

MODULARIZACE VÝUKY EVOLUČNÍ A EKOLOGICKÉ BIOLOGIE

CZ.1.07/2.2.00/15.0204



Metodologie molekulární fylogeneze a taxonomie hmyzu

Bi7770

Andrea Tóthová



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

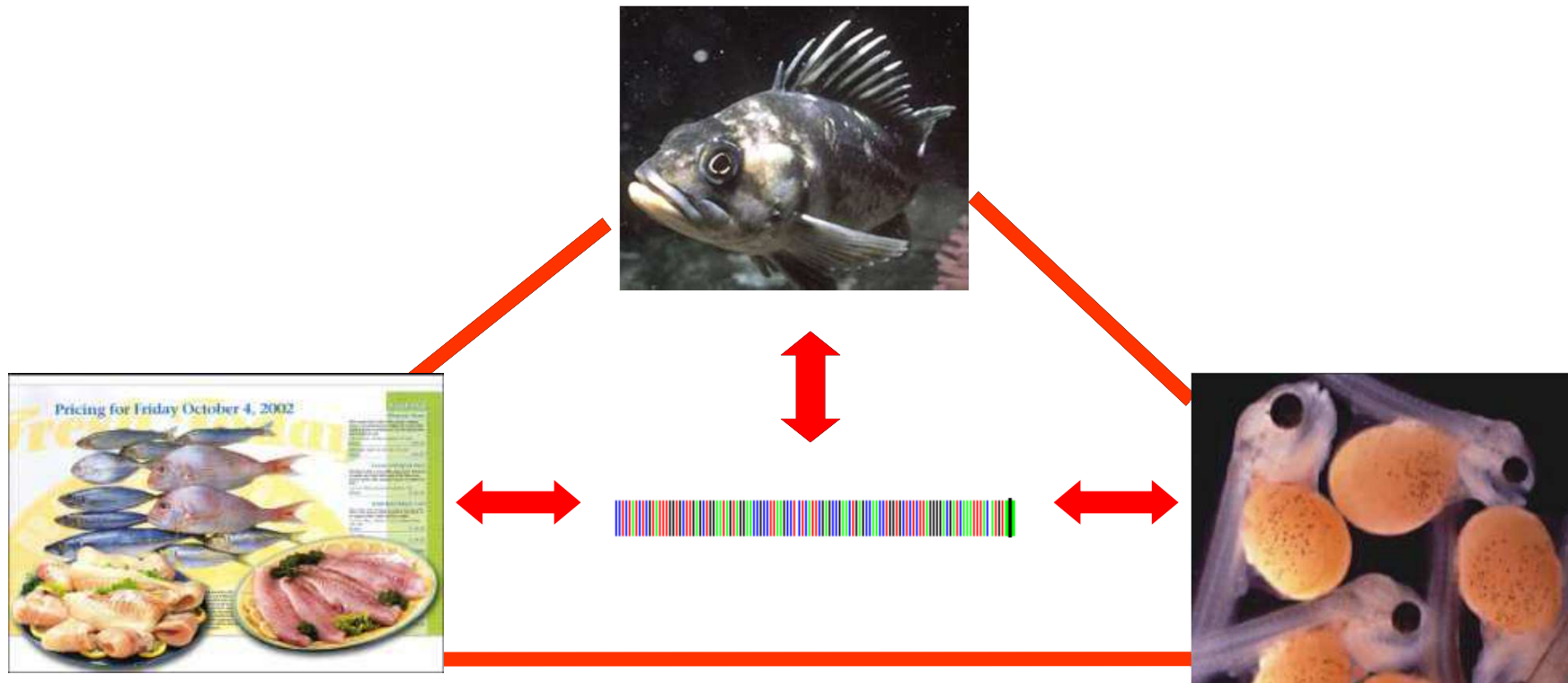
Dva pohledy genomiky

- Vertikální: kompletné genomy, hluboké, ale omezené informace
 - několik druhů a jedinců
 - úspěšně ukončeno u člověka, Drosophily a některých plodin
- Horizontální: krátké cílené sekvence, plytké, ale široké informace
 - mnoho druhů a jedinců
 - např. DNA barcoding

Historické výzvy

- Problémy s konceptem druhu a jeho aplikacemi
- Problémy s druhovou identifikací
- Systém znaků – morfologie, genetika, atd.
- Přístup k existujícím informacím
- Snižování odbornosti
- Snižování dostupných služeb
- Genomika a Internet nabízí nové možnosti

Asociace vývojových stádií, zpracovaných částí a dimorfických pohlaví



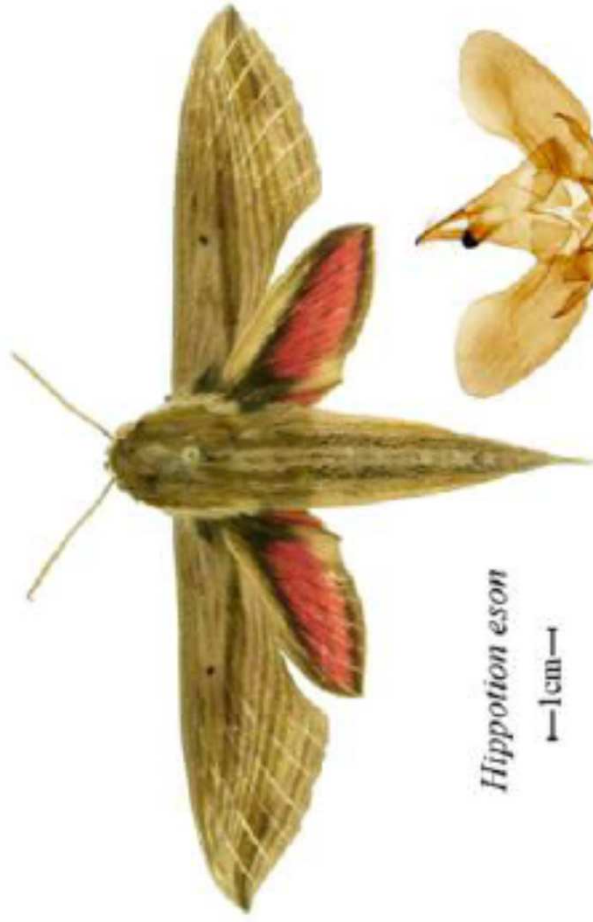
System čárového kódu

- **DNA barcode je krátká genová sekvence standardizované části genomu použitá k druhové identifikaci**



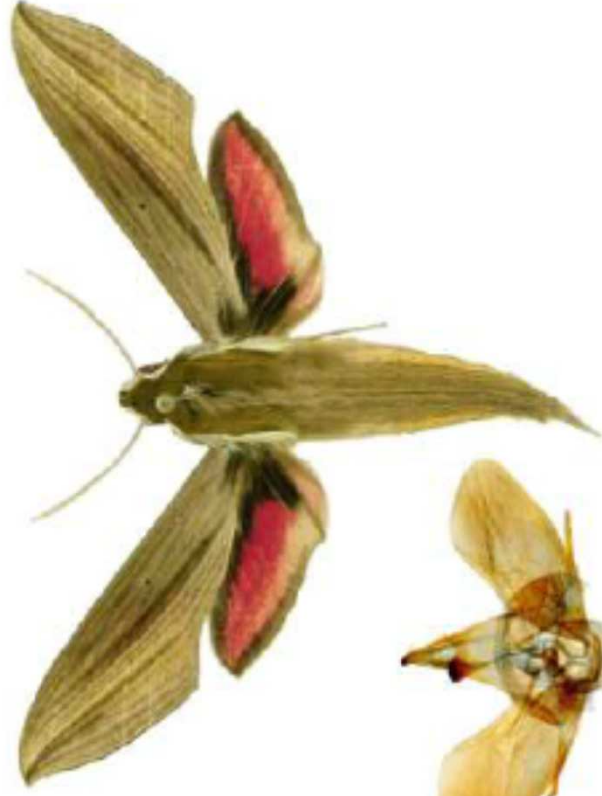
Příklad: Afriční motýli

- Tvoří species komplexy
- Jejich parazitoidi (Tachinidae) také (Dittrich et al 2006)
- *Leguminivora ptychora* na luštěninách je také species komplex
- Ale někteří škůdci jsou široko rozšíření, např. *Spoladea recurvalis* a *Maruca vitrata*



Hippotion eson

←1cm→



USNM slide 125245

USNM slide 125247

Rychlé a efektivní

- Čeleď Sphingidae – vzorky 49 druhů za 6 měsíců (téměř kompletní lokální fauna)
- DNA barcoding rozlišil druhy jak v lokálním, tak v globálním měřítku
- Místní knihovny mohou být rychle srovnány a přispět ke globálním knihovnám

Jak Barcoding funguje

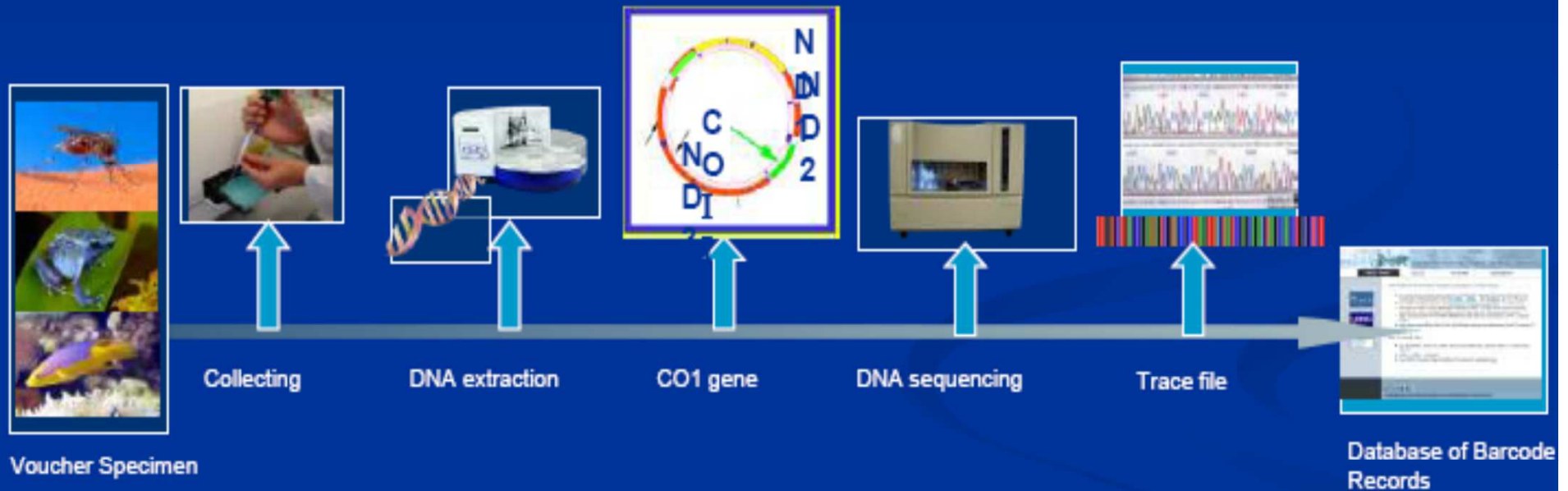
Tvorba referenční knihovny:

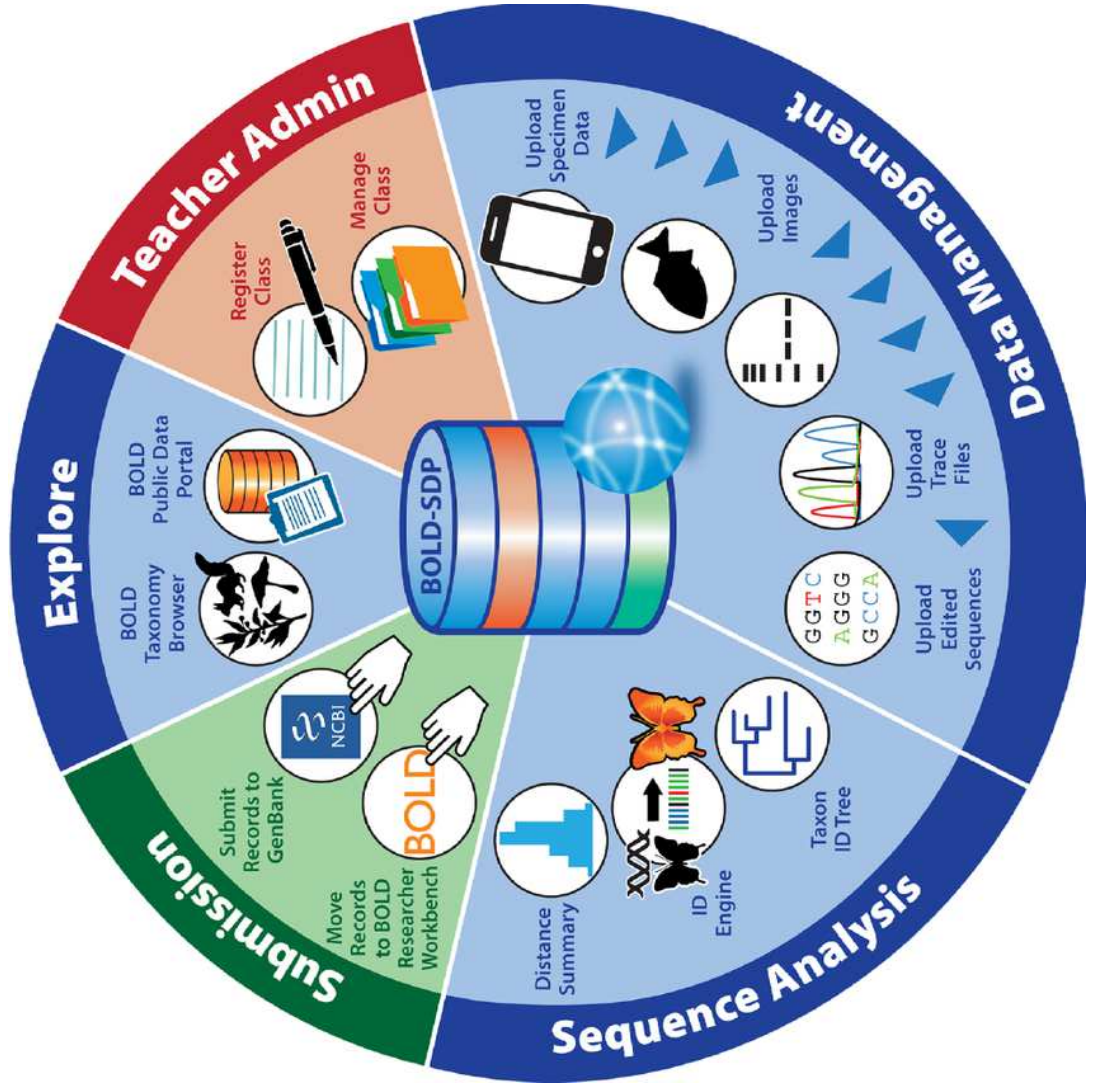
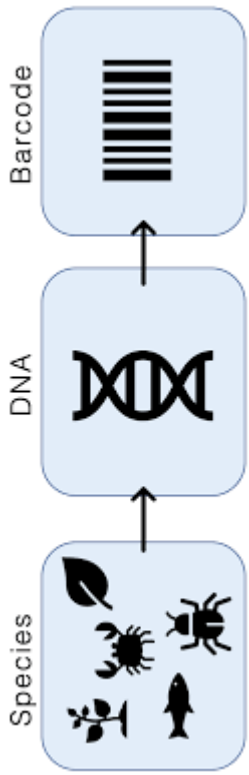
- Správně určený jedinec (vouchers)
- Vzorek tkáně
- DNA extrakce, PCR amplifikace
- DNA sekvenování
- Odeslání dat do GenBanku

Použití referenční knihovny :

- Neurčené druhy
- Tkáň, DNA, sekvenování
- Srovnání s referenčními sekvencemi

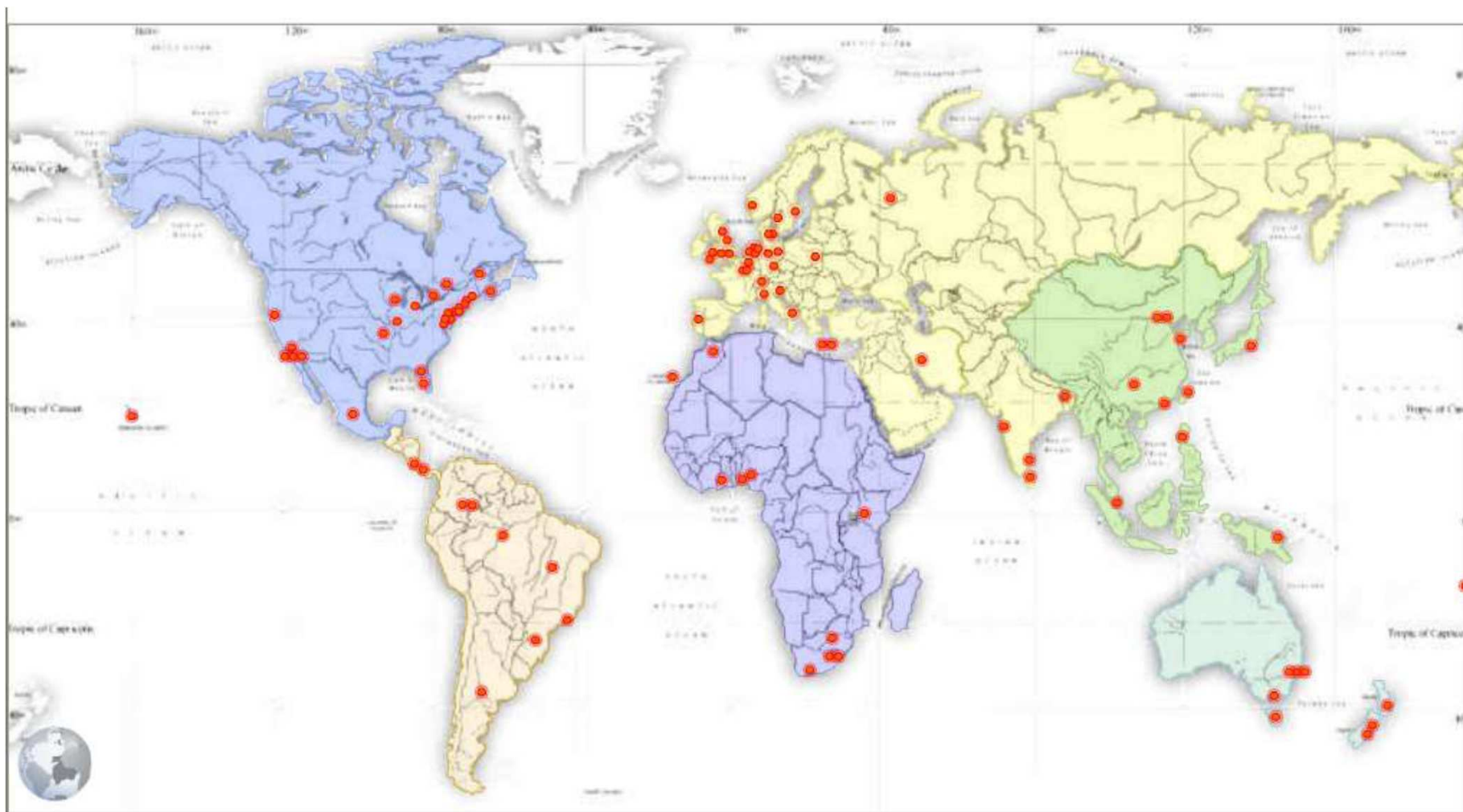
Jak se to vše děje od jedince přes sekvenci po druh?





CBOL – organizace členů od r. 2008

Více než 170 organizací z více než 50 zemí (z toho 54 organizací z 20 rozvojových zemí)



Mise CBO Lu: uvést DNA Barcoding jako globální standard

1. Vyvinout a zvednout standardy komunity
2. Barcode projekty plnit databáze
3. Globální participace a koordinace
4. Přijetí taxonomickou komunitou
5. Koordinace s jinými oblastmi vědy
6. Přijetí regulačními agenturami
7. Vyvíjení produktů soukromými společnostmi

Propojení GenBanku s vouchery

Registry of Biological Repositories

Institutional Acronyms and Collections Codes



Home Institutional Repositories Non-Institutional Repositories FAQ Contact Us

Institution

Search by

or Institution Name Acronym or Location search

Please find your institution and edit the associated data. The institution will be contacted and the new data will be confirmed before it is posted.

Click on the column header to sort institutions by Acronym, Name or Country

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

1 2 3 Next > Last >

Acronym	Institution Name	Address	City	Country	Status	Action
A	Arnold Arboretum, Harvard University	22 Divinity Avenue, Cambridge, Massachusetts, 02138		U.S.A.	Unconfirmed	Edit
AA	Ministry of Science, Academy of Sciences	44 Temirajzev Street, Alma-Ata 480070		Kazakhstan	Unconfirmed	Edit
AAAG	Alan and Anita Gilgoly				Unconfirmed	Edit
AAH	Arnold Arboretum, Harvard University	The Arborway, Jamaica Plain, Massachusetts, 02130		U.S.A.	Unconfirmed	Edit
AAPI	Plant Industry Laboratory	Alberta, Edmonton, Alberta Agriculture		Canada	Unconfirmed	Edit

On This Page

- Instructions for Users
- Searching datafields
- Sorting columns
- Alphabet Index
- Status column

Progress

- 6,836 Total Institutional Records
- 3 Confirmed Institutional Records
- 0 Confirmed Non-Institutional Records
- 2 Confirmed Collections


Sponsors



BOLD Specimen Webpage

BOLDSYSTEMS | Management & Analysis

Hesperia de la ACG 1 (CSOR)



BOLDSYSTEM | Management & Analysis


Hesperia de la ACG 1 (CSOR)

Specimens Identifiers

Sample ID:	02-2004-10276	Material ID:	02-2004-10276
Include / Field Name:		Collection Code:	
Display By:		Number of Fields:	3

Photographs

Dorsal View



BOLDSYSTEMS | Management & Analysis

Hesperia de la ACG 1 (CSOR)

Specimens Identifiers

Sample ID:	02-2004-10276	Material ID:	02-2004-10276
Include / Field Name:		Collection Code:	
Display By:		Number of Fields:	3

Taxonomy

Identifier:	13
kingdom:	Animalia
class:	Insecta
order:	Lepidoptera
family:	Hesperiidae
subfamily:	Pyraustinae
genus:	<i>Araucario</i>
species:	<i>Araucario oleria</i>

Specimen Details

Voucher Type:	
Issue Type:	Pyraustinae
Initials:	rs
Sex:	♂
Parasitoid:	
Life Stage:	

Collection Data

Collectors:	Proctor, Morison
Date Collected:	12-Jul-2002
Country:	Costa Rica
State/Province:	Guanacaste
Region:	Area de Conservacion 276
Sector:	Costa Rica
Elev. Site:	0 m
Latitude:	11.8291
Longitude:	-85.4732
Cont. Source:	
Derivation Depth:	200

Photographs

Dorsal View



Ventral View



IDENTIFICATION ENGINE

[ANIMAL IDENTIFICATION \[COI\]](#)
[FUNGAL IDENTIFICATION \[ITS\]](#)
[PLANT IDENTIFICATION \[RBCL & MATK\]](#)

The BOLD Identification System (IDS) for COI accepts sequences from the 5' region of the mitochondrial cytochrome c oxidase subunit I gene and returns a species-level identification when one is possible. Further validation with independent genetic markers will be desirable in some forensic applications.

Historical Databases: [Current](#) [Jul-2019](#) [Jul-2018](#) [Jul-2017](#) [Jul-2016](#) [Jul-2015](#) [Jul-2014](#) [Jul-2013](#) [Jul-2012](#) [Jul-2011](#) [Jul-2010](#) [Jul-2009](#)

Search Databases:

 All Barcode Records on BOLD (6,780,198 Sequences)

Every COI barcode record on BOLD with a minimum sequence length of 500bp (warning: unvalidated library and includes records without species level identification). This includes many species represented by only one or two specimens as well as all species with interim taxonomy. This search only returns a list of the nearest matches and does not provide a probability of placement to a taxon.

 Species Level Barcode Records (3,573,026 Sequences/211,694 Species/92,285 Interim Species)

Every COI barcode record with a species level identification and a minimum sequence length of 500bp. This includes many species represented by only one or two specimens as well as all species with interim taxonomy.

 Public Record Barcode Database (1,813,416 Sequences/127,891 Species/43,134 Interim Species)

All published COI records from BOLD and GenBank with a minimum sequence length of 500bp. This library is a collection of records from the published projects section of BOLD.

 Full Length Record Barcode Database (2,289,451 Sequences/189,716 Species/74,878 Interim Species)

Subset of the Species library with a minimum sequence length of 640bp and containing both public and private records. This library is intended for short sequence identification as it provides maximum overlap with short reads from the barcode region of COI.

Enter fasta formatted sequences in the forward orientation:

Globální projekty CBOL

- Fish Barcode of Life (FISH-BOL) - 30 000 mořských/sladkovodných druhů do r. 2010
- All Birds Barcoding Initiative (ABBI) - 10 000 druhů do r. 2010
- Tephritidae – 2 000 škůdců/prospěšných druhů do r. 2008
- Komáry - 3 300 druhů do r. 2008
- Ohrožené druhy
- Trees of the world

Staré a nové techniky

- CBOL staví na současných taxonomických poznatcích
- Sequence knihovny založeny na voucher jedincích, co dělá vědu opakovatelnou a testovatelnou
- Voucher jedince propojují historické, současné a budoucí výzkum

Příklad: CSIRO studie na bzučivkách mapující rezistence na insekticidy a zjišťování historie pomocí DNA z muzejních jedinců (PNAS 103: 8757)

DNA Taxonomie – pomoc při řešení problému nebo vnášení chaosu?

První záznamy COI do GenBanku – 1996, od té doby ca. 1000 sekvencí dvoukřídlých

V současnosti – ca. 153 000 popsanych druhů dipter – méně než 1% je zařazeno do „Barcoding procesu“

COI nevhodný pro odlišení blízkých druhů

Stanovení hranic druhu – podobnost sekvencí (pairwise distances) -
PROBLÉM

Fylogenetická rekonstrukce příbuzenských vztahů – možné řešení –
multigenový přístup

Světové sbírky hmyzu – nemožnost použít materiál pro analýzy - PROBLÉM

Taxonomie založená výlučně/převážně na DNA analýze – zkreslený pohled

Potřeba propojit s ostatními přístupy – INTEGRATIVNÍ TAXONOMIE

GenBank

- <http://www.ncbi.nlm.nih.gov/genbank/>
- Několik databází – Nucleotide, Protein, PubMed, CoreNucleotide, Structure, Genome, etc.
- Věrohodnost sekvencí vyšší než v databázích CBO Lu
- Součástí je BLAST - „multialign tool“

Po zadání hesla – Insect...

Search results for: **Insect**

Results by database

Results found in 33 databases

Literature	Genes	Proteins
Bookshelf	Gene	Conserved Domains
MeSH	GEO DataSets	Identical Protein Groups
NLM Catalog	GEO Profiles	Protein
PubMed	HomoloGene	Protein Clusters
PubMed Central	PopSet	Sparcle
		Structure
Genomes	Genetics	Chemicals
Assembly	ClinVar	BioSystems
BioCollections	dbGaP	PubChem BioAssay
BioProject	dbSNP	PubChem Compound
BioSample	dbVar	PubChem Substance
Genome	GTR	
Nucleotide	MedGen	
Probe	OMIM	
SRA		
Taxonomy		

Musca domestica cytochrome oxidase subunit II (COII) gene, partial cds; mitochondrial

GenBank: DQ133110.1

[FASTA](#) [Graphics](#) [PopSet](#)

[Go to:](#)

LOCUS DQ133110 663 bp DNA linear INV 13-DEC-2005
DEFINITION Musca domestica cytochrome oxidase subunit II (COII) gene, partial cds; mitochondrial.

ACCESSION DQ133110

VERSION DQ133110.1 GI:72398995

KEYWORDS .

SOURCE mitochondrion Musca domestica (house fly)

ORGANISM [Musca domestica](#)

[Eukaryota](#); [Metazoa](#); [Arthropoda](#); [Hexapoda](#); [Insecta](#); [Pterygota](#); [Neoptera](#); [Endopterygota](#); [Diptera](#); [Brachycera](#); [Muscomorpha](#); [Muscoidea](#); [Muscidae](#); [Musca](#).

REFERENCE 1 (bases 1 to 663)

AUTHORS Dittmar,K., Porter,M.L., Murray,S. and Whiting,M.F.

TITLE Molecular phylogenetic analysis of nycteriid and streblid bat flies (Diptera: Brachycera, Calyptratae): implications for host associations and phylogeographic origins

JOURNAL Mol. Phylogenet. Evol. 38 (1), 155-170 (2006)

PUBMED [16087354](#)

REFERENCE 2 (bases 1 to 663)

AUTHORS Dittmar,K., Porter,M.L., Murray,S. and Whiting,M.F.

TITLE Direct Submission

JOURNAL Submitted (18-JUL-2005) Integrative Biology, Brigham Young

University, 401 WIDB, Provo, UT 84602, USA

FEATURES Location/Qualifiers

source

1..663

/organism="Musca domestica"

/organelle="mitochondrion"

/mol_type="genomic DNA"

/specimen_voucher="Di180"

/db_xref="taxon:7370"

<1..>663

/gene="COII"

<1..>663

/gene="COII"

/codon_start=1

/transl_table=5

/product="cytochrome oxidase subunit II"

/protein_id="AAZ72903.1"

/db_xref="GI:72398996"

/translation="MSTWANLGLQDSSPLMEQLIEFFHDMILVMITVLVGYLMFT
LFFNKYVRYLLHGQTIEIWIILPAILLFFAPSLRLLYLLDEINEPVTIKRIGH
QWYSYEYSDFNNEVFDSYMIPINELPVDGFRLLDVRVVLPMNSQIRLLVTAADVI
HSWTVPALGVKVDGTPGRLNQINFLINRPLFYQCSEICGANHSFMPVIESIPVNY
FIK"

ORIGIN

1 atgtcaaat gagcaaat aggtttcaa gatagttctt ctccattaat agaacaatta
61 atttttttc atgatcatg attaataat ttagtaataa ttacagattt agtcgatat
121 ttaatgtta catatattt taataatat gttatcgtt atttattaca tggacaaca
181 attgaatta ttgaacta ttacactgca attattttt tattcttgc ttccctct
241 ttacgattt tatacttatt agatgaatt aatgaacct cagtaactt aaagctatt
301 ggtcatcaat gatattgaag ttatgaatat tcagatttta ataattgtga atttgattct

Customize view

Analyze this sequence

Run BLAST

Pick Primers

Find in this Sequence

Related information

Related Sequences

PopSet

Protein

PubMed

Taxonomy

Recent activity

Turn Off Clear

Musca domestica cytochrome oxidase subunit II (COII) gene, partial cds; mitochondrion... Nucleotide

Musca domestica-mitochondrial (488) Nucleotide

Musca domestica (1272) Nucleotide

insect (236769) Nucleotide

Nucleotide Sequence (1495 letters) BLAST

See more...

Nucleotide

Nucleotide

Limits Advanced

Search

Help

Display Settings: FASTA

Send:

Change region shown

Customize view

Musca domestica cytochrome oxidase subunit II (COII) gene, partial cds; mitochondrial

GenBank: DQ133110.1

[GenBank](#) [Graphics](#) [PopSet](#)

```
>gi|72398995|gb|DQ133110.1| Musca domestica cytochrome oxidase subunit II (COII)
gene, partial cds; mitochondrial
ATGTCACATGAGCAAATTTAGGTTTACAGATAGTCTTCCATTAAAGAACAAATTTTTC
ATGATCGCATTAAATTTAGTAATATACAGTATAGTCGGATATTTAAGTTTACATTAATTTT
TAATAAATAGTTAATCGGTTATTTATACATGACAAACAATGAAATTAITGAACTATTTACCTGCA
ATTAATTAATCAITGGTTCCCTTCTTACGATTAATTAATTAAGAAATTAAGAACCAI
CAGTAACTTTAAAGGCTAATGGTCAATCAATGATTAAGTATGAATTAATCAGATTTAATAATGTTGA
ATTTGAATCTAATAATCCTACAATGAATACCAGTAGACGGATTCGTTTATAGATGATGATAAT
CGAGTAGTTTTACCAATAAATCTCAAATTCGAAATTTAGTAATCTGCTGCTGATGTAATCAITCAIGAA
CTGTTCCGCTTTAGGTAAAGTTGATGGTACTCCTGGTCGCTAATCAAACAAATTTCTTAATTA
TCGACCAGGTTTATCTAIGGACAAIGTTCAGAAATTTGTGGAGCTAATCATAGTTTTATACCAAATGTA
ATTGAAAGTATTCCTGTAAATTAATTAATAG
```

Analyze this sequence

- Run BLAST
- Pick Primers
- Find in this Sequence

Related information

- Related Sequences
- PopSet
- Protein
- PubMed
- Taxonomy

Recent activity

- Turn Off Clear
- Musca domestica cytochrome oxidase subunit II (COII) gene, partial cds; mitochond... Nucleotide
- Musca domestica mitochondrial (188) Nucleotide
- Musca domestica (1272) Nucleotide

BLAST

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

The screenshot shows the BLAST website interface. At the top, there is a navigation bar with the NIH logo, "U.S. National Library of Medicine", "NCBI National Center for Biotechnology Information", and user links for "andritoth", "My NCBI", and "Sign Out". Below this is a secondary navigation bar with "BLAST" and links for "Home", "Recent Results", "Saved Strategies", and "Help".

The main content area features a "Basic Local Alignment Search Tool" section with a description: "BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance." A "Learn more" link is provided. To the right of this section is a "NEWS" box with the headline "End of updates for BLAST+ version 4 databases (dbV4)", the subtext "Start moving to the new version 5 databases!", and the date "Fri, 27 Sep 2019 16:00:00 EST". A "More BLAST news..." link is also present.

Below the main section is the "Web BLAST" section, which contains three main tool options:

- Nucleotide BLAST**: nucleotide ▶ nucleotide
- blastx**: translated nucleotide ▶ protein
- tblastn**: protein ▶ translated nucleotide
- Protein BLAST**: protein ▶ protein

At the bottom of the page is the "BLAST Genomes" section, which includes a search input field with the placeholder text "Enter organism common name, scientific name, or tax id" and a "Search" button. Below the input field are four radio button options: "Human", "Mouse", "Rat", and "Microbes".



Download BLAST

Get BLAST databases and executables



Use BLAST API

Call BLAST from your application



Use BLAST in the cloud

Start an instance at a cloud provider

Specialized searches

SmartBLAST

Find proteins highly similar to your query

Primer-BLAST

Design primers specific to your PCR template

Global Align

Compare two sequences across their entire span (Needleman-Wunsch)

CD-search

Find conserved domains in your sequence

GEO

Find matches to gene expression profiles

IgBLAST

Search immunoglobulins and T cell receptor sequences

VecScreen

Search sequences for vector contamination

CDART

Find sequences with similar conserved domain architecture

Targeted Loci

Search markers for phylogenetic analysis

Multiple Alignment

Align sequences using domain and protein constraints

BioAssay

Search protein or nucleotide targets in PubChem BioAssay

MOLE-BLAST

Establish taxonomy for uncultured or environmental sequences

NCBI/BLAST/blastn suite

blastn blastp blastx tblastn tblastx

FASTA nr/nrname search nucleotide databases using a nucleotide query motif

Reset Page Bookmark

Other reports: [Search Summary](#) [Taxonomy reports](#) [\[Distance tree of results\]](#)

Graphic Summary

Distribution of 100 Blast Hits on the Query Sequence



Descriptions

Legend for links to other resources: [U](#) UniGene [E](#) GEO [G](#) Gene [S](#) Structure [M](#) Map Viewer [P](#) PubChem BioAssay

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
GO455784.1	Musca domestica cytochrome oxidase subunit I (COI) gene, partial c	2761	2761	100%	0.0	100%	
AB479529.1	Musca domestica mitochondrial COI gene for cytochrome oxidase sub	2750	2750	100%	0.0	99%	
AB479528.1	Musca domestica mitochondrial COI gene for cytochrome oxidase sub	2750	2750	100%	0.0	99%	
AY526196.1	Musca domestica tRNA-Tyr gene, partial sequence; and cytochrome	2750	2750	100%	0.0	99%	
EU154477.1	Musca domestica mitochondrion, partial genome	2739	2739	100%	0.0	99%	
EU627696.1	Musca condegens cytochrome oxidase subunit I (COI) gene, partial c	2244	2244	99%	0.0	93%	
EU627694.1	Musca asiatica cytochrome oxidase subunit I (COI) gene, partial cds;	2222	2222	99%	0.0	93%	
EU627693.1	Musca sorbens cytochrome oxidase subunit I (COI) gene, partial cds;	2222	2222	99%	0.0	93%	
EU627700.1	Musca larvipara cytochrome oxidase subunit I (COI) gene, partial cds	2206	2206	99%	0.0	93%	
EU627698.1	Musca confiscata cytochrome oxidase subunit I (COI) gene, partial c	2178	2178	99%	0.0	92%	
AB479530.1	Musca crassirostris mitochondrial COI gene for cytochrome oxidase su	2167	2167	99%	0.0	92%	
AB479531.1	Musca crassirostris mitochondrial COI gene for cytochrome oxidase su	2161	2161	99%	0.0	92%	
EU627701.1	Musca convexifrons cytochrome oxidase subunit I (COI) gene, partial	2156	2156	99%	0.0	92%	
EU627702.1	Musca formosana cytochrome oxidase subunit I (COI) gene, partial c	2089	2089	99%	0.0	91%	
AB479533.1	Musca bezzii mitochondrial COI gene for cytochrome oxidase subunit	2061	2061	99%	0.0	91%	
EU627695.1	Musca crassirostris cytochrome oxidase subunit I (COI) gene, partial	2056	2056	99%	0.0	91%	
AB479532.1	Musca bezzii mitochondrial COI gene for cytochrome oxidase subunit	2056	2056	99%	0.0	91%	
EU815009.1	Musca domestica isolate JIA-A-1 cytochrome oxidase subunit I (COI)	2049	2049	74%	0.0	100%	
EU627699.1	Musca inferior cytochrome oxidase subunit I (COI) gene, partial cds;	2045	2045	99%	0.0	91%	
EU814999.1	Musca domestica isolate jia21 cytochrome oxidase subunit I (COI) ge	2045	2045	74%	0.0	100%	
EU814993.1	Musca domestica isolate jia14 cytochrome oxidase subunit I (COI) ge	2043	2043	73%	0.0	100%	
FJ114992.1	Musca domestica isolate iia13 cytochrome oxidase subunit I (COI) ad	2039	2039	73%	0.0	100%	

FLY TREE

2004-2008, 30 mil. USD, 649 taxonů, desítky tisíc bp



FLYTREE

Assembling the Diptera Tree of Life

FLYTREE

[Introduction](#)
[About this Grant](#)
[About & Contact Us](#)
[Opportunities](#)

Features

[About Flies](#)
[Pictures](#)
[Fly Morphology](#)
[Fly Nomenclature](#)
[Species Highlights](#)
[Phylogeny](#)
[Publications & Products](#)

News

[Press Releases](#)
[Talking About Flies](#)

Buzz About Flies

[Additional Buzz](#)
[Diptera.org](#)

Latest FLYTREE News:

[view all recent posts](#)

October 5, 2011

[Finding The Fly Tree of Life - The Poster!](#)

March 14, 2011

[Map of the Fly Tree of Life Published!](#)

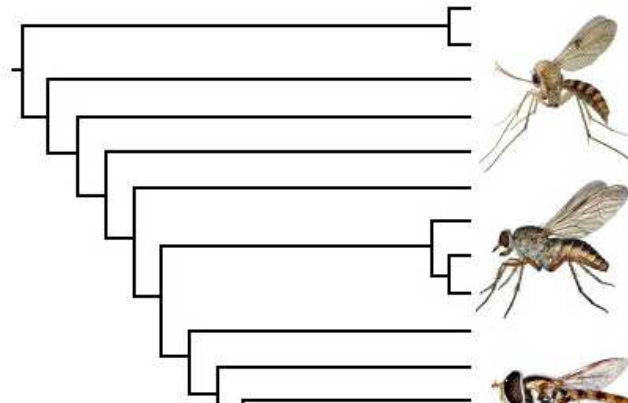
December 8, 2010

[Rediscovering World's Rarest Fly](#)

With over 158,000 described species, flies (the insect order Diptera), are among the most diverse groups of living organisms on the planet.

This diversity transcends simple species numbers and is demonstrated in the great breadth of morphological, ecological, and behavioral variation found in the group.

Flies have a deep evolutionary history that extends back to the Permian Period, over 250 million years ago.





Builders of the Dipteran Tree...

FLYTREE

[Introduction](#)
[About this Grant](#)
[About & Contact Us](#)
[Opportunities](#)

Features

[About Files](#)
[Pictures](#)
[Fly Morphology](#)
[Fly Nomenclature](#)
[Species Highlights](#)
[Phylogeny](#)
[Publications & Products](#)

News

[Press Releases](#)
[Talking About Files](#)

Buzz About Files

[Additional Buzz](#)
[Diptera.org](#)

EDIT Diptera

[//www.cals.ncsu.edu/entomology/wiegmann/](http://www.cals.ncsu.edu/entomology/wiegmann/)
Team of 136.



FLYTREE

Builders of the Dipteran Tree...



FLYTREE

[Introduction](#)
[About this Grant](#)
[About & Contact Us](#)
[Opportunities](#)

Features

[About Files](#)
[Pictures](#)
[Fly Morphology](#)
[Fly Nomenclature](#)
[Species Highlights](#)
[Phylogeny](#)
[Publications & Products](#)

News

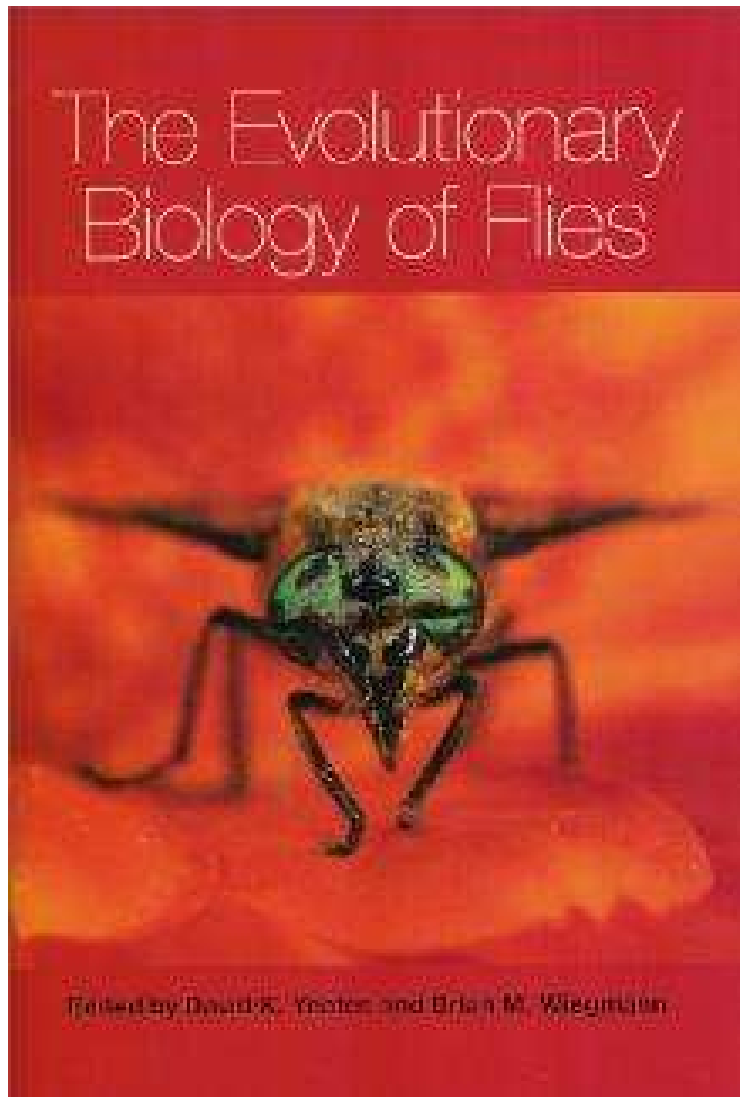
[Press Releases](#)
[Talking About Files](#)

Buzz About Files

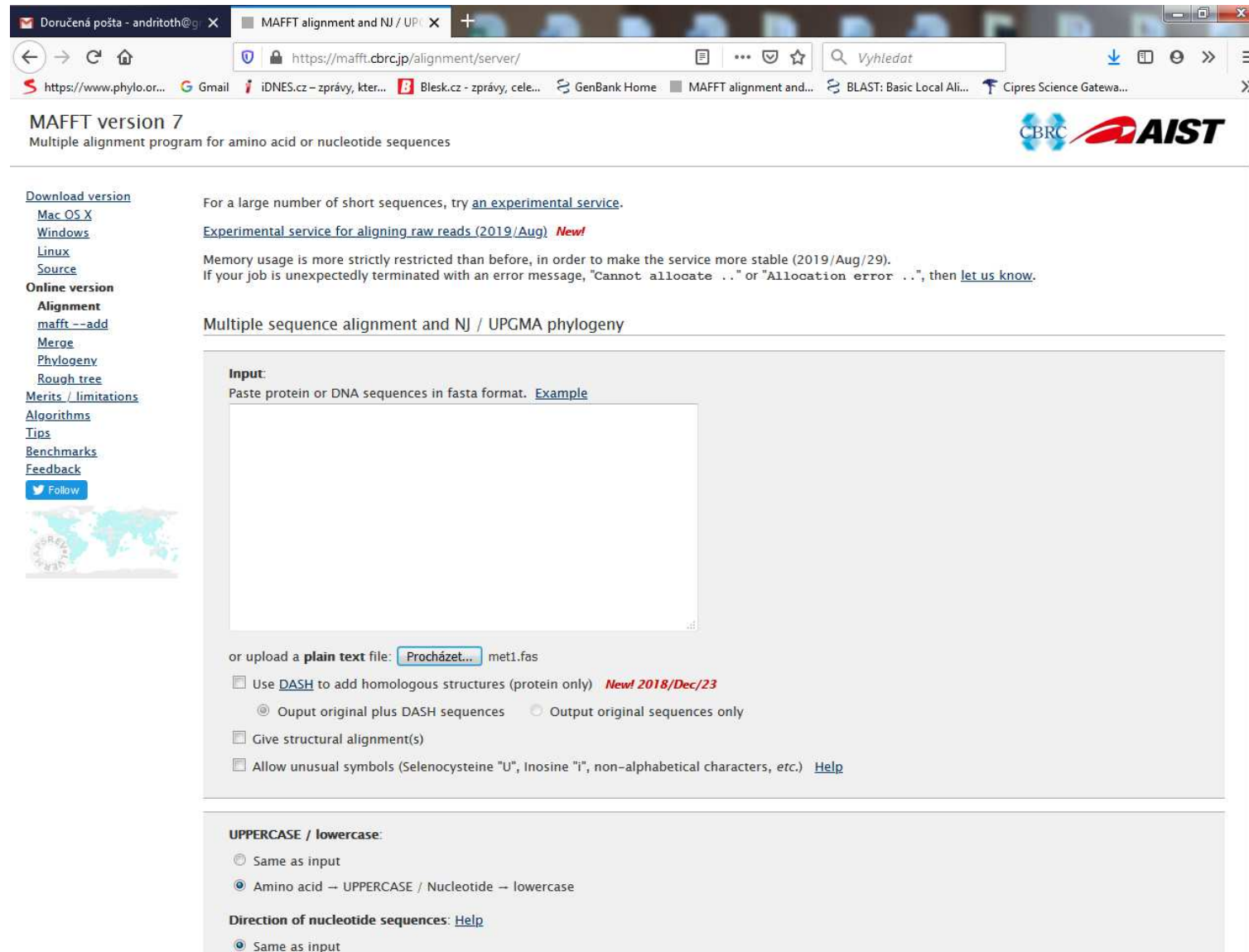
[Additional Buzz](#)
[Diptera.org](#)
Team of 136.



OUTPUTS



MAFFT v. 7 (<https://mafft.cbrc.jp/alignment/server/>)



The screenshot shows a web browser window with the URL <https://mafft.cbrc.jp/alignment/server/>. The page title is "MAFFT version 7" and the subtitle is "Multiple alignment program for amino acid or nucleotide sequences". The browser's address bar shows the URL and a search icon labeled "Vyhledat". The page features a navigation menu on the left with links for "Download version" (Mac OS X, Windows, Linux, Source), "Online version" (Alignment, mafft --add, Merge, Phylogeny, Rough tree, Merits / limitations, Algorithms, Tips, Benchmarks, Feedback), and a "Follow" button. The main content area includes a "Multiple sequence alignment and NJ / UPGMA phylogeny" section. It has an "Input" section with a text area for pasting sequences and a "Procházet..." button. Below this, there are options for "or upload a plain text file" and checkboxes for "Use DASH to add homologous structures (protein only)", "Give structural alignment(s)", and "Allow unusual symbols". There are also radio buttons for "Output original plus DASH sequences" and "Output original sequences only". At the bottom, there are options for "UPPERCASE / lowercase" (Same as input, Amino acid → UPPERCASE / Nucleotide → lowercase) and "Direction of nucleotide sequences" (Same as input).

MAFFT version 7
Multiple alignment program for amino acid or nucleotide sequences

Download version
[Mac OS X](#)
[Windows](#)
[Linux](#)
[Source](#)

Online version
Alignment
[mafft --add](#)
[Merge](#)
[Phylogeny](#)
[Rough tree](#)
[Merits / limitations](#)
Algorithms
[Tips](#)
[Benchmarks](#)
[Feedback](#)
[Follow](#)

For a large number of short sequences, try an [experimental service](#).

[Experimental service for aligning raw reads \(2019/Aug\)](#) **New!**

Memory usage is more strictly restricted than before, in order to make the service more stable (2019/Aug/29).
If your job is unexpectedly terminated with an error message, "Cannot allocate .." or "Allocation error ..", then [let us know](#).

Multiple sequence alignment and NJ / UPGMA phylogeny

Input:
Paste protein or DNA sequences in fasta format. [Example](#)

or upload a **plain text** file: [Procházet...](#) met1.fas

Use **DASH** to add homologous structures (protein only) **New! 2018/Dec/23**

Output original plus DASH sequences Output original sequences only

Give structural alignment(s)

Allow unusual symbols (Selenocysteine "U", Inosine "I", non-alphabetical characters, etc.) [Help](#)

UPPERCASE / lowercase:

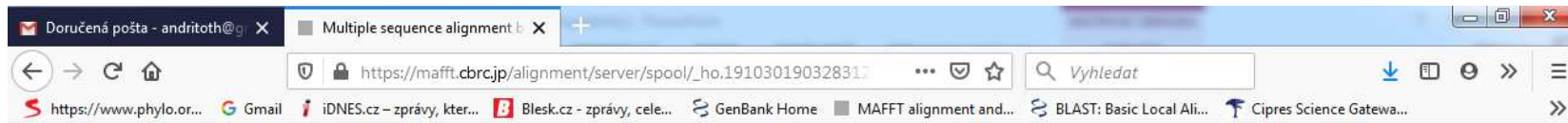
Same as input

Amino acid → UPPERCASE / Nucleotide → lowercase

Direction of nucleotide sequences: [Help](#)

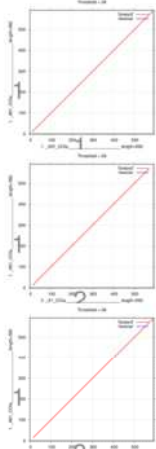
Same as input

MAFFT v. 7 (<https://mafft.cbrc.jp/alignment/server/>)



[LAST](#) hits (score>39) between the top sequence and the others.

[Open all plots](#)



Be careful if there are blue lines. By default, MAFFT considers similarities in forward strands (red) only, but ignores similarities in reverse strands (blue). If blue lines are seen around diagonal regions in the plots above, try the "Adjust direction" option in the input page.

Also note that MAFFT does not assume genomic rearrangements.

[Clustal format](#) | [Fasta format](#) | [MAFFT result](#) | [View](#) | [Tree](#) | [Refine dataset](#) | [Return to home](#)

to GCG, PHYLIP, MSF, NEXUS, uppercase/lowercase, etc. with Readseq

computes the residue-wise confidence scores and extracts well-aligned residues.

MAFFT-L-INS-i Result

CLUSTAL format alignment by MAFFT (v7.452)

```

A61_COIa      nnnnnnnnnGCTGAATAGTGGTACTTCCCTAAGAATTCTAATTCGTCGCGAATTAGGA
A1_COIa       GGTGCATGAGCTGGAATAGCTGGAACCTCTATAAGAATTTAATTCGAACAGAATTAGGA
A2_COIa       GGAGCGTGAGCAGGGAATAGTGGGTACATCCATAAGAATTTAATTCGAACCTGAATTAGGA
A3_COIa       GGAGCATGAGCAGGAATAGCAGGTACATCAATAAGAATTTAATTCGAACAGAATTAGGT
A4_COIa       GGAGCATGAGCAGGAATAGCTGGTACATCAATAAGAATTTCTTATTCGAACAGAATTAGGG
A8_COIa       GGAGCTTGAGCAGGAATAGCAGGAACCTCAATAAGAATTTAATTCGAACCTGAATTAGGT
A9_COIa       GGAGCATGAGCAGGAATAGCTGGTACATCAATAAGAATCCTTATTCGAACAGAATTAGGA
A10_COIa      GGAGCATGAGCAGGAATAGCCGGAACATCTATAAGAATTTCTTATTCGAACAGAATTGGGT
A13_COIa      GGAGCATGAGCAGGAATAGCAGGTACATCAATAAGAATTTCTTATTCGAACAGAATTAGGT
A14_COIa      GGAGCATGAGCAGGAATAGCTGGTACATCAATAAGAATTTCTTATTCGAACAGAATTAGGG
A15_COIa      GGAGCTTGAGCTGGAATAGCTGGAACATCAATAAGAATTTAATTCGAACAGAATTAGGA
A16_COIa      GGAGCATGAGCAGGAATAGCAGGAACCTCAATAAGAATTTCTTATTCGAACAGAATTGGGT
A17_COI_AJA   GGAGCATGAGCAGGAATAGTTGGAACCTCAATAAGAATTTCTTATTCGAACCTGAATTAGGT
A18_COIa      nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
A19_COIa      GGAGCATGAGCAGGAATAGTAGGACTTCAATAAGAATTTCTTATTCGAACCTGAATTAGGT
A20_COIa      nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
A21_COIa      GGTGCTTGAGCAGGAATAGTTGGAACCTCAATAAGAATTTAATTCGAGCAGAATTAGGT
A23          nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
A24_COIa      nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
A25_COIa      GGAGCTTGAGCAGGAATAGCTGGTACATCAATAAGAATTTCTTATTCGAACAGAATTGGGT
A26_COIa      GGAGCATGAGCAGGAATAGCAGGTACATCAATAAGAATTTCTAATTCGAACAGAATTAGGT
A27_COIa      GGAGCTTGAGCAGGAATAGCAGGAACCTCAATAAGAATTTCTAATTCGAACCTGAATTAGGT
A28_COIa      GGAGCATGAGCAGGAATAGCTGGTACATCTATAAGAATTTCTTATTCGAACAGAATTAGGA
A29          nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
A31_COIa      GGAGCATGAGCAGGTATAGCAGGAACATCAATAAGAATTTCTTATTCGAACCTGAATTAGGT
A32_COIa      GGAGCGTGGCAGGAATAGCCGGAACATCAATAAGAATTTCTTATTCGAACAGAATTGGGT
A33_COIa      GGAGCTTGAGCAGGAATAGCTGGTACATCAATAAGAATTTCTTATTCGAACAGAATTAGGT
A43_COIa      nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
A44_COIa      nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
A45_COIa      GGAGCATGAGCAGGAATAGCAGGAACCTCAATAAGAATTTCTTATTCGAACAGAATTAGGT
A46_COIa      GGAGCATGAGCAGGAATAGCAGGAACCTCAATAAGAATTTCTTATTCGAACCTGAATTAGGT
A47_COIa_R_OK GGAGCATGAGCAGGAATAGCAGGAACCTCAATAAGAATTTCTTATTCGAACAGAATTAGGT
A48_COIa      GGAGCATGAGCTGGAATAGCCGGAACGTCGATAAGAATTTCTTATTCGAACAGAATTAGGC
A49_COIa_F_OK nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
A50_COIa      GGAGCATGGGCAGGAATAGCAGGAACCTCAATAAGAATTTCTTATTCGAACCTGAATTAGGT
A51_COIa      GGAGCATGGGCAGGAATAGCTGGAACATCAATAAGAATTTCTTATTCGAACAGAATTAGGT
A52_COIa      GGGGCATGGGCAGGAATAGCCGGAACATCAATAAGAATTTCTTATTCGAACAGAATTGGGT
A53_COIa      GGAGCATGAGCAGGAATAGCTGGAACATCAATAAGAATTTCTTATTCGAACCTGAATTAGGT
    
```


Fylogenetické analýzy

Fylogenetický strom – hypotéza, která vznikla co nejlepším odhadem na základě omezeného zdroje informací

Metody FA – dva přístupy

1. Algoritmus – jde přímo k výsledku, co je jediný strom (odpadá srovnání vzájemně si konkurujících stromů) – metody shlukové analýzy (UPGMA), Neighbour-joining (NJ) – obě využívají data vzdáleností (distance)
2. Kritérium optimálnosti – **dva kroky** – definování kritéria, podle kterého je hodnocen každý strom určitým skóre, které se použije k následnému srovnání všech stromů
 - použití specifického algoritmu pro výpočet funkce (kritérium optimálnosti) a pro získání stromu s nejlepší hodnotou této funkce

Jaká by měla vybraná metoda být?

Výkonnost – „tempus fugit“ nebo „time is money“
pomoc – heuristické metody hledání v případě vyššího
počtu taxonů či znaků

Síla – kolik dat musíme shromáždit, aby byly výsledky správné

Konzistence – s přidáváním dalších znaků spějeme k správnému výsledku

Robustnost – do jaké míry vedou drobné odchýlky od vstupných
předpokladů k nesprávným závěrům

Falzifikovatelnost – určení nevhodnosti modelu na základě odchýlky
od
předpokladu

IDEÁLNÍ METODA NEEXISTUJE...

Metoda maximální parsimonie – úspornosti (MP)

Jedna z nejpoužívanějších metod - rychlá, jednoduchá preferuje jednodušší hypotézy před složitějšími (široká filozofická platnost), tzn. vybere možnost (strom) s minimálním počtem evolučních kroků nutných k vysvětlení vstupních dat
Ne všechny znaky jsou použitelné, parsimony - informative

+

Dobrá pochopitelnost, jednoduchost, rychlost, nízký počet předpokladů (předpokládá, že jakákoli evoluční změna je vzácná, takže MP strom se dá považovat za nejlepší odhad skutečné evoluce)

-

Nekonzistentnost, přitažlivost dlouhých větví (LBA)

Metoda maximální pravděpodobnosti (Maximum likelihood, ML)

- posuzují se jednotlivé hypotézy o evoluční historii zkoumaných taxonů z hlediska pravděpodobnosti, že jsou v souladu se získanými daty, výsledek – maximálně pravděpodobný odhad

Tři součásti - vstupné data

evoluční model

fylogenetický strom s topologií i délkou větví

+

Nízka náchylnost k chybě, robustnost vůči odchýlkám

-

Vysoká výpočetní náročnost při validaci vetví

Bayesian inference

Výpočet pravděpodobnosti na základě specifikovaného modelu a na základě toho, co jsme o charakteru dat zjistili

Základ – strom s danou topologií a délkami větví, model nukleotidových substitucí a rozložení substitučních frekvencí mezi jednotlivými nukleotidy

Princip přístupu jako u ML

VÝHODY – menší časová náročnost, strom zohledňující fylogenetický signál v datasetu, možnost použít i pro smíšený dataset

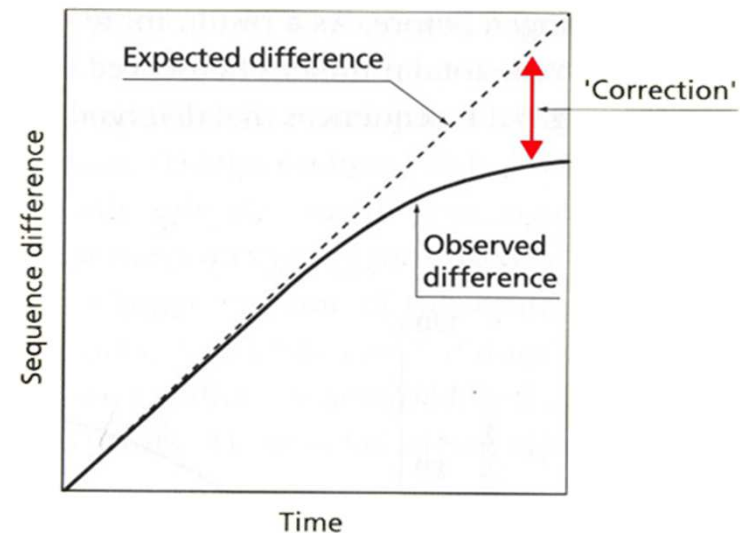
Distanční metody

Založené na podobnostech (vzdálenostech, rozdílech)

Poznání skutečné evoluční vzdálenosti mezi všemi členy studovaného souboru taxonů umožňuje velmi lehkou rekonstrukci evoluční historie těchto taxonů

Opakované změny
jednoho znaku –
korigované distance (jako
u pravděpodobnosti)

Nekorigovaná vzdálenost
– p-distance
Korekce: JC, F81, K2P,
F84, GTR



Předpříprava 😊

- Úprava sekvencí (Sequencher v4.8)
- Vytvoření alignmentu (**MEGA**) - .fas, .nex
- Alignment v MAFFT
- Analýza MP (**Paup**)
- Analýza NJ (Paup, MEGA)
- Vytvoření souboru pro MrModeltest (PAUP)
- MrModeltest
- Příprava souboru pro MrBayes a ML (RAxML vs GARLI)
- Využívání portálu CIPRES