Protein Databases for
Mass Spectrometric Analysis

Akhilesh Pandey, M.D., Ph.D.

http://pandeylab.igm.jhmi.edu/

Human Genome Annotation

- A case for proteomics-driven annotation of protein-coding regions

Genome Annotation by Mass Spectrometry: What Can We Gain?

- Assigning start codons
- Proteins isoforms (alternative splicing, novel exons)
- Novel genes (proteins less than 100 amino acids not predicted by programs)
- cSNPs
- Correction of incorrect gene predictions (50% of the genes in human are predicted)
- Validation of gene predictions

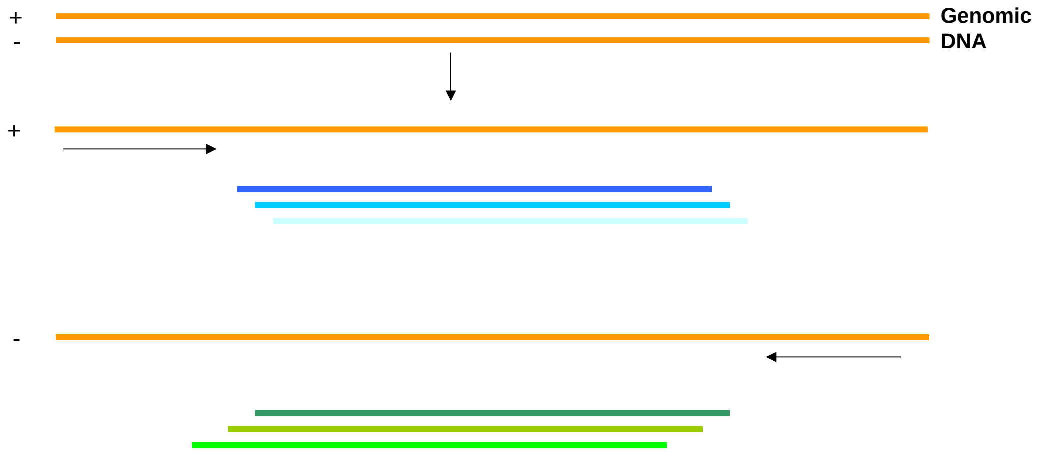
When is a peptide not identified from a database search?

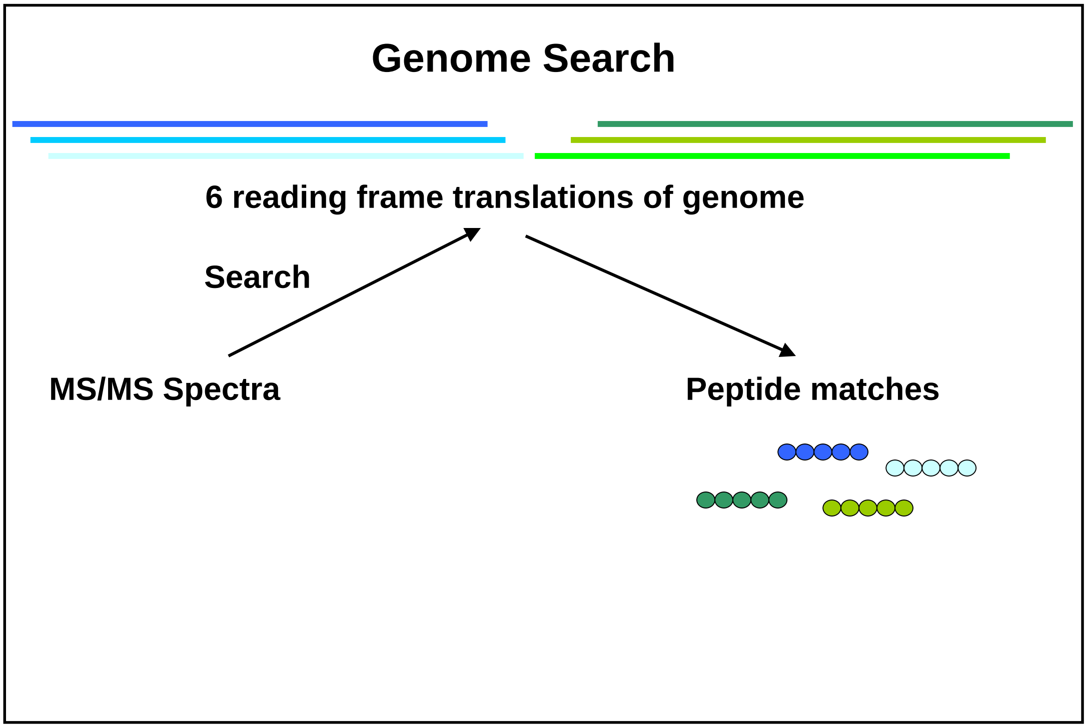
- Protein not described (i.e. novel protein)
- Polymorphisms
- Alternative splice forms
- Novel exon
- Wrong annotation

How do you identify such events?

- For novel genes and novel exons use the human genome sequence
- For polymorphisms and alternate splice forms, use a computational strategy

Genome Search





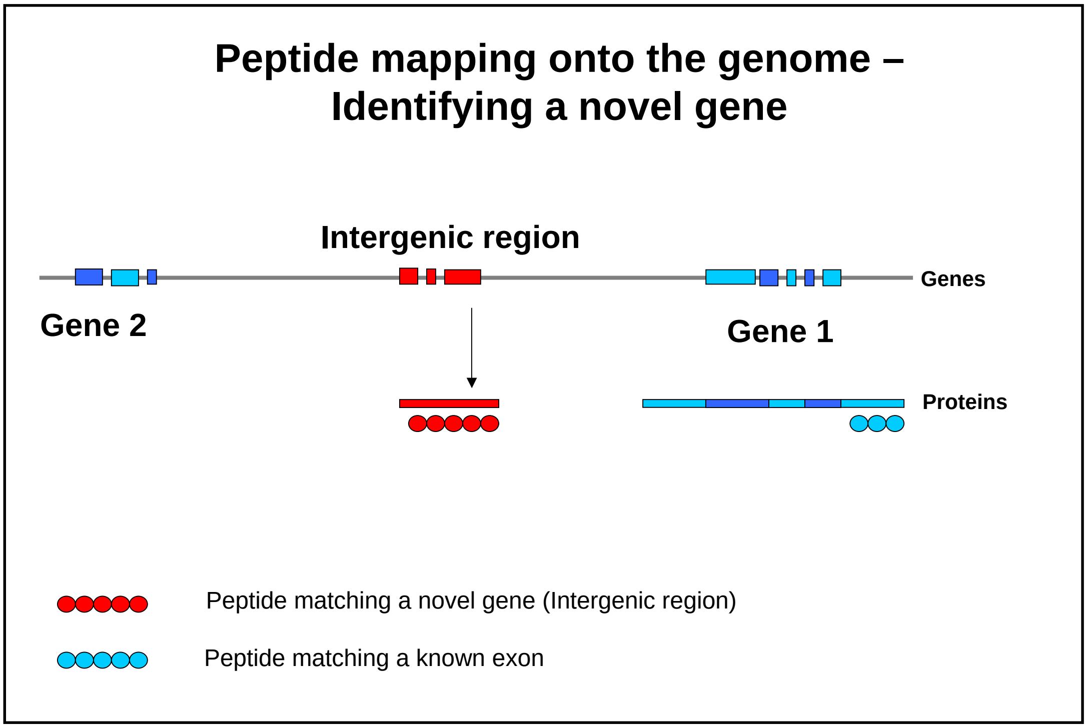
Genome Search

6 reading frame translations of genome

Search

MS/MS Spectra

Peptide matches

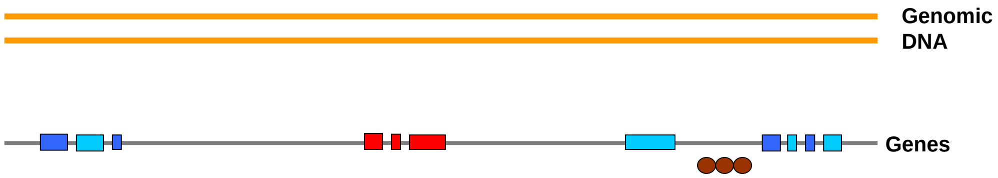


Peptide mapping onto the genome – Identifying a novel gene

🔴🔴🔴🔴🔴 Peptide matching a novel gene (Intergenic region)

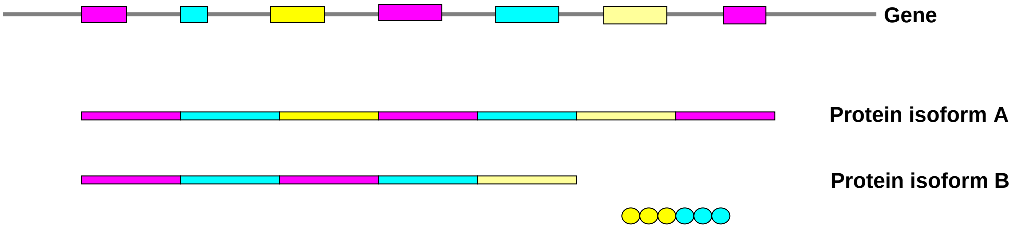
🔵🔵🔵🔵🔵 Peptide matching a known exon

Peptide mapping onto the genome – Identifying a novel exon



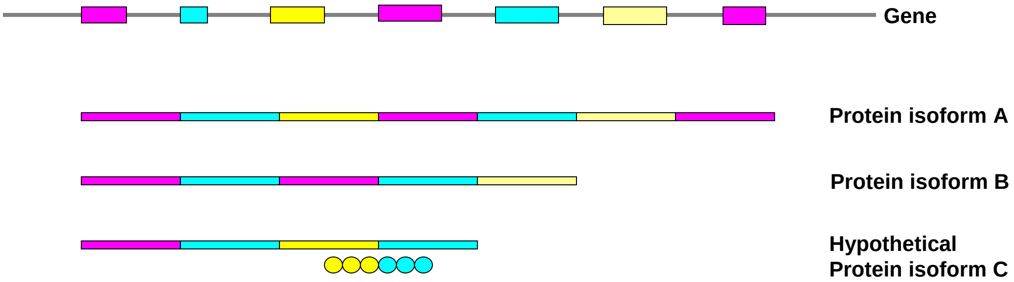
●●● Peptide matching intronic region

Alternate splice forms



No Match!!

Alternate splice forms



Gene

Protein isoform A

Protein isoform B

Hypothetical
Protein isoform C

Peptide matches!!

Alternate splice forms

Gene Symbol: HSPA8

```
NP_694881.1 MSKGPAVGIDLGTTYSCVGVFQHGKVEIIANDQGNRTTPSYVAFTDTERLIGDAAKNQVA 60
NP_006588.1 MSKGPAVGIDLGTTYSCVGVFQHGKVEIIANDQGNRTTPSYVAFTDTERLIGDAAKNQVA 60
            ************************************************************
NP_694881.1 MNPTNTVFDAKRLIGRRFDDAVVQSDMKHWPFMVVNDAGRPKVQVEYKGETKSFYPEEVS 120
NP_006588.1 MNPTNTVFDAKRLIGRRFDDAVVQSDMKHWPFMVVNDAGRPKVQVEYKGETKSFYPEEVS 120
            ************************************************************
NP_694881.1 SMVLTKMKEIAEAYLGKTVTNAVVTVPAYFNDSQRQATKDAGTIAGLNVLRIINEPTAAA 180
NP_006588.1 SMVLTKMKEIAEAYLGKTVTNAVVTVPAYFNDSQRQATKDAGTIAGLNVLRIINEPTAAA 180
            ************************************************************
NP_694881.1 IAYGLDKKVGAERNVLIFDLGGGTFDVSILTIEDGIFEVKSTAGDTHLGGEDFDNRMVNH 240
NP_006588.1 IAYGLDKKVGAERNVLIFDLGGGTFDVSILTIEDGIFEVKSTAGDTHLGGEDFDNRMVNH 240
            ************************************************************
NP_694881.1 FIAEFKRKHKKDISENKRAVRRLRTACERAKRTLSSSTQASIEIDSLYEGIDFYTSITRA 300
NP_006588.1 FIAEFKRKHKKDISENKRAVRRLRTACERAKRTLSSSTQASIEIDSLYEGIDFYTSITRA 300
            ************************************************************
NP_694881.1 RFEELNADLFRGTLDPVEKALRDAKLDKSQIHDIVLVGGSTRIPKIQKLLQDFFNGKELN 360
NP_006588.1 RFEELNADLFRGTLDPVEKALRDAKLDKSQIHDIVLVGGSTRIPKIQKLLQDFFNGKELN 360
            ************************************************************
NP_694881.1 KSINPDEAVAYGAAVQAAILSGDKSENVQDLLLLDVTPLSLGIETAGGVMTVLIKRNTTI 420
NP_006588.1 KSINPDEAVAYGAAVQAAILSGDKSENVQDLLLLDVTPLSLGIETAGGVMTVLIKRNTTI 420
            ************************************************************
NP_694881.1 PTKQTQTFTTYSDNQPGVLIQVYEGERAMTKDNNLLGKFELT------------------ 462
NP_006588.1 PTKQTQTFTTYSDNQPGVLIQVYEGERAMTKDNNLLGKFELTGIPPAPRGVPQIEVTFDI 480
            ******************************************
NP_694881.1 ------------------------------------------------------------
NP_006588.1 DANGILNVSAVDKSTGKENKITITNDKGRLSKEDIERMVQEAEKYKAEDEKQRDKVSSKN 540
NP_694881.1 ------------------------------------------------------------
NP_006588.1 SLESYAFNMKATVEDEKLQGKINDEDKQKILDKCNEIINWLDKNQTAEKEEFEHQQKELE 600
NP_694881.1 ---------------GMPGGMPGGFPGGGAPPSGGASSGPTIEEVD 493
NP_006588.1 KVCNPIITKLYQSAGGMPGGMPGGFPGGGAPPSGGASSGPTIEEVD 646
```

Alternate splice forms

Gene Symbol: OGT

```
NP_858059 MASSVGNVADSTG----------LAELAHREYQAGDFEAAERHCMQLWRQEPDNTGVLLL 50
NP_858058 MASSVGNVADSTEPTKRMLSFQGLAELAHREYQAGDFEAAERHCMQLWRQEPDNTGVLLL 60
          ************           *************************************

NP_858059 LSSIHFQCRRLDRSAHFSTLAIKQNPLLAEAYSNLGNVYKERGQLQEAIEHYRHALRLKP 110
NP_858058 LSSIHFQCRRLDRSAHFSTLAIKQNPLLAEAYSNLGNVYKERGQLQEAIEHYRHALRLKP 120
          ************************************************************

NP_858059 DFIDGYINLAAALVAAGDMEGAVQAYVSALQYNPDLYCVRSDLGNLLKALGRLEEAKACY 170
NP_858058 DFIDGYINLAAALVAAGDMEGAVQAYVSALQYNPDLYCVRSDLGNLLKALGRLEEAKACY 180
          ************************************************************

NP_858059 LKAIETQPNFAVAWSNLGCVFNAQGEIWLAIHHFEKAVTLDPNFLDAYINLGNVLKEARI 230
NP_858058 LKAIETQPNFAVAWSNLGCVFNAQGEIWLAIHHFEKAVTLDPNFLDAYINLGNVLKEARI 240
          ************************************************************


                         …..

NP_858059 TCLGCLELIAKNRQEYEDIAVKLGTDLEYLKKVRGKVWKQRISSPLFNTKQYTMELERLY 1010
NP_858058 TCLGCLELIAKNRQEYEDIAVKLGTDLEYLKKVRGKVWKQRISSPLFNTKQYTMELERLY 1020
          ************************************************************

NP_858059 LQMWEHYAAGNKPDHMIKPVEVTESA 1036
NP_858058 LQMWEHYAAGNKPDHMIKPVEVTESA 1046
          **************************
```

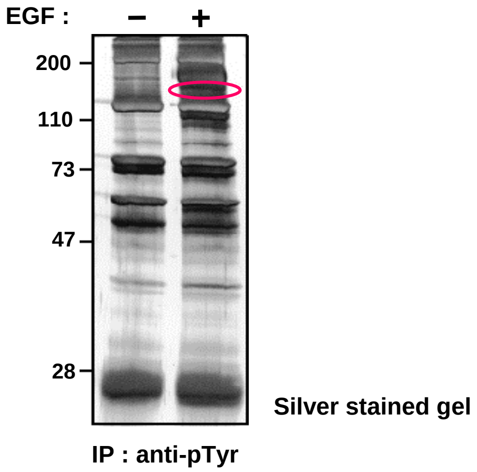
The Myth of Kozak's Consensus Sequence: Translation Initiation Codon

- **CCACCATGG**
- **Most upstream ATG used for translation initiation**
- **Biologists look for this sequence and annotate any ATG near the 5' end of the clone as the initiator methionine**

N-terminal Acetylation

- Perhaps the most common co-translational modification (60-85% of proteins in yeast)
- Usually, aminopeptidases cleave one or two N-terminal amino acids followed by acetylation of the 'mature' protein
- So, if you find an N-acetylated peptide, the initiation methionine can be established.

MS-Based Identification of a 130 kDa Protein in the EGF Receptor Signaling Pathway

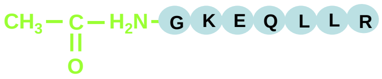


Assignment of the initiator methionine in a cDNA ‘fragment’ based on an N-terminal peptide

>KIAA0229 (1180 residues) FRAGMENT

SWGKGREGVVSPAGLGGALPGDGKFGSPSRLGCSLGEGVQRVAALGMGKEQ
ELLRAARTGHLPAVEKLLSGKRLSSGFGGGGGGGSGGGGGGSGGGGGGLGS
SSHPLSSLLSMWRGPNVNCVDSTGYTPLHHAALNGHHRRSSSSRSQDSAEGQ
DGQVPEQFSGLLHGSSPVCEVGQDPFQLLCTAGQSHPDGSPQQGACHKASM
QLEETGVHAPGASQPSALDQSKRVGYLTGLPTTNSRSHPETLTHTASPHPGGA
EEGDRSGAR

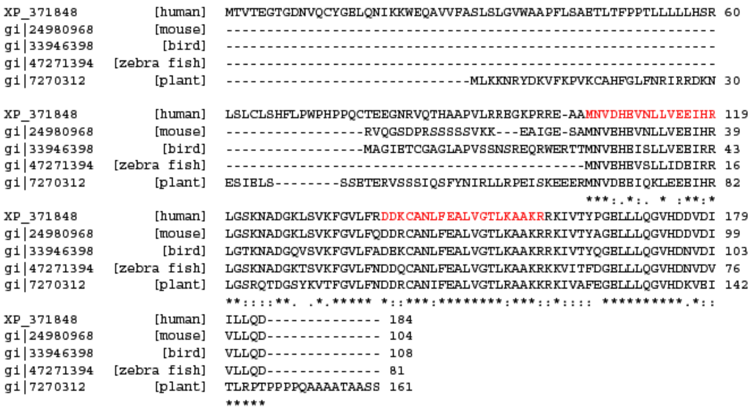
Assignment of the initiator methionine in a cDNA ‘fragment’ based on an N-terminal peptide



>KIAA0229 (1180 residues) FRAGMENT

SWGKGREGVVSPAGLGGALPGDGKFGSPSRLGCSLGEGVQRVAALG**MGKEQ**
LLRAARTGHLPAVEKLLSGKRLSSGFGGGGGGGSGGGGGGSGGGGGGLGSS
SHPLSSLLSMWRGPNVNCVDSTGYTPLHHAALNGHHRRSSSSRSQDSAEGQD
GQVPEQFSGLLHGSSPVCEVGQDPFQLLCTAGQSHPDGSPQQGACHKASMQL
EETGVHAPGASQPSALDQSKRVGYLTGLPTTNSRSHPETLTHTASPHPGGAEE
GDRSGAR

N-terminal Acetylated Peptide – Annotation of Start Codon



```
XP_371848          [human]  MTVTEGTGDNVQCYGELQNIKKWEQAVVFASLSLGVWAAPFLSAETLTFPPTLLLLLHSR 60
gi|24980968        [mouse]  ------------------------------------------------------------
gi|33946398         [bird]  ------------------------------------------------------------
gi|47271394   [zebra fish]  ------------------------------------------------------------
gi|7270312         [plant]  ------------------------------MLKKNRYDKVFKPVKCAHFGLFNRIRRDKN 30

XP_371848          [human]  LSLCLSHFLPWPHPPQCTEEGNRVQTHAAPVLRREGKPRRE-AAMNVDHEVNLLVEEIHR 119
gi|24980968        [mouse]  -----------------RVQGSDPRSSSSSVKK---EAIGE-SAMNVEHEVNLLVEEIHR 39
gi|33946398         [bird]  -----------------MAGIETCGAGLAPVSSNSREQRWERTTMNVEHEISLLVEEIRR 43
gi|47271394   [zebra fish]  --------------------------------------------MNVEHEVSLLIDEIRR 16
gi|7270312         [plant]  ESIELS--------SSETERVSSSIQSFYNIRLLRPEISKEEERMNVDEEIQKLEEEIHR 82
                                                                        ***:.*:. * :**:*
XP_371848          [human]  LGSKNADGKLSVKFGVLFRDDKCANLFEALVGTLKAAKRRKIVTYPGELLLQGVHDDVDI 179
gi|24980968        [mouse]  LGSKNADGKLSVKFGVLFQDDRCANLFEALVGTLKAAKRRKIVTYAGELLLQGVHDDVDI 99
gi|33946398         [bird]  LGTKNADGQVSVKFGVLFADEKCANLFEALVGTLKAAKRRKIVTYQGELLLQGVHDNVDI 103
gi|47271394   [zebra fish]  LGSKNADGKTSVKFGVLFNDDQCANLFEALVGTLKAAKRKKVITFDGELLLQGVHDNVDV 76
gi|7270312         [plant]  LGSRQTDGSYKVTFGVLFNDDRCANIFEALVGTLRAAKKRKIVAFEGELLLQGVHDKVEI 142
                            **::::**. .*.***** *::***:********:***::*:::: **********.*::
XP_371848          [human]  ILLQD-------------- 184
gi|24980968        [mouse]  VLLQD-------------- 104
gi|33946398         [bird]  VLLQD-------------- 108
gi|47271394   [zebra fish]  VLLQD-------------- 81
gi|7270312         [plant]  TLRPTPPPPQAAAATAASS 161
                            *****
```

Aligment of sequences from 5 species in databases. The sequence at the top (XP_371848) is the human protein predicted by gene prediction programs. Peptides identified by MS/MS are marked in bold red and conserved residues are marked with an asterisk. The open reading frame in the case of zebra fish was the only correctly annotated entry. The acetylated methionine in the case of the peptide provides clear evidence that this methionine residue marks the N-terminus of this family of proteins.

Protein Databases

- **Swiss-Prot**
- **nr (non-redundant protein database)**
- **RefSeq**
- **IPI (International Protein Index)**

Swiss-Prot

http://us.expasy.org/sprot/

- Swiss-prot is part of the ExPASy (Expert Protein Analysis System) proteomics server of the Swiss Institute of Bioinformatics.
- A highly curated protein sequence database with minimal redundancy
- Swiss-Prot currently contains 172,000 protein sequences representing 8,859 species
- 12,000 Human protein sequences

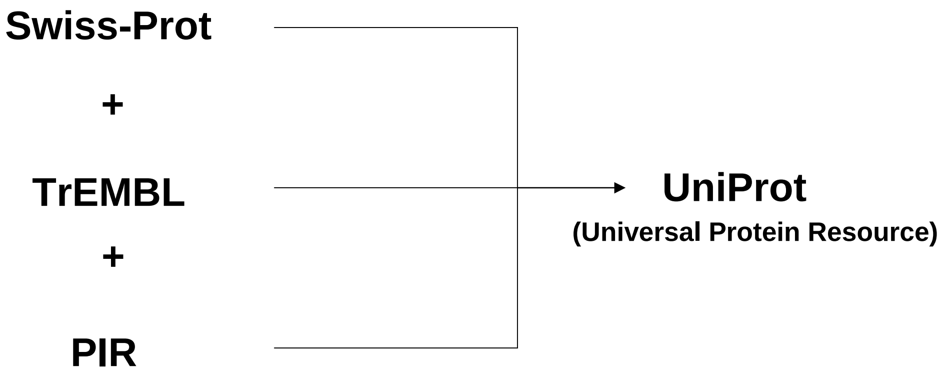
TrEMBL

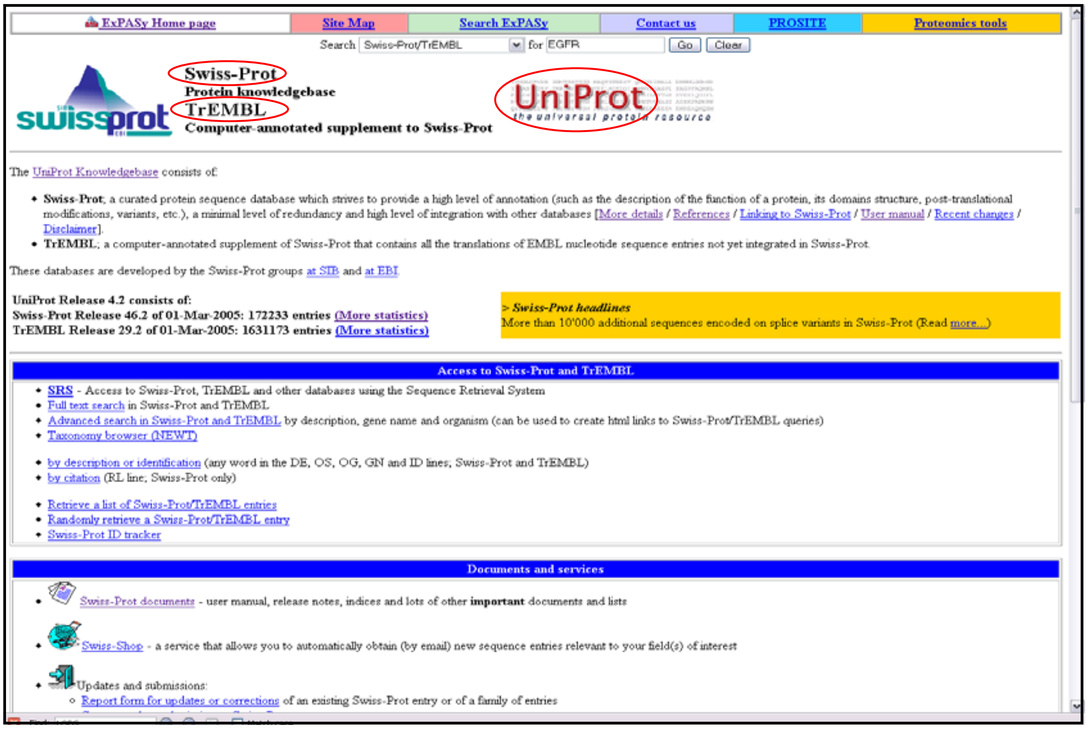
http://us.expasy.org/sprot/

- TrEMBL – A computer annotated supplement of Swiss-Prot containing all the translations of EMBL nucleotide sequence entries not yet integrated in Swiss-Prot
- TrEMBL can be considered as a preliminary section of Swiss-Prot
- TrEMBL is split in two main sections:
  SPTrEMBL and REMTrEMBL
- SPTrEMBL – All TrEMBL entries that should finally be upgraded to the standard Swiss-Prot quality, are assigned Swiss-Prot accessions
- REMTrEMBL – Remaining TrEMBL entries

UniProt

http://www.expasy.uniprot.org/





The UniProt Knowledgebase consists of:

- **Swiss-Prot**; a curated protein sequence database which strives to provide a high level of annotation (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases [More details / References / Linking to Swiss-Prot / User manual / Recent changes / Disclaimer].
- **TrEMBL**; a computer-annotated supplement of Swiss-Prot that contains all the translations of EMBL nucleotide sequence entries not yet integrated in Swiss-Prot.

These databases are developed by the Swiss-Prot groups at SIB and at EBI.

UniProt Release 4.2 consists of:
Swiss-Prot Release 46.2 of 01-Mar-2005: 172233 entries (More statistics)
TrEMBL Release 29.2 of 01-Mar-2005: 1631173 entries (More statistics)

> Swiss-Prot headlines
More than 10'000 additional sequences encoded on splice variants in Swiss-Prot (Read more...)

Access to Swiss-Prot and TrEMBL

- SRS - Access to Swiss-Prot, TrEMBL and other databases using the Sequence Retrieval System
- Full text search in Swiss-Prot and TrEMBL
- Advanced search in Swiss-Prot and TrEMBL by description, gene name and organism (can be used to create html links to Swiss-Prot/TrEMBL queries)
- Taxonomy browser (NEWT)

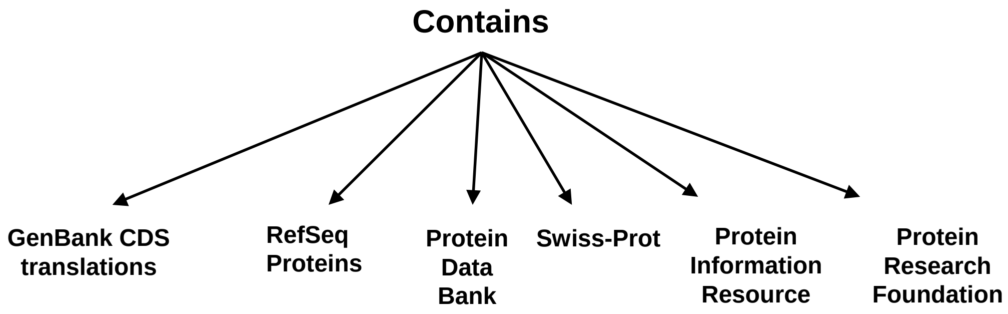
- by description or identification (any word in the DE, OS, OG, GN and ID lines; Swiss-Prot and TrEMBL)
- by citation (RL line; Swiss-Prot only)

- Retrieve a list of Swiss-Prot/TrEMBL entries
- Randomly retrieve a Swiss-Prot/TrEMBL entry
- Swiss-Prot ID tracker

Documents and services

- Swiss-Prot documents - user manual, release notes, indices and lots of other **important** documents and lists
- Swiss-Shop - a service that allows you to automatically obtain (by email) new sequence entries relevant to your field(s) of interest
- Updates and submissions:
  - Report form for updates or corrections of an existing Swiss-Prot entry or of a family of entries

nr (non-redundant) database



Contains

GenBank CDS translations

RefSeq Proteins

Protein Data Bank

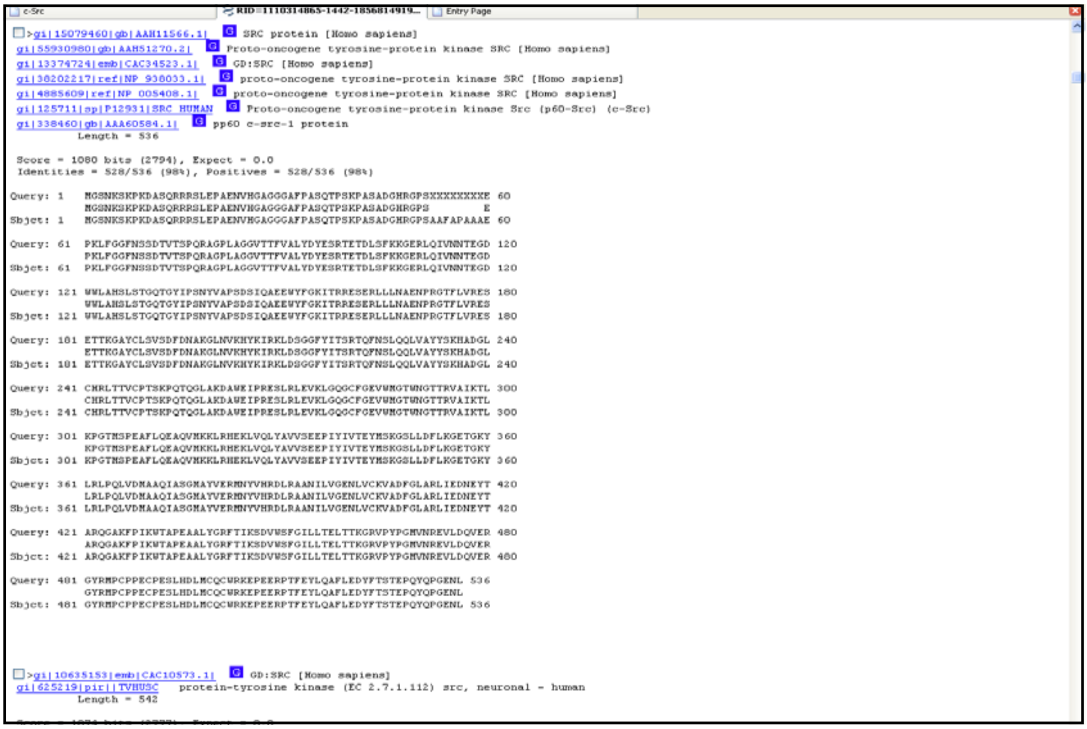
Swiss-Prot

Protein Information Resource

Protein Research Foundation

nr (non-redundant) database

- All identical sequences from any of the above databases are merged into a single entry
- It contains 1,800,000 protein sequences from 33,362 species
- Still NOT non-redundant (=VERY Redundant)



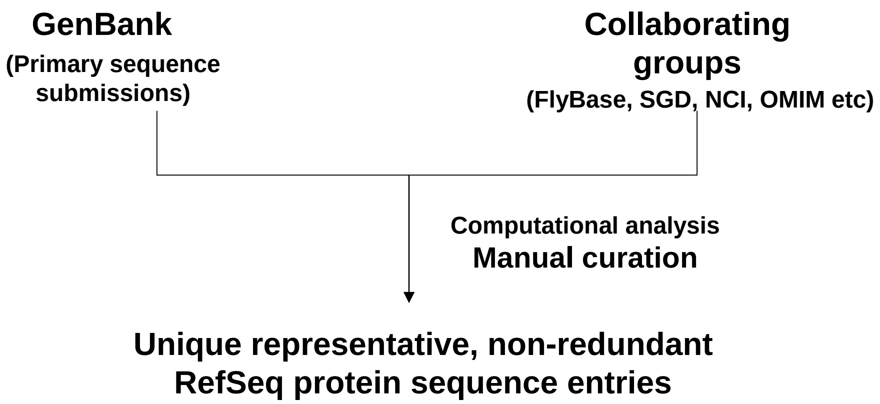
```
□>gi|15079460|gb|AAH11566.1|  G SRC protein [Homo sapiens]
 gi|55930980|gb|AAH51270.2| G Proto-oncogene tyrosine-protein kinase SRC [Homo sapiens]
 gi|13374724|emb|CAC34523.1| G GD:SRC [Homo sapiens]
 gi|38202217|ref|NP_938033.1| G proto-oncogene tyrosine-protein kinase SRC [Homo sapiens]
 gi|4885609|ref|NP_005408.1| G proto-oncogene tyrosine-protein kinase SRC [Homo sapiens]
 gi|125711|sp|P12931|SRC_HUMAN G Proto-oncogene tyrosine-protein kinase Src (p60-Src) (c-Src)
 gi|338460|gb|AAA60584.1| G pp60 c-src-1 protein
          Length = 536

 Score = 1080 bits (2794), Expect = 0.0
 Identities = 528/536 (98%), Positives = 528/536 (98%)

Query: 1   MGSNKSKPKDASQRRRSLEPAENVHGAGGGAFPASQTPSKPASADGHRGPSXXXXXXXXE 60
            MGSNKSKPKDASQRRRSLEPAENVHGAGGGAFPASQTPSKPASADGHRGPS        E
Sbjct: 1   MGSNKSKPKDASQRRRSLEPAENVHGAGGGAFPASQTPSKPASADGHRGPSAAFAPAAAE 60

Query: 61  PKLFGGFNSSDTVTSPQRAGPLAGGVTTFVALYDYESRTETDLSFKKGERLQIVNNTEGD 120
            PKLFGGFNSSDTVTSPQRAGPLAGGVTTFVALYDYESRTETDLSFKKGERLQIVNNTEGD
Sbjct: 61  PKLFGGFNSSDTVTSPQRAGPLAGGVTTFVALYDYESRTETDLSFKKGERLQIVNNTEGD 120

Query: 121 WWLAHSLSTGQTGYIPSNYVAPSDSIQAEEWYFGKITRRESERLLLNAENPRGTFLVRES 180
            WWLAHSLSTGQTGYIPSNYVAPSDSIQAEEWYFGKITRRESERLLLNAENPRGTFLVRES
Sbjct: 121 WWLAHSLSTGQTGYIPSNYVAPSDSIQAEEWYFGKITRRESERLLLNAENPRGTFLVRES 180

Query: 181 ETTKGAYCLSVSDFDNAKGLNVKHYKIRKLDSGGFYITSRTQFNSLQQLVAYYSKHADGL 240
            ETTKGAYCLSVSDFDNAKGLNVKHYKIRKLDSGGFYITSRTQFNSLQQLVAYYSKHADGL
Sbjct: 181 ETTKGAYCLSVSDFDNAKGLNVKHYKIRKLDSGGFYITSRTQFNSLQQLVAYYSKHADGL 240

Query: 241 CHRLTTVCPTSKPQTQGLAKDAWEIPRESLRLEVKLGQGCFGEVWMGTWNGTTRVAIKTL 300
            CHRLTTVCPTSKPQTQGLAKDAWEIPRESLRLEVKLGQGCFGEVWMGTWNGTTRVAIKTL
Sbjct: 241 CHRLTTVCPTSKPQTQGLAKDAWEIPRESLRLEVKLGQGCFGEVWMGTWNGTTRVAIKTL 300

Query: 301 KPGTMSPEAFLQEAQVMKKLRHEKLVQLYAVVSEEPIYIVTEYMSKGSLLDFLKGETGKY 360
            KPGTMSPEAFLQEAQVMKKLRHEKLVQLYAVVSEEPIYIVTEYMSKGSLLDFLKGETGKY
Sbjct: 301 KPGTMSPEAFLQEAQVMKKLRHEKLVQLYAVVSEEPIYIVTEYMSKGSLLDFLKGETGKY 360

Query: 361 LRLPQLVDMAAQIASGMAYVERMNYVHRDLRAANILVGENLVCKVADFGLARLIEDNEYT 420
            LRLPQLVDMAAQIASGMAYVERMNYVHRDLRAANILVGENLVCKVADFGLARLIEDNEYT
Sbjct: 361 LRLPQLVDMAAQIASGMAYVERMNYVHRDLRAANILVGENLVCKVADFGLARLIEDNEYT 420

Query: 421 ARQGAKFPIKWTAPEAALYGRFTIKSDVWSFGILLTELTTKGRVPYPGMVNREVLDQVER 480
            ARQGAKFPIKWTAPEAALYGRFTIKSDVWSFGILLTELTTKGRVPYPGMVNREVLDQVER
Sbjct: 421 ARQGAKFPIKWTAPEAALYGRFTIKSDVWSFGILLTELTTKGRVPYPGMVNREVLDQVER 480

Query: 481 GYRMPCPPECPESLHDLMCQCWRKEPEERPTFEYLQAFLEDYFTSTEPQYQPGENL 536
            GYRMPCPPECPESLHDLMCQCWRKEPEERPTFEYLQAFLEDYFTSTEPQYQPGENL
Sbjct: 481 GYRMPCPPECPESLHDLMCQCWRKEPEERPTFEYLQAFLEDYFTSTEPQYQPGENL 536



□>gi|10635153|emb|CAC10573.1|  G GD:SRC [Homo sapiens]
 gi|625219|pir||TVHUSC   protein-tyrosine kinase (EC 2.7.1.112) src, neuronal - human
          Length = 542
```

RefSeq (Reference Sequence) database

http://www.ncbi.nlm.nih.gov/RefSeq/

- **RefSeq database is a result of collaborative effort of NCBI and other groups and databases like TIGR, FlyBase, WormBase etc.**
- **A comprehensive, integrated and highly non-redundant curated protein sequence database**
- **28,000 Human protein sequences**
- **Contains protein sequences from all major research organisms**
- **Alternate splice forms listed individually**
- **Also contains predicted proteins translated from predicted transcripts (designated as XP_ entries)**

RefSeq (Reference Sequence) database



GenBank

(Primary sequence submissions)

Collaborating groups

(FlyBase, SGD, NCI, OMIM etc)

Computational analysis

Manual curation

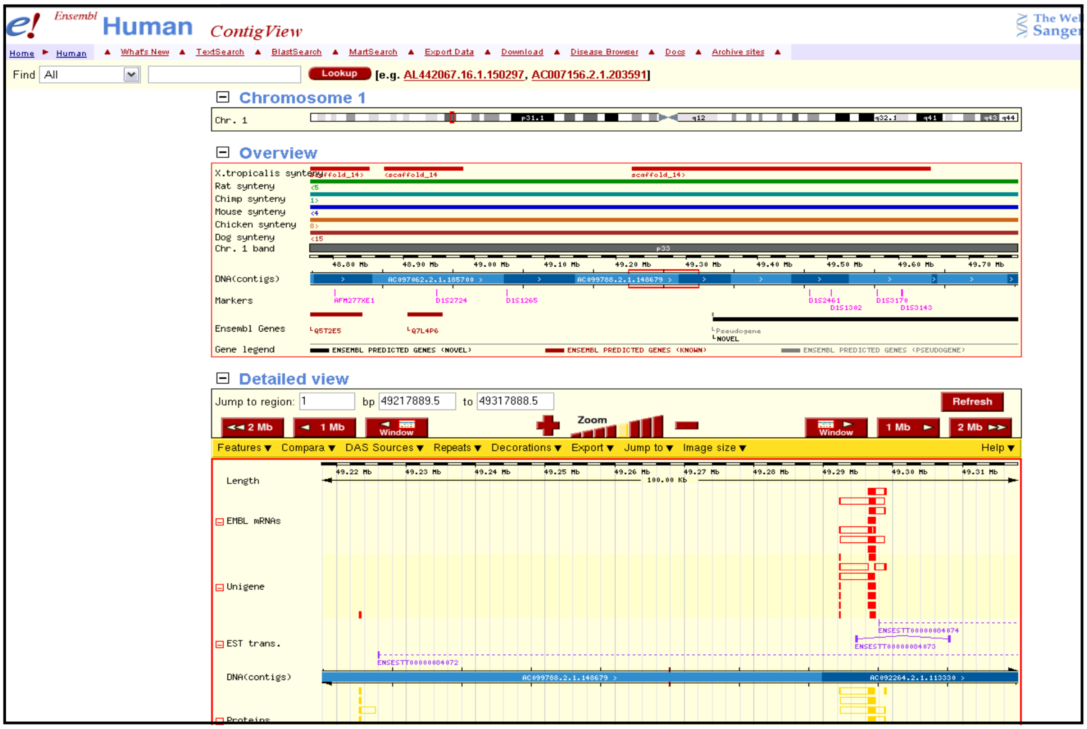
Unique representative, non-redundant RefSeq protein sequence entries

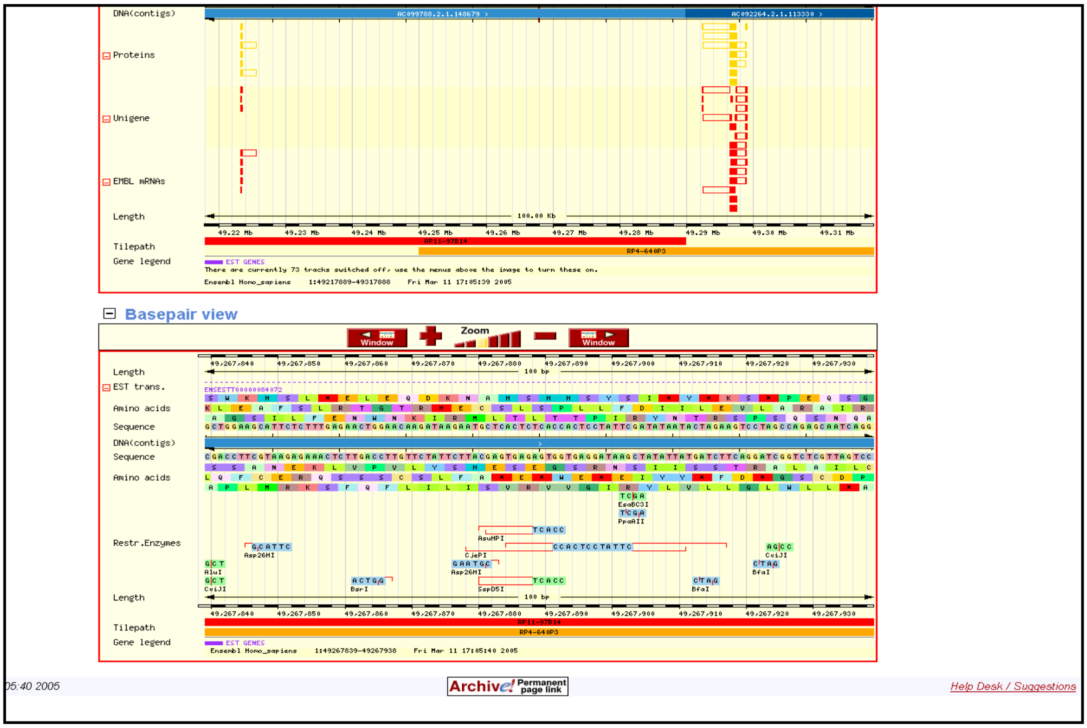
Ensembl database

http://www.ebi.ac.uk/ensembl/

- Ensembl is a joint project between the EMBL-EBI and the Wellcome Trust Sanger Institute that aims at developing a system that maintains automatic annotation of large eukaryotic genomes. database is a result of collaborative effort of NCBI and other groups and databases like TIGR, FlyBase, WormBase etc.

- It is a comprehensive source of stable annotation with confirmed gene predictions that have been integrated from external data sources.





There are currently 73 tracks switched off, use the menus above the image to turn these on.

Ensembl Homo_sapiens 1:49217889-49317888 Fri Mar 11 17:05:39 2005

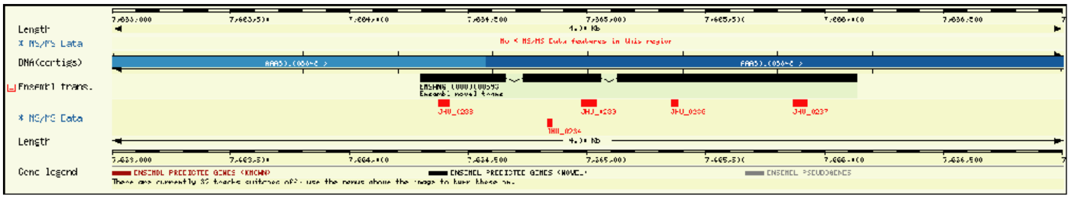
GCTGGAAGCATTCTCTTTGAGAACTGGAACAAGATAAGAATGCTCACTCTCACCACTCCTATTCGATATAATACTAGAAGTCCTAGCCAGAGCAATCAGG

CGACCTTCGTAAGAGAAACTCTTGACCTTGTTCTATTCTTACGAGTGAGAGTGGTGAGGATAAGCTATATTATGATCTTCAGGATCGGTCTCGTTAGTCC

Ensembl Homo_sapiens 1:49267839-49267938 Fri Mar 11 17:05:40 2005

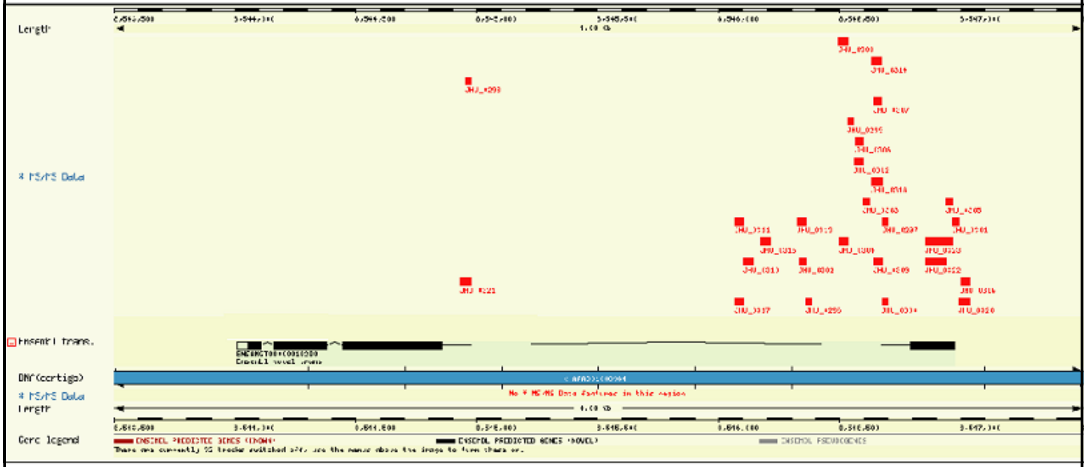
05:40 2005 Archive! Permanent page link Help Desk / Suggestions

Use of Ensembl Distributed Annotation System to Validate a Predicted Transcript



There are currently 32 tracks switched off; use the menus above the image to turn these on.

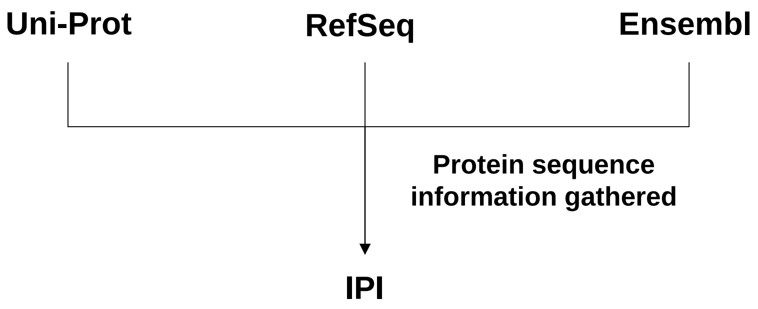
Correction of a Predicted Transcript



There are currently 32 tracks switched off; use the menus above the image to turn these on.

IPI (International Protein Index) database

http://www.ebi.ac.uk/IPI/



Uni-Prot

RefSeq

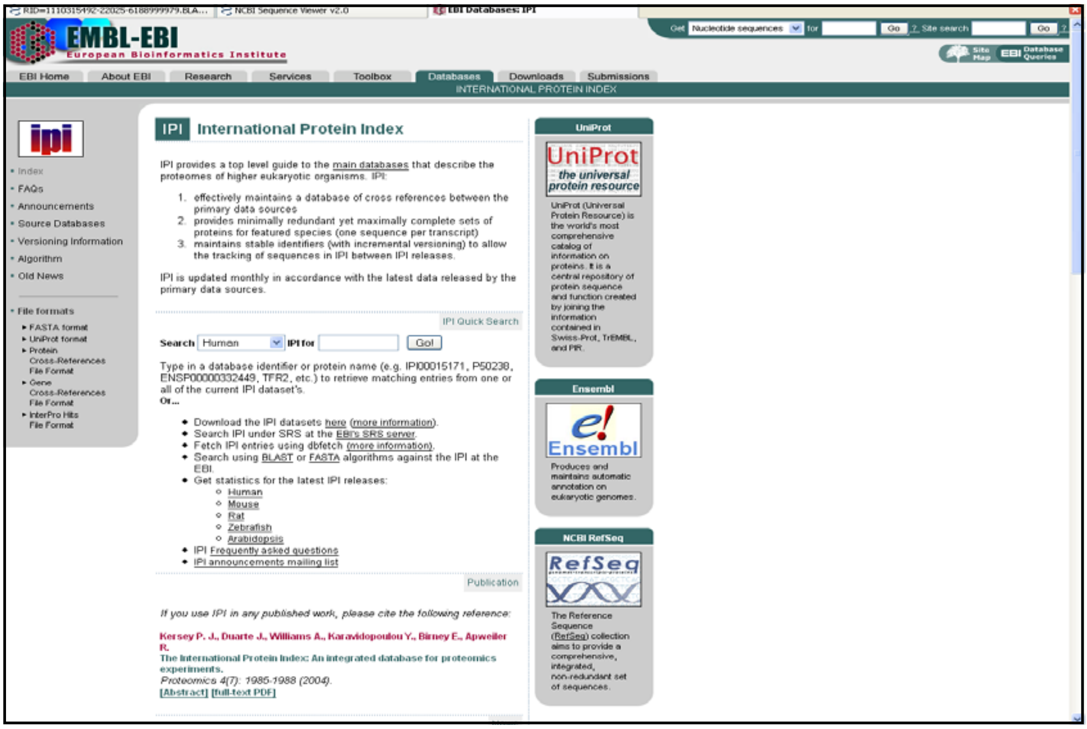
Ensembl

Protein sequence information gathered

IPI

IPI (International Protein Index) database

- IPI is a protein database from the European Bioinformatics Institute
- Has protein sequence information from Human, Mouse, Rat, Zebra fish and Arabidopsis species only
- 49,000 Human protein sequences
- A redundant database
- Has information on protein isoforms
- The sequence identifiers and sequence entries are not stable



IPI International Protein Index

IPI provides a top level guide to the main databases that describe the proteomes of higher eukaryotic organisms. IPI:

1. effectively maintains a database of cross references between the primary data sources
2. provides minimally redundant yet maximally complete sets of proteins for featured species (one sequence per transcript)
3. maintains stable identifiers (with incremental versioning) to allow the tracking of sequences in IPI between IPI releases.

IPI is updated monthly in accordance with the latest data released by the primary data sources.

Type in a database identifier or protein name (e.g. IPI00015171, P50238, ENSP00000332449, TFR2, etc.) to retrieve matching entries from one or all of the current IPI dataset's.

Or...

- Download the IPI datasets here (more information).
- Search IPI under SRS at the EBI's SRS server.
- Fetch IPI entries using dbfetch (more information).
- Search using BLAST or FASTA algorithms against the IPI at the EBI.
- Get statistics for the latest IPI releases:
  - Human
  - Mouse
  - Rat
  - Zebrafish
  - Arabidopsis
- IPI Frequently asked questions
- IPI announcements mailing list

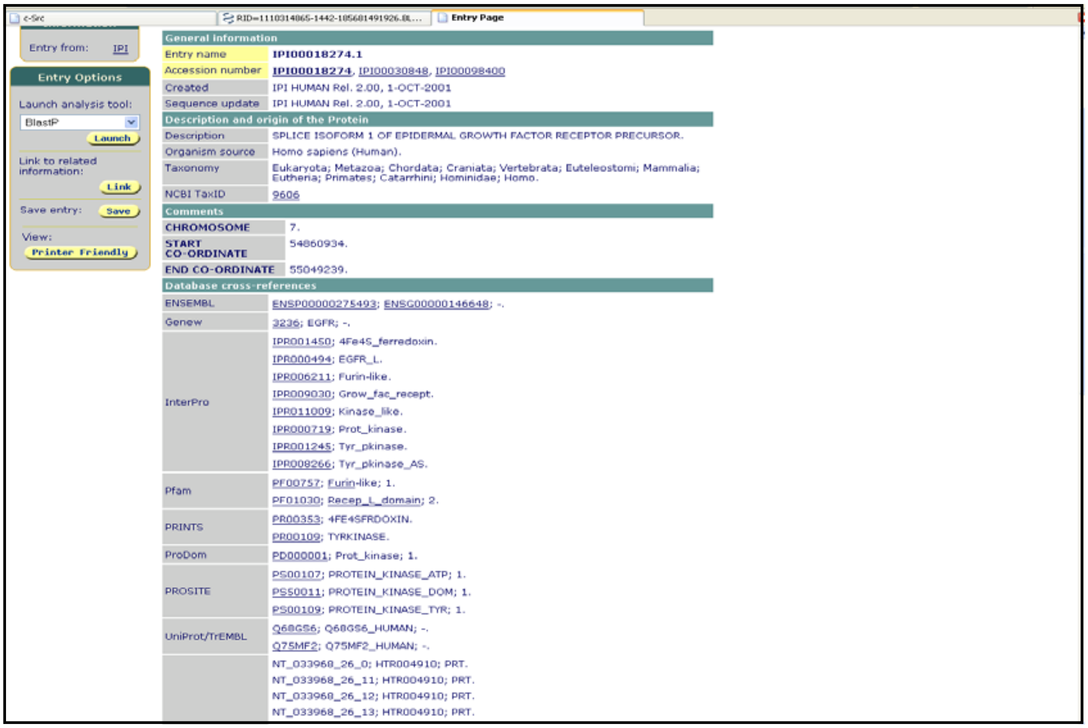
If you use IPI in any published work, please cite the following reference:

Kersey P. J., Duarte J., Williams A., Karavidopoulou Y., Birney E., Apweiler R.
The International Protein Index: An integrated database for proteomics experiments.
Proteomics 4(7): 1985-1988 (2004).
[Abstract] [full-text PDF]

UniProt (Universal Protein Resource) is the world's most comprehensive catalog of information on proteins. It is a central repository of protein sequence and function created by joining the information contained in Swiss-Prot, TrEMBL, and PIR.

Produces and maintains automatic annotation on eukaryotic genomes.

The Reference Sequence (RefSeq) collection aims to provide a comprehensive, integrated, non-redundant set of sequences.



Entry from: IPI

Launch analysis tool: BlastP

Link to related information:

Save entry:

View:

General information	
Entry name	**IPI00018274.1**
Accession number	**IPI00018274**, IPI00030848, IPI00098400
Created	IPI HUMAN Rel. 2.00, 1-OCT-2001
Sequence update	IPI HUMAN Rel. 2.00, 1-OCT-2001
Description and origin of the Protein	
Description	SPLICE ISOFORM 1 OF EPIDERMAL GROWTH FACTOR RECEPTOR PRECURSOR.
Organism source	Homo sapiens (Human).
Taxonomy	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
NCBI TaxID	9606
Comments	
CHROMOSOME	7.
START CO-ORDINATE	54860934.
END CO-ORDINATE	55049239.
Database cross-references	
ENSEMBL	ENSP00000275493; ENSG00000146648; -.
Genew	3236; EGFR; -.
InterPro	IPR001450; 4Fe4S_ferredoxin. IPR000494; EGFR_L. IPR006211; Furin-like. IPR009030; Grow_fac_recept. IPR011009; Kinase_like. IPR000719; Prot_kinase. IPR001245; Tyr_pkinase. IPR008266; Tyr_pkinase_AS.
Pfam	PF00757; Furin-like; 1. PF01030; Recep_L_domain; 2.
PRINTS	PR00353; 4FE4SFRDOXIN. PR00109; TYRKINASE.
ProDom	PD000001; Prot_kinase; 1.
PROSITE	PS00107; PROTEIN_KINASE_ATP; 1. PS50011; PROTEIN_KINASE_DOM; 1. PS00109; PROTEIN_KINASE_TYR; 1.
UniProt/TrEMBL	Q68GS6; Q68GS6_HUMAN; -. Q75MF2; Q75MF2_HUMAN; -.
	NT_033968_26_0; HTR004910; PRT. NT_033968_26_11; HTR004910; PRT. NT_033968_26_12; HTR004910; PRT. NT_033968_26_13; HTR004910; PRT.

Databases/Tools for protein information

Protein Information resources

Databases

- Swiss-Prot (http://us.expasy.org/sprot/)
- HPRD (Human Protein Reference Database) (http://www.hprd.org)

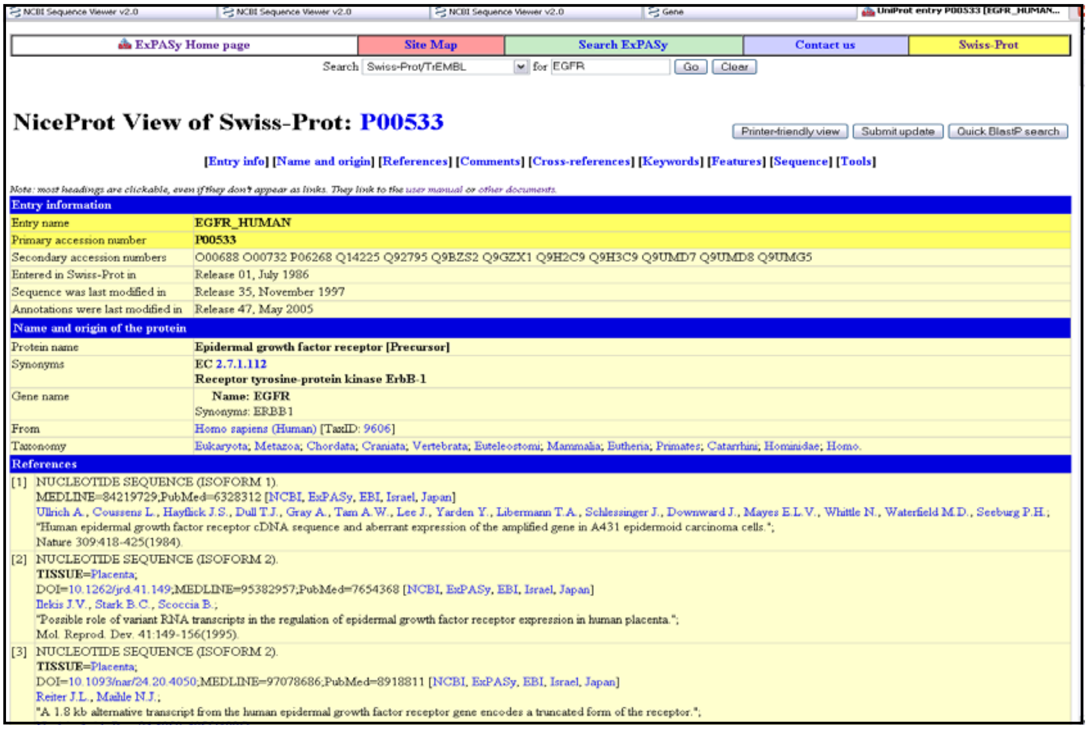
Tools

- SMART (http://smart.embl-heidelberg.de/)
- Pfam (http://www.sanger.ac.uk/Software/Pfam/)
- PSORT (http://psort.nibb.ac.jp/)

Swiss-Prot

Type of information than can be obtained for the protein of interest

- Function
- Architecture of protein (e.g. Domains, motifs)
- Post-translational modifications
- Alternate splice forms
- Localization
- Protein variants
- Cross-References to many other databases


NCBI Sequence Viewer v2.0
NCBI Sequence Viewer v2.0
NCBI Sequence Viewer v2.0
Gene
UniProt entry P00533 [EGFR_HUMAN...
ExPASy Home page
Site Map
Search ExPASy
Contact us
Swiss-Prot
Search Swiss-Prot/TrEMBL for EGFR Go Clear
Printer-friendly view
Submit update
Quick BlastP search

NiceProt View of Swiss-Prot: P00533

[Entry info] [Name and origin] [References] [Comments] [Cross-references] [Keywords] [Features] [Sequence] [Tools]

Note: most headings are clickable, even if they don't appear as links. They link to the user manual or other documents.

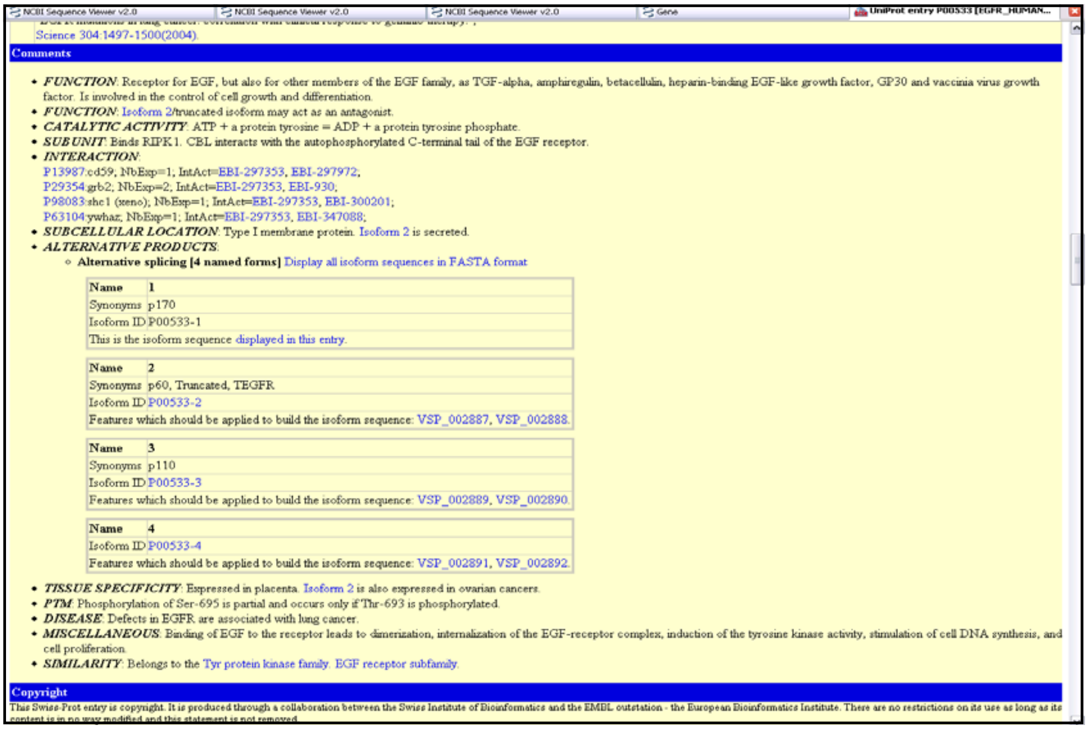
Entry information	
Entry name	**EGFR_HUMAN**
Primary accession number	**P00533**
Secondary accession numbers	O00688 O00732 P06268 Q14225 Q92795 Q9BZS2 Q9GZX1 Q9H2C9 Q9H3C9 Q9UMD7 Q9UMD8 Q9UMG5
Entered in Swiss-Prot in	Release 01, July 1986
Sequence was last modified in	Release 35, November 1997
Annotations were last modified in	Release 47, May 2005
Name and origin of the protein	
Protein name	**Epidermal growth factor receptor [Precursor]**
Synonyms	**EC 2.7.1.112** **Receptor tyrosine-protein kinase ErbB-1**
Gene name	**Name: EGFR** Synonyms: ERBB1
From	Homo sapiens (Human) [TaxID: 9606]
Taxonomy	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.

References

[1] NUCLEOTIDE SEQUENCE (ISOFORM 1).
MEDLINE=84219729;PubMed=6328312 [NCBI, ExPASy, EBI, Israel, Japan]
Ullrich A., Coussens L., Hayflick J.S., Dull T.J., Gray A., Tam A.W., Lee J., Yarden Y., Libermann T.A., Schlessinger J., Downward J., Mayes E.L.V., Whittle N., Waterfield M.D., Seeburg P.H.;
"Human epidermal growth factor receptor cDNA sequence and aberrant expression of the amplified gene in A431 epidermoid carcinoma cells.";
Nature 309:418-425(1984).

[2] NUCLEOTIDE SEQUENCE (ISOFORM 2).
TISSUE=Placenta;
DOI=10.1262/jrd.41.149;MEDLINE=95382957;PubMed=7654368 [NCBI, ExPASy, EBI, Israel, Japan]
Ilekis J.V., Stark B.C., Scoccia B.;
"Possible role of variant RNA transcripts in the regulation of epidermal growth factor receptor expression in human placenta.";
Mol. Reprod. Dev. 41:149-156(1995).

[3] NUCLEOTIDE SEQUENCE (ISOFORM 2).
TISSUE=Placenta;
DOI=10.1093/nar/24.20.4050;MEDLINE=97078686;PubMed=8918811 [NCBI, ExPASy, EBI, Israel, Japan]
Reiter J.L., Maihle N.J.;
"A 1.8 kb alternative transcript from the human epidermal growth factor receptor gene encodes a truncated form of the receptor.";


NCBI Sequence Viewer v2.0
NCBI Sequence Viewer v2.0
NCBI Sequence Viewer v2.0
Gene
UniProt entry P00533 [EGFR_HUMAN...

"EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy.";
Science 304:1497-1500(2004).

Comments

- ***FUNCTION***: Receptor for EGF, but also for other members of the EGF family, as TGF-alpha, amphiregulin, betacellulin, heparin-binding EGF-like growth factor, GP30 and vaccinia virus growth factor. Is involved in the control of cell growth and differentiation.
- ***FUNCTION***: Isoform 2/truncated isoform may act as an antagonist.
- ***CATALYTIC ACTIVITY***: ATP + a protein tyrosine = ADP + a protein tyrosine phosphate.
- ***SUBUNIT***: Binds RIPK1. CBL interacts with the autophosphorylated C-terminal tail of the EGF receptor.
- ***INTERACTION***:
  P13987:cd59; NbExp=1; IntAct=EBI-297353, EBI-297972;
  P29354:grb2; NbExp=2; IntAct=EBI-297353, EBI-930;
  P98083:shc1 (xeno); NbExp=1; IntAct=EBI-297353, EBI-300201;
  P63104:ywhaz; NbExp=1; IntAct=EBI-297353, EBI-347088;
- ***SUBCELLULAR LOCATION***: Type I membrane protein. Isoform 2 is secreted.
- ***ALTERNATIVE PRODUCTS***:
  - **Alternative splicing [4 named forms]** Display all isoform sequences in FASTA format

Name	**1**
Synonyms	p170
Isoform ID	P00533-1
This is the isoform sequence displayed in this entry.	

Name	**2**
Synonyms	p60, Truncated, TEGFR
Isoform ID	P00533-2
Features which should be applied to build the isoform sequence: VSP_002887, VSP_002888.	

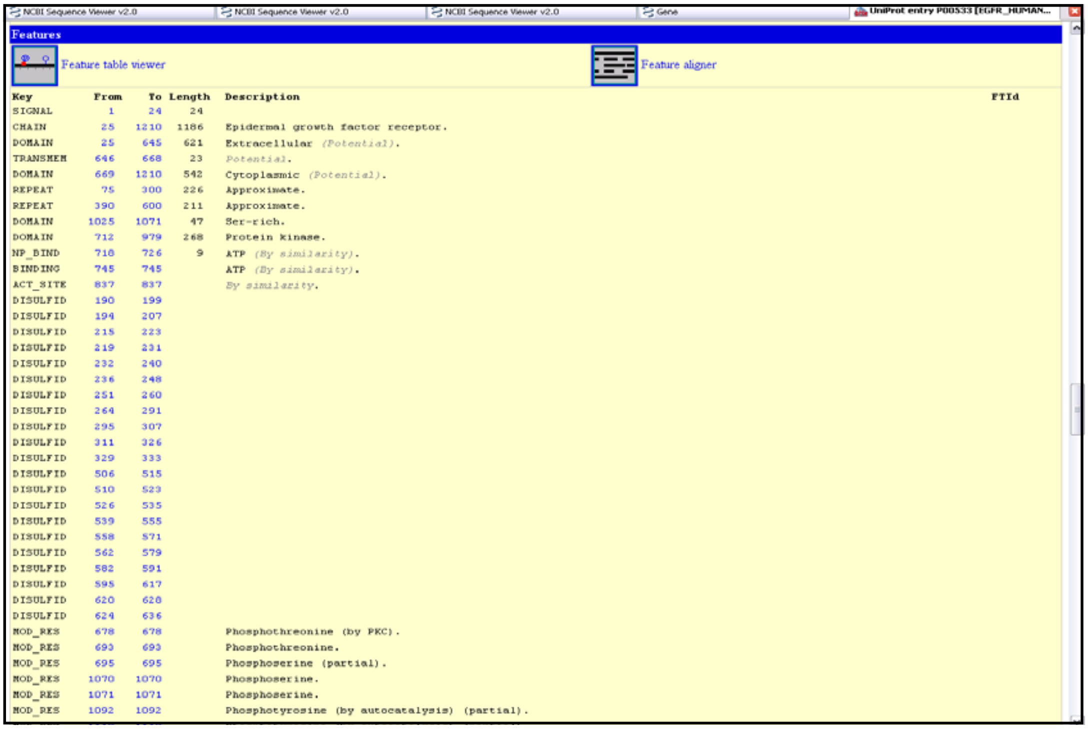
Name	**3**
Synonyms	p110
Isoform ID	P00533-3
Features which should be applied to build the isoform sequence: VSP_002889, VSP_002890.	

Name	**4**
Isoform ID	P00533-4
Features which should be applied to build the isoform sequence: VSP_002891, VSP_002892.	

- ***TISSUE SPECIFICITY***: Expressed in placenta. Isoform 2 is also expressed in ovarian cancers.
- ***PTM***: Phosphorylation of Ser-695 is partial and occurs only if Thr-693 is phosphorylated.
- ***DISEASE***: Defects in EGFR are associated with lung cancer.
- ***MISCELLANEOUS***: Binding of EGF to the receptor leads to dimerization, internalization of the EGF-receptor complex, induction of the tyrosine kinase activity, stimulation of cell DNA synthesis, and cell proliferation.
- ***SIMILARITY***: Belongs to the Tyr protein kinase family. EGF receptor subfamily.

Copyright

This Swiss-Prot entry is copyright. It is produced through a collaboration between the Swiss Institute of Bioinformatics and the EMBL outstation - the European Bioinformatics Institute. There are no restrictions on its use as long as its content is in no way modified and this statement is not removed.


NCBI Sequence Viewer v2.0
NCBI Sequence Viewer v2.0
NCBI Sequence Viewer v2.0
Gene
UniProt entry P00533 [EGFR_HUMAN...
Feature table viewer
Feature aligner

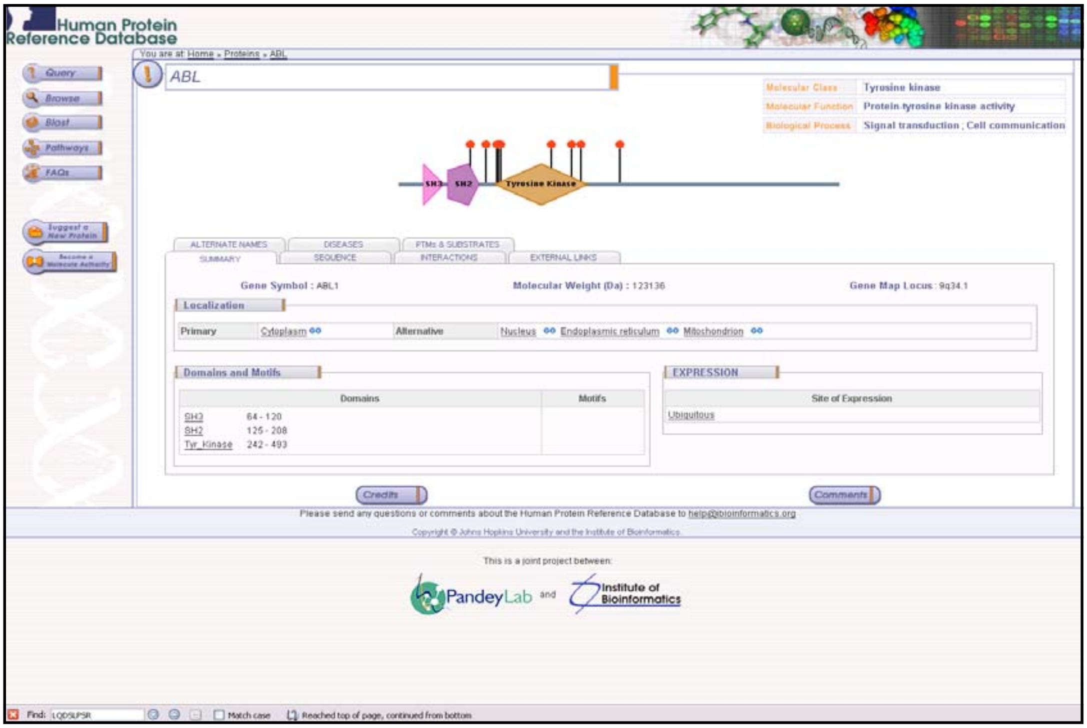
Features

Key	From	To	Length	Description	FTId
SIGNAL	1	24	24		
CHAIN	25	1210	1186	Epidermal growth factor receptor.	
DOMAIN	25	645	621	Extracellular *(Potential)*.	
TRANSMEM	646	668	23	*Potential*.	
DOMAIN	669	1210	542	Cytoplasmic *(Potential)*.	
REPEAT	75	300	226	Approximate.	
REPEAT	390	600	211	Approximate.	
DOMAIN	1025	1071	47	Ser-rich.	
DOMAIN	712	979	268	Protein kinase.	
NP_BIND	718	726	9	ATP *(By similarity)*.	
BINDING	745	745		ATP *(By similarity)*.	
ACT_SITE	837	837		*By similarity*.	
DISULFID	190	199			
DISULFID	194	207			
DISULFID	215	223			
DISULFID	219	231			
DISULFID	232	240			
DISULFID	236	248			
DISULFID	251	260			
DISULFID	264	291			
DISULFID	295	307			
DISULFID	311	326			
DISULFID	329	333			
DISULFID	506	515			
DISULFID	510	523			
DISULFID	526	535			
DISULFID	539	555			
DISULFID	558	571			
DISULFID	562	579			
DISULFID	582	591			
DISULFID	595	617			
DISULFID	620	628			
DISULFID	624	636			
MOD_RES	678	678		Phosphothreonine (by PKC).	
MOD_RES	693	693		Phosphothreonine.	
MOD_RES	695	695		Phosphoserine (partial).	
MOD_RES	1070	1070		Phosphoserine.	
MOD_RES	1071	1071		Phosphoserine.	
MOD_RES	1092	1092		Phosphotyrosine (by autocatalysis) (partial).	

HPRD (Human Protein Reference Database)

Type of information than can be obtained for the protein of interest

- **Function**
- **Architecture of protein (e.g. Domains, motifs)**
- **Post-translational modifications**
- **Expression**
- **Localization**
- **Disease associations**
- **Protein-protein interactions**



You are at: Home » Proteins » ABL

ABL

Molecular Class	Tyrosine kinase
Molecular Function	Protein-tyrosine kinase activity
Biological Process	Signal transduction ; Cell communication

Gene Symbol : ABL1 Molecular Weight (Da) : 123136 Gene Map Locus : 9q34.1

Localization

Primary	Cytoplasm	Alternative	Nucleus Endoplasmic reticulum Mitochondrion

Domains and Motifs

Domains		Motifs
SH3	64 - 120	
SH2	125 - 208	
Tyr_Kinase	242 - 493	

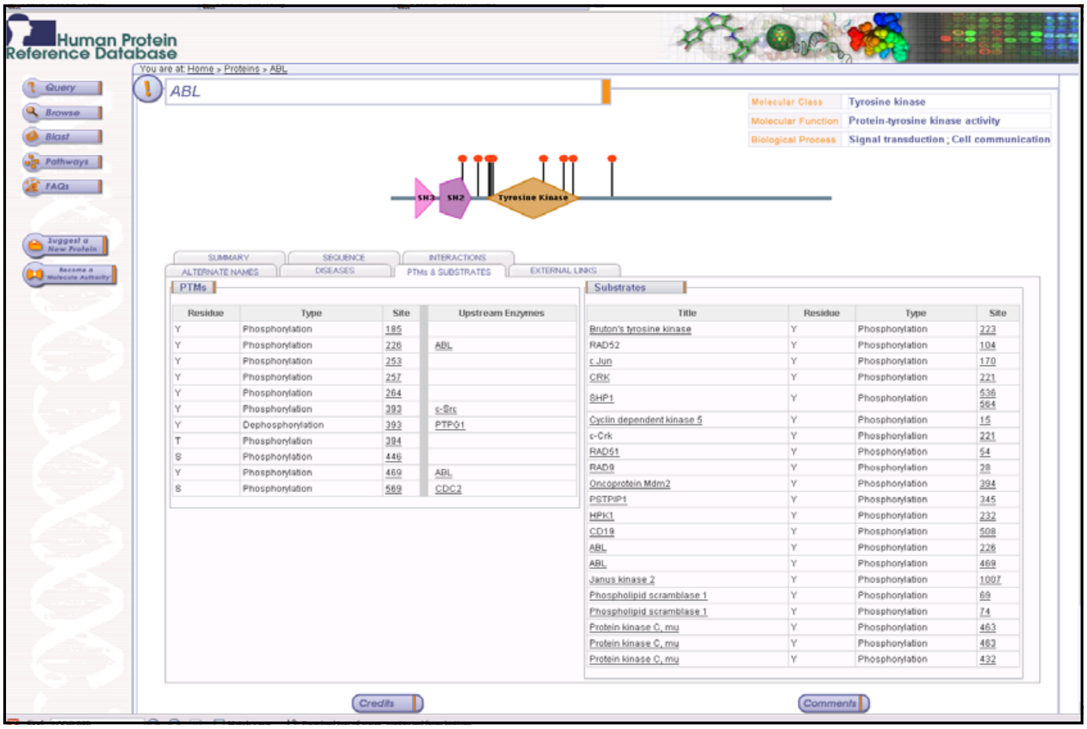
EXPRESSION

Site of Expression
Ubiquitous

Please send any questions or comments about the Human Protein Reference Database to help@bioinformatics.org

Copyright © Johns Hopkins University and the Institute of Bioinformatics

This is a joint project between:



ABL

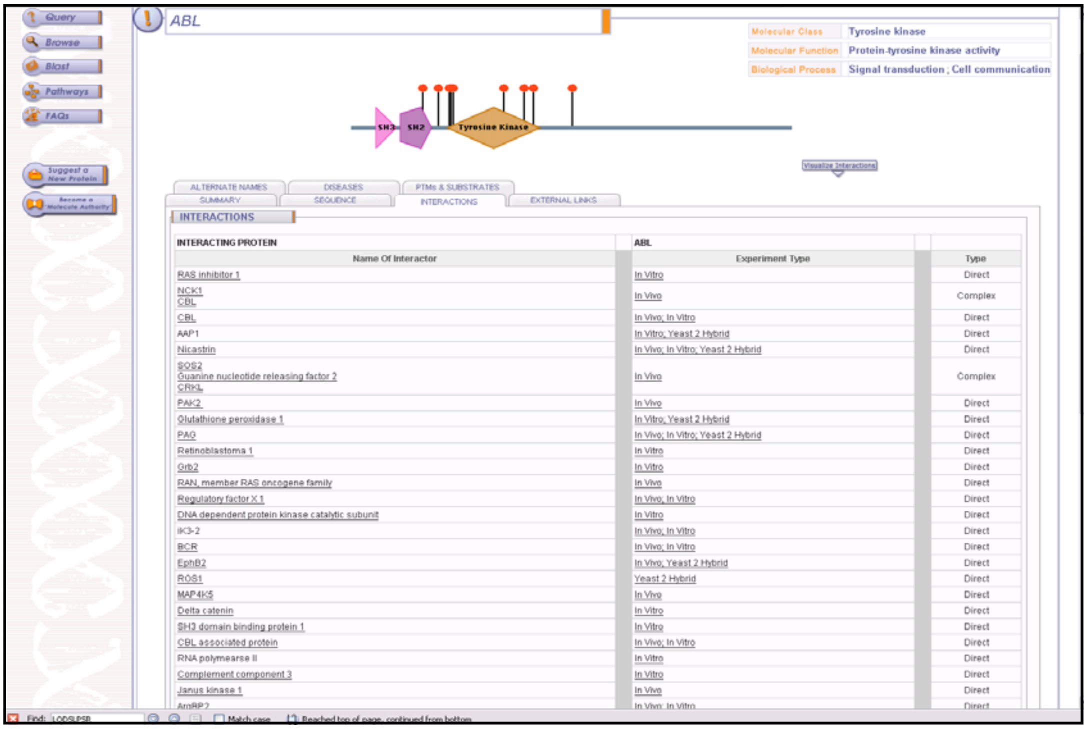
Molecular Class	Tyrosine kinase
Molecular Function	Protein-tyrosine kinase activity
Biological Process	Signal transduction ; Cell communication

PTMs

Residue	Type	Site	Upstream Enzymes
Y	Phosphorylation	185	
Y	Phosphorylation	226	ABL
Y	Phosphorylation	253	
Y	Phosphorylation	257	
Y	Phosphorylation	264	
Y	Phosphorylation	393	c-Src
Y	Dephosphorylation	393	PTPG1
T	Phosphorylation	394	
S	Phosphorylation	446	
Y	Phosphorylation	469	ABL
S	Phosphorylation	569	CDC2

Substrates

Title	Residue	Type	Site
Bruton's tyrosine kinase	Y	Phosphorylation	223
RAD52	Y	Phosphorylation	104
c Jun	Y	Phosphorylation	170
CRK	Y	Phosphorylation	221
SHP1	Y	Phosphorylation	536 564
Cyclin dependent kinase 5	Y	Phosphorylation	15
c-Crk	Y	Phosphorylation	221
RAD51	Y	Phosphorylation	54
RAD9	Y	Phosphorylation	28
Oncoprotein Mdm2	Y	Phosphorylation	394
PSTPIP1	Y	Phosphorylation	345
HPK1	Y	Phosphorylation	232
CD19	Y	Phosphorylation	508
ABL	Y	Phosphorylation	226
ABL	Y	Phosphorylation	469
Janus kinase 2	Y	Phosphorylation	1007
Phospholipid scramblase 1	Y	Phosphorylation	69
Phospholipid scramblase 1	Y	Phosphorylation	74
Protein kinase C, mu	Y	Phosphorylation	463
Protein kinase C, mu	Y	Phosphorylation	463
Protein kinase C, mu	Y	Phosphorylation	432

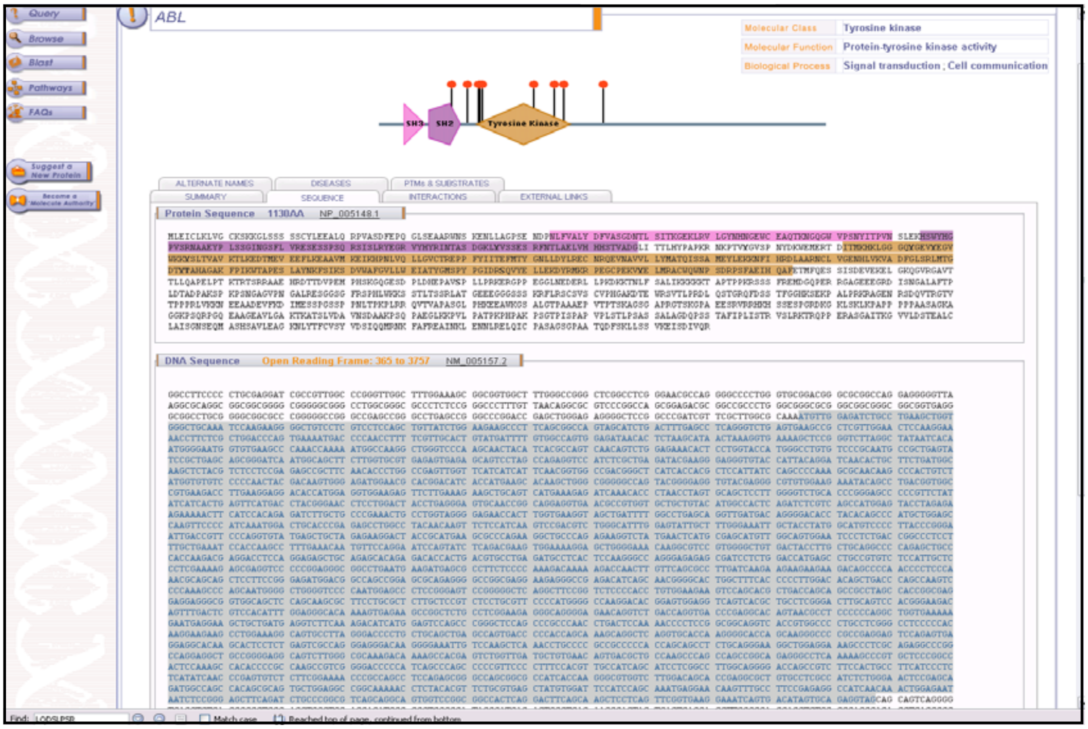


ABL

Molecular Class	Tyrosine kinase
Molecular Function	Protein-tyrosine kinase activity
Biological Process	Signal transduction ; Cell communication

INTERACTIONS

INTERACTING PROTEIN	ABL	
Name Of Interactor	Experiment Type	Type
RAS inhibitor 1	In Vitro	Direct
NCK1 CBL	In Vivo	Complex
CBL	In Vivo; In Vitro	Direct
AAP1	In Vitro; Yeast 2 Hybrid	Direct
Nicastrin	In Vivo; In Vitro; Yeast 2 Hybrid	Direct
SOS2 Guanine nucleotide releasing factor 2 CRKL	In Vivo	Complex
PAK2	In Vivo	Direct
Glutathione peroxidase 1	In Vitro; Yeast 2 Hybrid	Direct
PAG	In Vivo; In Vitro; Yeast 2 Hybrid	Direct
Retinoblastoma 1	In Vitro	Direct
Grb2	In Vitro	Direct
RAN, member RAS oncogene family	In Vivo	Direct
Regulatory factor X 1	In Vivo; In Vitro	Direct
DNA dependent protein kinase catalytic subunit	In Vitro	Direct
IK3-2	In Vivo; In Vitro	Direct
BCR	In Vivo; In Vitro	Direct
EphB2	In Vivo; Yeast 2 Hybrid	Direct
ROS1	Yeast 2 Hybrid	Direct
MAP4K5	In Vivo	Direct
Delta catenin	In Vitro	Direct
SH3 domain binding protein 1	In Vitro	Direct
CBL associated protein	In Vivo; In Vitro	Direct
RNA polymearse II	In Vitro	Direct
Complement component 3	In Vitro	Direct
Janus kinase 1	In Vivo	Direct
ArgBP2	In Vivo; In Vitro	Direct



ABL

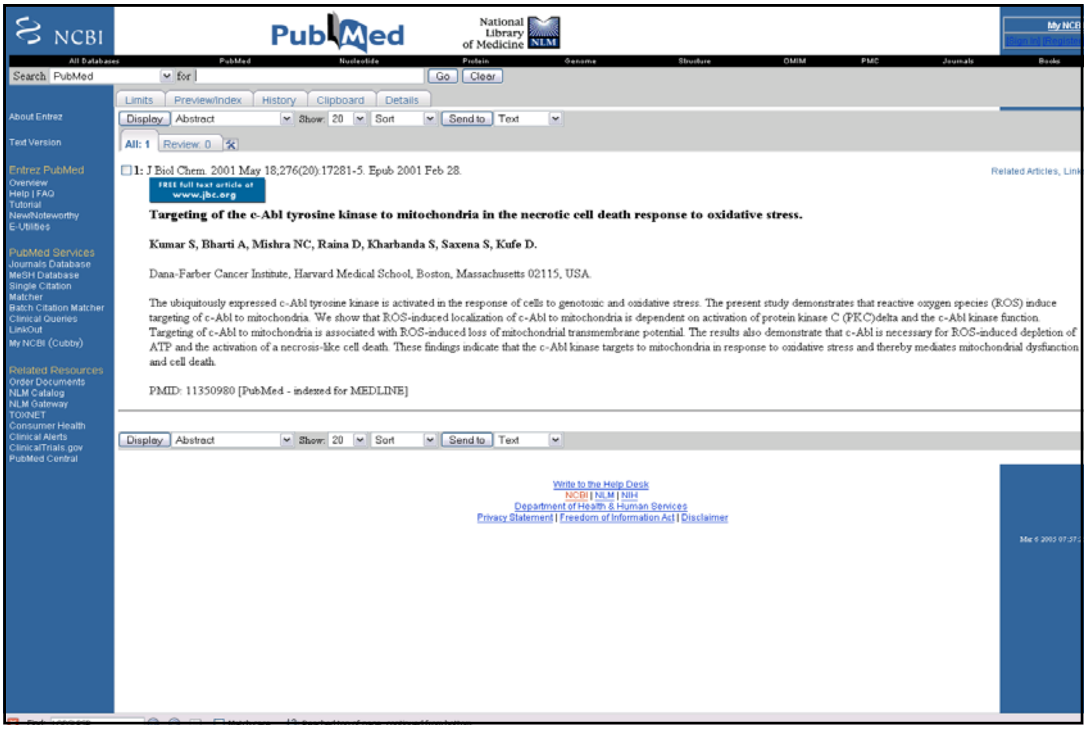
Molecular Class	Tyrosine kinase
Molecular Function	Protein-tyrosine kinase activity
Biological Process	Signal transduction ; Cell communication

Protein Sequence 1130AA NP_005148.1

```
MLEICLKLVG CKSKKGLSSS SSCYLEEALQ RPVASDFEPQ GLSEAARWNS KENLLAGPSE NDPNLFVALY DFVASGDNTL SITKGEKLRV LGYNHNGEWC EAQTKNGQGW VPSNYITPVN SLEKHSWYHG
PVSRNAAEYL LSSGINGSFL VRESESSPGQ RSISLRYEGR VYHYRINTAS DGKLYVSSES RFNTLAELVH HHSTVADGLI TTLHYPAPKR NKPTVYGVSP NYDKWEMERT DITMKHKLGG GQYGEVYEGV
WKKYSLTVAV KTLKEDTMEV EEFLKEAAVM KEIKHPNLVQ LLGVCTREPP FYIITEFMTY GNLLDYLREC NRQEVNAVVL LYMATQISSA MEYLEKKNFI HRDLAARNCL VGENHLVKVA DFGLSRLMTG
DTYTAHAGAK FPIKWTAPES LAYNKFSIKS DVWAFGVLLW EIATYGMSPY PGIDLSQVYE LLEKDYRMER PEGCPEKVYE LMRACWQWNP SDRPSFAEIH QAFETMFQES SISDEVEKEL GKQGVRGAVT
TLLQAPELPT KTRTSRRAAE HRDTTDVPEM PHSKGQGESD PLDHEPAVSP LLPRKERGPP EGGLNEDERL LPKDKKTNLF SALIKKKKKT APTPPKRSSS FREMDGQPER RGAGEEEGRD ISNGALAFTP
LDTADPAKSP KPSNGAGVPN GALRESGGSG FRSPHLWKKS STLTSSRLAT GEEEGGGSSS KRFLRSCSVS CVPHGAKDTE WRSVTLPRDL QSTGRQFDSS TFGGHKSEKP ALPRKRAGEN RSDQVTRGTV
TPPPRLVKKN EEAADEVFKD IMESSPGSSP PNLTPKPLRR QVTVAPASGL PHKEEAWKGS ALGTPAAAEP VTPTSKAGSG APRGTSKGPA EESRVRRHKH SSESPGRDKG KLSKLKPAPP PPPAASAGKA
GGKPSQRPGQ EAAGEAVLGA KTKATSLVDA VNSDAAKPSQ PAEGLKKPVL PATPKPHPAK PSGTPISPAP VPLSTLPSAS SALAGDQPSS TAFIPLISTR VSLRKTRQPP ERASGAITKG VVLDSTEALC
LAISGNSEQM ASHSAVLEAG KNLYTFCVSY VDSIQQMRNK FAFREAINKL ENNLRELQIC PASAGSGPAA TQDFSKLLSS VKEISDIVQR
```

DNA Sequence Open Reading Frame: 365 to 3757 NM_005157.2

```
GGCCTTCCCC CTGCGAGGAT CGCCGTTGGC CCGGGTTGGC TTTGGAAAGC GGCGGTGGCT TTGGGCCGGG CTCGGCCTCG GGAACGCCAG GGGCCCCTGG GTGCGGACGG GCGCGGCCAG GAGGGGGTTA
AGGCGCAGGC GGCGGCGGGG CGGGGGCGGG CCTGGCGGGC GCCCTCTCCG GGCCCTTTGT TAACAGGCGC GTCCCGGCCA GCGGAGACGC GGCCGCCCTG GGCGGGCGCG GGCGGCGGGC GGCGGTGAGG
GCGGCCTGCG GGGCGGCGCC CGGGGGCCGG GCCGAGCCGG GCCTGAGCCG GGCCCGGACC GAGCTGGGAG AGGGGCTCCG GCCCGATCGT TCGCTTGGCG CAAAATGTTG GAGATCTGCC TGAAGCTGGT
GGGCTGCAAA TCCAAGAAGG GGCTGTCCTC GTCCTCCAGC TGTTATCTGG AAGAAGCCCT TCAGCGGCCA GTAGCATCTG ACTTTGAGCC TCAGGGTCTG AGTGAAGCCG CTCGTTGGAA CTCCAAGGAA
AACCTTCTCG CTGGACCCAG TGAAAATGAC CCCAACCTTT TCGTTGCACT GTATGATTTT GTGGCCAGTG GAGATAACAC TCTAAGCATA ACTAAAGGTG AAAAGCTCCG GGTCTTAGGC TATAATCACA
ATGGGGAATG GTGTGAAGCC CAAACCAAAA ATGGCCAAGG CTGGGTCCCA AGCAACTACA TCACGCCAGT CAACAGTCTG GAGAAACACT CCTGGTACCA TGGGCCTGTG TCCCGCAATG CCGCTGAGTA
TCCGCTGAGC AGCGGGATCA ATGGCAGCTT CTTGGTGCGT GAGAGTGAGA GCAGTCCTAG CCAGAGGTCC ATCTCGCTGA GATACGAAGG GAGGGTGTAC CATTACAGGA TCAACACTGC TTCTGATGGC
AAGCTCTACG TCTCCTCCGA GAGCCGCTTC AACACCCTGG CCGAGTTGGT TCATCATCAT TCAACGGTGG CCGACGGGCT CATCACCACG CTCCATTATC CAGCCCCAAA GCGCAACAAG CCCACTGTCT
ATGGTGTGTC CCCCAACTAC GACAAGTGGG AGATGGAACG CACGGACATC ACCATGAAGC ACAAGCTGGG CGGGGGCCAG TACGGGGAGG TGTACGAGGG CGTGTGGAAG AAATACAGCC TGACGGTGGC
CGTGAAGACC TTGAAGGAGG ACACCATGGA GGTGGAAGAG TTCTTGAAAG AAGCTGCAGT CATGAAAGAG ATCAAACACC CTAACCTAGT GCAGCTCCTT GGGGTCTGCA CCCGGGAGCC CCCGTTCTAT
ATCATCACTG AGTTCATGAC CTACGGGAAC CTCCTGGACT ACCTGAGGGA GTGCAACCGG CAGGAGGTGA ACGCCGTGGT GCTGCTGTAC ATGGCCACTC AGATCTCGTC AGCCATGGAG TACCTAGAGA
AGAAAAACTT CATCCACAGA GATCTTGCTG CCCGAAACTG CCTGGTAGGG GAGAACCACT TGGTGAAGGT AGCTGATTTT GGCCTGAGCA GGTTGATGAC AGGGGACACC TACACAGCCC ATGCTGGAGC
CAAGTTCCCC ATCAAATGGA CTGCACCCGA GAGCCTGGCC TACAACAAGT TCTCCATCAA GTCCGACGTC TGGGCATTTG GAGTATTGCT TTGGGAAATT GCTACCTATG GCATGTCCCC TTACCCGGGA
ATTGACCGTT CCCAGGTGTA TGAGCTGCTA GAGAAGGACT ACCGCATGAA GCGCCCAGAA GGCTGCCCAG AGAAGGTCTA TGAACTCATG CGAGCATGTT GGCAGTGGAA TCCCTCTGAC CGGCCCTCCT
TTGCTGAAAT CCACCAAGCC TTTGAAACAA TGTTCCAGGA ATCCAGTATC TCAGACGAAG TGGAAAAGGA GCTGGGGAAA CAAGGCGTCC GTGGGGCTGT GACTACCTTG CTGCAGGCCC CAGAGCTGCC
CACCAAGACG AGGACCTCCA GGAGAGCTGC AGAGCACAGA GACACCACTG ACGTGCCTGA GATGCCTCAC TCCAAGGGCC AGGGAGAGAG CGATCCTCTG GACCATGAGC CTGCCGTGTC TCCATTGCTC
CCTCGAAAAG AGCGAGGTCC CCCGGAGGGC GGCCTGAATG AAGATGAGCG CCTTCTCCCC AAAGACAAAA AGACCAACTT GTTCAGCGCC TTGATCAAGA AGAAGAAGAA GACAGCCCCA ACCCCTCCCA
AACGCAGCAG CTCCTTCCGG GAGATGGACG GCCAGCCGGA GCGCAGAGGG GCCGGCGAGG AAGAGGGCCG AGACATCAGC AACGGGGCAC TGGCTTTCAC CCCCTTGGAC ACAGCTGACC CAGCCAAGTC
CCCAAAGCCC AGCAATGGGG CTGGGGTCCC CAATGGAGCC CTCCGGGAGT CCGGGGGCTC AGGCTTCCGG TCTCCCCACC TGTGGAAGAA GTCCAGCACG CTGACCAGCA GCCGCCTAGC CACCGGCGAG
GAGGAGGGCG GTGGCAGCTC CAGCAAGCGC TTCCTGCGCT CTTGCTCCGT CTCCTGCGTT CCCCATGGGG CCAAGGACAC GGAGTGGAGG TCAGTCACGC TGCCTCGGGA CTTGCAGTCC ACGGGAAGAC
AGTTTGACTC GTCCACATTT GGAGGGCACA AAAGTGAGAA GCCGGCTCTG CCTCGGAAGA GGGCAGGGGA GAACAGGTCT GACCAGGTGA CCCGAGGCAC AGTAACGCCT CCCCCCAGGC TGGTGAAAAA
GAATGAGGAA GCTGCTGATG AGGTCTTCAA AGACATCATG GAGTCCAGCC CGGGCTCCAG CCCGCCCAAC CTGACTCCAA AACCCCTCCG GCGGCAGGTC ACCGTGGCCC CTGCCTCGGG CCTCCCCCAC
AAGGAAGAAG CCTGGAAAGG CAGTGCCTTA GGGACCCCTG CTGCAGCTGA GCCAGTGACC CCCACCAGCA AAGCAGGCTC AGGTGCACCA AGGGGCACCA GCAAGGGCCC CGCCGAGGAG TCCAGAGTGA
GGAGGCACAA GCACTCCTCT GAGTCGCCAG GGAGGGACAA GGGGAAATTG TCCAAGCTCA AACCTGCCCC GCCGCCCCCA CCAGCAGCCT CTGCAGGGAA GGCTGGAGGA AAGCCCTCGC AGAGGCCCGG
CCAGGAGGCT GCCGGGGAGG CAGTCTTGGG CGCAAAGACA AAAGCCACGA GTCTGGTTGA TGCTGTGAAC AGTGACGCTG CCAAGCCCAG CCAGCCGGCA GAGGGCCTCA AAAAGCCCGT GCTCCCGGCC
ACTCCAAAGC CACACCCCGC CAAGCCGTCG GGGACCCCCA TCAGCCCAGC CCCCGTTCCC CTTTCCACGT TGCCATCAGC ATCCTCGGCC TTGGCAGGGG ACCAGCCGTC TTCCACTGCC TTCATCCCTC
TCATATCAAC CCGAGTGTCT CTTCGGAAAA CCCGCCAGCC TCCAGAGCGG GCCAGCGGCG CCATCACCAA GGGCGTGGTC TTGGACAGCA CCGAGGCGCT GTGCCTCGCC ATCTCTGGGA ACTCCGAGCA
GATGGCCAGC CACAGCGCAG TGCTGGAGGC CGGCAAAAAC CTCTACACGT TCTGCGTGAG CTATGTGGAT TCCATCCAGC AAATGAGGAA CAAGTTTGCC TTCCGAGAGG CCATCAACAA ACTGGAGAAT
AATCTCCGGG AGCTTCAGAT CTGCCCGGCG TCAGCAGGCA GTGGTCCGGC GGCCACTCAG GACTTCAGCA AGCTCCTCAG TTCGGTGAAG GAAATCAGTG ACATAGTGCA GAGGTAGCAG CAGTCAGGGG
```



☐ 1: J Biol Chem. 2001 May 18;276(20):17281-5. Epub 2001 Feb 28.

Targeting of the c-Abl tyrosine kinase to mitochondria in the necrotic cell death response to oxidative stress.

Kumar S, Bharti A, Mishra NC, Raina D, Kharbanda S, Saxena S, Kufe D.

Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts 02115, USA.

The ubiquitously expressed c-Abl tyrosine kinase is activated in the response of cells to genotoxic and oxidative stress. The present study demonstrates that reactive oxygen species (ROS) induce targeting of c-Abl to mitochondria. We show that ROS-induced localization of c-Abl to mitochondria is dependent on activation of protein kinase C (PKC)delta and the c-Abl kinase function. Targeting of c-Abl to mitochondria is associated with ROS-induced loss of mitochondrial transmembrane potential. The results also demonstrate that c-Abl is necessary for ROS-induced depletion of ATP and the activation of a necrosis-like cell death. These findings indicate that the c-Abl kinase targets to mitochondria in response to oxidative stress and thereby mediates mitochondrial dysfunction and cell death.

PMID: 11350980 [PubMed - indexed for MEDLINE]

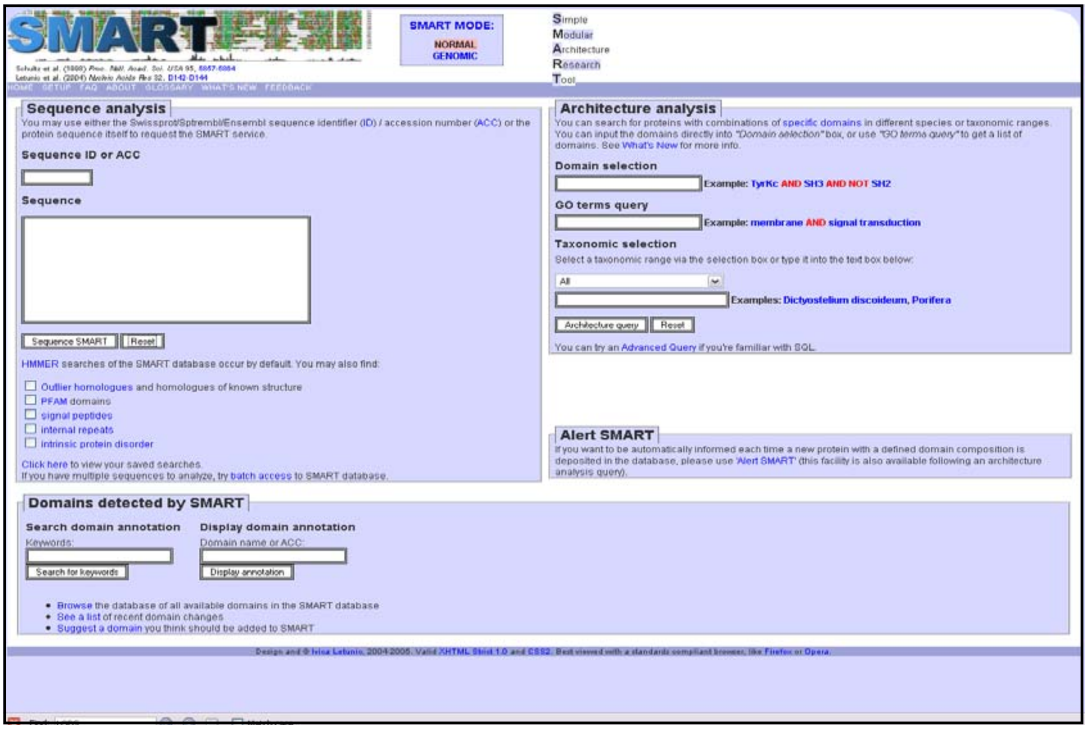
Status of HPRD

- Over 18,000 proteins annotated
- All 1,864 human disease genes in OMIM annotated
- Over 170,000 PubMed links provided (derived from reading of over 2,000,000 full-text articles)
- 25 types of post-translational modifications (PTM) annotated from literature
- 8,000 PTM sites annotated
- Over 24,000 binary interactions annotated
- Compatible with Gene Ontology, PSI-MI, Cytoscape

Protein Information resources

Tools

- SMART (http://smart.embl-heidelberg.de/)
- Pfam (http://www.sanger.ac.uk/Software/Pfam/)
- PSORT (http://psort.nibb.ac.jp/)



Schultz et al. (1998) Proc. Natl. Acad. Sci. USA 95, 5857-5864
Letunic et al. (2004) Nucleic Acids Res 32, D142-D144

Sequence analysis

You may use either the Swissprot/Sptrembl/Ensembl sequence identifier (ID) / accession number (ACC) or the protein sequence itself to request the SMART service.

Sequence ID or ACC

Sequence

Sequence SMART | Reset

HMMER searches of the SMART database occur by default. You may also find:

- □ Outlier homologues and homologues of known structure
- □ PFAM domains
- □ signal peptides
- □ internal repeats
- □ intrinsic protein disorder

Click here to view your saved searches.
If you have multiple sequences to analyze, try batch access to SMART database.

Architecture analysis

You can search for proteins with combinations of specific domains in different species or taxonomic ranges. You can input the domains directly into *"Domain selection"* box, or use *"GO terms query"* to get a list of domains. See What's New for more info.

Domain selection

Example: **TyrKc AND SH3 AND NOT SH2**

GO terms query

Example: **membrane AND signal transduction**

Taxonomic selection

Select a taxonomic range via the selection box or type it into the text box below:

All

Examples: **Dictyostelium discoideum, Porifera**

Architecture query | Reset

You can try an Advanced Query if you're familiar with SQL.

Alert SMART

If you want to be automatically informed each time a new protein with a defined domain composition is deposited in the database, please use 'Alert SMART' (this facility is also available following an architecture analysis query).

Domains detected by SMART

Search domain annotation

Keywords:

Search for keywords

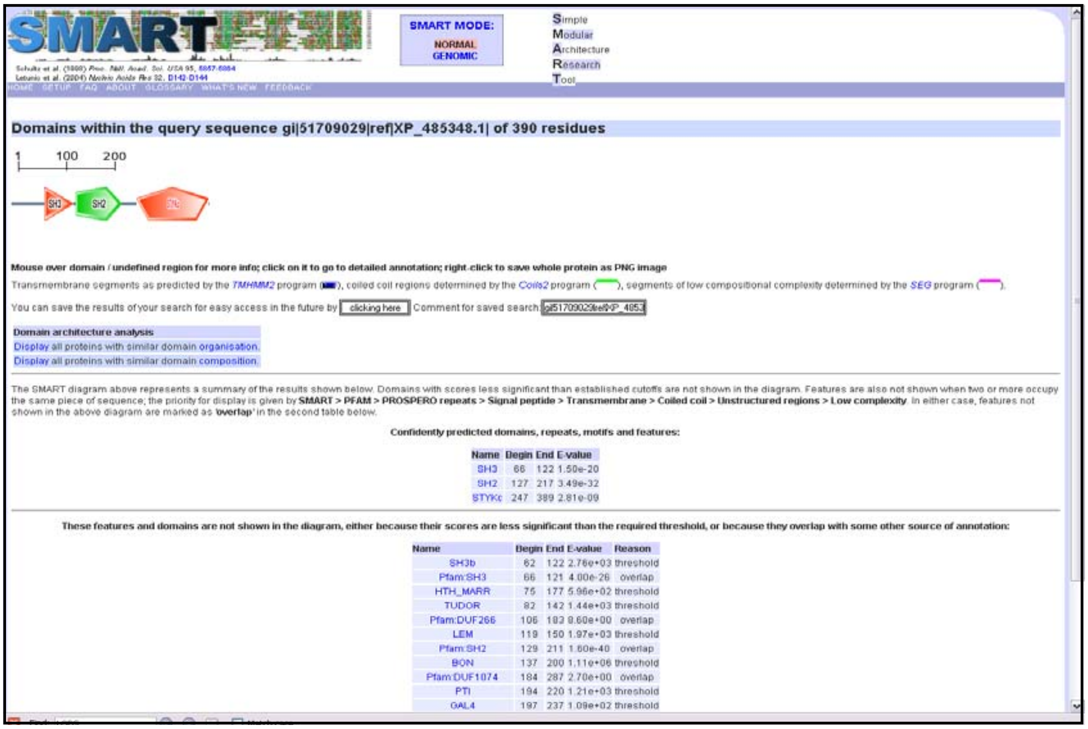
Display domain annotation

Domain name or ACC:

Display annotation

- Browse the database of all available domains in the SMART database
- See a list of recent domain changes
- Suggest a domain you think should be added to SMART

Design and © Ivica Letunic, 2004-2005. Valid XHTML Strict 1.0 and CSS2. Best viewed with a standards compliant browser, like Firefox or Opera.



Schultz et al. (1998) Proc. Natl. Acad. Sci. USA 95, 5857-5864
Letunic et al. (2004) Nucleic Acids Res 32, D142-D144

Domains within the query sequence gi|51709029|ref|XP_485348.1| of 390 residues

Mouse over domain / undefined region for more info; click on it to go to detailed annotation; right-click to save whole protein as PNG image

Transmembrane segments as predicted by the *TMHMM2* program (▬), coiled coil regions determined by the *Coils2* program (▬), segments of low compositional complexity determined by the *SEG* program (▬).

You can save the results of your search for easy access in the future by clicking here Comment for saved search: gi|51709029|ref|XP_4853

Domain architecture analysis
Display all proteins with similar domain organisation.
Display all proteins with similar domain composition.

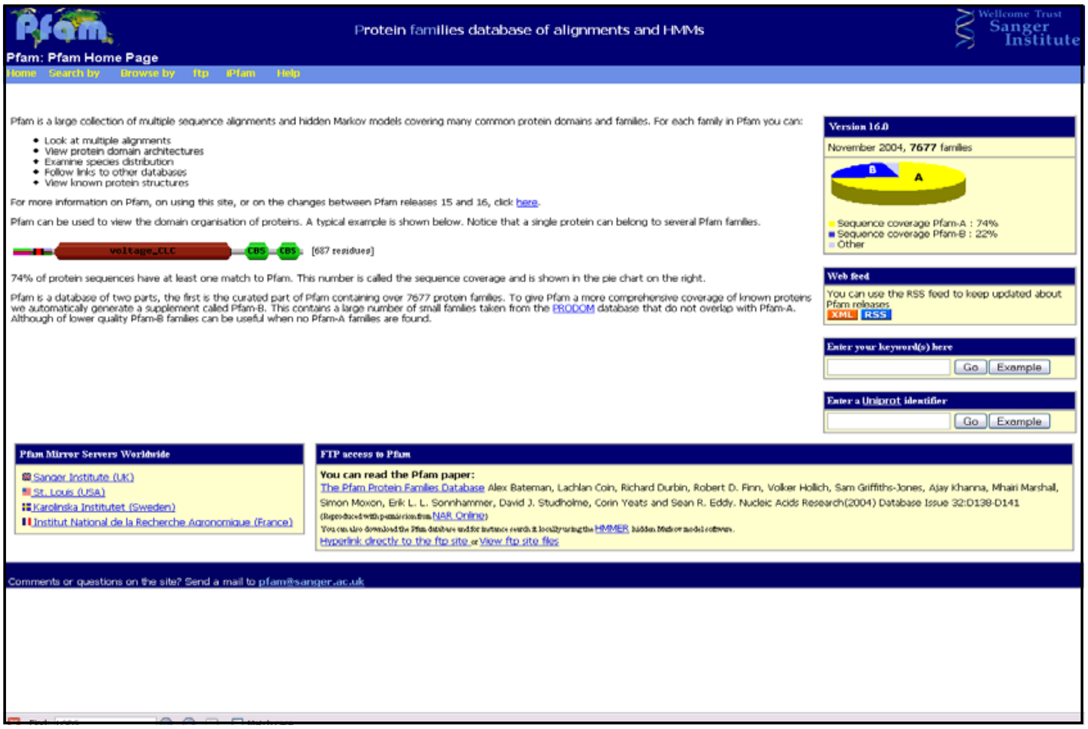
The SMART diagram above represents a summary of the results shown below. Domains with scores less significant than established cutoffs are not shown in the diagram. Features are also not shown when two or more occupy the same piece of sequence; the priority for display is given by **SMART > PFAM > PROSPERO repeats > Signal peptide > Transmembrane > Coiled coil > Unstructured regions > Low complexity**. In either case, features not shown in the above diagram are marked as **'overlap'** in the second table below.

Confidently predicted domains, repeats, motifs and features:

Name	Begin	End	E-value
SH3	66	122	1.50e-20
SH2	127	217	3.49e-32
STYKc	247	389	2.81e-09

These features and domains are not shown in the diagram, either because their scores are less significant than the required threshold, or because they overlap with some other source of annotation:

Name	Begin	End	E-value	Reason
SH3b	62	122	2.76e+03	threshold
Pfam:SH3	66	121	4.00e-26	overlap
HTH_MARR	75	177	5.96e+02	threshold
TUDOR	82	142	1.44e+03	threshold
Pfam:DUF266	106	183	8.60e+00	overlap
LEM	119	150	1.97e+03	threshold
Pfam:SH2	129	211	1.60e-40	overlap
BON	137	200	1.11e+06	threshold
Pfam:DUF1074	184	287	2.70e+00	overlap
PTI	194	220	1.21e+03	threshold
GAL4	197	237	1.09e+02	threshold



Pfam
Protein families database of alignments and HMMs
Wellcome Trust Sanger Institute
Pfam: Pfam Home Page
Home Search by Browse by ftp iPfam Help

Pfam is a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains and families. For each family in Pfam you can:

- Look at multiple alignments
- View protein domain architectures
- Examine species distribution
- Follow links to other databases
- View known protein structures

For more information on Pfam, on using this site, or on the changes between Pfam releases 15 and 16, click here.

Pfam can be used to view the domain organisation of proteins. A typical example is shown below. Notice that a single protein can belong to several Pfam families.

74% of protein sequences have at least one match to Pfam. This number is called the sequence coverage and is shown in the pie chart on the right.

Pfam is a database of two parts, the first is the curated part of Pfam containing over 7677 protein families. To give Pfam a more comprehensive coverage of known proteins we automatically generate a supplement called Pfam-B. This contains a large number of small families taken from the PRODOM database that do not overlap with Pfam-A. Although of lower quality Pfam-B families can be useful when no Pfam-A families are found.

Version 16.0

November 2004, **7677** families

- 🟨 Sequence coverage Pfam-A : 74%
- 🟦 Sequence coverage Pfam-B : 22%
- ▫ Other

Web feed

You can use the RSS feed to keep updated about Pfam releases

Enter your keyword(s) here

Enter a Uniprot identifier

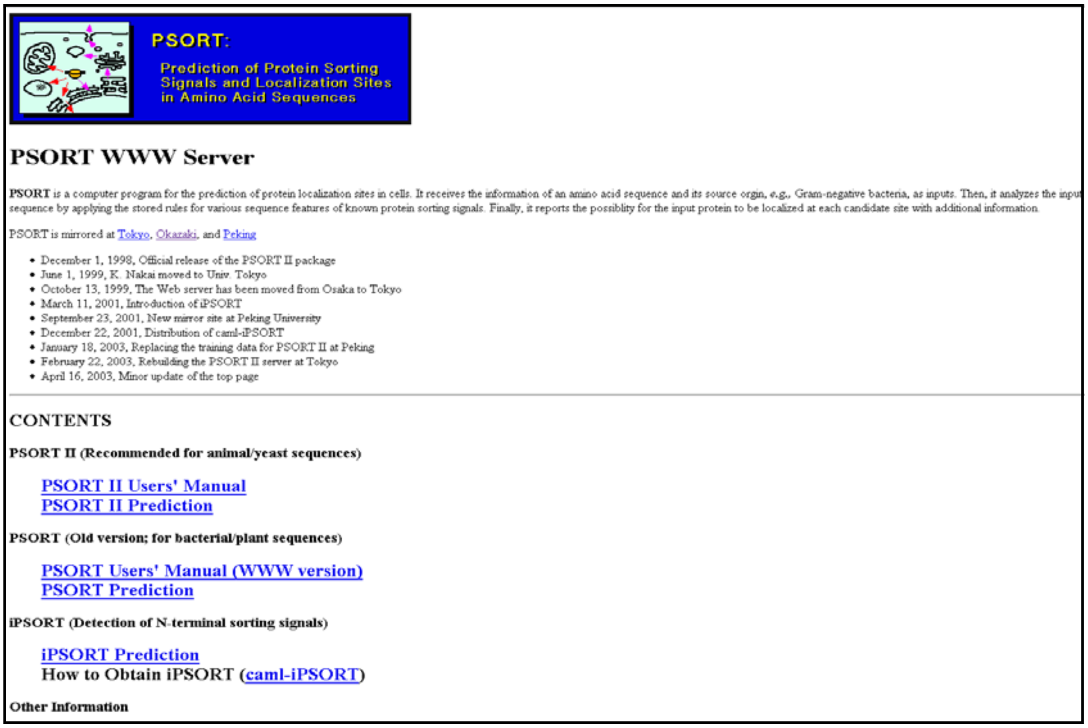
Pfam Mirror Servers Worldwide

- Sanger Institute (UK)
- St. Louis (USA)
- Karolinska Institutet (Sweden)
- Institut National de la Recherche Agronomique (France)

FTP access to Pfam

You can read the Pfam paper:
The Pfam Protein Families Database Alex Bateman, Lachlan Coin, Richard Durbin, Robert D. Finn, Volker Hollich, Sam Griffiths-Jones, Ajay Khanna, Mhairi Marshall, Simon Moxon, Erik L. L. Sonnhammer, David J. Studholme, Corin Yeats and Sean R. Eddy. Nucleic Acids Research(2004) Database Issue 32:D138-D141
(Reproduced with permission from NAR Online)
You can also download the Pfam database and/or instance search it locally using the HMMER hidden Markov model software.
Hyperlink directly to the ftp site or View ftp site files

Comments or questions on the site? Send a mail to pfam@sanger.ac.uk



PSORT:
Prediction of Protein Sorting Signals and Localization Sites in Amino Acid Sequences

PSORT WWW Server

PSORT is a computer program for the prediction of protein localization sites in cells. It receives the information of an amino acid sequence and its source orgin, *e.g.*, Gram-negative bacteria, as inputs. Then, it analyzes the input sequence by applying the stored rules for various sequence features of known protein sorting signals. Finally, it reports the possiblity for the input protein to be localized at each candidate site with additional information.

PSORT is mirrored at Tokyo, Okazaki, and Peking

- December 1, 1998, Official release of the PSORT II package
- June 1, 1999, K. Nakai moved to Univ. Tokyo
- October 13, 1999, The Web server has been moved from Osaka to Tokyo
- March 11, 2001, Introduction of iPSORT
- September 23, 2001, New mirror site at Peking University
- December 22, 2001, Distribution of caml-iPSORT
- January 18, 2003, Replacing the training data for PSORT II at Peking
- February 22, 2003, Rebuilding the PSORT II server at Tokyo
- April 16, 2003, Minor update of the top page

CONTENTS

PSORT II (Recommended for animal/yeast sequences)

PSORT II Users' Manual
PSORT II Prediction

PSORT (Old version; for bacterial/plant sequences)

PSORT Users' Manual (WWW version)
PSORT Prediction

iPSORT (Detection of N-terminal sorting signals)

iPSORT Prediction
How to Obtain iPSORT (caml-iPSORT)

Other Information