ttgtggaaca aaaggatccg aaattggcgt ttgtattccg 841
atgaatgcca cgtgtaatgg gatcaaggag tgtccactag gagatgacga gtcaaaacat 901  tgctccgaat gtgccagaaa gcgatgtgac cacacatgta tgaacactcc acacggggct 961
cgctgcattt gtcaagaagg atataagctt gccgatgacg gactcacttg cgaggatgaa 1021 gatgagtgtg caactcatgg gcacttgtgc cagcatttct gtgaagatcg tttgggttcc 1081
tttgcatgca aatgtgccaa cggttatgag cttgaaacgg atgggcattc ttgtaaatac 1141 gaggcaacca ctacgccaga aggatatttg ttcatcagtc ttggtggaga agttcgacag 1201
atgccattgg cagatttcac cgatggttca aattactcgg cgattcaaaa gtttgctggc 1261 cacggaacca tcagatcgat cgacttcatg catcgcaaca acaaaatgtt catgtcaatt 1321
tctgatgagc acggtgatcc aactggcgaa ttgtcagtgt ccgacaatgg attgatgaga 1381 gttcttcgag aaaatgtcat tggagtgagc aacgtggcag tcgactggat tggtggaaac 1441
gttttcttca cacaaaaatc tccatctcca agcgctggga tttccatctg cacaatgagc 1501 ggaatgttct gtcgccgagt tatcgaaggc aaagaacaag gacaatccta tcgtggtctt 1561
gttgttcacc cgatgcgcgg tctcatcatc tggatcgatt cttatcagaa atatcatcgc 1621 atcatgatgg ctaatatgga tgggtctcag gtcagaatcc ttctcgacaa caagttggaa 1681
gttccatcag ctcttgccat cgactacatc cgccacgatg tctattttgg agatgttgaa 1741 cgtcagttga tcgaaagagt caatatcgac acgaaagagc gccgcgtagt gatttcgaac 1801
ggagttcatc atccgtatga catggcttac ttcaatggtt tcctatactg ggcagattgg 1861 ggaagcgagt cattaaaggt tcaagagatg acccatcatc attcgagtcc tcaagtcatc 1921
catactttca atcgttatcc atatggtatt gctgtcaatc actcactcta ccagactggt 1981 cctccatcaa acccatgcct tgaactcgag tgcccatggc tctgcgttat tgtgccaaag 2041
agcgatttca ttatgactgc caagtgtgtc tgcccagacg gatacactca ttccgtcact 2101 gaaaactctt gcatcccgcc tgtgacgatt gaggacgagg agaaccttga gaagctttcc 2161
cacattggat ctgctttgat ggccgaatac tgcgaagctg gtgtcgcgtg tatgaatgga 2221 ggagcctgcc gtgaactaca aaatgagcac ggaagagctc atcgcatcgt ttgtgattgt 2281
gagggtccat atgacgggca atactgcgaa cggctcaatc cagagaagtt ctccgcaatg 2341 gaagaggaag attcgtcctt atggcttatc gttctgcttc tcatttttct catcatcgtt 2401
gcggtagtcg gaattattgc cttccttgg ttttctcaac aagagcatat gaaagatgtg 2461 atttccactg cccgtgtccg tgttgataac atggctagaa aagcggaaga tgctgcagct 2521
ccaattgtcg agaagttccg caaggtcact gataagcaga ggagcacgcc tcctagagaa 2581 ggttgtcaaa cggcaacaaa cgttgacttc gtttcctacg agacaaatgc tgagaaaaga 2641
attcggatgg actcttcgcc gacgtcatac ggaaacccca tgtacgatga agttcctgaa 2701 tcgtcaactg gtttcgtcag atcggcttcc gcaccattcg ctggagtcat tcgatttgag 2761
aacgacagct tgttgtga
```

Practical: *Gene*; *RNA*; *Protein*

AAD09364 (Protein)

```
  1 MRTMRLAWLL PLFIHILIKN TAQAPAVNNS TCDQAKEFDC GNGRLRCIPA EWQCDNVADC
 61 DKGRDESGCS YAHHCSTSFM LCKNGLCVAN EFKCDGEDDC RDGSDEQHCE YNILKSRFDG
121 SNPSAPTTFV GHNGPECHPP RLRCRSGQCI QPDLVCDGHQ DCSGGDDEVN CTRRGHENMQ
181 SSTDFHDDVH LVDPTFFANE DNKCRSGYTM CHSGDVCIPD SFLCDGDLDC DDASDEKNCQ
241 TNAPSEEEYL SGQADHMHSC SAAGMYSCGT KGSEIGVCIP MNATCNGIKE CPLGDDESKH
301 CSECARKRCD HTCMNTPHGA RCICQEGYKL ADDGLTCEDE DECATHGHLC QHFCEDRLGS
361 FACKCANGYE LETDGHSCKY EATTTPEGYL FISLGGEVRQ MPLADFTDGS NYSAIQKFAG
421 HGTIRSIDFM HRNNKMFMSI SDEHGDPTGE LSVSDNGLMR VLRENVIGVS NVAVDWIGGN
481 VFFTQKSPSP SAGISICTMS GMFCRRVIEG KEQGQSYRGL VVHPMRGLII WIDSYQKYHR
541 IMMANMDGSQ VRILLDNKLE VPSALAIDYI RHDVYFGDVE RQLIERVNID TKERRVVISN
601 GVHHPYDMAY FNGFLYWADW GSESLKVQEM THHHSSPQVI HTFNRYPYGI AVNHSLYQTG
661 PPSNPCLELE CPWLCVIVPK SDFIMTAKCV CPDGYTHSVT ENSCIPPVTI EDEENLEKLS
721 HIGSALMAEY CEAGVACMNG GACRELQNEH GRAHRIVCDC EGPYDGQYCE RLNPEKFSAM
781 EEEDSSLWLI VLLLIFLIIV AVVGIIAFLW FSQQEHMKDV ISTARVRVDN MARKAEDAAA
841 PIVEKFRKVT DKQRSTPPRE GCQTATNVDF VSYETNAEKR IRMDSSPTSY GNPMYDEVPE
901 SSTGFVRSAS APFAGVIRFE NDSLL
```

Practical: *Gene; RNA; Protein*

1. Download the sequences Gene, RNA and Protein

ANALYSIS:

1. Exon/intron organization.
   Use (1) BESTFIT & GAP ("gene" vs "rna")
   (2) SEARCH **U62639 in NCBI**
   CDS join(841..897,946..1497,1544..2393,2439..2629,2680..2888, 2933..3851)

2. Opening Reading Frame
   Use MAP to find the ORF
   Use TRANSLATE to write the ORF
   Compare your ORF with "protein"
3. Protein Domain Search (NCBI CD Search, Interpro)
4. Protein Sequence Analysis
   -use *ExPASY as a portal*

http://www.bioinformatics.org/sms2/

SMS

Format Conversion

-Combine FASTA
-EMBL to FASTA
-EMBL Feature Extractor
-EMBL Trans Extractor
-Filter DNA
-Filter Protein
-GenBank to FASTA
-GenBank Feature Extractor
-GenBank Trans Extractor
-One to Three
-Range Extractor DNA
-Range Extractor Protein
-Reverse Complement
-Split Codons
-Split FASTA
-Three to One
-Window Extractor DNA
-Window Extractor Protein

Sequence Manipulation Suite:

Version 2

- The Sequence Manipulation Suite is a collection of JavaScript programs for generating, formatting, and analyzing short DNA and protein sequences. It is commonly used by molecular biologists, for teaching, and for program and algorithm testing.
- See the about the Sequence Manipulation Suite page for more information about individual Sequence Manipulation Suite programs.
- You can easily mirror the Sequence Manipulation Suite on your own web site, or you can use it off-line.
- This version of the Sequence Manipulation Suite represents a complete re-write of the previous version. The new version is much faster and has many new features. The previous version of the Sequence Manipulation Suite can still be accessed.
- Send questions and comments to stothard@ualberta.ca.

new window | home | citation

Fri Jun 17 16:17:06 2016

Valid XHTML 1.0; Valid CSS

ASSIGNMENT 03

Download the file ex.fasta from website

1. Assemble the fragments
2. How many potential reading frames are there?
3. Give the names of these genes?
4. The identity and similarity of the last gene with H. sapiens?
   - nucleotide and amino acid sequence
5. MW, pI and potential post-translational modification sites of any ONE protein.

E-mail the ANSWER as attached files to
--petang@mail.cgu.edu.tw. Before nest Thursday 1200
****郵件主旨： ASS04 bioinfo – (學號)