Genome Annotation Pipeline

Anders Lanzén*, Svenn Helge Grindhaug*, Tom Oinn**, Pål Puntervoll*

* Computational Biology Unit, Bergen Center for Computational Science, Norway
** European Bioinformatics Institute, Hinxton, Cambridge, CB10 1SD, UK

Accurate prediction and functional annotation of novel genomes is of fundamental importance for understanding sequence data. To enable efficient analysis of this data and the information extracted from it, a suitable system for information storage and retrieval is also essential. We are currently setting up a pipeline for automatised and efficient handling and storage of all steps involved in sequencing, annotation and analysis of any genome. Important steps in this process include analysis of raw sequencing data, gene prediction, functional annotation and comparative genomic analysis.

We are currently implementing a pipeline / workflow for this purpose using Taverna [1]. Web Services are being set up for all of the individual tools and components needed in the genome annotation process. For the implementation of web services we are using Soaplab [2], where appropriate. In some cases, custom web services have been generated instead.

A critical aspect of the pipeline is that all steps in the process should require a minimum of manual interference or repetitive tasks. Yet, manual inspection and manipulation must be allowed at certain key steps. To enable user interaction inside a workflow while using Taverna for its execution, a user Interaction Server for Taverna is being developed as an informal collaboration between the CBU and the EBI together with other parties involved in the myGrid project. The Interaction Server will allow for including users, called expert reviewers, as components within a workflow.

A premature version of the annotation pipeline will be used and evaluated for the annotation of a number of viruses infecting green algae. The genome sequences of these viruses will be made available by the Dept. of Biology at the UoB. These viruses are expected to share a a number of features with two already annotated algal viruses, including, by viral standards, very large genomic size. Other expected properties is the possible occurrences of splice sites, tRNA genes and repeat regions. The gene prediction part of the pipeline has already been tested with encouraging results on the EhV86 genome, recently annotated by the Sanger Institute.

References

[1] Oinn T, Addis M, Ferris J, Marvin D, Senger M, Greenwood M, Carver T, Glover K, Pocock MR, Wipat A, Li P. Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics. 2004 Nov 22;20(17):3045-54.
[2] http://industry.ebi.ac.uk/soaplab/