# An Introduction to Molecular Phylogenetic Inference

**Kirill Vinnikov**
PhD Student,
Department of Biology

ITS Cyberinfrastructure Workshop Series
September 9, 2016

# 2. PRACTICAL LAB 1

## Outline

- CASE STUDY dataset
- Introduction to NCBI's sequence resources
- Using different alignment approaches
- Sequence file formats
- Other sequence databases

# 2. PRACTICAL LAB 1

## Reuired Software

- Text editor (e.g. TextWrangler, Notepad++)
- MEGA 7.0 (www.megasoftware.net)
- Sequence Matrix (gaurav.github.io/taxondna/ )

# CASE STUDY: Bears (family Ursidae)

ELSEVIER

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Molecular Phylogenetics and Evolution 32 (2004) 480–494

MOLECULAR PHYLOGENETICS AND EVOLUTION

www.elsevier.com/locate/ympev

## Phylogeny of the bears (Ursidae) based on nuclear and mitochondrial genes

Li Yu,[a,b,c] Qing-wei Li,[d] O.A. Ryder,[e] and Ya-ping Zhang[a,b,*]

[a] *Laboratory of Molecular Biology of Domestic Animals and Cellular and Molecular Evolution, Kunming Institute of Zoology, Kunming 650223, China*
[b] *Laboratory for Conservation and Utilization of Bio-resource, Yunnan University, Kunming 650091, China*
[c] *Graduate School, Chinese Academy of Sciences, Beijing China*
[d] *Liaoning Normal University, Dalian 116029, China*
[e] *Center for Reproduction of Endangered Species, Zoological Society of San Diego, P.O. Box 551, San Diego, CA 92112, USA*

Received 31 May 2003; revised 12 January 2004
Available online 1 April 2004

# CASE STUDY: extant bear species

Polar bear
*Ursus maritimus*

Brown bear
*Ursus arctos*

Sloth bear
*Melursus ursinus*

Panda
*Ailuropoda melanoleuca*

American black bear
*Ursus americanus*

Spectacled bear
*Tremarctos ornatus*

Sun bear
*Helarctos malayanus*

Asian black bear
*Ursus thibetanus*

Credit: Jón Baldur Hlíðberg (Iceland)

# CASE STUDY: molecular markers – cytB & IRBP

**Abstract**

The taxomic classification and phylogenetic relationships within the bear family remain argumentative subjects in recent years. Prior investigation has been concentrated on the application of different mitochondrial (mt) sequence data, herein we employ two nuclear single-copy gene segments, the partial exon 1 from gene encoding interphotoreceptor retinoid binding protein (IRBP) and the complete intron 1 from transthyretin (TTR) gene, in conjunction with previously published mt data, to clarify these enigmatic problems. The combined analyses of nuclear IRBP and TTR datasets not only corroborated prior hypotheses, positioning the spectacled bear most basally and grouping the brown and polar bear together but also provided new insights into the bear phylogeny, suggesting the sister-taxa association of sloth bear and sun bear with strong support. Analyses based on combination of nuclear and mt genes differed from nuclear analysis in recognizing the sloth bears as the earliest diverging species among the subfamily ursine representatives while the exact placement of the sun bear did not resolved. Asiatic and American black bears clustered as sister group in all analyses with moderate levels of bootstrap support and high posterior probabilities. Comparisons between the nuclear and mtDNA findings suggested that our combined nuclear dataset have the resolving power comparable to mtDNA dataset for the phylogenetic interpretation of the bear family. As can be seen from present study, the unanimous phylogeny for this recently derived family was still not produced and additional independent genetic markers were in need.


*Keywords:* Interphotoreceptor retinoid binding protein; Transthyretin; Evolution; Phylogenetic analysis

# NCBI: National Center for Biotechnology Information

www.ncbi.nlm.nih.gov

NCBI Resources How To Sign in to NCBI

NCBI National Center for Biotechnology Information

All Databases Search

- NCBI Home
- Resource List (A-Z)
- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy
- Training & Tutorials
- Variation

## Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | NCBI News | Blog

**Submit**
Deposit data or manuscripts into NCBI databases

**Download**
Transfer NCBI data to your computer

**Learn**
Find help documents, attend a class or watch a tutorial

**Develop**
Use NCBI APIs and code libraries to build applications

**Analyze**
Identify an NCBI tool for your data analysis task

**Research**
Explore NCBI research and collaborative projects

**Popular Resources**
- PubMed
- Bookshelf
- PubMed Central
- PubMed Health
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

**NCBI Announcements**

VAST+ update provides refined alignments
23 Aug 2016
The new version of VAST+ provides a refined structure-based alignment of

September 12th class at NLM: EDirect - Command Line Access to NCBI's Biomolecular Databases
22 Aug 2016

# NCBI: National Center for Biotechnology Information

**All Resources**

Click this button on the left

**Databases**

Scroll through all databases and read their descriptions

**Nucleotide database**

In the search line type `all[filter]`

**Ursidae**

In the 'species' section on the left click 'customize' and type `Ursidae`

# NCBI: nucleotide database

NCBI Resources How To Sign in to NCBI

Nucleotide | Nucleotide | all[filter] | Search

Create alert Advanced Help

**Species**
Animals (110,055,153)
Plants (34,873,683)
Fungi (6,106,011)
Protists (4,697,825)
Bacteria (26,489,369)
Archaea (526,967)
Viruses (2,287,703)
Ursidae (309,546)
Customize ...

**Molecule types**
genomic DNA/RNA (111,736,208)
mRNA (39,508,054)
rRNA (176,233)
Customize ...

**Source databases**
INSDC (GenBank) (172,853,511)
RefSeq (35,924,195)
Customize ...

**Genetic compartments**
Chloroplast (888,551)
Mitochondrion (3,689,309)
Plasmid (152,833)
Plastid (976,374)

**Sequence length**
Custom range...

**Release date**
Custom range...

**Revision date**
Custom range...

Clear all

Show additional filters

Summary 20 per page Sort by Default order Send to:

**Items: 1 to 20 of 209053924**

<< First < Prev Page 1 of 10452697 Next > Last >>

Found 324883430 nucleotide sequences. Nucleotide (209053924) EST (76257001) GSS (39572505)

1. Macellibacteroides sp. HH-ZS, whole genome shotgun sequencing project
4,081,830 rc other DNA
This entry is the master record for a whole genome shotgun sequencing project and contains no sequence data.
Accession: NZ_LZEK00000000.1 GI: 1056764880
GenBank

2. Macellibacteroides sp. HH-ZS contig_00067, whole genome shotgun sequence
348 bp linear DNA
Accession: NZ_LZEK01000067.1 GI: 1056764864
GenBank FASTA Graphics

3. Macellibacteroides sp. HH-ZS contig_00066, whole genome shotgun sequence
397 bp linear DNA
Accession: NZ_LZEK01000066.1 GI: 1056764861
GenBank FASTA Graphics

4. Macellibacteroides sp. HH-ZS contig_00065, whole genome shotgun sequence
397 bp linear DNA
Accession: NZ_LZEK01000065.1 GI: 1056764853
GenBank FASTA Graphics

5. Macellibacteroides sp. HH-ZS contig_00064, whole genome shotgun sequence
540 bp linear DNA
Accession: NZ_LZEK01000064.1 GI: 1056764850
GenBank FASTA Graphics

6. Macellibacteroides sp. HH-ZS contig_00063, whole genome shotgun sequence
629 bp linear DNA
Accession: NZ_LZEK01000063.1 GI: 1056764845
GenBank FASTA Graphics

**Filter your results:**
All (209053924)
Manage Filters

**Results by taxon**

**Find related data**
Database: Select
Find items

**Search details**
all[filter]
Search
See more...

**Recent activity**
Turn Off Clear
all[filter] (209053924) Nucleotide
all[filter] AND ("Ursidae"[ORGN]) (309546) Nucleotide
all (0) Nucleotide
all[filter] AND (animals[filter] AND biomol_genomic[PROP]) (47722949 Nucleotide
all[filter] AND (animals[filter]) (110055153) Nucleotide

# Search Field Descriptors for NCBI Nucleotide Database

| Search Field | Short | Description |
|---|---|---|
| **Accession** | ACCN | Sequence accession number in NCBI |
| **Author** | AUTH | Authors of the publication |
| **Feature Key** | FKEY | See Biological Features in the Feature Table |
| **Filter** | FILT | Filters subsets of the database |
| **Gene Name** | GENE | Gene name |
| **Issue** | ISS | Journal issue |
| **Journal** | JOUR | Journal name |
| **Keyword** | KYWD | Keywords |
| **Organism** | ORGN | Organism name (scientific or common) |
| **Properties** | PROP | Molecular type, source database, and other properties of the sequence record |
| **Protein Name** | PROT | Protein name |
| **Publication Date** | PDAT | Sequence publication date in NCBI |
| **Sequence Length** | SLEN | Total sequence length or the range |
| **Title** | TITL | Words and phrases in the title of the sequence record |
| **Volume** | VOL | Journal Volume Number |

# CASE STUDY: downloading the dataset

**Nucleotide database**

Type in the search line:
`"Yu L"[AUTH] AND "Ursidae"[ORGN] AND "irbp"[TITL]`

Save search results into ‘irbp-ursidae.fasta’ file

Type in the search line:
`“mitochondrion”[filter] AND "Ursidae"[ORGN] AND "Talbot SL"[Author] AND 1140[SLEN]`

Save search results into ‘cytb-ursidae.fasta’ file

# NCBI: GenBank file format

## Ursus maritimus interphotoreceptor retinoid binding protein (irbp) gene, partial cds

GenBank: AY303843.1

FASTA Graphics PopSet

Go to:

```
LOCUS       AY303843                1274 bp    DNA     linear   MAM 25-JUL-2016
DEFINITION  Ursus maritimus interphotoreceptor retinoid binding protein (irbp)
            gene, partial cds.
ACCESSION   AY303843
VERSION     AY303843.1  GI:34732767
KEYWORDS    .
SOURCE      Ursus maritimus (polar bear)
  ORGANISM  Ursus maritimus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Laurasiatheria; Carnivora; Caniformia; Ursidae;
            Ursus.
REFERENCE   1  (bases 1 to 1274)
  AUTHORS   Yu,L., Li,Q.W., Ryder,O.A. and Zhang,Y.P.
  TITLE     Phylogeny of the bears (Ursidae) based on nuclear and mitochondrial
            genes
  JOURNAL   Mol. Phylogenet. Evol. 32 (2), 480-494 (2004)
   PUBMED   15223031
REFERENCE   2  (bases 1 to 1274)
  AUTHORS   Yu,L., Li,Q.W., Ryder,O.A. and Zhang,Y.P.
  TITLE     Direct Submission
  JOURNAL   Submitted (20-MAY-2003) Kunming Institute of Zoology, 32 Jiaochang
            Eastern Road, Kunming, Yunnan 650223, China
FEATURES             Location/Qualifiers
     source          1..1274
                     /organism="Ursus maritimus"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:29073"
     gene            <1..>1274
                     /gene="irbp"
     mRNA            <1..>1274
                     /gene="irbp"
                     /product="interphotoreceptor retinoid binding protein"
     CDS             <1..>1274
                     /gene="irbp"
                     /codon_start=1
                     /product="interphotoreceptor retinoid binding protein"
                     /protein_id="AAQ81317.1"
                     /db_xref="GI:34732768"
                     /translation="PENLMGMQEAIEQATKSREILAISDPQTLAHVLTTGVQSSLNDP
                     RLVISYEPSTLEAPRQAPALTNLTQEELLARLQKGIRHEVLEGNVGYLRVDDIPGQEV
                     VSKLGGFLVASVWRKLMGTSALVLDLRHCTGGRISGIPYVISYLHPGNTVLHVDTIYD
                     RPSNTTTEIWTLPQVQGERYSADKDVVVLTSGHTGGVAEDITYILKQMRRAIVVGERT
                     VGGALDLQKLRIGQSDFFLTVPVSRSLGPLGGGSQTWEGSGVLPCVGTPAEQALEKAL
                     AILTLRRALPGIVRRLQEALQAYYTLVDRVPTLLHHLANMDFSAVVSQEDLVTKLNAG
```

Customize view

**Analyze this sequence**

- Run BLAST
- Pick Primers
- Highlight Sequence Features
- Find in this Sequence

**Related information**

- Full text in PMC
- PopSet
- Protein
- PubMed
- Taxonomy

**Recent activity**

Turn Off Clear

- Ursus maritimus interphotoreceptor retinoid binding protein (irbp) gene, partial... Nucleotide
- "Yu L"[AUTH] AND "Ursidae"[ORGN] AND "irbp"[TITL] (8) Nucleotide
- Type in the search line : "Yu L"[AUTH] AND "Ursidae"[ORGN] AND "i... (0) Nucleotide
- Tremarctos ornatus interphotoreceptor retinoid binding protein (irbp) gene, p Nucleotide
- Tremarctos ornatus irbp (2) Nucleotide

See more...

# CASE STUDY: finding the outgroup

Resources

Open the menu on the top

Sequence analysis

BLAST

blastn

Search the accession number for *Ursus maritimus* against the NCBI nucleotide database: `AY303843.1`

# CASE STUDY: finding the outgroup

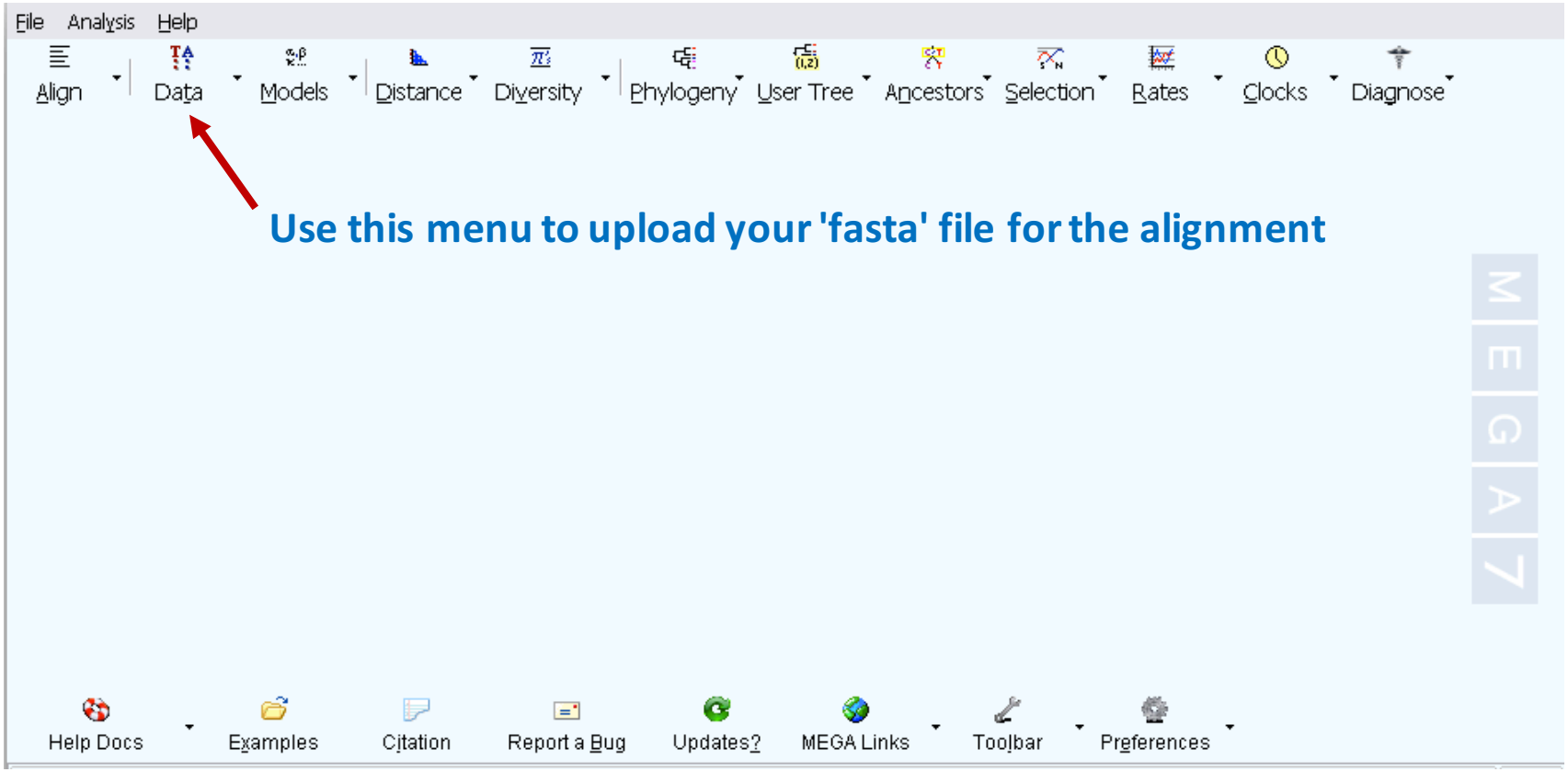
Nucleotide database

Type in the search line:
`“mitochondrion”[filter] AND ”Panthera pardus"[ORGN] AND ”complete genome”[TITL]`

Choose one sequence and save its 15000:16500 region into ‘cytb-outgroup.fasta’ file

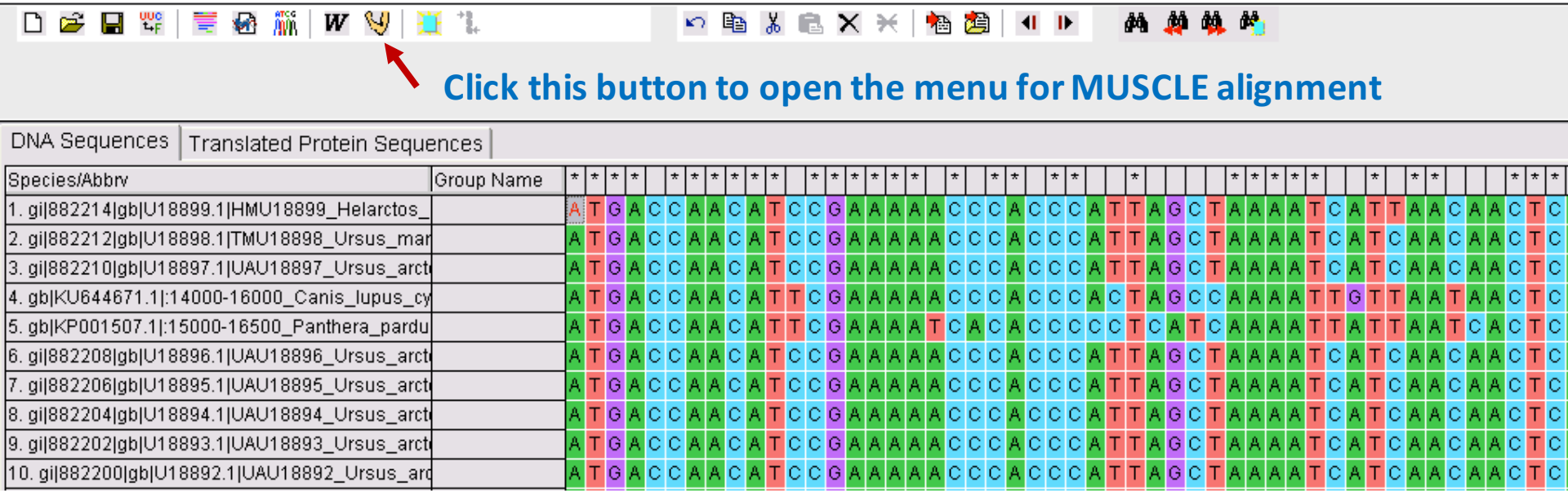
# CASE STUDY: multiple sequence alignment (MUSCLE)

## MEGA: Molecular Evolutionary Genetics Analysis



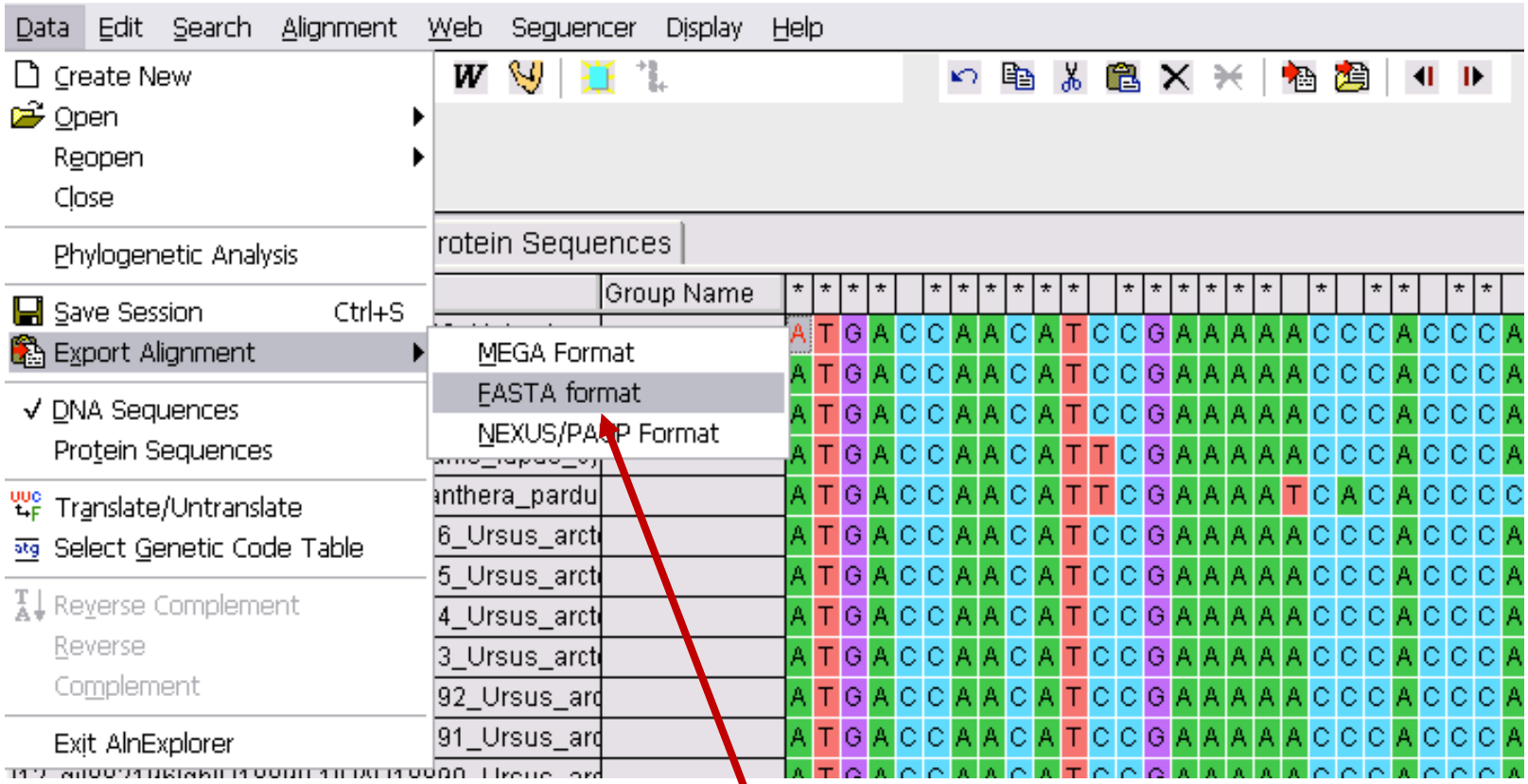
# CASE STUDY: multiple sequence alignment (MUSCLE)

## MEGA: alignment viewer and editor




1. **Trim columns with gaps on the both ends of your alignment!**
2. **Check and correct ORF by translating your DNA sequences**

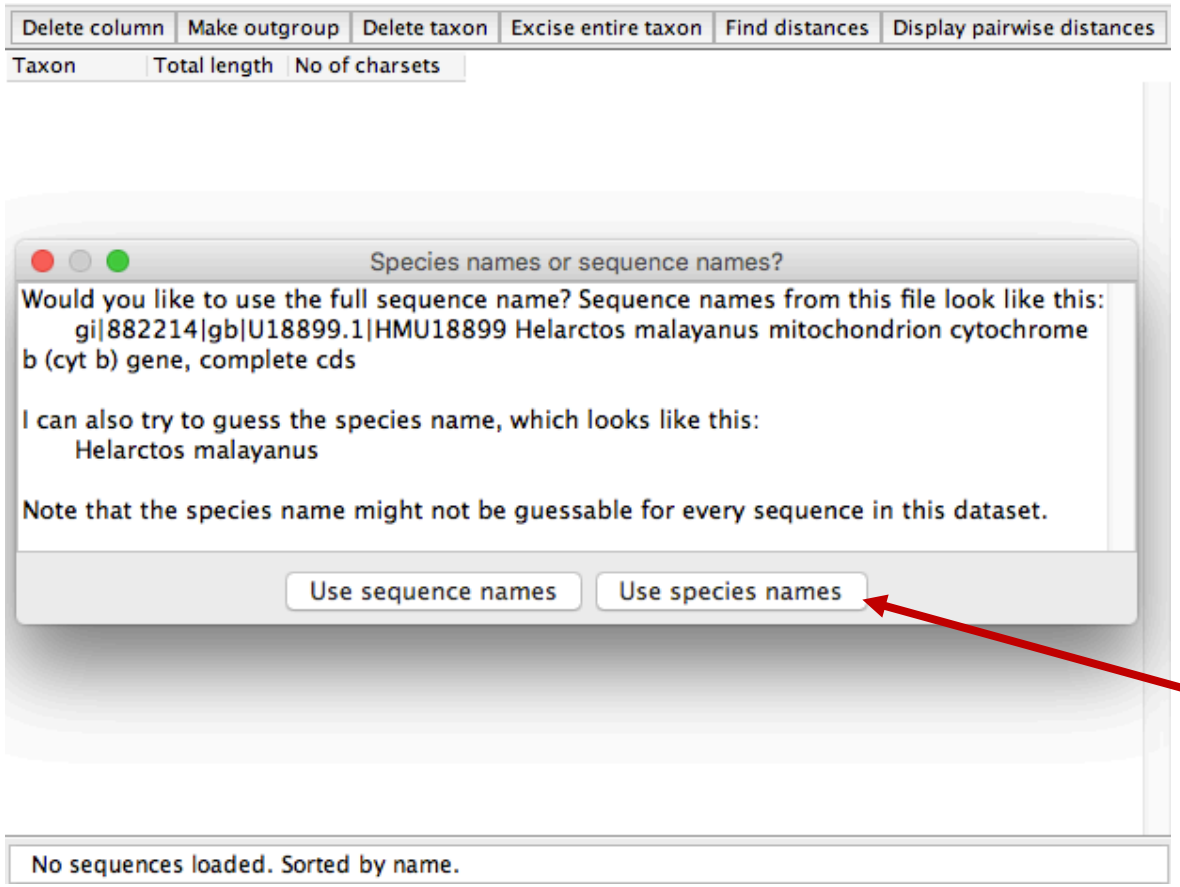
# CASE STUDY: multiple sequence alignment (MUSCLE)




**Export your alignment into fasta file**

# CASE STUDY: filter and concatenate alignment sequences

## Sequence Matrix




**Click to choose one sequence per species**

# CASE STUDY: filter and concatenate alignment sequences

## Sequence Matrix: export files

Delete column | Make outgroup | Delete taxon | Excise entire taxon | Find distances | Display pairwise distances

| Taxon | Total length | No of charsets | irbp_align.fasta |
|---|---|---|---|
| Ailuropoda melanoleuca | 1280 bp | 1 | 1280 (6 indels) |
| Canis lupus | 1280 bp | 1 | 1280 (6 indels) |
| Helarctos malayanus | 1280 bp | 1 | 1280 (6 indels) |
| Melursus ursinus | 1280 bp | 1 | 1280 (6 indels) |
| Panthera pardus | 1280 bp | 1 | 1280 (6 indels) |
| Tremarctos ornatus | 1280 bp | 1 | 1280 |
| Ursus americanus | 1280 bp | 1 | 1280 (6 indels) |
| Ursus arctos | 1280 bp | 1 | 1280 (6 indels) |
| Ursus maritimus | 1280 bp | 1 | 1280 (6 indels) |
| Ursus thibetanus | 1280 bp | 1 | 1280 (6 indels) |

10 taxa, 1 sets. Sorted by name.

Export:

- 'naked' nexus format: irbp.nex
- Phylip format (RAxML): irbp.phy

**Repeat the same format conversion for *cytochrome b***

# CASE STUDY: filter and concatenate alignment sequences

## Sequence Matrix: concatenate two alignments

Delete column | Make outgroup | Delete taxon | Excise entire taxon | Find distances | Display pairwise distances

| Taxon | Total length | No of charsets | cytb_align.fasta | irbp_align.fasta |
|---|---|---|---|---|
| Ailuropoda melanoleuca | 2420 bp | 2 | 1140 | 1280 (6 indels) |
| Canis lupus | 2420 bp | 2 | 1140 | 1280 (6 indels) |
| Helarctos malayanus | 2420 bp | 2 | 1140 | 1280 (6 indels) |
| Melursus ursinus | 2420 bp | 2 | 1140 | 1280 (6 indels) |
| Panthera pardus | 2420 bp | 2 | 1140 | 1280 (6 indels) |
| Tremarctos ornatus | 2420 bp | 2 | 1140 | 1280 |
| Ursus americanus | 2420 bp | 2 | 1140 | 1280 (6 indels) |
| Ursus arctos | 2420 bp | 2 | 1140 | 1280 (6 indels) |
| Ursus maritimus | 2420 bp | 2 | 1140 | 1280 (6 indels) |
| Ursus thibetanus | 2420 bp | 2 | 1140 | 1280 (6 indels) |

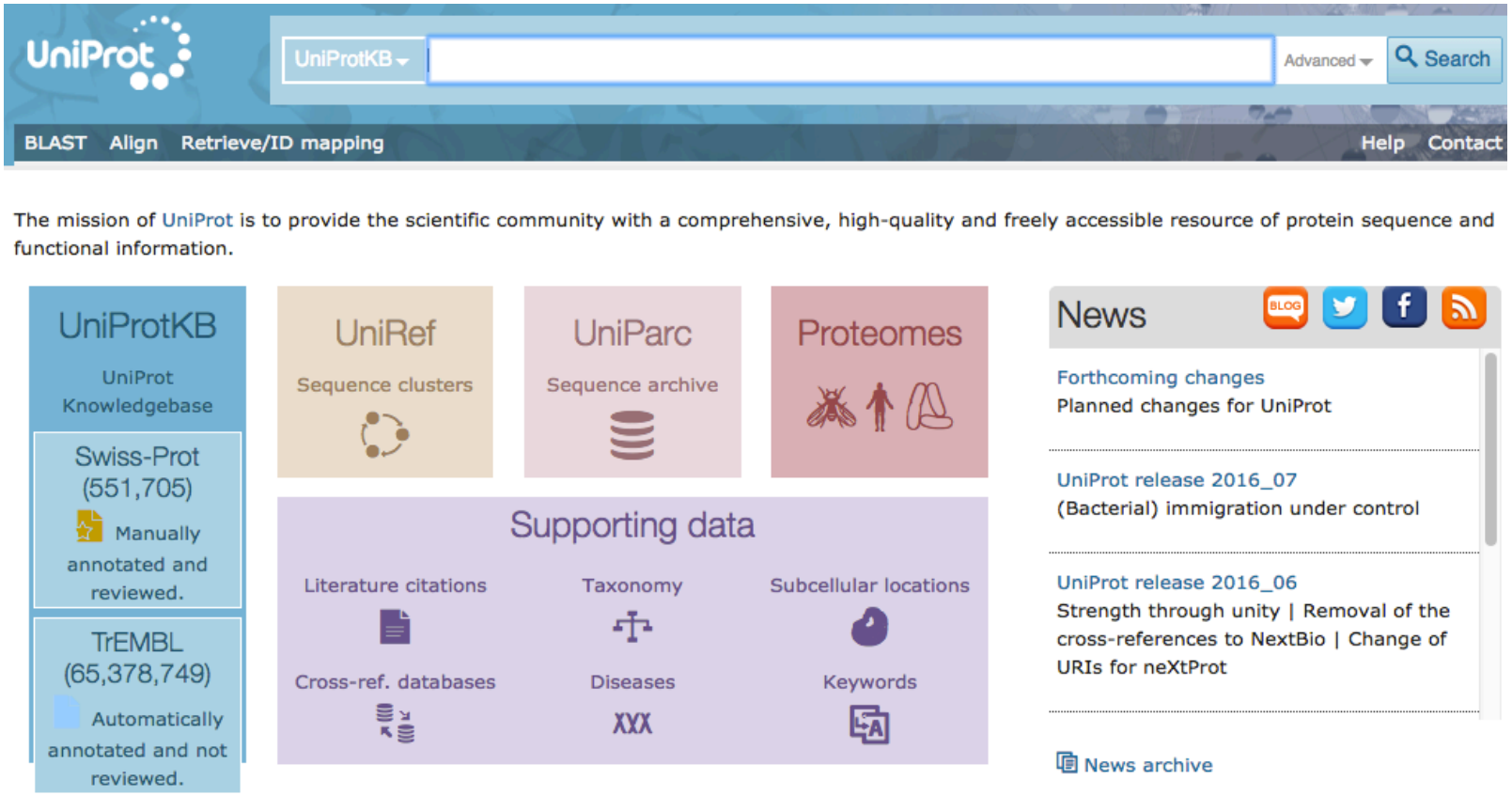
irbp_align.fasta: Ursus arctos. 10 taxa, 2 sets. Sorted by name.

Export:

- 'naked' nexus format: bears.nex
- Phylip format (RAxML): bears.phy

# Other databases: UniProt example

http://www.uniprot.org/



# Other databases: UniProt example

Type in the search line:
`name:"interphotoreceptor retinoid binding"`

Type in the search line:
`name:"interphotoreceptor retinoid binding" AND taxonomy:"Ursidae"`

## UniProtKB results

About UniProtKB | Basket

### Filter by

Reviewed (9)
Swiss-Prot

Unreviewed (4,256)
TrEMBL

### Popular organisms

Rat (7)
Mouse (3)
Zebrafish (2)
Bovine (1)
Human (1)
Other organisms

Go

### Search terms

Filter "interphotoreceptor retinoid binding" as:
protein name

### View by

BLAST | Align | Download | Add to basket | Columns

1 to 25 of 4,265 | Show 25

| Entry | Entry name | Protein names | Gene names | Organism | Length |
|---|---|---|---|---|---|
| P10745 | RET3_HUMAN | **Retinol-binding protein 3** | **RBP3** | Homo sapiens (Human) | 1,247 |
| Q7SZI7 | RET3_XENLA | **Retinol-binding protein 3** | **rbp3** | Xenopus laevis (African clawed frog) | 1,219 |
| P12661 | RET3_BOVIN | **Retinol-binding protein 3** | **RBP3** | Bos taurus (Bovine) | 1,286 |
| P49194 | RET3_MOUSE | **Retinol-binding protein 3** | **Rbp3** | Mus musculus (Mouse) | 1,234 |
| P12664 | RET3_RABIT | **Retinol-binding protein 3** | **RBP3** | Oryctolagus cuniculus (Rabbit) | 23 |
| P12662 | RET3_PIG | **Retinol-binding protein 3** | **RBP3** | Sus scrofa (Pig) | 25 |
| P12663 | RET3_SHEEP | **Retinol-binding protein 3** | **RBP3** | Ovis aries (Sheep) | 24 |
| P12666 | RET3_CAVPO | **Retinol-binding protein 3** | **RBP3** | Cavia porcellus (Guinea pig) | 19 |
| P12665 | RET3_CRISP | **Retinol-binding protein 3** | **RBP3** | Cricetidae sp. (Hamster) | 15 |
| Q9R0I0 | Q9R0I0_RAT | **Interphotoreceptor retinoid binding...** | **Rbp3** irbp | Rattus norvegicus (Rat) | 261 |
| O57689 | O57689_DANRE | **Interphotoreceptor retinoid-binding** | **irbp** IRBP | Danio rerio (Zebrafish) (Brachydanio rerio) | 628 |