An effective Workflow for automated comparative analysis

Miele M[1], Zappa A[2], Romano P[1]

Motivation

Workflow Management Systems, such as Taverna Workbench, may be used for automating processes that are applied to data in the life sciences. Workflows consist of a pre-defined series of inter-linked tasks that are performed by processors, i.e. local or remote elaboration units. A large number of processors are available for retrieving data and executing applications with various invocation mechanisms, often based on Web Services. Archives of available Web Services and workflows are now being built, showing a great interest in this technology. However, evidence of the effectiveness of workflows in real research environments is still lacking. Automated workflows could be extremely useful for comparative analysis of protein sequences among different species. It therefore seems that the development of new, effective workflows and the clear demonstration of their usefulness in real research settings could support the growing of the needed awareness among researchers. Here, we report on the development of a workflow aimed at undertaking a comparative analysis of proteins involved in the maintenance of stem cells. Stem cells have attracted great interest in many fields, particularly in medical areas, in which they have been heralded as potential therapeutic agents for many degenerative diseases. Both plants and animals contain stem cells, although they are especially prominent in plants. Indeed several studies clearly show that totipotency in plants is not restricted to the zygote, but it is also present in somatic differentiated cells. However, it is only recently that the notion of stem cell reversal has been recognized for animal somatic stem cells that can be induced to revert to a pluripotent stage by forced expression of defined proteins. Plant and other metazoan transcriptional factors have also been shown to play a role in the maintaining and restoring of pluripotency, but little is known on the evolution of these factors among different eukaryotes. In the present study, we developed an automated workflow for building a phylogenetic tree aimed at investigating relationships between the main factors involved in the maintenance of the stem cell totipotency among eukaryotes.

Methods

The workflow was developed by using Taverna Workbench 1.7. It implements "parallel" multiple sequence alignments and then it builds bootstrapped phylogenetic trees for each of them. Protein identifiers or sequences can be used

[1] National Institute for Cancer Research, Genoa, Italy [2] Department of Informatics, Systems and Telematics, University of Genoa, Genoa, Italy

as input. A series of alignments, with relative results, phylogenetic distance files, tree and proteins, with their descriptions, constitute the output. For an improved visualization and for further evaluation of the data, output files can be represented by using FigTree. Web Services allowing access to four of the main alignment tools (emma-ClustalW, ClustalW2, Muscle and T-coffee) and to EMBOSS-PHYLIP via a Soaplab-based server available at the EBI are used. Some local processors are also defined, including beanshell scripts. The workflow was made available on myExperiment at: http://www.myexperiment.org/workflows/1097/. All Web Services are registered and annotated in BioCatalogue.

Results

The effectiveness of the workflow was tested using, as input, IDs of plant, drosophila, planarian and mammalian transcriptional factors that have been shown to have a role in the maintenance and the restoration of totipotency. Proteins selected included DNA binding proteins that regulate transcription at nucleus level and gene silencing RNA-mediated proteins acting in cytoplasm. Plant, drosophila and mammalian proteins were selected from Uni-Prot by specifying stem cell maintenance and nucleic acid binding as reference terms in the ontology search. Planarian proteins were mainly obtained from the most recent literature. The relationship between selected proteins, obtained using genetic distance matrix based on similarity of the sequences, was produced as output by the workflow. Sequence alignment was performed with Muscle and maximum-likelihood was used for boostrap analysis. The tree obtained could be divided into 2 large clades. RNA-binding proteins were located in 2 different subgroups of the former clade: one included only animal PUM proteins and another one included animal and plant PIWI proteins. A third subgroup, including both animal and plant DNA-binding proteins, was also located in the former clade. In the latter clade 2 subgroups including plant and animal proteins were also located. Data regarding PUM and PIWI domain conservation among eukaryotes agreed to what reported in literature. This, together with the finding of NANOG and WUS families in the same clade, suggested that any clear separation between animal and plant clades is not evident. In conclusion, the use of the workflow was particularly effective. It allowed automation of repetitive procedures and provided results in different formats. It can therefore, represent a useful and time-saving tool for the study of phylogenetic relationship among of different organisms.

Availability

http://www.myexperiment.org/workflows/1097/

Contact e-mail

achille.zappa@istge.it