



COAST: A bioinformatic tool to identify the closest proteomes

Diogo Macedo (1), Sílvio Santos (2), Óscar Dias (2)

(1) Departamento de Informática, Universidade do Minho

(2) Centro de Engenharia Biológica, Universidade do Minho

Recent advances in sequencing technologies, DNA manipulation and synthetic biology approaches, together with their potential applications in many fields of research, led to an increasing amount of DNA sequence and whole-genome data contributing to the accumulation of enormous raw data sets. Such massive data *per se* is meaningless and requires the development of efficient bioinformatics tools to analyse and mine data from it, constituting a major challenge, mainly if we want to look globally at whole genomes.

Identifying homologous genes and analysing their synteny are essential to understand the rules of genome structure and predict gene or protein structure and function. Besides the growing development of bioinformatics, tools able to perform comparative analysis between whole genomes is still scarce, and the existing ones rely on a manual, awkward and slow process. Moreover, there is a lack of tools that identify the closest genomes and/or proteomes from discrete sequence datasets, a gap that needs to be fulfilled.

We developed COAST, The Comparative Omics Alignment Search Tool, considering all this knowledge. COAST is a command-line tool that exploits existing alignment algorithms and databases to provide a simple search tool capable of identifying organism with similar proteomes or taxonomy closeness. The algorithm scores closeness using scores based on the Average Amino acid Identity (AAI), a well-recognised and used relatedness score.(1) These organisms will be required for further downstream comparative analysis, improving the annotation of new genomes by identifying and characterising the functions for newly found putative genes based on the homologs, leading to a deeper understanding of one organism in the context of another and giving insights into the structure and function of genes and genomes aiding on the taxonomic and phylogenetic studies of new organisms. The tool is available for the Galaxy platform (2,3), through a simple webapp that can be integrated into larger Galaxy workflows.

References:

1. Konstantinidis KT, Tiedje JM. Towards a Genome-Based Taxonomy for Prokaryotes. *J Bacteriol.* 2005 Sep 15;187(18):6258–64.
2. Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Čech M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research.* 2018 Jul 2;46(W1):W537–44.
3. Blankenberg D, Von Kuster G, Bouvier E, Baker D, Afgan E, Stoler N, et al. Dissemination of scientific software with Galaxy ToolShed. *Genome Biology.* 2014 Feb 20;15(2):403.

