Retrieval & Alignment Tools

Sequence and Metadata Retrieval

- **DataRetrieval**

Multiple Alignments

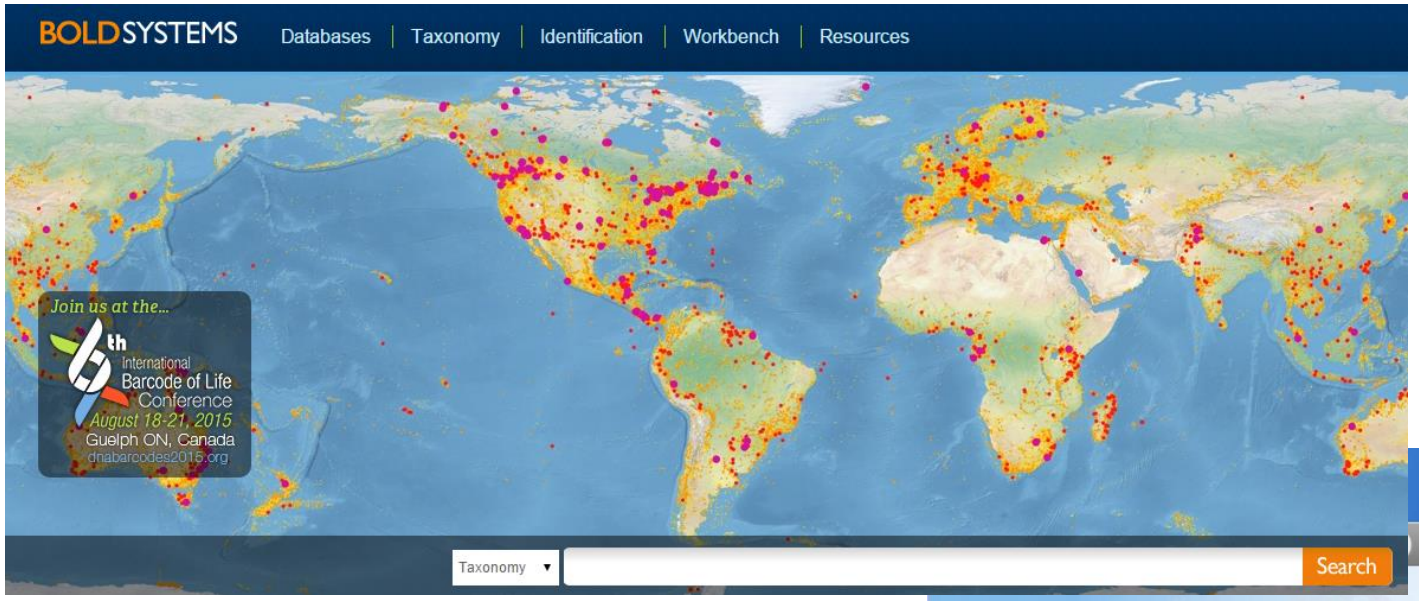
- **MSA-PAD**

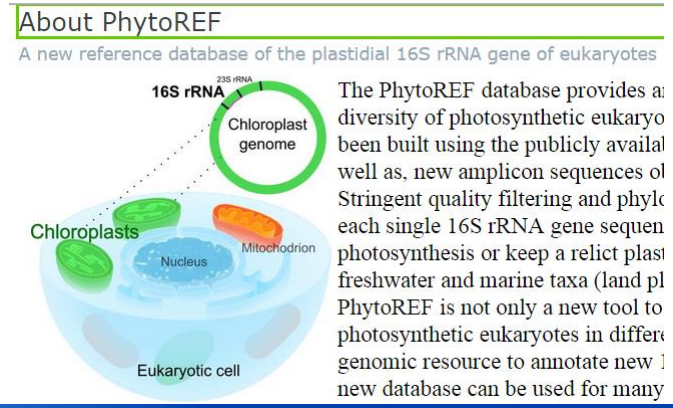
LifeWatch
E-Science European Infrastructure for Biodiversity and Ecosystem Research

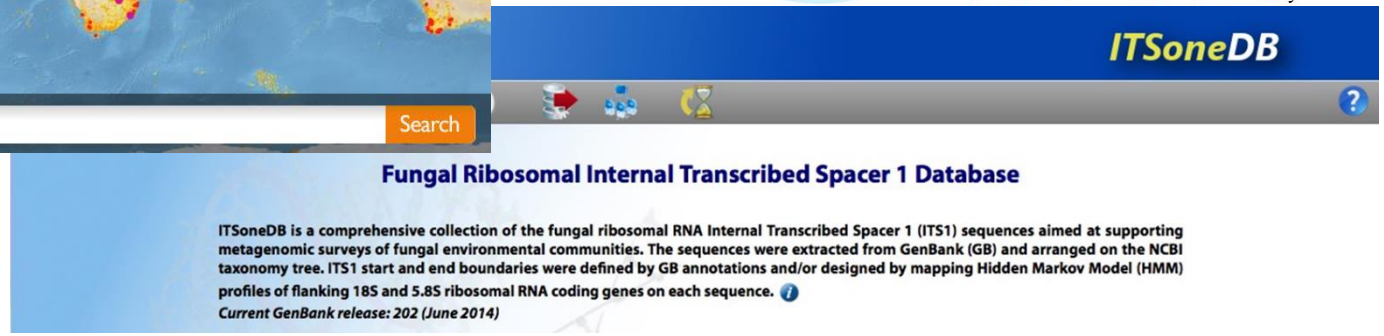
DataRetrieval: DNA Barcode Sequence and Metadata Retrieval

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







NCBI
National Center for Biotechnology Information

DataRetrieval: Arguments and Execution

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

LifeWatch
E-Science European Infrastructure for Biodiversity and Ecosystem Research

DataRetrieval Outputs

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
- Lattitude
- Longitude
- Genbank accession
- Nucleotides
- sequencing primers

DataRetrieval
in-progress

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**

Multiple Sequence Alignments

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

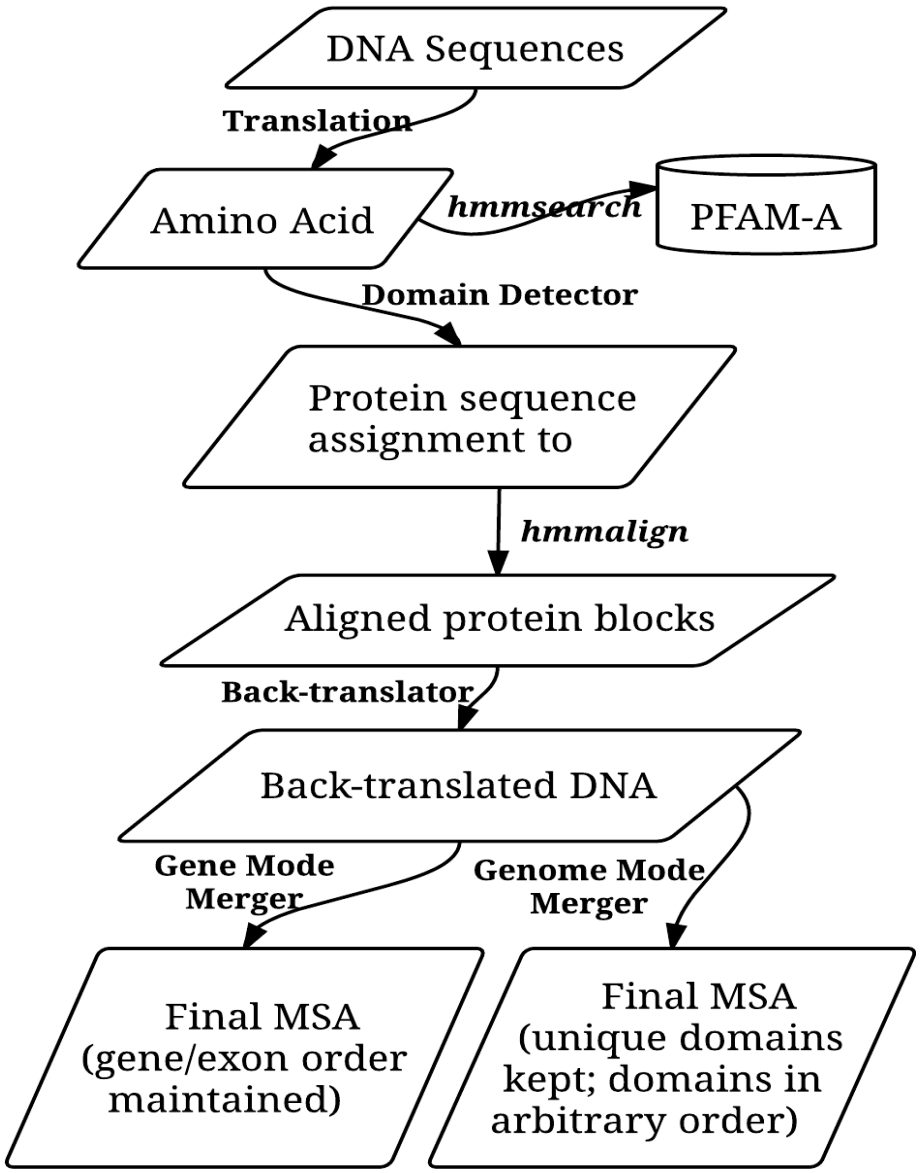
Bachir Balech[1], Saverio Vicario[2], Giacinto Donvito[3], Alfonso Monaco[3], Pasquale Notarangelo[3] and Graziano Pesole[1,4,*]

- 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)

LifeWatch
E-Science European Infrastructure for Biodiversity and Ecosystem Research

Lower layer implementation

MSA-PAD concept



REST call of JST webservice: Input upload and execution

Wrappers

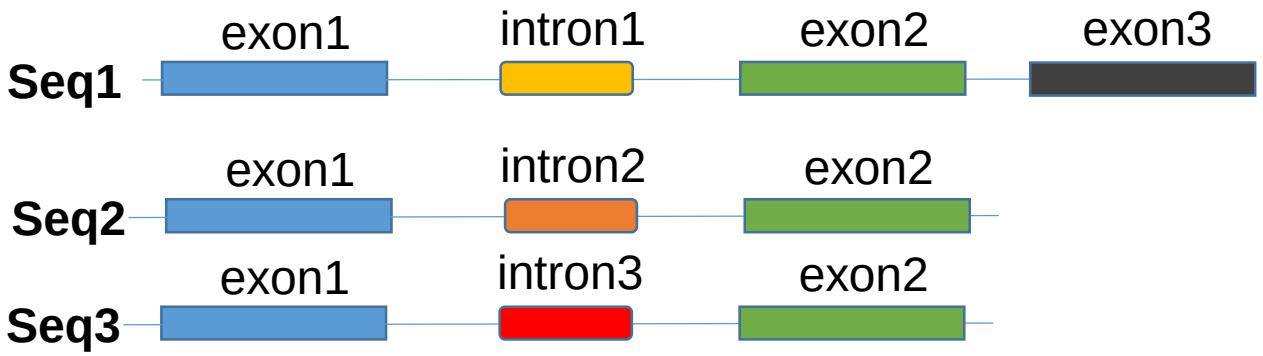
HMMer3.0:
hmmsearch
hmmalign

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

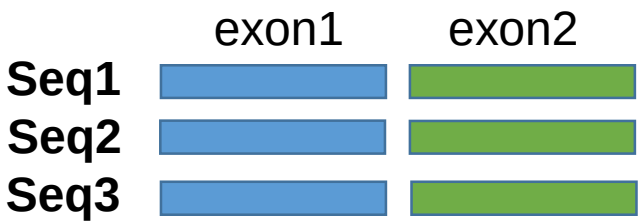
Email Client Answer: Output retrieval

MSA-PAD @ work – Use Cases

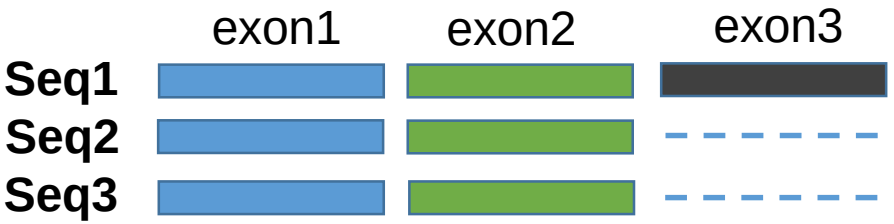
Intron occurence



Gene Mode Alignment

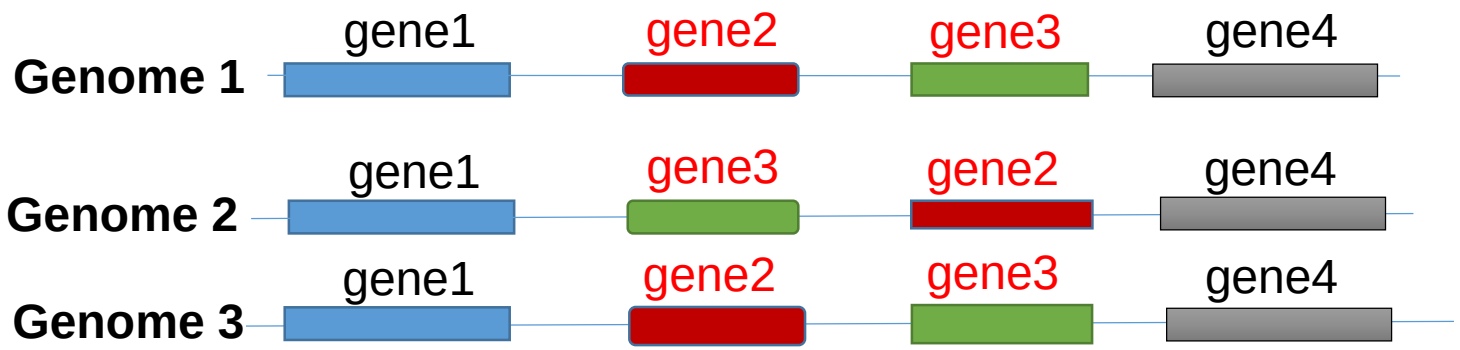


Genome Mode Alignment

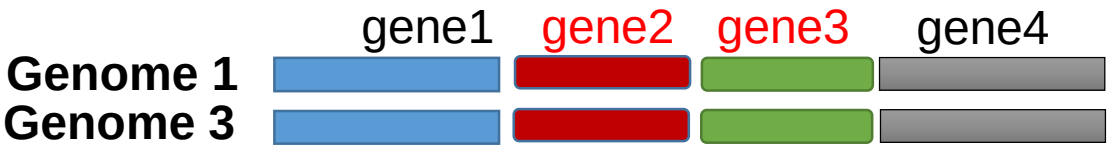


MSA-PAD @ work – Use Cases

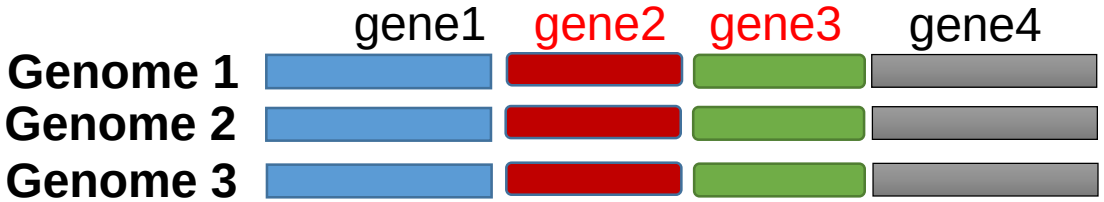
Genomes Rearragements



Gene Mode Alignment

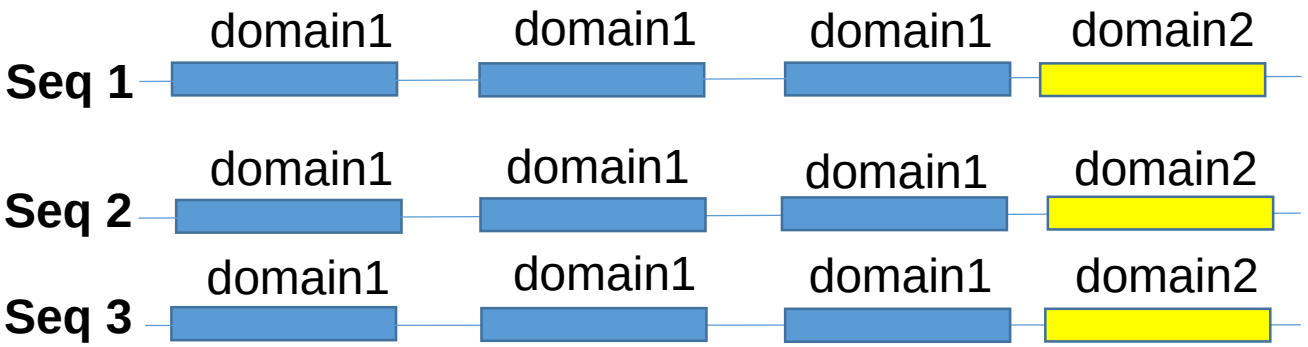


Genome Mode Alignment

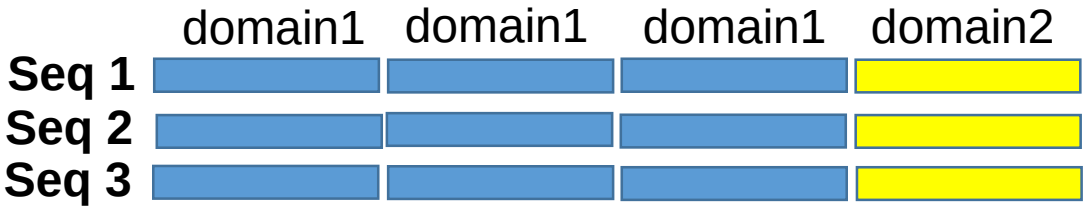


MSA-PAD @ work – Use Cases

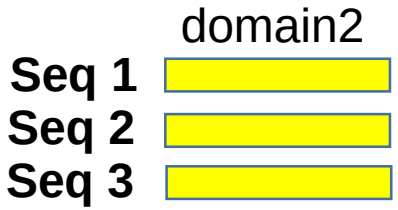
Repeated Domains



Gene Mode Alignment



Genome Mode Alignment



LifeWatch
E-Science European Infrastructure for Biodiversity and Ecosystem Research

MSA-PAD Outputs

Main outputs:

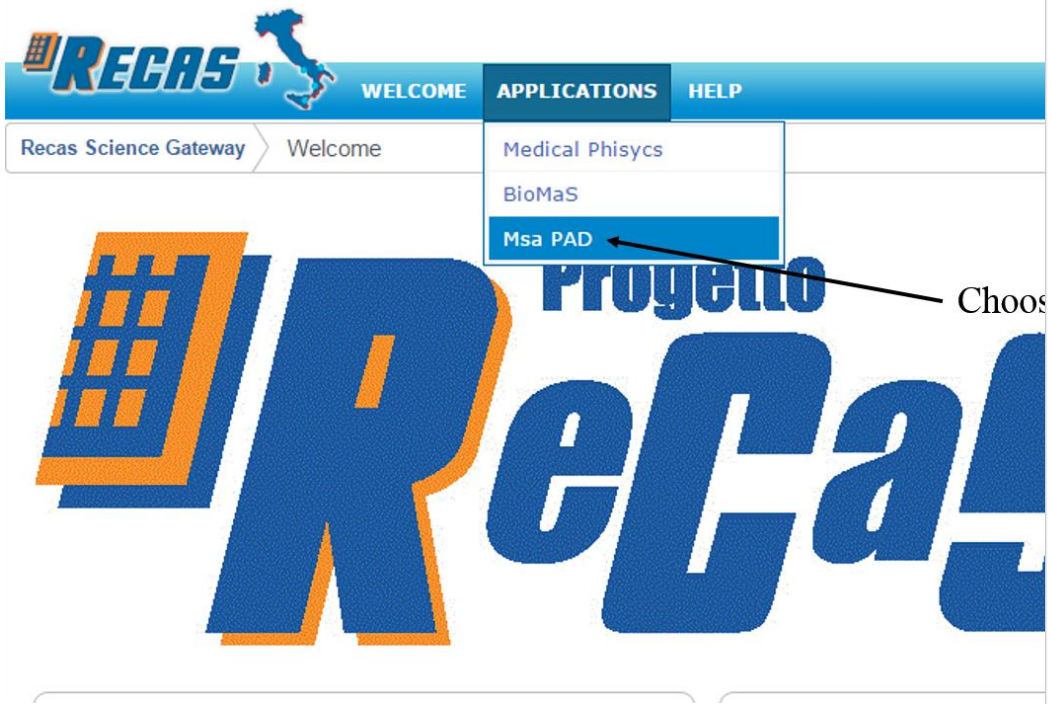
- **Final multiple DNA alignment** -> FASTA format
- **AlignmentDomainsPartitions** -> the coordinates of each protein domain in the final MSA
- **ExcludedSequencesIDs** -> sequence IDs (separated by comma) not present in the final MSA

Additional outputs:

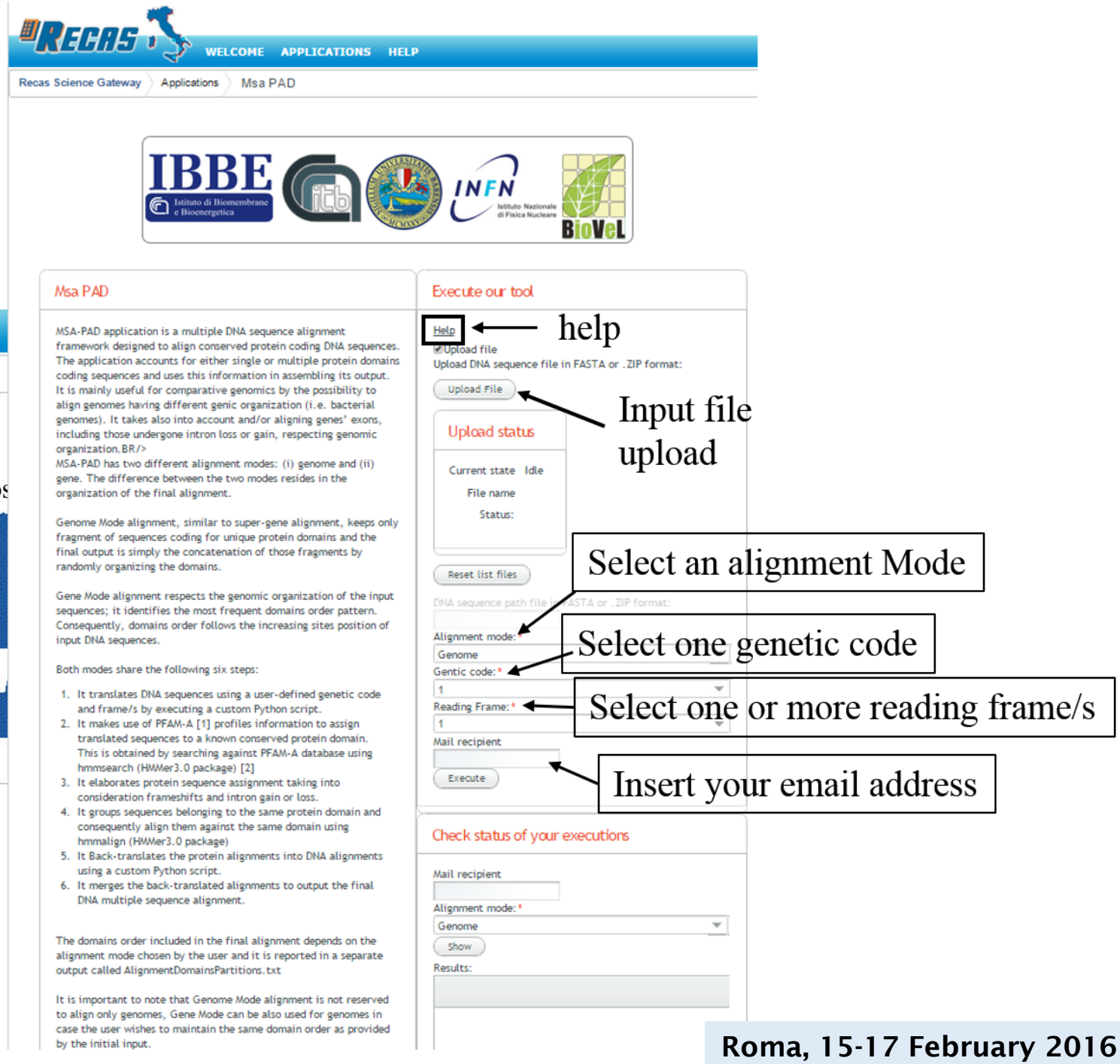
- ✓ **File/s with *hmmAligned* suffices ->** alignments (STOCKHOLM format) of each protein sequences block with PFAM profile as prefix
- ✓ **File/s with *Backaligned.fasta* suffices ->** alignments of each back-translated DNA sequences block
- ✓ **MissingSites_Report** -> DNA sites position missing from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

LifeWatch
E-Science European Infrastructure for Biodiversity and Ecosystem Research

MSA-PAD: Web Application

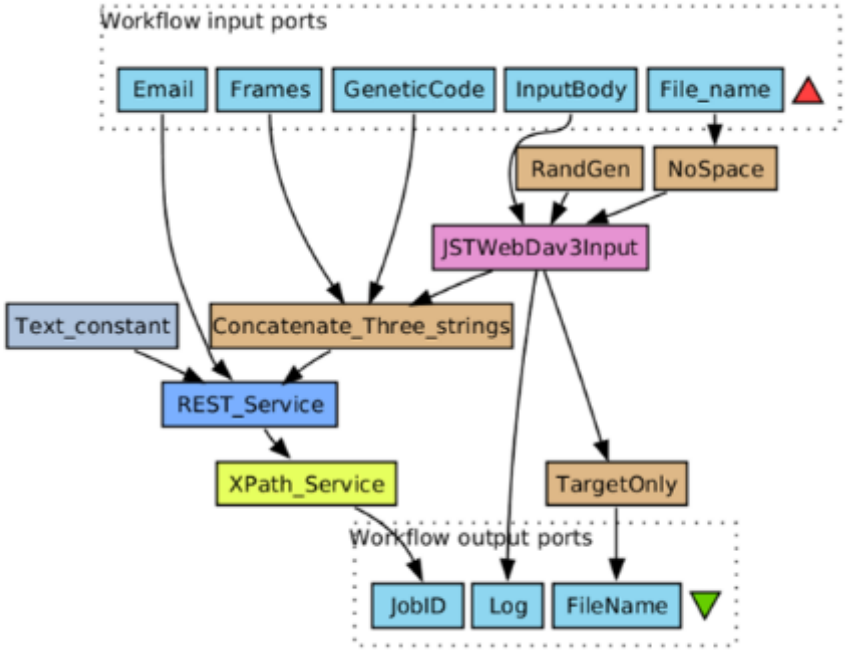


https://recasgateway.ba.infn.it/



MSA-PAD: Taverna Workflows

MsaPAD workflow in Taverna Workbench Biodiversity 2.5



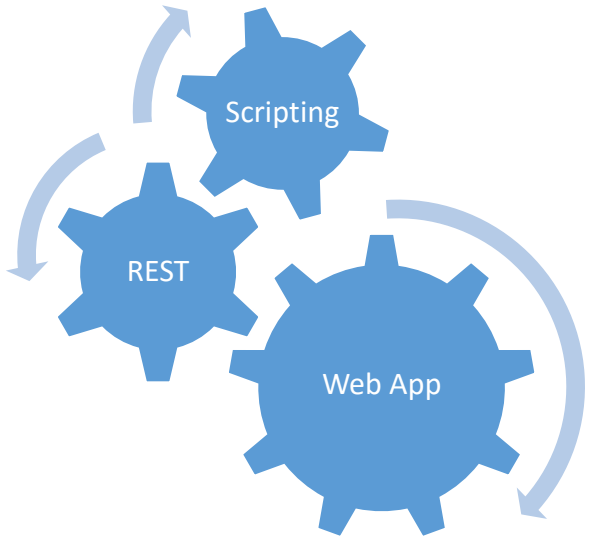
myExperiment addresses:

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

LifeWatch
E-Science European Infrastructure for Biodiversity and Ecosystem Research

MSA-PAD: Next Release

In-progress

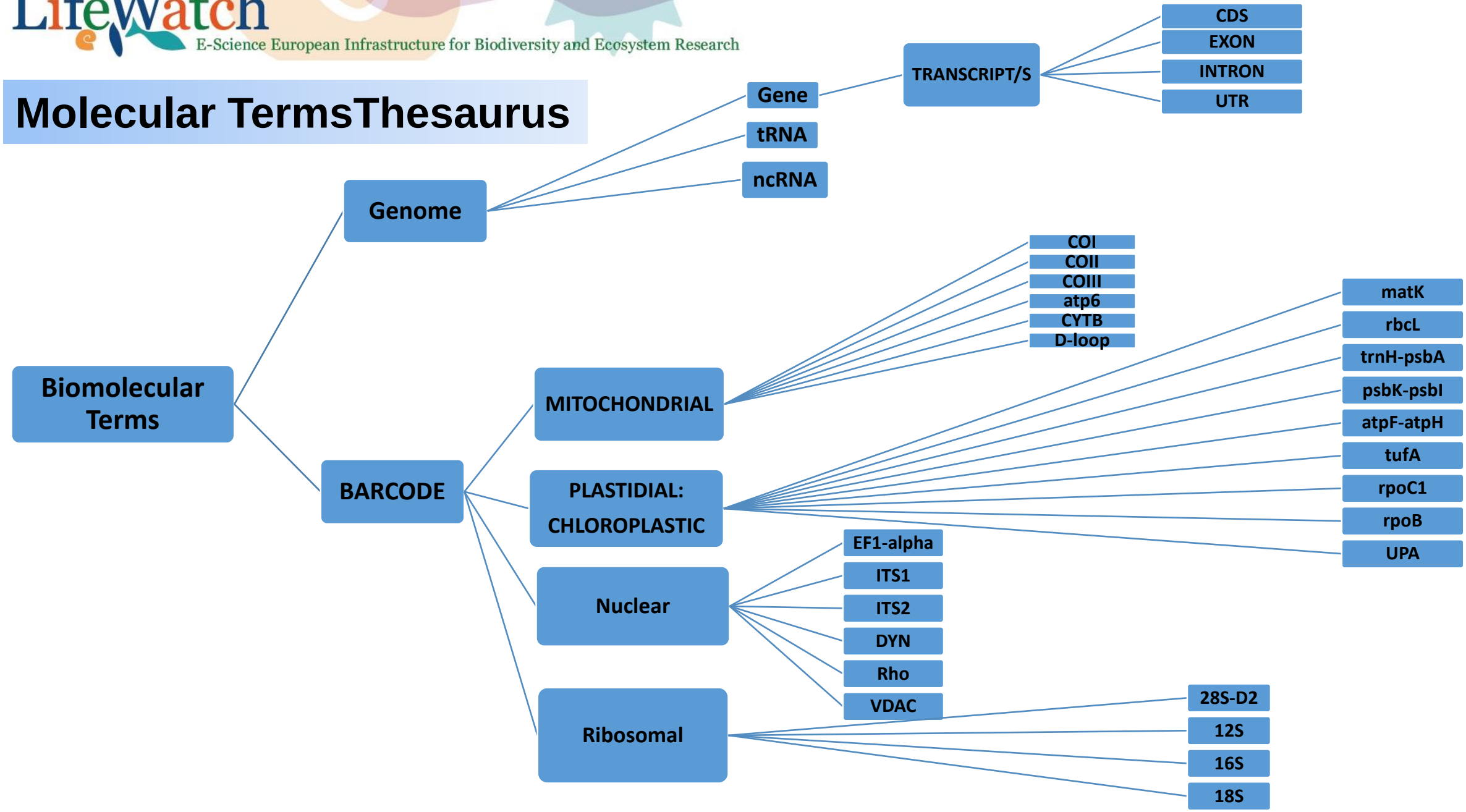


Additional options:

- 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

LifeWatch
E-Science European Infrastructure for Biodiversity and Ecosystem Research

Molecular TermsThesaurus



Staff

Personnel & contact details

Collaborations

INFN ISTITUTO NAZIONALE DI FISICA NUCLEARE

Consiglio Nazionale delle Ricerche

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

LifeWatch