OSCARS

Open Science Clusters' Action for Research & Society

Funded Project

FAIRFUN4Biodiversity

Rosa Fernández, Institute of Evolutionary Biology (CSIC-UPF), 0000-0002-4719-6640
Ana Rojas, Andalusian Center for Developmental Biology (CSIC-UPO), 0000-0003-0750-9099
Aureliano Bombarely, Institute of Plant Molecular and Cellular Biology (CSIC),0000-0001-6257-8914

Implemented by

IBE CSIC upf.

Centro Andaluz de Biología del Desarrollo

ibmcp

CSIC
Consejo Superior de Investigaciones Científicas

Funded by the European Union

What problem(s) are you going to solve?

- Functional annotation of coding genes is currently done with homology-based method (i.e. sequence similarity).
- A large percentage of coding genes in nonmodel organisms do not have any functional annotation based on these methods (e.g., more than 50% in invertebrates)(this is what we call the 'dark proteome').
- We recently developed a pipeline, FANTASIA (Functional ANoTAtion based on embedding space SImilArity) to leverage Natural Language Processing (NLP) models. It annotates function in virtually all genes.
- We want to illuminate the function of the 'dark proteome' of thousands of genomes of nonmodel organisms, leveraging resources produced by the scientific community working in biodiversity genomics.

SOLUTION

What are you planning to do to solve the problem?

- We aim at:
  (1) Make FANTASIA FAIR and interoperable between science clusters.
  (2) Improve FANTASIA and add information on structure and more NLP models.
  (3) Apply FANTASIA to thousands of genomes of animals, plants, protists and fungi, to illuminate their ‘dark proteome’.
  (4) Train the biodiversity genomics community across Europe on this tool.

RESULTS

OSCARS

What will be the results and how do you plan to make them available to the broader community?

- Results:

(1) A FAIR tool, easy to use by everyone.

(2) Millions of functional annotations of the ‘dark proteome’ (one per gene, with thousands of genes in each genome annotated).

(3) Training material readily available for everyone.

- The tool and the functional annotations will be deposited in open access repositories, with the help of LifeWatch and ELIXIR.

RISKS

What risks could limit the success of the project, and how can they be mitigated

- The main risk is running into problems with the FAIRification of the tool. This can be mitigated with the help of collaborators and new team members with the proper expertise.
- A minor risk is having enough storage space for the millions of annotation that will be generated, and that we want to deposit in open access databases. The help of the appropriate people from the science clusters will be key to overcome this limitation - but we are confident there is an easy solution.

TEAM

Who is doing it? (OPTIONAL)

- Rosa Fernández, Institute of Evolutionary Biology (CSIC-UPF), 0000-0002-4719-6640
- Ana Rojas, Andalusian Center for Developmental Biology (CSIC-UPO), 0000-0003-0750-9099
- Aureliano Bombarely, Institute of Plant Molecular and Cellular Biology (CSIC),0000-0001-6257-8914
- Belén Carbonetto (postdoc, Rosa Fernández's team)
- Fran Pérez Canales (software engineer, Ana Roja's team)