

Bio Applications on NW-GRID

The list of applications for computational biology installed on NW-GRID includes: BLAST; mipBLAST; InterProScan; Exonerate; and a range of codes for bio-chemical analysis.

NCBI BLAST is typically used in molecular genetics to search for similarities between query sequences and database records. These can be either DNA or amino acid (protein) sequences. A variety of search and translation algorithms are available. A parallel version is implemented in mpiBLAST which is available as open source and used on the Grid clusters.

InterProScan is an integrated framework for a number of sequence-searching applications. It combines different protein signature recognition methods, native to the InterPro member databases. It is also freely available and used to annotate genomes and cross-reference corresponding databases. Current work is addressing 942 bacterial, 607 eukaryotic and 56 archaeal sequences.

Exonerate is a generic tool for pair-wise sequence comparison and is available as source or a pre-compiled binary. Genome annotation heavily depends on homology searches. It is for instance used to search similar gene sequences between species in order to predict the unknown gene structure from known ones. Such prediction needs some validation and this is currently using 20 million EST sequences to validate the genome annotation. In this case Exonerate performs better than a simple BLAST search.

To make life easier for end users, portlet interfaces have been written for Exonerate and InterProScan which will be integrated with the other tools in the NW-GRID VRE.

Figure 1: BioPortal interface to InterProScan

Bio-chemical analysis is being applied to determine the structure and function of proteins and enzymes. A range of codes is used for modelling large biochemical systems, efficiently and accurately. These highly parallelised techniques make it possible to construct very realistic models. Information on biological systems at a biological level is essential to understanding experimental data and in drug design, for instance enzyme inhibition.

Some of this work is carried out in the e-Fungi project <http://www.cs.man.ac.uk/~cornell/eFungi>. This is a joint research project between the School of Computer Science and the Faculty of Life Sciences at The University of Manchester and the

Departments of Computer Science and Biological Sciences at the University of Exeter.

As part of the e-Fungi project a data warehouse is being developed that integrates data from multiple fungal genomes in a way that facilitates the systematic comparative study of those genomes. Less well understood species can be studied with reference to model organisms with more fully explored functional characteristics.

To support comparative functional genomics, a library of bio-informatics queries is provided. These analyses can be combined in different ways to conduct studies of pathogenicity and evolution.

The current e-Fungi release provides integrated access to genomic and functional data of 34 fungal genomes

and 2 oomycetes. The analysis library for comparative genomics consists of a number of queries that can be parameterised by the user. The queries are organised in different groups according to their scope and the data analysed. The groups of queries include EST analysis, Mcl and OrthoMcl Cluster analysis, Essential yeast genes cluster analysis, Transcript abundance, Cellular localisation analysis and Secretome analysis.

[1] Haizhou Tang *et. al.*, *Characterising alternate splicing and tissue specific expression in the chicken from ESTs*, in *Cytogenetic and Genome Research* (2006) in press

[2] Intikhab Alam, Simon J. Hubbard, Stephen G. Oliver and Magnus Rattray, *A kingdom-specific protein domain HMM library for improved annotation of fungal genomes*, *BMC Genomics* (April 2007) 8:97

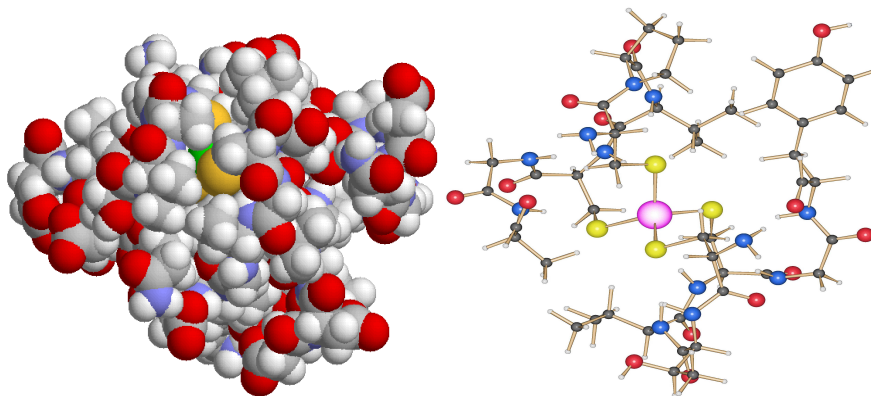


Figure 2: Iron proteins are found in a variety of organism and are vital to enabling essential bio-chemical reactions