Enabling incremental updates for bioinformatics workflows

Edvard Pedersen[1], Nils Peder Willassen[2], Lars Ailo Bongo[1]

[1]University of Tromsø, Department of Computer Science

[2]University of Tromsø, Department of Chemistry

A large up-to-date compendium of integrated genomic data is often required for biological data analysis. The compendium can be tens of terabytes in size, and must often be frequently updated with new experimental or meta-data. Manual compendium update is cumbersome, requires a lot of unnecessary computation, and it may result in errors or inconsistencies in the compendium. We propose a transparent file based approach for adding incremental update capabilities to unmodified genomics data analysis tools and pipeline workflow managers. We implemented this approach in the GeStore system, and integrated it with the Galaxy platform and two in-house workflow managers, using three different integration approaches. We evaluated GeStore using several real world genomics compendia. Our preliminary results show that it is easy to add incremental updates to genomics data processing pipelines, and that incremental updates can reduce the computation time such that it becomes practical to maintain an up-to-date genomics compendia on small clusters.