Poster A-18

3D-ANNOTATE: A Web Portal for Structure-based Protein Annotation using Grid Technology

Authors:
Shikta Das (*Imperial College London*)

Short Abstract: We present a stand-alone web application, back ended by a robustly designed database, called 3D-ANNOTATE, a part of the collaborative work of the e-Protein, that allows the user to select a wide range of homology based analytical tools and automatically apply them to proteins in the database.

Long Abstract:
Since the completion of the Human Genome Project in 2003, scientists have been working hard to identify functions for all of the proteins that are encoded in the Human Genome. However, to date only 30% of the proteins that exist in the human genome are known. Whilst it is possible to determine the function of the remaining proteins in a number of ways experimentally, these methods are time consuming, expensive or both. The potential to use annotation procedure that can filter the genomic data to predict the three-dimensional structure of proteins directly offers a valuable alternative. Currently, most protein structure prediction is done using a method called comparative modelling, where researchers use the structures of homologs to match the structure of a new, related gene of interest. Knowledge of the structure is vital as the structural data exposes the mechanisms behind biological function and the evolutionary relationships that are hidden at the sequence level but may be revealed at the structural level.

We present a stand-alone web application, back-ended by a robustly designed database, called 3D-ANNOTATE that allows a user to select a wide range of homology and fold recognition based analytical tools and automatically applies them to proteins in the database. This is an automated system for protein annotation applied via Grid technology for a fully developed analysis system. The work proposed here is a part of a collaborative project called e-Protein, (http://www.e-Protein.org/) which aims to provide a distributed pipeline for large-scale structural and functional annotation. The e-Protein project is an on going initiative for academics to functionally annotate protein sequences using bioinformatics applications to assign structures to the proteins and generate three-dimensional models. It utilises Grid technology in combining heterogeneous resources at multiple sites to execute the protein annotation pipelines.

At Imperial College London, the e-Protein system was developed by collaborative work between Structural Bioinformatics Group (SBG) and London e-Science Centre (LeSC). The SBG developed a homology and fold recognition pipeline along with a protein sequence database, called 3D-GENOMICS and LeSC has been developing a Grid enabled web portal, called 3D-ANNOTATE. Together, the automated system allows users to select a wide range of sequence analysis tools and automatically apply them to each protein provided by the user or selected from the database. The 3D-GENOMICS database is a database resource for storing protein of sequenced genomes using the perl based annotation pipeline. It allows

queries to the database in order to find genes within selected organisms that encode proteins with particular three-dimensional folds. The database is intended for protein sequences and currently stores 261 genomes for various organisms. Homology models are available for the protein sequences of 13 of these genomes. The 3D-ANNOTATE system provides a web interface to the annotation pipeline. The underlying perl pipeline is defined as a set of components or jobs which is executed on the Grid and is implemented by a control program which submits and monitors them. This forms an excellent bridge between the user and Grid for transparently submitting jobs on multiple resources and abstracts the user from the need to understand the architecture or the workflow requirements for performing the annotation.