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] College of Biotechnology, Portuguese Catholic University, Rua Dr. António Bernardino de Almeida, P-4200-072 Porto, Portugal.
[2] Cranfield Health, Cranfield University, Silsoe, Bedfordshire MK45 4DT, England, U.K.

* prmoreira@mail.esb.ucp.pt

Introduction

With the increasing volume of biological data available growing in a daily basis, researchers working with nucleotide or amino acid sequences need computational methods for data organization and analysis. Frequently the researcher needs to perform successively steps using bioinformatics applications and tools available in the internet. In order to speed up the process a **workflow** for general protein sequence analysis was developed, which could integrate the available applications, make the data flow throughout them without almost any user intervention and present the result in a user-friendly way.

Materials and Methods

The aim of this research effort was to build up a workflow able to perform a generic analysis of an unknown protein sequence. A variety of processors from **Taverna** (Onim et al., 2005) were used to build the workflow since each processor has to be linked to each other according to the input and output requirements. As a result of the input specifications some of the processors could not be directly linked, and thus linking processors had to be created with Beanshell Scripting, written with Java-syntax.

For the development of **Workflow for Protein Sequence Analysis (WPSA)** the different analysis were built separately and finally joined together in a single workflow.

Results and Discussion

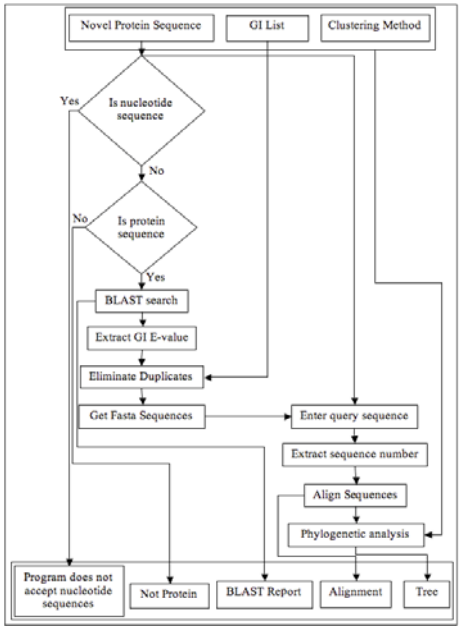

Figure 1. Flowchart showing the control level of the WPSA workflow.

The workflow tentatively named Workflow for Protein Sequence Analysis (WPSA), 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.

Once the workflow has started in Taverna interface, all the steps of the workflow can be followed and the progress of the web services can be visualised.
After running WPSA with the proper inputs, query protein sequence, list of known protein identifiers and choice of the clustering method to build the tree (Neighbor Joining or UPGMA) seven different results are given in Taverna results window.
For the homology search analysis, a BLAST report is given. Multiple alignments are visualised as image with the aligned sequences and the consensus pattern.
The workflow designed gives, in particular, **fast runs (i.e. 5 to 10 min)** and informational and significant responses on the sequence entered. The resulting workflow, Workflow for Protein Sequence Analysis may be accessed over the web http://www.myexperiment.org/workflows/124;download?version=1.

Conclusions

The workflow designed gives a **fast and significant answer** to the user about the sequence entered. It takes about 5 to 10 minutes to run or more, depending on the Internet connection and also on the web services. WPSA **eliminates the need for the user to cut and paste its data throughout web applications**.
Although the workflow implements all required tasks in an acceptable fashion, **several improvements** aiming at a better performance were identified for posterior development, specially for the abovementioned problems.

References

Onim, T., Greenwood, M., Addis, M., Alpdemir, M.N., Ferris, J., Glover, K., Goble, C., Goderis, A., Hull, D., Marvin, D., Li, P., Lord, P., Pocock, M.R., Senger, M., Stevens, R., Wipat, A., Wroe, C. Taverna: lessons in creating a workflow environment for the life science. Concurrency and Computation: Practice and Experience, 18(10):1067-1100 (2005)

Acknowledgments

Author P.R. Moreira acknowledges Fundação para a Ciência e a Tecnologia (Portugal) for a postdoctoral grant (ref. SFRH/BPD/26527/2006).