Poster B-53

GAFFA: Genome Annotation Framework for Flexible Analysis

Authors:
Anders Lanzén (*Computational Biology Unit, Bergen Center for Computational Science, University of Bergen*)
Svenn Helge Grindhaug (*Computational Biology Unit, Bergen Center for Computational Science, University of Bergen*)
Pål Puntervoll (*Computational Biology Unit, Bergen Center for Computational Science, University of Bergen*)

Short Abstract: The aim of GAFFA is to provide a flexible and modular framework for integrating all steps involved in a genome analysis project. Further, it provides a system for keeping track of the analysis process and results generated. GAFFA is developed using a service-oriented architecture.

Long Abstract:

The need for efficient and accurate analysis of genomic data increases with the speed of genome sequencing. Recent years have seen an impressive growth of genome data as well as new analysis methods, applications and frameworks integrating multiple applications.

As the analysis methods are improved and become increasingly more complex, the primary investigator in a genome analysis project can no longer evaluate and optimise every step in the process, but may instead benefit from using proven protocols or workflows combining a number of tools in a specified manner. However, the choice of analysis methods and how to combine them into workflows varies from project to project, depending on factors such as the aim of the project, the organisms targeted, findings during the progress of the project, and personal preferences. To meet the needs of different projects in a flexible manner, while still providing a rigid and user-friendly framework for integration and data storage, is a challenge that we feel is not yet fulfilled in existing frameworks. To this end, we are developing the integration framework GAFFA.

Hence, the aim of GAFFA is to provide a flexible and modular framework for integrating all steps involved in genome analysis, from raw sequence data to gene prediction, functional annotation and comparative genomics. The framework will target a wide range of project types in genomics including EST sequencing, metagenomics, whole-genome sequencing both in prokaryotes and eukaryotes. In addition to providing a flexible framework for connecting the different analysis steps together into workflows, GAFFA will provide a system for keeping track of the analysis process and results generated. The goal is to eliminate the need for repetitive manual tasks and housekeeping as far as possible. Yet, to ensure high quality results manual inspection and manipulation must be allowed where needed.

In order to address issues such as interoperability, operating system- and language independency and modularity, GAFFA is developed using a service-oriented architecture as far as possible. The different programs and workflows used for specific tasks in the

annotation process are made available as Web Services.

A preliminary version of GAFFA is currently being used and evaluated in a number of genome/sequence analysis projects. This version includes several workflows for gene annotation, BLAST searches using a large computer cluster, EST and whole genome pre-assembly and assembly. A database schema that encompasses raw data, analysis data and metadata has been developed, and has been partially instantiated for EST analysis. Also, a user-friendly web application presenting EST assembly contigs and their sequence similarities to public database sequences has also been developed.

The development of GAFFA is targeted against several projects including full-genome sequencing of algae viruses, EST sequencing in cod and halibut and sequencing of metagenomics libraries targeting crenarchaeote. Collaborating partners include the Department of Biology at the University of Bergen, the Norwegian Institute of Marine Research and The National Institute of Nutrition and Seafood Research of Norway. By developing GAFFA in close collaboration with these partners, we believe that we will be able to better target real needs. Ultimately, we want to provide a more useful and flexible framework for scientists involved in genome annotation and analysis projects than is currently available.