



# **MSA-PAD: Novel DNA Multiple Sequence Alignment Guided by PFAM Conserved Domains**

**Bachir Balech**

**Institute of Biomembranes and Bioenergetics  
(IBBE-CNR)**

**NETTAB 2015  
Bari October 14<sup>th</sup> 2015**

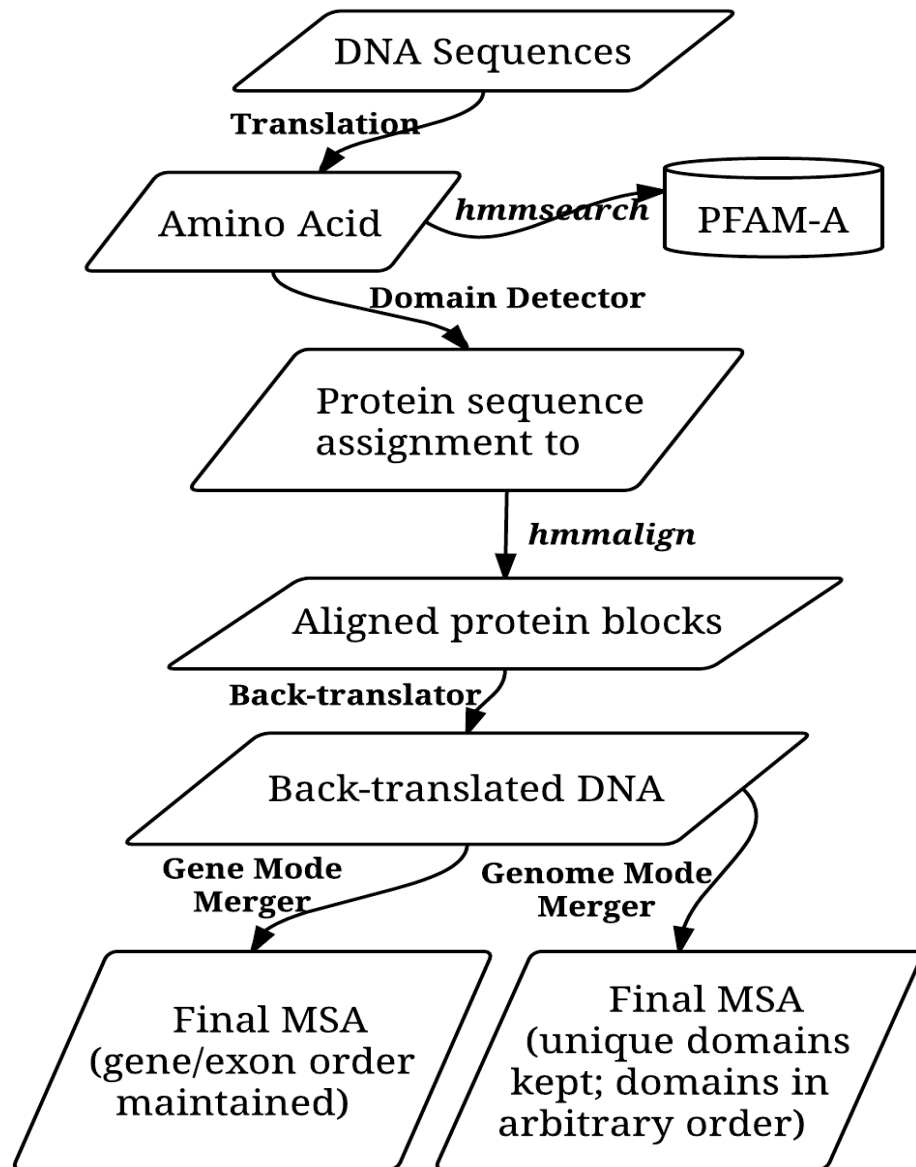
- MSA is a fundamental step in sequence analysis applications such as phylogenetic inference and comparative genomics
- Higher MSA precision is often accomplished with protein alignments
- Protein alignments can be exploited to guide a DNA ones (e.g. translatorX, tranalign)

## Shortcomings of available methods

- ✓ They do not make use of information embedded in protein domains, intron occurrence and gene order variations (e.g. mitochondrial genomes)
- ✓ Input protein alignments should be provided

## MSA-PAD: DNA Multiple Sequence Alignment Framework based on PFAM Accessed Domain Information

How it accounts for those shortcomings?



**REST call of JST webservice:** Input upload and execution

## Wrappers

### HMMer3.0:

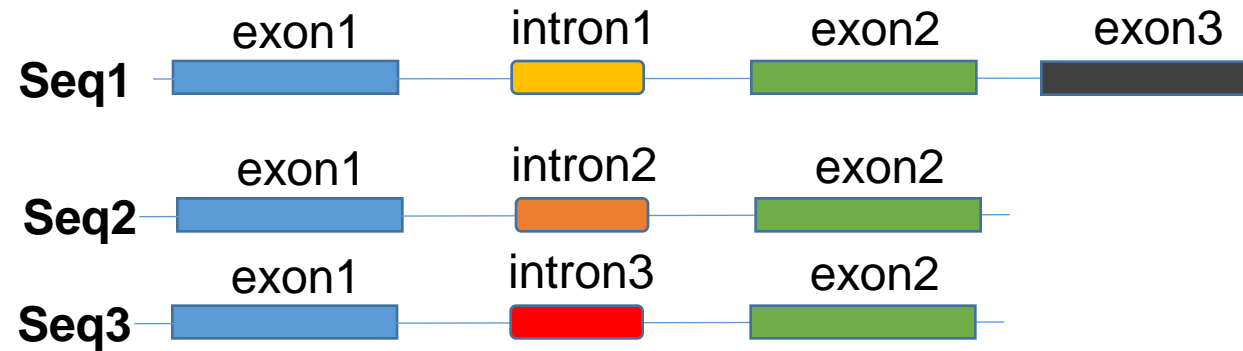
*hmsearch*  
*hmmalign*

### Python parsers:

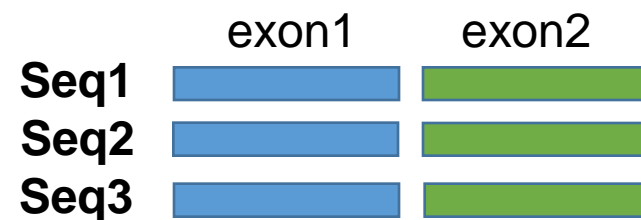
*Translator.py*  
*Backaligner.py*  
*Merger.py*

**Email Client Answer:** Output retrieval

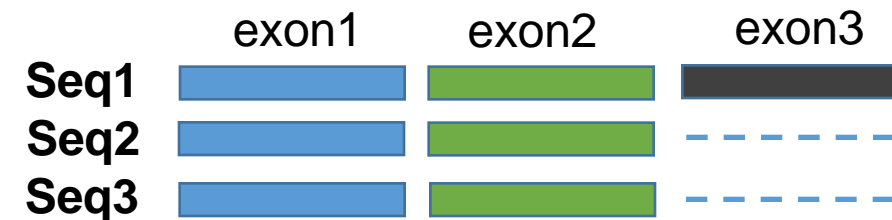
## Intron occurrence



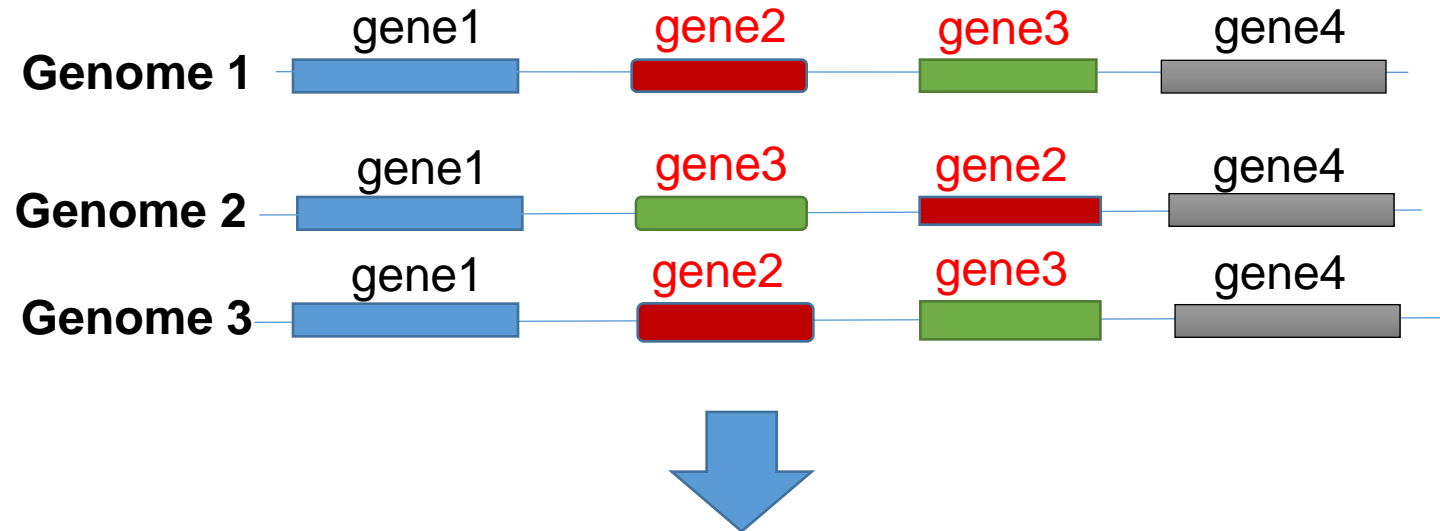
## Gene Mode Alignment



## Genome Mode Alignment

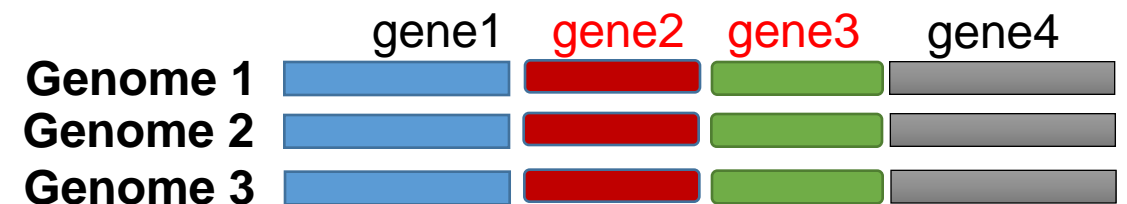
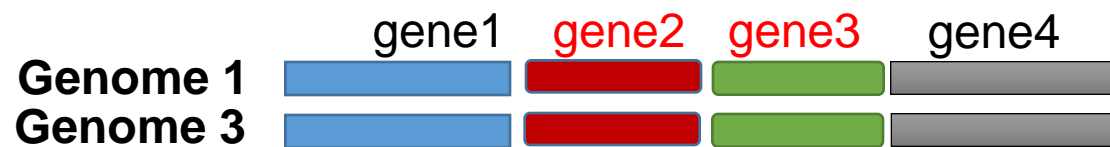


## Genomes Rearrangements

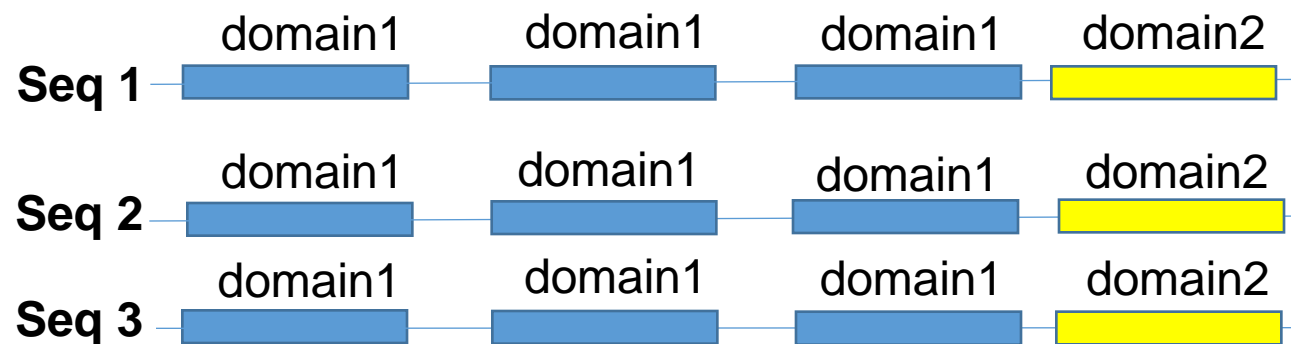


## Gene Mode Alignment

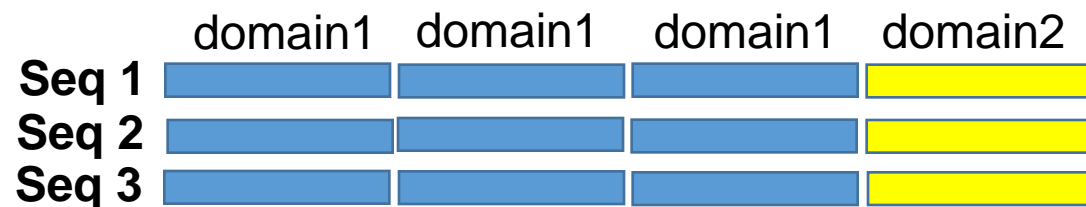
## Genome Mode Alignment



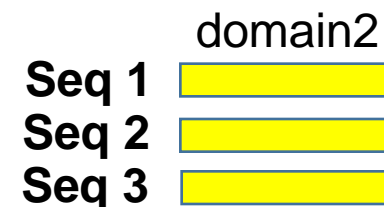
## Repeated Domains



### Gene Mode Alignment



### Genome Mode Alignment



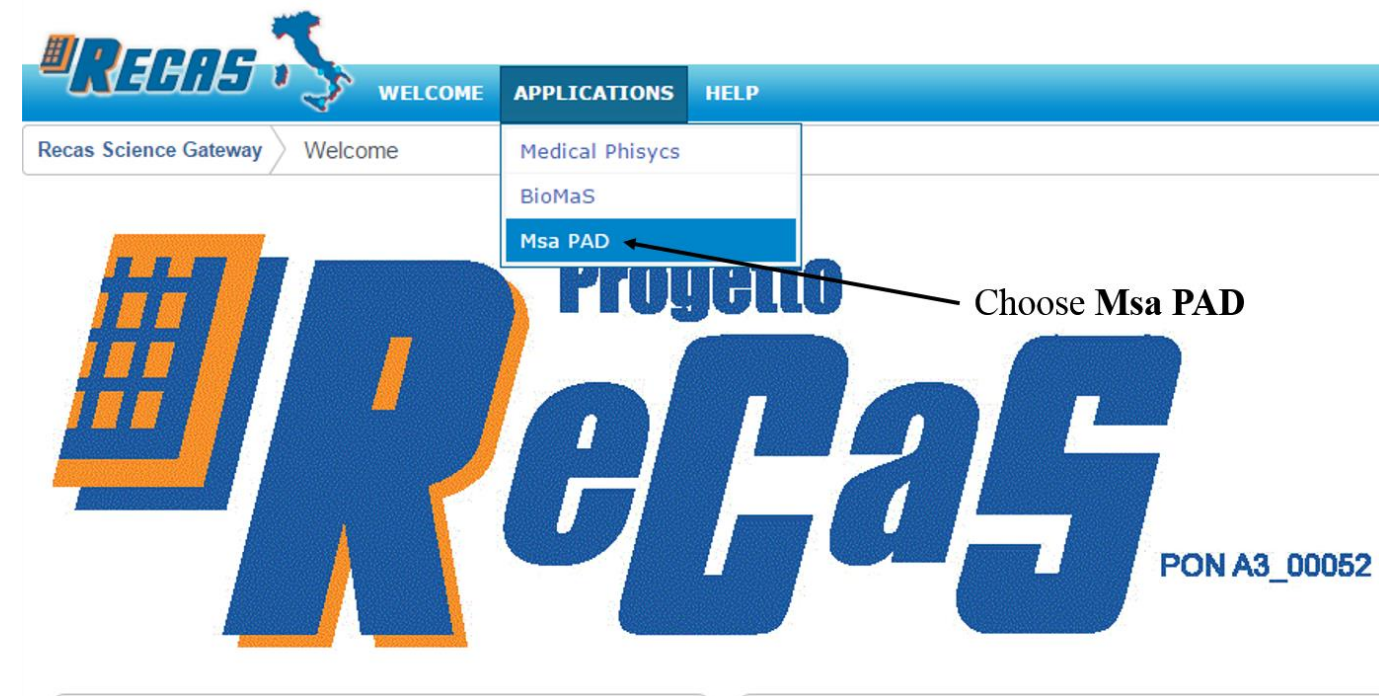
## Main outputs:

- **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



The screenshot shows the Recas Science Gateway website. The navigation bar includes 'WELCOME', 'APPLICATIONS', and 'HELP'. The 'APPLICATIONS' menu is open, showing 'Medical Phisycs', 'BioMaS', and 'Msa PAD'. An arrow points to 'Msa PAD' with the text 'Choose Msa PAD'. The main logo features a grid icon, the word 'Progetto' above 'Recas', and 'PON A3\_00052' below it.

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



# MSA-PAD: Web Application

Choose



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



## Msa PAD

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 hmmersearch (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.

## 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:

Select an alignment Mode

Reset list files

DNA sequence path file in FASTA or .ZIP format:

Alignment mode:

Genome

Select one genetic code

Genetic code: \*

1

Select one or more reading frame/s

Reading Frame: \*

1

Mail recipient

Execute

Insert your email address

## Check status of your executions

Mail recipient

Alignment mode: \*

Genome

Show

Results:

Workflow Entry: MsaPAD: Multiple Sequence Alignment - Input Submission and email notification

Download workflow

Version 1 (of 1)

Workflow Type: Taverna 2

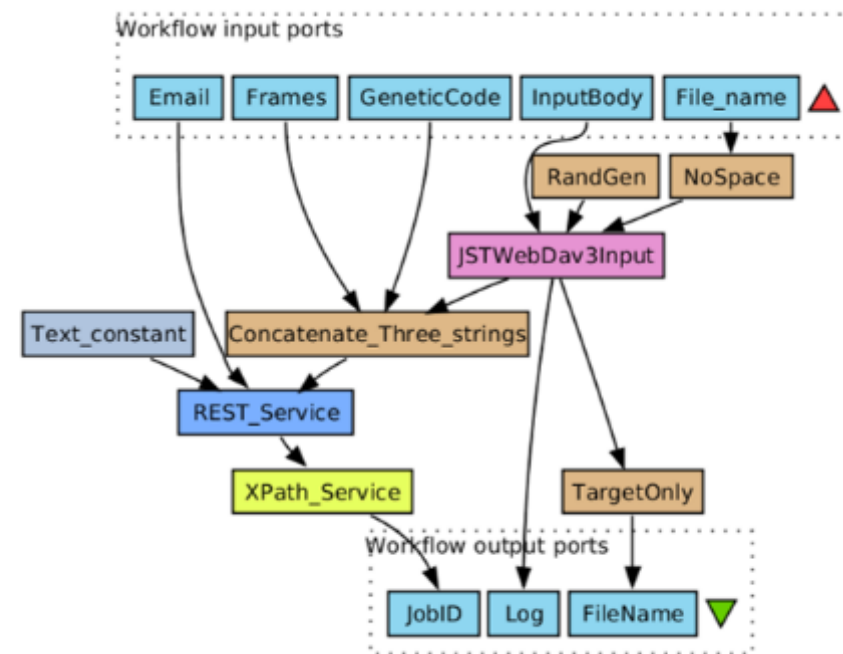
Original Uploader: Bachrib

License: All versions of this Workflow are licensed under: CC BY-NC-SA

Credits (4) (People/Groups): Bachrib, Giacinto Donvito, Saverio Vicario, Pasquale Notarangelo

Attributions (1) (Workflows/Files)

## 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>

---

Sequence analysis

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

Bachir Balech<sup>1</sup>, Saverio Vicario<sup>2</sup>, Giacinto Donvito<sup>3</sup>, Alfonso Monaco<sup>3</sup>,  
Pasquale Notarangelo<sup>3</sup> and Graziano Pesole<sup>1,4,\*</sup>

### **Additional options in the coming release:**

- 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

**For more information and bug report please write us at**

**[balechbachir@gmail.com](mailto:balechbachir@gmail.com) and/or [g.pesole@ibbe.cnr.it](mailto:g.pesole@ibbe.cnr.it)**

## *Acknowledgments*

### ***IBBE – CNR***

- Prof. Graziano Pesole

### ***ITB – CNR***

- Saverio Vicario

### ***INFN***

- Giacinto Donvito
- Alfonso Monaco
- Pasquale Notarangelo