Bergen Center for Computational Science
COMPUTATIONAL BIOLOGY UNIT

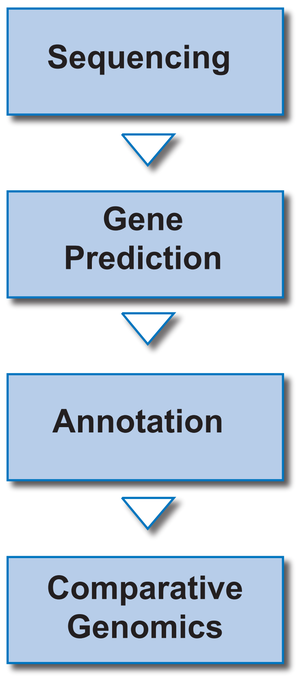
An Automatic Genome Annotation Pipeline

Anders Lanzén[1], Svenn Helge Grindhaug[1], Tom Oinn[2] and Pål Puntervoll[1]
[1]Computational Biology Unit, Bergen Center for Computational Science, University of Bergen, Norway
[2]European Bioinformatics Institute - EBI, United Kingdom

Purpose

Functional annotation of novel genomes is of fundamental importance for understanding sequence data. To enable for an efficient analysis of this data and for accurate predictions to be made from it, a suitable set of tools is necessary. However, without a suitable interface to these tools, a system for keeping track of the process and results generated, as well as a flexible architecture for connecting the different processes together, i.e. a pipeline, much time will unavoidably be wasted on manual tasks and manual housekeeping.

The purpose of the Annotation Pipeline is to be an efficient tool for automatised and efficient handling and information 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.



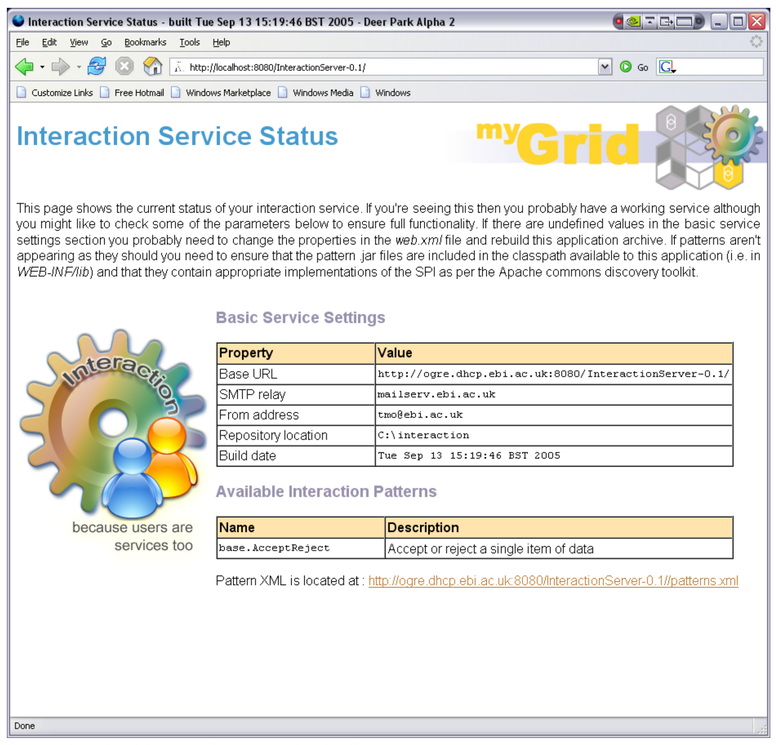
A critical aspect of the pipeline is that all steps in the process should require a minimum of manual interference or repetitive tasks. Yet, it must allow for manual inspection at certain steps. Results and metadata from all steps involved should be efficiently stored and easy to retrieve.

Design

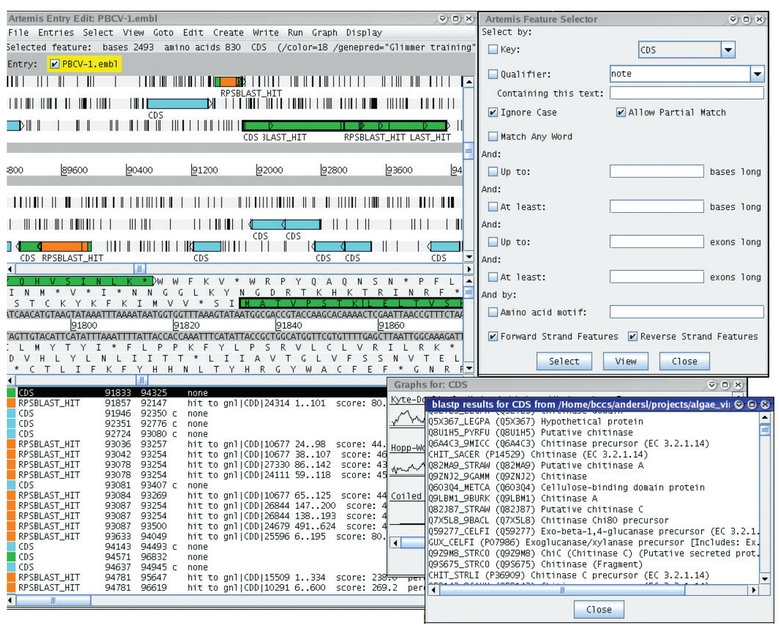
In order to address issues such as interoperability, OS- and language independency and modularity, we have tried to developed the Annotation Pipeline using a service oriented architecture (SOA) as far as possible. The services, i.e. the different tools and programs used for specific tasks in the annotation process, are made available as Web Services.

An important requirement of a useful annotation pipeline is the ability to easily change the services / components that builds it up, their behaviour and how these will be coordinated and combined. This work is carried out by a workflow management module. For this we use the Taverna workbench[1], which is part of myGrid - a UK eScience pilot project [2]. It can also be used to execute and monitor the progress in a workflow while it is running.

Another important part of the pipeline, that has yet to be developed, is the project management module. This module will keep track of and store the results and metadata of the various annotation projects that have been carried out, such as the specific workflow used and the various results generated, how they have been modified and their interdependency.



Screenshot from the status page of the Taverna Interaction Server



Screenshot from the annotation editor Artemis, developed by the Sanger Institute.

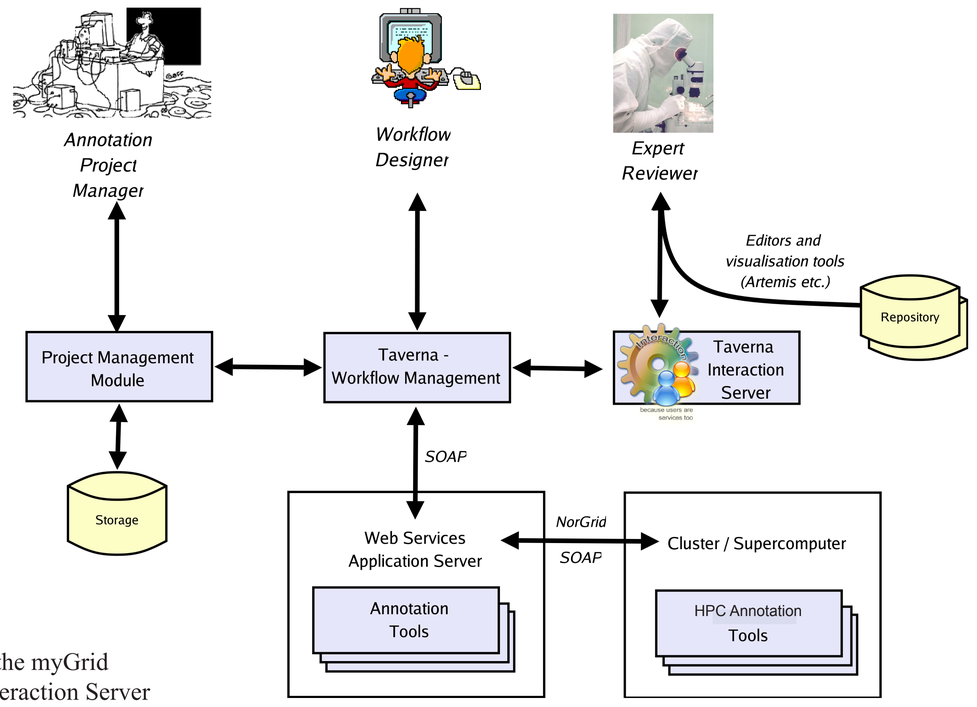
Implementation of Web Services

For the implementations of web services we are using Soaplab [3], where appropriate. In addition to this, a more advanced web service wrapping the tool HMMER [4] has been implemented. This fulfils the WSRF standard by the Globus Alliance [5], which specifies useful standards for instantiation and states of services. It is also enabled for parallelisation, such that a portal web service splits up a demanding request and spawns multiple parallel instances of a second web service that do "the heavy work". For annotation projects by the CBU, this service will be run on a cluster and be accessed through the Norgrid Portal - a standardised access point to distributed computing and grid resources in Norway.

The Taverna Interaction Server - Enabling Human Inspection and Intervention

Manual inspection and manipulation must be allowed at certain key steps of a workflow. A program called the Taverna Interaction Server is being developed for this purpose by the EBI together with other parties involved in the myGrid project, as an informal collaboration with the CBU. The Interaction Server can be deployed to a Java servlet container such as Tomcat. It allows for including users, called expert reviewers, as components within a workflow.

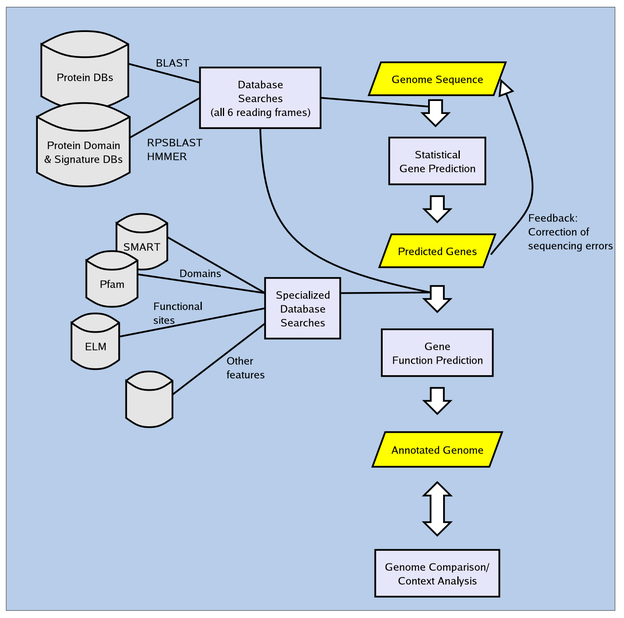
The basic workflow of the server is to send an email including a hyperlink to an applet or application to launch. This application then downloads the relevant data and displays it to the expert reviewer, who is presented with a number of choices; typically to modify the result and resubmit it, accept it as is or reject it. The behaviour in which the Interaction Server interacts with the expert reviewer, such as the application to launch, should obviously depend on the type of data and its relation to the workflow. This behaviour is defined by an Interaction Pattern, implementations of which can be added to the server at runtime. We have implemented one such interaction pattern for genome annotation "flatfiles", using the genome browser and annotation tool Artemis [6]. A modified version of Artemis has also been developed for this purpose. This is launched as a Java Web Start application and is capable of downloading the relevant data from an interaction server and uploading any changes made by the reviewer.

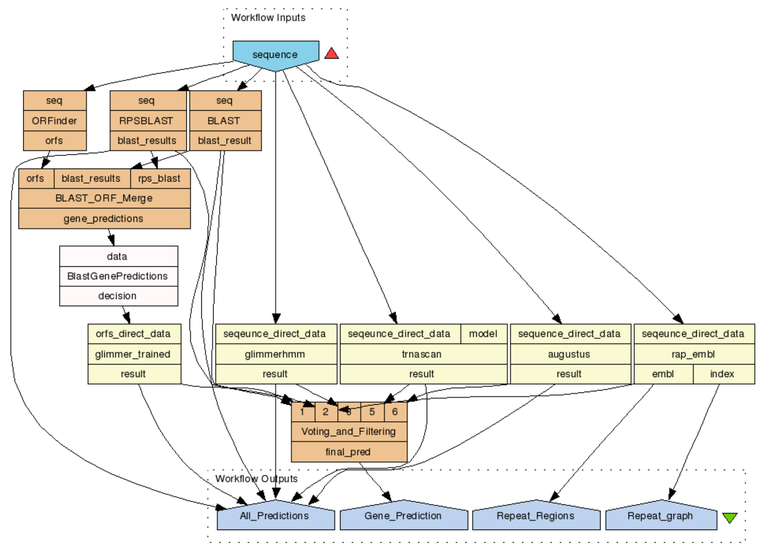


Gene prediction in algal viruses - a case study

A premature version of the annotation pipeline will be developed, 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 University of Bergen. These viruses are expected to share a number of features with each other as well as with three already annotated algal viruses, including, by viral standards, very large genomic sizes. Other expected properties are the possible occurrences of splice sites, tRNA genes and repeat regions.

The first step in a genome annotation project is to make a prediction of the gene structure that is as accurate as possible. In absence of experimental expression data, the most accurate gene predictions of protein coding genes rely on similarity to genes from other organisms. However, we do not expect this strategy to be sufficient, particularly in viral genomes, since not all genes are expected to show homology to genes currently present in available databases. Thus ab initio prediction will be necessary. The specificity of such tools when not sufficiently supervised or trained is generally unsatisfying for reliable annotation. Therefore, such tools will only be used in combination with comparative prediction methods in the gene prediction. We have developed a workflow for this purpose that includes components such as Blast, RPS-BLAST against the CDD database[7], Glimmer[8], tRNAScan-SE, RAP for identification of repeats and HMMPfam. Web services have been developed for these tools and the workflow has been used for benchmark studies on a number of previously annotated algae viruses.





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://www.mygrid.org.uk/

[3] http://industry.ebi.ac.uk/soaplab/

[4] Bateman A, Birney E, Durbin R, Eddy SR, Finn RD, Sonnhammer EL. Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucleic Acids Res.* 1999 Jan 1;27(1):260-2.

[5] http://www.globus.org/wsrf/

[6] K Rutherford, J Parkhill, J Crook, T Horsnell, P Rice, MA Rajandream, and B Barrell. Artemis: sequence visualization and annotation. *Bioinformatics*, 16(10):944-5, Oct 2000. *Nucleic Acids Res*, 27(23):4636-41, Dec 1999.

[7] Aron Marchler-Bauer, John B Anderson, Praveen F Cherukuri, Carol DeWeese-Scott, Lewis Y Geer, Marc Gwadz, Siqian He, David I Hurwitz, John D Jackson, Zhaoxi Ke, Christopher J Lanczycki, Cynthia A Liebert, Chunlei Liu, Fu Lu, Gabriele H Marchler, Mikhail Mullokandov, Benjamin A Shoemaker, Vahan Simonyan, James S Song, Paul A Thiessen, Roxanne A Yamashita, Jodie J Yin, Dachuan Zhang, and Stephen H Bryant. CDD: a Conserved Domain Database for protein classification. *Nucleic Acids Res*, 33 Database Issue:D192-6, Jan 2005.

[8] AL Delcher, D Harmon, S Kasif, O White, and SL Salzberg. Improved microbial gene identification with GLIMMER.

Main sponsoring institutions

UNIVERSITAS BERGENSIS

FUGE
Functional Genomics
Norwegian Research Council