The Nordic BioGrid project – Bioinformatics for the grid

Joel Hedlund[1], Olli Tourunen[2], Josva Kleist[2], Michael Grønager[2], Steffen Möller[3], Erik Sonnhammer[4], Ann-Charlotte Berglund Sonnhammer[5], Inge Jonassen[6], and Bengt Persson[1,7,8]

[1] IFM Bioinformatics, Linköping University, S-581 83 Linköping, Sweden
[2] Nordic Data Grid Facility, Kastruplundgade 22, DK-2770 Kastrup, Denmark
[3] Institut für Neuro- und Bioinformatik, University of Lübeck, D-23538 Lübeck, Germany
[4] Stockholm Bioinformatics Center (SBC), Stockholm University, S-106 91 Stockholm, Sweden
[5] Linnaeus Center for Bioinformatics (LCB), Uppsala University, S-751 05 Uppsala, Sweden
[6] CBU, Bergen Centre for Computational Science, N-5020 Bergen, Norway
[7] Dept of Cell and Molecular Biology, Karolinska Institutet, S-171 77 Stockholm, Sweden
[8] National Supercomputer Centre (NSC), S-581 83 Linköping, Sweden

E-mail: yohell@ifm.liu.se

Project web page: http://wiki.ndgf.org/display/ndgfwiki/BioGrid
Software web page: http://wiki.ndgf.org/display/ndgfwiki/BioGrid+software
Open source license: MIT license

Life sciences have undergone an immense transformation during the recent years, where advances in genomics, proteomics and other high-throughput techniques produce floods of raw data that need to be stored, analysed and interpreted in various ways. Biology and medicine have become information sciences and new areas of comparative biology have opened. Bioinformatics is crucial by providing tools to efficiently utilize these gold mines of data in order to better understand the roles of proteins and genes and to spark ideas for new experiments.

BioGrid is an effort to establish a Nordic grid infrastructure for bioinformatics, supported by NDGF (Nordic DataGrid Facility). BioGrid aims both to gridify computationally heavy tasks and to coordinate bioinformatic infrastructure efforts in order to use the Nordic resources more efficiently. Hitherto, the widely used bioinformatic software packages BLAST and HMMer have been gridified. Furthermore, the multiple sequence alignment programs ClustalW, MAFFT and MUSCLE have been made available on the grid. Regarding databases, the frequently used databases UniProtKB and UniRef have been made available on the distributed and cached storage system within the Nordic grid. A system for database updating has been deployed in a virtual machine hosted by NDGF. The database PairsDB updates are currently being run on BioGrid & M-grid resources. Further applications are in the pipeline to be gridified including molecule dynamics and phylogeny calculations.

The BioGrid has already contributed to provide computational power for analysis of the medium-chain dehydrogenase/reductase (MDR) superfamily. The size and complexity of this superfamily has recently been shown to far surpass the means of subclassification that have traditionally been employed for this task. Instead, more computationally demanding methods must be employed, such as profile Hidden Markov Models, implemented in the HMMer package.

The presentation will describe BioGrid from a user perspective.