View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Repositorio Institucional de la Universidad de Córdoba

P. 29

DEVELOPMENT OF A WORKFLOW FOR GENERAL PROTEIN SEQUENCE ANALYSIS BASED ON THE TAVERNA WORKBENCH® SOFTWARE

Mariana B. Monteiro [1,2], Manuela E. Pintado[1], Ian Shadforth [2], F. Xavier Malcata[1] and Patrícia R. Moreira[1*]

[1]Escola Superior de Biotecnologia, Universidade Católica Portuguesa,
R. Dr. António Bernardino de Almeida, Porto, P-4200-072 Porto, Portugal;
[2]Cranfield Health, Cranfield University, Silsoe, Bedfordshire MK45 4DT, England, U.K.;
*Corresponding author. Fax (+351) 225 090 351; e-mail: prmoreira@mail.esb.ucp.pt

The aim of this research effort was to build up a workflow able to perform a generic analysis of an unknown protein sequence. Recall that workflows permit processing of large amounts of data, which can efficiently flow throughout different and complex tasks in a time-saving, user-friendly way.

The implementation of said workflow was based on Taverna Workbench® software (http://taverna.sourceforge.net/). Said software provides computational resources (i.e. web services) to develop distinct workflow steps, taking advantage of a user-friendly interface while providing shim services to develop own scripts (if required).

The workflow developed was tentatively named Workflow for Protein Sequence Analysis (WPSA), and included: an initial homology search; a multiple sequence alignment; and construction of phylogenetic trees. WPSA accepts three types of input, and retrieves several outputs. The inputs that the user needs to provide include: a query protein sequence; a list of known protein identification numbers; and a choice of method to build the tree. The outputs generated include: a BLAST report; a description of different protein sequences; an image of the multiple sequence alignment; two different output files from the clustering method used; two types of trees; and conditional outputs, according to the query sequence entered. For each type of analysis, distinct web services from as many alternative sources were used. The workflow designed gives, in particular, fast runs (i.e. 5 to 10 min) and informational and significant responses on the sequence entered.

Although the workflow implements all required tasks in an acceptable fashion, several improvements aiming at a better performance were identified for posterior development.