

Retrieval & Alignment Tools

Sequence and Metadata Retrieval

- **DataRetrieval**

Multiple Alignments

- **MSA-PAD**

DataRetrieval: DNA Barcode Sequence and Metadata Retrieval

DataRetrieval is a RESTfull service able to query Public primary and specialized databases to extract DNA barcode sequences and their relevant metadata corresponding to a given taxon

Target DBs

BOLD SYSTEMS Databases | Taxonomy | Identification | Workbench | Resources

Join us at the...
10th International Barcode of Life Conference
August 18-21, 2015
Guelph ON, Canada
dnabarcodes2015.org

Taxonomy [dropdown] [input] Search

About PhytoREF

A new reference database of the plastidial 16S rRNA gene of eukaryotes

The PhytoREF database provides a diversity of photosynthetic eukaryotes. It has been built using the publicly available well as, new amplicon sequences of Stringent quality filtering and phylogenetic analysis of each single 16S rRNA gene sequence. PhytoREF is not only a new tool to photosynthetic eukaryotes in different genomic resource to annotate new species. This new database can be used for many...

ITSoneDB

Navigation icons: Home, Databases, Search, Help

NCBI
National Center for
Biotechnology Information

Fungal Ribosomal Internal Transcribed Spacer 1 Database

ITSoneDB is a comprehensive collection of the fungal ribosomal RNA Internal Transcribed Spacer 1 (ITS1) sequences aimed at supporting metagenomic surveys of fungal environmental communities. The sequences were extracted from GenBank (GB) and arranged on the NCBI taxonomy tree. ITS1 start and end boundaries were defined by GB annotations and/or designed by mapping Hidden Markov Model (HMM) profiles of flanking 18S and 5.8S ribosomal RNA coding genes on each sequence. [?](#)
Current GenBank release: 202 (June 2014)

How to call **DataRetrieval** REST web service?

Required arguments:

- TaxonName
- Kingdom name: Animals, Fungi, Protists, Plants

Address

[http://alicegrid17.ba.infn.it:8080/INFN.Grid.RestFrontEnd/services/QueryJob/InsertJobs?NAME=**Data Retrieval**&arguments=**Cydia pomonella Animals**&mail=e-mail@mail.com](http://alicegrid17.ba.infn.it:8080/INFN.Grid.RestFrontEnd/services/QueryJob/InsertJobs?NAME=Data Retrieval&arguments=Cydia pomonella Animals&mail=e-mail@mail.com)

Current online version: BOLD query

Summary data

- NUM Publicbins
- NUM Sequenced specimens
- NUM Publicrecords
- NUM Public marker sequences
- NUM Barcode specimens
- NUM Specimen records
- Geographical distribution

Outputs

Full data & Metadata: tsv format

- RecordID
- Taxonomy
- institution_storing
- tissue_type
- Collectors
- Collection date
- Life stage
- Sex reproduction
- Latitude
- Longitude
- Genbank accession
- Nucleotides
- sequencing primers

Next release version: Implementation steps

- ✓ Integrating the query over both BOLD and NCBI
- ✓ **Eliminate redundancy: based on the accession numbers** (if NCBI accession number is present in both BOLD and NCBI, that of BOLD is preferred)
- ✓ **DNA sequences mapping against a generic COXI profile to ensure their identity as COXI barcode region**
- ✓ **Taxonomy will be provided with Sequence ID at the six main levels in the following form:**
 - **>TaxonName_AccessionNum;Phylum;Class;Order;Family;Genus;Species**
- ✓ **PCR primers, the retrieved sequences were amplified with, will be provided in a tab-limited format:**
 - **Primers_1** **Id1,Id2,....**
 - **Primers_2** **Id3, Id4,....**
 - **...**
- ✓ **Tool extension to other main Barcode Markers**

Bioinformatics, 31(15), 2015, 2571–2573

doi: 10.1093/bioinformatics/btv141

Advance Access Publication Date: 26 March 2015

Applications Note

OXFORD

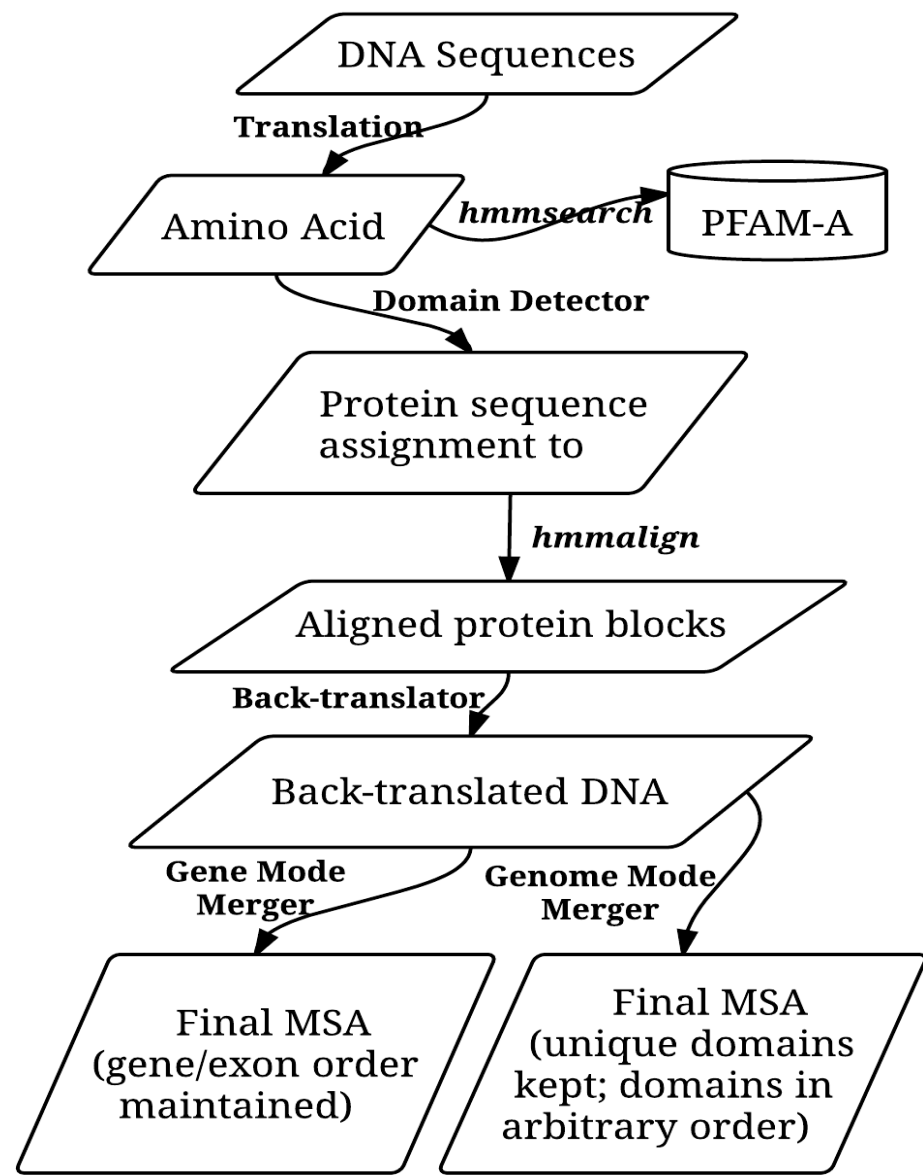
Sequence analysis

MSA-PAD: DNA multiple sequence alignment framework based on PFAM accessed domain information

Bachir Balech¹, Saverio Vicario², Giacinto Donvito³, Alfonso Monaco³, Pasquale Notarangelo³ and Graziano Pesole^{1,4,*}

- It aligns DNA sequences encoding either single or multiple protein domains by two alignment options: **Gene and Genome**
- It makes use of information embedded in protein domains (PFAM domains), intron occurrence and gene order variations (e.g. mitochondrial genomes)

MSA-PAD concept



Lower layer implementation

REST call of JST webservice: Input upload and execution

Wrappers

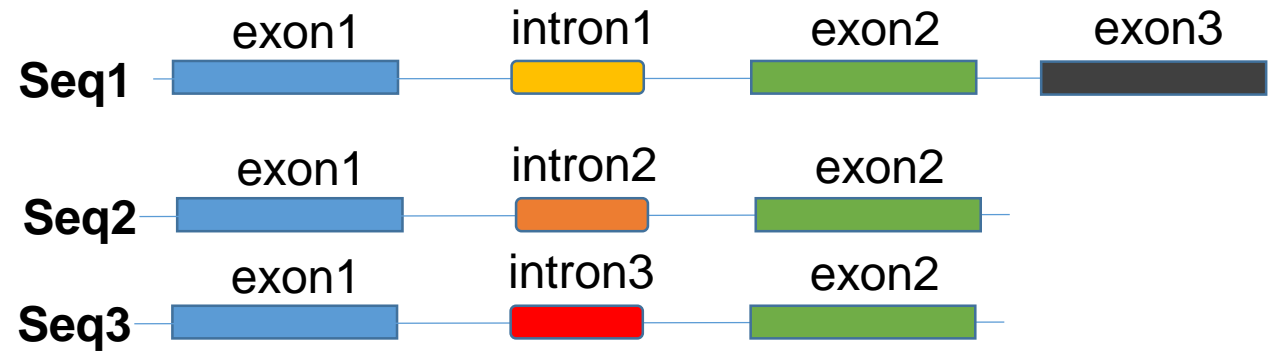
HMMer3.0:
hmmsearch
hmmalign

Python parsers:
Translator.py
Backaligner.py
Merger.py

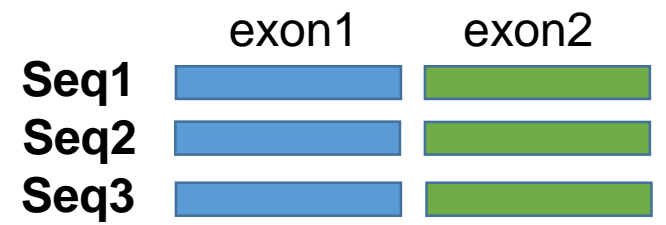
Email Client Answer: Output retrieval

MSA-PAD @ work – Use Cases

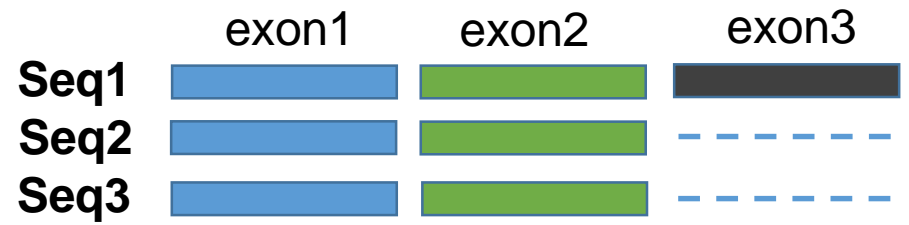
Intron occurrence



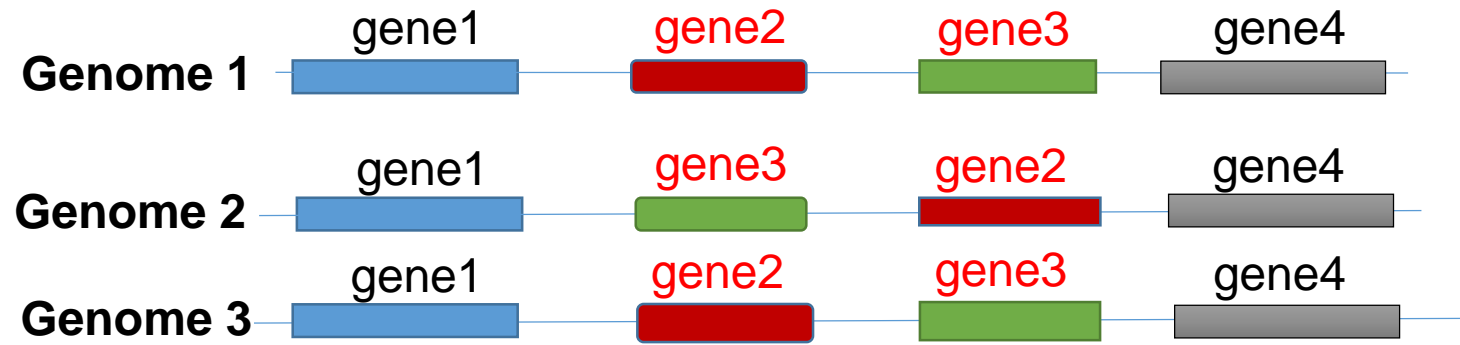
Gene Mode Alignment



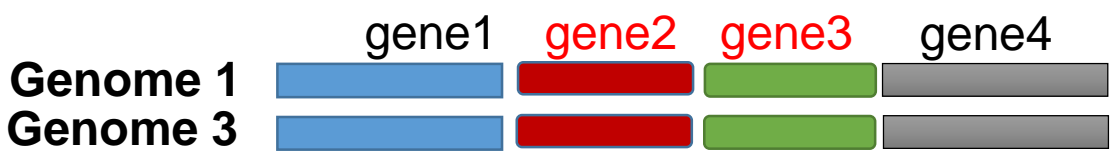
Genome Mode Alignment



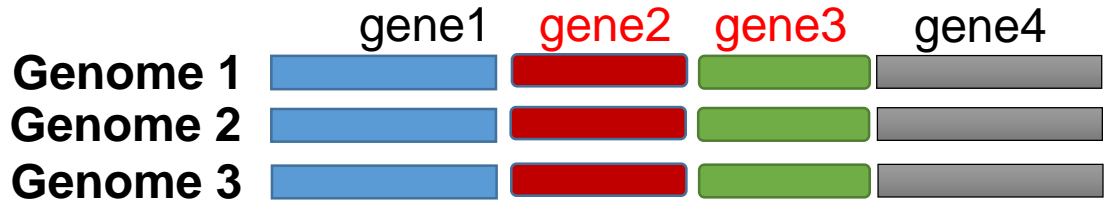
Genomes Rearrangements



Gene Mode Alignment

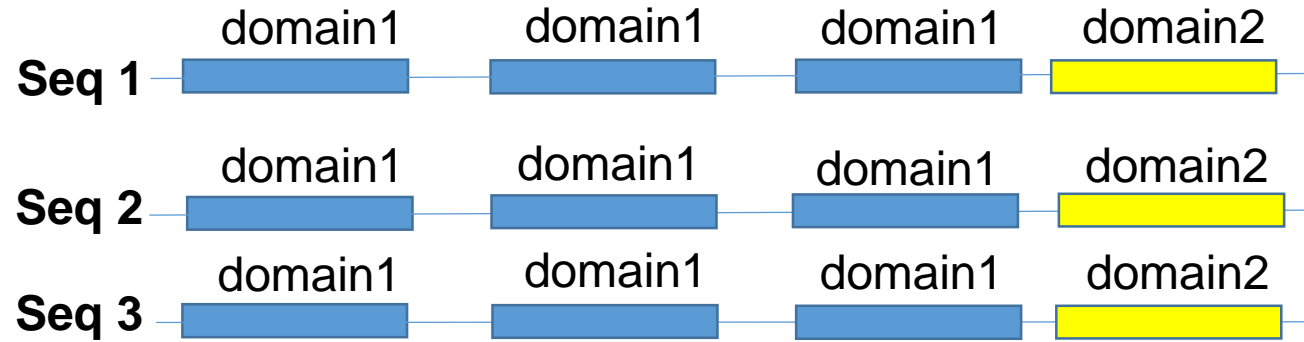


Genome Mode Alignment

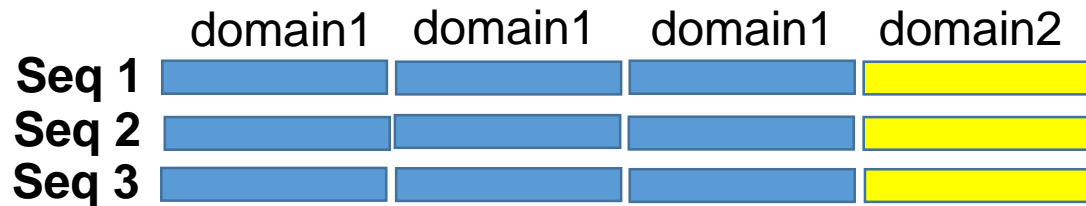


MSA-PAD @ work – Use Cases

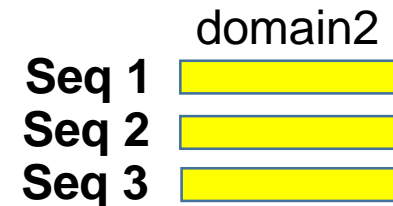
Repeated Domains



Gene Mode Alignment



Genome Mode Alignment



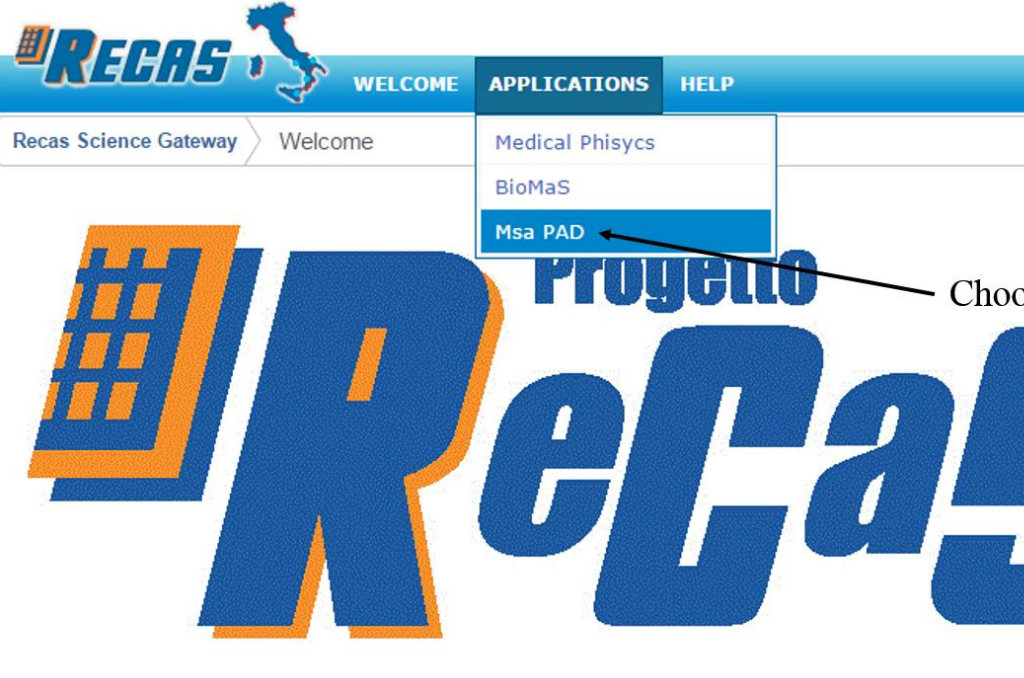
Main outputs:

- **Final multiple DNA alignment** -> FASTA format
- **AlignmentDomainsPartitions** -> the coordinates of each protein domain in the final MSA
- **ExcludedSequencesIDs** -> sequence IDs (separated by comma) not present in the final MSA

Additional outputs:

- ✓ **File/s with *hmmAligned* suffices** -> alignments (STOCKHOLM format) of each protein sequences block with PFAM profile as prefix
- ✓ **File/s with *Backaligned.fasta* suffices** -> alignments of each back-translated DNA sequences block
- ✓ **MissingSites_Report** -> DNA sites position missing from the final MSA

MSA-PAD: Web Application



RECAS Science Gateway | Welcome | **APPLICATIONS** | HELP

- Medical Physics
- BioMaS
- Msa PAD** ← Choose

<https://recasgateway.ba.infn.it/>



Msa PAD

Execute our tool

Help ← help

Upload file
Upload DNA sequence file in FASTA or .ZIP format:

Upload File ← Input file upload

Upload status

Current state: Idle

File name: _____

Status: _____

Reset list files

Select an alignment Mode

DNA sequence path file: _____

Alignment mode: ← Select one genetic code

Genetic code: * ← Select one or more reading frame/s

Reading Frame: *

Mail recipient: _____

Execute ← Insert your email address

Check status of your executions

Mail recipient: _____

Alignment mode: *

Genome

Show

Results: _____

Msa PAD application is a multiple DNA sequence alignment framework designed to align conserved protein coding DNA sequences. The application accounts for either single or multiple protein domains coding sequences and uses this information in assembling its output. It is mainly useful for comparative genomics by the possibility to align genomes having different genic organization (i.e. bacterial genomes). It takes also into account and/or aligning genes' exons, including those undergone intron loss or gain, respecting genomic organization.BR/>

Msa-PAD has two different alignment modes: (i) genome and (ii) gene. The difference between the two modes resides in the organization of the final alignment.

Genome Mode alignment, similar to super-gene alignment, keeps only fragment of sequences coding for unique protein domains and the final output is simply the concatenation of those fragments by randomly organizing the domains.

Gene Mode alignment respects the genomic organization of the input sequences; it identifies the most frequent domains order pattern. Consequently, domains order follows the increasing sites position of input DNA sequences.

Both modes share the following six steps:

1. It translates DNA sequences using a user-defined genetic code and frame/s by executing a custom Python script.
2. It makes use of PFAM-A [1] profiles information to assign translated sequences to a known conserved protein domain. This is obtained by searching against PFAM-A database using hmmsearch (HMMer3.0 package) [2]
3. It elaborates protein sequence assignment taking into consideration frameshifts and intron gain or loss.
4. It groups sequences belonging to the same protein domain and consequently align them against the same domain using hmalign (HMMer3.0 package)
5. It Back-translates the protein alignments into DNA alignments using a custom Python script.
6. It merges the back-translated alignments to output the final DNA multiple sequence alignment.

The domains order included in the final alignment depends on the alignment mode chosen by the user and it is reported in a separate output called AlignmentDomainsPartitions.txt

It is important to note that Genome Mode alignment is not reserved to align only genomes, Gene Mode can be also used for genomes in case the user wishes to maintain the same domain order as provided by the initial input.

MSA-PAD: Taverna Workflows



<http://www.myexperiment.org/workflows/4549.html>

Biodiversity Virtual e-Laboratory
 on myExperiment

Home | BioVeL | **Workflows** | Files | Packs

New Workflow - GO

Home > Workflows > MsaPAD: Multiple Sequence Alignment - Input Submission and email notification


Download Workflow | Open in OnlineHPC

Workflow Entry: MsaPAD: Multiple Sequence Alignment - Input Submission and email notification
 Created at: 30/07/14 @ 15:02:56 | Last updated: 30/07/14 @ 15:07:25
 License | Credits (4) | Contributions (1) | Tags (4) | Featured in Packs (0) | Ratings (0) | Attributed By (0) | Favoured By (0) | Citations (0) | Version History | Reviews (0) | Comments (0)

Download workflow
 Version 1 (of 1)
 Version created on: 30/07/14 @ 15:02:53 by: Bachirb
 Last edited on: 30/07/14 @ 15:07:23 by: Bachirb

Title: MsaPAD: Multiple Sequence Alignment - Input Submission and email notification
Type: Taverna 2

Original Uploader: Bachirb

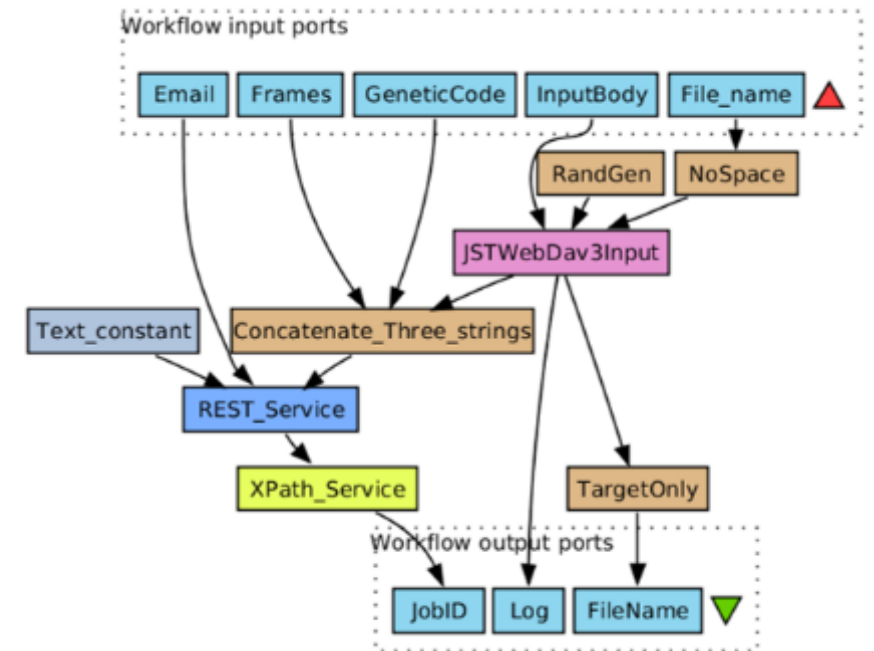
License: All versions of this Workflow are licensed under: 

Credits (4) (People/Groups)
 Bachirb
 Giacinto Donvito
 Saverio Vicario
 Pasquale Notarangelo

Attributions (1) (Workflows/Files)

Workflow Inputs: Email, Frames, GeneticCode, InputBody, File_name
Workflow Outputs: JobID, Log, FileName

MsaPAD workflow in Taverna Workbench Biodiversity 2.5

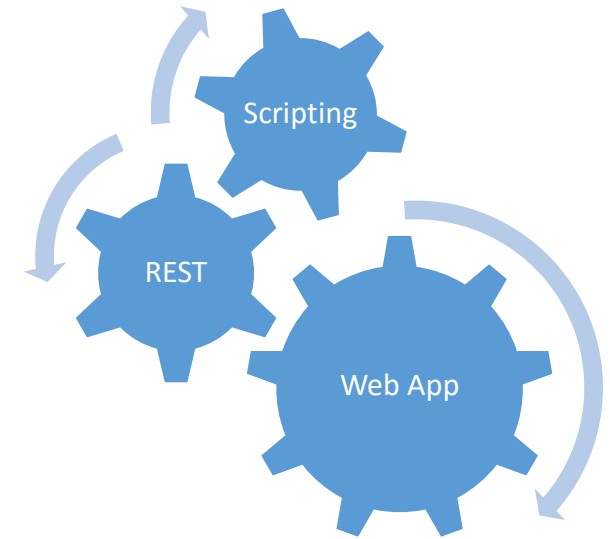


myExperiment addresses:

- ❖ **GeneMode:** <http://www.myexperiment.org/workflows/4549.html>
- ❖ **GenomeMode:** <http://www.myexperiment.org/workflows/4551.html>

MSA-PAD: Next Release

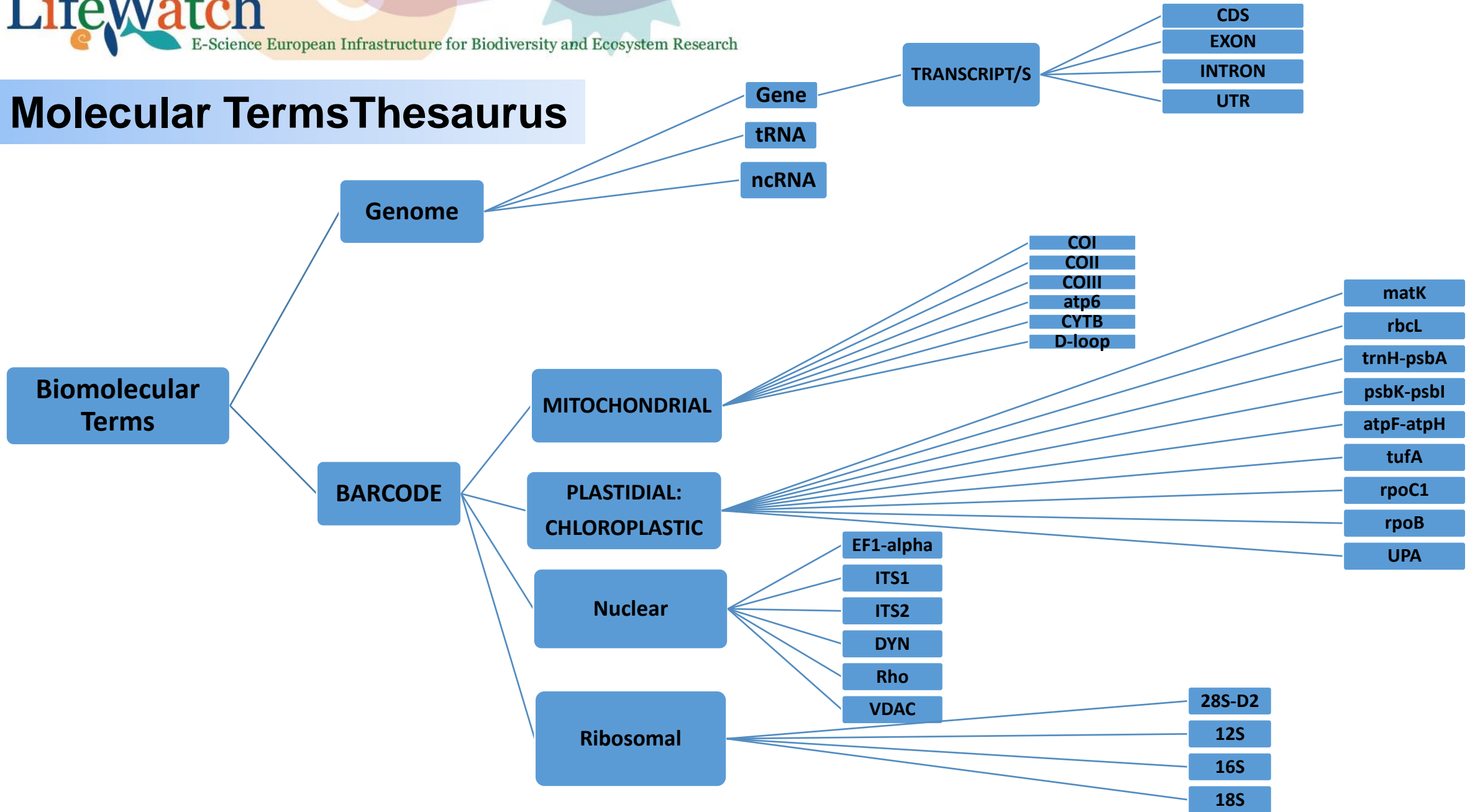
In-progress



Additional options:

- Possibility to **upload** a private user profile domain and add it to PFAM database
- Possibility to **run** the alignment on a **pre-selected** PFAM/private profile domain

Molecular Terms Thesaurus

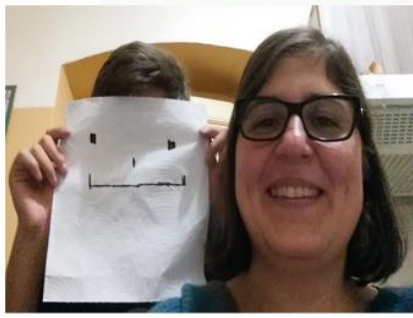
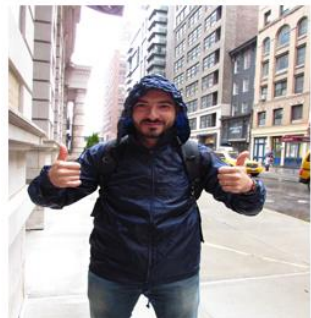




Staff



Personnel & contact details



Collaborations

