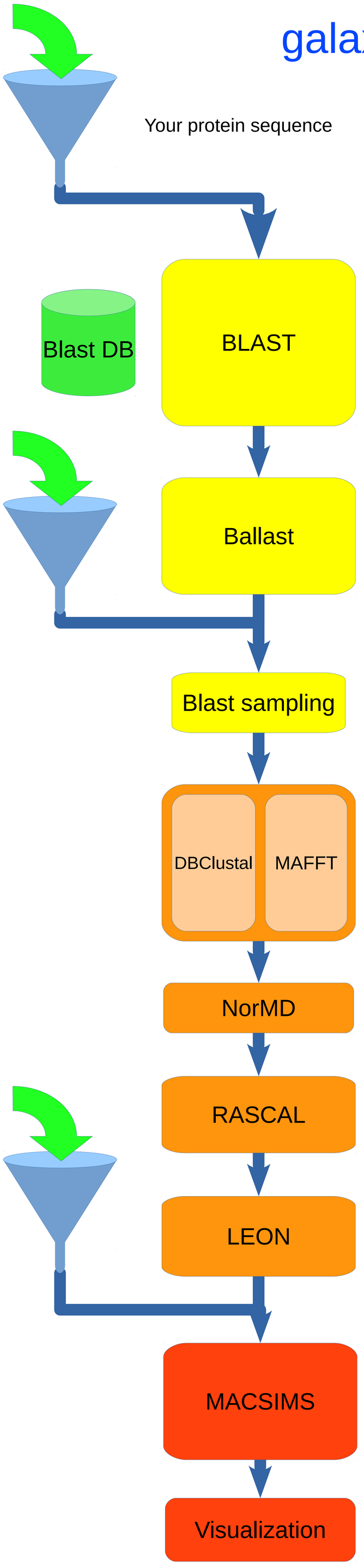
PipeAlign in Galaxy

Workflows for comparative protein analysis

Arnaud KRESS, Luc THOMES, Raymond RIPP, Julie THOMPSON
LBGI, Faculté de médecine, Strasbourg

galaxy.bioinfo-bistro.fr galaxy.lbgi.fr



PipeAlign is a tool for comparative or evolutionary analysis of a protein family, starting from a protein sequence to visualization of results.

Homology search

First, perform a Blast search of the selected database using your sequence

BLAST post-processing

Ballast processes the results of a BLAST database search to identify local maximum segments (LMSs) that are conserved between the query sequences and the BLAST hits. Ballast also constructs a list of pairwise 'anchors' based on the LMSs, which can be used to guide a multiple sequence alignment.

Sequence selection

The blast sampling module selects the sequences with an e-value threshold, a maximum number of sequences, and also three differents filters :
- **Strips**, a selection by subdividing the log curve of e-values and selecting the best entries in each strip
- **Taxonomy**, selecting hits belonging to one of the specified taxon IDs. It uses the NCBI taxonomic tree references.
- **FDD** performs a second derivative of the log curve of e-values to select sequences at inflexion points

Multiple sequence alignment

- DbClustal takes a list of sequences as input, together with a list of anchors from the Ballast program, and constructs a MSA.
- MAFFT performs fast MSA using a fast Fourier transform

Evaluation of alignment quality

NorMD is an objective scoring function for multiple sequence alignments.

Alignment rapid scanning and correction

RASCAL takes a multiple sequence alignment as input, scans the alignment for local errors and tries to correct them by realigning segments of the sequences.

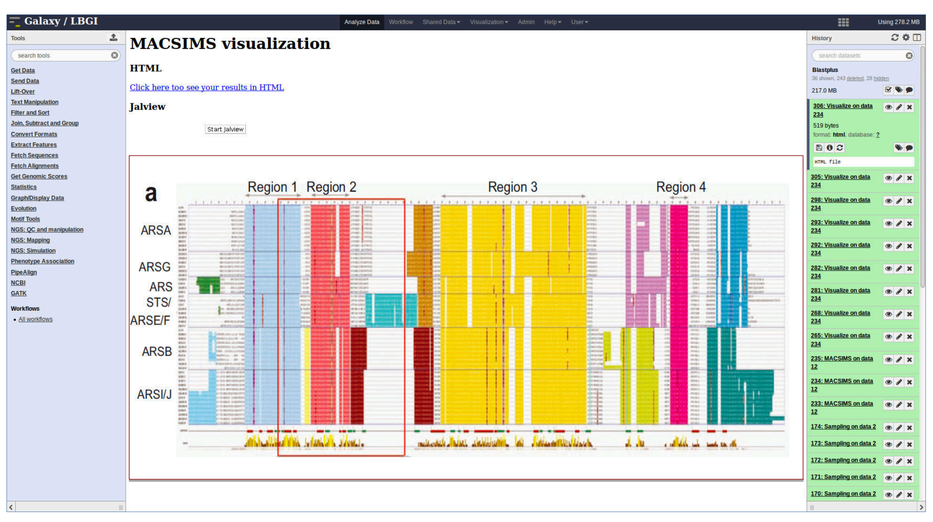
Removal of unrelated sequences

LEON takes a multiple sequence alignment as input and identifies sequence segments that are homologous to the query sequence specified by the user. Sequences with no homologous regions are removed from the alignment.

Structural/functional annotation

MACSIMS annotates a MSA with structural and/or functional information. Information is retrieved from public databases (Uniprot, Interpro, PDB, …), validated, propagated from known to unknown sequences, and then stored in an XML format file.

Jalview or HTML



All the tools have been packaged in Galaxy

Galaxy packages are available at
http://toolshed.galaxeast.fr

References :

- Plewniak F., et al. (2003) PipeAlign: a new toolkit for protein family analysis. Nucleic Acids Research, **31**, 3829–3832.
- Plewniak,F., Thompson,J.D. and Poch,O. (2000) Ballast: blast post-processing based on locally conserved segments. *Bioinformatics*, **9**, 750–759.
- Thompson,J.D., Plewniak,F., Thierry,J. and Poch,O. (2000) DbClustal: rapid and reliable global multiple alignments of protein sequences detected by database searches. *Nucleic Acids Res.*, **15**, 2919–2926.
- Thompson,J.D., Plewniak,F., Thierry,J. and Poch,O. (2003) RASCAL: Rapid scanning and correction of multiple sequence alignment programs.*Bioinformatics*, **19**, 1155-61.
- Thompson,J.D., Plewniak,F., Ripp,R., Thierry,J.C. and Poch,O. (2001) Towards a reliable objective function for multiple sequence alignments. *J. Mol. Biol.*, **4**, 937–951.
- Thompson, J.D., Prigent, V., Poch, O. (2004) LEON: multiple aLignment Evaluation Of Neighbours. *Nucleic Acids Res.* **32**, 1298-307.
- Thompson JD1, Muller A, Waterhouse A, Procter J, Barton GJ, Plewniak F, Poch O. (2006) MACSIMS: multiple alignment of complete sequences information management system. *BMC Bioinformatics, **7**, 318.*
- Waterhouse, A.M., Procter, J.B., Martin, D.M.A, Clamp, M. and Barton, G. J. (2009) Jalview Version 2 - a multiple sequence alignment editor and analysis workbench. *Bioinformatics, **25,*** 1189-1191.