Protein Clustering with CD-HIT on the EGEE

J.L. Vazquez-Poletti, E. Huedo, R.S. Montero, I.M. Llorente
Universidad Complutense de Madrid, Spain
J.M. Fernandez, A. Valencia
Centro Nacional de Investigaciones Oncologicas, Spain

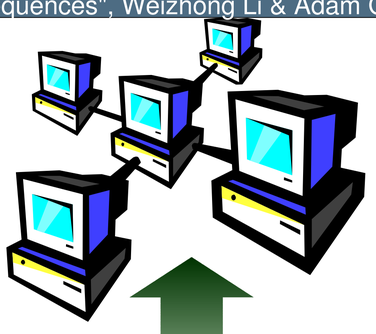
Protein Clustering

Process where sequences from a protein database are organized into groups which resembles protein families in order to capture protein similarity. It can be applied to protein families analysis, protein function prediction and annotation, homology modeling, etc...

CD-HIT

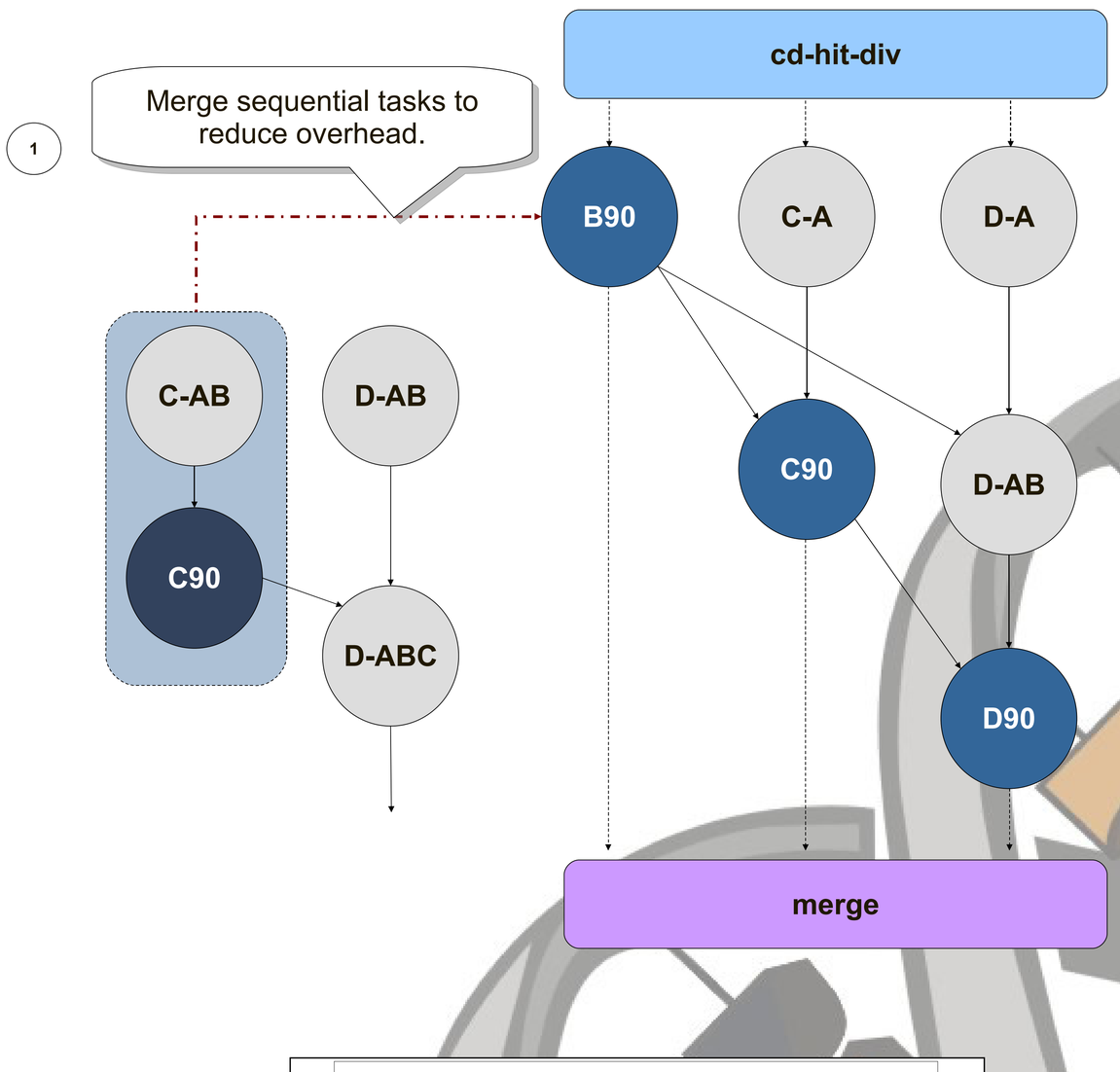
Toolkit for clustering large protein databases at high sequence identity thresholds. Redundant sequences over a given level are removed and a database of only representatives is generated.

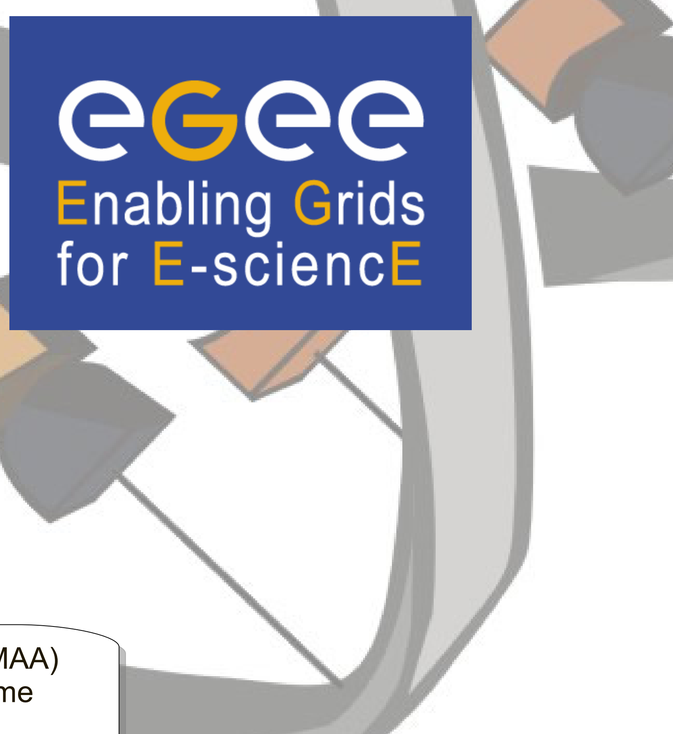
- "CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences", Weizhong Li & Adam Godzik Bioinformatics, (2006) 22:1658-9.



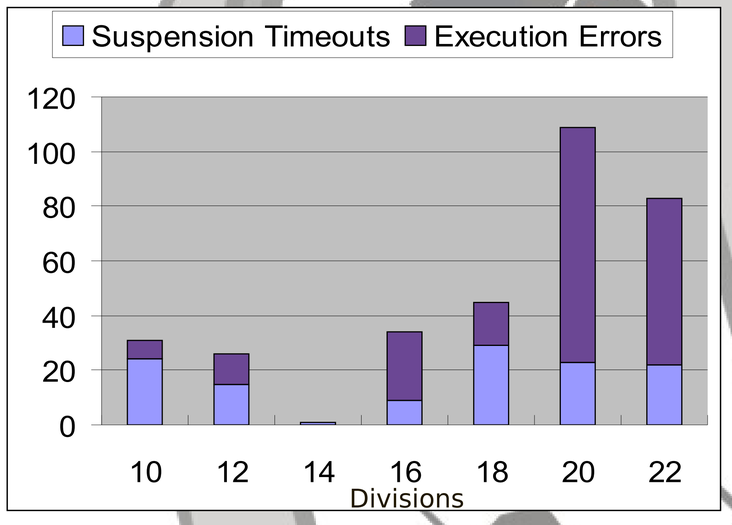
Test Database

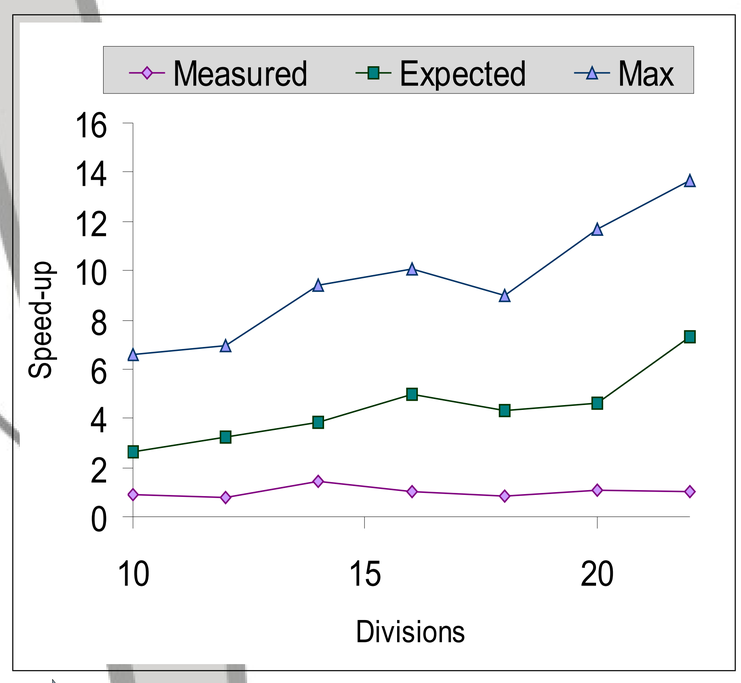
504,876 proteins (435MB) provided by National Center for Biotechnology Information

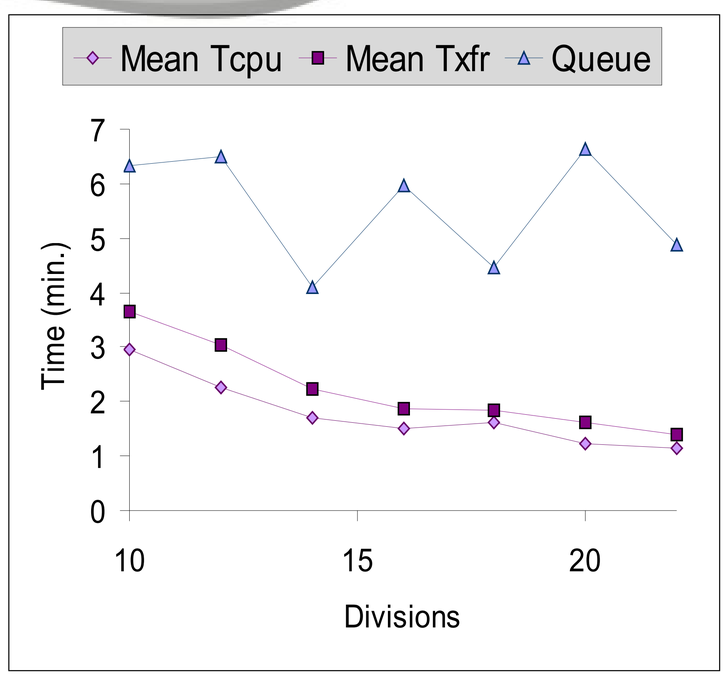




Results



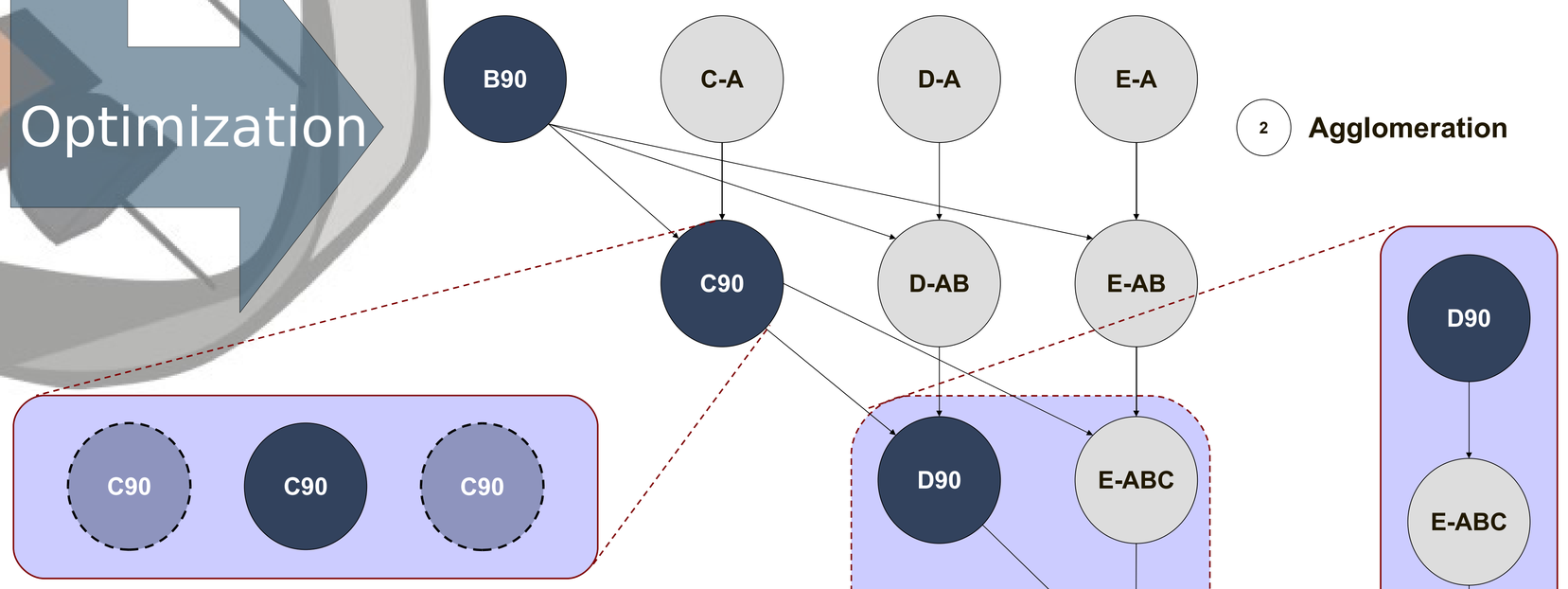




GridWay (http://www.gridway.org/)

- Metascheduler standing on top of Globus services.
- Handles DAG based workflows.
- Allows advanced flow structures (loops, branches).
- Implements the Distributed Resource Management Application API (DRMAA) which is an OGF Standard.
- Considers static and dynamic resource information.
- Offers automatic staging mechanisms.
- Provides fault tolerance mechanisms (network outage, remote and local machine crash):
  - Tries task execution/file transfer on the same resource.
  - Submits failed task to an alternate resource.
  - Failed tasks are moved transparently to other resources.

Optimization



New Input Database (Prod.)

4,186,284 proteins (1.7GB) provided by Centro Nacional de Investigaciones Oncológicas ...and growing!