EBI Patent Sequence Services

- PIUG 2012 Biotechnology
- Workshop
- February 6th, 2012
- Boston

Jennifer McDowall

EMBL-EBI

EBI is an Outstation of the European Molecular Biology Laboratory.

Overview

1) Know the data

- **European Nucleotide Archive**
- **UniProt**
- **Non-redundant patent sequence DB**

2) The Toolbox

- **EBI search**
- **SRS advanced text search**
- **Sequence searching**

1) Know the Data

Know the data

- Many databases, each getting bigger
- Efficient searching requires knowledge of what data is stored in a database
  - Don’t assume annotation can be transferred because of a good match
- Databases can contain errors
- Data can change
  - Deletions, sequence modifications
  - Daily updates, identifier changes…

Major sequence databases

European Nucleotide Archive

- >170 million sequences
  - (~42 million non-redundant)
- release every 3 months, daily updates

UniProt

- >30.1 million non-redundant sequences
- monthly release, daily updates

Additional sequence data

Specialized databases

- Immunoglobulins: IMGT/HLA, IMGT/LIGM
- Immunopolymorphisms : IPD-KIR , IPD-MHC
- Variation: HGVBase, dbSNP
- Alternative splicing: ASTD
- Completed genomes: Ensembl, Integr8
- Structure: PDB, Structural Genomics targets
- Interaction : IntAct

Patent Sequences

Patent sequences can be found in the following databases:

ENA

- Patent nucleotides

UniProt Archive

- Patent proteins

NR patent sequences

- Patent nucleotides and proteins

Which database do you use?

let's take a look…

European nucleotide archive

UniProt

Non-redundant patent sequence databases

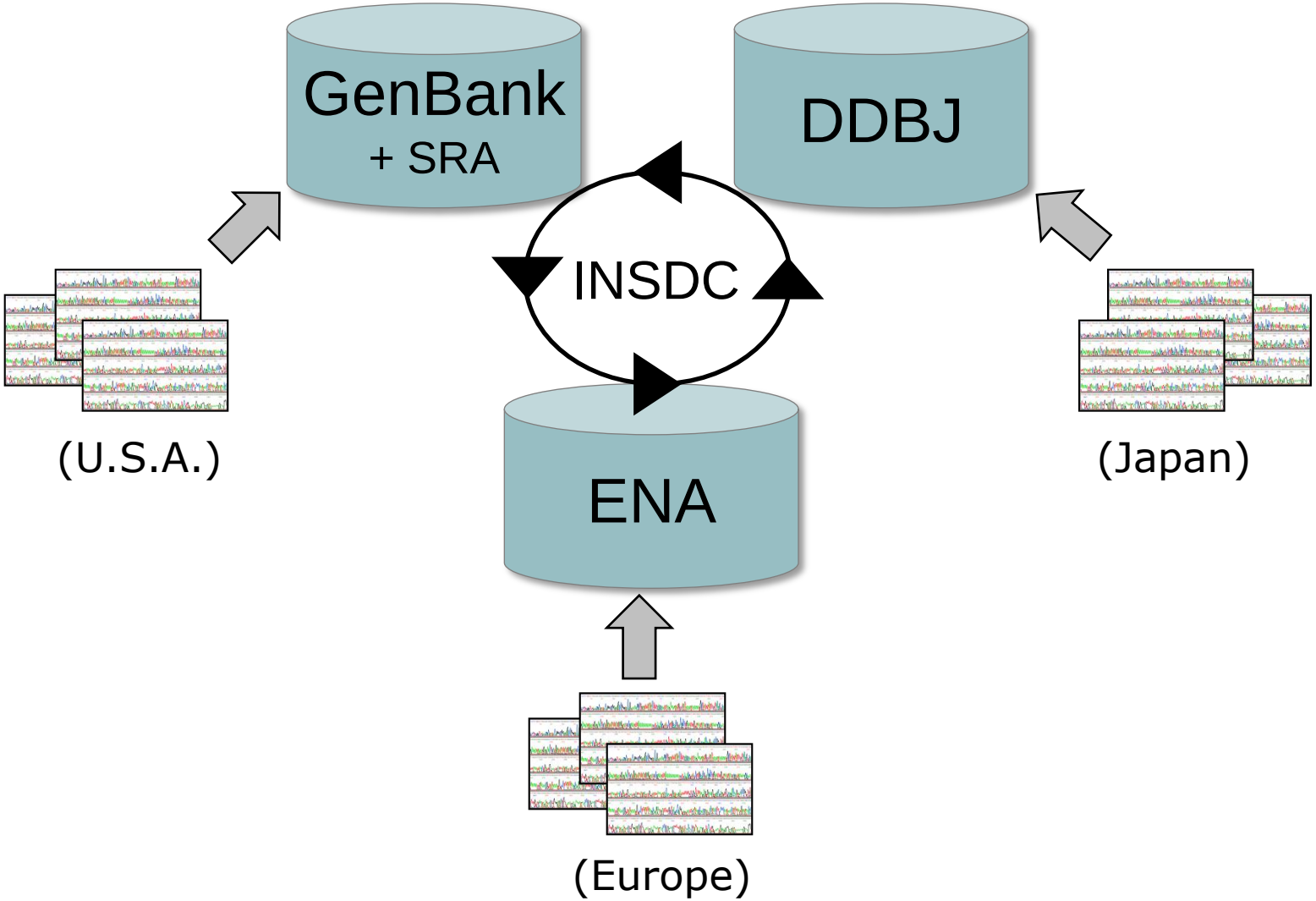
European nucleotide archive

UniProt

Non-redundant patent sequence databases

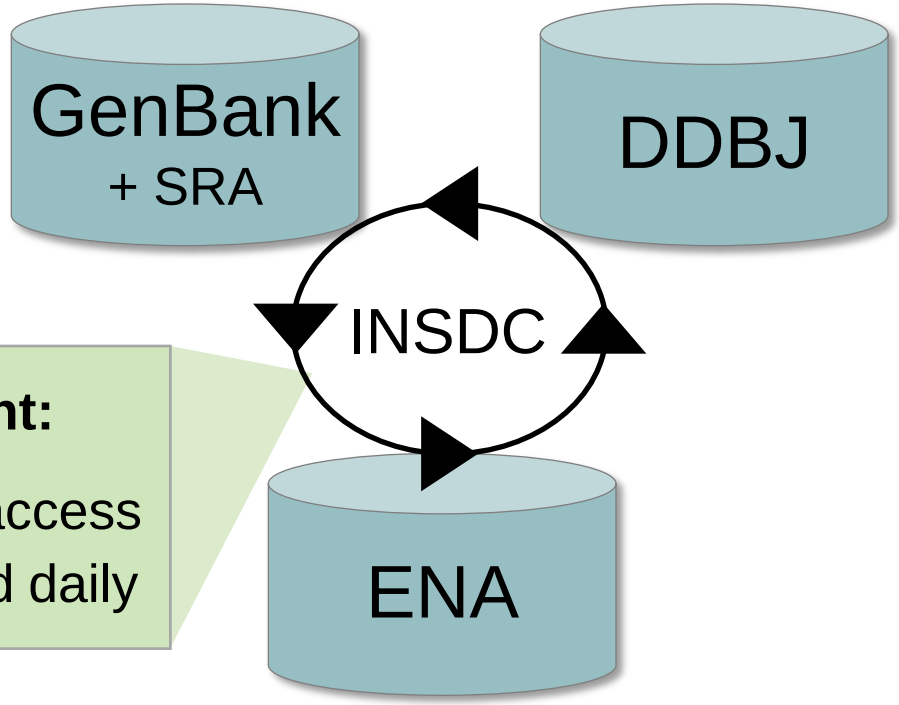
Primary sequence databases

Primary data submitted to databases



Primary sequence databases

Primary data submitted to databases



INSDC agreement:

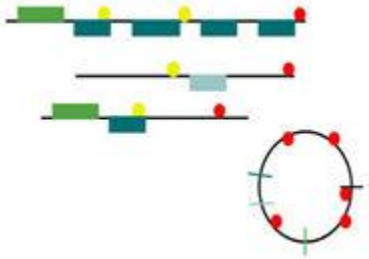
- Free unrestricted access
- All data exchanged daily

How do they differ?

- organization of data
- tools and database links

ENA has a 3-tiered structure

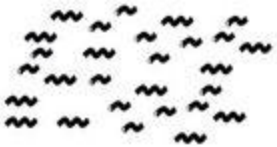
Feature annotation



Assembly information

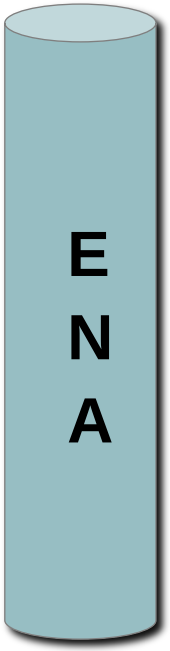
1) **EMBL-Bank**

Sequencing & sampling information



2) **Sequence Read Archive**
(Next Gen sequencing)

3) **Trace Archive**
(Capillary sequencing)



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

How is the data organised?

Data in EMBL-Bank is divided in 2 ways:

1) Data classes

- Type of data <u>or</u> methodology used to obtain data
- Each entry belongs to one data class

2) Taxonomic Divisions

- Each entry belongs to one taxonomic division

EMBL-Bank data classes

CON	Constructed from sequence assemblies
EST	Expressed Sequence Tag (cDNA)
GSS	Genome Survey Sequence (high-throughput short sequence)
HTC	High-Throughput cDNA (unfinished)
HTG	High-Throughput Genome sequencing (unfinished)
MGA	Mass Genome Annotation
PAT	Patent sequences
SRA	Sequence Read Archive (both databank and data class)
STS	Sequence Tagged Site (short unique genomic sequences)
STD	Standard (high quality annotated sequence)
TSA	Transcriptome Shotgun Assembly (computational assembly)
WGS	Whole Genome Shotgun

EMBL-Bank data classes

Data is always changing

- **Assembly** of sequences into larger fragments
- Suppression of **obsolete** entries (i.e. once assembled)
- Sequence **modifications**
- Daily **updates**
- Identifier **changes**
- **Corrections** (databases can contain errors)
- etc…

EMBL-Bank data classes

Data assembly can affect entries

Example:

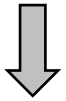
WGS **Shotgun**

- Fragments in separate entries

- Join to make new CON entries

CON **Constructed**

Old WGS entries archived



- Join into large STD entry (e.g. completed genome)
- Add annotation

STD **Standard**

Old CON entries archived

ENA taxonomy

All INSDC databases use NCBI taxonomy

Only **sequenced** organisms represents

Divisions

HUM	Human
MUS	Mouse
ROD	Rodent
MAM	Mammal
VRT	Vertebrate
FUN	Fungi
INV	Invertebrate
PLN	Plant
PRO	Prokaryote
PHG	Phage
VIR	Viral

Other:

ENV	Environmental
SYN	Synthetic
TGN	Transgenic
UNC	Unclassified

ENA taxonomy

Some species EXCLUDED from certain taxonomic ranges

ROD Rodent → **excludes mouse**

MAM Mammal → **excludes**
- **human**
- **mouse**
- **rodent**

VRT Vertebrate → **excludes**
- **human**
- **mouse**
- **rodent**
- **mammal**

Applies to **ftp files** and **sequence search tools** but not to ENA browser

ENA taxonomy

Sometimes there is no taxonomic data

Environmental

- *Genus species* = ‘uncultivated bacterium’ or ‘unspecified’

Synthetic

- *Genus species* = ‘synthetic construct’

Transgenic

- Taxonomy for recipient and donor organisms

Patent

- Exempt from requiring *Genus species*

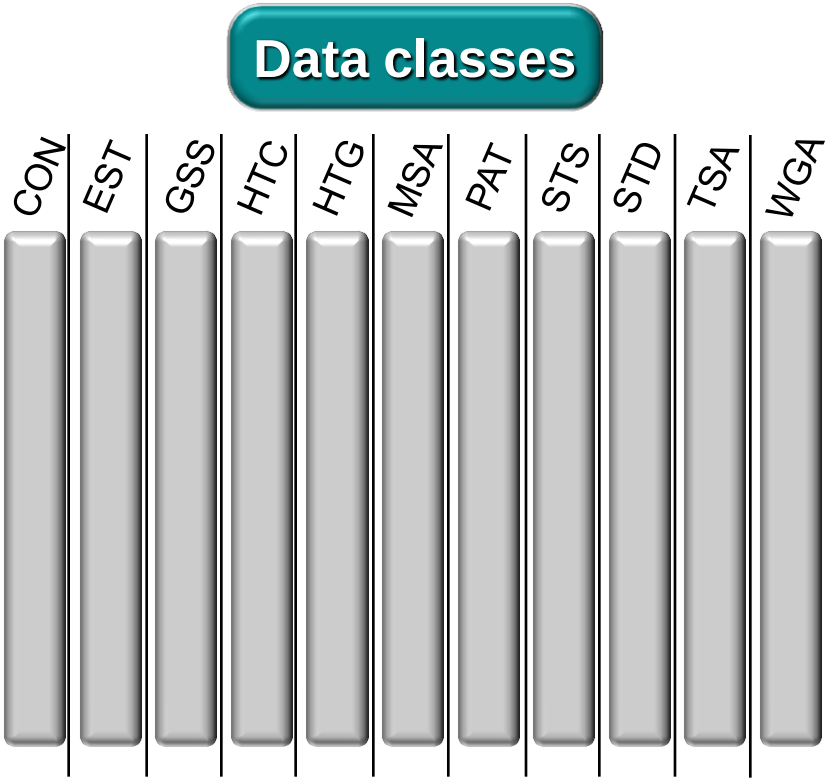
Database structure

EMBL-Bank:

ENA Database

Database structure

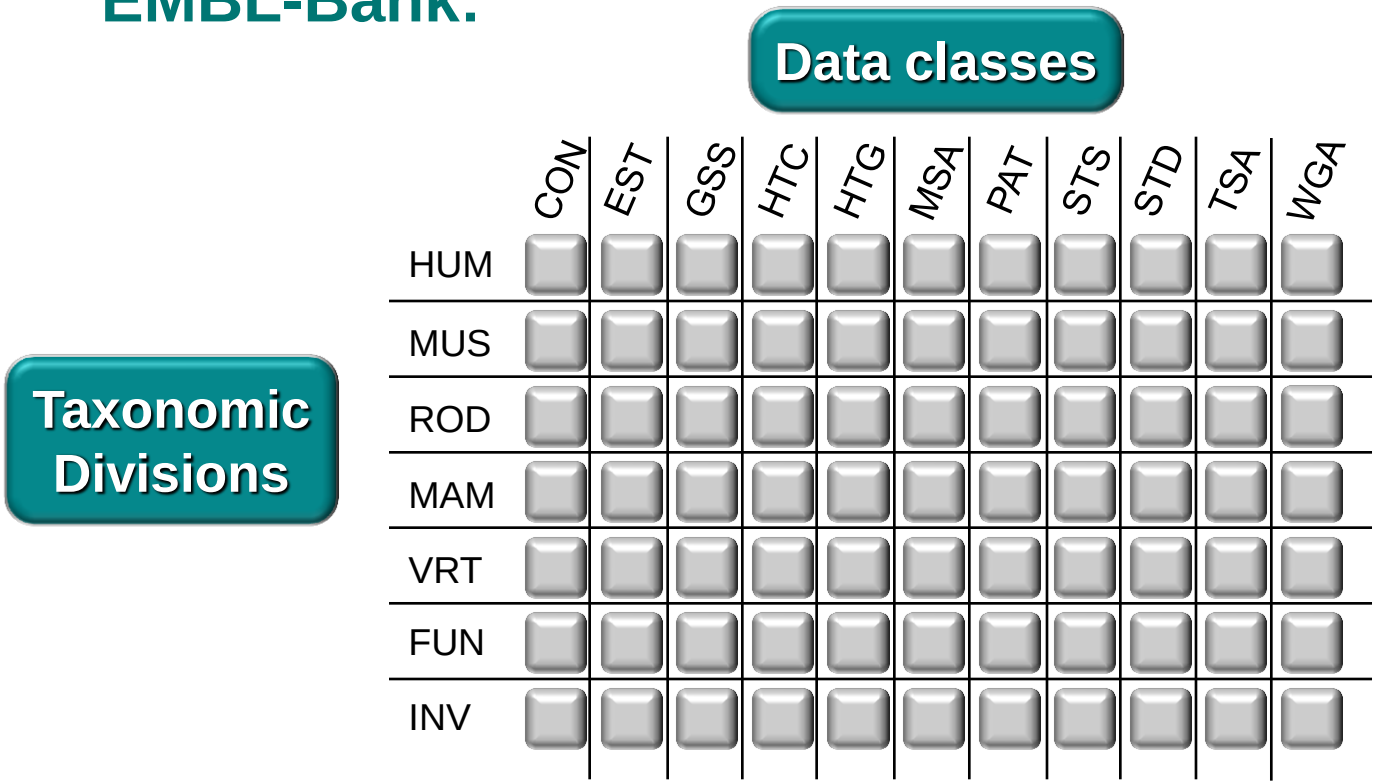
EMBL-Bank:



1st: Data split into classes

Database structure

EMBL-Bank:



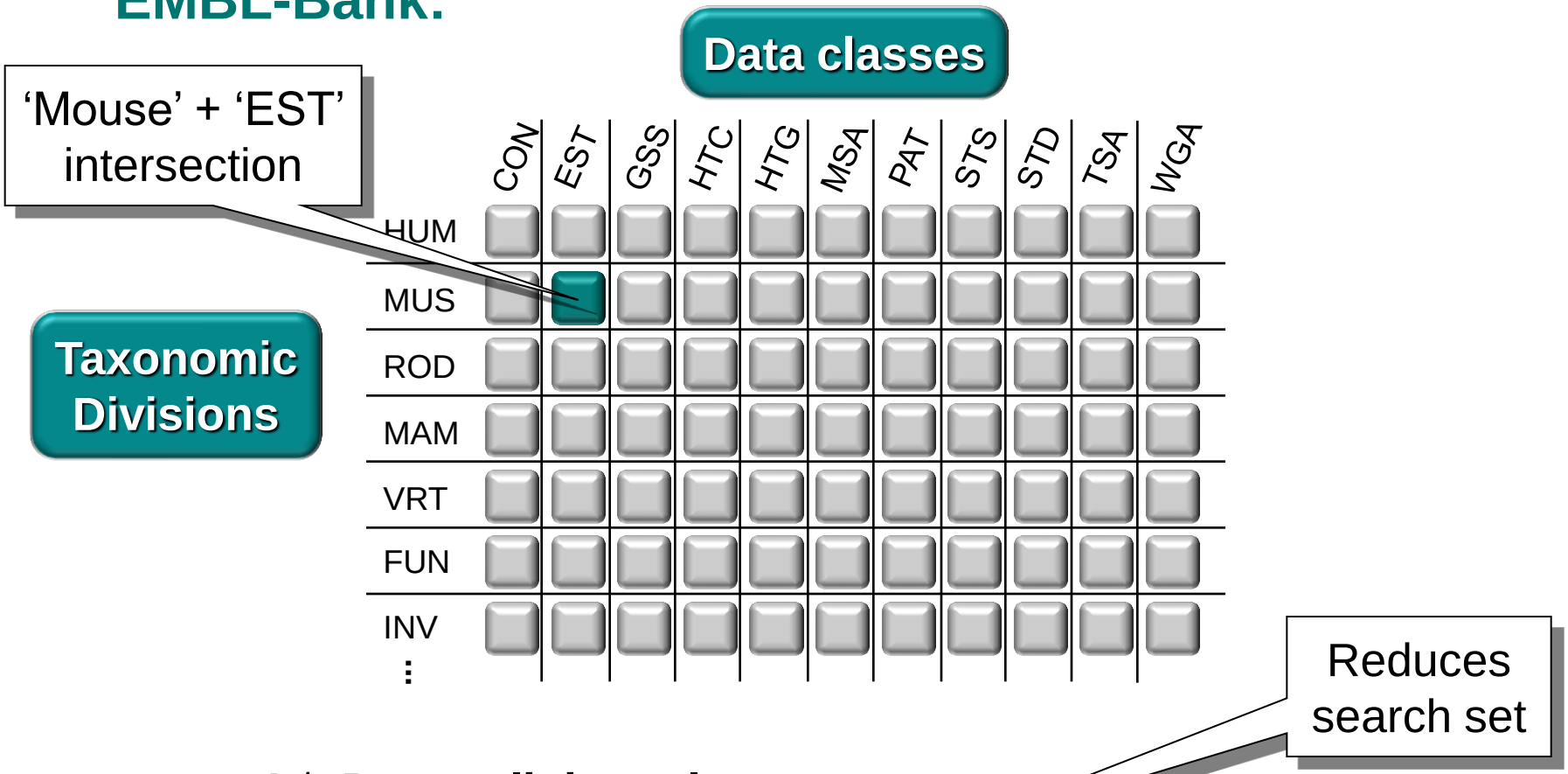
Reduces search set

1st: Data split into classes

2nd: Data split into <u>intersecting</u> slices by taxonomy

Database structure

EMBL-Bank:

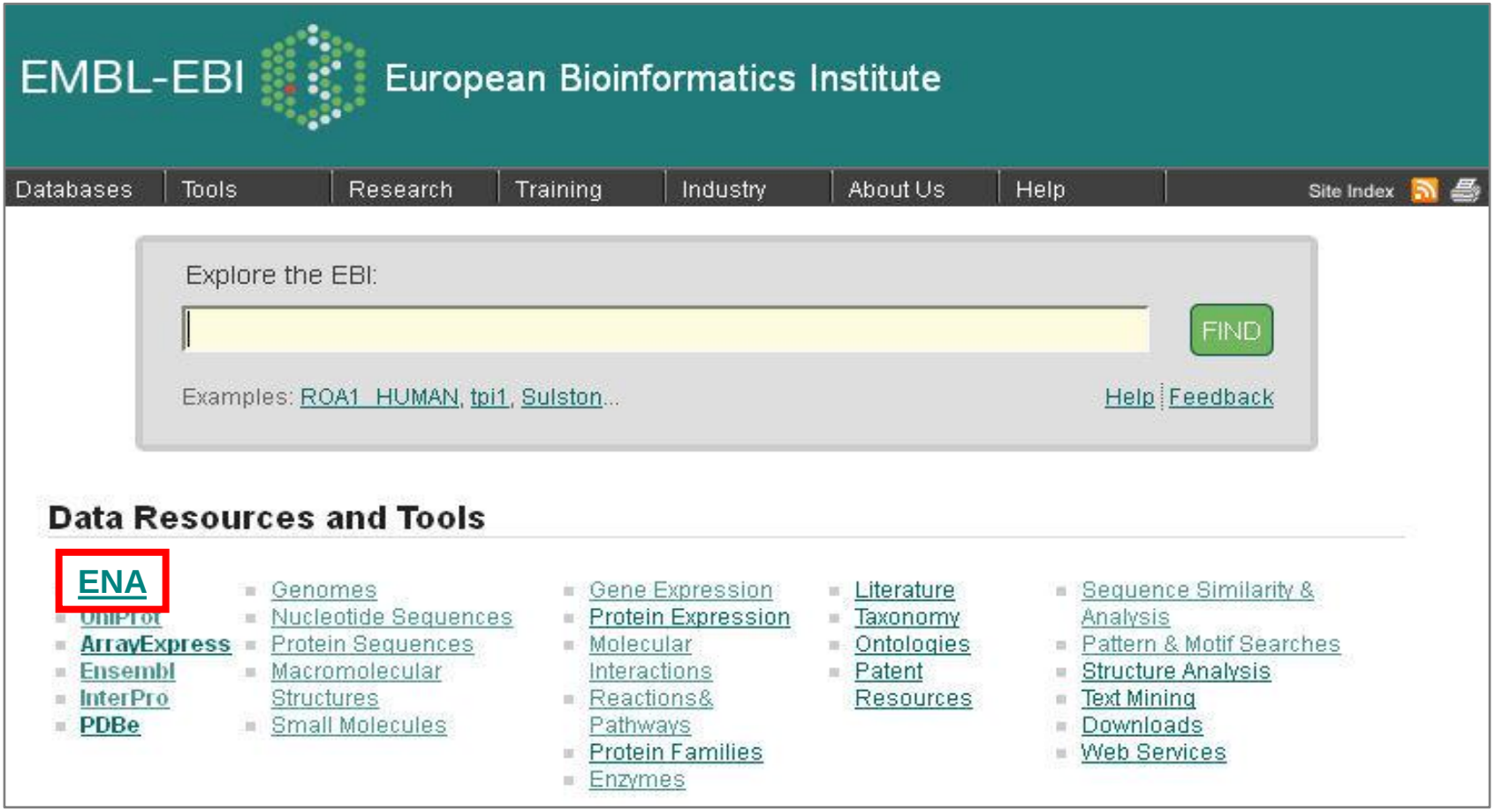


1st: Data split into classes

2nd: Data split into <u>intersecting</u> slices by taxonomy

European Nucleotide Archive

ENA is accessible from the EBI homepage



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

ENA homepage

EMBL-EBI

Find Help

Databases | Tools | Research | Training | Industry | About Us | Help | Site Index

ENA

EBI Home » ENA Home ›

- ENA Home
- Search & Browse
- Submit & Update
- About ENA
- Contact

NEWS AND ANNOUNCEMENTS

CRAM toolkit 0.3 released
3 Aug 2011

CRAM toolkit version 0.3 has been released with support for read quality masking (selective preservation of base quality scores). More information with download, installation and usage instructions are available here.

ArchiveBAM 1.0 specification
16 Mar 2011

The ArchiveBAM 1.0

European Nucleotide Archive

The European Nucleotide Archive (ENA) provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation ... more

Access to ENA data is provided though the browser, through search tools, large scale file download and through the API.

ENA European Nucleotide Archive

Text search

Enter search query, for example: BN000065 | Search

Sequence Search Advanced Search

Enter or paste a nucleotide sequence or accession

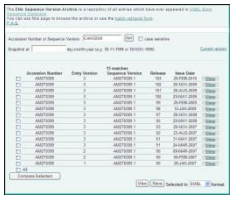
- Text search
- Sequence search
- Programmatic access

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

Patent sequence record in EMBL-Bank

Sequence version

Navigate to related data **e.g. Version archive**



Patent reference

Sequence

EMBL-Bank : A39349 .1 : Sequence 15 from Patent WO9413821.

View: TEXT XML FASTA Download: TEXT XML FASTA

Overview | Source Feature | Other Features | References | Comments | Sequence | Send feedback

Molecule type	Topology	Data class	Sequence length	Sequence Version	First public	Last updated
unassigned DNA	linear	PAT	300	1	1997-03-05(Release 51)	2006-11-14(Release 89)

Description: Sequence 15 from Patent WO9413821.

Navigation

Taxon :	4911 : Kluyveromyces marxianus
SVA:	Historical versions of this EMBL-Bank record are available here: A39349
CABRI:	CBS 6556
StrainInfo:	455808

Overview — Top

Visible feature range: 1 – 300 Apply

Overview: Forward strand 300 bp — A39349.1

Features: Forward strand 300 bp; 1 bp – 300 bp

Source: Kluyveromyces marxianus

Genes: INU1

CDS: INU1

Source Feature(s) — Top

source	1..300
Taxon:	4911 : Kluyveromyces marxianus
strain	CBS 6556
clone	PUR2415

Other Features — Top

Visible feature range: 1 – 300 Filter: Default / Main Apply

CDS	8..295
partial	
gene	INU1
product	INULINASE
protein_id	CAA02439
translation	KAITGTFSLNRPSVHFTPSHGWMNDPNGLWYDAKEEDWHLYYQYNPAATIWGTPLYWGHAVSKDLTSWTDYGASLGPGS DDAGAFSGSMVIDYNN
EMBL-CDS	CAA02439

References — Top

[1] THE USE OF THE KLUYVEROMYCES MARXIANUS INULINASE GENE PROMOTER FOR PROTEIN PRODUCTION
CHAPMAN John William, MUSTERS Wouter, ROUWENHORST Robert Jan, TOSCHKA Holger York, VERBAKEL Johannes Maria A
QUEST INTERNATIONAL B.V.
(PubMed:WO9413821) abstract
CiteXplore

Comments — Top

Other publication AU 5811194 940704

Sequence — Top

Visible sequence range: 1 – 1000 Apply

```
>ENA|A39349|A39349.1 Sequence 15 from Patent WO9413821. : Location:1..10
00
ggtacccaaggctatcaccggcaccactttcagtttgaacagaccttctgtgcatttcac
tccatcccatggttggatgaacgatccaaatggtttgtggtacgatgccaaggaagaaga
ctggcatttgtactaccagtacaacccagcagccacgatctggggtactccattgtactg
gggtcacgctgtttccaaggatttgacttcttggacagattacggtgcttccttgggccc
aggttccgacgacgctggtgcgttcagtggtagtatggtcatagactacaacaacacggg
```

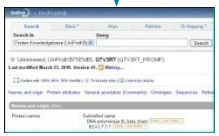
Download data

Dates (first public and last updated)

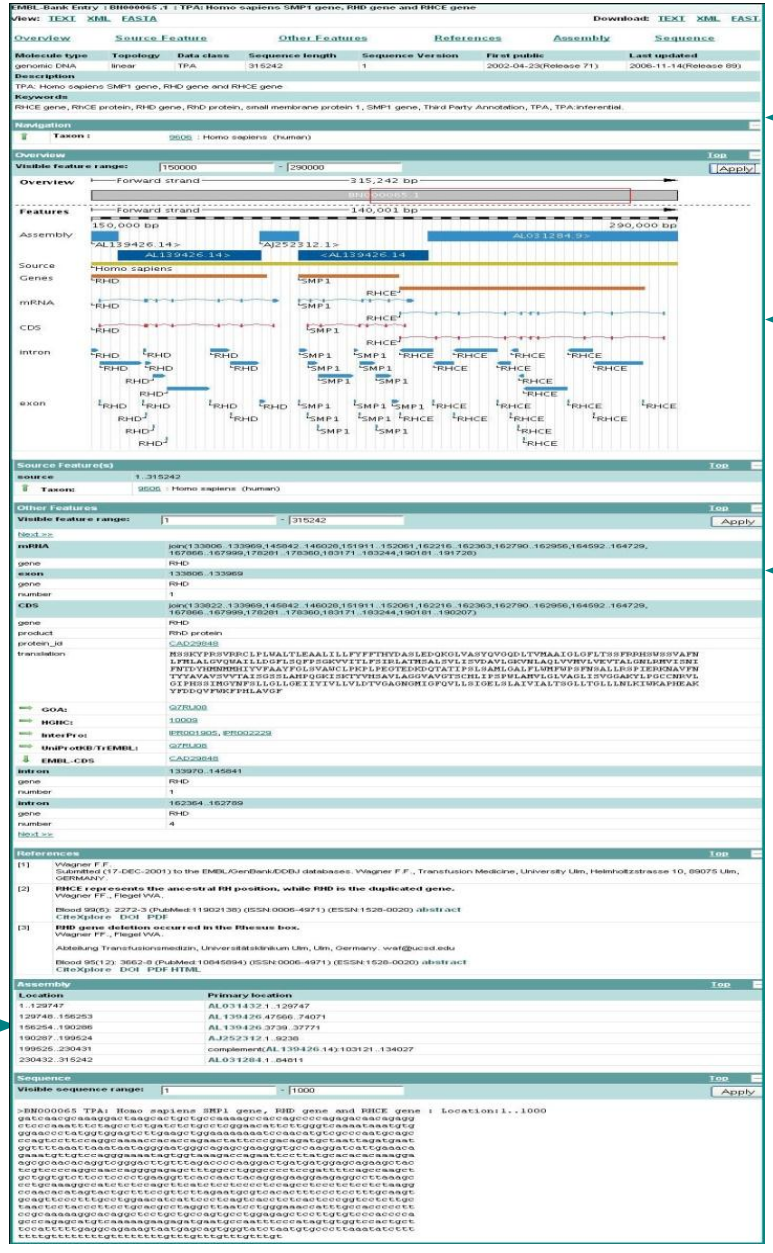
Graphical viewer

DNA source

Navigate to external data sources **e.g. UniProt**



Non-patent entry in EMBL-Bank



General information

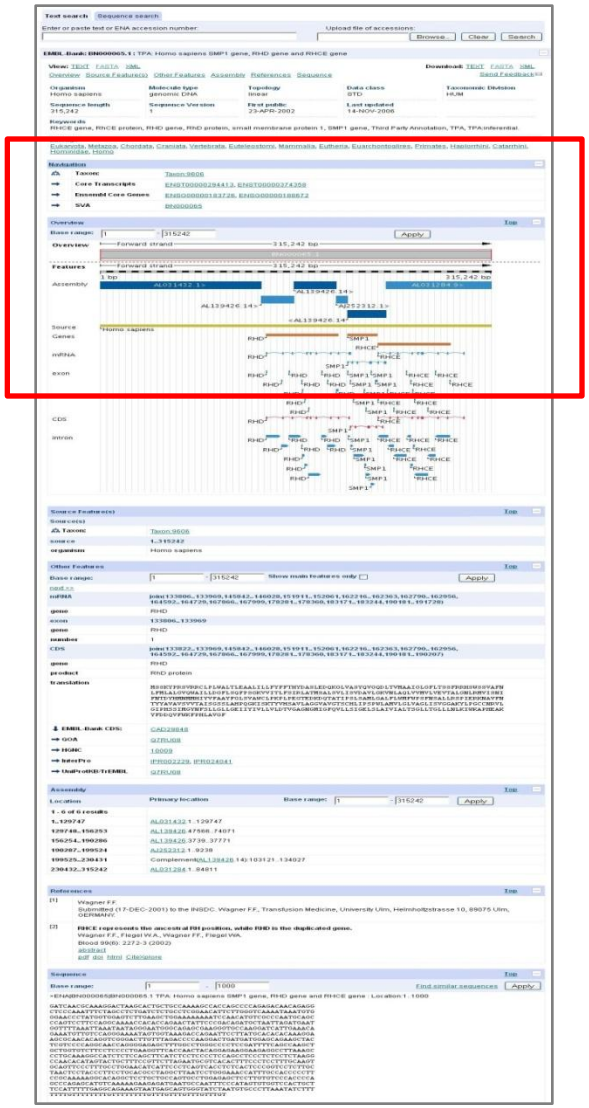
More detailed graphical view

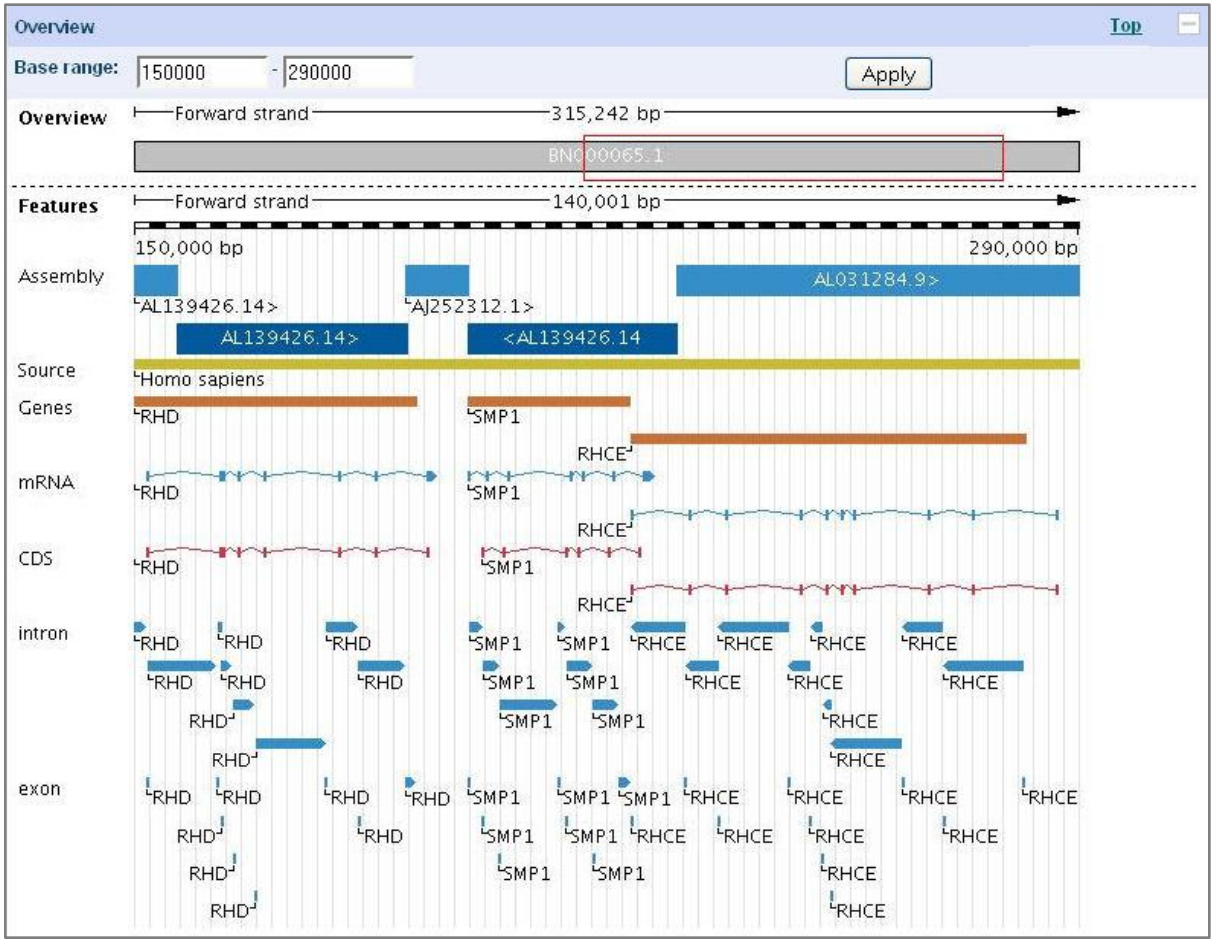
Genome annotation

Additional information

Assembly information

ENA graphical viewer





Patent sequences in ENA

- **Comprehensive** archive with >160 million entries
- **Release** every 3 months, daily updates
- **EMBL-Bank** contains **>22 million patent** entries
- Patent sequences from **EPO, USPTO, JPO** and **KIPO**
- Sequence **redundancy** arises from different patents claiming same sequence
- Redundancy may be useful for studies in variation

…but a **problem for patents**

ENA

ENA is a comprehensive resource for

nucleotide sequence data,

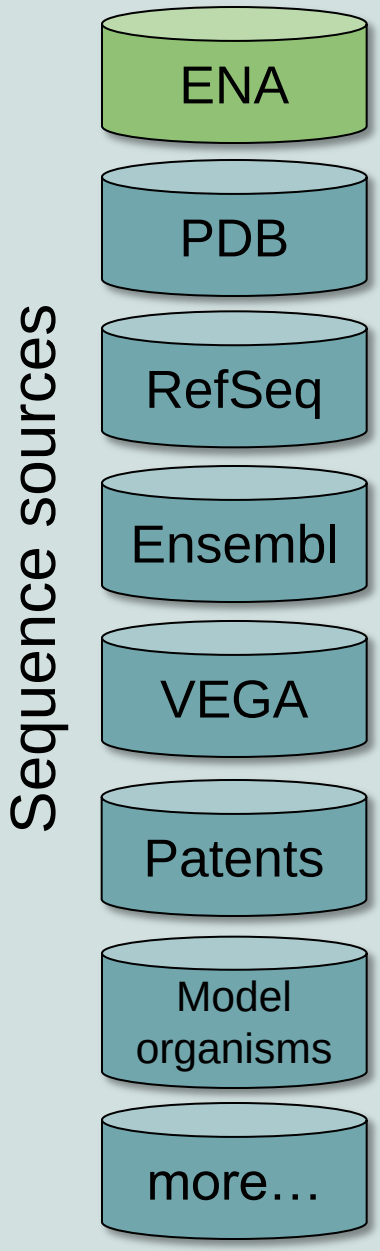
but it is better for **non-patent data**

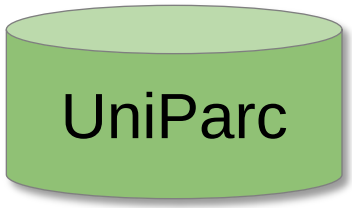
European nucleotide archive

UniProt

Non-redundant patent sequence databases

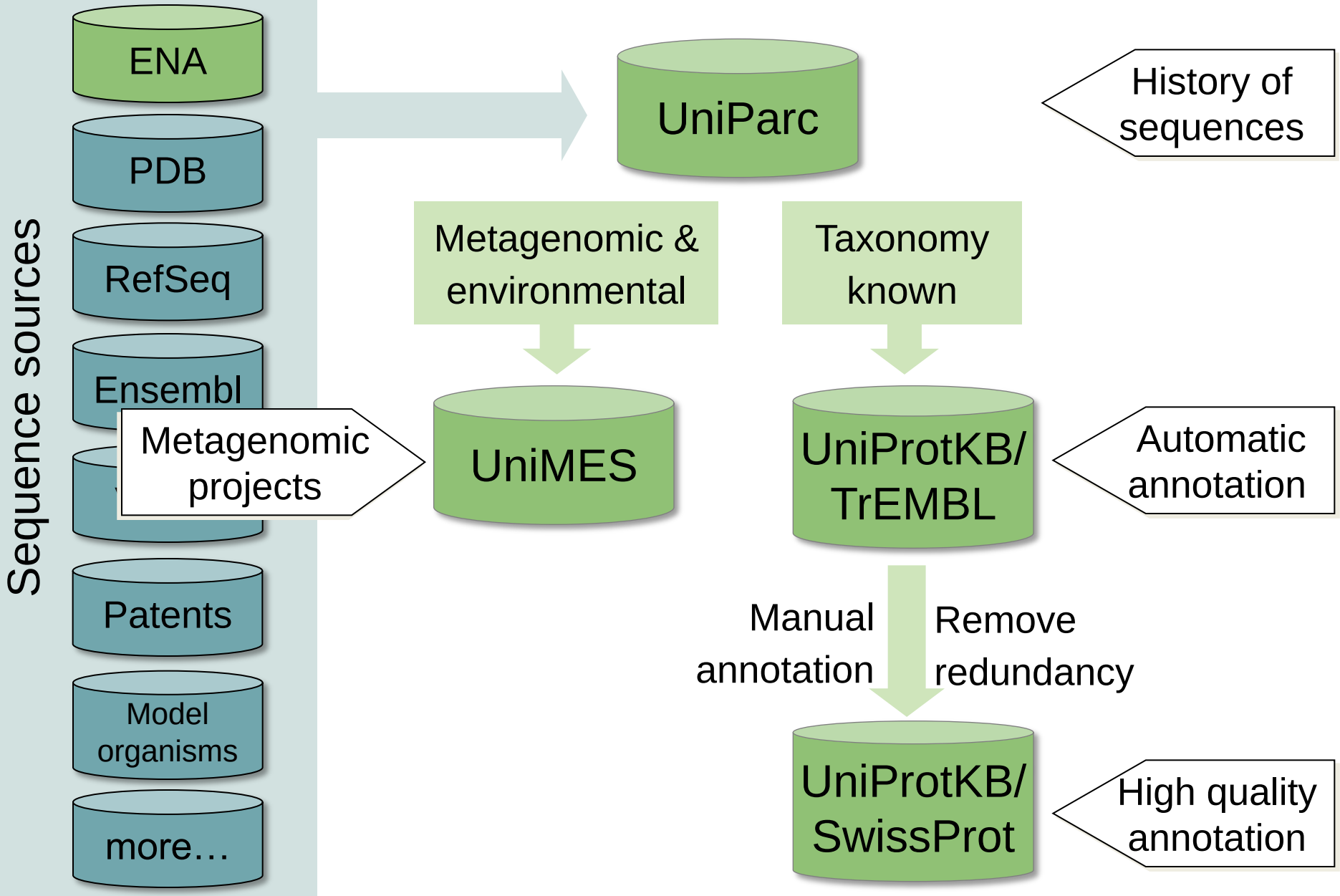
Where does the data come from?



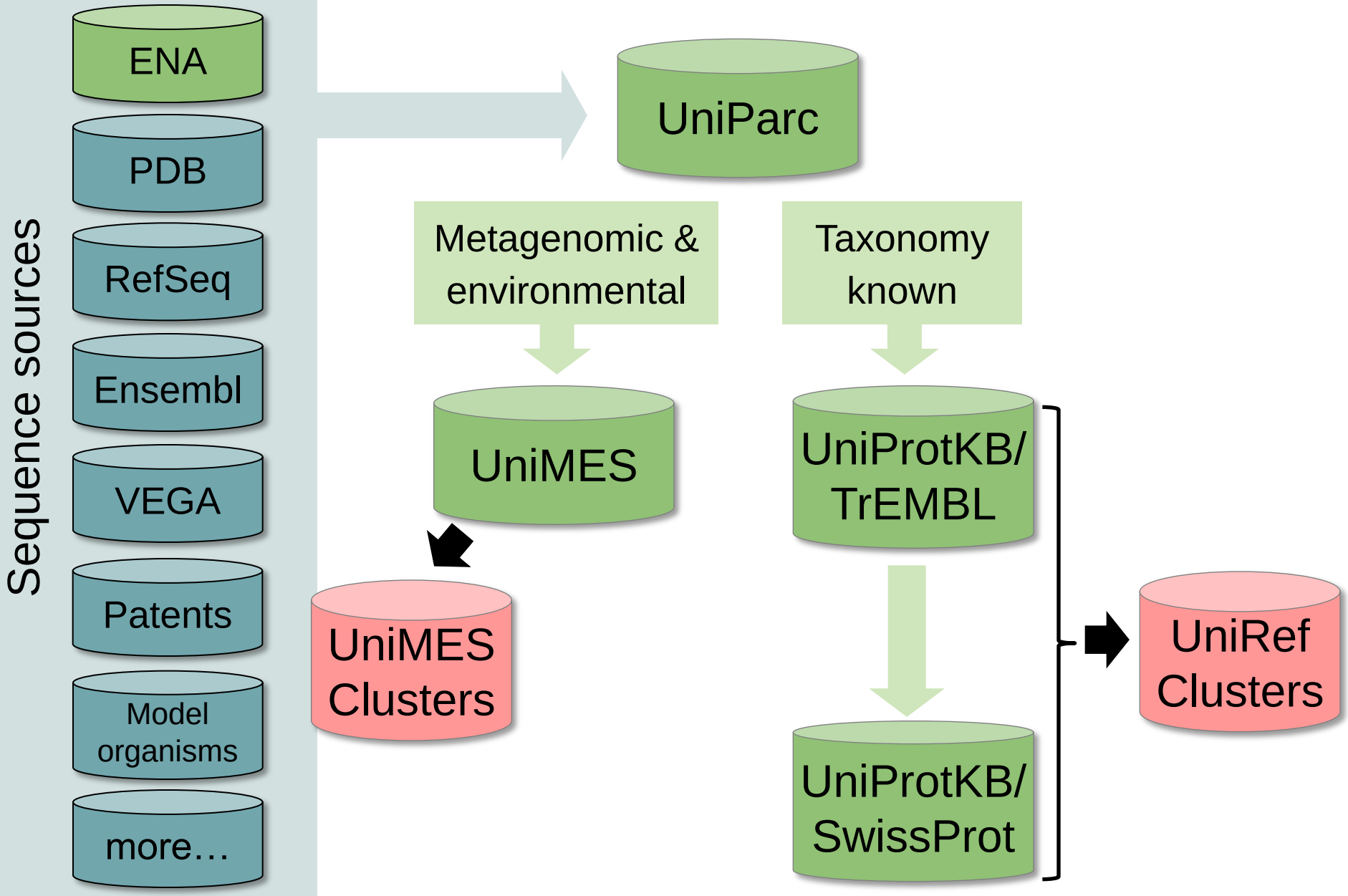


exchange data daily

UniProt has a 3-tiered structure



UniProt has a 3-tiered structure



UniProt has a 3-tiered structure

UniParc

- Complete history of sequences *(no annotation)*
- Cross-links to external sequence sources

UniProtKB

- *Swiss-Prot*: non-redundant, manual annotation
- *TrEMBL*: redundant, automatic annotation

UniMES

- Sequences from metagenomic projects

UniRef

- Combines sequences (speed searching)
- UniRef100, UniRef90, UniRef50

Patent sequence record in UniParc

Accession

Patents not found in UniProtKB

List of databases containing sequence

Navigate to individual entries

Sequence **UPI000002B7A2**

xml rdf/xml fasta tab

This sequence was excluded from UniProtKB because it is a patent sequence.

Customize display | Entries · Sequence

Entries

Found in 4 databases: REMTREMBL, EMBL CDS, USPTO and EPO

› Hide inactive entries (1)

Database	Entry	Version	Organism	First seen	Last seen	Active
EMBL CDS	CAA00041	1	Homo sapiens (Human)	2003-03-12	2010-06-06	Yes
REMTREMBL	E961026			2003-03-28	2003-11-17	No
EPO	A00210	1	Homo sapiens (Human)	2003-03-26	2010-06-10	Yes
USPTO	AAN19596	1		2003-03-26	2010-06-16	Yes
USPTO	AAQ71736	1		2004-07-05	2010-06-16	Yes

Sequence

Sequence	Length	Mass (Da)
UPI000002B7A2	124	13,013

Checksum: 3359AA86A31AF06F

```
        10         20         30         40         50         60
MVVSEVDIAK ADPAAASHPL LLNGDATVAQ KNPGSVAENN LCSQYEEKVR PCIDLIDSLR

        70         80         90        100        110        120
ALGVEQDLAL PAIAVIGDQS SGKSSVLEAL SGVALPRGSG IVTRCPLVLK LKKLVNEDKW

RGKV
```

Download data

Deleted entries identified (greyed out)

Sequence

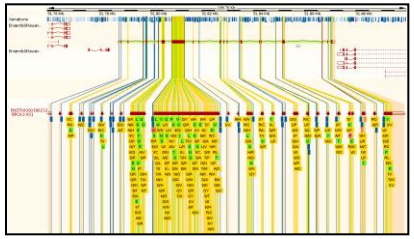
Browsing a UniProtKB/SwissProt entry

Names (synonyms) and taxonomy

Protein attributes

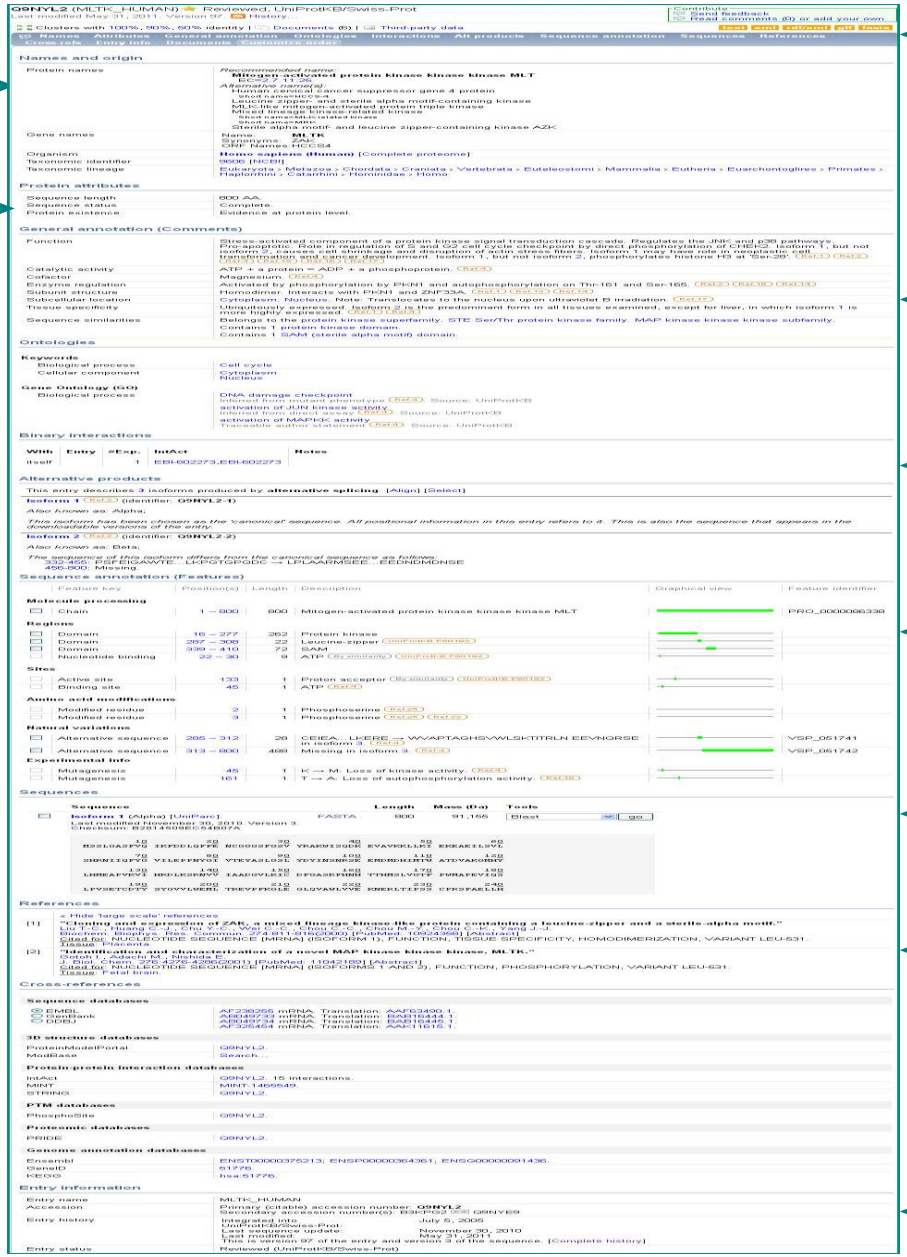
Ontologies

Splice variants

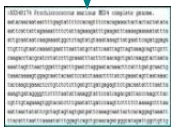


Navigate to external data sources

e.g. Ensembl



Download data



Annotation

Protein interactions

Sequence features

Sequence

References

General information

Browsing a UniProtKB/TrEMBL entry

Name
(could be clone name)

Taxonomy

Automatic annotation
(derived from InterPro)

Ontologies
(both automatic and manual curation)

Names and origin

Protein names	*Submitted name:* Lsk protein EMBL AAB30995.1
Gene names	Name:**lsk** EMBL AAB30995.1
Organism	**Homo sapiens (Human)** EMBL AAB30995.1
Taxonomic identifier	9606 [NCBI]
Taxonomic lineage	Eukaryota › Metazoa › Chordata › Craniata › Vertebrata › Euteleostomi › Mammalia › Eutheria › Euarchontoglires › Primates › Haplorrhini › Catarrhini › Hominidae › Homo

General annotation (Comments)

Catalytic activity	ATP + a [protein]-L-tyrosine = ADP + a [protein]-L-tyrosine phosphate. SAAS SAAS000980
Sequence similarities	Contains 1 SH2 domain. SAAS SAAS020635 Contains 1 SH3 domain. SAAS SAAS020635 Contains 1 protein kinase domain. SAAS SAAS020635

Ontologies

Keywords

Ligand	ATP-binding SAAS SAAS020635 Nucleotide-binding
Molecular function	Kinase Transferase Tyrosine-protein kinase SAAS SAAS000980

Gene Ontology (GO)

Biological process	protein phosphorylation Inferred from electronic annotation. Source: InterPro signal transduction Inferred from electronic annotation. Source: InterPro

Browsing a UniRef90 entry

Faster and more sensitive sequence search with no loss of information

Map to UniProtKB or UniParc

Download

Page 1 of 1

Results Customize

› Retrieve Q9NYL2

	Accession	Status	Cluster name	Size	Members	Organisms	Length	Identity
☐	UniRef90_Q9NYL2	★	Cluster: Mitogen-activated protein kinase kinase kinase MLT	16	Q9NYL2 UPI0001CE0AC4 D2HJF2 Q9ESL4 A2AR41 E1B8N9 E2REE9 B8ZZU2 UPI00005A58D7 +6	Homo sapiens (Human) Oryctolagus cuniculus (Rabbit) Ailuropoda melanoleuca (Giant panda) Mus musculus (Mouse) Bos taurus (Bovine) Canis familiaris (Dog) (Canis lupus familiaris) Equus caballus (Horse) Callithrix jacchus (White-tufted-ear marmoset) Macaca mulatta (Rhesus macaque) +1	800	90%

Status (SwissProt and/or TrEMBL)

Cluster name

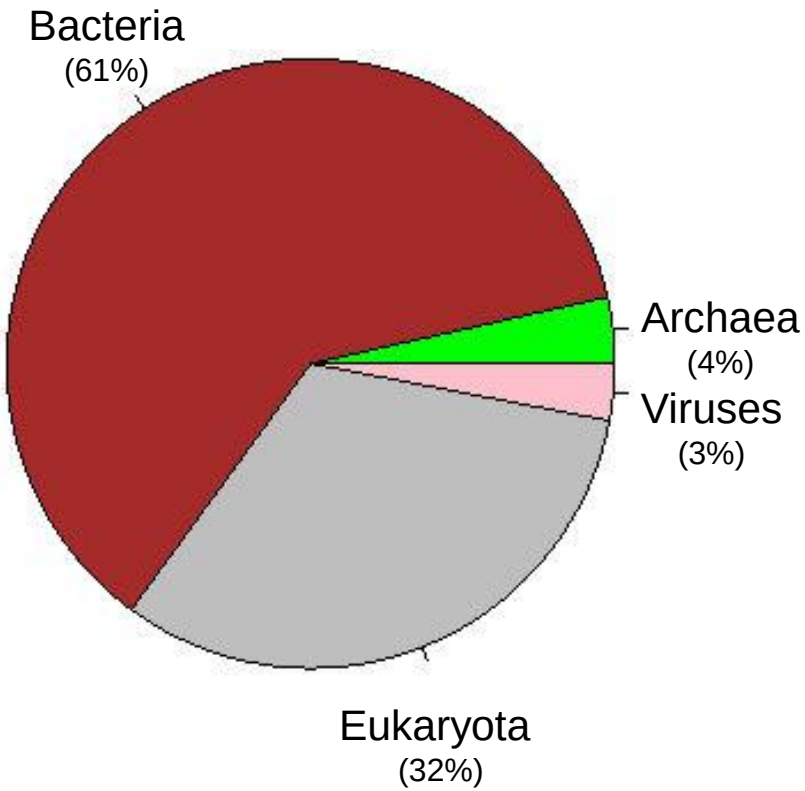
List of entries in cluster

Taxonomy of each entry

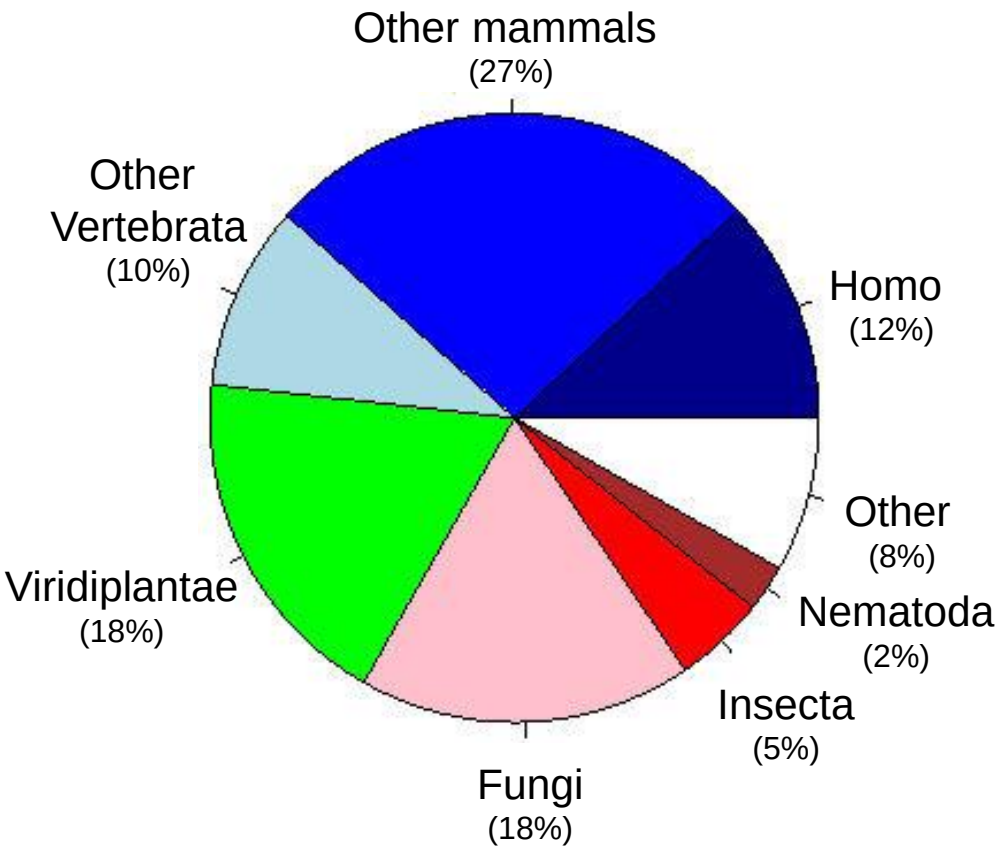
% identity of sequences in cluster

Taxonomic distribution of species

All kingdoms:



Within Eukaryota:



SwissProt – most represented species

Mainly model organisms

2.2 Table of the most represented species

Number	Frequency	Species
1	20331	Homo sapiens (Human)
2	16072	Mus musculus (Mouse)
3	7723	Arabidopsis thaliana (Mouse-ear cress)
4	7269	Rattus norvegicus (Rat)
5	6552	Saccharomyces cerevisiae (Baker's yeast)
6	5559	Bos taurus (Bovine)
7	4752	Schizosaccharomyces pombe (Fission yeast)
8	4342	Escherichia coli (strain K12)
9	3576	Bacillus subtilis
10	3410	Dictyostelium discoideum (Slime mold)
11	3212	Caenorhabditis elegans
12	2921	Xenopus laevis (African clawed frog)
13	2883	Drosophila melanogaster (Fruit fly)
14	2374	Danio rerio (Zebrafish) (Brachydanio rerio)
15	2184	Pongo abelii (Sumatran orangutan)
16	2089	Gallus gallus (Chicken)
17	1981	Oryza sativa subsp. japonica (Rice)
18	1974	Escherichia coli O157:H7
19	1782	Methanocaldococcus jannaschii (Methanococcus jannaschii)
20	1773	Haemophilus influenzae

Protein existence tag

!! Not sequence validation !!

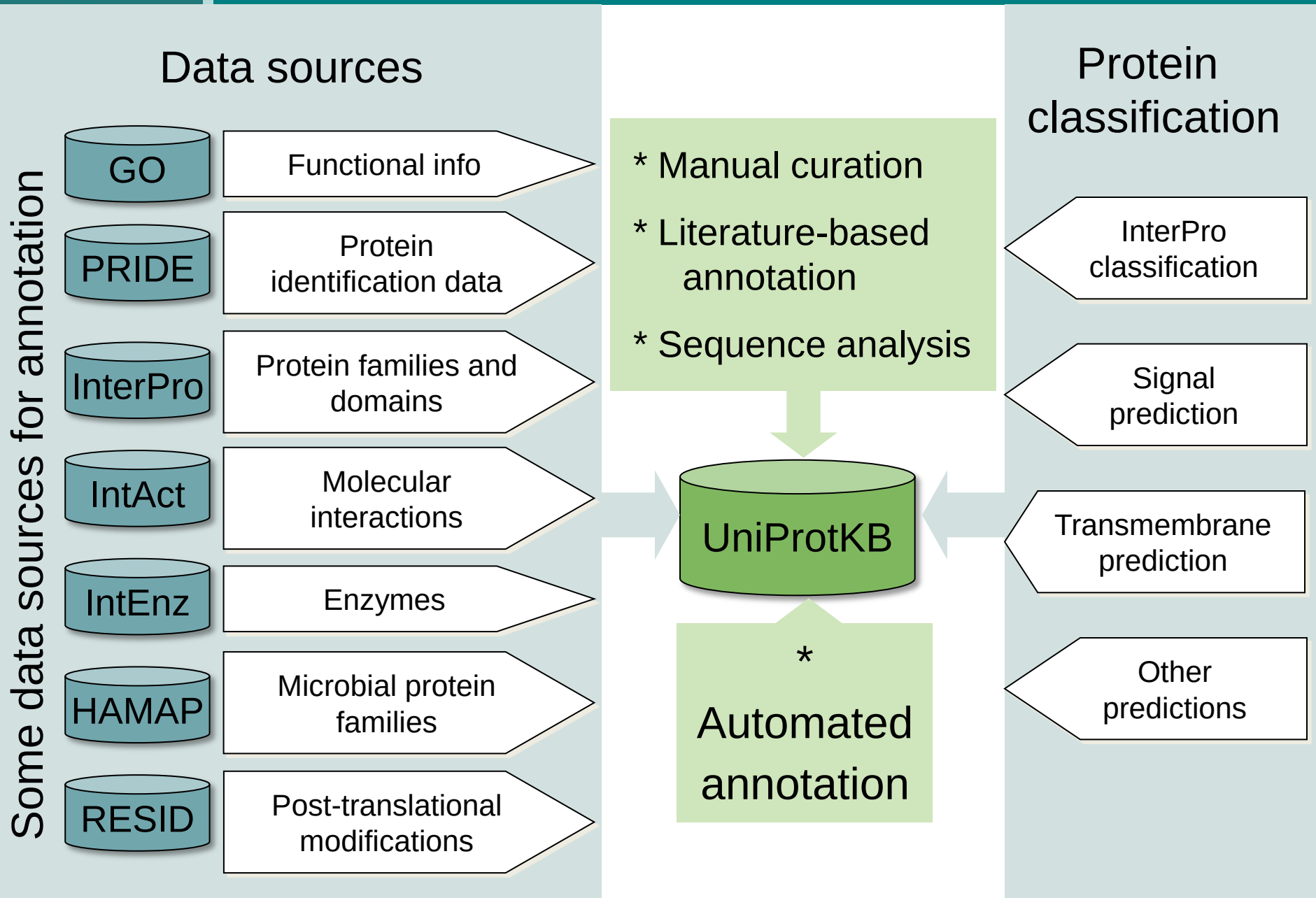
Protein existence level:	Total
Evidence at protein level	13%
Evidence at transcript level	12%
Inferred from homology	70%
Predicted	5%
Uncertain (mainly TrEMBL)	-

Protein existence tag

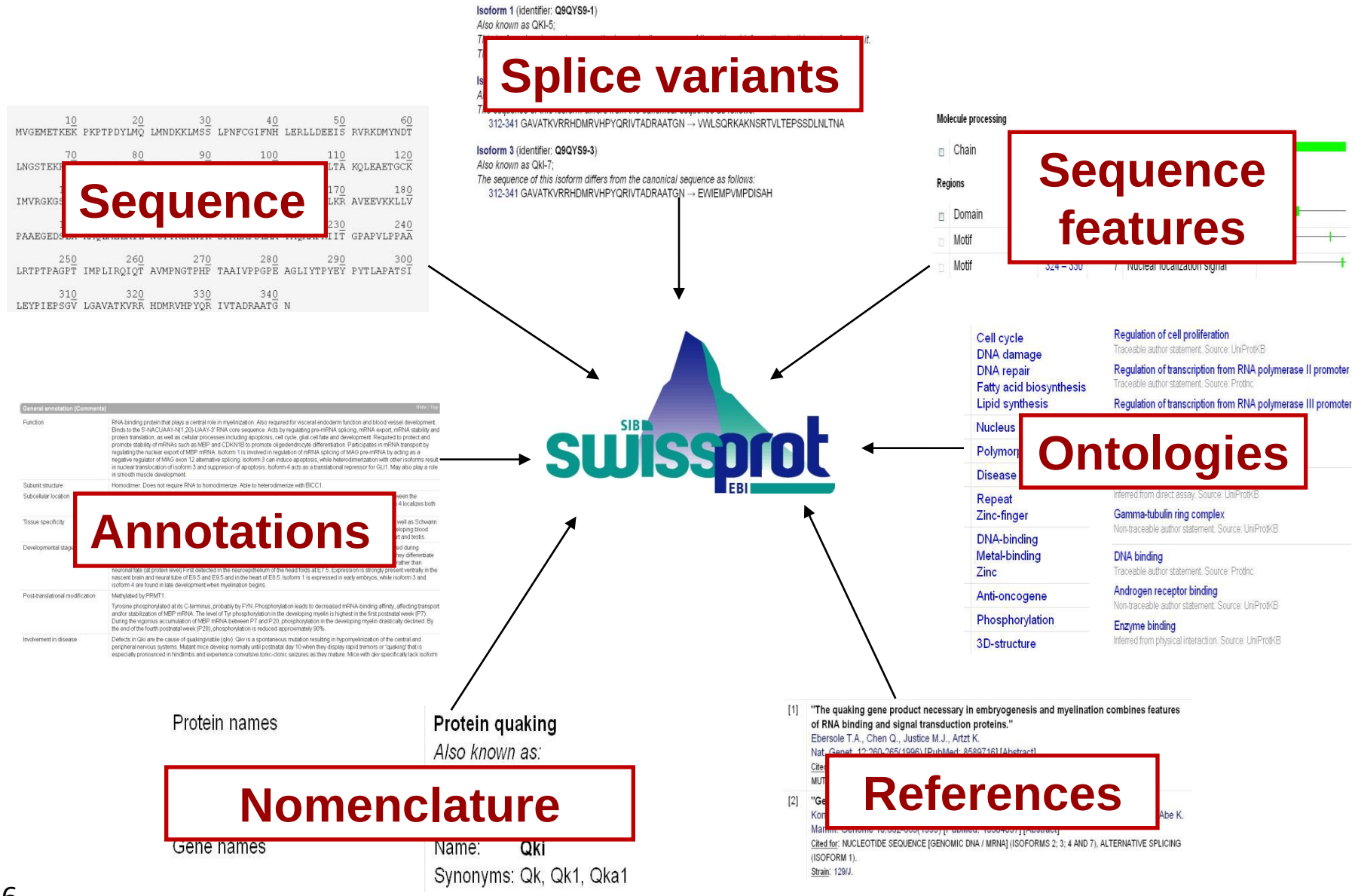
!! Not sequence validation !!

Protein existence level:	Human
Evidence at protein level	59%
Evidence at transcript level	37.5%
Inferred from homology	1%
Predicted	0.5%
Uncertain (mainly TrEMBL)	2%

Annotation sources for UniProtKB



Features of UniProtKB



A wealth of external links

125 links!

Organism-specific DBs

DictyBase, AGD, EchoBASE, CGD, EcoGene, CTD, euHCVdb, CYGD, FlyBase, HGNC, GeneCards, HPA, GeneFarm, MGI, Gramene, MIM, H-InvDB, RGD, LegioList, SGD, Leproma, TAIR, ListiList, ZFIN, MaizeGDB, MypuList, Orphanet, PharmGKB, PhotoList, PseudoCAP, SagaList, SubtiList, TubercuList, WormBase, WormPep, Xenbase, GeneDB_Spombe, ArachnoServer, BuruList

Enzyme & pathway DBs

BioCyc, BRENDA, Reactome, Pathway_Interaction_DB

Proteomic DBs

PeptideAtlas, PRIDE, ProMEX

Genome annotation DBs

Ensembl, KEGG, GeneID, NMPDR, VectorBase, UCSC, GenomeReviews, TIGR

Family and domain DBs

Gene3D, PIRSF, HAMAP, PRINTS, InterProProDom, PANTHER, PROSITE, Pfam, TIGRFAMs, SMART

Phylogenomic DBs

HOGENOM, OMA, HOVERGEN, PhylomeDB, InParanoid, OrthoDB

Polymorphism DBs

dbSNP

Ontologies

GO

2D gel DBs

2DBase-Ecoli, ANU-2DPAGE, Aarhus/Ghent-2DPAGE *(no server)*, COMPLUYEAST-2DPAGE, Cornea-2DPAGE, DOSAC-COBS-2DPAGE, ECO2DBASE *(no server)*, HSC-2DPAGE, OGP, PHCI-2DPAGE, PMMA-2DPAGE, Rat-heart-2DPAGE, REPRODUCTION-2DPAGE, Siena-2DPAGE, SWISS-2DPAGE, World-2DPAGE

3D structure DBs

DisProt, HSSP, PDB, PDBsum, SMR

Gene expression DBs

ArrayExpress, Bgee, GermOnline, CleanEx, Genevestigator

Protein family/group DBs

CAZy, MEROPS, PeroxiBase, REBASE, PptaseDB, TCDB

PTM DBs

GlycoSuiteDB, PhosphoSite, PhosSite

Others

BindingDB, PMAP-CutDB, DrugBank, NextBio

Sequence DBs

EMBL, IPI, PIR, RefSeq, UniGene

Protein-protein interaction DBs

DIP, IntAct, STRING

SwissProt manual annotation

1. Protein sequence

- Merge available CDS (coding sequence)
- Annotate sequence discrepancies
- Report sequencing errors...

2. Biological information

- Extract literature information
- Orthologue data propagation
- Protein sequence analysis...

Merge available CDS

1 SwissProt entry = 1 gene (1 species)

Entry information	Hide | Top
Entry name	MDM2_HUMAN
Accession	Primary (citable) accession number: **Q00987** Secondary accession number(s): Q13226, Q13297, Q13298, Q13299, Q13300, Q13301, Q53XW0, Q71TW9, Q8WYJ1, Q8WYJ2, Q9UGI3, Q9UMT8
Entry history	Integrated into UniProtKB/Swiss-Prot: April 1, 1993 Last sequence update: April 1, 1993 Last modified: December 4, 2007 This is version 107 of the entry and version 1 of the sequence. [Complete history]
Entry status	Reviewed (UniProtKB/Swiss-Prot)
Annotation project	HPI (Human Proteome Initiative)

→Merge TrEMBL entries representing the same protein

→ Manually analyze and annotate the differences

Annotate sequence discrepancies

Identification of amino acid variants

☐	Natural variant	35	1	L → F in a liver tumor.
☐	Natural variant	43	1	L → S in a renal tumor.
☐	Natural variant	47	1	P → S: dbSNP rs1800371.
☐	Natural variant	53	1	W → C in a leukemia and a lymphoma.
☐	Natural variant	60	1	P → S in a leukemia and a lymphoma.
☐	Natural variant	72	1	P → R: dbSNP rs1042522.

....and of PTMs

☐	Modified residue	46	1	Phosphoserine; by HIPK2
☐	Modified residue	55	1	Phosphothreonine; by TAF1
☐	Modified residue	99	1	Phosphoserine
☐	Modified residue	305	1	N6-acetyllysine
☐	Modified residue	315	1	Phosphoserine; by CDC2
☐	Modified residue	373	1	N6-acetyllysine

Report sequencing errors

Cross-references

Sequence databases

EMBL	AF497978 mRNA. Translation: AAM83402.1. Frameshift. AK027356 mRNA. Translation: BAB55058.1. AF212241 mRNA. Translation: AAK14926.1. Different initiation. AF109358 mRNA. Translation: AAQ13506.1. Frameshift. AF172818 mRNA. Translation: AAQ13612.1. Different initiation. BC011885 mRNA. Translation: AAH11885.1.

Caution	Was initially thought to constitute the ortholog of prokaryotic IF-2 (infB) protein.
Sequence caution	AAM83402.1 sequence differs from that shown. Reason: Frameshift at positions 82 and 90. AAQ13506.1 sequence differs from that shown. Reason: Frameshift at position 513.

Other examples of sequencing errors include:

premature stop codons, read-throughs, erroneous initiator methionines

SwissProt manual annotation

1. Protein sequence

- Merge available CDS (coding sequence)
- Annotate sequence discrepancies
- Report sequencing errors...

2. Biological information

- Extract literature information
- Orthologue data propagation
- Protein sequence analysis...

Sources of annotated information

UniProtKB/SwissProt gathers information from multiple sources:

- Publications (literature/PubMed)
- Prediction proteins (Prosite, Anabelle)
- Contact with experts
- Other databases
- Nomenclature committees

Nomenclature

Names and origin

Protein names	*Recommended name:* **Glutamate carboxypeptidase 2** EC=3.4.17.21 *Alternative name(s):* Cell growth-inhibiting gene 27 protein Folate hydrolase 1 Folylpoly-gamma-glutamate carboxypeptidase Short name=FGCP Glutamate carboxypeptidase II Short name=GCPII Membrane glutamate carboxypeptidase Short name=mGCP N-acetylated-alpha-linked acidic dipeptidase I Short name=NAALADase I Prostate-specific membrane antigen Short name=PSM Short name=PSMA Pteroylpoly-gamma-glutamate carboxypeptidase
Gene names	Name: **FOLH1** Synonyms: FOLH, NAALAD1, PSM, PSMA ORF Names: GIG27

Synonyms useful for literature searching

Nomenclature

Names and origin

Protein names	*Recommended name:* **Arginine biosynthesis bifunctional protein ArgJ** Cleaved into the following 2 chains: 1. ***Arginine biosynthesis bifunctional protein ArgJ alpha chain*** 2. ***Arginine biosynthesis bifunctional protein ArgJ beta chain*** Including the following 2 domains: 1. ***Glutamate N-acetyltransferase*** EC=2.3.1.35 *Alternative name(s):* Ornithine acetyltransferase Short name=OATase Ornithine transacetylase 2. ***Amino-acid acetyltransferase*** EC=2.3.1.1 *Alternative name(s):* N-acetylglutamate synthase Short name=AGS
Gene names	Name: **argJ** Ordered Locus Names: BSU11200

Provides synonyms and cleavage products of bifunctional proteins

Molecule processing

Transit peptide	1 – 8	8	Mitochondrion	
Chain	9 – 214	206	Arginine biosynthesis bifunctional protein ARG7 alpha chain	
Chain	215 – 441	227	Arginine biosynthesis bifunctional protein ARG7 beta chain	

Sites

Site	214 – 215	2	Cleavage; by autolysis	

Annotation comments

>30 comment fields

General annotation (Comments)

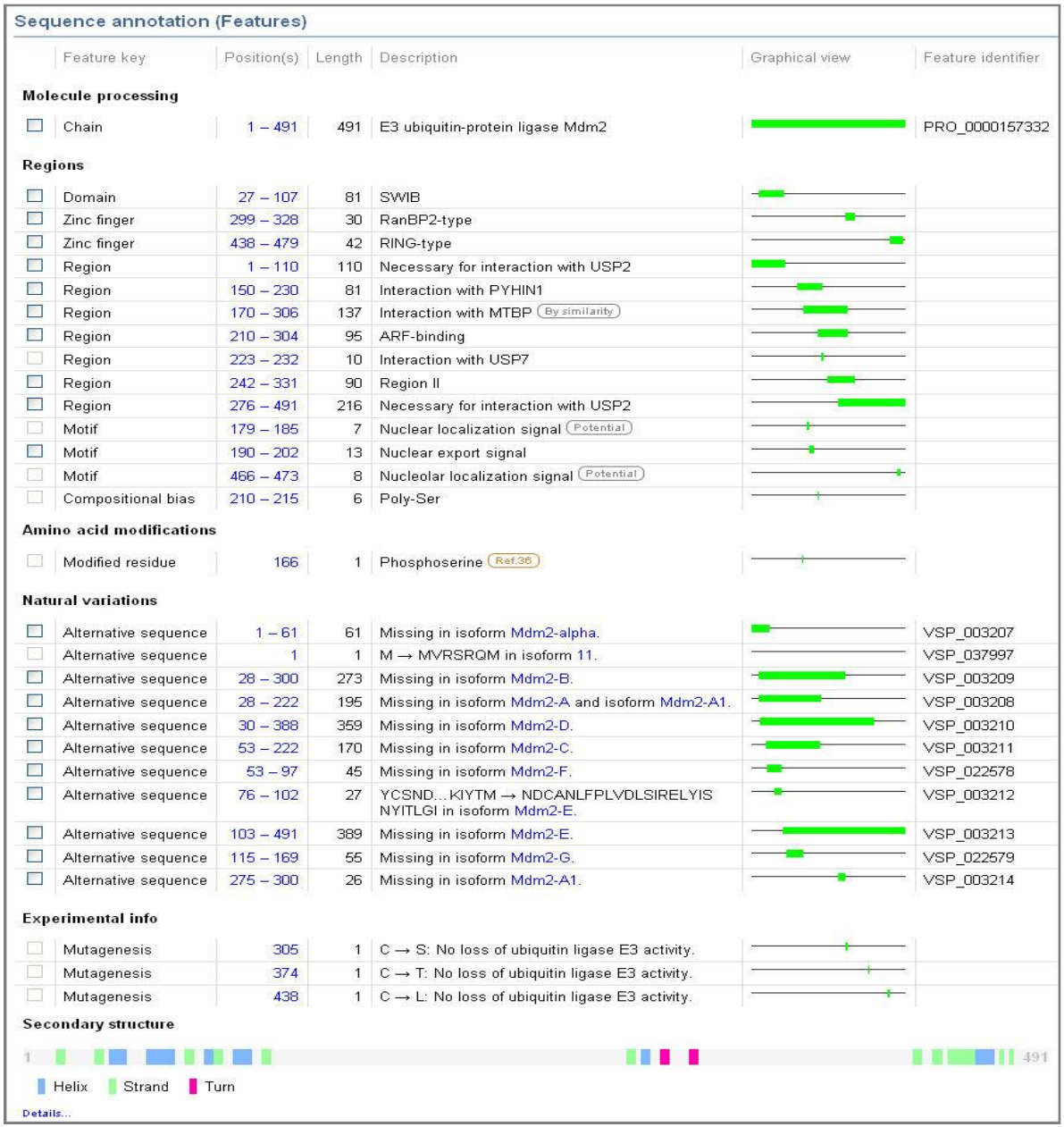
Function	Catalyzes two activities which are involved in the cyclic version of arginine biosynthesis: the synthesis of acetylglutamate from glutamate and acetyl-CoA, and of ornithine by transacetylation between acetylornithine and glutamate. Ref.5
Catalytic activity	N(2)-acetyl-L-ornithine + L-glutamate = L-ornithine + N-acetyl-L-glutamate. Acetyl-CoA + L-glutamate = CoA + N-acetyl-L-glutamate.
Enzyme regulation	Inhibited by ornithine. Ref.5
Pathway	Amino-acid biosynthesis; L-arginine biosynthesis; L-ornithine and N-acetyl-L-glutamate from L-glutamate and N(2)-acetyl-L-ornithine (cyclic): step 1/1. Amino-acid biosynthesis; L-arginine biosynthesis; N(2)-acetyl-L-ornithine from L-glutamate: step 1/4.
Subunit structure	Heterodimer of an alpha and a beta chain.
Subcellular location	Mitochondrion matrix Ref.7.
Post-translational modification	The alpha and beta chains are autoproteolytically processed from a single precursor protein within the mitochondrion.
Sequence similarities	Belongs to the ArgJ family.
Biophysicochemical properties	Kinetic parameters: K_M=8.4 mM for glutamate Ref.8 K_M=2.8 mM for N-acetylornithine

Controlled vocabularies used whenever possible…

Sequence annotation (Features)

…enable researchers to obtain a summary of what is known about a protein…

…including domain annotation, identifying binding sites…



Sequence annotation (Features)

	Feature key	Position(s)	Length	Description	Graphical view	Feature identifier
Molecule processing						
☐	Chain	1 – 491	491	E3 ubiquitin-protein ligase Mdm2		PRO_0000157332
Regions						
☐	Domain	27 – 107	81	SWIB		
☐	Zinc finger	299 – 328	30	RanBP2-type		
☐	Zinc finger	438 – 479	42	RING-type		
☐	Region	1 – 110	110	Necessary for interaction with USP2		
☐	Region	150 – 230	81	Interaction with PYHIN1		
☐	Region	170 – 306	137	Interaction with MTBP (By similarity)		
☐	Region	210 – 304	95	ARF-binding		
☐	Region	223 – 232	10	Interaction with USP7		
☐	Region	242 – 331	90	Region II		
☐	Region	276 – 491	216	Necessary for interaction with USP2		
☐	Motif	179 – 185	7	Nuclear localization signal (Potential)		
☐	Motif	190 – 202	13	Nuclear export signal		
☐	Motif	466 – 473	8	Nucleolar localization signal (Potential)		
☐	Compositional bias	210 – 215	6	Poly-Ser		
Amino acid modifications						
☐	Modified residue	166	1	Phosphoserine (Ref.36)		
Natural variations						
☐	Alternative sequence	1 – 61	61	Missing in isoform Mdm2-alpha.		VSP_003207
☐	Alternative sequence	1	1	M → MVRSRQM in isoform 11.		VSP_037997
☐	Alternative sequence	28 – 300	273	Missing in isoform Mdm2-B.		VSP_003209
☐	Alternative sequence	28 – 222	195	Missing in isoform Mdm2-A and isoform Mdm2-A1.		VSP_003208
☐	Alternative sequence	30 – 388	359	Missing in isoform Mdm2-D.		VSP_003210
☐	Alternative sequence	53 – 222	170	Missing in isoform Mdm2-C.		VSP_003211
☐	Alternative sequence	53 – 97	45	Missing in isoform Mdm2-F.		VSP_022578
☐	Alternative sequence	76 – 102	27	YCSND...KIYTM → NDCANLFPLVDLSIRELYIS NYITLGI in isoform Mdm2-E.		VSP_003212
☐	Alternative sequence	103 – 491	389	Missing in isoform Mdm2-E.		VSP_003213
☐	Alternative sequence	115 – 169	55	Missing in isoform Mdm2-G.		VSP_022579
☐	Alternative sequence	275 – 300	26	Missing in isoform Mdm2-A1.		VSP_003214
Experimental info						
☐	Mutagenesis	305	1	C → S: No loss of ubiquitin ligase E3 activity.		
☐	Mutagenesis	374	1	C → T: No loss of ubiquitin ligase E3 activity.		
☐	Mutagenesis	438	1	C → L: No loss of ubiquitin ligase E3 activity.		

Secondary structure

1 … 491

Helix Strand Turn

Details...

Sequence annotation (Features)

Feature (e.g. domain) highlighted on sequence

```
        10         20         30         40         50         60
MCNTNMSVPT DGAVTTSQIP ASEQETLVRP KPLLLKLLKS VGAQKDTYTM KEVLFYLGQY

        70         80         90        100        110        120
IMTKRLYDEK QQHIVYCSND LLGDLFGVPS FSVKEHRKIY TMIYRNLVVV NQQESSDSGT

       130        140        150        160        170        180
SVSENRCHLE GGSDQKDLVQ ELQEEKPSSS HLVSRPSTSS RRRAISETEE NSDELSGERQ

       190        200        210        220        230        240
RKRHKSDSIS LSFDESLALC VIREICCERS SSSESTGTPS NPDLDAGVSE HSGDWLDQDS

       250        260        270        280        290        300
VSDQFSVEFE VESLDSEDYS LSEEGQELSD EDDEVYQVTV YQAGESDTDS FEEDPEISLA

       310        320        330        340        350        360
DYWKCTSCNE MNPPLPSHCN RCWALRENWL PEDKGKDKGE ISEKAKLENS TQAEEGFDVP

       370        380        390        400        410        420
DCKKTIVNDS RESCVEENDD KITQASQSQE SEDYSQPSTS SSIIYSSQED VKEFEREETQ

       430        440        450        460        470        480
DKEESVESSL PLNAIEPCVI CQGRPKNGCI VHGKTGHLMA CFTCAKKLKK RNKPCPVCRQ

       490
PIQMIVLTYF P
```

Sequence annotation (Features)

	Feature key	Position(s)	Length	Description	Graphical view	Feature identifier
	lecule processing					
	Chain	1 – 491	491	E3 ubiquitin-protein ligase Mdm2		PRO_0000157332
	ions					
	Domain	27 – 107	81	SWIB		
	Zinc finger	299 – 328	30	RanBP2-type		
	Zinc finger	438 – 479	42	RING-type		
	Region	1 – 110	110	Necessary for interaction with USP2		
	Region	150 – 230	81	Interaction with P53/hIN1		
	Region	170 – 306	137	Interaction with MTBP (By similarity)		
			5	binding		
			0	Interaction with USP7		
			0	Region II		
			6	Necessary for interaction with USP2		
			7	Nuclear localization signal (Potential)		
			8	Nuclear export signal		
			8	Nucleolar localization signal (Potential)		
			6	Poly-Ser		
			1	Phosphoserine (Ref.36)		
			1	Missing in isoform Mdm2-alpha.		VSP_003207
			1	M → MVRSRQM in isoform 11.		VSP_037997
			8	Missing in isoform Mdm2-B.		VSP_003209
			5	Missing in isoform Mdm2-A and isoform Mdm2-A1.		VSP_003208
			9	Missing in isoform Mdm2-D.		VSP_003210
			0	Missing in isoform Mdm2-C.		VSP_003211
			5	Missing in isoform Mdm2-F.		VSP_022578
			7	YCSND...KIYTM → NDCANLFPLVDLSIRELYIS NYITLGI in isoform Mdm2-E.		VSP_003212
			9	Missing in isoform Mdm2-E.		VSP_003213
			5	Missing in isoform Mdm2-G.		VSP_022579
			6	Missing in isoform Mdm2-A1.		VSP_003214
			1	C → S: No loss of ubiquitin ligase E3 activity.		
			1	C → T: No loss of ubiquitin ligase E3 activity.		
			1	C → L: No loss of ubiquitin ligase E3 activity.		

1 491

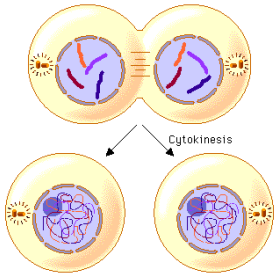
Helix Strand Turn

Details...

Gene Ontology

1. Biological Process

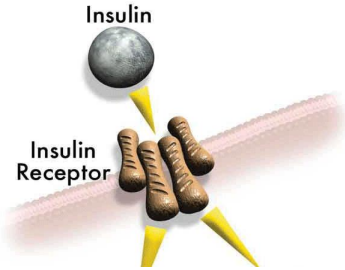
A commonly recognized series of events



- Cell division
- Mitosis
- Organelle fission

2. Molecular Function

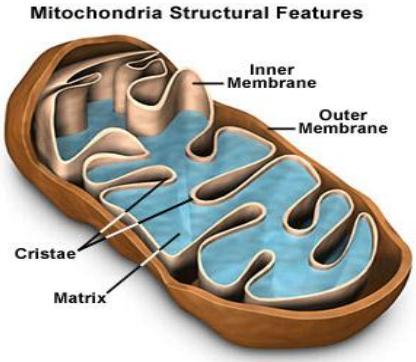
An elemental activity or task or job



- Protein kinase activity
- Insulin binding
- Insulin receptor activity

3. Cellular Component

Where a gene product is located



- Mitochondrion
- Mitochondrial matrix
- Mitochondrial membrane

Gene Ontology

Annotation for human Rhodopsin:

Gene Ontology (GO)

Biological process	phototransduction, visible light Traceable author statement. Source: ProtInc protein-chromophore linkage Inferred from electronic annotation. Source: UniProtKB-KW rhodopsin mediated signaling pathway Traceable author statement. Source: ProtInc
Cellular component	integral to plasma membrane Traceable author statement. Source: ProtInc
Molecular function	G-protein coupled receptor activity Traceable author statement. Source: ProtInc photoreceptor activity Inferred from electronic annotation. Source: UniProtKB-KW protein binding Inferred from physical interaction. Source: IntAct

Complete GO annotation...

Imported annotation

Binary interactions are taken from the IntAct database

Interactors of human p53

Binary interactions

With	Entry	#Exp.	IntAct	Notes
	P03070	4	EBI-366083,EBI-617698	From a different organism.
ATR	Q13535	1	EBI-366083,EBI-968983	
BARD1	Q99728	1	EBI-366083,EBI-473181	
Bard1	O70445	1	EBI-366083,EBI-1790207	From a different organism.
BCL2L1	Q07817-1	2	EBI-366083,EBI-287195	
BMX	P51813	1	EBI-366083,EBI-696657	
BTBD2	Q9BX70	1	EBI-366083,EBI-710091	
CABLES1	Q8TDN4	1	EBI-366083,EBI-604615	
Cables1	Q9ESJ1	1	EBI-366083,EBI-604411	From a different organism.
CEBPB	P17676	3	EBI-366083,EBI-969696	
CREBBP	Q92793	3	EBI-366083,EBI-81215	
Crebbp	P45481	2	EBI-366083,EBI-296306	From a different organism.
CSE1L	P55060	4	EBI-366083,EBI-286709	
CUL7	Q14999	1	EBI-366083,EBI-308606	

Evidence for annotation

P16671 (CD36_HUMAN) Reviewed, UniProtKB/Swiss-Prot

General annotation (Comments)

Function	Seems to have numerous potential physiological functions. Binds to collagen, thrombospondin, anionic phospholipids and oxidized LDL. May function as a cell adhesion molecule. Directly mediates cytoadherence of Plasmodium falciparum parasitized erythrocytes. Binds long [illegible] and may function in the transport and/or as a regulator of fatty acid transport. Ref.19 Ref.21
Subcellular location	Membrane; Multi-pass membrane protein.
Post-translational modification	N-glycosylated and O-glycosylated with a ratio of 2:1 Ref.8 Ref.13 Ref.14 Ref.15
Polymorphism	Genetic variation in CD36 influences the severity and outcome of malaria infection.
Involvement in disease	Defects in CD36 are the cause of platelet glycoprotein IV deficiency (PG4D)[MIM:608404]; also known as CD36 deficiency. Platelet glycoprotein IV deficiency can be divided into 2 subgroups. The type I phenotype is characterized by platelets and monocytes/macrophages exhibiting complete CD36 deficiency. The type II phenotype lacks the surface expression of CD36 in platelets, but expression in monocytes/macrophages is near normal. Ref.16 Ref.20

Proven

Proven

Sequence annotation (Features)

Feature key	Position(s)	Length	Description
Amino acid modifications			
Lipidation	3	1	S-palmitoyl cysteine Ref.12
Lipidation	7	1	S-palmitoyl cysteine Ref.12
Lipidation	464	1	S-palmitoyl cysteine Ref.12
Lipidation	466	1	S-palmitoyl cysteine Ref.12
Glycosylation	79	1	N-linked (GlcNAc...) Potential
Glycosylation	102	1	N-linked (GlcNAc...) Potential
Glycosylation	134	1	N-linked (GlcNAc...) Potential
Glycosylation	163	1	N-linked (GlcNAc...) Potential
Disulfide bond	243 ↔ 311		By similarity
Disulfide bond	272 ↔ 333		By similarity
Disulfide bond	313 ↔ 322		By similarity

Proven

Potential

By similarity

Sources references included

[21] **"Bcl6 and P53 gene mutations in tonsillar B cells."**
Yavuz A.S., Farner N.L., Yavuz S., Grammer A.C., Girschick H.J., Lipsky P.E.
Submitted (MAR-2000) to the EMBL/GenBank/DDBJ databases
Cited for: NUCLEOTIDE SEQUENCE OF 225-260.

[22] **"Allelic variant in intron 9 of TP53 gene."**
Pinto E.M., Mendonca B.B., Latronico A.C.
Submitted (APR-2003) to the EMBL/GenBank/DDBJ databases
Cited for: NUCLEOTIDE SEQUENCE OF 332-366.

[23] **"The tumor suppressor p53 is bound to RNA by a stable covalent linkage."**
Samad A., Carroll R.B.
Mol. Cell. Biol. 11:1598-1606(1991) [PubMed: 1705009] [Abstract]
Cited for: RNA-BINDING.

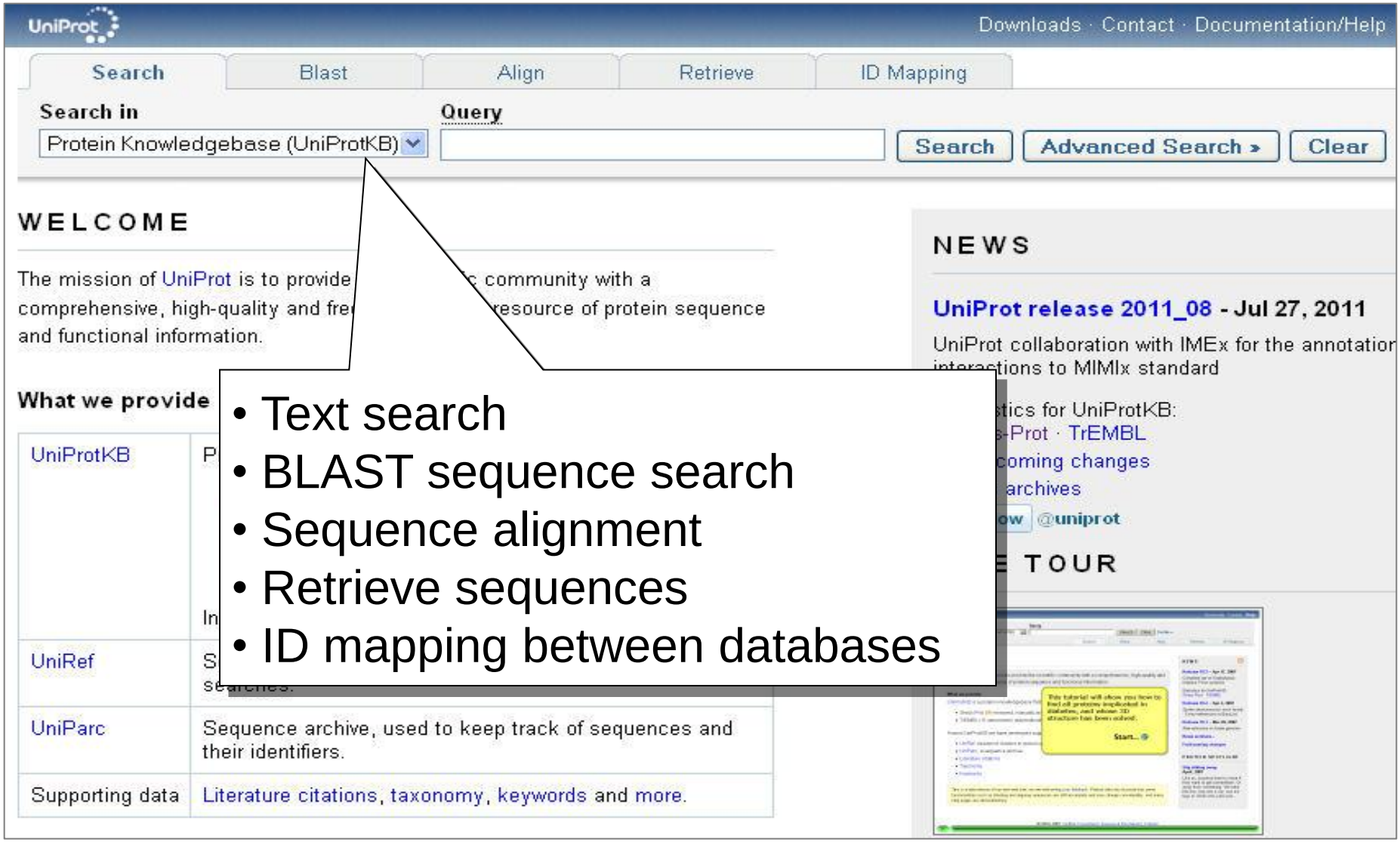
[24] **"The human tumour suppressor gene p53 is alternatively spliced in normal cells."**
Flaman J.-M., Waridel F., Estreicher A., Vannier A., Limacher J.-M., Gilbert D., Iggo R., Frebourg T.
Oncogene 12:813-818(1996) [PubMed: 8632903] [Abstract]
Cited for: ALTERNATIVE SPLICING.

[25] **"The p53 nuclear localisation signal is structurally linked to a p34cdc2 kinase motif."**
Addison C., Jenkins J.R., Sturzbecher H.-W.
Oncogene 5:423-426(1990) [PubMed: 2156209] [Abstract]
Cited for: NUCLEAR LOCALIZATION SIGNAL.

[26] **"p53 amino acids 339-346 represent the minimal p53 repression domain."**
Hong T.M., Chen J.J., Peck K., Yang P.C., Wu C.W.
J. Biol. Chem. 276:1510-1515(2001) [PubMed: 11007800] [Abstract] [Article from publisher]
Cited for: MINIMAL REPRESSION DOMAIN.

\+ Additional computationally mapped references.

UniProt homepage



- Text search
- BLAST sequence search
- Sequence alignment
- Retrieve sequences
- ID mapping between databases

http://www.uniprot.org/

Patent sequences in UniProt

- **Comprehensive** archive consisting of specialised databases
- **Release** every month, daily updates
- UniProtKB/SwissProt annotation-rich, but has no patent data
- Patent sequences only found in **UniParc** as an archived list
- UniParc is **non-redundant** but contains no annotation

…therefore patent information **limited**

UniProt

UniProt is an excellent source of quality protein sequence and annotation data, but it is better for **non-patent data**

Sequence archives

Old entries accessible in both ENA and UniProt

- ENA nucleotide sequence version archive
  www.ebi.ac.uk/embl/sva

The **ENA Sequence Version Archive** is a repository of all entries which have ever appeared in EMBL-Bank Sequence Database.
You can use this page to browse the archive or use the batch retrieval form.
F.A.Q.

Accession Number or Sequence Version: Go!

day-month-year (*e.g.* 30-11-1998 *or* 30-NOV-1998) Current version

Search by accession → get all records

Search by date → get specific record

- UniSave – UniProt sequence/annotation version archive
  www.ebi.ac.uk/uniprot/unisave

The UniProtKB Sequence/Annotation Version Archive (UniSave) is a repository of UniProtKB/Swiss-Prot and UniProtKB/TrEMBL entry versions.

Primary accession number or entry name: Go!

Date: day-month-year (*e.g.* 30-11-1998 *or* 30-NOV-1998) or year-month-day.

Show Help

Provides complete version list

The **ENA Sequence Version Archive** is a repository of all entries which have ever appeared in EMBL-Bank Sequence Database.
You can use this page to browse the archive or use the batch retrieval form.

Sequence Version: AF527840 Go! case sensitive

day-month-year (*e.g.* 30-11-1998 *or* 30-NOV-1998) Curre

31 matches, 1 to 20 shown

Show all on one page | Next page

Accession Number	Entry Version	Sequence Version	Release	Issue Date	
AF527840	2	AF527840.1	108	26-MAY-2011	View
AF527840	2	AF527840.1	105	27-AUG-2010	View
AF527840	2	AF527840.1	104	27-MAY-2010	View
AF527840	2	AF527840.1	103	26-FEB-2010	View
AF527840	2	AF527840.1	102	26-NOV-2009	View
AF527840	2	AF527840.1	101	28-AUG-2009	View
AF527840	2	AF527840.1	100	29-MAY-2009	View
AF527840	2	AF527840.1	99	26-FEB-2009	View
AF527840	2	AF527840.1	98	12-JAN-2009	View
AF527840	2	AF527840.1	97	28-NOV-2008	View
AF527840	2	AF527840.1	96	28-AUG-2008	View
AF527840	2	AF527840.1	95	29-MAY-2008	View
AF527840	2	AF527840.1	93	29-NOV-2007	View
AF527840	2	AF527840.1	90	08-MAR-2007	View
AF527840	2	AF527840.1	89	30-NOV-2006	View
AF527840	2	AF527840.1	89	14-NOV-2006	View
AF527840	1	AF527840.1	87	19-JUN-2006	View
AF527840	1	AF527840.1	86	21-MAR-2006	View
AF527840	1	AF527840.1	85	16-DEC-2005	View
AF527840	1	AF527840.1	84	12-SEP-2005	View

All

Compare Selected | View | Save | **Selected in** EMBL **format.**

Select and **compare** versions

View specific old entry

Tracks all changes to an entry

View old entries

The **ENA Sequence Version Archive** is a repository of all entries which have ever appeared in EMBL-Bank Sequence Database.
You can use this page to browse the archive or use the batch retrieval form.
F.A.Q.

Accession Number or Sequence Version: AX429748 Go! ☐ case sensitive

Snapshot at day-month-year (*e.g.* 30-11-1998 *or* 30-NOV-1998) Current version

Back to List

```
ID   AX429748   standard; DNA; HUM; 100 BP.
XX
AC   AX429748;
XX
SV   AX429748.1
XX
DT   21-JUN-2002 (Rel. 72, Created)
DT   21-JUN-2002 (Rel. 72, Last updated, Version 1)
XX
DE   Sequence 13 from Patent EP1203827.
XX
KW   .
XX
OS   Homo sapiens (human)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Primates; Catarrhini; Hominidae; Homo.
XX
RN   [1]
RA   Smith J.C.;
RT   "Polymorphisms in the human kdr gene";
RL   Patent number EP1203827-A/13, 08-MAY-2002.
RL   AstraZeneca AB  (SE).
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..100
FT                   /db_xref="taxon:9606"
FT                   /organism="Homo sapiens"
XX
SQ   Sequence 100 BP; 14 A; 29 C; 36 G; 20 T; 1 other;
     gcgcgatggc gaagagggtc ctgcactttg acgcgcctgg tgagggagcg stgctcttcg        60
     cagcgctcct ggtgatgctc cccaaatttc ggggaccggc                             100
//
```

Compare different versions

The **ENA Sequence Version Archive** is a repository of all entries which have ever appeared in EMBL-Bank Sequence Database.
You can use this page to browse the archive or use the batch retrieval form.
F.A.Q.

Accession Number or Sequence Version: AX429748 Go! ☐ case sensitive

Snapshot at day-month-year (*e.g.* 30-11-1998 *or* 30-NOV-1998) Current version

Differences for AX429748 30-MAY-2002 / 28-NOV-2008 Back to List

Lines unchanged **Lines removed** **Lines inserted**

```
ID   AX429748; SV 1; linear; unassigned DNA; PAT; HUM; 100 BP.
ID   AX429748   standard; DNA; HUM; 100 BP.
XX
AC   AX429748;
XX
SV   AX429748.1
XX
DT   21-JUN-2002 (Rel. 72, Created)
DT   21-JUN-2002 (Rel. 72, Last updated, Version 1)
XX
DE   Sequence 13 from Patent EP1203827.
XX
KW   .
XX
OS   Homo sapiens (human)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae;
OC   Homo.
OC   Eutheria; Primates; Catarrhini; Hominidae; Homo.
XX
RN   [1]
RA   Smith J.C.;
RT   "Polymorphisms in the human kdr gene";
RL   Patent number EP1203827-A2/13, 08-MAY-2002.
RL   Patent number EP1203827-A/13, 08-MAY-2002.
RL   AstraZeneca AB  (SE).
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..100
FT                   /organism="Homo sapiens"
FT                   /mol_type="unassigned DNA"
FT                   /db_xref="taxon:9606"
FT                   /organism="Homo sapiens"
XX
SQ   Sequence 100 BP; 14 A; 29 C; 36 G; 20 T; 1 other;
     gcgcgatggc gaagagggtc ctgcactttg acgcgcctgg tgagggagcg stgctcttcg      60
     cagcgctcct ggtgatgctc cccaaatttc ggggaccggc                           100
//
```

Changes highlighted

Why do we need a new sequence database for patents?

To address the following problems:

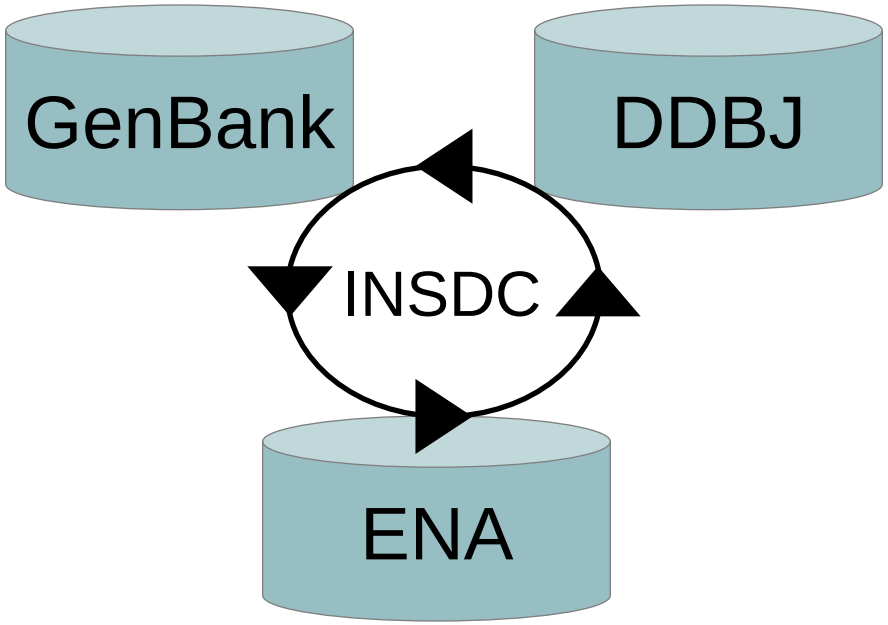
- **redundancy**
- **lack of patent-specific annotation**

European nucleotide archive

UniProt

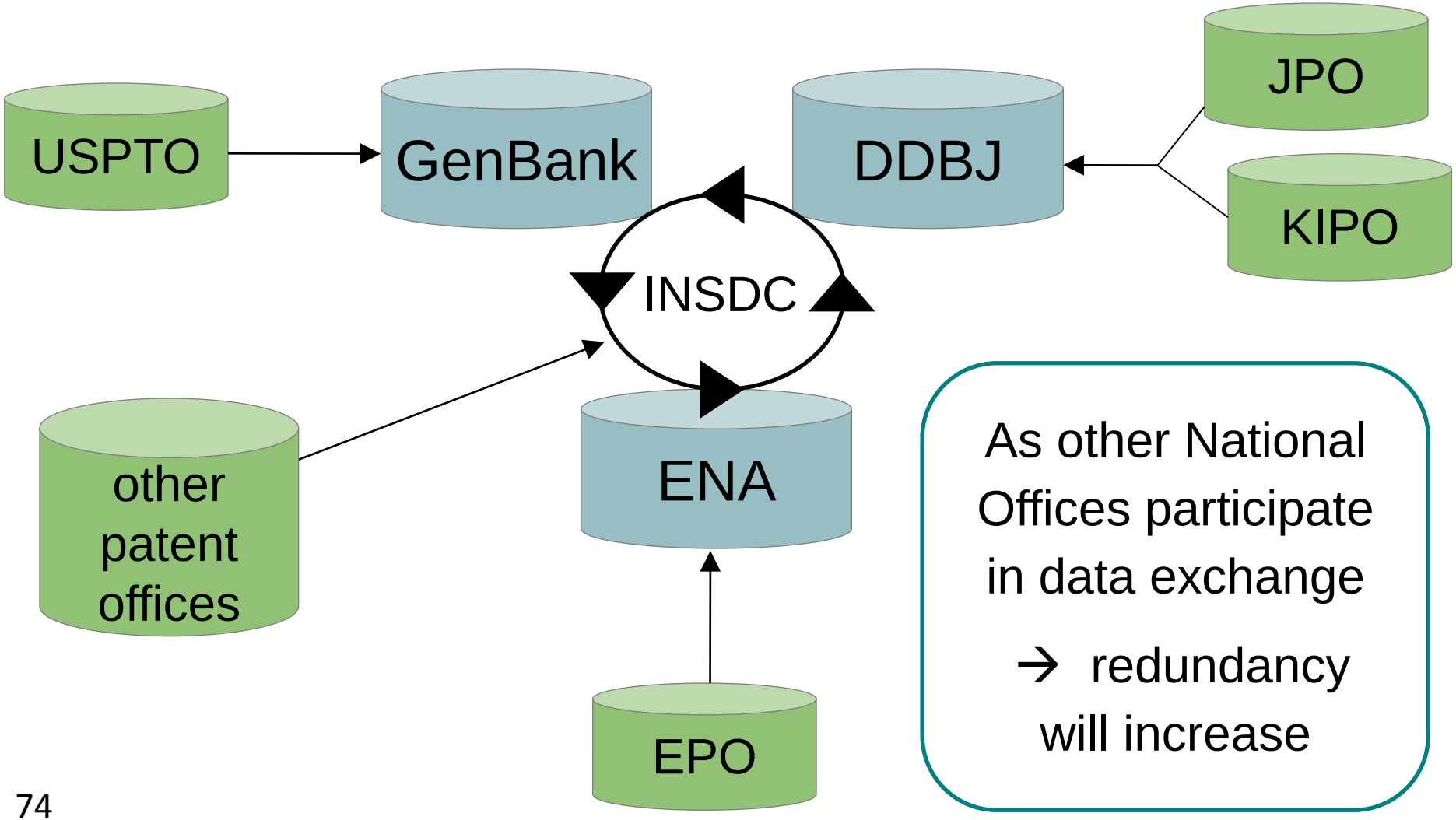
Non-redundant patent sequence databases

Distributing patent sequences



Distributing patent sequences

Redundancy: a consequence of the international cooperation

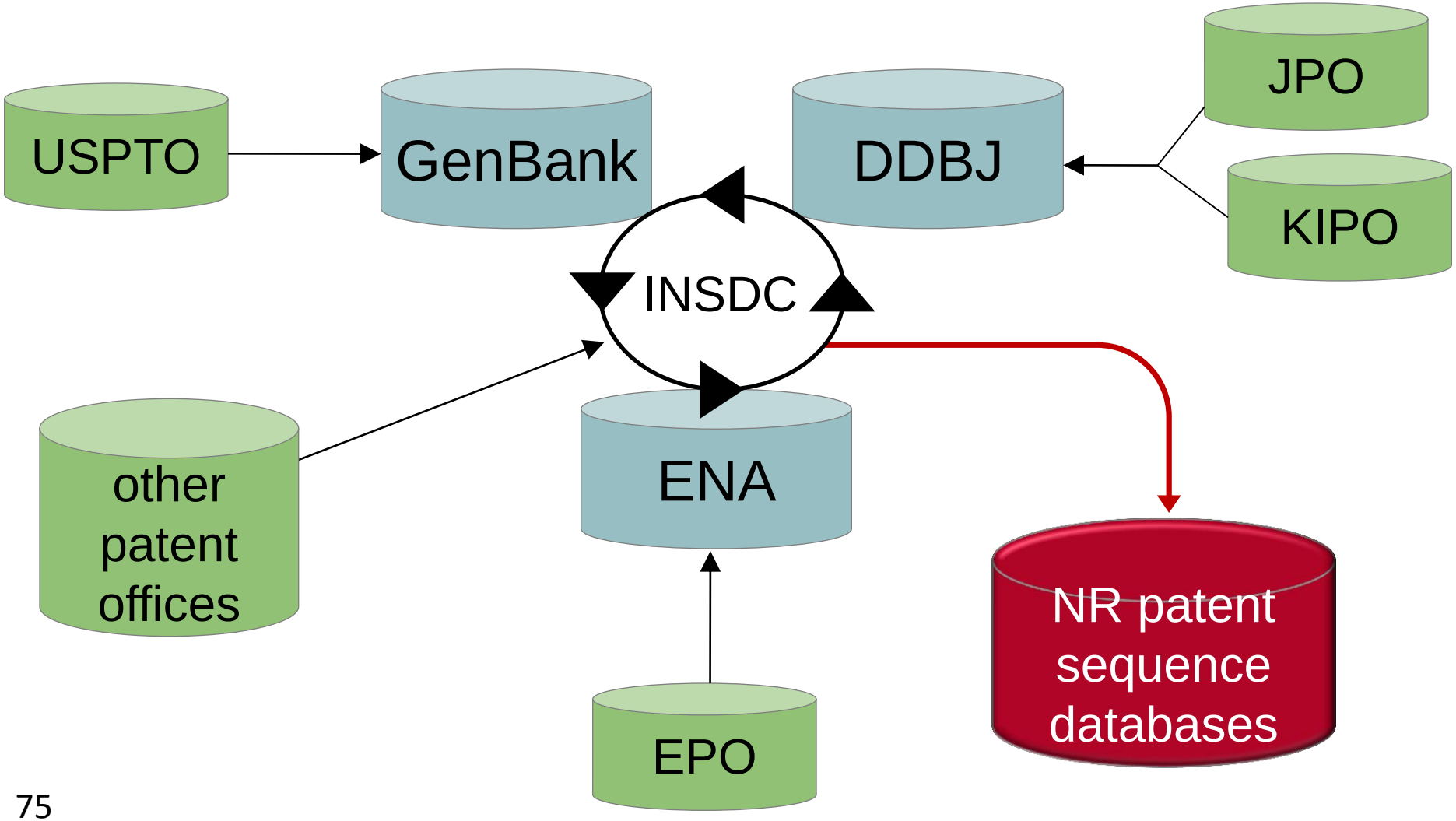


As other National Offices participate in data exchange

→ redundancy will increase

Distributing patent sequences

NR patent databases remove redundancy



EBI-EPO collaboration

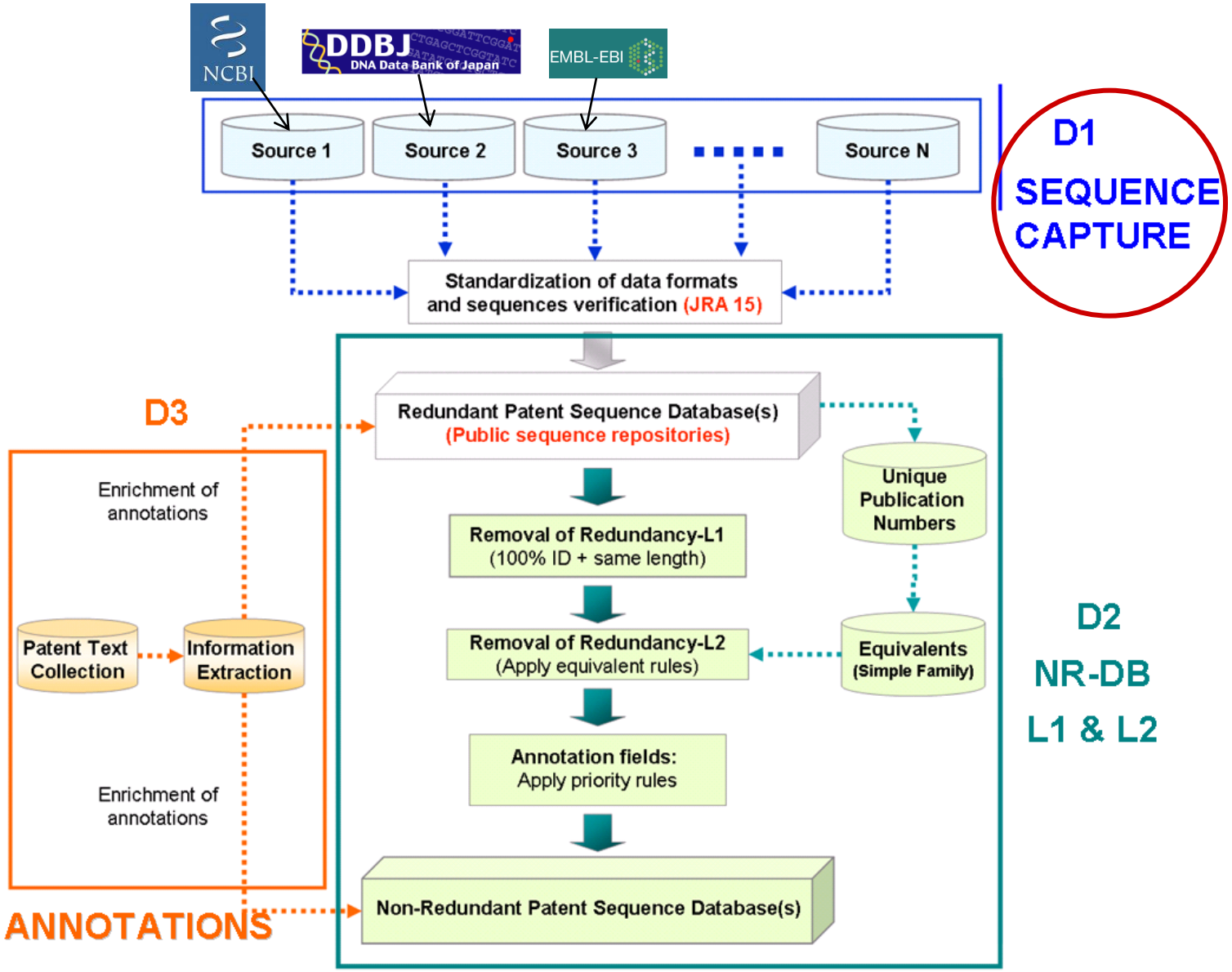
Collaboration between

EBI and **EPO**

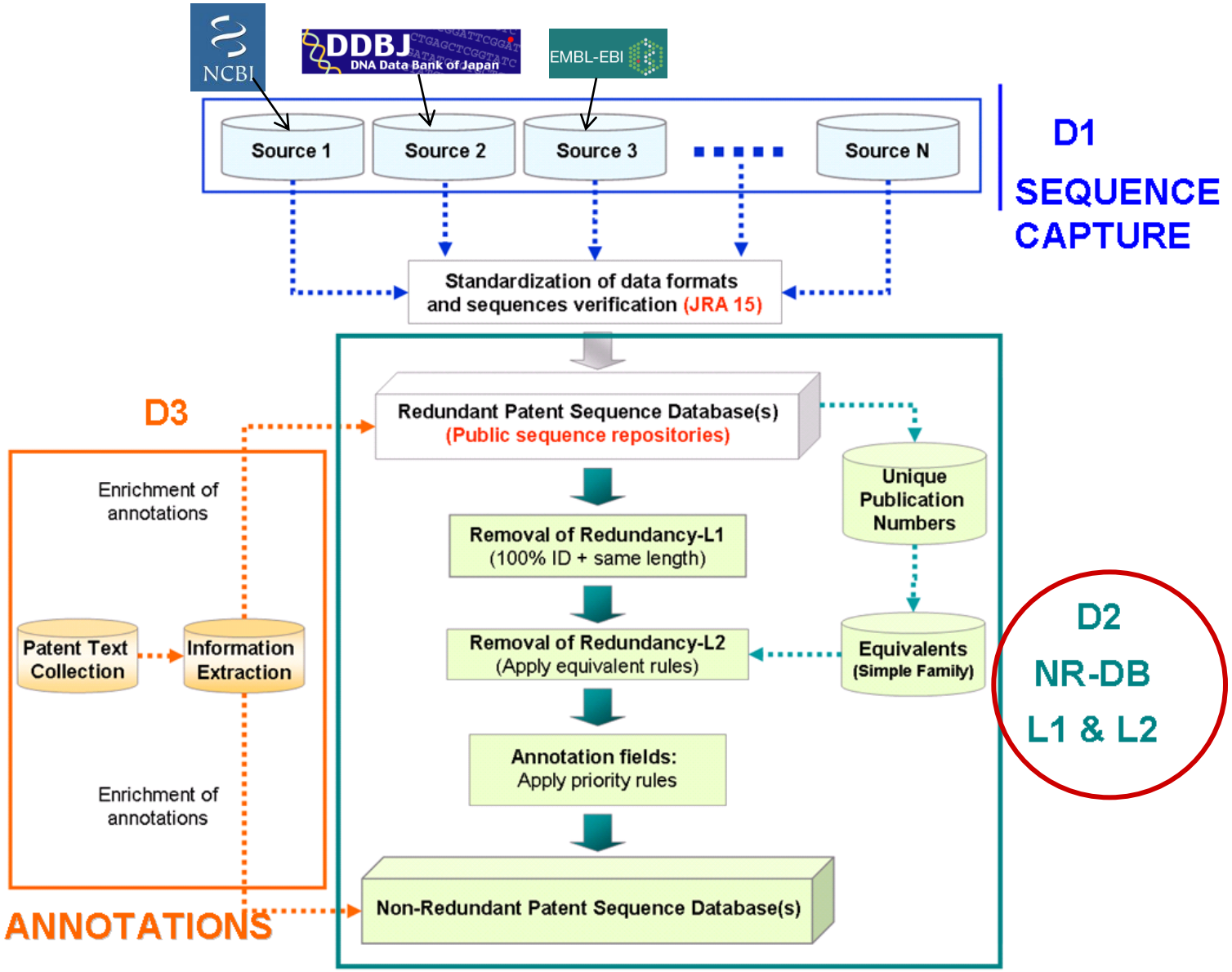
- Database development and maintenance
- Link to EBI search engines
- Link to EBI analysis tools
- Link to EBI databases

- Acquire patent sequences
- Link to patent literature
- Extract patent annotation
- Collate patent family information

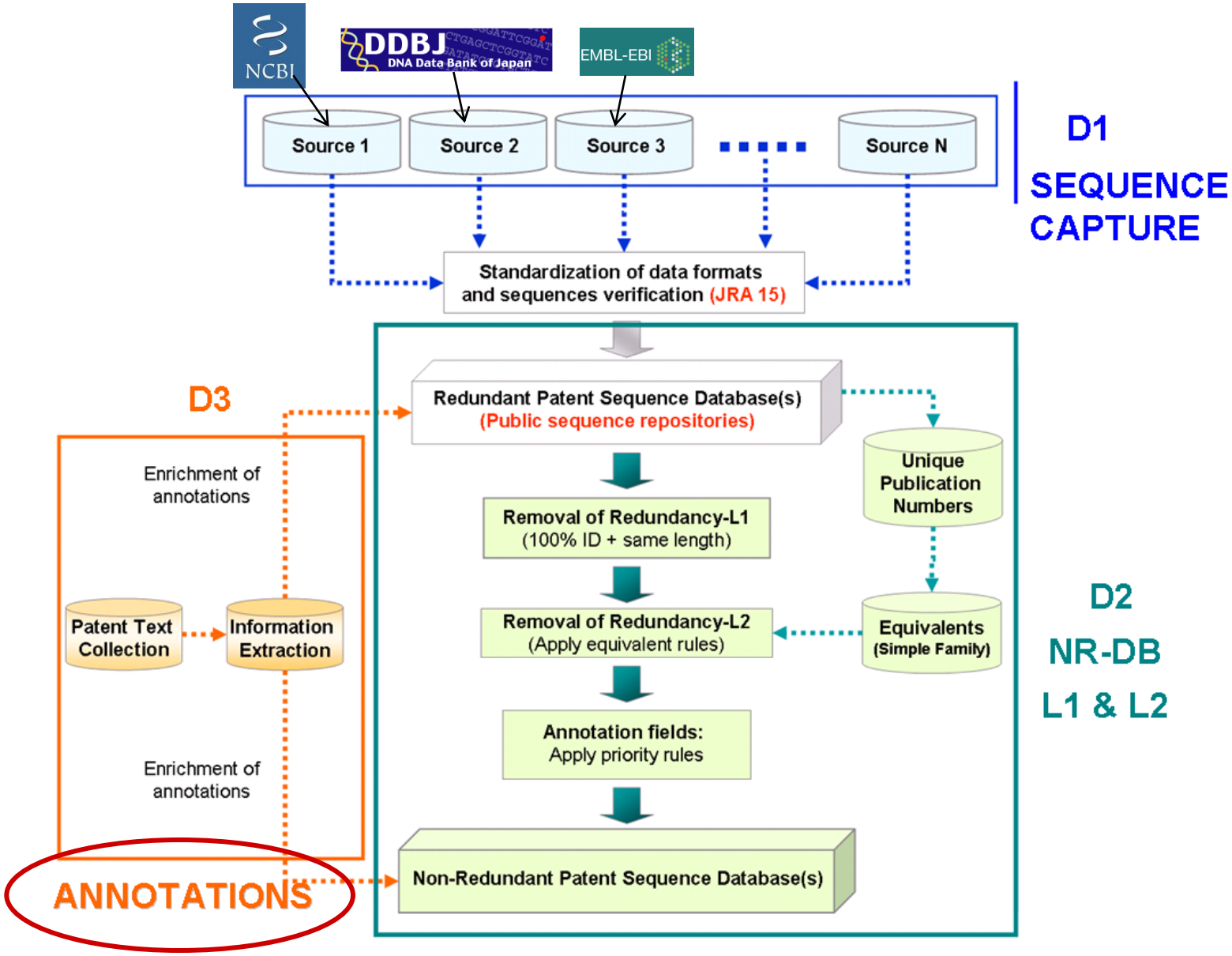
Creating a non-redundant database



Creating a non-redundant database

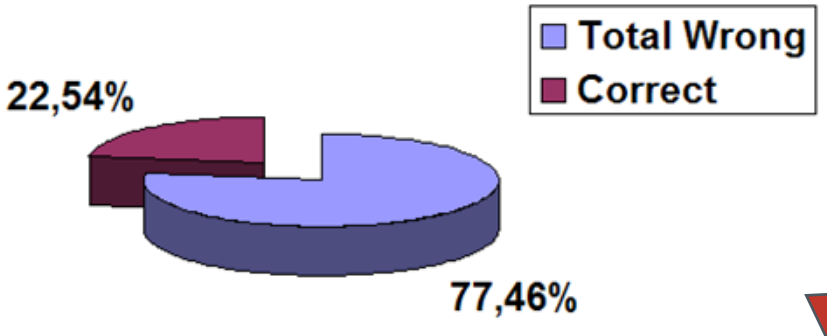


Creating a non-redundant database



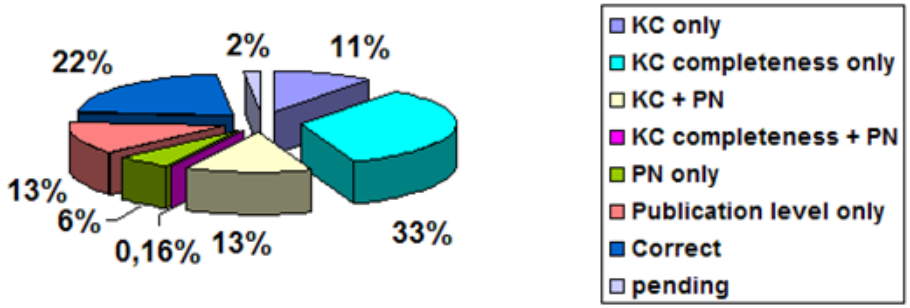
Patent data correction

A)



Correction of Publication Numbers and kind Codes

B)




```
DR   USPOP:ABZ68249;
DE   Sequence 8 from patent US 7326554.
PN   US7326554-A/8, 05-FEB-2008
PN   US2004175376 A1 09-SEP-2004
CC   First level of publication supplied by the EPO
```

```
DR   USPOP:AAO99687;
DE   Sequence 8 from patent US 6514495.
PN   US6514495-A/8, 04-FEB-2003
PN   US6514495 B1 04-FEB-2003
CC   Adapted Kind Code supplied by the EPO
```

```
DR   JPOP:BD555512;
DE   Phytase variants.
PN   JP2002507412-A/9, 12-MAR-2002
PN   JP2002507412T T 12-MAR-2002
CC   Adapted Patent Number supplied by the EPO
```

```
DR   KPOP:DI578933;
DE   Phytase Variants.
PN   KR1020007010543-A/8, 23-SEP-2000
CC   Patent Number could not be successfully verified
```

EMBL-EBI

Patent resources at EBI

Resources	Description	Access
Patent Abstracts	Patent abstracts contain the biology-related abstracts of patent applications derived from data products of the European Patent Office (EPO). Patent documents from Europe (EP), USA (US) and World (WO) are included. Patent abstracts can be accessed via the SRS, EB-eye and the literature search engine CiteXplore.	CiteXplore EB-eye SRS
Patent Chemical Compounds	Patent chemical compounds are available in the ChEBI database which is a dictionary of molecular entities focused on 'small' chemical compounds. You can search the patent chemical compounds using the ChEBI Advanced Search page by narrowing down your search to the Patent Database.	ChEBI ChEBI Advanced
Patent Sequences	Multiple sets of patent sequences are available at EBI. 1. Patent proteins cover sequences of EPO (European Patent Office) proteins, JPO (Japan Patent Office) proteins, KIPO (Korean Intellectual Property Office) proteins and USPTO (United States Patent and Trademark Office) proteins. 2. Patent nucleotides contain the patent class data in the EMBL-Bank. 3. Non-redundant patent sequences consist of 2 levels databases. Level-1 non-redundant patent sequences are 100% identical over the same length; Level-2 non-redundant patent sequences are identical and belong to a same patent family (a same invention).	Patent proteins Patent nucleotides Non-redundant patent sequences
Patent Equivalent Data	A "patent family" can be defined as all patent equivalents for a single invention. All of the published patent applications from various countries and the subsequent granted patents on an invention are commonly referred to as patent equivalents. They are not "true equivalents" in that each country may have different regulations for filing and different interpretations of the invention. It may include multiple patents in some countries because of differences in patent laws (e.g., how much new technology can be included in a single patent).	SRS

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

More information provided

Patent proteins:

EPO | USPTO | JPO | KIPO

Patent nucleotides:

ENA
(EPO, USPTO, JPO, KIPO)

Resources	Description	Access
Patent Sequences	Multiple sets of patent sequences are available at EBI. 1. Patent proteins cover sequences of EPO (European Patent Office) proteins, JPO (Japan Patent Office) proteins, KIPO (Korean Intellectual Property Office) proteins and USPTO (United States Patent and Trademark Office) proteins. 2. Patent nucleotides contain the patent class data in the EMBL-Bank. 3. Non-redundant patent sequences consist of 2 levels databases. Level-1 non-redundant patent sequences are 100% identical over the same length; Level-2 non-redundant patent sequences are identical and belong to a same patent family (a same invention).	Patent proteins Patent nucleotides Non-redundant patent sequences

- **Complete sequences** (EPO, USPTO, JPO, KIPO)
- **Non-redundant** sequence data
- Patent **family classification**
- Enriched with **patent information**

let’s look at an example…

Searching redundant databases

Protein

Example:

Search patent protein sequence

FASTA/SSEARCH/GGSEARCH/GLSEARCH - Protein

This tool provides sequence similarity searching against protein da
programs. FASTA provides a heuristic search with a protein query. F
Optimal searches are available with SSEARCH (local), GGSEARCH
local database).

Use this tool

STEP 1 - Select your databases

PROTEIN DATABASES

4 Databanks Selected — X Clear Sele

- Patents
  - EPO Patent Protein Sequences
  - JPO Patent Protein Sequences
  - KIPO Patent Protein Sequences
  - USPTO Patent Protein Sequences

STEP 2 - Enter your input sequence

Enter or paste a PROTEIN sequence in any supported forn

```
MASLGQILFWSIISIIIILAGAIALIIGFGISGRHSITVTTVASAGNIGEDGILSCTFEP
DIKLSDIVIQWLKEGVLGLVHEFKEGKDELSEQDEMFRGRIAVFADQVIVGNASLRLKNV
QLTDAGTYKCYIITSKGKGNANLEYKTGAFSMPEVNVDYNASSETLRCEAPRWFPQPTVV
WASQVDQGANFSEVSNTSFELNSENVTMKVVSVLYNVTINNTYSCMIENDIAKATGDIKV
TESEIKRRSHLQLLNSKASLCVSSFFAISWALLPLSPYLMLK
```

or Upload a file: Browse...

STEP 3 - Set your parameters

PROGRAM

FASTA

The default settings will fulfill the needs of most users and, for that reason, are not visible.

More options... *(Click here, if you want to view or change the default settings.)*

STEP 4 - Submit your job

Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

Submit

4 Databanks Selected — X Clear Selection

- UniProt Knowledgebase
- UniProtKB/Swiss-Prot
- UniProtKB/Swiss-Prot isoforms
- UniProtKB/TrEMBL
- UniProtKB Taxonomic Subsets
- UniProt Clusters
- Patents
  - EPO Patent Protein Sequences
  - JPO Patent Protein Sequences
  - KIPO Patent Protein Sequences
  - USPTO Patent Protein Sequences
  - NR Patent Proteins Level-1
  - NR Patent Proteins Level-2
- Structure
- Other Protein Databases

Patent proteins

http://www.ebi.ac.uk/Tools/sss/

Results from redundant databases

FASTA Results

Summary Table | Tool Output | Visual Output | Submission Details | Submit Another Job

Alignments

Selection: Show Annotations | Hide Annotations | Show Alignments | Hide Alignments

Download in fasta format

Clear Selection | Select All | Invert Selection

Align.	DB:ID	Source	Length	Score	Identities	Positives	E()
1	USPOP:ABH98446	ABH98446 Sequence 291 from patent US 7070955.	282	1808	100.0	100.0	3.5E-148
2	USPOP:ABP26275	ABP26275 Sequence 60 from patent US 7193047.	282	1808	100.0	100.0	3.5E-148
3	USPOP:ABY13906	ABY13906 Sequence 218 from patent US 7304145.	282	1808	100.0	100.0	3.5E-148
4	USPOP:ABP13991	ABP13991 Sequence 60 from patent US 7186801.	282	1808	100.0	100.0	3.5E-148
5	USPOP:ACQ12151	ACQ12151 Sequence 60 from patent US 7507404.	282	1808	100.0	100.0	3.5E-148
6	USPOP:ABX91442	ABX91442 Sequence 60 from patent US 7285624.	282	1808	100.0	100.0	3.5E-148
7	USPOP:ACH00887	ACH00887 Sequence 60 from patent US 7390876.	282	1808	100.0	100.0	3.5E-148
8	USPOP:ACK09334	ACK09334 Sequence 60 from patent US 7435798.	282	1808	100.0	100.0	3.5E-148
9	USPOP:ABY00585	ABY00585 Sequence 60 from patent US 7294690.	282	1808	100.0	100.0	3.5E-148
10	EPOP:HB433204	HB433204 Sequence 56 from Patent WO2009073533.	282	1808	100.0	100.0	3.5E-148
11	USPOP:ABC11735	ABC11735 Sequence 393 from patent US 6962980.	282	1808	100.0	100.0	3.5E-148
12	USPOP:ABU37148	ABU37148 Sequence 60 from patent US 7244817.	282	1808	100.0	100.0	3.5E-148
13	USPOP:ACE15168	ACE15168 Sequence 218 from patent US 7365156.	282	1808	100.0	100.0	3.5E-148
14	USPOP:ABZ22951	ABZ22951 Sequence 218 from patent US 7312316.	282	1808	100.0	100.0	3.5E-148
15	JPOP:BD750823	BD750823 JP 2004507204-A/116: SECRETED AND TRANSMEMBRANE POLYPEPTIDES AND NUCLEIC ACIDS ENCODING THE SAME.	282	1808	100.0	100.0	3.5E-148
16	USPOP:ADA48774	ADA48774 Sequence 5 from pa[illegible] 7622565.					
17	USPOP:ABU37060	ABU37060 Sequence 291 from [illegible]t US 7244816.					
⋮							
264	USPOP:ABJ18761	ABJ18761 Sequence 2[illegible]					
265	USPOP:ACW23786	ACW23786 Sequence 21[illegible]					
266	USPOP:ABX97150	ABX97150 Sequence 218 from patent US 7291700.	282	1808	100.0	100.0	3.5E-148
267	USPOP:ABN22111	ABN22111 Sequence 218 from patent US 7153939.	282	1808	100.0	100.0	3.5E-148
268	EPOP:CQ875685	CQ875685 Sequence 8 from Patent EP1445317.	282	1808	100.0	100.0	3.5E-148

>260 identical results

→ too much to analyze

NR patent sequence databases

LEVEL1 NR patent sequence database

removes redundancy

→fewer results to analyze, less chance

of missing important results

Searching patent sequences

NR patent Level-1

Example:

Search patent protein sequence

FASTA/SSEARCH/GGSEARCH/GLSEARCH - Protein

This tool provides sequence similarity searching against protein da
programs. FASTA provides a heuristic search with a protein query. F
Optimal searches are available with SSEARCH (local), GGSEARCH
local database).

Use this tool

STEP 1 - Select your databases

PROTEIN DATABASES

1 Databank Selected — X Clear Selec

- KIPO Patent Protein Sequences
- USPTO Patent Protein Sequences
- NR Patent Proteins Level-1
- NR Patent Proteins Level-2

Structure

Other Protein Databases

STEP 2 - Enter your input sequence

Enter or paste a PROTEIN sequence in any supported form

```
MASLGQILFWSIISIIIILAGAIALIIGFGISGRHSITVTTVASAGNIGEDGILSCTFEP
DIKLSDIVIQWLKEGVLGLVHEFKEGKDELSEQDEMFRGRIAVFADQVIVGNASLRLKNV
QLTDAGTYKCYIITSKGKGNANLEYKTGAFSMPEVNVDYNASSETLRCEAPRWFPQPTVV
WASQVDQGANFSEVSNTSFELNSENVTMKVVSVLYNVTINNTYSCMIENDIAKATGDIKV
TESEIKRRSHLQLLNSKASLCVSSFFAISWALLPLSPYLMLK
```

or Upload a file: Browse...

STEP 3 - Set your parameters

PROGRAM

FASTA

The default settings will fulfill the needs of most users and, for that reason, are not visible.

More options... *(Click here, if you want to view or change the default settings.)*

STEP 4 - Submit your job

Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

Submit

1 Databanks Selected — X Clear Selection

- UniProt Knowledgebase
- UniProtKB/Swiss-Prot
- UniProtKB/Swiss-Prot isoforms
- UniProtKB/TrEMBL

UniProtKB Taxonomic Subsets

UniProt Clusters

Patents

- EPO Patent Protein Sequences
- JPO Patent Protein Sequences
- KIPO Patent Protein Sequences
- USPTO Patent Protein Sequences
- NR Patent Proteins Level-1
- NR Patent Proteins Level-2

Structure

Other Protein Databases

NR patent level-1

http://www.ebi.ac.uk/Tools/sss/

Results from level-1 database

FASTA Results

Summary Table | Tool Output | Visual Output | Submission Details | Submit Another Job

Alignments

Selection: Show Annotations | Hide Annotations | Show Alignments | Hide Alignments

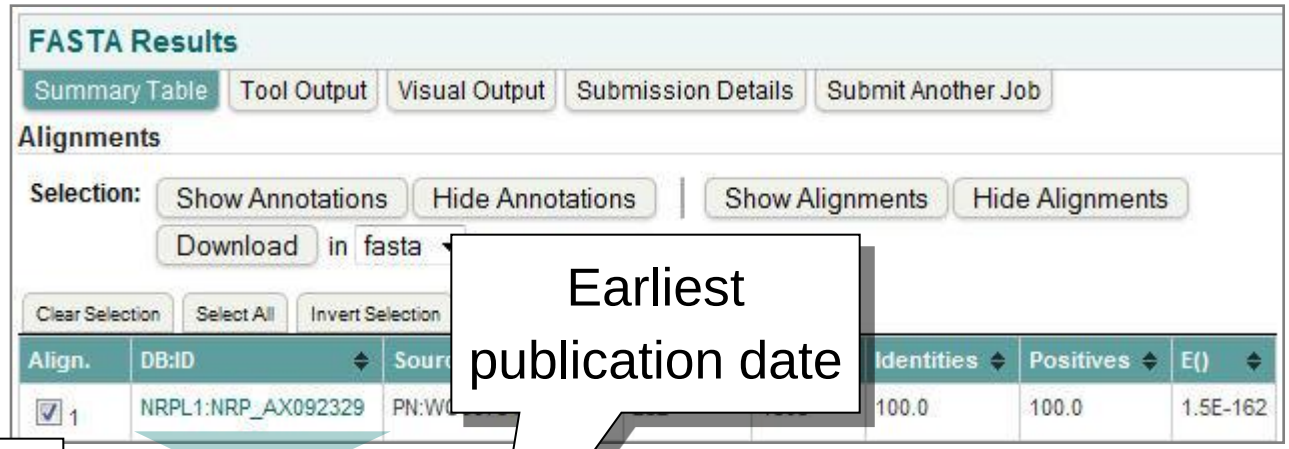
Download in fasta format

Clear Selection | Select All | Invert Selection

Align.	DB:ID	Source	Length	Score	Identities	Positives	E()
1	NRPL1:NRP_AX092329	PN:WO0073454 A1	282	1808	100.0	100.0	4.1E-162
2	NRPL1:NRP_HI379470	PN:EP2104513 A2	290	1808	100.0	100.0	
3	NRPL1:NRP_AX156352	PN:WO0140269 A2	309	1808	100.0	100.0	
4	NRPL1:NRP_JA171544	PN:US2003124128 A1	282	1801	99.6	99.6	1.9E-161
5	NRPL1:NRP_ADA40459	PN:JP2007504280T T	306	1801	99.6	99.6	2.2E-161
6	NRPL1:NRP_ADA40466	PN:JP2007504280T T	309	1801	99.6	99.6	2.2E-161
7	NRPL1:NRP_HI962287	PN:US2005238649 A1	282	1798	99.6	99.6	3.6E-161
8	NRPL1:NRP_JA171549	PN:US2003124140 A1	282	1789	98.6	100.0	2.6E-160
9	NRPL1:NRP_JA171551	PN:WO2011026122 A2	282	1780	98.2	99.6	1.9E-159
10	NRPL1:NRP_JA171547	PN:WO2011026122 A2	278	1727	99.3	100.0	1.9E-154
11	NRPL1:NRP_AX403050	PN:WO0202624 A2	258	1665	100.0	100.0	1.3E-148
12	NRPL1:NRP_JA171542	PN:WO2011026122 A2	274	1656	94.5	97.8	1.0E-147
13	NRPL1:NRP_JA171586	PN:WO2011026122 A2	259	1653	100.0	100.0	1.8E-147
14	NRPL1:NRP_JA171543	PN:WO2011026122 A2	255	1650	100.0	100.0	3.4E-147

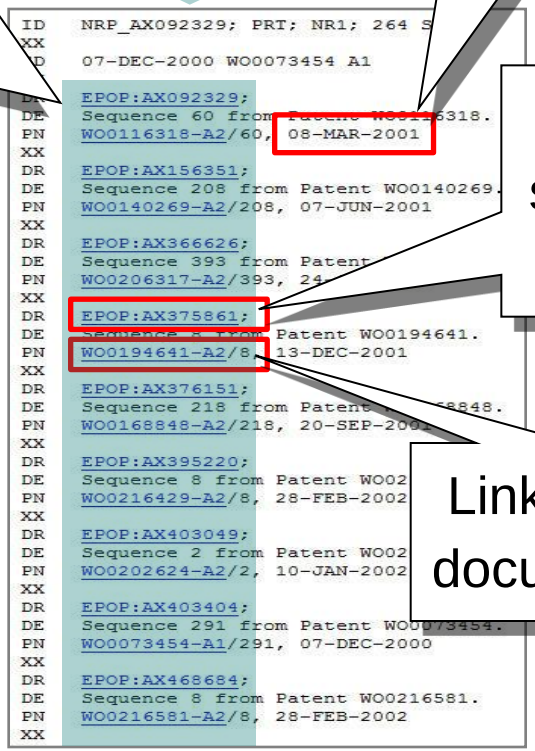
Each hit unique

Results from level-1 database



Earliest publication date

List of **all** patents containing the sequence



Link to sequence entry

Link to patent documentation

Patent sequence record in NRNL1

EMBL-EBI EB-eye Search All Databases Enter Text Here Go Reset Advanced Search Give us feedback

Databases | Tools | EBI Groups | Training | Industry | About Us | Help | Site Index

Quick Search | Library Page | Query Form | Tools | Results | Projects | Views | Databanks | HELP

Text Entry | NRL1_Entry_View | Related Data

Reset

Entry Information

Entry from: Patent DNA NRL1

Entry Options

Launch analysis tool: NCBI BLASTN

Launch

1. **NRNL1:NRN_AX429748 08-MAY-2002**

Sequence Entries:

Accession	Description	Patent	Pub. Date
EM_PAT:AX429748	Sequence 13 from Patent EP1203827.	EP1203827-A2/13	08-MAY-2002
EM_PAT:CQ891794	Sequence 13 from Patent EP1469084.	EP1469084-A2/13	20-OCT-2004
EM_PAT:BD183760	Chemical compounds.	JP2002345489-A/13	03-DEC-2002

Sequence:

Sequence Length: 100; Checksum: eb9ae8a024ff652a05c71b40e3695af0;

```
>EM_PAT:AX429748
gcgcgatggcgaagagggtcctgcactttgacgcgcctggtgagggagcgstgctcttcg
cagcgctcctggtgatgctccccaaatttcggggaccggc
```

Patents containing 100% identical sequence

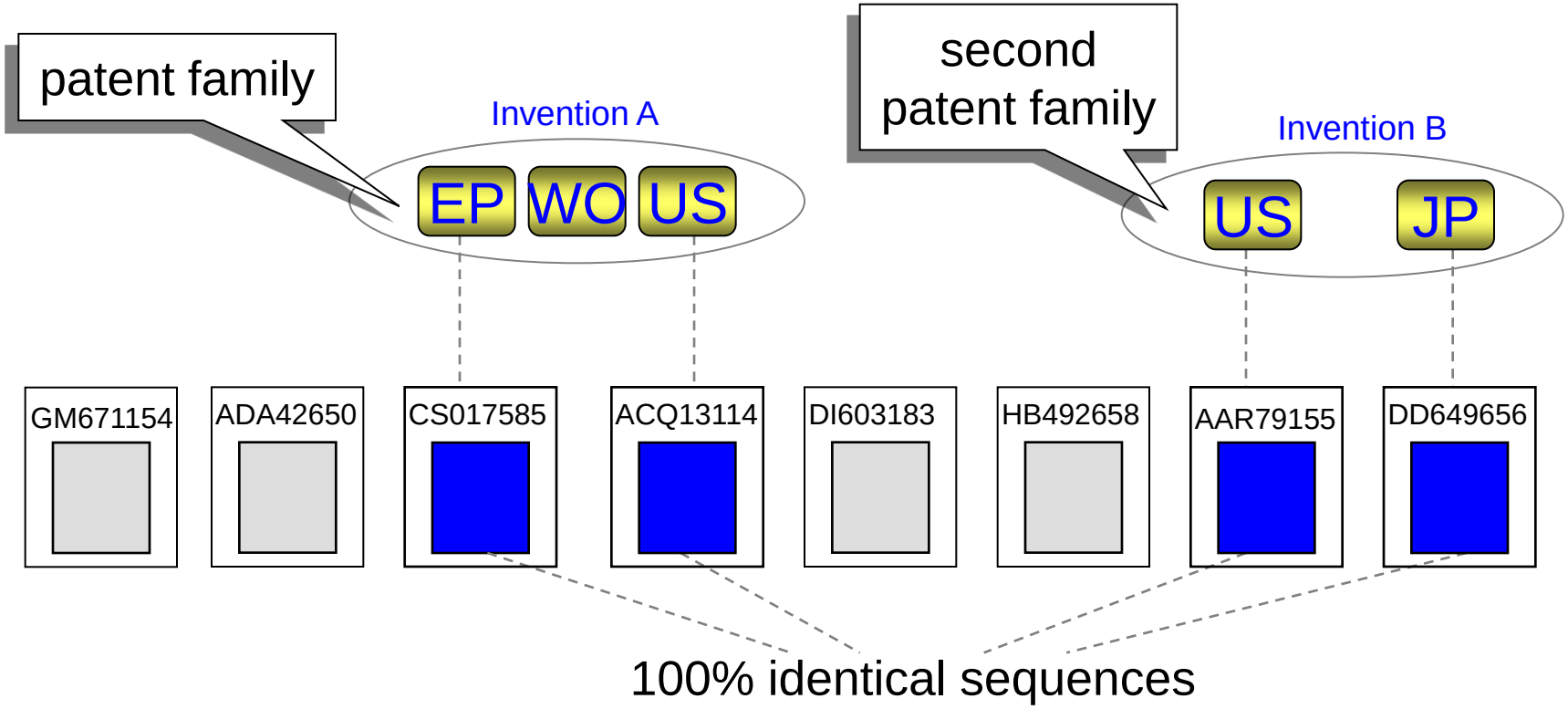
Sequence

Patent families

Simple **Patent Family** is a group of patents that relate to the same invention, and are based on the same originating application

- They arise when an invention is patented in multiple countries
- Grouping patents into families reduces multi-national results down to a representative member

Patent families



Same sequence can appear multiple times in a database due to:

- Same invention filed multiple times in different offices *(same patent family)*
- Different inventors use the same sequence in different contexts *(different patent families)*

NR patent sequence databases

LEVEL2 NR patent sequence database

groups identical sequences by ***patent family***

→ provides **earliest priority date** for family

Searching patent sequences

NR patent Level-2

Example:

Search patent protein sequence

FASTA/SSEARCH/GGSEARCH/GLSEARCH - Prote

This tool provides sequence similarity searching against protein d
programs. FASTA provides a heuristic search with a protein query
Optimal searches are available with SSEARCH (local), GGSEARC
local database).

Use this tool

STEP 1 - Select your databases

PROTEIN DATABASES

1 Databank Selected — X Clear Sel

- JPO Patent Protein Sequences
- KIPO Patent Protein Sequences
- USPTO Patent Protein Sequences
- NR Patent Proteins Level-1
- NR Patent Proteins Level-2
- Structure

STEP 2 - Enter your input sequence

Enter or paste a PROTEIN sequence in any supported fo

```
MASLGQILFWSIISIIIILAGAIALIIGFGISGRHSITVTTVASAGNIGEDGILSCTFEP
DIKLSDIVIQWLKEGVLGLVHEFKEGKDELSEQDEMFRGRIAVFADQVIVGNASLRLKNV
QLTDAGTYKCYIITSKGKGNANLEYKTGAFSMPEVNVDYNASSETLRCEAPRWFPQPTVV
WASQVDQGANFSEVSNTSFELNSENVTMKVVSVLYNVTINNTYSCMIENDIAKATGDIKV
TESEIKRRSHLQLLNSKASLCVSSFFAISWALLPLSPYLMLK
```

or Upload a file: Browse...

STEP 3 - Set your parameters

PROGRAM

FASTA

The default settings will fulfill the needs of most users and, for that reason, are not visible.

More options... *(Click here, if you want to view or change the default settings.)*

STEP 4 - Submit your job

Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

Submit

1 Databanks Selected — X Clear Selection

- UniProt Knowledgebase
- UniProtKB/Swiss-Prot
- UniProtKB/Swiss-Prot isoforms
- UniProtKB/TrEMBL
- UniProtKB Taxonomic Subsets
- UniProt Clusters
- Patents
  - EPO Patent Protein Sequences
  - JPO Patent Protein Sequences
  - KIPO Patent Protein Sequences
  - USPTO Patent Protein Sequences
  - NR Patent Proteins Level-1
  - NR Patent Proteins Level-2
- Structure
- Other Protein Databases

NR patent level-2

http://www.ebi.ac.uk/Tools/sss/

Results from level-2 database

FASTA Results

Summary Table | Tool Output | Visual Output | Submission Details | Submit Another Job

Alignments

Selection: Show Annotations | Hide Annotations | Show Alignments | Hide Alignments

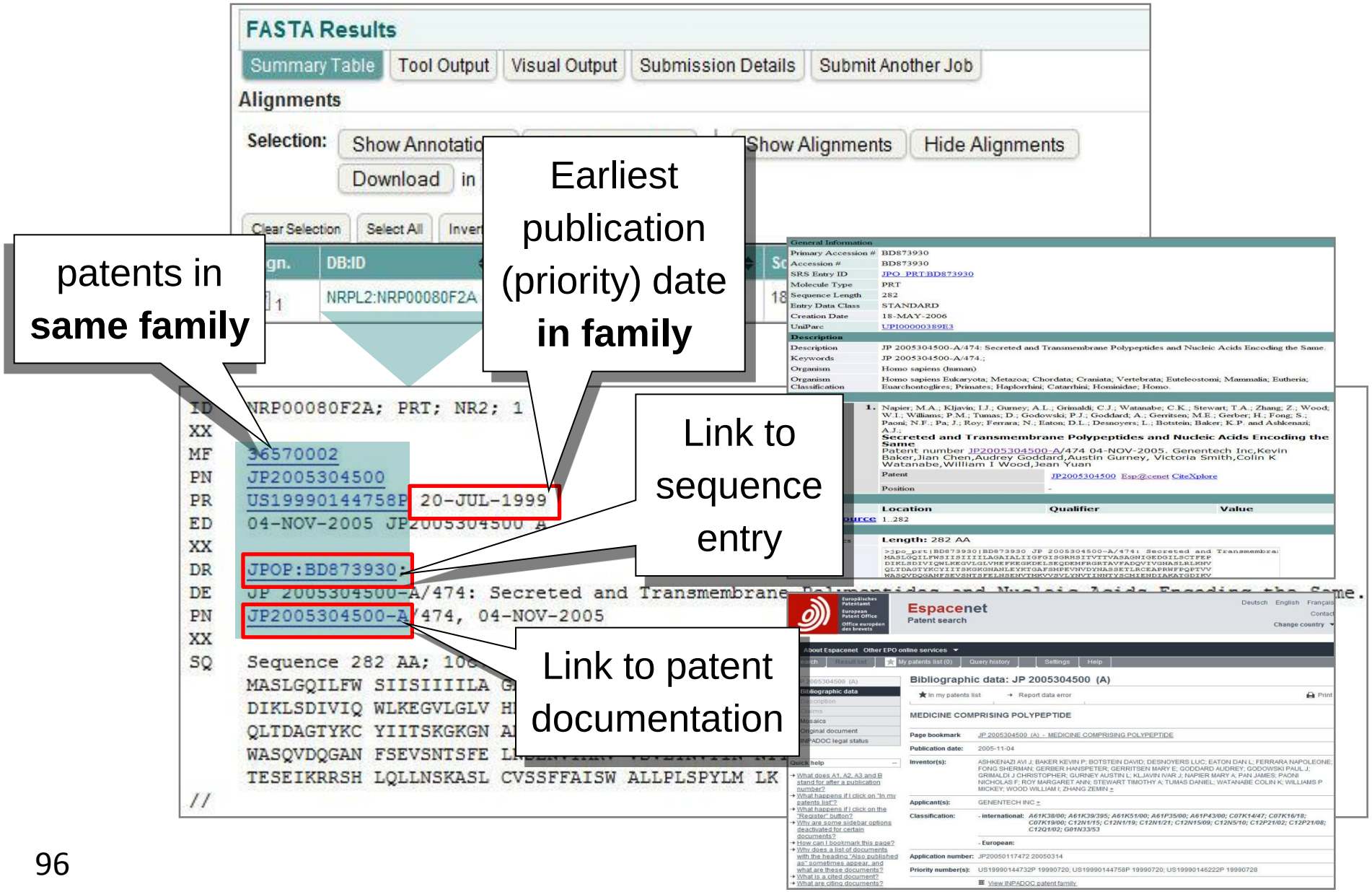
Download in fasta format

Clear Selection | Select All | Invert Selection

Align.	DB:ID	Source	Length	Score	Identities	Positives	E()
1	NRPL2:NRP00080F2A	PN:JP2005304500 A	282	1808	100.0	100.0	4.7E-157
2	NRPL2:NRP00080F2B	PN:US7056736 B2	282	1808	100.0		
3	NRPL2:NRP00080F88	PN:US7339034 B2	282	1808	100.0		
4	NRPL2:NRP00080F89	PN:US7345145 B2	282	1808	100.0	100.0	4.7E-157
5	NRPL2:NRP00080F2C	PN:US7060812 B2	282	1808	100.0	100.0	4.7E-157
6	NRPL2:NRP00080F2D	PN:US2003181679 A1	282	1808	100.0	100.0	4.7E-157
7	NRPL2:NRP00080F2E	PN:US7083978 B2	282	1808	100.0	100.0	4.7E-157
8	NRPL2:NRP00080F2F	PN:US7084255 B2	282	1808	100.0	100.0	4.7E-157
9	NRPL2:NRP00080F30	PN:US7074897 B2	282	1808	100.0	100.0	4.7E-157
10	NRPL2:NRP00080F8E	PN:US7358340 B2	282	1808	100.0	100.0	4.7E-157
11	NRPL2:NRP00080F8F	PN:US7365156 B2	282	1808	100.0	100.0	4.7E-157
12	NRPL2:NRP00080F31	PN:WO2006098887 A2	282	1808	100.0	100.0	4.7E-157
13	NRPL2:NRP00080F32	PN:US7101687 B2	282	1808	100.0	100.0	4.7E-157
14	NRPL2:NRP00080F92	PN:US7378486 B2	282	1808	100.0	100.0	4.7E-157

Each hit = one family

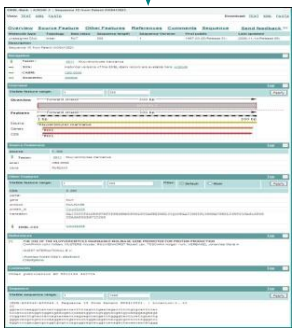
Results from level-2 database



Patent sequence record in NRNL2

Patent equivalents

Sequence record in ENA



Sequence

EMBL-EBI | EB-eye Search | All Databases | Enter Text Here | Go | Reset | Advanced Search | Give us feedback

Databases | Tools | EBI Groups | Training | Industry | About Us | Help | Site Index

Quick Search | Library Page | Query Form | Tools | Results | Projects | Views | Databanks | HELP

Text Entry | NRL2_Entry_View | Related Data

Reset

Previous Entry | Entry **44 of 251** from Query 1 | Next Entry

Entry Information

Entry from: Patent DNA NRL2

Entry Options

Launch analysis tool: NCBI BLASTN — Launch

Link to related information: Link

Save entry: Save

View: Printer Friendly

1. NRNL2:NRN00197A41 28-JUN-2001

Patent Family:

Family	Priority Number	Priority date
26867389	US19990171746P	22-DEC-1999

Sequence Entries:

Accession	Description	Patent	Date
EM_PAT:AX181616;	Sequence 404 from Patent WO0146262.	WO0146262-A2/404	28-JUN-2001
EM_PAT:AX181640;	Sequence 428 from Patent WO0146262.	WO0146262-A2/428	28-JUN-2001

Features:

Key	Location	Qualifier	Value
source	1..649	organism	Mus musculus domesticus
		mol_type	unassigned DNA
		note	Taxon = 10092; clone = OR2M; Accession DDBJ/EMBL/GenBank = AF0739
		db_xref	taxon:10092
cds	<2..>649	product	olfactory receptor; region between transmembrane domain s
		protein_id	CAC50154.1
		translation	>CAC50154.1 FTDLCFSSVTMPKLLQNMQSQVPSIPYAGCLTQMYFFLFF FPLHYTSIMSPRLCVSLVLLSWLLTMSHSMLHTLLLTRLS ACSDIHINELVILIIGGLVVILPFLLVTVSYARIISSILK VSLFYGTIIGLYLCPSANNSTLKDTVMSLMYTVVTP
		protein_id	CAC50071.1 {AX181640}
misc_feature	1..649	note	AAMTv0.9:AX181640
		note	Taxon = 10092; clone = OR9M; Accession DDBJ/EMBL/GenBank = AF0739

Sequence:

Sequence Length: 649; Checksum: 21041042261dcdce529ca6533d359fff;

```
>NRNL2:NRN00197A41 PN:WO0146262 A2
cttcactgacctctgcttttcctctgtcacaatgcccaagttgctgcagaacatgcagag
ccaagttccttcaatcccctatgcaggctgcctgacacaaatgtacttcttttgttttt
tggagatcttgagagcttcctccttgtggccatggcctatgaccgatatgtagccatctg
cttccctcttcattacaccagcattatgagccccaggctctgtgtgagtcttgtgctgct
gtcctggttgctgaccatgtcccattccatgctgcacactttgctcttaactaggttgtc
tttctgtgaaaacaatgtgatcccccatttttctgtgatctgtctgctctgctgaagct
ggcctgctctgatattcacattaatgaattggtgatattgatcataggagggcttgttgt
tatacttccatttctactcgtcacagtgtcttatgcacgcatcatctcctccattctcaa
ggtcccttcaactcgaggcatccacaaggtcttctccacttgtggttctcacctgtctgt
ggtgtcactgttctatgggacaattattggcctctacttatgtccatctgctaataactc
tactctaaaggacactgtcatgtctctgatgtacactgtggtaactccc
```

Priority number and date

Patent literature

Translation

Non-redundant patent databases

	Patent nucleotides	Patent proteins	
Level-1	**NRNL1** (Non-redundant nucleotide level-1)	**NRPL1** (Non-redundant protein level-1)	➢ **Groups together 100% identical patent sequences**
Level-2	**NRNL2** (Non-redundant nucleotide level-2)	**NRPL2** (Non-redundant protein level-2)	➢ **Groups together identical sequences by patent family**

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

Non-redundant patent databases

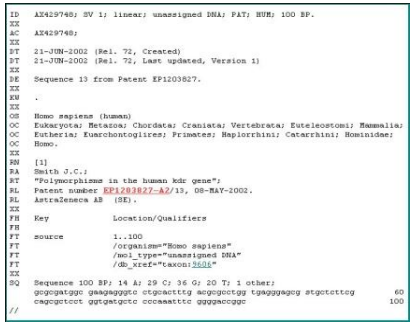
ENA
(redundant)

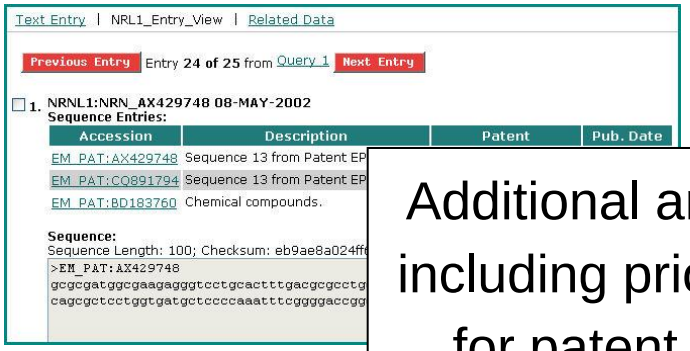
Remove
sequence redundancy

Level-1 NR

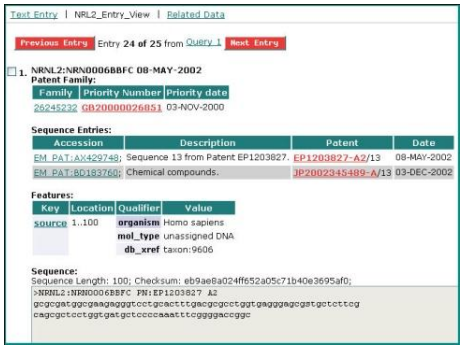
Group by
patent families

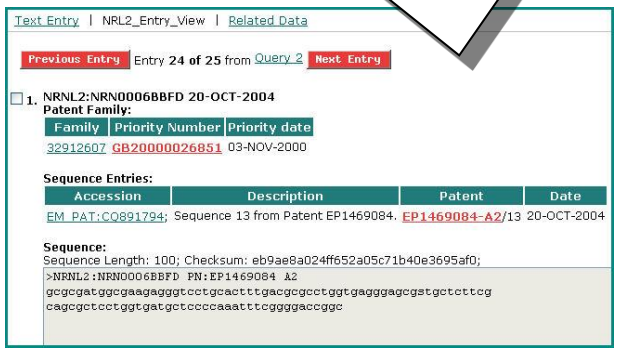
Level-2 NR





Additional annotation,
including priority dates
for patent families





Patent sequence records at EBI

Nucleotide

ENA ~23.1 M PAT sequences

NRNL1 ~11.9 M sequences

NRNL2 ~15.0 M sequences

Protein

Patent Proteins ~6.3 M PRT sequences

NRPL1 ~2.5 M sequences

NRPL2 ~3.8 M sequences

NR Patent Sequence Databases

- Sequence searches against a non-redundant database is faster and avoids overlooking data

- These databases are the first non-redundant collection take account of both sequence and family concepts

- Publication corrections significantly increase data quality

- Collation of biological features in a single record enables understanding of biological concept in which the sequence is being used

2) The Toolbox

EBI search

SRS advanced search

Sequence search

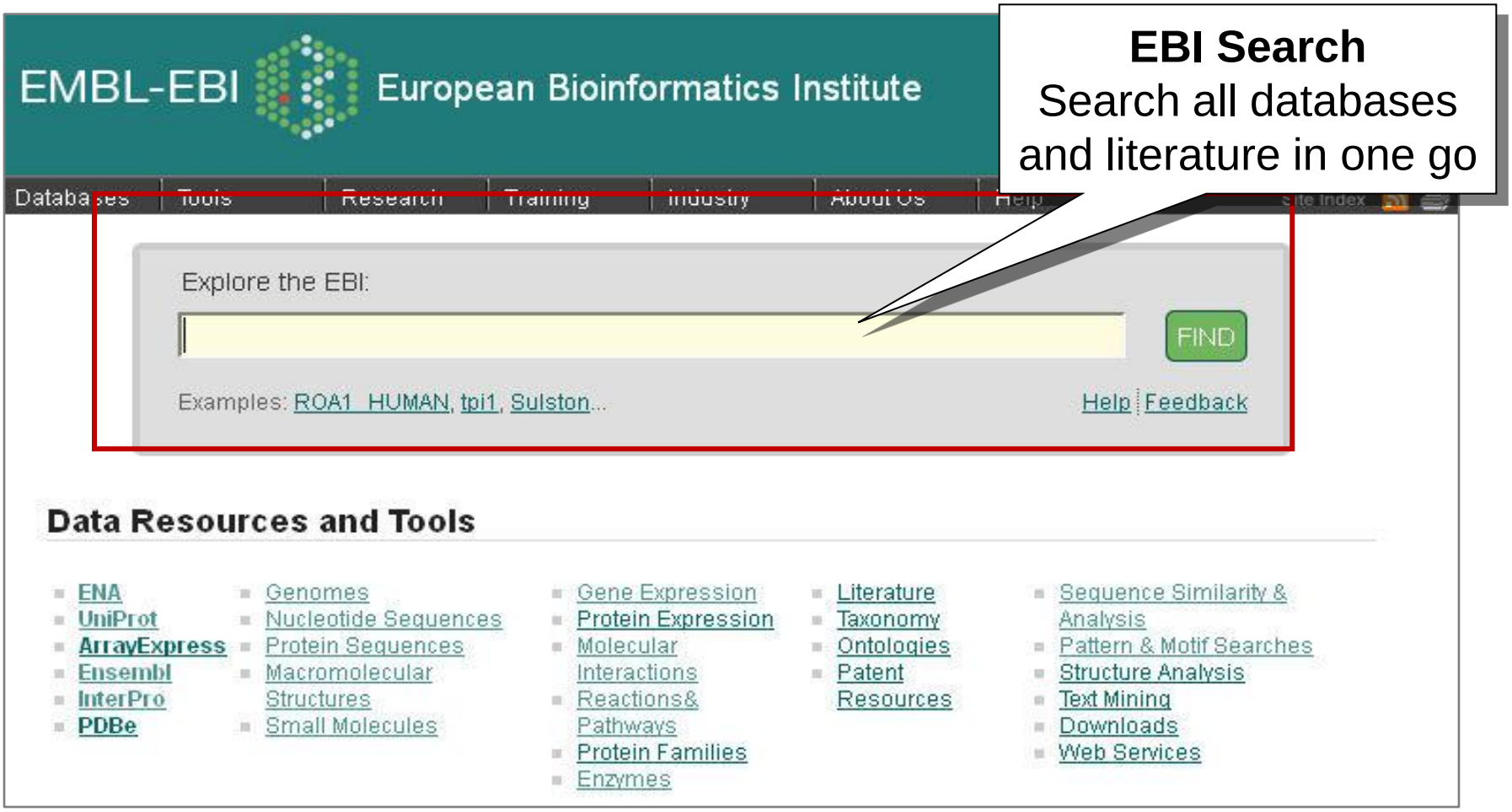
EBI search

SRS advanced search

Sequence search

EBI search

EBI-Search accessible from any EBI page

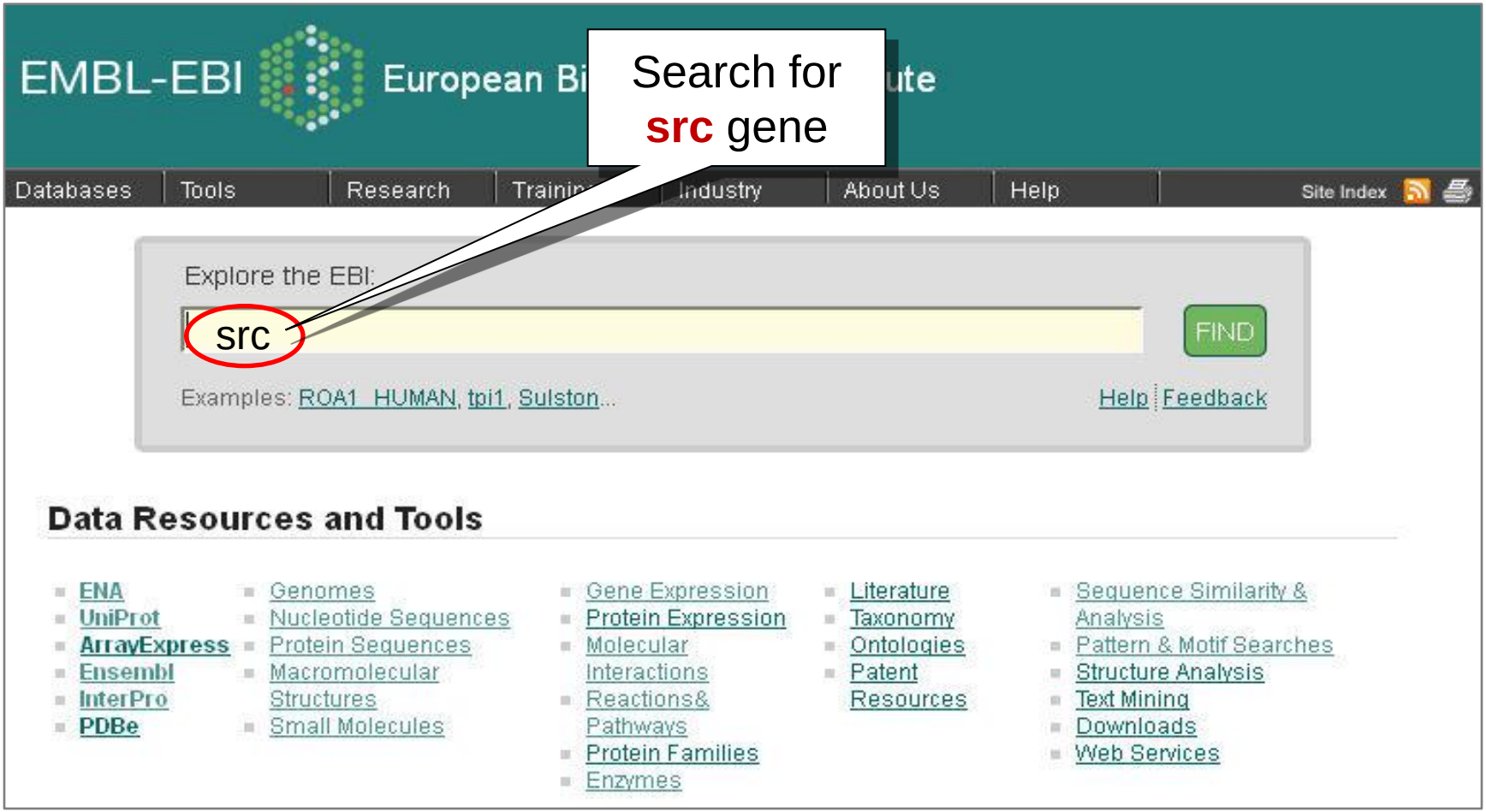


EBI Search
Search all databases and literature in one go

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

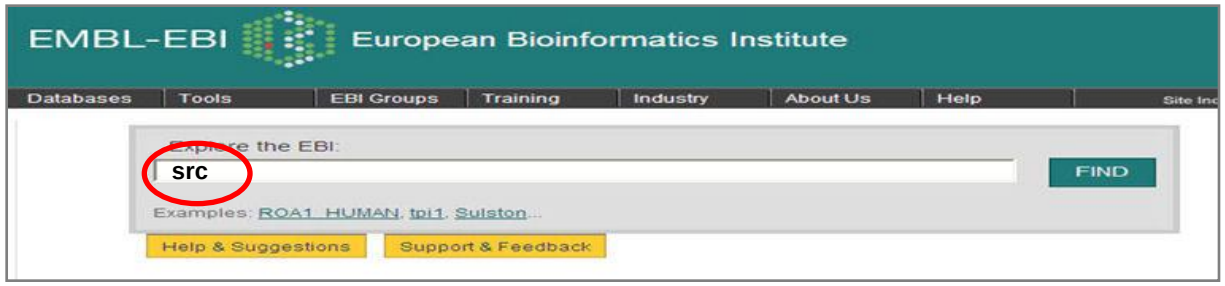
EBI search

EBI-Search by **gene name**



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

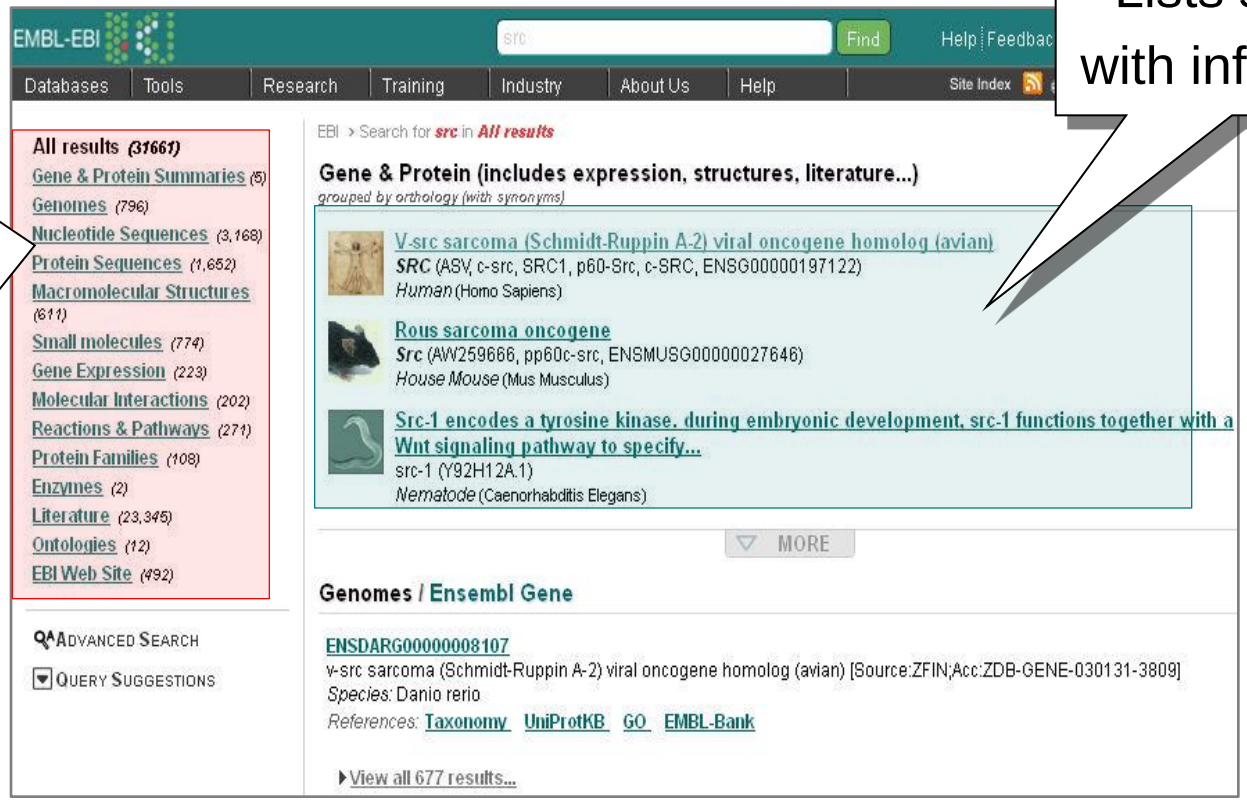
EBI search by gene name



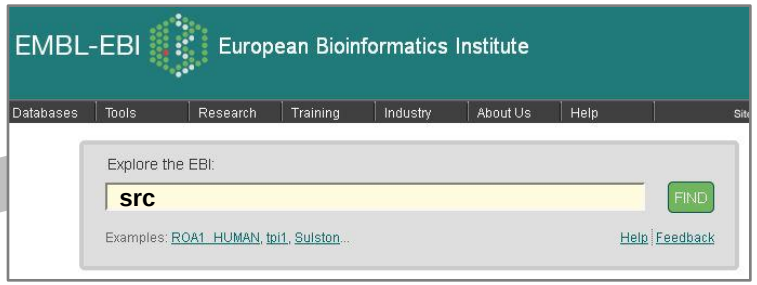
query

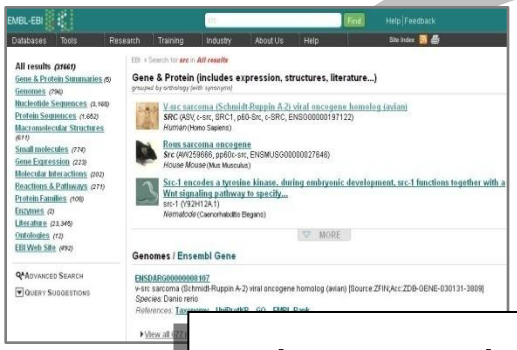
Lists species with information

Lists relevant entries in <u>all</u> EBI resources



EBI search by gene name

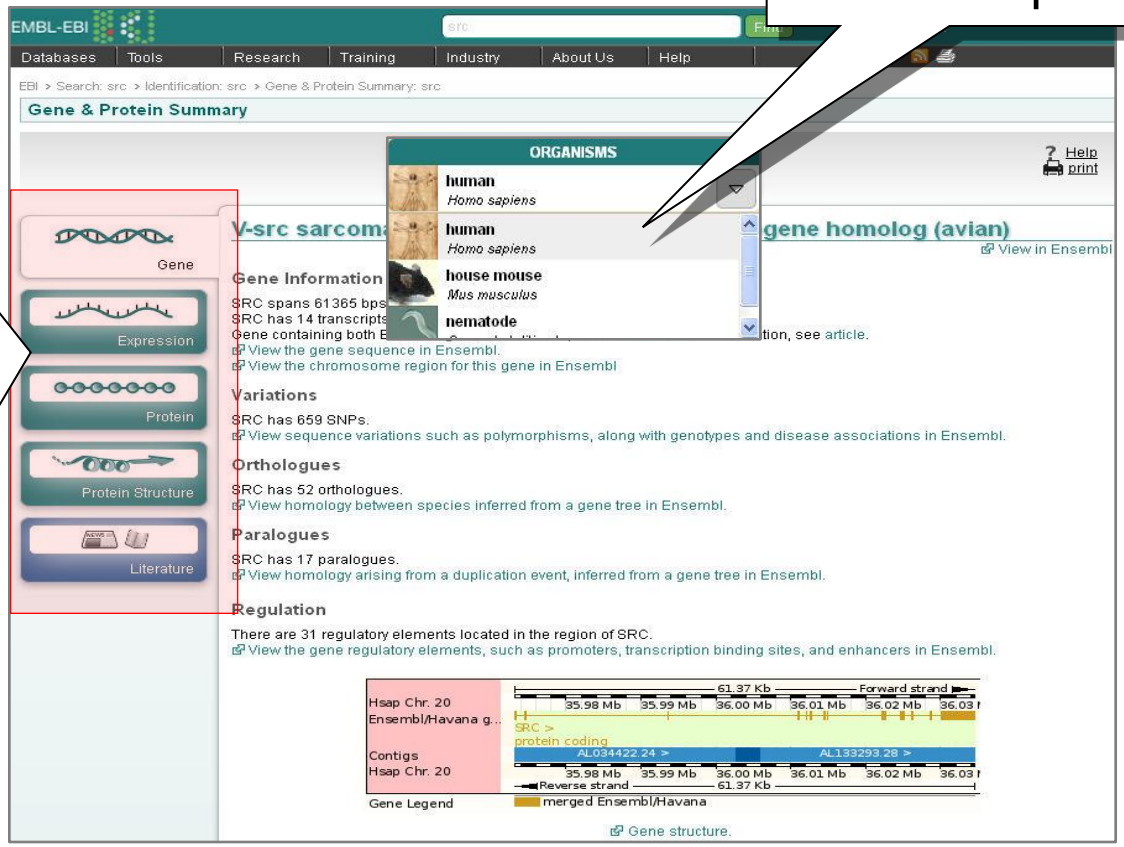




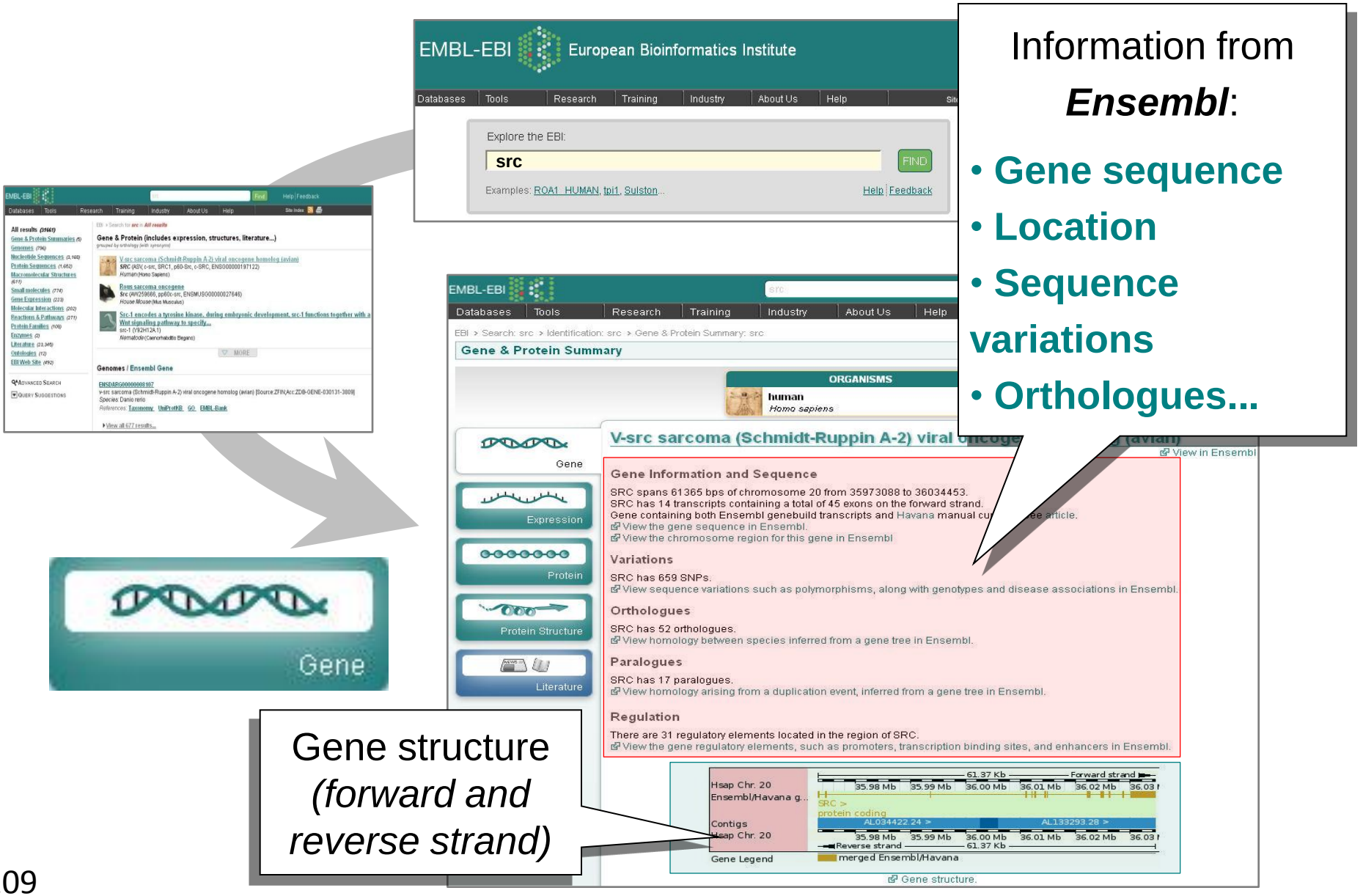
Easy to change between species

Tabs organise data by:

- gene
- expression
- protein
- structure
- literature



EBI search by gene name



Information from ***Ensembl***:

- **Gene sequence**
- **Location**
- **Sequence variations**
- **Orthologues...**

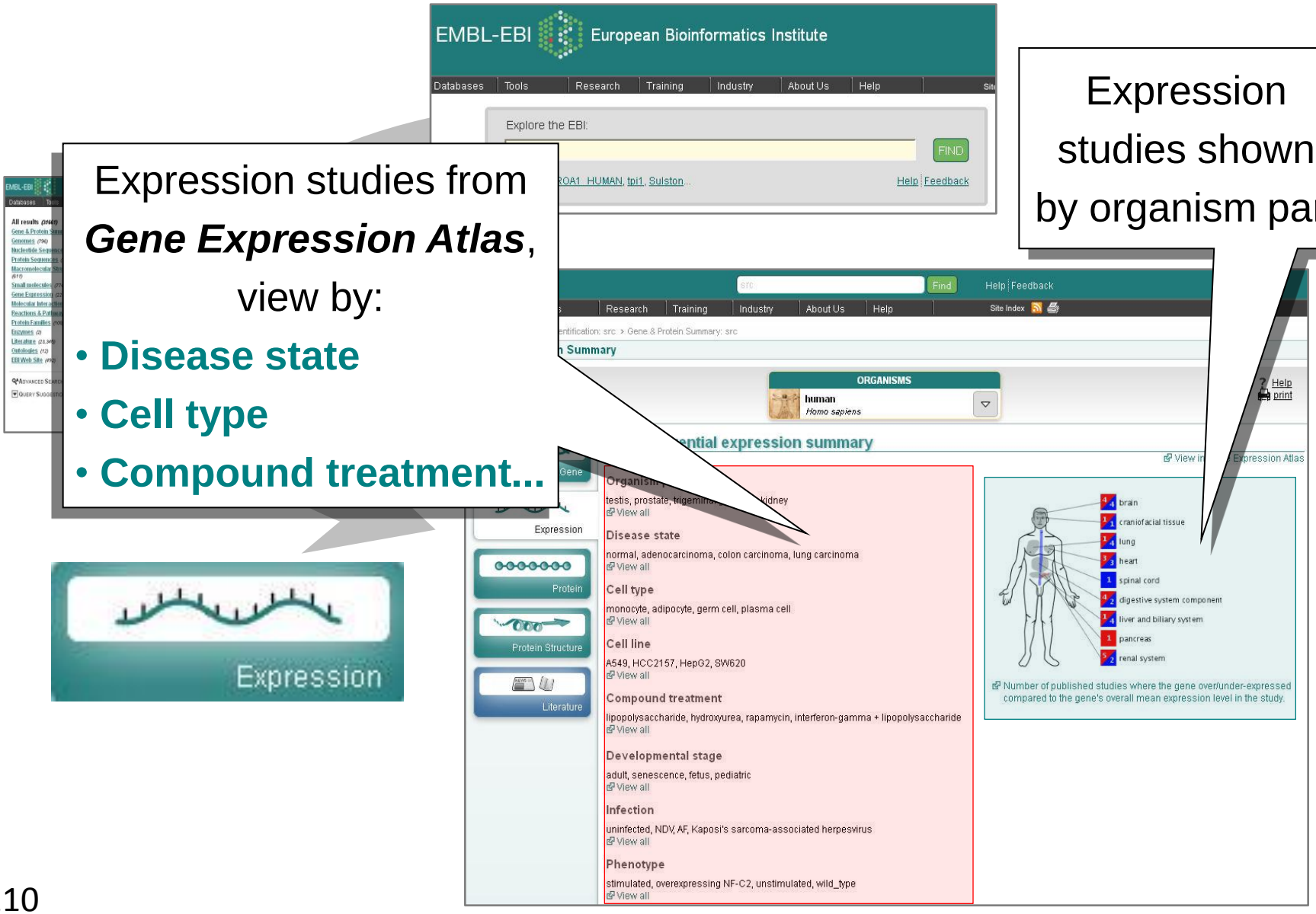
Gene structure *(forward and reverse strand)*

EBI search by gene name

Expression studies from ***Gene Expression Atlas***, view by:

- **Disease state**
- **Cell type**
- **Compound treatment...**

Expression studies shown by organism part



EBI search by gene name

Information from ***UniProt***:

- **Function**
- **Gene Ontology**
- **Isoforms**
- **Sequence...**

InterPro domain architecture

IntAct protein interaction data

EBI search by gene name

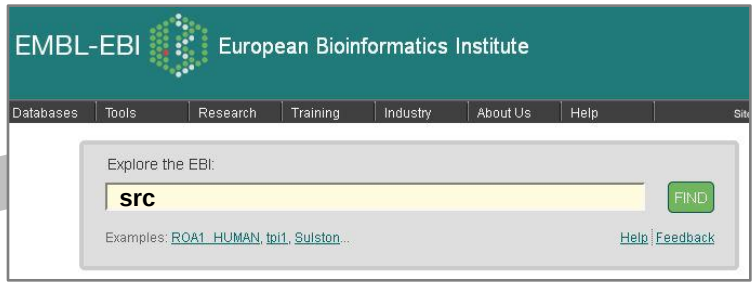
Information from ***PDBe***:

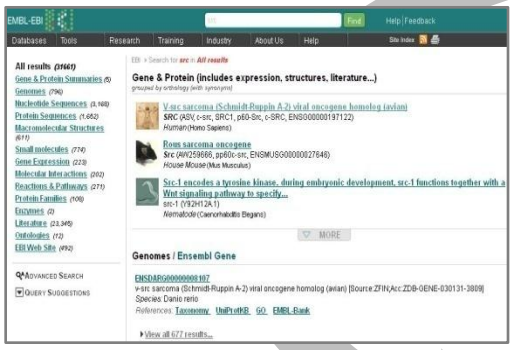
- **Chain information**
- **Structural domains**
- **Citations...**

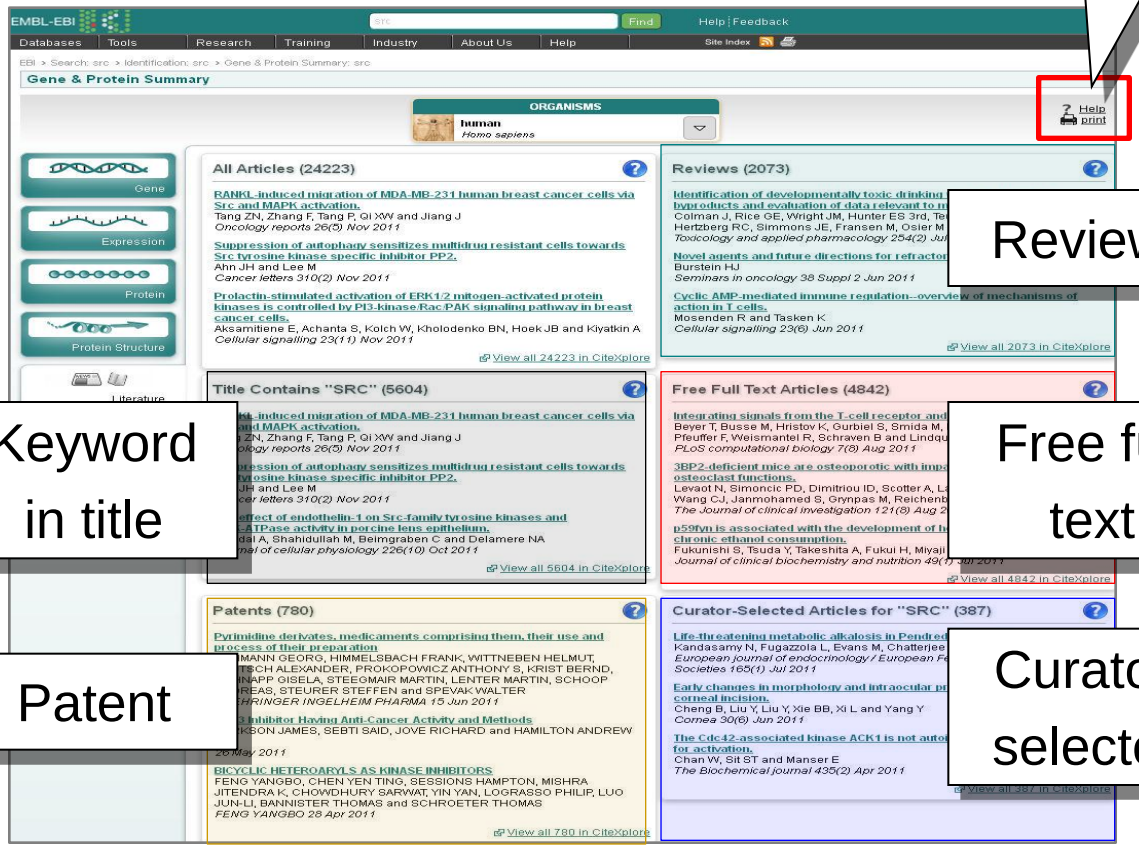
View additional structures

View structure

EBI search by gene name







Can print full summary of any page

Reviews

Keyword in title

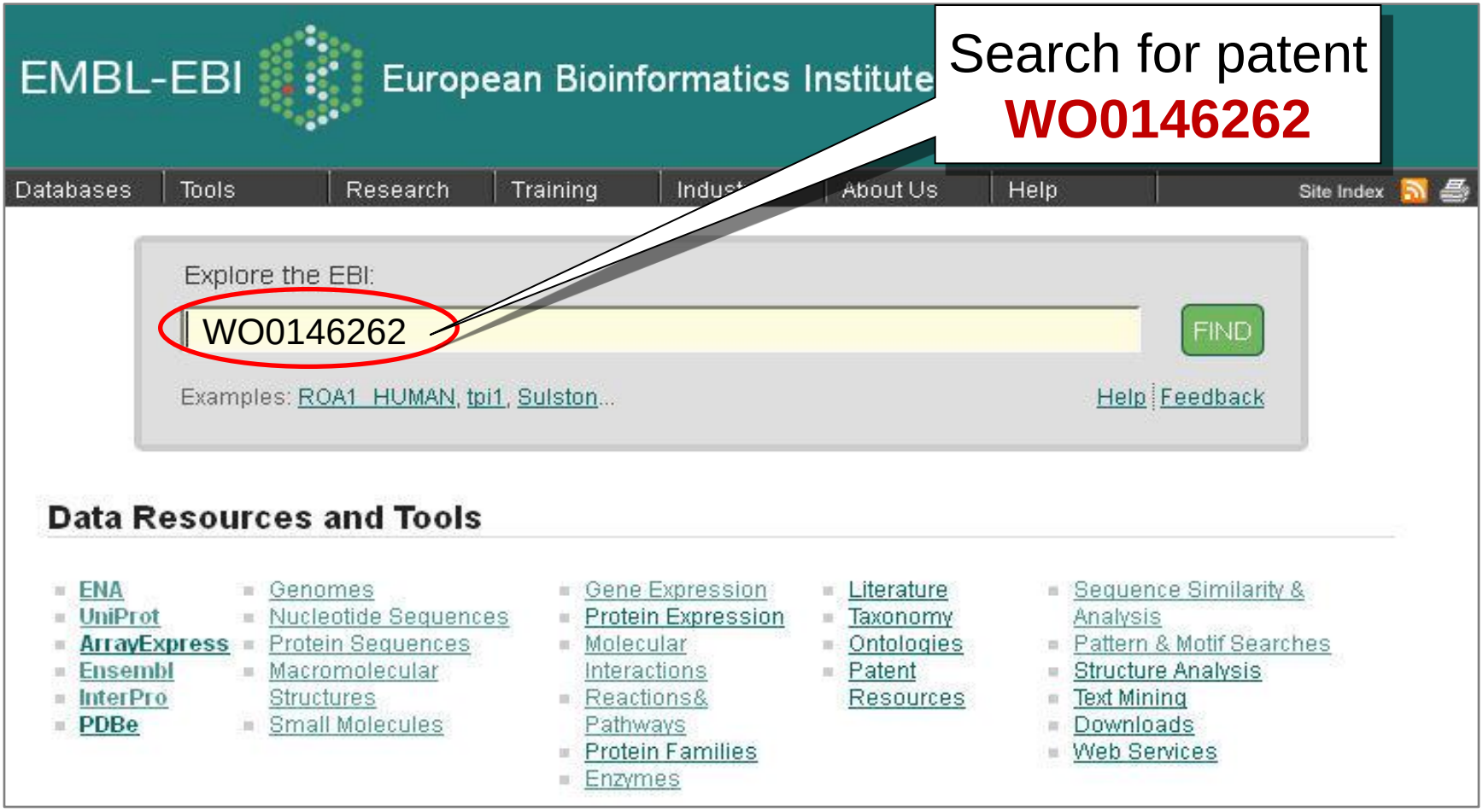
Free full text

Patent

Curator-selected

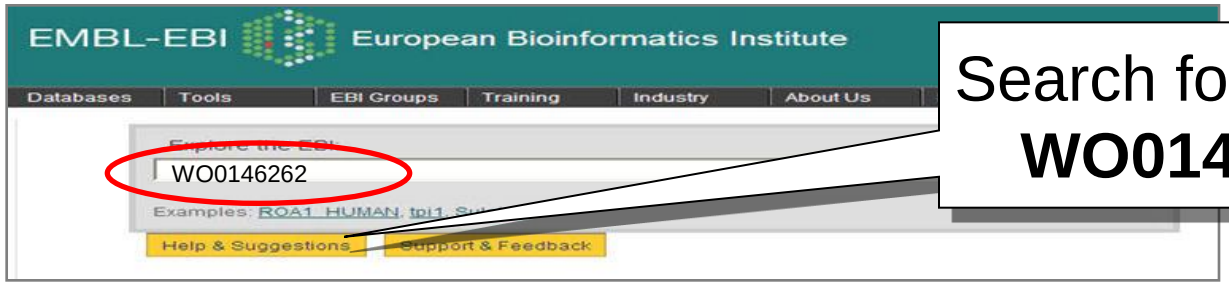
EBI search

EBI-Search for **patent** information



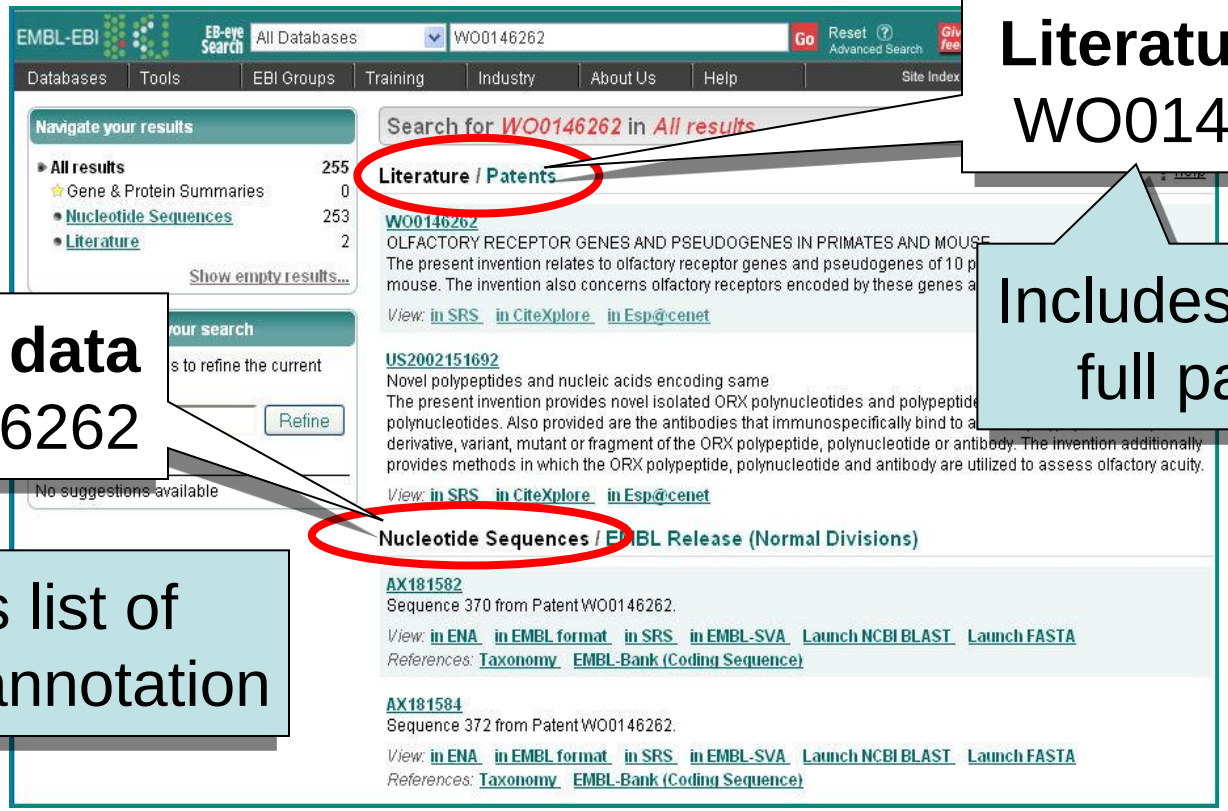
http://www.ebi.ac.uk/

EBI Search



Search for patent **WO0146262**

query



Literature for WO0146262

Includes link to full paper

Sequence data for WO0146262

Includes list of additional annotation

EBI search

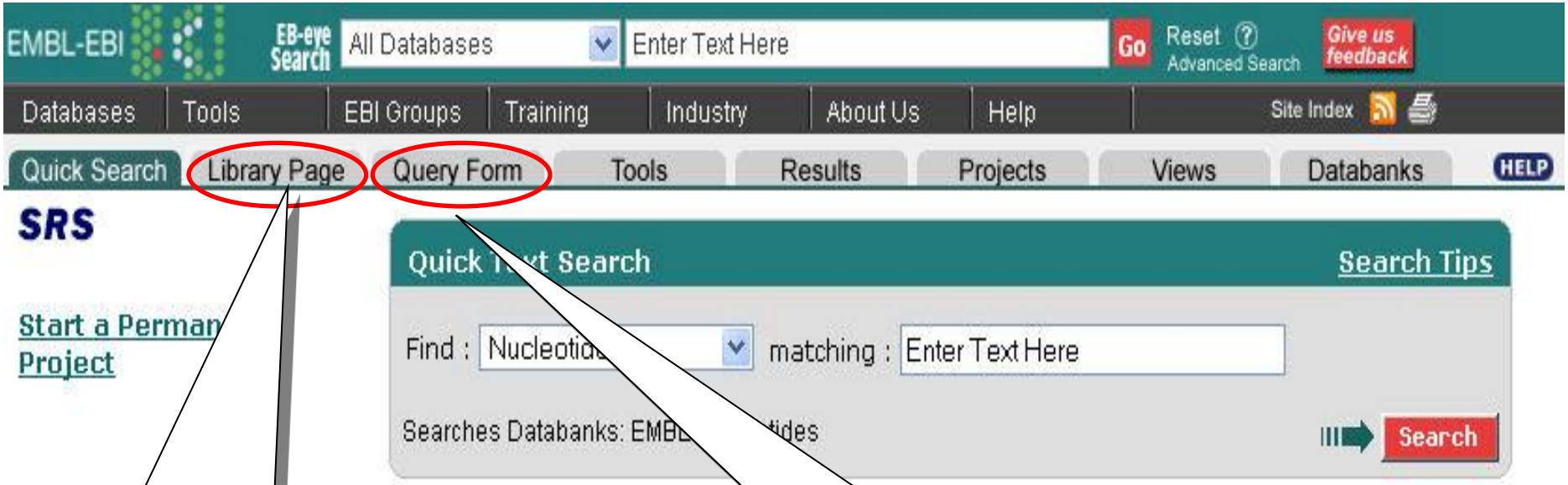
EBI search is a quick way to find literature and sequences (in ENA and UniProt) associated with a patent

EBI search

SRS advanced search

Sequence search

SRS: advanced text search

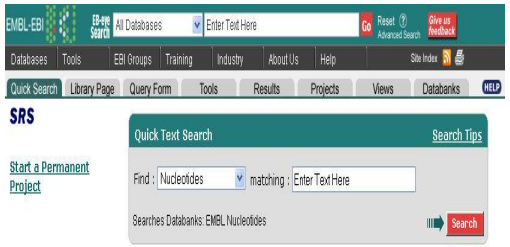


1st: Select resources to search

2nd: Create query

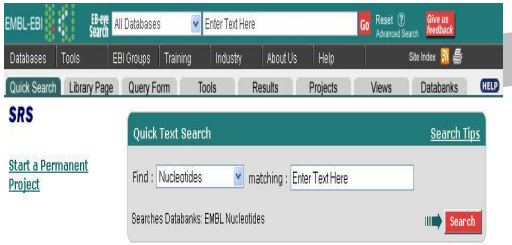
http://www.ebi.ac.uk/srs/

SRS: advanced text search



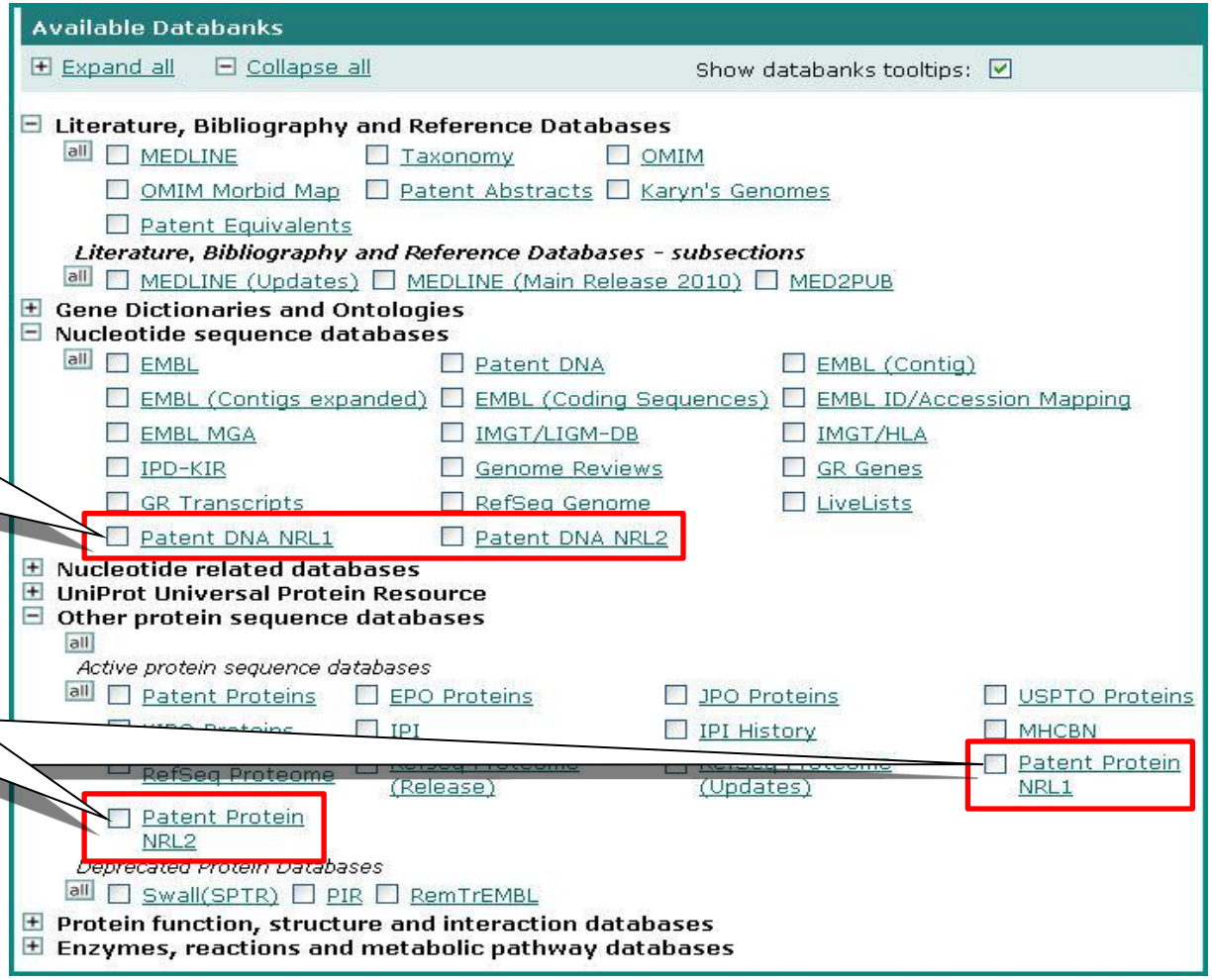
Select **library** tab

SRS: advanced text search



Select **library** tab

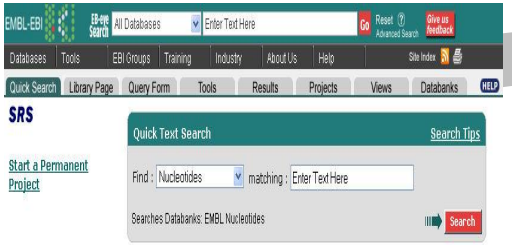
Search >100 databases



NR patent DNA
(*NRNL1 & NRNL2*)

NR patent proteins
(*NRPL1 & NRPL2*)

SRS: advanced text search

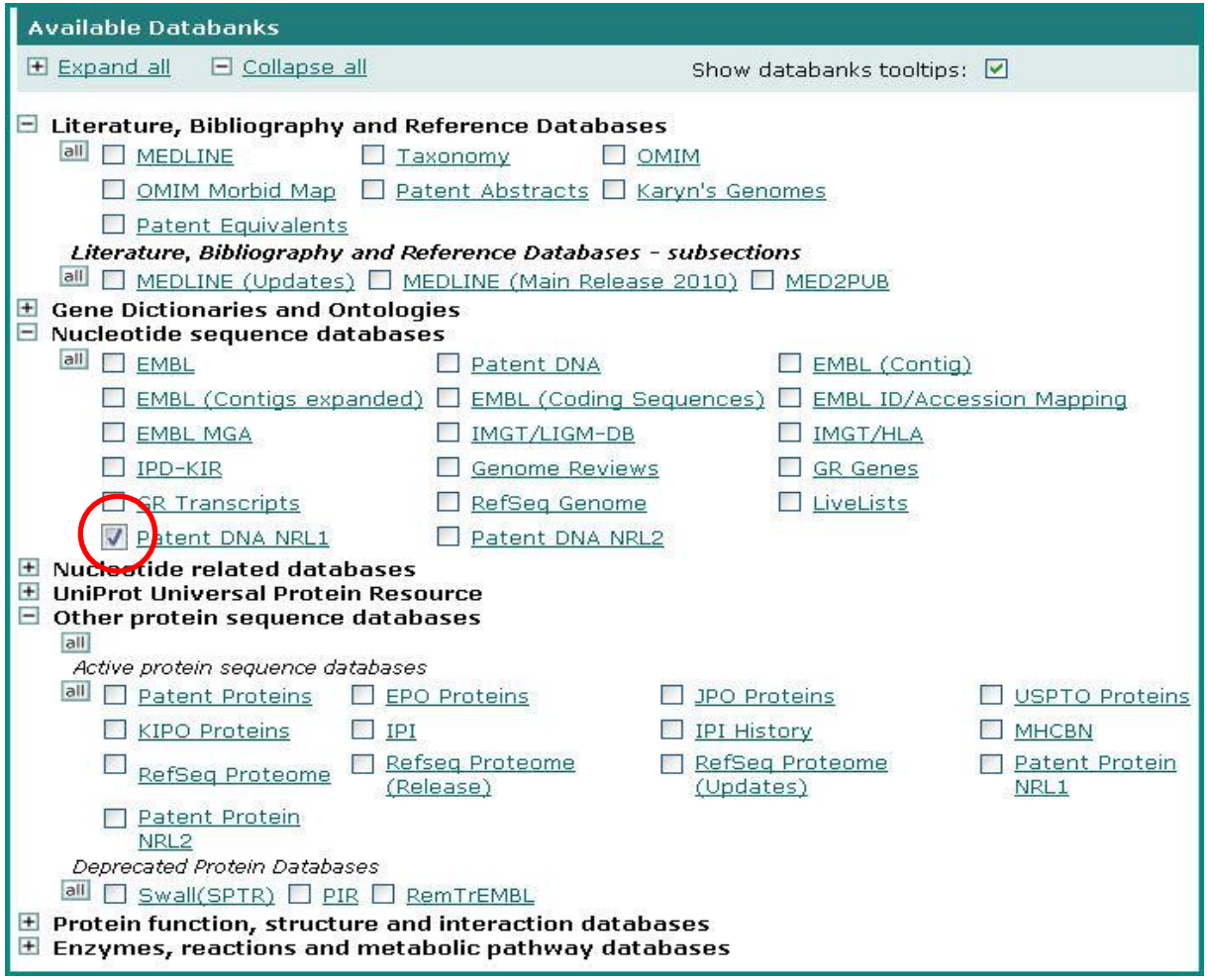


Select **library** tab

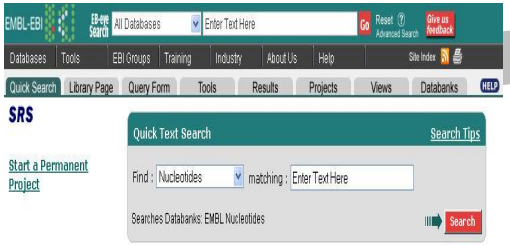
Search >100 databases

Example:

Selected to search NR level-1 patent DNA database



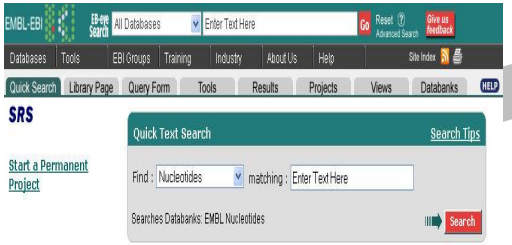
SRS: advanced text search



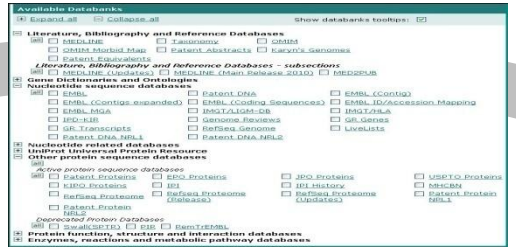
Select **library** tab

Select resources to search

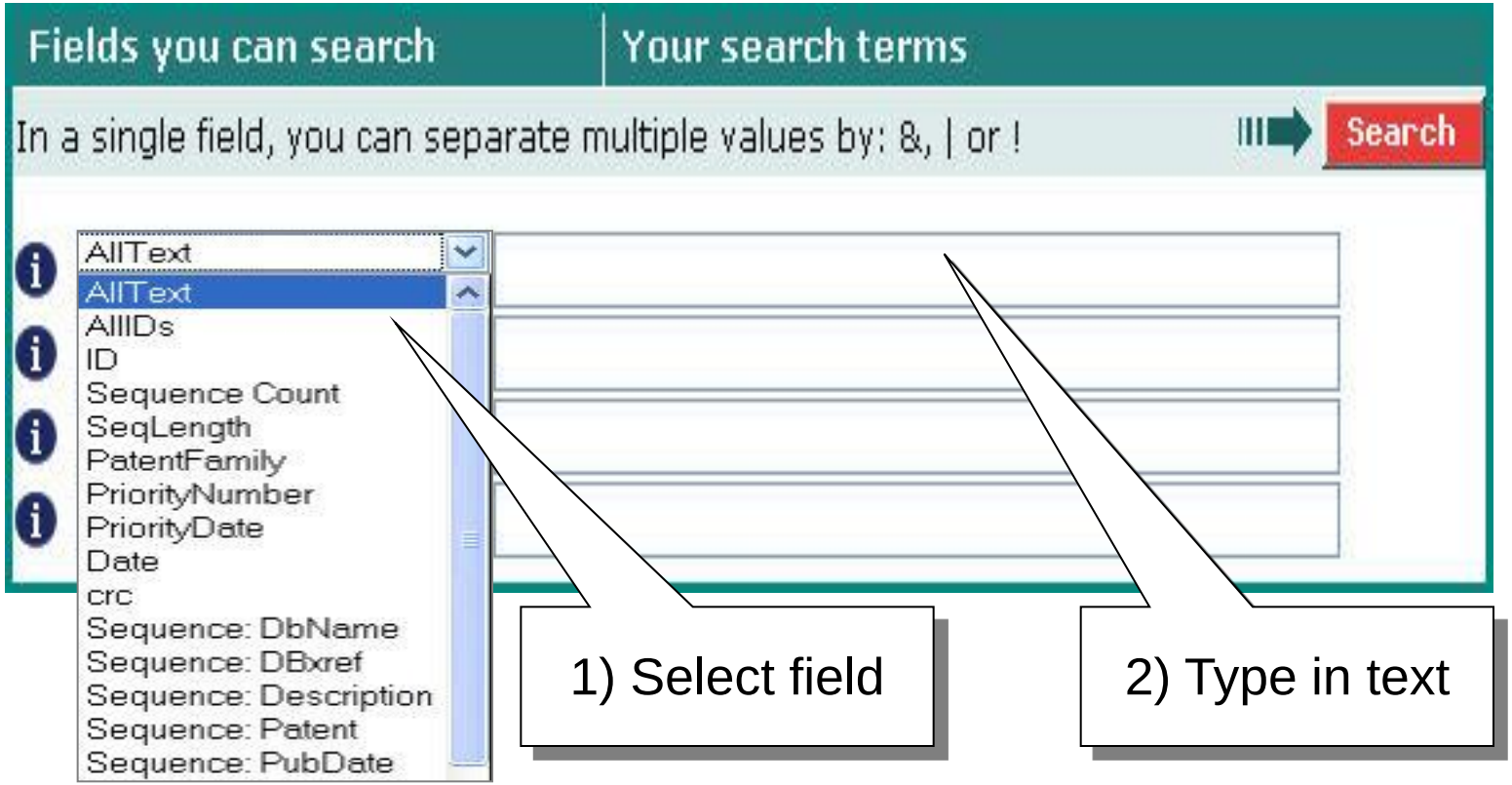
SRS: advanced text search



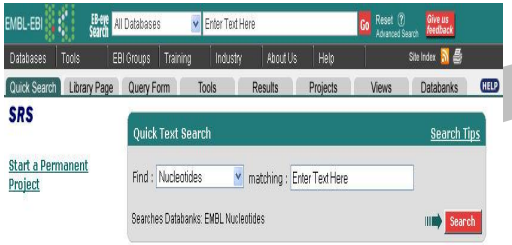
Select **library** tab



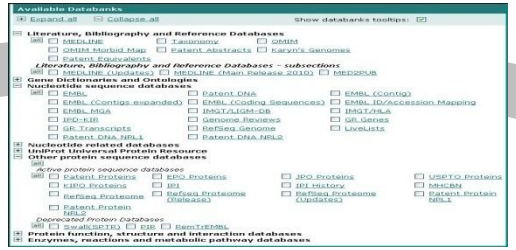
Select resources to search



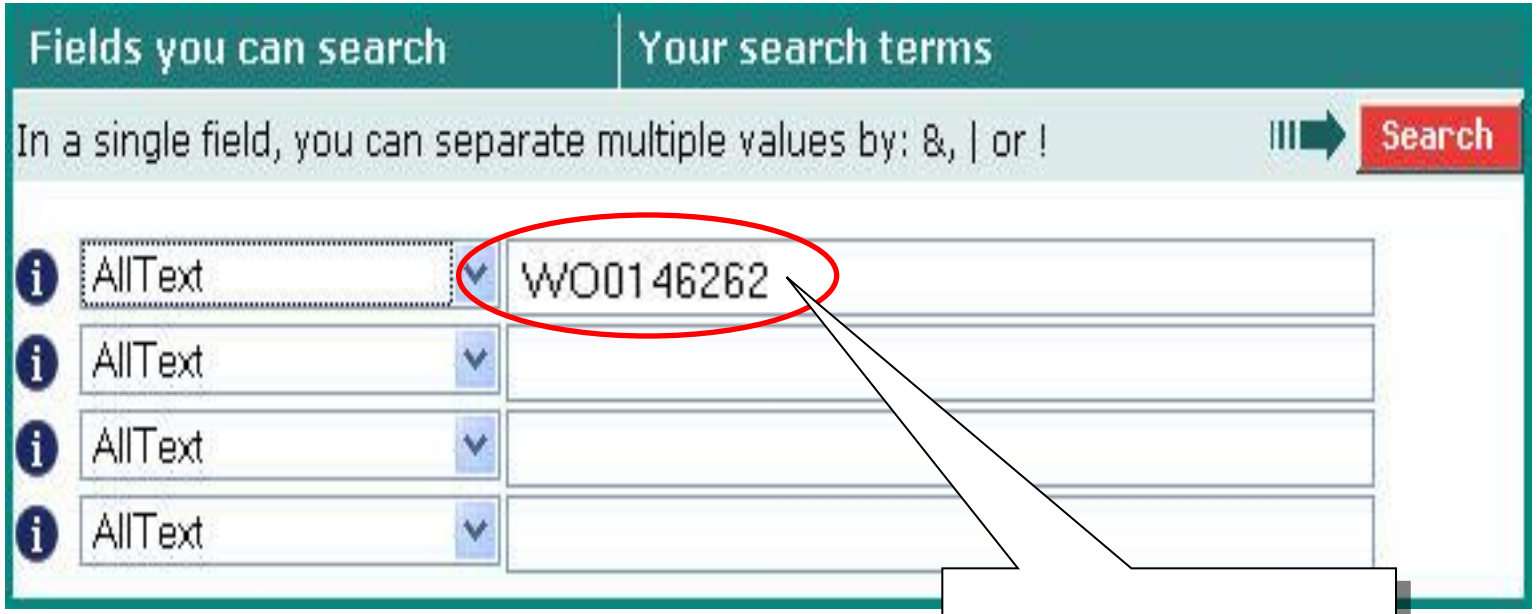
SRS: advanced text search



Select **library** tab

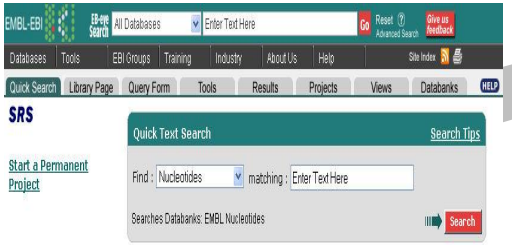


Select resources to search

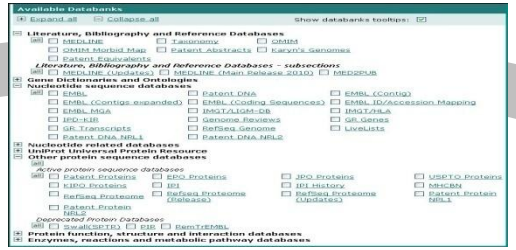


Here, selected patent number

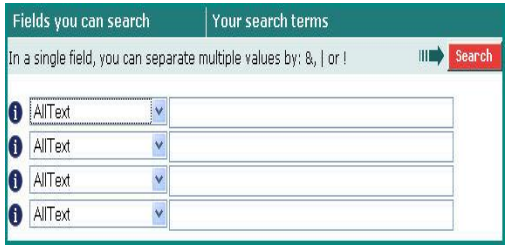
SRS: advanced text search



Select **library** tab

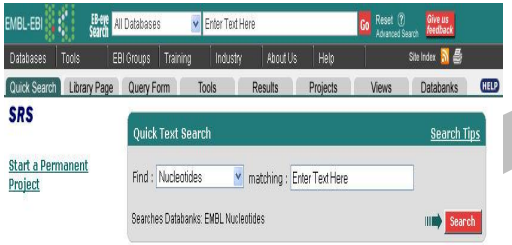


Select resources to search



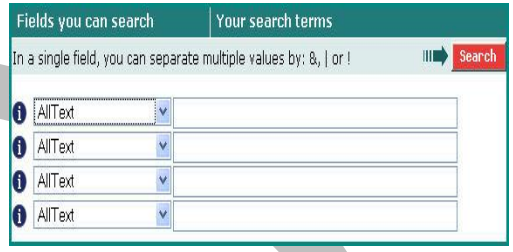
Create query

SRS: advanced text search

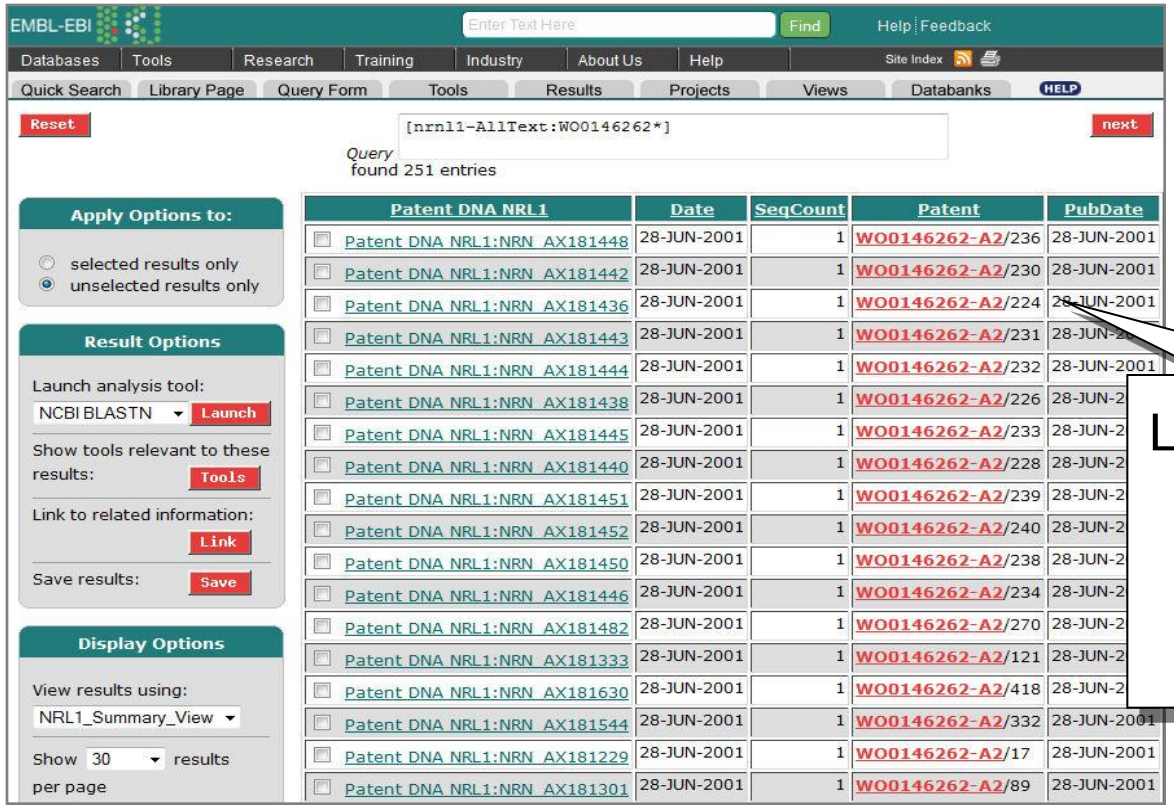


Select **library** tab

Select resources to search



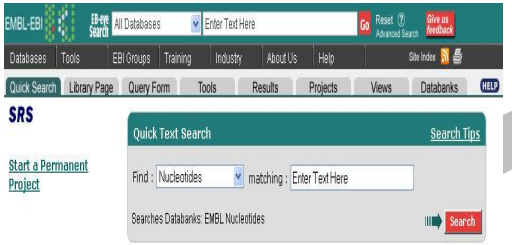
Create query



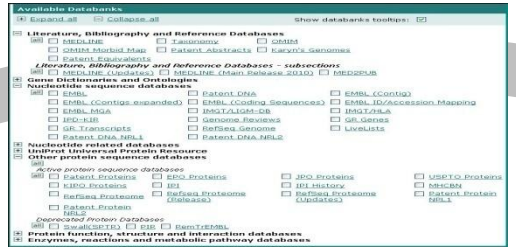
Patent DNA NRL1	Date	SeqCount	Patent	PubDate
Patent DNA NRL1:NRN_AX181448	28-JUN-2001	1	WO0146262-A2/236	28-JUN-2001
Patent DNA NRL1:NRN_AX181442	28-JUN-2001	1	WO0146262-A2/230	28-JUN-2001
Patent DNA NRL1:NRN_AX181436	28-JUN-2001	1	WO0146262-A2/224	28-JUN-2001
Patent DNA NRL1:NRN_AX181443	28-JUN-2001	1	WO0146262-A2/231	28-JUN-2
Patent DNA NRL1:NRN_AX181444	28-JUN-2001	1	WO0146262-A2/232	28-JUN-2001
Patent DNA NRL1:NRN_AX181438	28-JUN-2001	1	WO0146262-A2/226	28-JUN-2
Patent DNA NRL1:NRN_AX181445	28-JUN-2001	1	WO0146262-A2/233	28-JUN-2
Patent DNA NRL1:NRN_AX181440	28-JUN-2001	1	WO0146262-A2/228	28-JUN-2
Patent DNA NRL1:NRN_AX181451	28-JUN-2001	1	WO0146262-A2/239	28-JUN-2
Patent DNA NRL1:NRN_AX181452	28-JUN-2001	1	WO0146262-A2/240	28-JUN-2
Patent DNA NRL1:NRN_AX181450	28-JUN-2001	1	WO0146262-A2/238	28-JUN-2
Patent DNA NRL1:NRN_AX181446	28-JUN-2001	1	WO0146262-A2/234	28-JUN-2
Patent DNA NRL1:NRN_AX181482	28-JUN-2001	1	WO0146262-A2/270	28-JUN-2
Patent DNA NRL1:NRN_AX181333	28-JUN-2001	1	WO0146262-A2/121	28-JUN-2
Patent DNA NRL1:NRN_AX181630	28-JUN-2001	1	WO0146262-A2/418	28-JUN-2
Patent DNA NRL1:NRN_AX181544	28-JUN-2001	1	WO0146262-A2/332	28-JUN-2001
Patent DNA NRL1:NRN_AX181229	28-JUN-2001	1	WO0146262-A2/17	28-JUN-2001
Patent DNA NRL1:NRN_AX181301	28-JUN-2001	1	WO0146262-A2/89	28-JUN-2001

Lists non-redundant nucleotide sequences from WO0146262

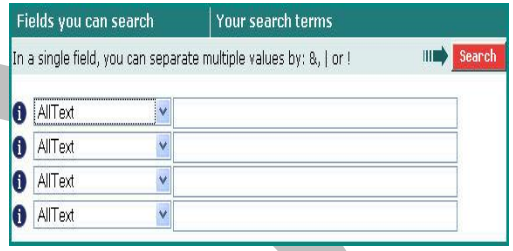
SRS: advanced text search



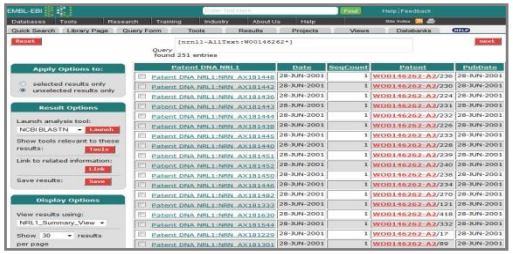
Select **library** tab



Select resources to search

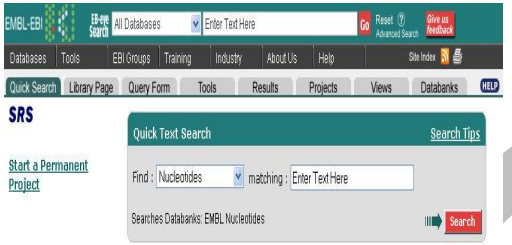


Create query

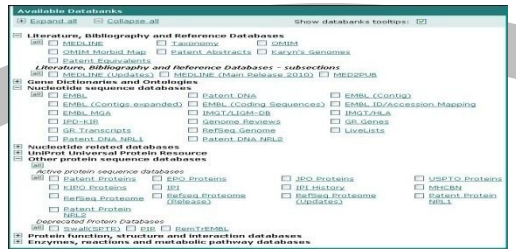


WO0146262 sequences

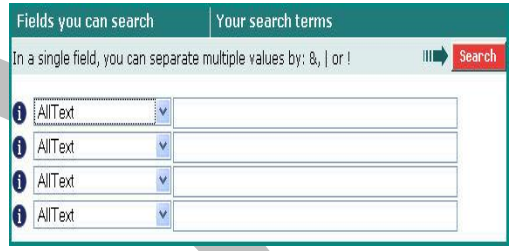
SRS: advanced text search



Select **library** tab

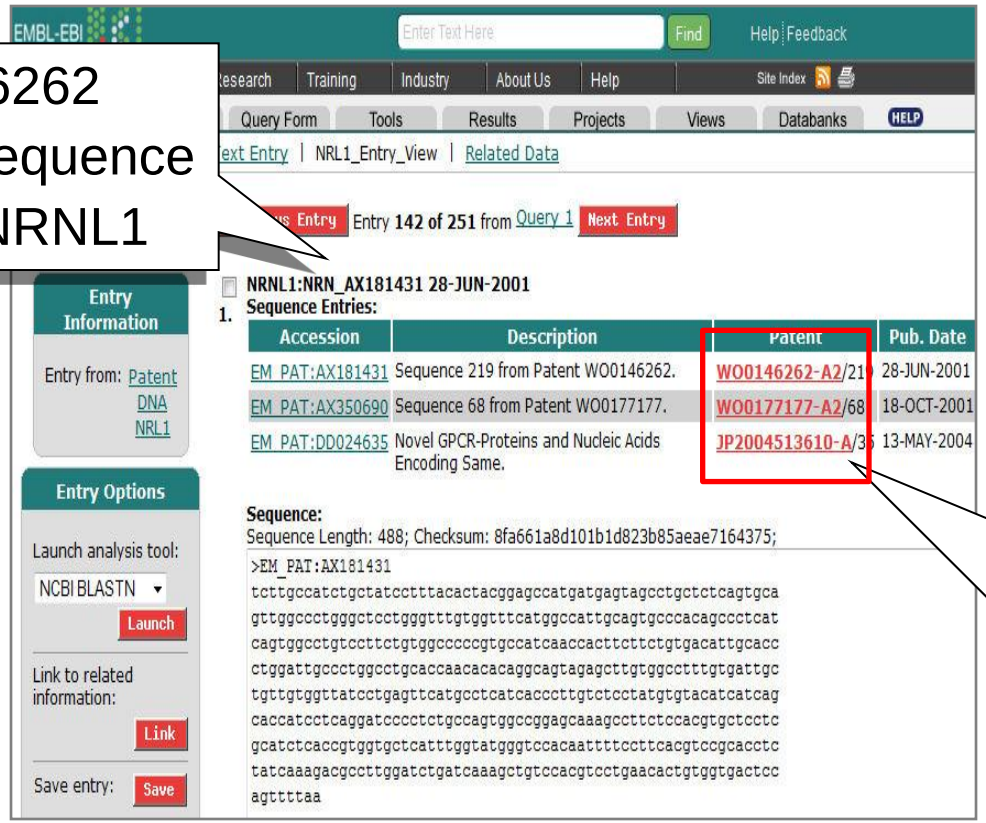


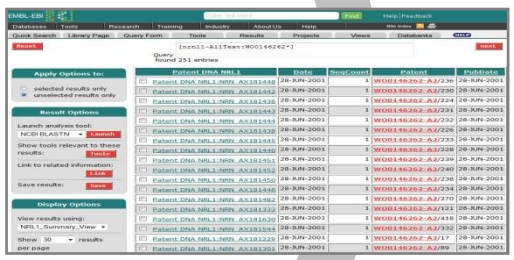
Select resources to search



Create query

WO0146262 nucleotide sequence record in NRNL1

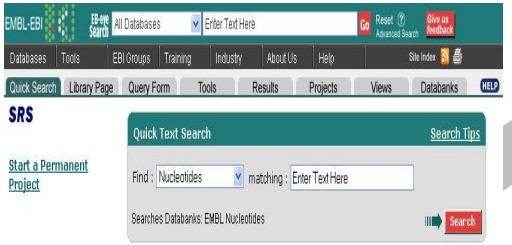




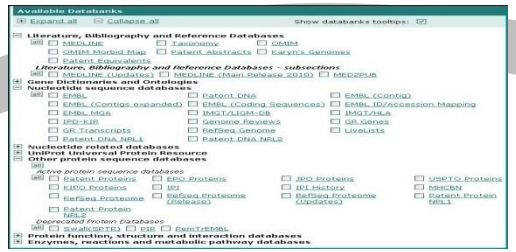
WO0146262 sequences

Details which other patents also claim this sequence *(with NRNL2, would see family grouping)*

SRS: advanced text search

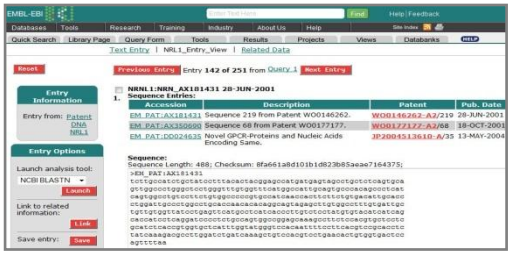


Select **library** tab

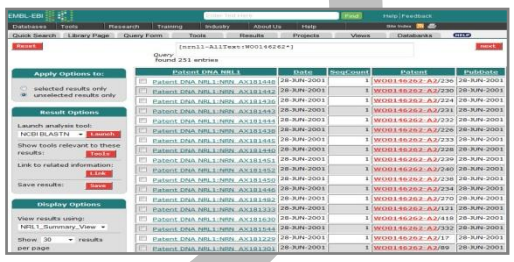


Select resources to search

Create query

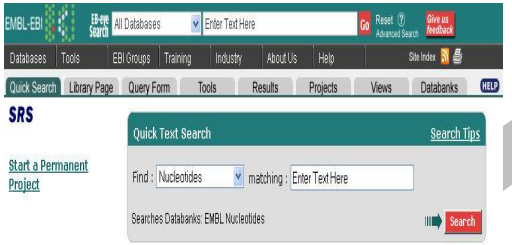


NRNL1 **sequence** record

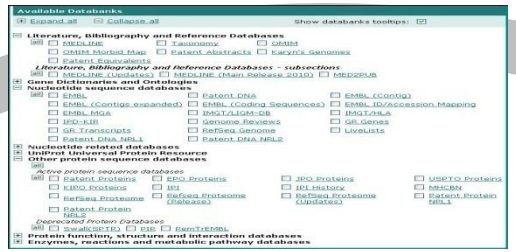


WO0146262 sequences

SRS: advanced text search



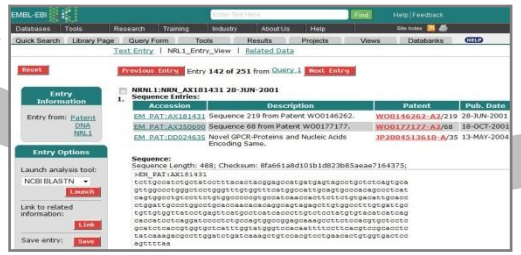
Select **library** tab



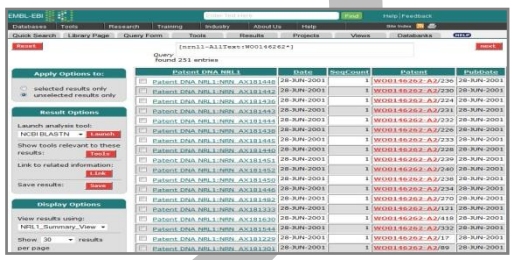
Select resources to search

Create query

WO0146262 **literature**



NRNL1 **sequence** record



WO0146262 sequences

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

SRS: advanced text search

ENA

Find all sequences associated with a patent

NRNL1

Find all sequences associated with a patent

+ *identify **all patents** associated with each sequence*

NRNL2

Find all sequences associated with a patent

+ *identify **all patents in the same family** associated with each sequence*

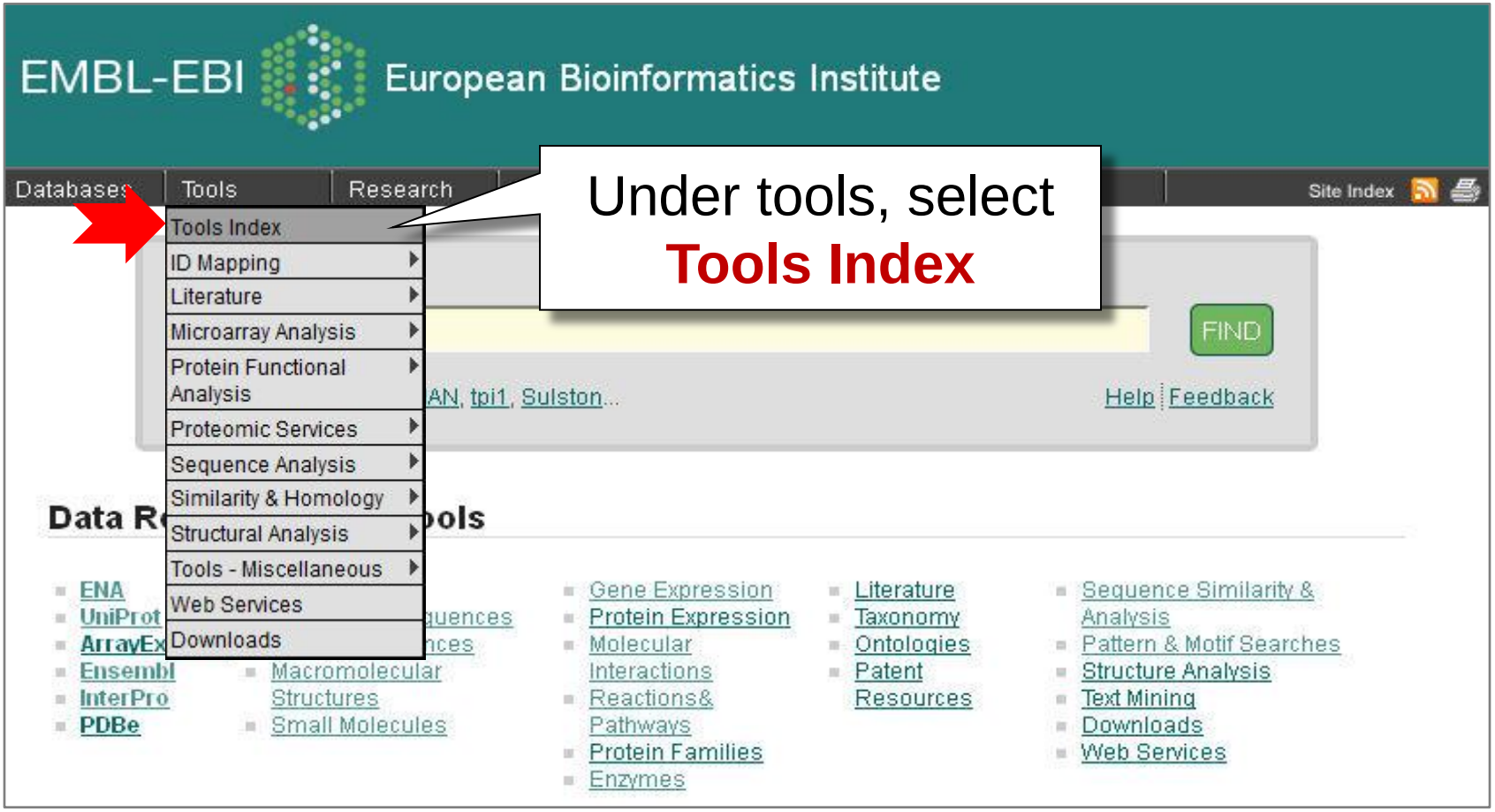
EBI search

SRS advanced search

Sequence search

What's available at EBI

Tools are accessible from the EBI homepage



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

What's available at EBI

EMBL-EBI

Enter Text Here | Find | Help | Feedback

Databases | Tools | Research | Training | Industry | About Us | Help | Site Index

- Tools Home
- Tools A-Z
- ID Mapping
- Literature
- Microarray Analysis
- Protein Functional Analysis
- Proteomic Services
- Sequence Analysis
- Similarity & Homology
- Structural Analysis
- Tools - Miscellaneous
- Web Services

- Databases
- Downloads

EBI > Tools

Tools at the EBI

We provide a comprehensive range of bioinformatics tools.

This page shows a selection of those that are used most fre e analysis and comparison of nucleotide and protein sequences, data ents, text mining of the scientific literature and tools for determination lar structures.

All these tools can be accessed over the web and most pr SOAP or REST APIs.

Link to list of all tools

Nucleotide and Protein sequence searching

Nucleotide sequence searches

The sequence databases that can be searched with the tools outlined below include EMBL-Bank, Coding Sequences, immunoglobulins and High throughput cDNA:

- ENA Search
- BLAST Nucleotide

Protein sequence searches

The protein sequence databases available to search below include UniProtKB, sequences derived from macro molecular structures, immunoglobulins and sequences from patents:

- BLAST Protein
- PSI-Search

Multiple Sequence Alignment

Alignment of three or more sequences to identify regions of conservation which may indicate functional constraints and infer evolutionary relationships.

- Clustal Omega
- ClustalW2
- WebPRANK
- T-Coffee
- MUSCLE

Pairwise Sequence Alignments

Alignment of two sequences to identify regions where the sequence is conserved and conversely regions where the sequence is not conserved.

- Needle
- LAlign

Most popular tools are listed

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

What's available at EBI

EMBL-EBI

Enter Text Here | Find | Help | Feedback

Databases | Tools | Research | Training | Industry | About Us | Help | Site Index

- Tools Home
- Tools A-Z
- Web Services
- Download

EBI > Tools > Sequence Similarity Searching

Sequence Similarity Searching

Sequence Similarity Searching is a method of searching sequence databases by using al
sequence. By statistically assessing how well database and query sequences match one
and transfer information to the query sequence.

The tools can be launched with different form pre-sets using the buttons - these can be cha
page as well.

BLAST

NCBI BLAST — NCBI BLAST (blastall) is the most commonly used sequence simil
heuristics to perform fast **local** alignment searches.
Protein | Nucleotide | Vectors

WU-BLAST — WU-BLAST is similar to NCBI BLAST but combines multiple parameter
simpler 'sensitivity' setting.
Protein | Nucleotide | Parasites

PSI-BLAST — PSI-BLAST allows users to construct and perform a BLAST search with a custom, position-specific, scoring matrix which can help find distant evolutionary relationships. PHI-BLAST functionality is also available to restrict results using patterns.
Protein

FASTA

FASTA — FASTA is another commonly used sequence similarity search tool which uses heuristics for fast **local** alignment searching.
Protein | Nucleotide | Proteomes | Genomes | Whole Genome Shotgun | ASD Protein | ASD Nucleotide | LGIC Protein | LGIC Nucleotide

SSEARCH — SSEARCH is an optimal (as opposed to heuristics-based) **local** alignment search tool using the Smith-Waterman algorithm. Optimal searches guarantee you find the best alignment score for your given parameters.
Protein | Nucleotide | Proteomes | Genomes | Whole Genome Shotgun | ASD Protein | ASD Nucleotide | LGIC Protein | LGIC Nucleotide

PSI-Search — PSI-Search combines the sensitivity of the Smith-Waterman search algorithm (SSEARCH) with the PSI-BLAST profile construction strategy to find distantly related protein sequences.
Protein

GGSEARCH — GGSEARCH performs optimal **global-global** alignment searches using the Needleman-Wunsch algorithm.
Protein | Nucleotide

GLSEARCH — GLSEARCH performs an optimal sequence search using alignments that are **global** in the query but **local** in the database sequence. This can be useful when you want to match all of a short query sequence to part of a larger database sequence.
Protein | Nucleotide

FASTM/S/F — These specialist programs allow searches of databases using a group of short peptides as the query.
Protein | Nucleotide | Proteomes | ASD Protein | LGIC Protein

ENA SEQUENCE SEARCH

The EBI has a new search tool which is far faster than BLAST, with only a marginal loss in search sensitivity. Try it out at ENA Sequence Search.

The tools described on this page are provided using our **new bioinformatics analysis tools framework**. If you have any feedback or encounter any issues please let us know via EBI support.

Full list of sequence search tools

http://www.ebi.ac.uk/Tool/ssss

STEP 1:

Choose a search algorithm

Choosing the right search engine

BLAST – Fast search (better for proteins than nucleotides)

FASTA – Fast search

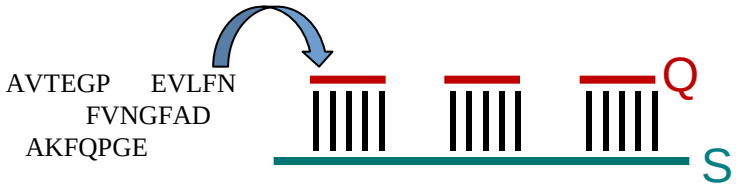
PSI-SEARCH – Finding remote homologues

SSEARCH – Sensitive but slow; good for short sequences

GGSEARCH – Force full-length matches

GLSEARCH – Match domains/patterns to protein; oligo-to-gene

FASTM – Multi-peptide search



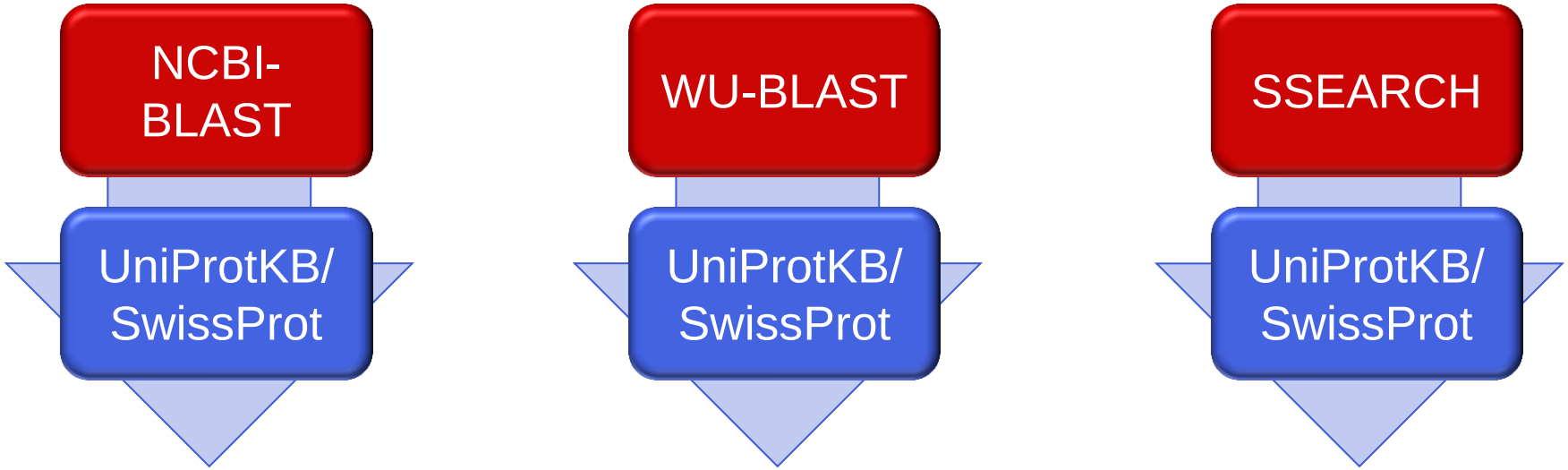
Example:

Comparing search engines using a short peptide query sequence

query sequence: **RPPSWIPK**

Comparing search engines

query : **RPPSWIPK**



No hits found

hit	length	e()
1: TY01_PHYAZ	61	5.4

hit	length	e()
1: TY01_PHYAZ	61	0.42
2: BRK5_PHYNO	8	4.8
3: BRK_AMICA	9	9.2
5: BRK_LEPOS	9	9.2

Look at the difference in **e-values**

Comparing search engines

SSEARCH is a sensitive search engine

suitable for short sequences

(may be too slow for longer sequences)

Comparing search engines - specialised

query : **RPPSWIPK**

SSEARCH

UniProtKB SwissProt

GLEARCH

iProtKB/ wissProt

GGEARCH

iProtKB/ wissProt

GLSEARCH has a preference for long hits

GGSEARCH limited to similar sized hits

hit	length	e()
1: TY01_PHYAZ	61	0.42
2: BRK5_PHYNO	8	4.8
3: BRK_AMICA	9	9.2
5: BRK_LEPOS	9	9.2

hit	length	e()
1: TY01_PHYAZ	61	e-16
2: BRK5_PHYNO	8	1.9e-11
3: TY01_PHYBU		e-8
4: BRK_ONCMY	10	5.6e-8
5: BRK_LEPOS	9	1e-7
8: B4GT2_HUMAN	372	5.8e-5
...	...	...
39: DNAA_PROM	465	0.024
40: DNAA_PROM	199	

hit	length	e()
1: BRK5_PHYNO	8	e-7
2: TY01_PHYBU	8	2e-5
3: BRK_LEPOS	9	5.4e-4
4: BRK_AMICA	9	5.4e-4
5: BRK4_PHAJA	8	0.0087
6: BRK_PHYHY	8	0.0087
...	...	...
34: TY51_LITRU	7	8.3

Comparing search engines - specialised

GGSEARCH finds **similar length** sequences;

GLSEARCH matches entire sequence to

any length sequences

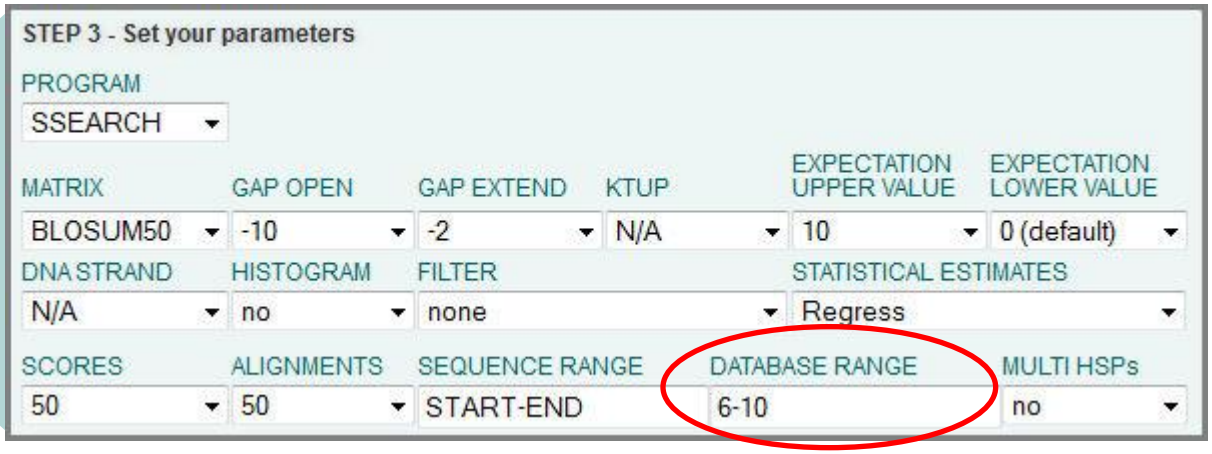
Restricting length of matches

query : **RPPSWIPK**

SSEARCH

Database range 6-10

UniProtKB/ SwissProt



Restricting length of matches

query : **RPPSWIPK**

SSEARCH

Database range 6-10

UniProtKB/ SwissProt

GGEARCH

Limiting database range limits size of hits, but stricter than GGSEARCH

SSEARCH

UniProtKB/ SwissProt

hit	length	e()
1: BRK5_PHYNO	8	0.34
2: BRK_LEPOS	9	0.57
3: BRK_AMICA	9	0.57
4: TY01_PHYBU	8	0.61
5: BRK_ONCMY	10	0.63
6: BRK4_PHAJA	8	5.0
...	...	...
18: BRK3_PELRI	9	9.8

	length	e()
1: BRK5_PHYNO	8	8e-7
2: TY01_PHYBU	8	2e-5
3: BRK_LEPOS	9	5.4e-4
4: BRK_AMICA	9	5.4e-4
5: BRK4_PHAJA	8	0.0087
6: BRK_PHYHY	8	0.0087
...	...	...
34: TY01_LITRU	7	8.3

hit	length	e()
1: TY01_PHYAZ	61	0.42
2: BRK5_PHYNO	8	4.8
3: BRK_AMICA	9	9.2
4: BRK_LEPOS	9	9.2

STEP 2:

Choose a database to search

Several databases available

Protein

FASTA/SSEARCH/GGSEARCH/GLSEARCH - Protein S

This tool provides sequence similarity searching against protein datab

Use this tool

STEP 1 - Select your databases

PROTEIN DATABASES

1 Databank Selected — Clear Selectio

- UniProt Knowledgebase
- UniProtKB/Swiss-Prot
- UniProtKB/Swiss-Prot isoforms
- UniProtKB/TrEMBL
- UniProtKB Taxonomic Subsets
- UniProt Clusters

STEP 2 - Enter your input sequence

0 Databank Selected — Clear Selection

- UniProt Knowledgebase
- UniProtKB/Swiss-Prot
- UniProtKB/Swiss-Prot isoforms
- UniProtKB/TrEMBL
- UniProtKB Taxonomic Subsets
- UniProt Clusters
- Patents
  - EPO Patent Protein Sequences
  - JPO Patent Protein Sequences
  - KIPO Patent Protein Sequences
  - USPTO Patent Protein Sequences
  - NR Patent Proteins Level-1
  - NR Patent Proteins Level-2
- Structure
- Other Protein Databases
  - UniProt Archive
  - International Protein Index
  - IntAct
  - IMGT/HLA
  - IPD-KIR
  - MACiE Annot Pub

Note: these databases cover the same sequences

- NR level-1
- NR level-2
- EPO + JPO + KIPO + USPTO

Patent databases

The default settings will fulfill the needs of most users and, for that reason, are not visible.

More options... *(Click here, if you want to view or change the default settings.)*

STEP 4 - Submit your job

Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

Submit

Several databases available

Nucleotide

FASTA/SSEARCH/GGSEARCH/GLSEARCH - Nucleotid

This tool provides sequence similarity searching against nucleotide da
programs. FASTA provides a heuristic search with a nucleotide query.
database for searching with a protein query. Optimal searches are ava
GGSEARCH (global) and GLSEARCH (global query, local database).

Use this tool

STEP 1 - Select your databases

NUCLEOTIDE DATABASES

107 Databanks Selected — Clear Selectio

or upload a file: Browse...

STEP 3 - Set your parameters

PROGRAM

FASTA

The default setti ... and, for that rea

More options... ... the defau

STEP 4 - Submit your job

Be notified by email (Tick this box if you want to be notified by ema

Submit

Note: these databases cover the same sequences

- NR level-1
- NR level-2
- EMBL Patents

Patent data

0 Databank Selected — Clear Selection

- EMBL-Bank
  - EMBL Release
    - EMBL Environmental
    - EMBL Fungi
    - EMBL Human
    - EMBL Invertebrate
    - EMBL Mammal
    - EMBL Mouse
    - EMBL Phage
    - EMBL Plant
    - EMBL Prokaryote
    - EMBL Rodent
    - EMBL Synthetic
    - EMBL Transgenic
    - EMBL Unclassified
    - EMBL Viral
    - EMBL Vertebrate
      - EMBL EST Vertebrate
      - EMBL GSS Vertebrate
      - EMBL HTC Vertebrate
      - EMBL HTG Vertebrate
      - EMBL Patent Vertebrate
      - EMBL Standard Vertebrate
      - EMBL STS Vertebrate
      - EMBL TSA Vertebrate
  - EMBL Updates
  - EMBL Coding Sequence
  - Others
    - EMBL Expressed Sequence Tag
    - EMBL Genome Survey Sequence
    - EMBL High Throughput cDNA
    - EMBL High Throughput Genome
    - EMBL Patent
    - EMBL Standard
    - EMBL Sequence Tagged Site
    - EMBL Transcriptome Shotgun Assembly
    - EMBL Release and Updates
    - EMBL Vectors
- IMGT
- HGV
- Patents
  - NR Patent DNAs Level-1
  - NR Patent DNAs Level-2
- Structure

Database size is important

The larger the database searched,
the higher (less significant) the resulting e-values

Most sequence databases are large...
...and growing every day:

- ENA-Annotation >160 million entries
- UniParc (non-redundant) >30 million entries

Example:

Comparing database size when searching with a short peptide query sequence

query sequence: **RPPSWIPK**

Comparing databases size

query : **RPPSWIPK**

Database size decreasing

SSEARCH	SSEARCH	SSEARCH
UniParc	UniProtKB	UniProtKB/ SwissProt

No hits found

hit	e()
1: TY01_PHYAZ	3.6

hit	e()
1: TY01_PHYAZ	0.068
2: BRK5_PHYNO	1.9
3: TY01_PHYBU	7.7
4: BRK_AMICA	8.7
5: BRK_LEPOS	8.7
6: BRK_ONCMY	9.7

Look at the difference in **e-values**

Comparing databases size

The larger the database searched,

the higher (less significant) the resulting e-values

- Search the **smallest** database likely to contain your sequence
- You can also run a second search of the entire database, or run multiple small searches

Is it best to search a protein or a nucleotide database?

Is it best to search a protein or a nucleotide database?

2 issues are worth considering…

1) Codon degeneracy

Because amino acids are encoded by different codons, there can be **more variability between CDSs** than between proteins

	U	C	A	G
U	UUU, UUC — Phe UUA, UUG — Leu	UCU, UCC, UCA, UCG — Ser	UAU, UAC — Tyr UAA, UAG — *Stop*	UGU, UGC — Cys UGA — *Stop* UGG — Trp
C	CUU, CUC, CUA, CUG — Leu	CCU, CCC, CCA, CCG — Pro	CAU, CAC — His CAA, CAG — Gln	CGU, CGC, CGA, CGG — Arg
A	AUU, AUC, AUA — Ile AUG — Met	ACU, ACC, ACA, ACG — Thr	AAU, AAC — Asn AAA, AAG — Lys	AGU, AGC — Ser AGA, AGG — Arg
G	GUU, GUC, GUA, GUG — Val	GCU, GCC, GCA, GCG — Ala	GAU, GAC — Asp GAA, GAG — Glu	GGU, GGC, GGA, GGG — Gly

Amino acids

Ser
Ser

match

Nucleotides

UCU
AGC

mismatch

1) Codon degeneracy

Human CKS1B kinase

Zebra finch CDC28 kinase 1B

Proteins

```
SeqA Name                        Len(aa)  SeqB Name                        Len(aa)  Score
==================================================================================================
1    sp|P61024|CKS1_HUMAN        79       2    tr|B5FXN4|B5FXN4_TAEGU      79       96
==================================================================================================
```

```
                10        20        30        40        50        60        70
MSHKQIYYSDKYDDEEFEYRHVMLPKDIAKLVPKTHLMSESEWRNLGVQQSQGWVHYMIHEPEPHILLFRRPLPKKPKK
MAHKQIYYSDKYDDEEFEYRHVMLPKDIAKLVPKTHLLSESEWRNLGVQQSQGWVHYMIHEPEPHILLFRRPLPKKPEK
```

Nucleotides

```
SeqA Name                        Len(nt)  SeqB Name                        Len(nt)  Score
==================================================================================================
1    ENA|CAA38702|CAA38702.1     240      2    ENA|ACH43795|ACH43795.1     240      81
==================================================================================================
```

```
        10        20        30        40        50        60        70        80        90        100       110       120
ATGTCGCACAAACAAATTTACTATTCGGACAAATACGACGACGAGGAGTTTGAGTATCGACATGTCATGCTGCCCAAGGACATAGCCAAGCTGGTCCCTAAAACCCATCTGATGTCTGAAT
ATGGCGCACAAACAGATCTACTACTCCGACAAGTACGACGATGAGGAGTTTGAGTACCGGCACGTGATGCTGCCCAAGGACATCGCCAAGCTGGTGCCCAAGACCCACCTGCTGTCGGAGT

        130       140       150       160       170       180       190       200       210       220       230
CTGAATGGAGGAATCTTGGCGTTCAGCAGAGTCAGGGATGGGTCCATTATATGATCCATGAACCAGAACCTCACATCTTGCTGTTCCGGCGCCCACTACCCAAGAAACCAAAGAAATGA
CGGAGTGGCGGAACCTGGGGGTGCAGCAGAGCCAGGGCTGGGTCCACTACATGATCCACGAGCCAGAGCCCCACATCCTGCTCTTCCGACGGCCACTGCCGAAGAAGCCGGAAAAGTGA
```

1) Codon degeneracy

Sequence conservations is

more stringent at the protein level,

than at the level of the

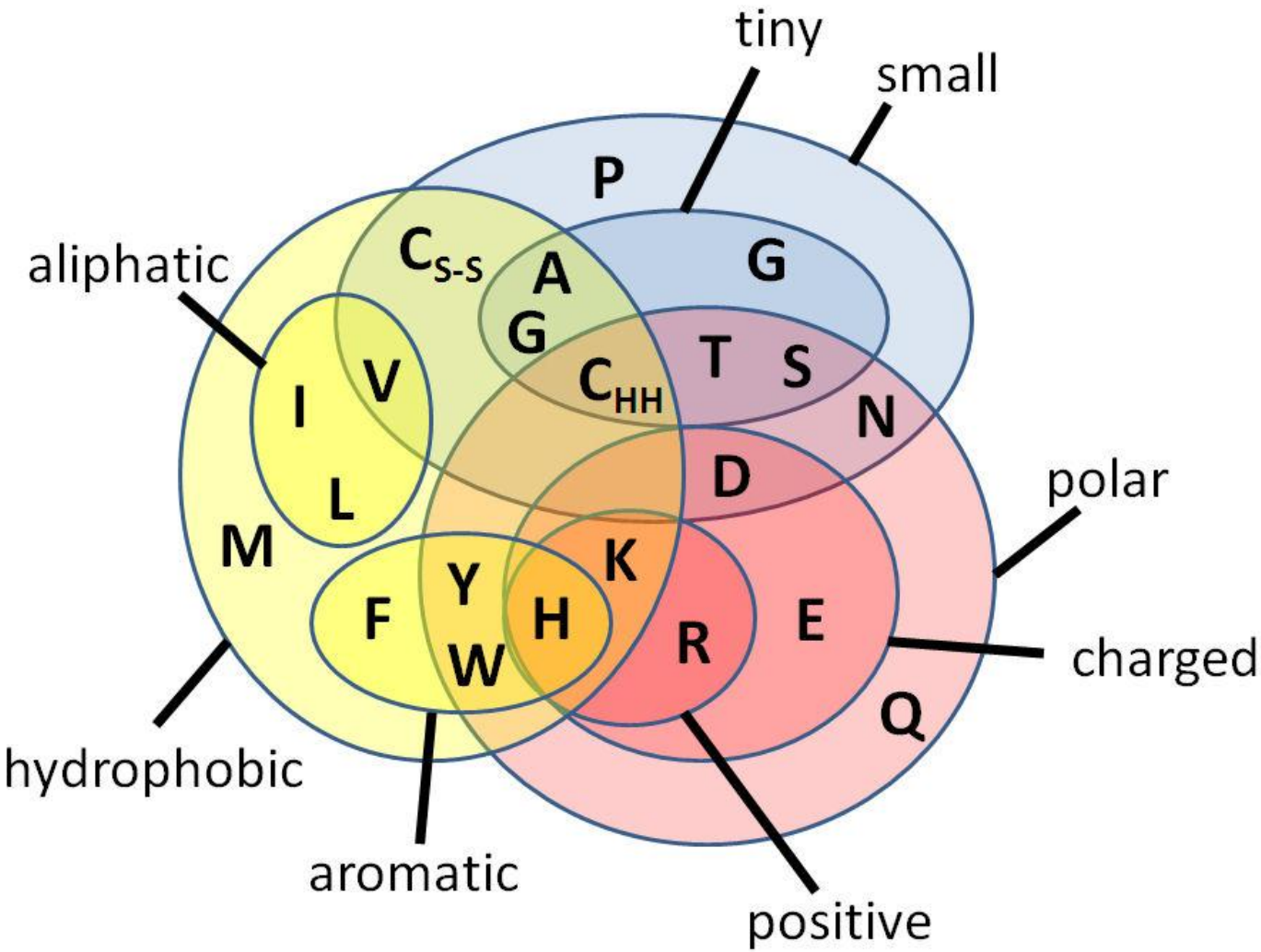
nucleotide coding sequence

2) Amino acid similarity

Protein sequence searches can distinguish between **exact, similar** and **dissimilar** matches

2) Amino acid similarity

Amino acids grouped by physical & chemical properties



2) Amino acid similarity

Protein searches take account of amino acid similarities

	highly conserved	weakly conserved	not conserved
Amino acids	Ser | Ser	Ser | Asn	Ser | Leu
	identical	**similar**	**mismatch**
Nucleotides	UCU | | | AGC	UCU | | | AAC	UCU | | | CUC
	mismatch	**mismatch**	**mismatch**

No distinction

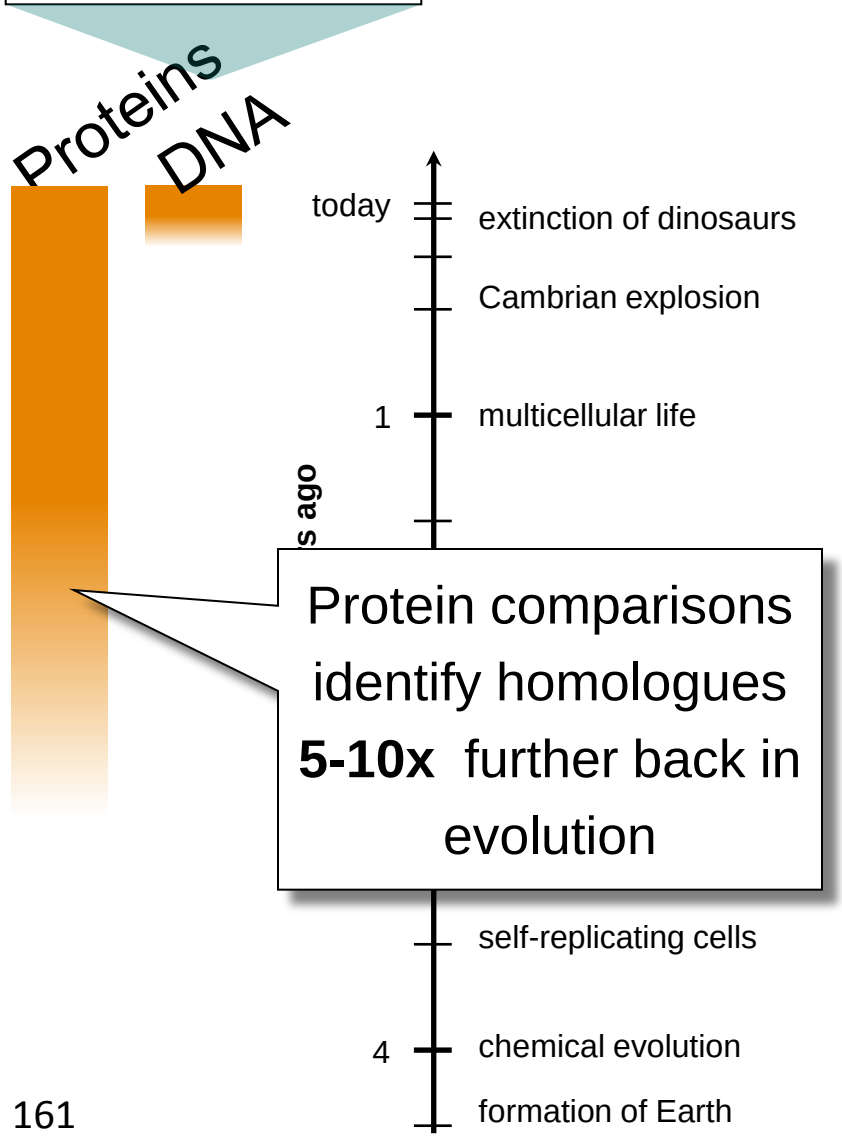
2) Amino acid similarity

- *Protein alignments* can score a conservative amino acid substitution differently from a non-conservative one through the use of **scoring matrices**

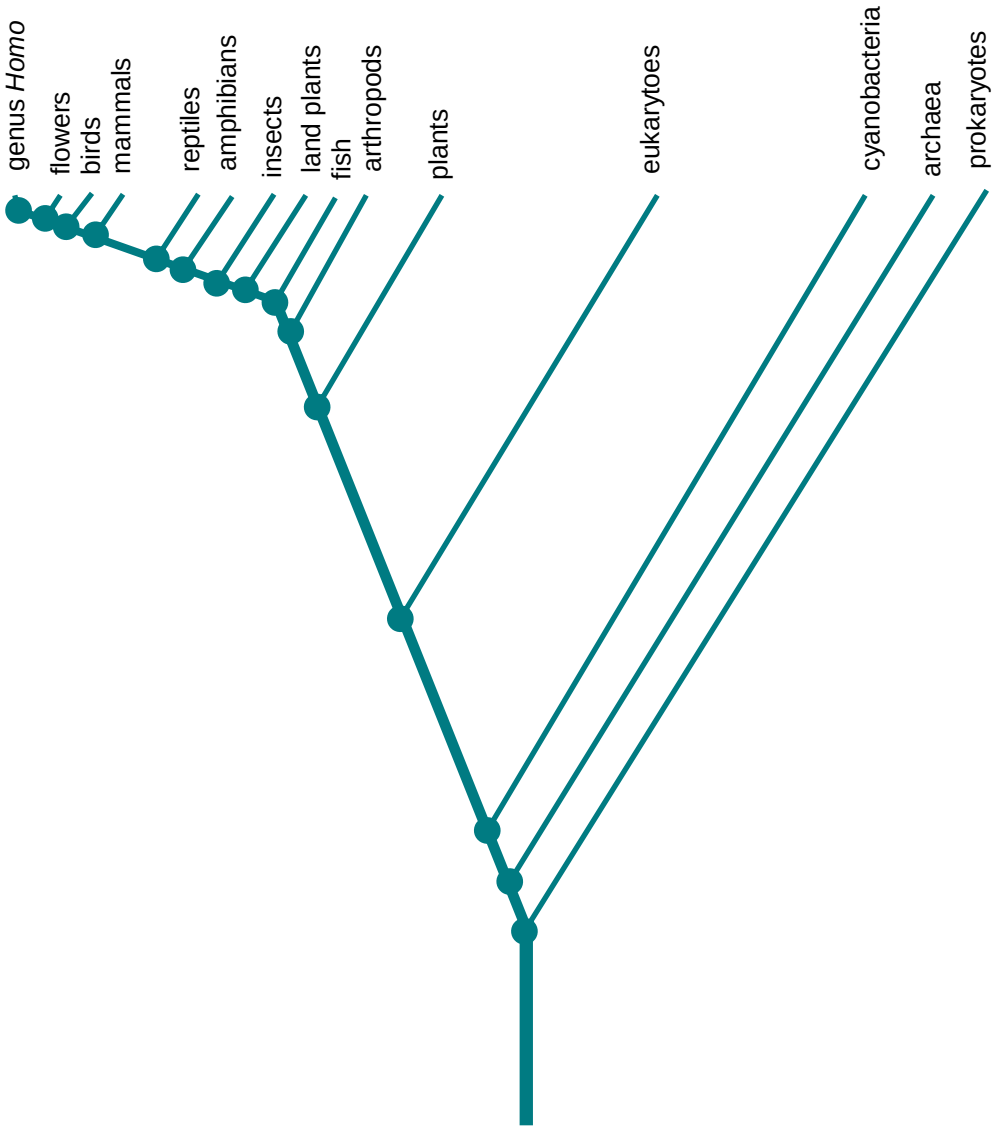
- By contrast, *nucleotide alignments* use **over-simple** (less sensitive) match/mismatch scoring

Protein v nucleotide search

Identify homologues searching:



Protein comparisons identify homologues **5-10x** further back in evolution



Protein v nucleotide: *example*

100% identity → more significant e-value for DNA match because a longer sequence

Codon degeneracy and **simple scoring** give rise to less significant e-values for DNA matches

			DNA e-value	Protein e-value
Human			4.0e-199	1.8e-92
Bovine		*nsferase P*	5.9e-154	8.0e-82
Mouse	GSTP2	*Glutathione S-transferase P2*	4.4e-133	4.7e-79
Toad	GSTP1	*Glutathione S-transferase P*	1.5e-55	9.8e-59
Frog	GSTP1	*Glutathione S-transferase P*	8.9e-33	2.6e-45
Nematode		*sferase P*	8.5e-10	4.2e-32
Rabbit		*sferase Mu*	1.3e-07	1.3e-20
Bovine			4.3	4.4e-17
Liver fluke		*nsferase MU*	-	3.2e-15
Hornworm		*nsferase 2*	-	3.2e-12
Fruit fly		*nsferase S1*	-	2.3e-08
Slime mold	GSTA2	*Glutathione S-transferase a2*	-	2.6e-05
Maize	GST1	*Glutathione S-transferase 1*	-	3.0e-01
Wheat	GSTA1	*Glutathione S-transferase 1*	-	1.9

Protein v nucleotide search

…therefore, if a patent claims both a nucleotide CDS and a protein sequence, the protein sequence could pull out many more homologues than the nucleotide CDS

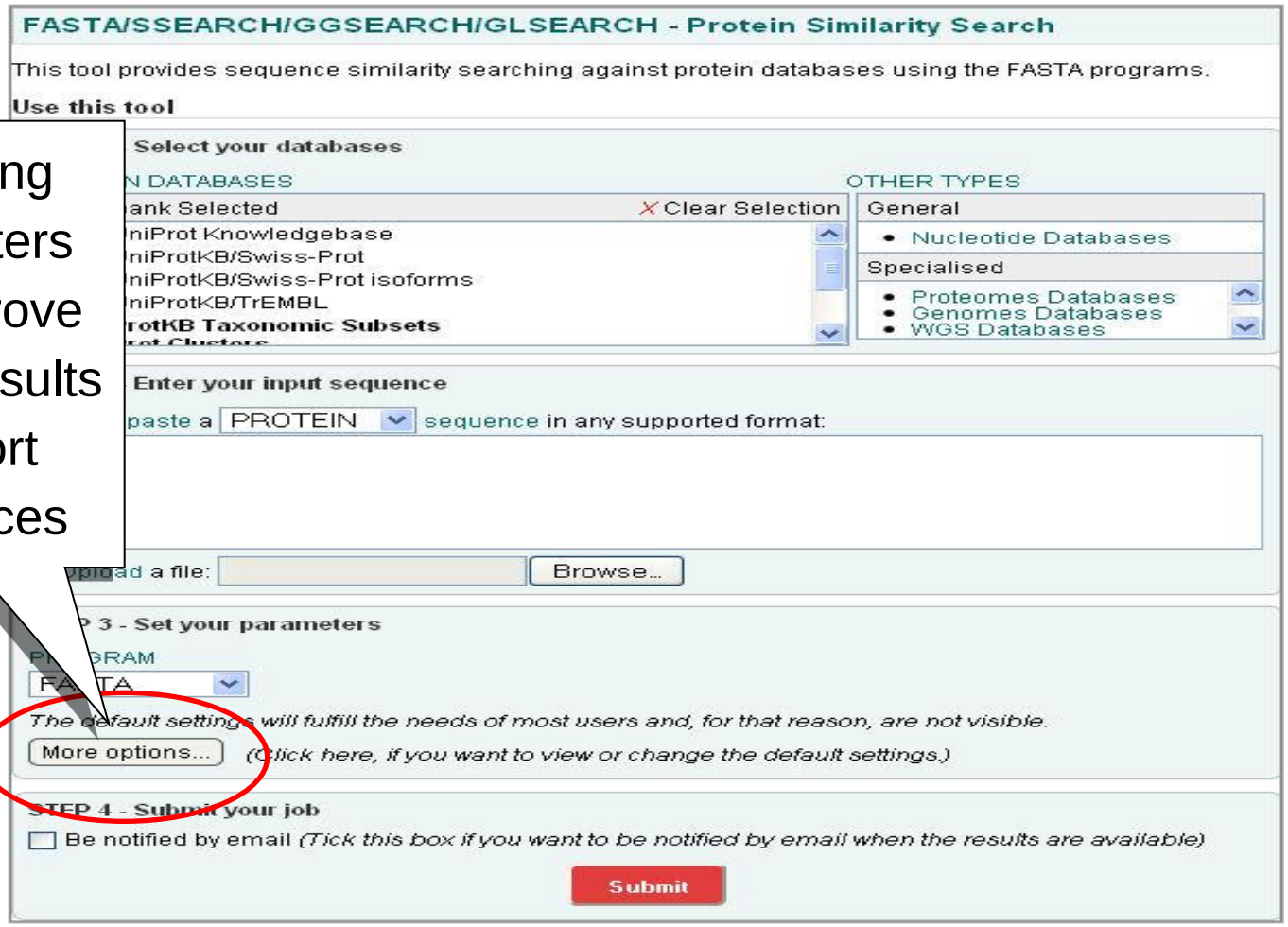
STEP 3:

Choosing search parameters to fit the task

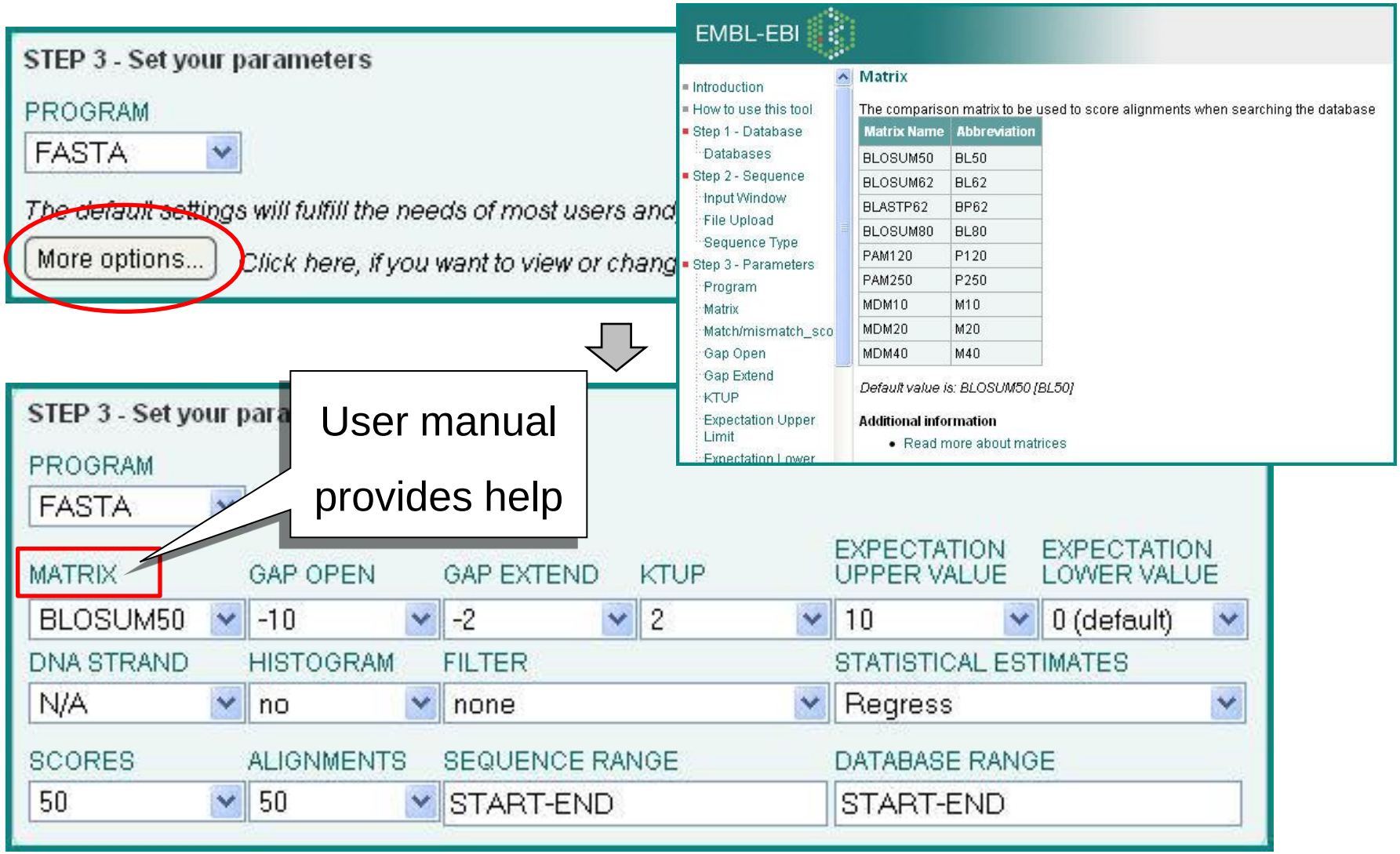
Choosing parameters

Parameters are set for searching a full-length protein or gene

Changing parameters can improve search results for short sequences

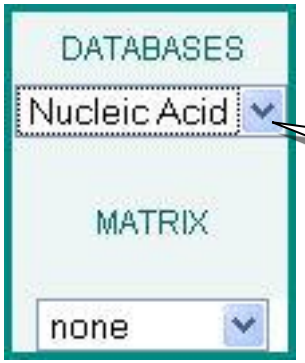


How to optimise parameters?

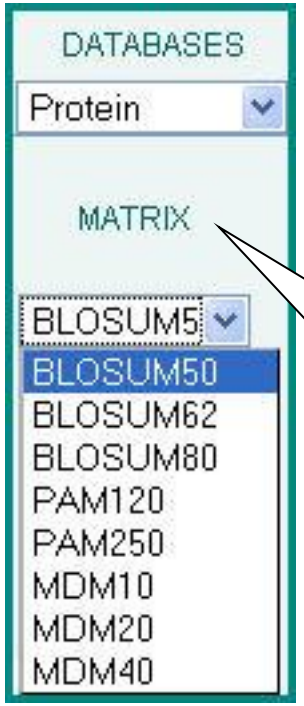


Which parameters to choose?

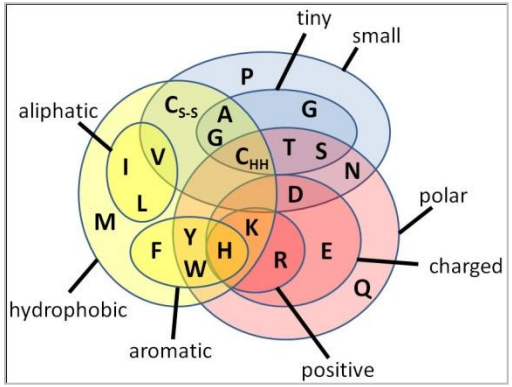
Matrix



Nucleotide search 'simpler' - only match/mismatch



Protein search uses substitution matrix tables (based on amino acid similarities and rate of change)



Protein matrices

Choice of <u>matrix</u> depends on:

1. strictness of search

BLOSUM 80	BLOSUM 62	BLOSUM 45
PAM 1	PAM 120	PAM 250

Less divergent ←——→ *More divergent*

2. length of query sequence

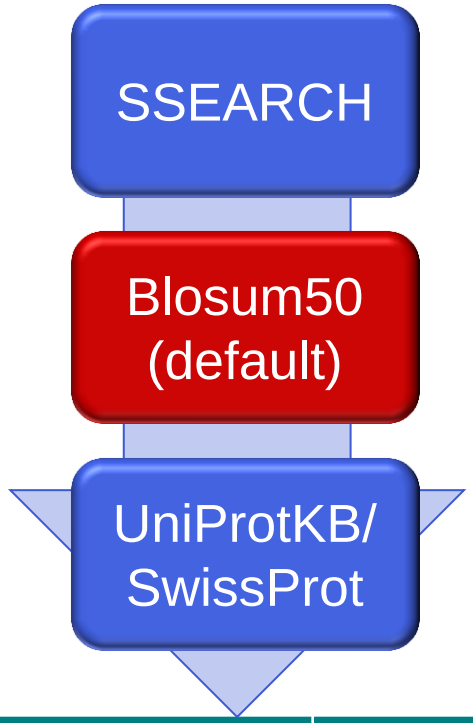
QUERY LENGTH	MATRIX	open	ext
>300	BLOSUM50	-10	-2
85-300	BLOSUM62	-7	-1
50-85	BLOSUM80	-16	-4
>300	PAM250	-10	-2
85-300	PAM120	-16	-4
35-85	MDM40	-12	-2
<=35	MDM20	-22	-4
<=10	MDM10	-23	-4

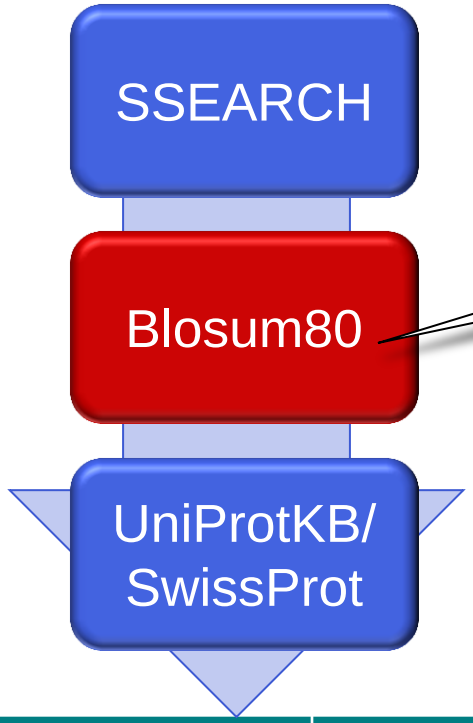
Example:

Comparing matrices when searching with a short peptide query sequence

query sequence: **MDM2_HUMAN**

Comparing matrices





Blosum80 *stricter* than Blosum50

hit	e()
MDM2_XENLA	4e-70
MDM4_BOVIN	9e-48
MDM4_DANRE	6e-17
XB34_ORYSJ	0.01
RN157_MOUSE	0.19
MGRN1_MOUSE	1.2
MGRN1_HUMAN	2.4

hit	e()
MDM2_XENLA	6e-109
MDM4_BOVIN	9e-51
MDM4_DANRE	5e-24
XB34_ORYSJ	0.05
RN157_MOUSE	
MGRN1_MOUSE	6.0
MGRN1_HUMAN	6.4

Blosom80 more significant (close) match

Blosom50 more significant (distant) match

Comparing matrices

Matrices:

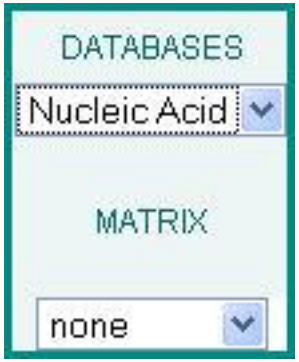
Use a high Blosum to find **close** matches,

a low Blosum to find **distant** matches;

Use MDM to find **longer** matches

Nucleotide match/mismatch

Matrix - protein

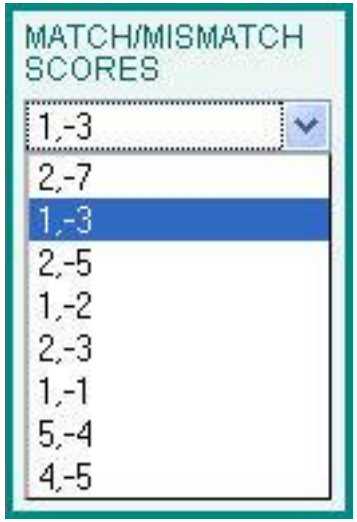


...instead have...

Match/mismatch - nucleotide

FASTA

BLAST



Nucleotide match/mismatch

- “Reward” for match, “penalty” for mismatch
- Reward/penalty ratio:
  - Increase ratio to find more divergent sequences:
  - Ratio of 0.33 (1/-3) for 99% conserved
  - Ratio of 0.5 (1/-2) for 95% conserved
  - Ratio of 1 (1/-1) for 75% conserved

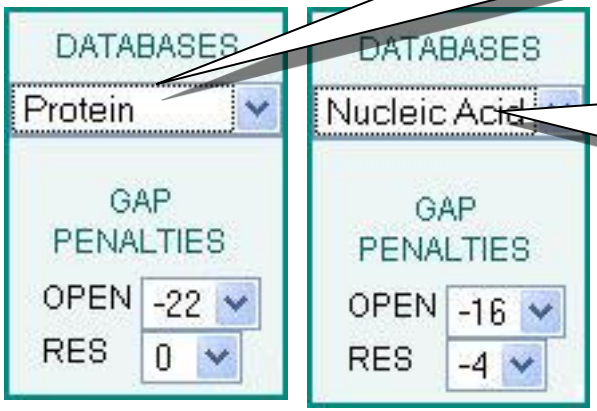
Gap penalties

gap penalties

Protein search

gap open = 0 to -23

Gap extension = 0 to -8



Nucleotide search

gap open = -2 to -16

Gap extension = 0 to -4

Gap penalties

Choice of <u>gap penalties</u> depends on:

1. strictness of search
   - larger penalty → fewer gaps

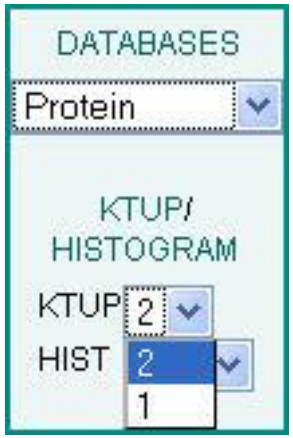
2. to match scoring matrix

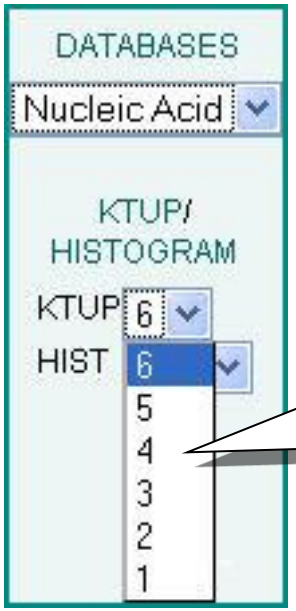
QUERY LENGTH	MATRIX	open	ext
>300	BLOSUM50	-10	-2
85-300	BLOSUM62	-7	-1
50-85	BLOSUM80	-16	-4
>300	PAM250	-10	-2
85-300	PAM120	-16	-4
35-85	MDM40	-12	-2
<=35	MDM20	-22	-4
<=10	MDM10	-23	-4

Word length - ktup

KTUP (word length)

- KTUP = 'word-length' of search
- Large word-length → less sensitive
  → faster





Nucleotide search: fewer bases than amino acids → higher KTUP

Example:

Comparing ktup when searching with a short RNA query sequence

query sequence: **23bp RNA**

Comparing ktup

query : **23bp RNA**

FASTA → ktup6 (default) → EMBL release

No hits found

FASTA → ktup3 → EMBL release

hit	
1: AB334817	0.12
2: AY238603	0.14
3: AC101743	0.16
4: AC115920	0.18

FASTA → ktup2 E=50 → EMBL release

hit	e()
1: AB334817	0.074
2: AY238603	0.092
3: AC101743	0.11
4: AC115920	0.12
5: BC098485	22
...	...
11: AL591512	32

Lower ktup is more sensitive

E=50 extends e-value cut-off

Comparing ktup

- Lowering ktup makes the search more sensitive
- Increase e-value cut-off for short sequences
- Increase match/mismatch score (+5/-4 for FASTA)
- Increase gap penalties

Masking

Do I mask my sequence?

Low complexity regions should be masked to avoid spurious results

- CA repeats
- poly-A tails
- proline-rich regions

****Be careful you don't mask what you are looking for**

Parameters for short sequences

What do I use for <u>short</u> sequences?

- use strict matrices
- use high gap penalties
- avoid masking
- allow high e-values

Adding value to your search results

Sequence search results page

Actions on all results

On selected results

Enter Text Here | Find | Help | Feedback

Research | Training | Industry | About Us | Help | Site Index

EBI > Tools > Sequence Similarity Searching > FASTA

FASTA Results

Summary Table | Tool Output | Visual Output | Functional Predictions | Submission Details | Submit Another Job

Alignments

Selection: Show Annotations | Hide Annotations | Show Alignments | Hide Alignments

Download in fasta format

Clear Selection | Select All | Invert Selection

Align.	DB:ID	Source	Length	Score	Identities	Positives	E()
☑ 1	NRPL2:NRP001827BA	PN:US6407209 B1	124	794	100.0	100.0	2.6E-49
☑ 2	NRPL1:NRP_A00210	PN:EP0242329 A2	124	794	100.0	100.0	2.6E-49
☑ 3	NRPL2:NRP001827B8	PN:EP0242329 A2	124	794	100.0	100.0	2.6E-49
☑ 4	NRPL2:NRP001827B9	PN:US5466585 A	124	794	100.0	100.0	2.6E-49
☑ 5	TR:C9JUL1_HUMAN	Uncharacterized protein OS=Homo sapiens GN=MX1 PE=3 SV=1 *Cross-references and related information in:* ▶ Nucleotide Sequences ▶ Genomes ▶ Ontologies ▶ Protein Families ▶ Literature	197	794	100.0	100.0	4.4E-49
☑ 6	TR:C9JN19_HUMAN	Uncharacterized protein OS=Homo sapiens GN=MX1 PE=3 SV=1 *Cross-references and related information in:* ▶ Nucleotide Sequences ▶ Genomes ▶ Ontologies ▶ Protein Families ▶ Literature	201	794	100.0	100.0	4.5E-49
☑ 7	NRPL2:NRP00221FC9	PN:US2007218466 A1	508	794	100.0	100.0	1.2E-48
☑ 8	NRPL1:NRP_ACH06771	PN:US2007218466 A1	508	794	100.0	100.0	1.2E-48

Visual output

Graphical display of results

Enter Text Here | Find | Help | Feedback

Research | Training | Industry | About Us | Help | Site Index

EBI > Tools > Sequence Similarity Searching > FASTA

FASTA Results

Summary Table | Tool Output | Visual Output | Functional Predictions | Submission Details | Submit Another Job

Alignments

Show Annotations | Hide Annotations | Show Alignments | Hide Alignments

Download in fasta format

Selection | Select All | Invert Selection

Align	DB:ID	Source	Length	Score	Identities	Positives	E()
1	NRPL2:NRP001827BA	PN:US6407209 B1	124	794	100.0	100.0	2.6E-49
2	NRPL1:NRP_A00210	PN:EP0242329 A2	124	794	100.0	100.0	2.6E-49
3	NRPL2:NRP001827B8	PN:EP0242329 A2	124	794	100.0	100.0	2.6E-49
4	NRPL2:NRP001827B9	PN:US5466585 A	124	794	100.0	100.0	2.6E-49
5	TR:C9JUL1_HUMAN	Uncharacterized protein OS=Homo sapiens GN=MX1 PE=3 SV=1 *Cross-references and related information in:* ▶ Nucleotide Sequences ▶ Genomes ▶ Ontologies ▶ Protein Families ▶ Literature	197	794	100.0	100.0	4.4E-49
6	TR:C9JN19_HUMAN	Uncharacterized protein OS=Homo sapiens GN=MX1 PE=3 SV=1 *Cross-references and related information in:* ▶ Nucleotide Sequences ▶ Genomes ▶ Ontologies ▶ Protein Families ▶ Literature	201	794	100.0	100.0	4.5E-49
7	NRPL2:NRP00221FC9	PN:US2007218466 A1	508	794	100.0	100.0	1.2E-48
8	NRPL1:NRP_ACH06771	PN:US2007218466 A1	508	794	100.0	100.0	1.2E-48

Visual output

Select sequence

View alignment

```
>>SP:GSTMU_RABIT P46409 Glutathione S-transferase Mu 1
OS=Oryctolagus cuniculus PE=1 SV=2 (218 nt)
 initn: 295 initl: 152 opt: 391  Z-score: 344.2  bits: 72.5 E(528048
Smith-Waterman score: 391; 33.2% identity (63.5% similar) in 211 nt

                40        70        100                 130       160
X06547 PYTVVYFPVRGRCAALRMLLADQGQSWKEEVVTV--------ETWQEGSLKASCLYGQLP
       :.:. :. :::    .::::     :..:.  :.          :   ..  .  . .::
SP:GST PMTLGYWDVRGLALPIRMLLEYTDTSYEEKKYTMGDAPNYDQSKWLSEKFTLGLDFPNLP
                10        20        30        40        50        60

            190       220       250       280       310       340
X06547 KFQDGDLTLYQSNTILRHLGRTLGLYGKDQQEAALVDMVNDGVEDLRCKYISLIYT-NYE
        . ::   : :::.:::.:.:  :: :. ..:   ::.... . : : . ... :. ..:
SP:GST YLIDGTHKLTQSNAILRYLARKHGLCGETEEERIRVDILENQLMDNRFQLVNVCYSPDFE
                70        80        90       100       110       120

            370       400       430       460       490       520
X06547 AGKDDYVKALPGQLKPFETLLSQNQGGKTFIVGDQISFADYNLLDLLLIHEVLAPGCLDA
        : .:.:.:: .:.    : ::  :.  ...::.:.:::. . :.:  .....::::::
SP:GST KLKPEYLKGLPEKLQ----LYSQFLGSLPWFAGDKITFADFLVYDVLDQNRIFVPGCLDA
            130           140       150       160       170
```

Visual output

Visual output provides an at-a-glance view of the length and position of all matches

e-values or % identity?

Enter Text Here | Find | Help | Feedback

Research | Training | Industry | About Us | Help | Site Index

EBI > Tools > Sequence Similarity Searching > FASTA

FASTA Results

Summary Table | Tool Output | Visual Output | Functional Predictions | Submission Details | Submit Another Job

Alignments

Selection: Show Annotations | Hide Annotations | Show Alignments | Hide Alignments

Download in fasta format

Clear Selection | Select All | Invert Selection

Align.	DB:ID	Source	Length	Score	Identities	Positives	E()
1	NRPL2:NRP001827BA	PN:US6407209 B1	124	794	100.0	100.0	2.6E-49
2	NRPL1:NRP_A00210	PN:EP0242329 A2	124	794	[illegible]	100.0	2.6E-49
3	NRPL2:NRP001827B8	PN:EP0242329 A2	[illegible]	[illegible]	100.0	100.0	2.6E-49
4	NRPL2:NRP[illegible]	[illegible]	[illegible]	794	100.0	100.0	2.6E-49
5	TR:C9JUL1[illegible]	[illegible] sapiens [illegible] information in: [illegible] Genomes [illegible] ▶ Literature	197	794	100.0	100.0	4.4E-49
6	TR:C9JN19[illegible]	[illegible] Homo sapiens GN=MX1 PE=3 SV=1 Cross-references and related information in: ▶ Nucleotide Sequences ▶ Genomes ▶ Ontologies ▶ Protein Families ▶ Literature	201	794	100.0	100.0	4.5E-49
7	NRPL2:NRP00221FC9	PN:US2007218466 A1	508	794	100.0	100.0	1.2E-48
8	NRPL1:NRP_ACH06771	PN:US2007218466 A1	508	794	100.0	100.0	1.2E-48

Do you use e-values or % identity?

e-values or % identity?

e-value is a better estimate of similarity than % identity, but patents use % identity

e-value → Estimates statistical significance of matches

- Default = 10 → expect 10 matches found by chance
- E() = 1-10 → frequently related
- E() = <0.01 → usually homologous

% identity → % of positions identical between query and match sequence

e-values or % identity?

example...

DB:ID	Source	Length	Identity%	Similar%	Overlap	E()
SW:ATP6_XENLA	ATP synthase subunit a OS=Xenopus	226	100.0	100.0	226	8.6e-85
SW:ATP6_DROYA	ATP synthase subunit a OS=Drosoph	224	36.6	70.3	232	9.4e-20
SW:ATP6_YEAST	ATP synthase subunit a OS=Sacchar	259	31.8	73.2	198	3.8e-16
SW:ATP6_EMENI	ATP synthase subunit a OS=Emerice	256	27.7	69.1	191	1.5e-11
SW:ATP6_COCHE	ATP synthase subunit a OS=Cochlio	257	27.6	67.3	196	1.7e-11
SW:ATP6_TRITI	ATP synthase subunit a OS=Triticu	386	32.9	70.5	146	1.2e-10
SW:ATPI_TOBAC	ATP synthase subunit a, chloropla	247	24.3	55.0	222	0.13
SW:ATP6_ECOLI	ATP synthase subunit a OS=Escheri	271	26.1	63.0	165	0.0084
SW:GTR5_SHEEP	Solute carrier family 2, facilita	501	26.3	57.1	156	2.9

Similar % identity scores

Different e-values

e-values or % identity?

example...

Pattern of conservation indicates homology

DB:ID
SW:ATP6_XENLA
SW:ATP6_DROYA
SW:ATP6_YEAST
SW:ATP6_EMENI
SW:ATP6_COCHE
SW:ATP6_TRITI
SW:ATPI_TOBAC
SW:ATP6_ECOLI
SW:GTR5_SHEEP

```
PTPLIPVLIIIETISLFIRPLALGVRLTANLT     LIQLIAT
PAILMPFMVCIETISNIIRPGTLAVRLTANMI     LLLTLLG-
PLPLVPLLVIIETLSYFARAISLGLRLGSNIL     LLMVILAG
PLGLLPLLVLIEFISYLSRNVSLGLRLAANIL     HMLLSILSG
PLGLLPLLVLIEFISYLARNVSLGLRLAANIL     HMLLNILSG
PLPLAPFLVLLELISYCFRALSLGIRLFANMM     AGHSLVKILSG
PLPLAPFLVLLELIPYCFRALSSGIRLFANMM     AGHSSVKILSG
HWAFIPVNLILEGVSLLSKPVSLGLRLFGNMYAGELIFILIAG
VISWMPYVSIACVISYVIGHALGPSPIPALLVTEIFLQSSRPA
```

No evidence of homology

Check alignments

Enter Text Here | Find | Help | Feedback

Research | Training | Industry | About Us | Help | Site Index

EBI > Tools > Sequence Similarity Searching > FASTA

FASTA Results

Summary Table | Tool Output | Visual Output | Functional Predictions | Submission Details | Sub

Alignments

Selection: Show Annotations | Hide Annotations | Show Alignments | Hide Alignments

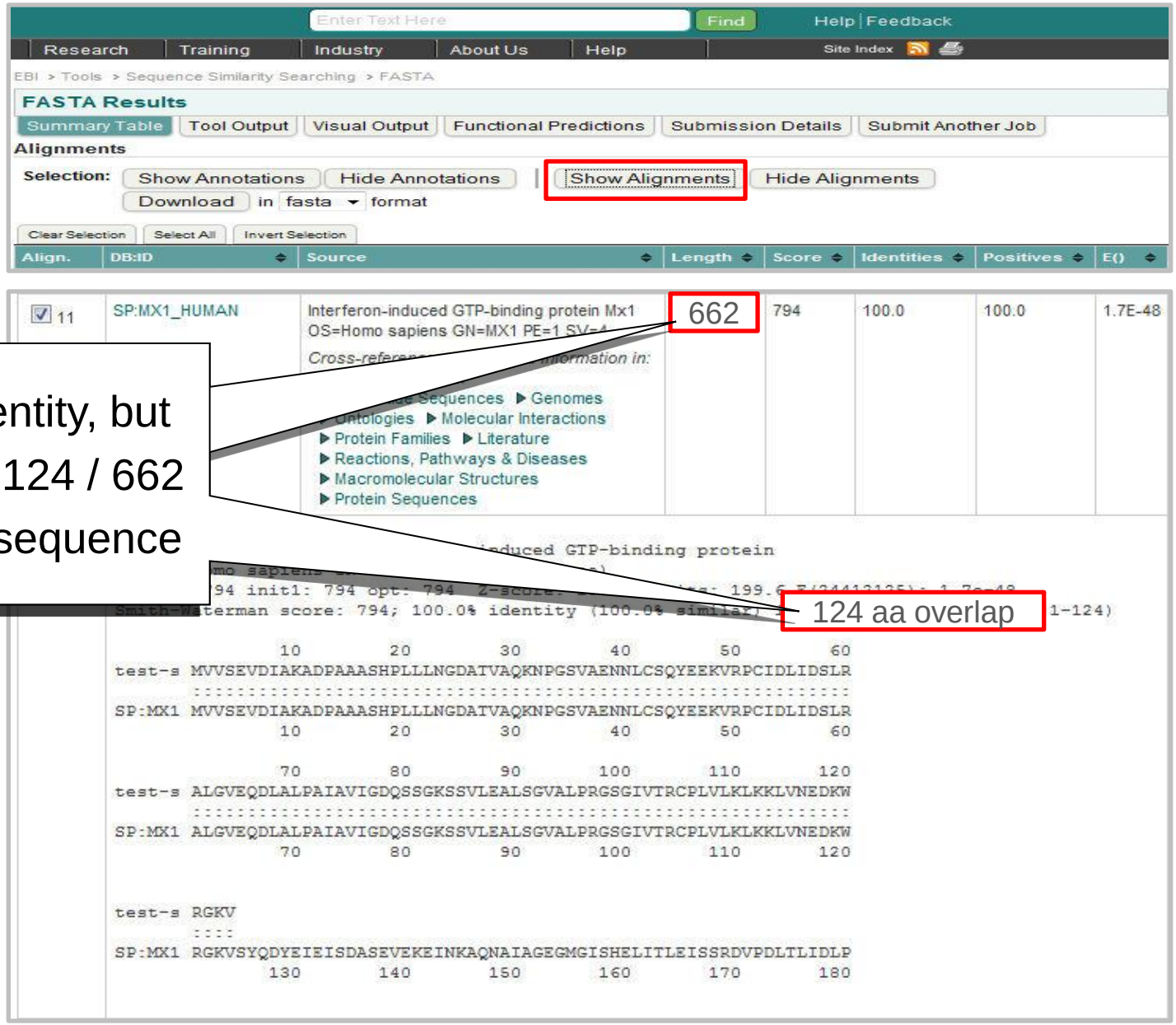
Download in fasta format

Clear Selection | Select All | Invert Selection

Align.	DB:ID	Source	Length	Score	Identities	Positives	E()
☑ 1	NRPL2:NRP001827BA	PN:US6407209 B1	124	794	100.0	100.0	2.6E-49
☑ 2	NRPL1:NRP_A00210	PN:EP0242329 A2	124	794	100.0	100.0	2.6E-49
☑ 3	NRPL2:NRP001827B8	PN:EP0242329 A2	124	794	100.0	100.0	2.6E-49
☑ 4	NRPL2:NRP001827B9	PN:US5466585 A	124	794	100.0	100.0	2.6E-49
☑ 5	TR:C9JUL1_HUMAN	Uncharacterized protein OS=Homo sapiens GN=MX1 PE=3 SV=1 *Cross-references and related information in:* ▶ Nucleotide Sequences ▶ Genomes ▶ Ontologies ▶ Protein Families ▶ Literature	197	794	100.0	100.0	4.4E-49
☑ 6	TR:C9JN19_HUMAN	Uncharacterized protein OS=Homo sapiens GN=MX1 PE=3 SV=1 *Cross-references and related information in:* ▶ Nucleotide Sequences ▶ Genomes ▶ Ontologies ▶ Protein Families ▶ Literature	201	794	100.0	100.0	4.5E-49
☑ 7	NRPL2:NRP00221FC9	PN:US2007218466 A1	508	794	100.0	100.0	1.2E-48
☑ 8	NRPL1:NRP_ACH06771	PN:US2007218466 A1	508	794	100.0	100.0	1.2E-48

Check alignments, especially if using a local-local algorithm

Check alignments

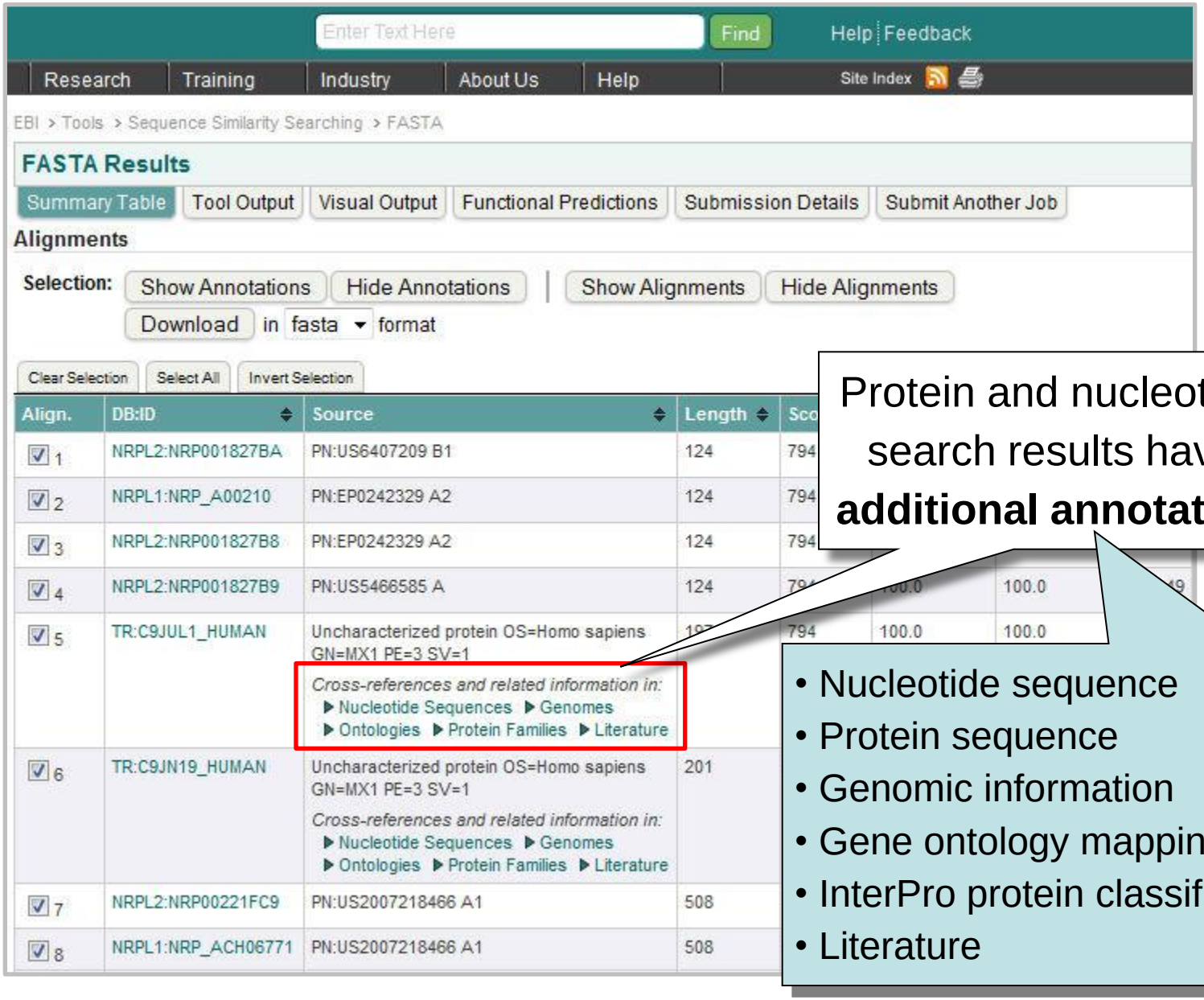


100% identity, but only over 124 / 662 (**20%**) of sequence

Check alignments

Always **check alignments** to see where and to what extent the query & target sequences match

Additional annotation



Protein and nucleotide search results have **additional annotation**

- Nucleotide sequence
- Protein sequence
- Genomic information
- Gene ontology mapping
- InterPro protein classification
- Literature

Additional annotation

Research | Training

EBI > Tools > Sequence Similarity S

FASTA Results

Summary Table | Tool Output

Alignments

Selection: Show Annotation

Download in f

Clear Selection | Select All | Invert S

Align.	DB:ID						
1	NRPL2:NRP001827BA						
2	NRPL1:NRP_A00210						
3	NRPL2:NRP001827B8						
4	NRPL2:NRP001827B9						
5	TR:C9JUL1_HUMAN	Uncharacterized pro…=Homo sapiens GN=MX1 PE=3 SV=… *Cross-references and related information in:* ▶ Nucleotide Sequences ▶ Genomes ▶ Ontologies ▶ Protein Families ▶ Literature	137	794	100.0	100.0	4.4E-45
6	TR:C9JN19_HUMAN	Uncharacterized protein OS=Homo sapiens GN=MX1 PE=3 SV=1 *Cross-references and related information in:* ▶ Nucleotide Sequences ▶ Genomes ▶ Ontologies ▶ Protein Families ▶ Literature	201	794	100		
7	NRPL2:NRP00221FC9	PN:US2007218466 A1	508	794	100.0	100.0	1.2E-48
8	NRPL1:NRP_ACH06771	PN:US2007218466 A1	508	794	100.0	100.0	1.2E-48

Nucleotide Sequences references for entry *C9JUL1_HUMAN* from *UniProtKB*

This entry has other cross references in: **InterPro - Medline - Taxonomy - GO - Ensembl - HGNC - EMBL-Bank**

3 results found in EMBL Release (Normal Divisions)

AP001610
Homo sapiens genomic DNA, chromosome 21, clone:CIT2533B8, MX1-D21S171 region.
View: ▶ in ENA ▶ in EMBL format ▶ in SRS ▶ in EMBL-SVA
References: ▶ EMBL-Bank ▶ Taxonomy ▶ InterPro ▶ UniProtKB ▶ PDBe ▶ HGNC

AL773577
Homo sapiens chromosome 21 from cosmid LL21NC02-14C10 map 21q22.2,D21S349-MX1
View: ▶ in ENA ▶ in EMBL format ▶ in SRS ▶ in EMBL-SVA
References: ▶ EMBL-Bank ▶ Taxonomy ▶ UniProtKB ▶ PDBe ▶ InterPro ▶ HGNC

AL773578
Homo sapiens chromosome 21 from PAC RP1-265B9 map 21q22.2,D21S349-MX1
View: ▶ in ENA ▶ in EMBL format ▶ in SRS ▶ in EMBL-SVA
References: ▶ EMBL-Bank ▶ Taxonomy ▶ InterPro ▶ UniProtKB ▶ HGNC

View in the UniProt website: ▶ C9JUL1_HUMAN

e.g. related ENA nucleotide entries

Functional predictions

Protein search results have 'Function Prediction'

Enter Text Here | Find | Help | Fe

Research | Training | Industry | About Us | Help | Site Inde

EBI > Tools > Sequence Similarity Searching > FASTA

FASTA Results

Summary Table | Tool Output | Visual Output | Functional Predictions | Submission Details | Submit Another Job

Alignments

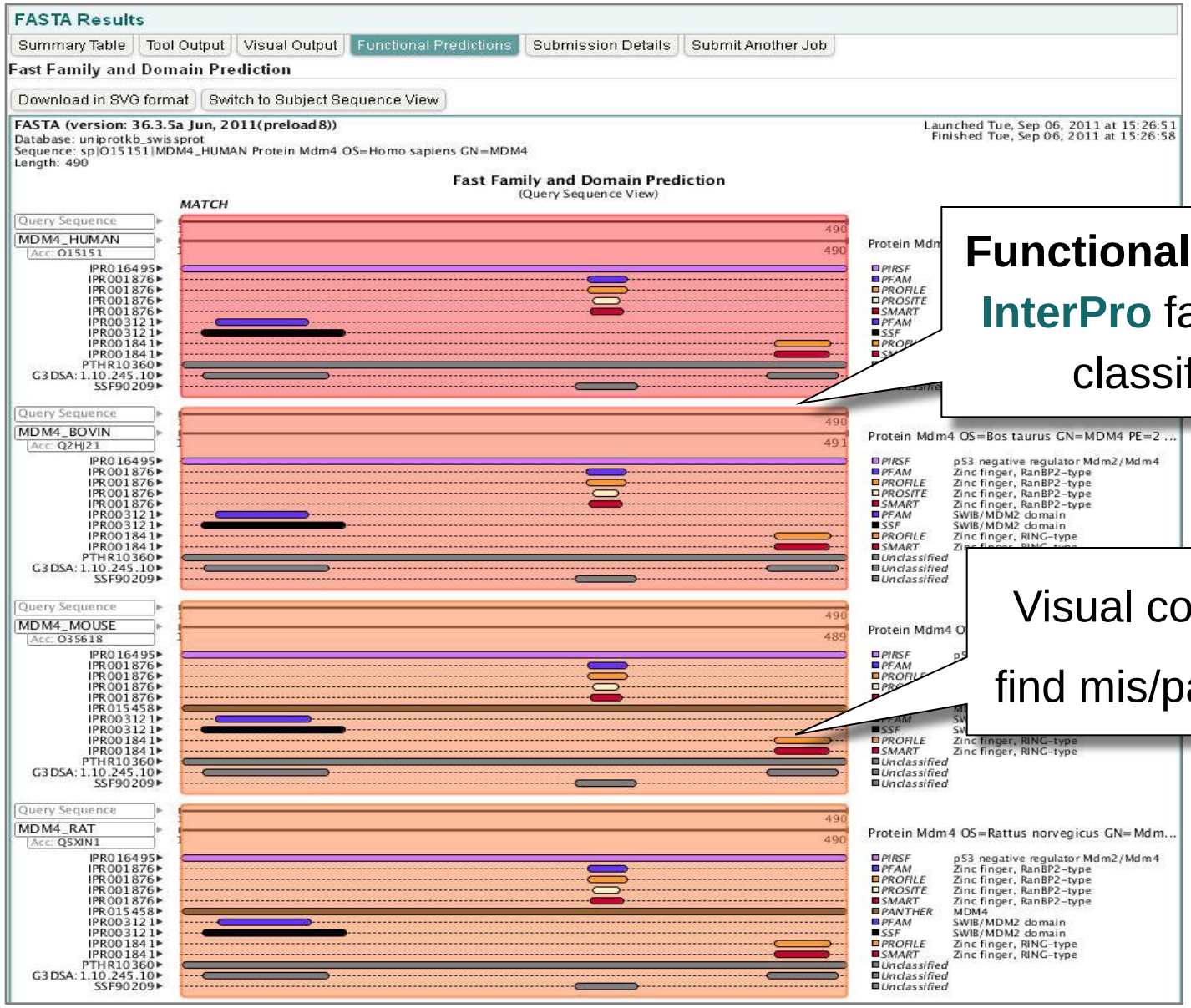
Selection: Show Annotations | Hide Annotations | Show Alignments | Hide Alignments

Download in fasta format

Clear Selection | Select All | Invert Selection

Align.	DB:ID	Source	Length	Score	Identities	Positives	E()
☑ 1	NRPL2:NRP001827BA	PN:US6407209 B1	124	794	100.0	100.0	2.6E-49
☑ 2	NRPL1:NRP_A00210	PN:EP0242329 A2	124	794	100.0	100.0	2.6E-49
☑ 3	NRPL2:NRP001827B8	PN:EP0242329 A2	124	794	100.0	100.0	2.6E-49
☑ 4	NRPL2:NRP001827B9	PN:US5466585 A	124	794	100.0	100.0	2.6E-49
☑ 5	TR:C9JUL1_HUMAN	Uncharacterized protein OS=Homo sapiens GN=MX1 PE=3 SV=1 *Cross-references and related information in:* ▶ Nucleotide Sequences ▶ Genomes ▶ Ontologies ▶ Protein Families ▶ Literature	197	794	100.0	100.0	4.4E-49
☑ 6	TR:C9JN19_HUMAN	Uncharacterized protein OS=Homo sapiens GN=MX1 PE=3 SV=1 *Cross-references and related information in:* ▶ Nucleotide Sequences ▶ Genomes ▶ Ontologies ▶ Protein Families ▶ Literature	201	794	100.0	100.0	4.5E-49
☑ 7	NRPL2:NRP00221FC9	PN:US2007218466 A1	508	794	100.0	100.0	1.2E-48
☑ 8	NRPL1:NRP_ACH06771	PN:US2007218466 A1	508	794	100.0	100.0	1.2E-48

Public site: function prediction

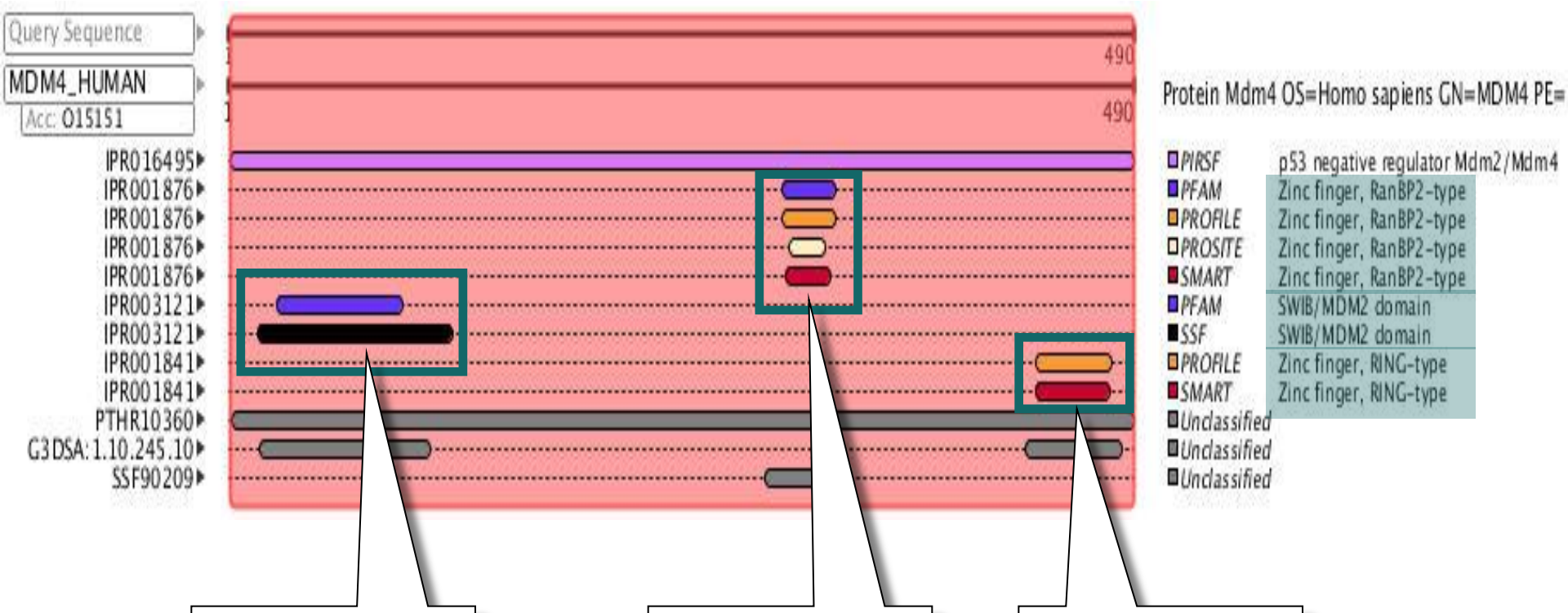


Functional predictions: **InterPro** family/domain classifications

Visual comparison → find mis/partial matches

InterPro annotation

Domain annotation



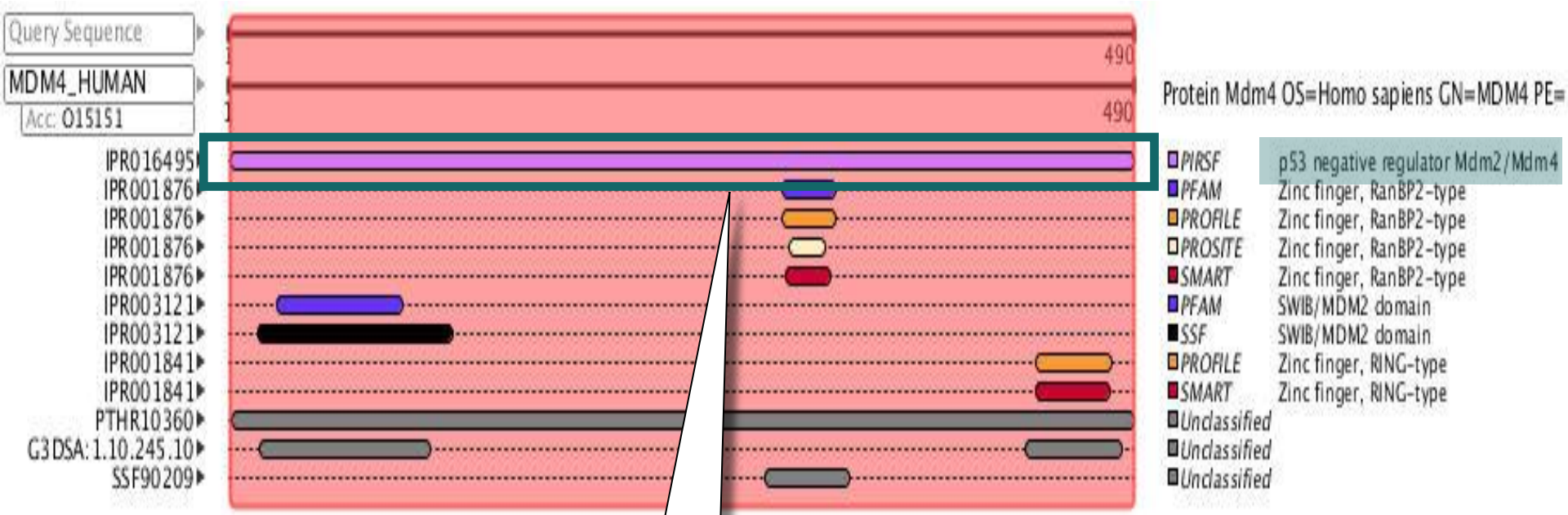
SWIB/MDM2 domain

RanBP2-type zinc finger

RING-type zinc finger

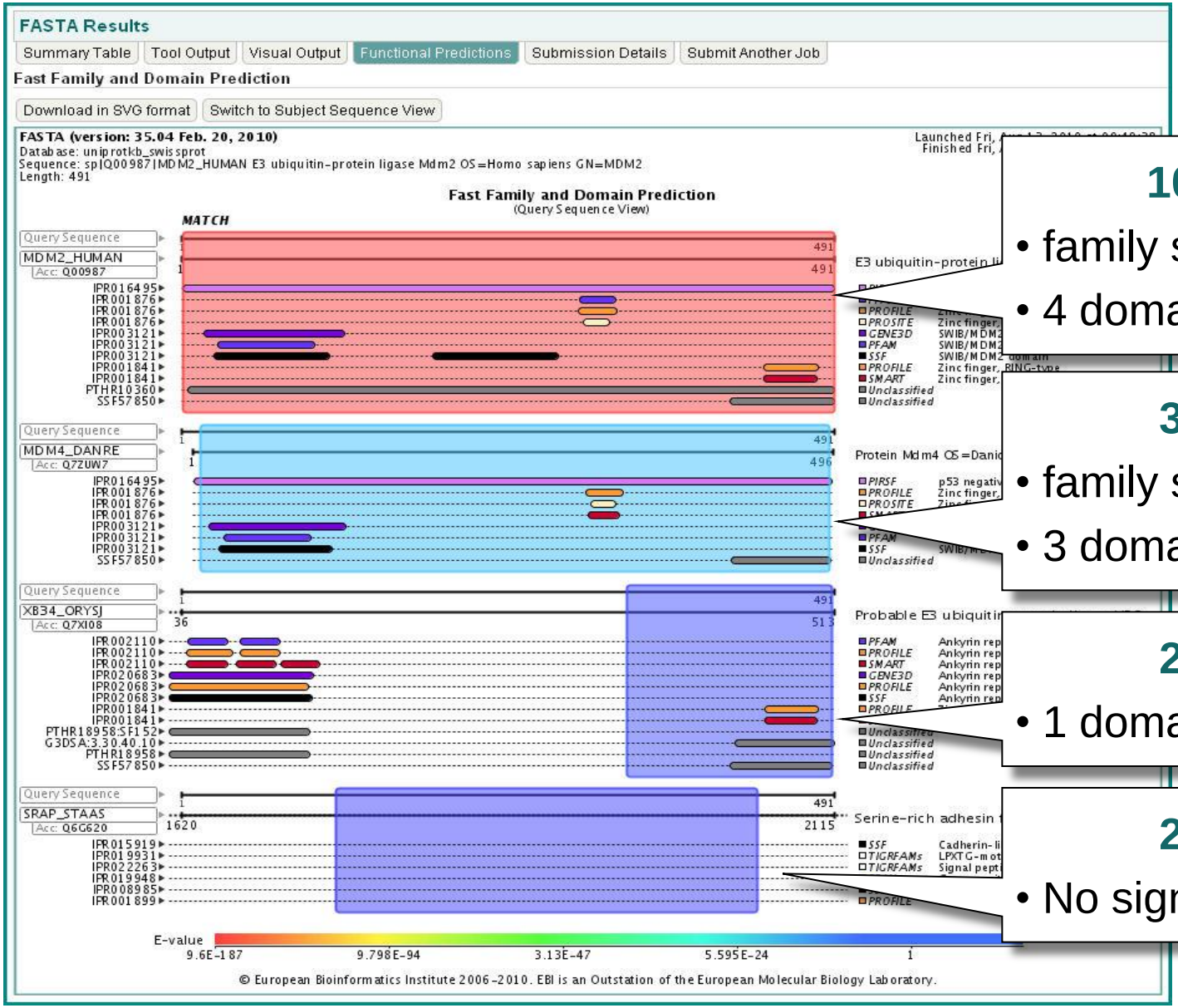
InterPro annotation

Family classification



Mdm2/Mdm4 family

Public site: function prediction



100% ID

- family signature
- 4 domain signatures

34% ID

- family signature
- 3 domain signatures

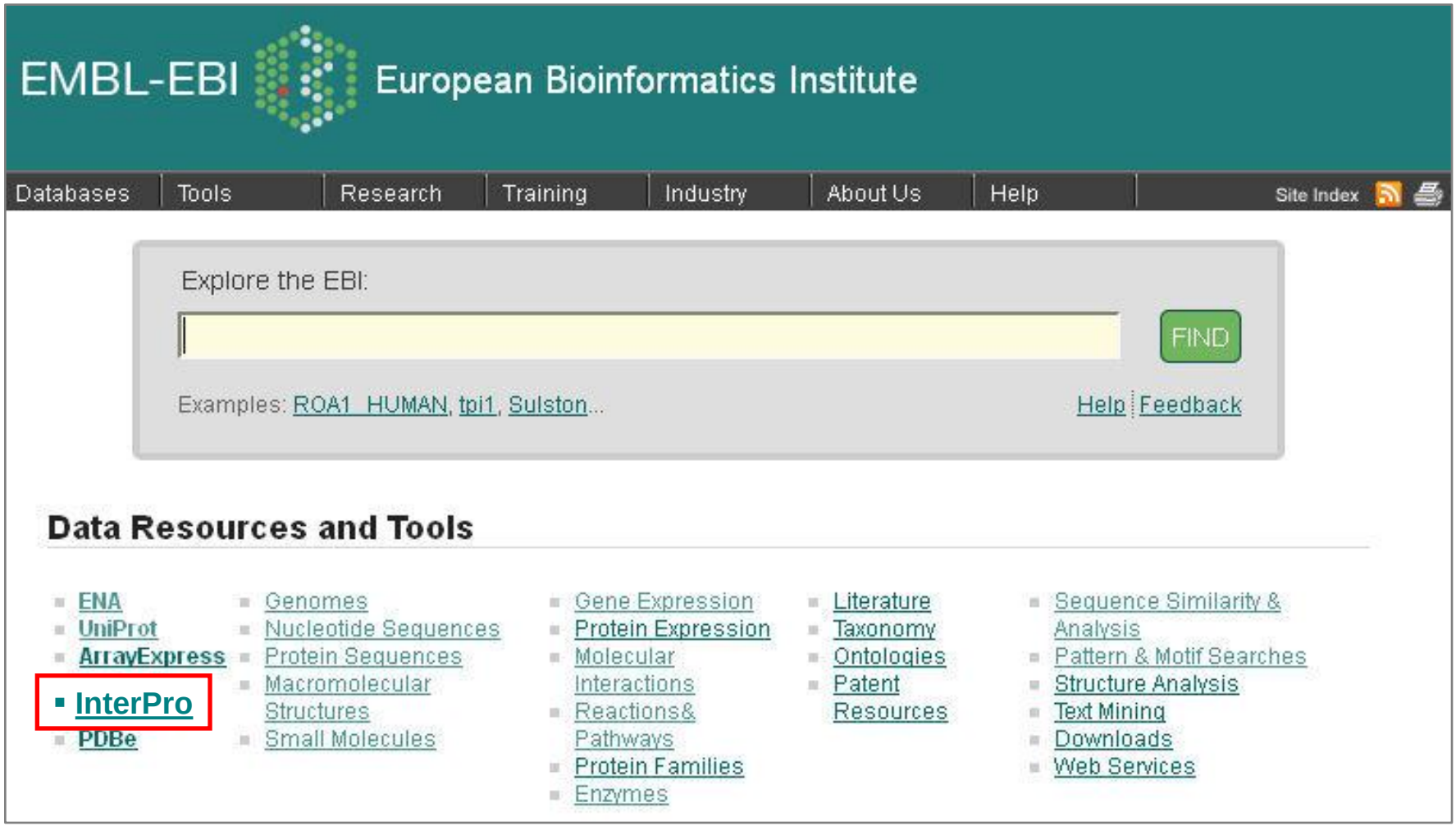
28% ID

- 1 domain signature

24% ID

- No signatures

InterPro: access directly



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

InterPro homepage

EMBL-EBI

Enter Text Here | Find | Help | Feedback

Databases | Tools | Research | Training | Industry | About Us | Help | Site Index

- InterPro:Home
  - Advanced Search
    - InterProScan
    - BioMart
  - Help / Documentation
    - About InterPro
    - Release Notes
    - BioMart Manual
    - Tutorial
    - Publications
    - Contributors
  - Web Services

EBI > Databases > InterPro

TRY NEW Interpro Beta!

InterPro protein sequence analysis & classification

InterPro is an integrated database of predictive protein "signatures" used for the classification and automatic annotation of proteins and genomes. InterPro classifies sequences at superfamily, family and subfamily levels, predicting the occurrence of functional domains, repeats and important sites. InterPro adds in-depth annotation, including GO terms, to the protein signatures.

Current release: **33.0 4th July 2011** (see Release Notes for further details)

Search InterPro: >

Do a sequence search of InterPro, via InterProScan

Extract large datasets by querying our BioMart

You can access our data programmatically, via Web Services

Use the updated InterProScan Web Service

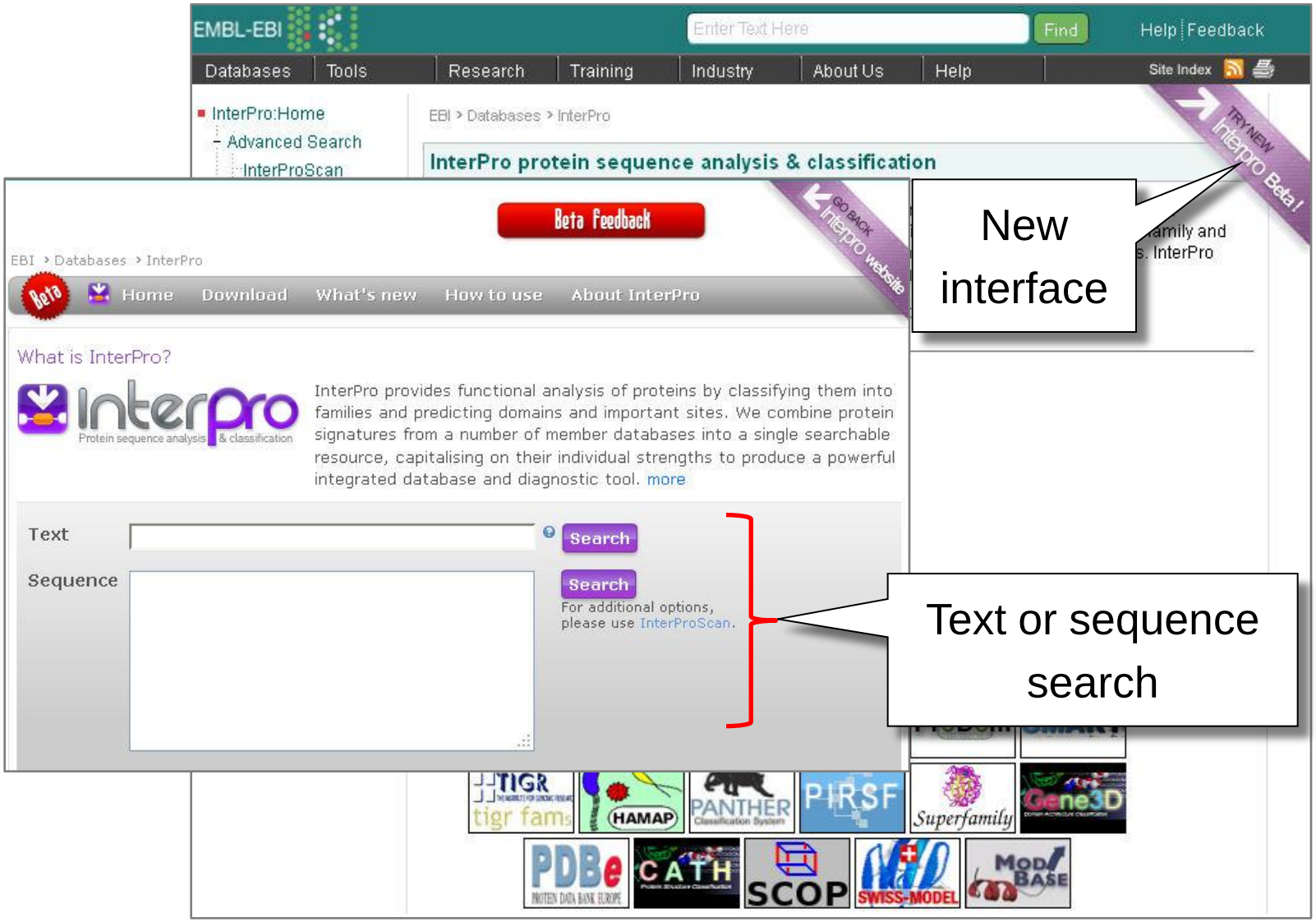
UniProt Universal Protein Resource | prosite | Pfam | PRINTS | ProDom | SMART | TIGR tigr fams | HAMAP | PANTHER | PIRSF | Superfamily | Gene3D | PDBe | CATH | SCOP | SWISS-MODEL | ModBase

>60,000 protein signatures from 11 member databases

Text or sequence search

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

InterPro homepage



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

InterProScan sequence search

EMBL-EBI

Enter Text Here | Find | Help | Feedback

Databases | Tools | Research | Training | Industry | About Us | Help | Site Index

- InterProScan
  - Help
  - Programmatic Access
- Download
- UniProt
- UniParc
- Similar Applications

InterProScan related literature

Search for InterProScan related literature in Medline...
more

EBI > Tools > Protein Functional Analysis > InterProScan

InterProScan Sequence Search

This form allows you to query your sequence against InterPro. For more detailed information see the documentation for the perl stand-alone InterProScan package (Readme file or FAQ's), or the InterPro manual or help pages.

Use this tool

STEP 1 - Enter your input sequence

Enter or paste a **PROTEIN** sequence in any supported format:

```
MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRY
YETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSV
SSINRVLRNLASEKQQMGADGMYDKLRMLNGQTGSWGTRPGWYPGTSVPGQPTQDGCQQQ
EGGGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQEQIEALEKEFERTHYPDVFAR
ERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRRQASNTPSHIPISSSFSTSVYQPIP
QPTTPVSSFTSGSMLGRTDTALTNTYSALPPMPSFTMANNLPMQPPVPSQTSSYSCMLPT
SPSVNGRSYDTYTPPHMQTHMNSQPMGTSGTTSTGLISPGVSVPVQVPGSEPDMSQYWPR
LQ
```

Or, upload a file: Browse...

STEP 2 - Select the applications to run

Select All Clear All

☑ BlastProDom	☑ FPrintScan	☑ HMMPIR	☑ HMMPfam	☑ HMMSmart
☑ HMMTigr	☑ ProfileScan	☑ HAMAP	☑ PatternScan	☑ SuperFamily
☑ SignalPHMM	☑ TMHMM	☑ HMMPanther	☑ Gene3D	

STEP 3 - Submit your job

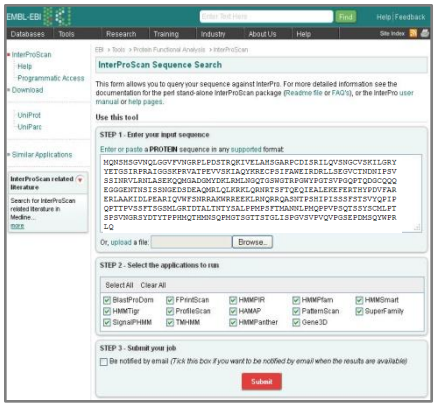
☐ Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

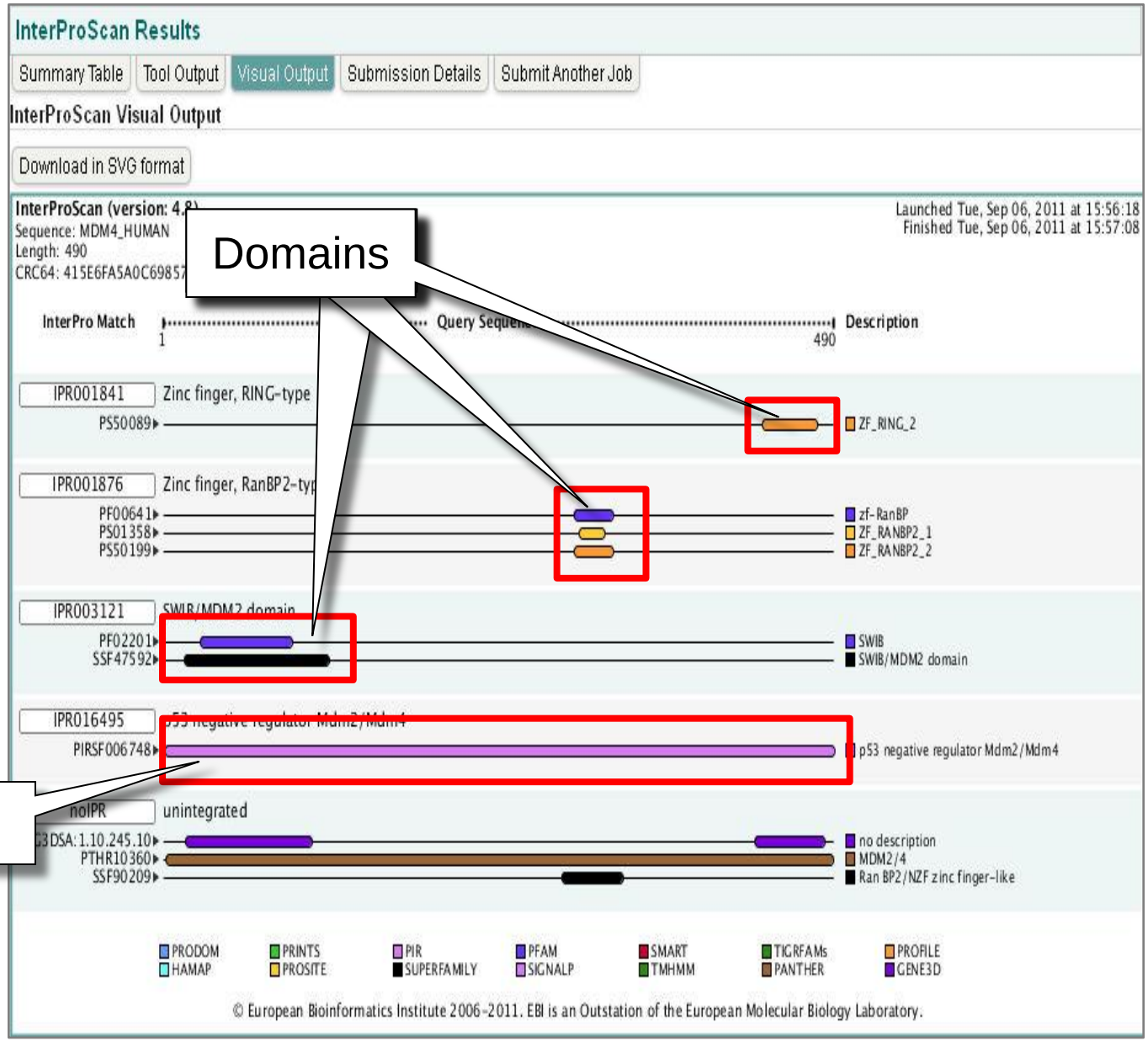
Submit

Download version takes both **protein and nucleotide** sequence

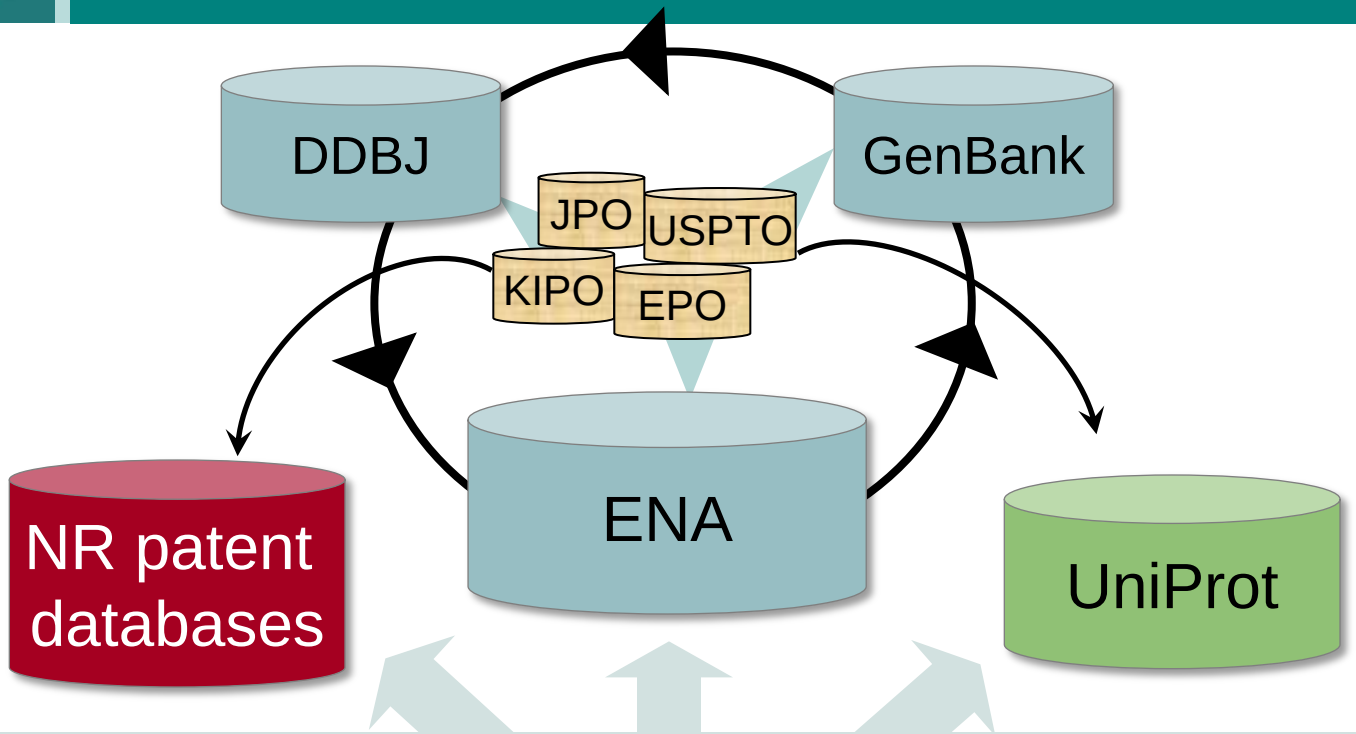
All search engines of member databases

InterProScan results





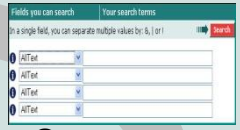
Recap



SRS text search

Select library tab

Select resources


Create query

patent **literature**

Sequence record

Sequence list

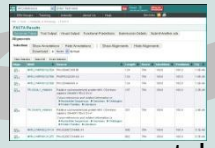
Sequence search

Navigate to sss tools

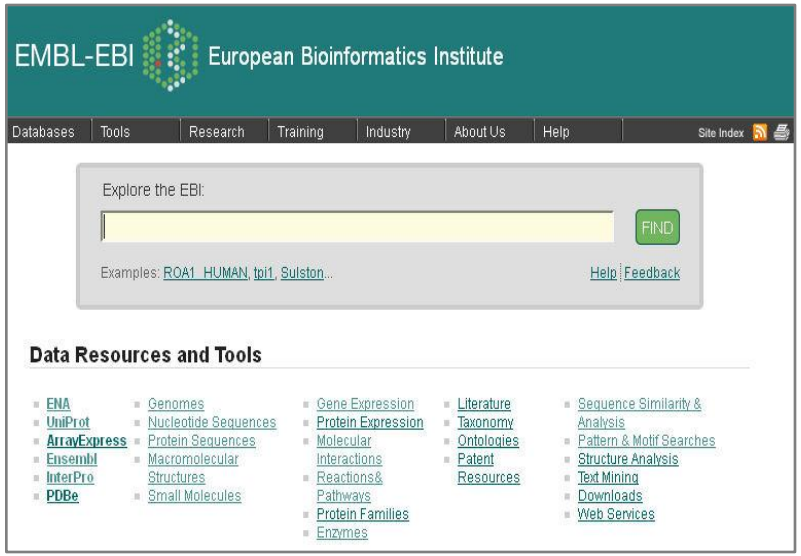
Select search tool

Create search

Functional predictions


Sequence matches

Summary



EBI provides free access to >100 databases

EBI provides specialised patent sequence databases

EBI provides multiple sequence search options

EBI provides advanced text search options

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

Help

Contacts:

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