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 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

- GenBank + SRA
- DDBJ
- ENA
- INSDC

(U.S.A.)

(Japan)

(Europe)
Primary sequence databases

Primary data submitted to databases

- GenBank + SRA
- DDBJ
- ENA

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

1) EMBL-Bank

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

Old **WGS** entries archived
Old **CON** entries archived

**STD**

- Standard
All INSDC databases use NCBI taxonomy

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

Only **sequenced** organisms represents

Some species EXCLUDED from certain taxonomic ranges

**ENA taxonomy**

- **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
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*
EMBL-Bank:

ENA Database
EMBL-Bank:

<table>
<thead>
<tr>
<th>Data classes</th>
<th>CON</th>
<th>EST</th>
<th>GSS</th>
<th>HTC</th>
<th>HTG</th>
<th>MSA</th>
<th>PAT</th>
<th>STS</th>
<th>STD</th>
<th>TSA</th>
<th>WGA</th>
</tr>
</thead>
</table>

1st: Data split into classes
**EMBL-Bank:**

Database structure

<table>
<thead>
<tr>
<th>Taxonomic Divisions</th>
</tr>
</thead>
<tbody>
<tr>
<td>HUM</td>
</tr>
<tr>
<td>MUS</td>
</tr>
<tr>
<td>ROD</td>
</tr>
<tr>
<td>MAM</td>
</tr>
<tr>
<td>VRT</td>
</tr>
<tr>
<td>FUN</td>
</tr>
<tr>
<td>INV</td>
</tr>
</tbody>
</table>

**Data classes**

<table>
<thead>
<tr>
<th></th>
<th>CON</th>
<th>EST</th>
<th>GSS</th>
<th>HTC</th>
<th>HTG</th>
<th>MSA</th>
<th>PAT</th>
<th>STS</th>
<th>STD</th>
<th>TSA</th>
<th>WGA</th>
</tr>
</thead>
<tbody>
<tr>
<td>HUM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MUS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ROD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VRT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FUN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1\textsuperscript{st}: Data split into classes

2\textsuperscript{nd}: Data split into intersecting slices by taxonomy

Reduces search set
# Database structure

**EMBL-Bank:**

- **Data classes**
- **Taxonomic Divisions**

<table>
<thead>
<tr>
<th></th>
<th>CON</th>
<th>EST</th>
<th>GSS</th>
<th>HTC</th>
<th>HTG</th>
<th>MSA</th>
<th>PAT</th>
<th>STS</th>
<th>STD</th>
<th>TSA</th>
<th>WGA</th>
</tr>
</thead>
<tbody>
<tr>
<td>HUM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MUS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ROD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VRT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FUN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **1<sup>st</sup>: Data split into classes**
- **2<sup>nd</sup>: Data split into intersecting slices by taxonomy**

---

**Click to edit Master text styles**

**Second level**

**Third level**

**Fourth level**

**Fifth level**

---

'**Mouse’ + ‘EST’ intersection**'

**Reduces search set**
European Nucleotide Archive

ENA is accessible from the EBI homepage

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

ENA homepage

- 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
- **Download data**
- **Dates (first public and last updated)**
- **Graphical viewer**
- **DNA source**
- **Navigate to external data sources** e.g. UniProt

**EMBL-EBI**

**Overview**

- **Source Feature**
- **Other Features**
- **References**

**Molecule type**
- **Topology**
- **Data class**
- **Sequence length**
- **Sequence Version**
- **First public**
- **Last updated**

**Navigation**

- **Taxon**
- **StrainInfo**
- **Visible feature range**

**Source Feature(s)**

- **Source**
- **Taxon**
- **strain**
- **name**

**Other Features**

- **Visible feature range**
- **Gene**
- **product**
- **protein**

**References**

- **Related publication**

**Sequence**

- **Visible sequence range**

---

Non-patent entry in EMBL-Bank

<table>
<thead>
<tr>
<th>Assembly information</th>
<th>General information</th>
<th>More detailed graphical view</th>
<th>Genome annotation</th>
</tr>
</thead>
</table>

Non-patent entry in EMBL-Bank

Additional information
ENA graphical viewer
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 is a comprehensive resource for nucleotide sequence data, but it is better for **non-patent data**
European nucleotide archive

Non-redundant patent sequence databases
Where does the data come from?

Sequence sources:
- ENA
- PDB
- RefSeq
- Ensembl
- VEGA
- Patents
- Model organisms
- more...

UniParc

exchange data daily
UniProt has a 3-tiered structure

Sequence sources

- ENA
- PDB
- RefSeq
- Ensembl

UniParc

- Metagenomic & environmental
- Taxonomy known

UniMES

- Automatic annotation
- Manual annotation
- Remove redundancy

UniProtKB/SwissProt

- History of sequences
- High quality annotation

Metagenomic projects

Sequence sources include:
- ENA
- PDB
- RefSeq
- Ensembl

Additional sources:
- Patents
- Model organisms
- More...

UniParc acts as a central repository for sequence data, with different tiers representing different types of data and annotation methods.
UniProt has a 3-tiered structure

Sequence sources:

- ENA
- PDB
- RefSeq
- Ensembl
- VEGA
- Patents
- Model organisms
- more…

UniParc

UniMES

Metagenomic & environmental

Taxonomy known

UniProtKB/TrEMBL

UniProtKB/SwissProt

UniRef Clusters

UniMES Clusters
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
### Patents not found in UniProtKB

- **Accession**
- **List of databases containing sequence**
- **Navigate to individual entries**

### Download data

### Deleted entries identified (greyed out)

### Sequence

**Sequence**

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Length</th>
<th>Mass (Da)</th>
</tr>
</thead>
<tbody>
<tr>
<td>UPI00002B7A2</td>
<td>124</td>
<td>13,013</td>
</tr>
</tbody>
</table>

**Chromat:** 3356A8E3A4AF65F

```
   10  20  30  40  50  60  70  80  90  100  110  120
  MVGSVEVDAGDPAAASQELALMDDATPAQKNEEESVAEKNLCSQFEKVRQPDGIDQSL
  ALQMEIQALPAVLNSGSITMLSDROSALSGVALFSLPSGIVTRCPFVLRKTXLKLHEEQV
```

### Customize display

- **Entries**
- **Sequence**

**Entries**

- **Found in 4 databases: REMTREMBL, EMBL CDS, USPTO and EPD**
- **Hide inactive entries (1)**

<table>
<thead>
<tr>
<th>Database</th>
<th>Entry</th>
<th>Version</th>
<th>Organism</th>
<th>First seen</th>
<th>Last seen</th>
<th>Active</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMBL CDS</td>
<td>CAA00041</td>
<td>1</td>
<td>Homo sapiens (Human)</td>
<td>2003-03-12</td>
<td>2010-06-06</td>
<td>Yes</td>
</tr>
<tr>
<td>REMTREMBL</td>
<td>E38125</td>
<td></td>
<td></td>
<td>2003-03-28</td>
<td>2003-11-17</td>
<td>No</td>
</tr>
<tr>
<td>EPO</td>
<td>A000210</td>
<td>1</td>
<td>Homo sapiens (Human)</td>
<td>2003-03-26</td>
<td>2010-06-10</td>
<td>Yes</td>
</tr>
<tr>
<td>USPTO</td>
<td>AAN19596</td>
<td>1</td>
<td></td>
<td>2003-03-26</td>
<td>2010-06-16</td>
<td>Yes</td>
</tr>
<tr>
<td>USPTO</td>
<td>AAC71736</td>
<td>1</td>
<td></td>
<td>2004-07-05</td>
<td>2010-06-16</td>
<td>Yes</td>
</tr>
</tbody>
</table>
## Browsing a UniProtKB/TrEMBL entry

### Names and origin

<table>
<thead>
<tr>
<th>Protein names</th>
<th>Submitted name: Lsk protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene names</td>
<td>Name: Lsk</td>
</tr>
<tr>
<td>Organism</td>
<td>Homo sapiens (Human)</td>
</tr>
<tr>
<td>Taxonomic identifier</td>
<td>9606 [NCBI]</td>
</tr>
</tbody>
</table>

### Taxonomic lineage

- Eukaryota
- Metazoa
- Chordata
- Craniata
- Vertebrata
- Euteleostomi
- Mammalia
- Eutheria
- Euarchontoglires
- Primates
- Haplorhini
- Catarrhini
- Hominidae
- Homo

### General annotation (Comments)

<table>
<thead>
<tr>
<th>Catalytic activity</th>
<th>ATP + a [protein]-L-tyrosine = ADP + a [protein]-L-tyrosine phosphate, ([SAAS SAAS020985])</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence similarities</td>
<td>Contains 1 SH2 domain: ([SAAS SAAS020985])&lt;br&gt;Contains 1 SH3 domain: ([SAAS SAAS020985])&lt;br&gt;Contains 1 protein kinase domain: ([SAAS SAAS020985])</td>
</tr>
</tbody>
</table>
Browsing a UniRef90 entry

Faster and more sensitive sequence search with no loss of information

<table>
<thead>
<tr>
<th>Accession</th>
<th>Status</th>
<th>Cluster name</th>
</tr>
</thead>
<tbody>
<tr>
<td>UniRef90_G9NYL2</td>
<td></td>
<td>Cluster: Mitogen-activated protein kinase kinase MLT</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Size</th>
<th>Members</th>
<th>Organisms</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>Q9NYL2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UP10001C4EAC4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>D2HJF2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Q9ESL4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>A2AR41</td>
<td></td>
</tr>
<tr>
<td></td>
<td>E1B8N9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>E2REE9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B8ZZU2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UP100005A58D7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+5</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Homo sapiens (Human)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oryctolagus cuniculus (Rabbit)</td>
</tr>
<tr>
<td>Ailuropoda melanoleuca (Giant panda)</td>
</tr>
<tr>
<td>Mus musculus (Mouse)</td>
</tr>
<tr>
<td>Bos taurus (Bovine)</td>
</tr>
<tr>
<td>Canis familiaris (Dog)</td>
</tr>
<tr>
<td>(Canis lupus familiaris)</td>
</tr>
<tr>
<td>Equus caballus (Horse)</td>
</tr>
<tr>
<td>Callithrix jacchus (White-tufted-ear marmoset)</td>
</tr>
<tr>
<td>Macaca mulatta (Rhesus macaque)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Length</th>
<th>Identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>600</td>
<td>50%</td>
</tr>
</tbody>
</table>
**All kingdoms:**

- **Bacteria** (61%)
- **Archaea** (4%)
- **Viruses** (3%)
- **Eukaryota** (32%)

**Within Eukaryota:**

- **Fungi** (18%)
- **Viridiplantae** (18%)
- **Insecta** (5%)
- **Other** (8%)
- **Nematoda** (2%)
- **Homo** (12%)
- **Other mammals** (27%)
- **Other vertebrata** (10%)
- **Other** (8%)
- **Vertebrata** (10%)

**Taxonomic distribution of species**
### Mainly model organisms

#### 2.2 Table of the most represented species

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

<table>
<thead>
<tr>
<th>Evidence at protein level</th>
<th>13%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evidence at transcript level</td>
<td>12%</td>
</tr>
<tr>
<td>Inferred from homology</td>
<td>70%</td>
</tr>
<tr>
<td>Predicted</td>
<td>5%</td>
</tr>
<tr>
<td>Uncertain (mainly TrEMBL)</td>
<td>-</td>
</tr>
</tbody>
</table>

!! Not sequence validation !!
Protein existence tag

!! Not sequence validation !!

<table>
<thead>
<tr>
<th>Protein existence level:</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evidence at protein level</td>
<td>59%</td>
</tr>
<tr>
<td>Evidence at transcript level</td>
<td>37.5%</td>
</tr>
<tr>
<td>Inferred from homology</td>
<td>1%</td>
</tr>
<tr>
<td>Predicted</td>
<td>0.5%</td>
</tr>
<tr>
<td>Uncertain (mainly TrEMBL)</td>
<td>2%</td>
</tr>
</tbody>
</table>
Annotation sources for UniProtKB

Data sources

- GO
  - Functional info

- PRIDE
  - Protein identification data

- InterPro
  - Protein families and domains

- IntAct
  - Molecular interactions

- IntEnz
  - Enzymes

- HAMAP
  - Microbial protein families

- RESID
  - Post-translational modifications

Some data sources for annotation

Protein classification

- InterPro classification
- Signal prediction
- Transmembrane prediction
- Other predictions

- Manual curation
- Literature-based annotation
- Sequence analysis

UniProtKB

* Automated annotation
Features of UniProtKB

**Sequence Features**
- Splice variants

**Annotations**
- Protein names
- Gene names

**Nomenclature**
- Name: Qk1
  - Synonyms: Qk, Qk1, Qka1

**Ontologies**
- Splice variants

**References**

**Sequence**

**Features of UniProtKB**
- Sequence
- Annotations
- Nomenclature
- References
A wealth of external links

**Organism-specific DBs**
- DictyBase
- EchoBASE
- EcoGene
- euHCvdb
- FlyBase
- GeneCards
- GeneFarm
- Gramene
- H-InvDB
- LegioList
- Leproma
- ListiList ZFIN
- MaizeGDB
- Orphanet
- PhotoList
- SagaList
- TuberculList
- WormPep
- GeneDB_Spombe
- ArachnoServer
- BuruList

**Enzyme & pathway DBs**
- BioCyc
- BREnda
- Reactome
- Pathway_Interaction_DB

**Proteomic DBs**
- PeptideAtlas
- PRIDE
- ProMEX

**Proteome annotation DBs**
- Ensembl
- KEGG
- GeneID
- NMPDR
- UCSC
- GenomeReviews

**Genome annotation DBs**
- Gene3D
- PIRSF
- HAMAP
- PRINTS
- InterPro
- ProDom
- PANTHER
- PROSITE
- Pfam
- TIGRFAMs
- SMART

**Family and domain DBs**
- HOGENOM
- OMA
- HOVERGEN
- PhylomeDB
- InParanoid
- OrthoDB

**Phylogenetic DBs**
- dbSNP

**Polymorphism DBs**
- BindingDB
- PMAP
- CutDB
- DrugBank
- NextBio

**Gene expression DBs**
- ArrayExpress
- Bgee
- GermOnline
- CleanEx
- Genevestigator

**3D structure DBs**
- DisProt
- HSSP
- PDB
- PDBsum
- SMR

**Gene family/group DBs**
- CAZy
- MEROPS
- PeroxiBase
- REBASE
- PptaseDB
- TCDB

**PTM DBs**
- GlycoSuiteDB
- PhosphoSite

**Sequence DBs**
- EMBL
- IPI
- PIR
- RefSeq
- UniGene

**Protein-protein interaction DBs**
- DIP
- IntAct
- STRING

**EMBL-EBI**

125 links!
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)

<table>
<thead>
<tr>
<th>Entry information</th>
<th>Hide</th>
<th>Top</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entry name</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accession</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Entry history</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Entry status</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Annotation project</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **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**

<table>
<thead>
<tr>
<th>Natural variant</th>
<th>35</th>
<th>1</th>
<th>L → F in a liver tumor.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural variant</td>
<td>43</td>
<td>1</td>
<td>L → S in a renal tumor.</td>
</tr>
<tr>
<td>Natural variant</td>
<td>47</td>
<td>1</td>
<td>P → S: dbSNP rs1800371.</td>
</tr>
<tr>
<td>Natural variant</td>
<td>53</td>
<td>1</td>
<td>W → C in a leukemia and a lymphoma.</td>
</tr>
<tr>
<td>Natural variant</td>
<td>60</td>
<td>1</td>
<td>P → S in a leukemia and a lymphoma.</td>
</tr>
<tr>
<td>Natural variant</td>
<td>72</td>
<td>1</td>
<td>P → R: dbSNP rs1042522.</td>
</tr>
</tbody>
</table>

....and of PTMs

<table>
<thead>
<tr>
<th>Modified residue</th>
<th>46</th>
<th>1</th>
<th>Phosphoserine; by HIPK2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modified residue</td>
<td>55</td>
<td>1</td>
<td>Phosphothreonine; by TAF1</td>
</tr>
<tr>
<td>Modified residue</td>
<td>99</td>
<td>1</td>
<td>Phosphoserine</td>
</tr>
<tr>
<td>Modified residue</td>
<td>305</td>
<td>1</td>
<td>N6-acetylysine</td>
</tr>
<tr>
<td>Modified residue</td>
<td>315</td>
<td>1</td>
<td>Phosphoserine; by CDC2</td>
</tr>
<tr>
<td>Modified residue</td>
<td>373</td>
<td>1</td>
<td>N6-acetylysine</td>
</tr>
</tbody>
</table>
Other examples of sequencing errors include: premature stop codons, read-throughs, erroneous initiator methionines
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
<table>
<thead>
<tr>
<th>Protein names</th>
<th>Recommended name: Glutamate carboxypeptidase 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EC=3.4.17.21</td>
</tr>
<tr>
<td>Alternative name(s):</td>
<td></td>
</tr>
<tr>
<td>Cell growth-inhibiting gene 27 protein</td>
<td></td>
</tr>
<tr>
<td>Folate hydrolase 1</td>
<td></td>
</tr>
<tr>
<td>Poly-L-poly-gamma-glutamate carboxypeptidase</td>
<td></td>
</tr>
<tr>
<td>Short name=FGCP</td>
<td></td>
</tr>
<tr>
<td>Glutamate carboxypeptidase II</td>
<td></td>
</tr>
<tr>
<td>Short name=GCPII</td>
<td></td>
</tr>
<tr>
<td>Membrane glutamate carboxypeptidase</td>
<td></td>
</tr>
<tr>
<td>Short name=mGCP</td>
<td></td>
</tr>
<tr>
<td>N-acetylated-alpha-linked acidic dipeptidase I</td>
<td></td>
</tr>
<tr>
<td>Short name=NAALADase I</td>
<td></td>
</tr>
<tr>
<td>Prostate-specific membrane antigen</td>
<td></td>
</tr>
<tr>
<td>Short name=PSM</td>
<td></td>
</tr>
<tr>
<td>Short name=PSMA</td>
<td></td>
</tr>
<tr>
<td>Pteroylpoly-gamma-glutamate carboxypeptidase</td>
<td></td>
</tr>
</tbody>
</table>

| Gene names                                        | Name: FOLH1                                    |
|                                                  | Synonyms: FOLH, NAALAD1, PSM, PSMA             |
|                                                  | ORF Names:GIG27                                |
Nomenclature

Provides synonyms and cleavage products of bifunctional proteins

<table>
<thead>
<tr>
<th>Gene names</th>
<th>Name: argJ Ordered Locus Names: BSU11200</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Molecule processing</strong></td>
<td></td>
</tr>
<tr>
<td>□ Transit peptide</td>
<td>1 – 8</td>
</tr>
<tr>
<td>□ Chain</td>
<td>9 – 214</td>
</tr>
<tr>
<td>□ Chain</td>
<td>215 – 441</td>
</tr>
<tr>
<td>□ Site</td>
<td>214 – 215</td>
</tr>
</tbody>
</table>
**Controlled vocabularies used whenever possible…**

### 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 acetylpornithine and glutamate.  
| 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.  
| 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.  
| 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:  
| | $K_M = 8.4 \text{ mM}$ for glutamate  
| | $K_M = 2.8 \text{ mM}$ for N-acetylpornithine  

>30 comment fields

**Annotation comments**
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 (e.g. domain) highlighted on sequence
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
## Annotation for human Rhodopsin:

<table>
<thead>
<tr>
<th>Gene Ontology (GO)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Biological process</strong></td>
<td><strong>phototransduction, visible light</strong></td>
</tr>
<tr>
<td></td>
<td>Traceable author statement. Source: ProtInc</td>
</tr>
<tr>
<td></td>
<td><strong>protein-chromophore linkage</strong></td>
</tr>
<tr>
<td></td>
<td>Inferred from electronic annotation. Source: UniProtKB-KW</td>
</tr>
<tr>
<td></td>
<td><strong>rhodopsin mediated signaling pathway</strong></td>
</tr>
<tr>
<td></td>
<td>Traceable author statement. Source: ProtInc</td>
</tr>
<tr>
<td><strong>Cellular component</strong></td>
<td><strong>integral to plasma membrane</strong></td>
</tr>
<tr>
<td></td>
<td>Traceable author statement. Source: ProtInc</td>
</tr>
<tr>
<td><strong>Molecular function</strong></td>
<td><strong>G-protein coupled receptor activity</strong></td>
</tr>
<tr>
<td></td>
<td>Traceable author statement. Source: ProtInc</td>
</tr>
<tr>
<td></td>
<td><strong>photoreceptor activity</strong></td>
</tr>
<tr>
<td></td>
<td>Inferred from electronic annotation. Source: UniProtKB-KW</td>
</tr>
<tr>
<td></td>
<td><strong>protein binding</strong></td>
</tr>
<tr>
<td></td>
<td>Inferred from physical interaction. Source: IntAct</td>
</tr>
</tbody>
</table>

*Complete GO annotation...*
Binary interactions are taken from the database Interactors of human p53

<table>
<thead>
<tr>
<th>With</th>
<th>Entry</th>
<th>#Exp.</th>
<th>IntAct</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>P03070</td>
<td>4</td>
<td>EBI-366083,EBI-617698</td>
<td>From a different organism.</td>
<td></td>
</tr>
<tr>
<td>ATR</td>
<td>Q13535</td>
<td>1</td>
<td>EBI-366083,EBI-968983</td>
<td></td>
</tr>
<tr>
<td>BARD1</td>
<td>Q99728</td>
<td>1</td>
<td>EBI-366083,EBI-473181</td>
<td></td>
</tr>
<tr>
<td>Bard1</td>
<td>O70445</td>
<td>1</td>
<td>EBI-366083,EBI-1790207</td>
<td>From a different organism.</td>
</tr>
<tr>
<td>BCL2L1</td>
<td>Q07817-1</td>
<td>2</td>
<td>EBI-366083,EBI-287195</td>
<td></td>
</tr>
<tr>
<td>BMX</td>
<td>P51813</td>
<td>1</td>
<td>EBI-366083,EBI-696657</td>
<td></td>
</tr>
<tr>
<td>BTBD2</td>
<td>Q9BX70</td>
<td>1</td>
<td>EBI-366083,EBI-710091</td>
<td></td>
</tr>
<tr>
<td>CABLES1</td>
<td>Q8TDN4</td>
<td>1</td>
<td>EBI-366083,EBI-604615</td>
<td></td>
</tr>
<tr>
<td>Cables1</td>
<td>Q9ESJ1</td>
<td>1</td>
<td>EBI-366083,EBI-604411</td>
<td>From a different organism.</td>
</tr>
<tr>
<td>CEBPB</td>
<td>P17676</td>
<td>3</td>
<td>EBI-366083,EBI-969696</td>
<td></td>
</tr>
<tr>
<td>CREBBP</td>
<td>Q92793</td>
<td>3</td>
<td>EBI-366083,EBI-81215</td>
<td></td>
</tr>
<tr>
<td>Crebbp</td>
<td>P45481</td>
<td>2</td>
<td>EBI-366083,EBI-296306</td>
<td>From a different organism.</td>
</tr>
<tr>
<td>CSE1L</td>
<td>P55060</td>
<td>4</td>
<td>EBI-366083,EBI-286709</td>
<td></td>
</tr>
<tr>
<td>CUL7</td>
<td>Q14999</td>
<td>1</td>
<td>EBI-366083,EBI-308606</td>
<td></td>
</tr>
</tbody>
</table>
**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 to LDL receptors and may function in the transport and/or as a regulator of fatty acid transport.  
  - Proven: Ref.19, Ref.21

- **Subcellular location**: Membrane; Multi-pass membrane protein.
  - Proven: Ref.8, Ref.19, Ref.13, Ref.14, Ref.15

- **Post-translational modification**: N-glycosylated and O-glycosylated with a ratio of 2:1.  
  - Proven: 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:606404]; 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.  
  - Proven: Ref.19, Ref.20

**Sequence annotation (Features)**

- **Amino acid modifications**
<table>
<thead>
<tr>
<th>Reference</th>
<th>Citation</th>
</tr>
</thead>
</table>

+ 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 is an excellent source of quality protein sequence and annotation data, but it is better for **non-patent data**.
Old entries accessible in both ENA and UniProt

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

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

Select and compare versions

View specific old entry

Tracks all changes to an entry

<table>
<thead>
<tr>
<th>Accession Number</th>
<th>Entry</th>
<th>Version</th>
<th>Sequence Version</th>
<th>Release</th>
<th>Issue Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>AF527840</td>
<td>2</td>
<td></td>
<td>AF527840.1</td>
<td>108</td>
<td>28-MAY-2011</td>
</tr>
<tr>
<td>AF527840</td>
<td>2</td>
<td></td>
<td>AF527840.1</td>
<td>105</td>
<td>27-AUG-2010</td>
</tr>
<tr>
<td>AF527840</td>
<td>2</td>
<td></td>
<td>AF527840.1</td>
<td>104</td>
<td>27-MAY-2010</td>
</tr>
<tr>
<td>AF527840</td>
<td>2</td>
<td></td>
<td>AF527840.1</td>
<td>103</td>
<td>26-FEB-2010</td>
</tr>
<tr>
<td>AF527840</td>
<td>2</td>
<td></td>
<td>AF527840.1</td>
<td>102</td>
<td>26-NOV-2009</td>
</tr>
<tr>
<td>AF527840</td>
<td>2</td>
<td></td>
<td>AF527840.1</td>
<td>101</td>
<td>28-AUG-2009</td>
</tr>
<tr>
<td>AF527840</td>
<td>2</td>
<td></td>
<td>AF527840.1</td>
<td>100</td>
<td>29-MAY-2009</td>
</tr>
<tr>
<td>AF527840</td>
<td>2</td>
<td></td>
<td>AF527840.1</td>
<td>99</td>
<td>26-FEB-2009</td>
</tr>
<tr>
<td>AF527840</td>
<td>2</td>
<td></td>
<td>AF527840.1</td>
<td>98</td>
<td>12-JAN-2009</td>
</tr>
<tr>
<td>AF527840</td>
<td>2</td>
<td></td>
<td>AF527840.1</td>
<td>97</td>
<td>28-NOV-2008</td>
</tr>
<tr>
<td>AF527840</td>
<td>2</td>
<td></td>
<td>AF527840.1</td>
<td>96</td>
<td>20-AUG-2008</td>
</tr>
<tr>
<td>AF527840</td>
<td>2</td>
<td></td>
<td>AF527840.1</td>
<td>95</td>
<td>29-MAY-2008</td>
</tr>
<tr>
<td>AF527840</td>
<td>2</td>
<td></td>
<td>AF527840.1</td>
<td>93</td>
<td>29-NOV-2007</td>
</tr>
<tr>
<td>AF527840</td>
<td>2</td>
<td></td>
<td>AF527840.1</td>
<td>90</td>
<td>08-MAR-2007</td>
</tr>
<tr>
<td>AF527840</td>
<td>2</td>
<td></td>
<td>AF527840.1</td>
<td>89</td>
<td>30-NOV-2006</td>
</tr>
<tr>
<td>AF527840</td>
<td>2</td>
<td></td>
<td>AF527840.1</td>
<td>89</td>
<td>14-NOV-2006</td>
</tr>
<tr>
<td>AF527840</td>
<td>1</td>
<td></td>
<td>AF527840.1</td>
<td>87</td>
<td>19-JUN-2006</td>
</tr>
<tr>
<td>AF527840</td>
<td>1</td>
<td></td>
<td>AF527840.1</td>
<td>86</td>
<td>21-MAR-2006</td>
</tr>
<tr>
<td>AF527840</td>
<td>1</td>
<td></td>
<td>AF527840.1</td>
<td>85</td>
<td>16-DEC-2005</td>
</tr>
<tr>
<td>AF527840</td>
<td>1</td>
<td></td>
<td>AF527840.1</td>
<td>84</td>
<td>12-SEP-2005</td>
</tr>
</tbody>
</table>
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].

**Accession Number or Sequence Version:** AX429748

**Snapshot at:** [day-month-year (e.g. 30-11-1998 or 30-NOV-1998)]

<table>
<thead>
<tr>
<th>ID</th>
<th>AX429748 standard: DNA; HUM; 100 BP.</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC</td>
<td>AX429748;</td>
</tr>
<tr>
<td>SV</td>
<td>AX429748.1</td>
</tr>
<tr>
<td>DT</td>
<td>21-JUN-2002 (Rel. 72, Created)</td>
</tr>
<tr>
<td>DT</td>
<td>21-JUN-2002 (Rel. 72, Last updated, Version 1)</td>
</tr>
<tr>
<td>KW</td>
<td></td>
</tr>
<tr>
<td>OS</td>
<td>Homo sapiens (human)</td>
</tr>
<tr>
<td>UC</td>
<td>Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;</td>
</tr>
<tr>
<td>OC</td>
<td>Eutheria; Primates; Catarrhini; Hominidae; Homo.</td>
</tr>
<tr>
<td>RN</td>
<td>[1]</td>
</tr>
<tr>
<td>RA</td>
<td>Smith J.C.;</td>
</tr>
<tr>
<td>RT</td>
<td>&quot;Polymorphisms in the human kdr gene&quot;;</td>
</tr>
<tr>
<td>RL</td>
<td>AstraZeneca AB (SE).</td>
</tr>
</tbody>
</table>

| XX  |                                      |
| FH  | Key Location/Qualifiers               |
| FT  | source 1..100                         |
| FT  | /db xref="taxon:9606"                 |
| FT  | /organism="Homo sapiens"              |

| SQ  | Sequence 100 BP; 14 A; 29 C; 36 G; 20 T; 1 other; |
|     | gcgccccgcc tcgacacttg gacggcctgg gtaagggagcg stgtcttccc gcgcttttt ggtgatgcc cccaaatttt cggyggacccg // |

[Current version]
Compare different versions

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

Accession Number or Sequence Version: AX429748

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

Differences for AX429748 30-MAY-2002 / 28-NOV-2008

<table>
<thead>
<tr>
<th>ID</th>
<th>AX429748; SV 1; linear; unassigned DNA; PAT; HUM; 100 BP.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>AX429748 standard; DNA; HUM; 100 BP.</td>
</tr>
<tr>
<td>XX</td>
<td>AX429748;</td>
</tr>
<tr>
<td>XX</td>
<td>AX429748.1</td>
</tr>
<tr>
<td>DX</td>
<td>Sequence 13 from Patent EP1203827.</td>
</tr>
<tr>
<td>FW</td>
<td>.</td>
</tr>
<tr>
<td>OS</td>
<td>Homo sapiens (human)</td>
</tr>
<tr>
<td>OC</td>
<td>Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; H Chain;</td>
</tr>
<tr>
<td>OC</td>
<td>Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominoidea;</td>
</tr>
<tr>
<td>OC</td>
<td>Homo.</td>
</tr>
<tr>
<td>OC</td>
<td>Eutheria; Primates; Catarrhini; Hominoidea; Homo.</td>
</tr>
<tr>
<td>RN</td>
<td>[1]</td>
</tr>
<tr>
<td>RA</td>
<td>Smith J.C.:</td>
</tr>
<tr>
<td>RT</td>
<td>&quot;Polymorphisms in the human kdr gene&quot;;</td>
</tr>
<tr>
<td>RL</td>
<td>AstraZeneca AB (SE).</td>
</tr>
<tr>
<td>FH</td>
<td>Key Location/Qualifiers</td>
</tr>
<tr>
<td>FT</td>
<td>source 1..100</td>
</tr>
<tr>
<td>FT</td>
<td>/organism=&quot;Homo sapiens&quot;</td>
</tr>
<tr>
<td>FT</td>
<td>/mol_type=&quot;unassigned DNA&quot;</td>
</tr>
<tr>
<td>FT</td>
<td>/db_xref=&quot;taxon:9606&quot;</td>
</tr>
<tr>
<td>FT</td>
<td>/organism=&quot;Homo sapiens&quot;</td>
</tr>
<tr>
<td>SQ</td>
<td>Sequence 100 BP: 14 A; 29 C; 36 G; 20 T; 1 other:</td>
</tr>
<tr>
<td></td>
<td>ggcgcagtgcc gacagaggtgc ctgaaccttg aggcgcttg tggggacgc etgcttcttg 60</td>
</tr>
<tr>
<td></td>
<td>cagcgcctct ggtgatgcgc cccaaatttc ggggagcgcg 100</td>
</tr>
</tbody>
</table>
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

GenBank → INSDC → ENA → DDBJ
Distributing patent sequences

Redundancy: a consequence of the international cooperation

- USPTO
- GenBank
- DDBJ
- JPO
- KIPO
- other patent offices
- ENA
- EPO
- INSDC

As other National Offices participate in data exchange → redundancy will increase
Distributing patent sequences

NR patent databases remove redundancy

NR patent sequence databases

USPTO ➔ GenBank ➔ DDBJ ➔ KIPO ➔ JPO

other patent offices

ENA

INSDC

EPO
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

Sequence Searching Tools

- Source 1
- Source 2
- Source 3
- Source N

Standardization of data formats and sequences verification (JRA 15)

- Redundant Patent Sequence Database(s) (Public sequence repositories)
- Removal of Redundancy-L1 (100% ID + same length)
- Removal of Redundancy-L2 (Apply equivalent rules)

Unique Publication Numbers
Equivalents (Simple Family)
Annotation fields: Apply priority rules
Non-Redundant Patent Sequence Database(s)

Enrichment of annotations
- Patent Text Collection
- Information Extraction

Annotations

D1
SEQUENCE CAPTURE

D2
NR-DB
L1 & L2
Creating a non-redundant database

D1
SEQUENCE CAPTURE

D2
NR-DB
L1 & L2

D3
Enrichment of annotations

Enrichment of annotations

ANNOTATIONS

Source 1 Source 2 Source 3 Source N

Standardization of data formats and sequences verification (JRA 15)

Redundant Patent Sequence Database(s)
(Public sequence repositories)

Removal of Redundancy-L1 (100% ID + same length)

Removal of Redundancy-L2 (Apply equivalent rules)

Equivalents (Simple Family)

Unique Publication Numbers

Annotation fields: Apply priority rules

Non-Redundant Patent Sequence Database(s)
Creating a non-redundant database

D1
SEQUENCE CAPTURE

D2
NR-DB
L1 & L2

D3
ANNOTATIONS

Source 1, Source 2, Source 3, Source N

Standardization of data formats and sequences verification (JRA15)

Redundant Patent Sequence Database(s)
(Public sequence repositories)

Removal of Redundancy-L1
(100% ID + same length)

Removal of Redundancy-L2
(Apply equivalent rules)

Equivalents
(Simple Family)

Unique Publication Numbers

Annotation fields:
Apply priority rules

Non-Redundant Patent Sequence Database(s)

Enrichment of annotations

Patent Text Collection
Information Extraction

Enrichment of annotations
Patent data correction

A) 22.54% 77.46%

- Total Wrong
- Correct

B) 22% 2% 11% 13% 33%

- KC only
- KC completeness only
- KC + PN
- KC completeness + PN
- PN only
- Publication level only
- Correct
- pending

Correction of Publication Numbers and kind Codes

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:AA099687;
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:DT578933;
DE  Phytase Variants.
CC  Patent Number could not be successfully verified
## Patent resources at EBI

| Resources          | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | Access    |
|--------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Patent Chemical    | 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 |
| Compounds          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |            |
| Patent Sequences   | Multiple sets of patent sequences are available at EBI.  
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  | 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        |
| Data               |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |            |

[http://www.ebi.ac.uk/patentdata/](http://www.ebi.ac.uk/patentdata/)
More information provided

**Patent proteins:**
- EPO
- USPTO
- JPO
- KIPO

**Patent nucleotides:**
- ENA (EPO, USPTO, JPO, KIPO)

<table>
<thead>
<tr>
<th>Resources</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Patent Sequences</strong></td>
<td>Multiple sets of patent sequences are available at EBI.</td>
</tr>
<tr>
<td>2. Patent nucleotides</td>
<td>Contain the patent class data in the EMBL-Bank.</td>
</tr>
<tr>
<td>3. Non-redundant patent sequences</td>
<td>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).</td>
</tr>
</tbody>
</table>

- **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

Example:
Search patent protein sequence

http://www.ebi.ac.uk/Tools/sss/
Results from redundant databases

>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

**Example:**
Search patent protein sequence

http://www.ebi.ac.uk/Tools/sss/
Results from level-1 database

<table>
<thead>
<tr>
<th>Align.</th>
<th>DB:ID</th>
<th>Source</th>
<th>Length</th>
<th>Score</th>
<th>Identities</th>
<th>Positives</th>
<th>E()</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NRPL1:NRP_AX092329</td>
<td>PN:WO0073454 A1</td>
<td>282</td>
<td>1808</td>
<td>100.0</td>
<td>100.0</td>
<td>4.1E-162</td>
</tr>
<tr>
<td>2</td>
<td>NRPL1:NRP_HI379470</td>
<td>PN:EP2104513 A2</td>
<td>290</td>
<td>1808</td>
<td>100.0</td>
<td>100.0</td>
<td>2.2E-161</td>
</tr>
<tr>
<td>3</td>
<td>NRPL1:NRP_AX156352</td>
<td>PN:WO00140269 A2</td>
<td>309</td>
<td>1808</td>
<td>100.0</td>
<td>100.0</td>
<td>2.2E-161</td>
</tr>
<tr>
<td>4</td>
<td>NRPL1:NRP_JA171544</td>
<td>PN:US2003124128 A1</td>
<td>282</td>
<td>1801</td>
<td>99.6</td>
<td>99.6</td>
<td>1.9E-161</td>
</tr>
<tr>
<td>5</td>
<td>NRPL1:NRP_ADA40459</td>
<td>PN:JP2007504280T T</td>
<td>306</td>
<td>1801</td>
<td>99.6</td>
<td>99.6</td>
<td>2.2E-161</td>
</tr>
<tr>
<td>6</td>
<td>NRPL1:NRP_ADA40466</td>
<td>PN:JP2007504280T T</td>
<td>309</td>
<td>1801</td>
<td>99.6</td>
<td>99.6</td>
<td>2.2E-161</td>
</tr>
<tr>
<td>7</td>
<td>NRPL1:NRP_HI962287</td>
<td>PN:US2005238649 A1</td>
<td>282</td>
<td>1798</td>
<td>99.6</td>
<td>99.6</td>
<td>3.6E-161</td>
</tr>
<tr>
<td>8</td>
<td>NRPL1:NRP_JA171549</td>
<td>PN:US2003124140 A1</td>
<td>282</td>
<td>1789</td>
<td>98.6</td>
<td>100.0</td>
<td>2.6E-160</td>
</tr>
<tr>
<td>9</td>
<td>NRPL1:NRP_JA171551</td>
<td>PN:WO2011026122 A2</td>
<td>282</td>
<td>1780</td>
<td>98.2</td>
<td>99.6</td>
<td>1.9E-159</td>
</tr>
<tr>
<td>10</td>
<td>NRPL1:NRP_JA171547</td>
<td>PN:WO2011026122 A2</td>
<td>278</td>
<td>1727</td>
<td>99.3</td>
<td>100.0</td>
<td>1.9E-154</td>
</tr>
<tr>
<td>11</td>
<td>NRPL1:NRP_ADA403050</td>
<td>PN:WO0020624 A2</td>
<td>258</td>
<td>1665</td>
<td>100.0</td>
<td>100.0</td>
<td>1.3E-148</td>
</tr>
<tr>
<td>12</td>
<td>NRPL1:NRP_JA171542</td>
<td>PN:WO2011026122 A2</td>
<td>274</td>
<td>1656</td>
<td>94.5</td>
<td>97.8</td>
<td>1.0E-147</td>
</tr>
<tr>
<td>13</td>
<td>NRPL1:NRP_JA171586</td>
<td>PN:WO2011026122 A2</td>
<td>259</td>
<td>1653</td>
<td>100.0</td>
<td>100.0</td>
<td>1.8E-147</td>
</tr>
<tr>
<td>14</td>
<td>NRPL1:NRP_JA171543</td>
<td>PN:WO2011026122 A2</td>
<td>255</td>
<td>1650</td>
<td>100.0</td>
<td>100.0</td>
<td>3.4E-147</td>
</tr>
</tbody>
</table>

Each hit unique
Results from level-1 database

List of all patents containing the sequence

Earliest publication date

Link to sequence entry

Link to patent documentation
Patent sequence record in NRNL1

Patents containing 100% identical sequence

Sequence:
```
gcg cgatg gct gagggctcct gcaact tggacg cggagccttg gtggaggg gagg cggagctgc tctctg ccctt ggtttccggg gaccggc
```
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.
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)*
LEVEL2 NR patent sequence database
groups identical sequences by *patent family*

→ provides *earliest priority date* for family
Searching patent sequences

Example:
Search patent protein sequence

http://www.ebi.ac.uk/Tools/sss/
## Results from level-2 database

### FASTA Results

<table>
<thead>
<tr>
<th>Align.</th>
<th>DB:ID</th>
<th>Source</th>
<th>Length</th>
<th>Score</th>
<th>Identities</th>
<th>Positives</th>
<th>E()</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NRPL2:NRP00080F2A</td>
<td>PN:JP2005304500 A</td>
<td>282</td>
<td>1808</td>
<td>100.0</td>
<td>100.0</td>
<td>4.7E-157</td>
</tr>
<tr>
<td>2</td>
<td>NRPL2:NRP00080F2B</td>
<td>PN:US7056736 B2</td>
<td>282</td>
<td>1808</td>
<td>100.0</td>
<td>100.0</td>
<td>4.7E-157</td>
</tr>
<tr>
<td>3</td>
<td>NRPL2:NRP00080F88</td>
<td>PN:US7339034 B2</td>
<td>282</td>
<td>1808</td>
<td>100.0</td>
<td>100.0</td>
<td>4.7E-157</td>
</tr>
<tr>
<td>4</td>
<td>NRPL2:NRP00080F89</td>
<td>PN:US7345145 B2</td>
<td>282</td>
<td>1808</td>
<td>100.0</td>
<td>100.0</td>
<td>4.7E-157</td>
</tr>
<tr>
<td>5</td>
<td>NRPL2:NRP00080F2C</td>
<td>PN:US7060812 B2</td>
<td>282</td>
<td>1808</td>
<td>100.0</td>
<td>100.0</td>
<td>4.7E-157</td>
</tr>
<tr>
<td>6</td>
<td>NRPL2:NRP00080F2D</td>
<td>PN:US2003181679 A1</td>
<td>282</td>
<td>1808</td>
<td>100.0</td>
<td>100.0</td>
<td>4.7E-157</td>
</tr>
<tr>
<td>7</td>
<td>NRPL2:NRP00080F2E</td>
<td>PN:US7083978 B2</td>
<td>282</td>
<td>1808</td>
<td>100.0</td>
<td>100.0</td>
<td>4.7E-157</td>
</tr>
<tr>
<td>8</td>
<td>NRPL2:NRP00080F2F</td>
<td>PN:US7084255 B2</td>
<td>282</td>
<td>1808</td>
<td>100.0</td>
<td>100.0</td>
<td>4.7E-157</td>
</tr>
<tr>
<td>9</td>
<td>NRPL2:NRP00080F30</td>
<td>PN:US7074897 B2</td>
<td>282</td>
<td>1808</td>
<td>100.0</td>
<td>100.0</td>
<td>4.7E-157</td>
</tr>
<tr>
<td>10</td>
<td>NRPL2:NRP00080F8E</td>
<td>PN:US7356340 B2</td>
<td>282</td>
<td>1808</td>
<td>100.0</td>
<td>100.0</td>
<td>4.7E-157</td>
</tr>
<tr>
<td>11</td>
<td>NRPL2:NRP00080F8F</td>
<td>PN:US7365156 B2</td>
<td>282</td>
<td>1808</td>
<td>100.0</td>
<td>100.0</td>
<td>4.7E-157</td>
</tr>
<tr>
<td>12</td>
<td>NRPL2:NRP00080F31</td>
<td>PN:WO2006098887 A2</td>
<td>282</td>
<td>1808</td>
<td>100.0</td>
<td>100.0</td>
<td>4.7E-157</td>
</tr>
<tr>
<td>13</td>
<td>NRPL2:NRP00080F32</td>
<td>PN:US7101687 B2</td>
<td>282</td>
<td>1808</td>
<td>100.0</td>
<td>100.0</td>
<td>4.7E-157</td>
</tr>
<tr>
<td>14</td>
<td>NRPL2:NRP00080F92</td>
<td>PN:US7378486 B2</td>
<td>282</td>
<td>1808</td>
<td>100.0</td>
<td>100.0</td>
<td>4.7E-157</td>
</tr>
</tbody>
</table>

Each hit = one family
Results from level-2 database

- Patents in same family
- Earliest publication (priority) date in family
- Link to sequence entry
- Link to patent documentation
Patent sequence record in NRNL2

- Patent equivalents
- Sequence record in ENA
- Sequence
- Priority number and date
- Patent literature
- Translation
Non-redundant patent databases

- **Patent nucleotides**
  - **NRNL1** (Non-redundant nucleotide level-1)
  - **NRNL2** (Non-redundant nucleotide level-2)

- **Patent proteins**
  - **NRPL1** (Non-redundant protein level-1)
  - **NRPL2** (Non-redundant protein level-2)

- Groups together 100% identical patent sequences
- 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 that 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
Sequence Searching Tools

EBI Search

EBI-Search accessible from any EBI page

http://www.ebi.ac.uk/
EBI-Search by **gene name**

**Search for src gene**

[Image of EBI search interface]

http://www.ebi.ac.uk/
EBI search by gene name

Lists species with information

Lists relevant entries in all EBI resources

Gene & Protein (includes expression, structures, literature...)
grouped by ontology (with synonyms)

src sarcoma (Schmidt-Ruppin A.2) viral oncogene homolog (avian)

- SRC (ASV c-src, SRC1, p60-src, c-SRC, ENSG00000161122)
- Human (Homo Sapiens)

- Rous sarcoma oncogene
  - Src (AVG 59666, pp60 c-src, ENSMUSG00000027644)
  - Horse (Equus caballus)

- Src-1 encodes a tyrosine kinase. During embryonic development, src-1 functions together with a
  Wnt signaling pathway to specify...
  - src-1 (YR120124-1)
  - Astyanax (Astyanax bicolor)

Genomes / Ensembl Gene

- ENSDARP00000000107
  - src sarcoma (Schmidt-Ruppin A.2) viral oncogene homolog (avian) [Source:ZFIN,Acc:ZDB-GENE-000131-3809]
  - Species: Danio rerio
  - References: Taxonomy, UniProt, GO, EMBL Bank

- View all 677 results...
EBI search by gene name

Tabs organise data by:
- gene
- expression
- protein
- structure
- literature

Easy to change between species
EBI search by gene name

Information from *Ensembl*:
- Gene sequence
- Location
- Sequence variations
- Orthologues...

Gene structure (forward and reverse strand)
Expression studies from *Gene Expression Atlas*, view by:

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

View additional structures
EBI search by gene name

Can print full summary of any page

Reviews

Keyword in title

Free full text

Patent

Curator-selected
EBI-Search for **patent** information

**EBI search**

[Image of EBI search interface]

Search for patent **WO0146262**

[Evaluation of protein expression](http://www.ebi.ac.uk/)
EBI Search

Search for patent WO0146262

Sequence data for WO0146262

Includes list of additional annotation

Literature for WO0146262

Includes link to full paper

Literature

OLFACTORY RECEPTOR GENES AND PSEUDOGENES IN PRIMATES AND MOUSE

The present invention refers to olfactory receptor genes and pseudogenes of 16 species of mouse. The invention also concerns olfactory receptors encoded by these genes and antibodies specific for the same.

Sequence data for WO0146262

Includes list of additional annotation

Literature

Literature / Patents

Search for WO0146262 in All results

WO0146262

AX116162

Sequence 3' flanked from Patent WO0146262.

View in EBI, in EMBL format, in SRS, in EMBL-SVA, Launch NCEBLAST, Launch FASTA

References: Taxonomy, EMBL-Bank (Coding Sequence)
**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**

**Search >100 databases**

**Available Databases**

<table>
<thead>
<tr>
<th>Literature, Bibliography and Reference Databases</th>
</tr>
</thead>
<tbody>
<tr>
<td>- MEDLINE</td>
</tr>
<tr>
<td>- Taxonomy</td>
</tr>
<tr>
<td>- OMIM</td>
</tr>
<tr>
<td>- OMIM Morbid Map</td>
</tr>
<tr>
<td>- Patent Abstracts</td>
</tr>
<tr>
<td>- Karyn's Genomes</td>
</tr>
<tr>
<td>- Patent Equivalents</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Literature, Bibliography and Reference Databases - subsections</th>
</tr>
</thead>
<tbody>
<tr>
<td>- MEDLINE (Updates)</td>
</tr>
<tr>
<td>- MEDLINE (Main Release 2010)</td>
</tr>
<tr>
<td>- MED2PUB</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gene Dictionaries and Ontologies</th>
</tr>
</thead>
<tbody>
<tr>
<td>- EMBL</td>
</tr>
<tr>
<td>- EMBL (Contigs expanded)</td>
</tr>
<tr>
<td>- EMBL (Coding Sequences)</td>
</tr>
<tr>
<td>- EMBL (Contig)</td>
</tr>
<tr>
<td>- EMBL ID/accession Mapping</td>
</tr>
<tr>
<td>- IMGT/LIGM-DB</td>
</tr>
<tr>
<td>- Genome Reviews</td>
</tr>
<tr>
<td>- IMGT/HLA</td>
</tr>
<tr>
<td>- GR Genes</td>
</tr>
<tr>
<td>- LiveLists</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nucleotide sequence databases</th>
</tr>
</thead>
<tbody>
<tr>
<td>- IPD-KB</td>
</tr>
<tr>
<td>- GR Transcripts</td>
</tr>
<tr>
<td>- RefSeq Genome</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Patent DNA (NRNL1 &amp; NRNL2)</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Patent DNA NR1L2</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>NR patent DNA (NRNL1 &amp; NRNL2)</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>NR patent proteins (NRPL1 &amp; NRPL2)</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Patent Protein NPL1</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Patent Protein NPL2</th>
</tr>
</thead>
</table>

**Select library tab**

**NR patent DNA (NRNL1 & NRNL2)**

**NR patent proteins (NRPL1 & NRPL2)**

**EMBL-EBI**
**SRS: advanced text search**

Select **library** tab

Example:
Selected to search NR level-1 patent DNA database

Search >100 databases

Available Databases

- **Literature, Bibliography and Reference Databases**
  - MEDLINE
  - Taxonomy
  - OMIM
  - OMIM Morbid Map
  - Patent Abstracts
  - Karyn’s Genomes
  - Patent Equivalents
  - MEDLINE (Updates)
  - MEDLINE (Main Release 2010)
  - MED2PUB

- **Gene Dictionaries and Ontologies**
  - Patent DNA
  - EMBL DNA
  - EMBL (Coding Sequences)
  - EMBL (Contig)
  - EMBL ID/Accession Mapping
  - IMGT/HLA
  - Genome Reviews
  - GenBank

- **Nucleotide sequence databases**
  - Patent DNA NRL2
  - Patent DNA NRL1

- **Nucleotide related databases**
  - UniProt Universal Protein Resource

- **Other protein sequence databases**
  - Active protein sequence databases
  - Patent Proteins
  - PIR Proteins
  - IPI

- **Protein function, structure and interaction databases**
  - Swall (SPTR)
  - PIR
  - RemTriEMBL

- **Enzymes, reactions and metabolic interaction databases**
  - SubAll
SRS: advanced text search

Select **library** tab

Select resources to search
SRS: advanced text search

Select library tab

Select resources to search

1) Select field

2) Type in text
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

Lists non-redundant nucleotide sequences from WO0146262
SRS: advanced text search

Select library tab
Select resources to search
Create query

WO0146262 sequences
Select **library** tab

**WO0146262**

nucleotide sequence record in NRNL1

Select resources to search

Create query

**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

Find all sequences associated with a patent

Find all sequences associated with a patent + identify all patents associated with each sequence

Find all sequences associated with a patent + identify all patents in the same family associated with each sequence
What’s available at EBI

Tools are accessible from the EBI homepage

Under tools, select **Tools Index**

http://www.ebi.ac.uk/
What’s available at EBI

**Sequence Searching Tools**

What’s available at EBI

Most popular tools are listed

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

Link to list of all tools
What’s available at EBI

http://www.ebi.ac.uk/Tool/ssss
STEP 1:

Choose a search algorithm
### Choosing the right search engine

<table>
<thead>
<tr>
<th>Tool</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLAST</td>
<td>Fast search (better for proteins than nucleotides)</td>
</tr>
<tr>
<td>FASTA</td>
<td>Fast search</td>
</tr>
<tr>
<td>PSI-SEARCH</td>
<td>Finding remote homologues</td>
</tr>
<tr>
<td>SSEARCH</td>
<td>Sensitive but slow; good for short sequences</td>
</tr>
<tr>
<td>GGSEARCH</td>
<td>Force full-length matches</td>
</tr>
<tr>
<td>GLSEARCH</td>
<td>Match domains/patterns to protein; oligo-to-gene</td>
</tr>
<tr>
<td>FASTM</td>
<td>Multi-peptide search</td>
</tr>
</tbody>
</table>

*Example usage:*

**Query:**

- AVTEGP
- EVLNF
- FVNGFAD
- AKFQPGE

**Subject:**

- Q
- S
Comparing search engines using a short peptide query sequence

query sequence: RPPSWIPK
Comparing search engines

query: RPPSWIPK

<table>
<thead>
<tr>
<th>hit</th>
<th>length</th>
<th>e()</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: TY01_PHYAZ</td>
<td>61</td>
<td>5.4</td>
</tr>
</tbody>
</table>

Look at the difference in e-values

<table>
<thead>
<tr>
<th>hit</th>
<th>length</th>
<th>e()</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: TY01_PHYAZ</td>
<td>61</td>
<td>0.42</td>
</tr>
<tr>
<td>2: BRK5_PHYNO</td>
<td>8</td>
<td>4.8</td>
</tr>
<tr>
<td>3: BRK_AMICA</td>
<td>9</td>
<td>9.2</td>
</tr>
<tr>
<td>5: BRK_LEPOS</td>
<td>9</td>
<td>9.2</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>hit</th>
<th>length</th>
<th>e()</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: TY01_PHYAZ</td>
<td>61</td>
<td>0.42</td>
</tr>
<tr>
<td>2: BRK5_PHYNO</td>
<td>8</td>
<td>4.8</td>
</tr>
<tr>
<td>3: BRK_AMICA</td>
<td>9</td>
<td>9.2</td>
</tr>
<tr>
<td>5: BRK_LEPOS</td>
<td>9</td>
<td>9.2</td>
</tr>
</tbody>
</table>

### GLEARCH

<table>
<thead>
<tr>
<th>hit</th>
<th>length</th>
<th>e()</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: TY01_PHYAZ</td>
<td>61</td>
<td>e-16</td>
</tr>
<tr>
<td>2: BRK5_PHYNO</td>
<td>8</td>
<td>1.9e-11</td>
</tr>
<tr>
<td>3: TY01_PHYBU</td>
<td>9</td>
<td>e-8</td>
</tr>
<tr>
<td>4: BRK_ONCMY</td>
<td>10</td>
<td>5.6e-8</td>
</tr>
<tr>
<td>5: BRK_LEPOS</td>
<td>9</td>
<td>1e-7</td>
</tr>
<tr>
<td>8: B4GT2_HUMAN</td>
<td>372</td>
<td>5.8e-5</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>39: DNAA_PROM</td>
<td>...</td>
<td>0.024</td>
</tr>
<tr>
<td>40: DNAA_PROM</td>
<td>199</td>
<td></td>
</tr>
</tbody>
</table>

### GGEARCH

<table>
<thead>
<tr>
<th>hit</th>
<th>length</th>
<th>e()</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: BRK5_PHYNO</td>
<td>8</td>
<td>e-7</td>
</tr>
<tr>
<td>2: TY01_PHYBU</td>
<td>8</td>
<td>2e-5</td>
</tr>
<tr>
<td>3: BRK_LEPOS</td>
<td>9</td>
<td>5.4e-4</td>
</tr>
<tr>
<td>4: BRK_AMICA</td>
<td>9</td>
<td>5.4e-4</td>
</tr>
<tr>
<td>5: BRK4_PHAJA</td>
<td>8</td>
<td>0.0087</td>
</tr>
<tr>
<td>6: BRK_PHYHY</td>
<td>8</td>
<td>0.0087</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>34: TY51_LITRU</td>
<td>7</td>
<td>8.3</td>
</tr>
</tbody>
</table>

GLSEARCH has a preference for long hits

GGSEARCH limited to similar sized hits

---

**GLSEARCH**: limited to similar sized hits

**GGSEARCH**: limited to similar sized hits
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**

**STEP 3 - Set your parameters**

<table>
<thead>
<tr>
<th>PROGRAM</th>
<th><strong>SSEARCH</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MATRIX</strong></td>
<td><strong>BLOSUM50</strong></td>
</tr>
<tr>
<td></td>
<td>-10</td>
</tr>
<tr>
<td><strong>DNA STRAND</strong></td>
<td><strong>HISTOGRAM</strong></td>
</tr>
<tr>
<td></td>
<td>no</td>
</tr>
<tr>
<td><strong>SEQUENCE RANGE</strong></td>
<td><strong>DATABASE RANGE</strong></td>
</tr>
<tr>
<td></td>
<td>6-10</td>
</tr>
<tr>
<td><strong>SCORING</strong></td>
<td><strong>50</strong></td>
</tr>
<tr>
<td></td>
<td>50</td>
</tr>
</tbody>
</table>
### Restricting length of matches

**query**: RPPSWIPK

<table>
<thead>
<tr>
<th>hit</th>
<th>length</th>
<th>e()</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: BRK5_PHYNO</td>
<td>8</td>
<td>0.34</td>
</tr>
<tr>
<td>2: BRK_LEPOS</td>
<td>9</td>
<td>0.57</td>
</tr>
<tr>
<td>3: BRK_AMICA</td>
<td>9</td>
<td>0.57</td>
</tr>
<tr>
<td>4: TY01 PHYBU</td>
<td>8</td>
<td>0.61</td>
</tr>
<tr>
<td>5: BRK_ONCMY</td>
<td>10</td>
<td>0.63</td>
</tr>
<tr>
<td>6: BRK4 PHAJA</td>
<td>8</td>
<td>5.0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>18: BRK3 PELRI</td>
<td>9</td>
<td>9.8</td>
</tr>
</tbody>
</table>

### Database range 6-10

- **SSEARCH**
  - UniProtKB/SwissProt
  - Database range 6-10

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

<table>
<thead>
<tr>
<th>hit</th>
<th>length</th>
<th>e()</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: BRK5_PHYNO</td>
<td>8</td>
<td>8e-7</td>
</tr>
<tr>
<td>2: TY01 PHYBU</td>
<td>8</td>
<td>2e-5</td>
</tr>
<tr>
<td>3: BRK_LEPOS</td>
<td>9</td>
<td>5.4e-4</td>
</tr>
<tr>
<td>4: BRK_AMICA</td>
<td>9</td>
<td>5.4e-4</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>34: TY01 LITRU</td>
<td>7</td>
<td>8.3</td>
</tr>
</tbody>
</table>

### SSEARCH

- **UniProtKB/SwissProt**
STEP 2:

Choose a database to search
Several databases available

Note: these databases cover the same sequences
- NR level-1
- NR level-2
- EPO + JPO + KIPO + USPTO
Several databases available

Nucleotide

Note: these databases cover the same sequences
- NR level-1
- NR level-2
- EMBL Patents

Patent data
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

<table>
<thead>
<tr>
<th>hit</th>
<th>e()</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: TY01_PHYAZ</td>
<td>3.6</td>
</tr>
</tbody>
</table>

Look at the difference in e-values

<table>
<thead>
<tr>
<th>hit</th>
<th>e()</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: TY01_PHYAZ</td>
<td>0.068</td>
</tr>
<tr>
<td>2: BRK5_PHYNO</td>
<td>1.9</td>
</tr>
<tr>
<td>3: TY01_PHYBU</td>
<td>7.7</td>
</tr>
<tr>
<td>4: BRK_AMICA</td>
<td>8.7</td>
</tr>
<tr>
<td>5: BRK_LEPOS</td>
<td>8.7</td>
</tr>
<tr>
<td>6: BRK_ONCMY</td>
<td>9.7</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>Amino acids</th>
<th>Nucleotides</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ser</td>
<td>UCU</td>
</tr>
<tr>
<td>Ser</td>
<td>AGC</td>
</tr>
</tbody>
</table>

**UCU** match

**AGC** mismatch
1) Codon degeneracy

**Proteins**

Human CKS1B kinase

Zebra finch CDC28 kinase 1B

**Nucleotides**

Human CKS1B kinase

Zebra finch CDC28 kinase 1B

<table>
<thead>
<tr>
<th>SeqA Name</th>
<th>Len(aa)</th>
<th>SeqB Name</th>
<th>Len(aa)</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>sp</td>
<td>P61024</td>
<td>CKS1_HUMAN</td>
<td>79</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SeqA Name</th>
<th>Len(nt)</th>
<th>SeqB Name</th>
<th>Len(nt)</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENA</td>
<td>CAA38702</td>
<td>CAA38702.1</td>
<td>240</td>
<td>2</td>
</tr>
</tbody>
</table>
1) Codon degeneracy

Sequence conservations is more stringent at the protein level, than at the level of the nucleotide coding sequence
Protein sequence searches can distinguish between **exact**, **similar** and **dissimilar** matches
2) Amino acid similarity

Amino acids grouped by physical & chemical properties

- Amino acids grouped by physical & chemical properties
  - aliphatic
  - hydrophobic
  - aromatic
  - polar
  - charged
  - tiny
  - small
  - positive
2) Amino acid similarity

Protein searches take account of amino acid similarities

<table>
<thead>
<tr>
<th>Amino acids</th>
<th>Nucleotides</th>
<th>identical</th>
<th>similar</th>
<th>mismatch</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ser</td>
<td>UCU</td>
<td>Ser</td>
<td>Ser</td>
<td>UCU</td>
</tr>
<tr>
<td>Ser</td>
<td>AGC</td>
<td>Ser (Asn)</td>
<td>Asn</td>
<td>AAC</td>
</tr>
<tr>
<td>Ser</td>
<td></td>
<td>Ser (Leu)</td>
<td>Leu</td>
<td>CUC</td>
</tr>
</tbody>
</table>

- **highly conserved**: Ser - Ser
- **weakly conserved**: Ser (Asn) - Asn
- **not conserved**: Ser (Leu) - Leu

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:

Proteins

DNA

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

today

Cambrian explosion

extinction of dinosaurs

multicellular life

self-replicating cells

chemical evolution

formation of Earth

4

1

s. ago

Billions of years ago

formation of Earth

extinction of dinosaurs

Cambrian explosion

multicellular life

self-replicating cells

chemical evolution

formation of Earth

Proteins

DNA
### Protein v nucleotide: example

<table>
<thead>
<tr>
<th>Protein</th>
<th>DNA e-value</th>
<th>Protein e-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>4.3</td>
<td>4.4e-17</td>
</tr>
<tr>
<td>Bovine</td>
<td>1.3e-07</td>
<td>1.3e-20</td>
</tr>
<tr>
<td>Mouse</td>
<td>4.4e-133</td>
<td>4.7e-79</td>
</tr>
<tr>
<td>Toad</td>
<td>8.9e-33</td>
<td>2.6e-45</td>
</tr>
<tr>
<td>Frog</td>
<td>8.5e-10</td>
<td>4.2e-32</td>
</tr>
<tr>
<td>Nematode</td>
<td>-</td>
<td>3.2e-12</td>
</tr>
<tr>
<td>Rabbit</td>
<td>3.2e-15</td>
<td>3.2e-12</td>
</tr>
<tr>
<td>Bovine</td>
<td>-</td>
<td>2.3e-08</td>
</tr>
<tr>
<td>Liver fluke</td>
<td>-</td>
<td>3.2e-15</td>
</tr>
<tr>
<td>Hornworm</td>
<td>-</td>
<td>2.3e-08</td>
</tr>
<tr>
<td>Fruit fly</td>
<td>-</td>
<td>3.0e-01</td>
</tr>
<tr>
<td>Slime mold</td>
<td>-</td>
<td>1.9</td>
</tr>
</tbody>
</table>

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

- DNA match because a longer sequence
- Codon degeneracy and simple scoring give rise to less significant e-values for DNA matches
...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?

User manual provides help

<table>
<thead>
<tr>
<th>PROGRAM</th>
<th>MATRIX</th>
<th>GAP OPEN</th>
<th>GAP EXTEND</th>
<th>KTUP</th>
<th>EXPECTATION UPPER VALUE</th>
<th>EXPECTATION LOWER VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>FASTA</td>
<td>BLOSUM50</td>
<td>-10</td>
<td>-2</td>
<td>2</td>
<td>10</td>
<td>0 (default)</td>
</tr>
<tr>
<td></td>
<td>DNA STRAND</td>
<td>N/A</td>
<td>no</td>
<td>none</td>
<td>Regress</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SCORES</td>
<td>50</td>
<td>50</td>
<td>START-END</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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)
### Choice of Matrix depends on:

#### Databases

<table>
<thead>
<tr>
<th>Protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matrix</td>
</tr>
<tr>
<td>BLOSUM50</td>
</tr>
<tr>
<td>BLOSUM50</td>
</tr>
<tr>
<td>BLOSUM62</td>
</tr>
<tr>
<td>BLOSUM80</td>
</tr>
<tr>
<td>PAM120</td>
</tr>
<tr>
<td>PAM250</td>
</tr>
<tr>
<td>MDM10</td>
</tr>
<tr>
<td>MDM20</td>
</tr>
<tr>
<td>MDM40</td>
</tr>
</tbody>
</table>

#### Matrix

<table>
<thead>
<tr>
<th>Length of Query Sequence</th>
<th>BLOSUM 80</th>
<th>BLOSUM 62</th>
<th>BLOSUM 45</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;300</td>
<td>BLOSUM50</td>
<td>-10</td>
<td>-2</td>
</tr>
<tr>
<td>85–300</td>
<td>BLOSUM62</td>
<td>-7</td>
<td>-1</td>
</tr>
<tr>
<td>50–85</td>
<td>BLOSUM80</td>
<td>-16</td>
<td>-4</td>
</tr>
<tr>
<td>&gt;300</td>
<td>PAM250</td>
<td>-10</td>
<td>-2</td>
</tr>
<tr>
<td>85–300</td>
<td>PAM120</td>
<td>-16</td>
<td>-4</td>
</tr>
<tr>
<td>35–85</td>
<td>MDM40</td>
<td>-12</td>
<td>-2</td>
</tr>
<tr>
<td>&lt;=35</td>
<td>MDM20</td>
<td>-22</td>
<td>-4</td>
</tr>
<tr>
<td>&lt;=10</td>
<td>MDM10</td>
<td>-23</td>
<td>-4</td>
</tr>
</tbody>
</table>

### Choice of Matrix

1. **Strictness of Search**

   - **Less divergent**
   - **More divergent**

2. **Length of Query Sequence**

   - **QUERY LENGTH**
     - >300
     - 85–300
     - 50–85
     - >300
     - 85–300
     - 35–85
     - <=35
     - <=10
   - **MATRIX**
     - BLOSUM50
     - BLOSUM62
     - BLOSUM80
     - PAM250
     - PAM120
     - MDM40
     - MDM20
     - MDM10
   - **open**
     - 10
     - 7
     - 16
     - 10
     - 16
     - 12
     - 22
     - 23
   - **ext**
     - 2
     - 1
     - 4
     - 2
     - 4
     - 2
     - 4
     - 4
Example:
Comparing matrices when searching with a short peptide query sequence

query sequence: MDM2_HUMAN
Comparing matrices

**SSEARCH**

- **Blosum50 (default)**
  - UniProtKB/SwissProt

- **Blosum80**
  - UniProtKB/SwissProt

### Comparisons

<table>
<thead>
<tr>
<th>Hit</th>
<th>e()</th>
</tr>
</thead>
<tbody>
<tr>
<td>MDM2_XENLA</td>
<td>4e-70</td>
</tr>
<tr>
<td>MDM4_BOVIN</td>
<td>9e-48</td>
</tr>
<tr>
<td>MDM4_DANRE</td>
<td>6e-17</td>
</tr>
<tr>
<td>XB34_ORYSJ</td>
<td>0.01</td>
</tr>
<tr>
<td>RN157_MOUSE</td>
<td>0.19</td>
</tr>
<tr>
<td>MGRN1_MOUSE</td>
<td>1.2</td>
</tr>
<tr>
<td>MGRN1_HUMAN</td>
<td>2.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hit</th>
<th>e()</th>
</tr>
</thead>
<tbody>
<tr>
<td>MDM2_XENLA</td>
<td>6e-109</td>
</tr>
<tr>
<td>MDM4_BOVIN</td>
<td>9e-51</td>
</tr>
<tr>
<td>MDM4_DANRE</td>
<td>5e-24</td>
</tr>
<tr>
<td>XB34_ORYSJ</td>
<td>0.05</td>
</tr>
<tr>
<td>RN157_MOUSE</td>
<td>6.0</td>
</tr>
<tr>
<td>MGRN1_MOUSE</td>
<td>6.4</td>
</tr>
<tr>
<td>MGRN1_HUMAN</td>
<td>6.4</td>
</tr>
</tbody>
</table>

**Note:**
- Blosum80 stricter than Blosum50
- Blosum80 more significant (close) match
- Blosum50 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

match/mismatch - nucleotide

...instead have...

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

**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**

- Database:
  - Protein
  - Nucleic Acid
Gap penalties

Choice of gap penalties depends on:

1. strictness of search
   - larger penalty → fewer gaps

2. to match scoring matrix

<table>
<thead>
<tr>
<th>QUERY LENGTH</th>
<th>MATRIX</th>
<th>open</th>
<th>ext</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;300</td>
<td>BLOSUM50</td>
<td>-10</td>
<td>-2</td>
</tr>
<tr>
<td>85–300</td>
<td>BLOSUM62</td>
<td>-7</td>
<td>-1</td>
</tr>
<tr>
<td>50–85</td>
<td>BLOSUM80</td>
<td>-16</td>
<td>-4</td>
</tr>
<tr>
<td>&gt;300</td>
<td>PAM250</td>
<td>-10</td>
<td>-2</td>
</tr>
<tr>
<td>85–300</td>
<td>PAM120</td>
<td>-16</td>
<td>-4</td>
</tr>
<tr>
<td>35–85</td>
<td>MDM40</td>
<td>-12</td>
<td>-2</td>
</tr>
<tr>
<td>&lt;=35</td>
<td>MDM20</td>
<td>-22</td>
<td>-4</td>
</tr>
<tr>
<td>&lt;=10</td>
<td>MDM10</td>
<td>-23</td>
<td>-4</td>
</tr>
</tbody>
</table>
**Sequence Searching Tools**

- KTUP = ‘word-length’ of search
- Large word-length $\rightarrow$ less sensitive $\rightarrow$ faster

**Nucleotide search:** fewer bases than amino acids $\rightarrow$ higher KTUP
Example:
Comparing ktup when searching with a short RNA query sequence

query sequence: 23bp RNA
Comparing ktup

query: 23bp RNA

E=50 extends e-value cut-off

ktup6 (default)
FASTA
EMBL release
No hits found

ktup3
EMBL release
hit: 1: AB334817 0.12
2: AY238603 0.14
3: AC101743 0.16
4: AC115920 0.18

ktup2 E=50
FASTA
EMBL release
hit: 1: AB334817 0.074
2: AY238603 0.092
3: AC101743 0.11
4: AC115920 0.12
5: BC098485 22
... 32
11: AL591512

Lower ktup is more sensitive
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
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 short 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
Visual output

Graphical display of results
Select sequence

View alignment

Visual output
Visual output provides an at-a-glance view of the length and position of all matches.
Do you use e-values or % identity?

<table>
<thead>
<tr>
<th>Align.</th>
<th>DB:ID</th>
<th>Source</th>
<th>Length</th>
<th>Score</th>
<th>Identities</th>
<th>Positives</th>
<th>E()</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NRPL2:NRP001827BA</td>
<td>PN:US6407209 B1</td>
<td>124</td>
<td>794</td>
<td>100.0</td>
<td>100.0</td>
<td>2.6E-49</td>
</tr>
<tr>
<td>2</td>
<td>NRPL1:NRP_A00210</td>
<td>PN:EP024329 A2</td>
<td>124</td>
<td>794</td>
<td>100.0</td>
<td>100.0</td>
<td>2.6E-49</td>
</tr>
<tr>
<td>3</td>
<td>NRPL2:NRP001827BA</td>
<td>PN:EP024329 A2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>NRPL2:NRP00221FC9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>TR:C9JUL1</td>
<td>Homo sapiens</td>
<td>197</td>
<td>794</td>
<td>100.0</td>
<td>100.0</td>
<td>4.4E-49</td>
</tr>
<tr>
<td>6</td>
<td>TR:C9JN19</td>
<td></td>
<td>201</td>
<td>794</td>
<td>100.0</td>
<td>100.0</td>
<td>4.5E-49</td>
</tr>
<tr>
<td>7</td>
<td>NRPL2:NRP00221FC9</td>
<td>PN:US2007218466 A1</td>
<td>508</td>
<td>794</td>
<td>100.0</td>
<td>100.0</td>
<td>1.2E-48</td>
</tr>
<tr>
<td>8</td>
<td>NRPL1:NRP_ACH06771</td>
<td>PN:US2007218466 A1</td>
<td>508</td>
<td>794</td>
<td>100.0</td>
<td>100.0</td>
<td>1.2E-48</td>
</tr>
</tbody>
</table>
**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...**

<table>
<thead>
<tr>
<th>DB:ID</th>
<th>Source</th>
<th>Length</th>
<th>Identity%</th>
<th>Similar%</th>
<th>Overlap</th>
<th>E()</th>
</tr>
</thead>
<tbody>
<tr>
<td>SW:ATP6_XENLA</td>
<td>ATP synthase subunit a OS=Xenopus</td>
<td>226</td>
<td>100.0</td>
<td>100.0</td>
<td>226</td>
<td>8.6e-85</td>
</tr>
<tr>
<td>SW:ATP6_DROYA</td>
<td>ATP synthase subunit a OS=Drosoph</td>
<td>224</td>
<td>36.6</td>
<td>70.3</td>
<td>232</td>
<td>9.4e-20</td>
</tr>
<tr>
<td>SW:ATP6_YEAST</td>
<td>ATP synthase subunit a OS=Sacchar</td>
<td>259</td>
<td>31.8</td>
<td>73.2</td>
<td>198</td>
<td>3.8e-16</td>
</tr>
<tr>
<td>SW:ATP6_EMENI</td>
<td>ATP synthase subunit a OS=Emerice</td>
<td>256</td>
<td>27.7</td>
<td>69.1</td>
<td>191</td>
<td>1.5e-11</td>
</tr>
<tr>
<td>SW:ATP6_COCHE</td>
<td>ATP synthase subunit a OS=Cochlio</td>
<td>257</td>
<td>27.6</td>
<td>67.3</td>
<td>196</td>
<td>1.7e-11</td>
</tr>
<tr>
<td>SW:ATP6_TRITI</td>
<td>ATP synthase subunit a OS=Triticu</td>
<td>386</td>
<td>32.9</td>
<td>70.5</td>
<td>146</td>
<td>1.2e-10</td>
</tr>
<tr>
<td>SW:ATP1_TOBAC</td>
<td>ATP synthase subunit a, chloroplast</td>
<td>247</td>
<td>24.3</td>
<td>55.0</td>
<td>222</td>
<td>0.13</td>
</tr>
<tr>
<td>SW:ATP6_ECOLI</td>
<td>ATP synthase subunit a OS=Escheri</td>
<td>271</td>
<td>26.1</td>
<td>63.0</td>
<td>165</td>
<td>0.0084</td>
</tr>
<tr>
<td>SW:GTR5_SHEEP</td>
<td>Solute carrier family 2, facilita</td>
<td>501</td>
<td>26.3</td>
<td>57.1</td>
<td>156</td>
<td>2.9</td>
</tr>
</tbody>
</table>

**Similar % identity scores**

**Different e-values**
e-values or % identity?

**example...**

<table>
<thead>
<tr>
<th>DB:ID</th>
<th>Sequence Searching Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>SW:ATP6_XENLA</td>
<td></td>
</tr>
<tr>
<td>SW:ATP6_DROYA</td>
<td></td>
</tr>
<tr>
<td>SW:ATP6_YEAST</td>
<td></td>
</tr>
<tr>
<td>SW:ATP6_EMENI</td>
<td></td>
</tr>
<tr>
<td>SW:ATP6_COCHE</td>
<td></td>
</tr>
<tr>
<td>SW:ATP6_TRITI</td>
<td></td>
</tr>
<tr>
<td>SW:ATP1_TOBAC</td>
<td></td>
</tr>
<tr>
<td>SW:ATP6_ECOLI</td>
<td></td>
</tr>
<tr>
<td>SW:GTR5_SHEEP</td>
<td></td>
</tr>
</tbody>
</table>

Pattern of conservation indicates homology

No evidence of homology
Check alignments, especially if using a local-local algorithm.
100% identity, but only over 124 / 662 (20%) of sequence
Always **check alignments** to see where and to what extent the query & target sequences match.
Protein and nucleotide search results have additional annotation:

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

Nucleotide Sequences references for entry C9JUL1_HUMAN from UniProtKB

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]

- **Nucleotide Sequences**
- **Genomes**
- **Ontologies**
- **Protein Families**
- **Literature**

e.g. related ENA nucleotide entries
**Functional predictions**

Protein search results have ‘Function Prediction’
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

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)

>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

Download version takes both protein and nucleotide sequence

All search engines of member databases

STEP 1. Enter your input sequence

Enter or paste a PROTEIN sequence in any supported format:

MQNSHGSRQLGGVFGVRFRPLRQSEVHLASEGASDRPCADSIIRRQLNVSGCVSKLGRY
YTGSIFRPAEESRPQWVSKIAQKRONPSCFIAWEIRDLLESGVCTINDRPSV
SITNRLVNLASEKQCMAGDMYDKLRMLNGTQGWSVTRPGYPTVPVQPTQDQCGQQ
EYGENTNISIASSREDSQDAQMLQRQLRJQRTSFTQEIEQIAKFLNKEREHTHYFVDF
ERRSQKIDLRPILQTVVFQSNHRKAKRKWREELNQRQASLNPHSIPSSFSFSTI
QTDQFPVST5FS5TSMGLDSIATLNTGYEALPE5FTAMANLLPQFPVF5QPV535Y5S5MLP
TLPSPVNSGRSYTDYTPHMQTHMNQSPMMGTSGTTSTGLGPSVGSPVQPVGPSEPDSMSQYWNPR
LQ

Or, upload a file: [Browse...]

STEP 2. Select the applications to run

SelectAll  Clear All

- BlastProDom
- FPrintScan
- HMMPIR
- HMMFam
- HMMSSmart
- 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
Recap

SRS text search
- Select library tab
- Select resources
- Create query
- Navigate to sss tools
- Functional predictions
- Create search
- Sequence matches

Sequence search
- Select library tab
- Select resources
- Create query
- Navigate to sss tools
- Functional predictions
- Create search
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/
Contacts:

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