Protein Clustering on the EGEE

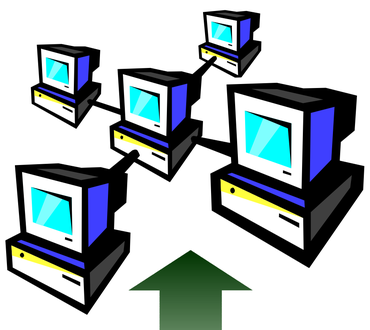
J.L. Vazquez-Poletti, E. Huedo, R.S. Montero, I.M. Llorente
Universidad Complutense de Madrid, Spain
J.M. Fernandez, A. Valencia
Centro Nacional de Investigaciones Oncologicas, Spain

Protein Clustering

Process where protein databases are organized into groups or families in order to capture protein similarity. It can be applied in domain analysis, large protein database organization and search improvement.

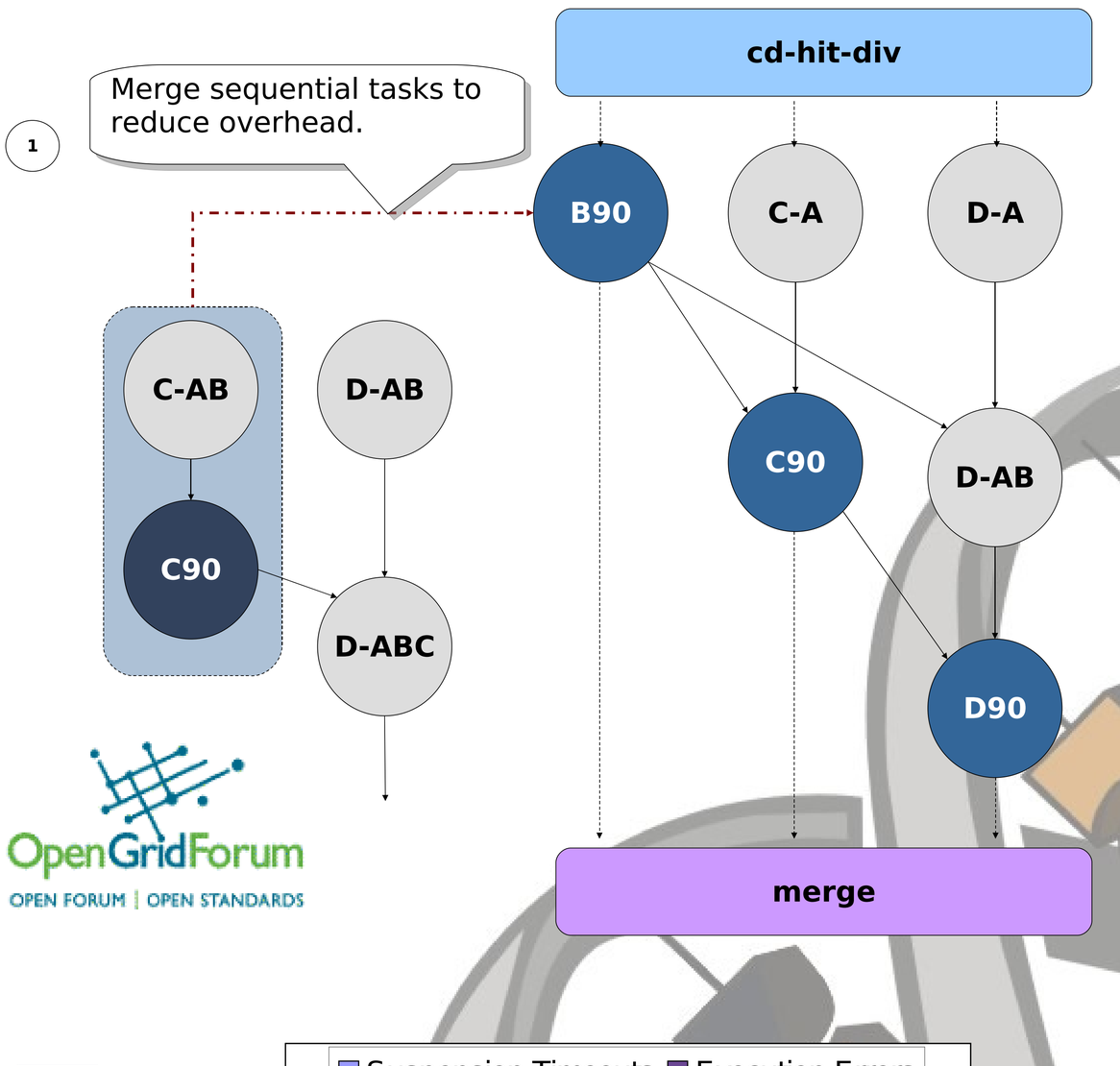
CD-HIT

Toolkit for clustering large protein databases at high sequence identity threshold. Redundant sequences are removed an a database of only representatives is generated.

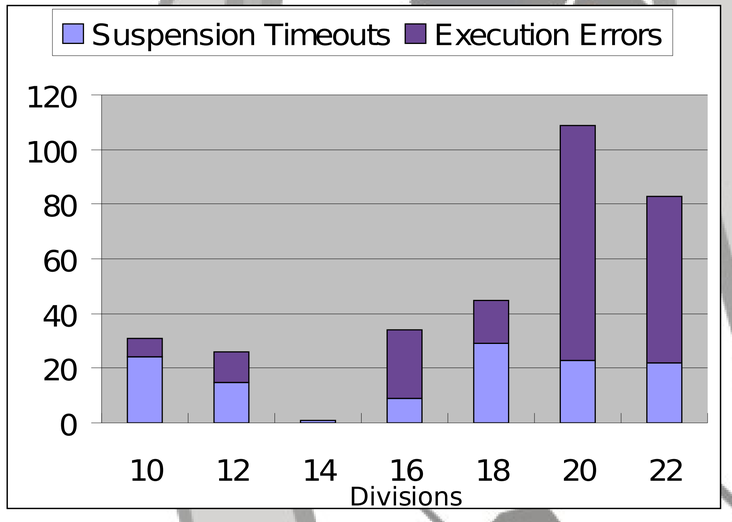


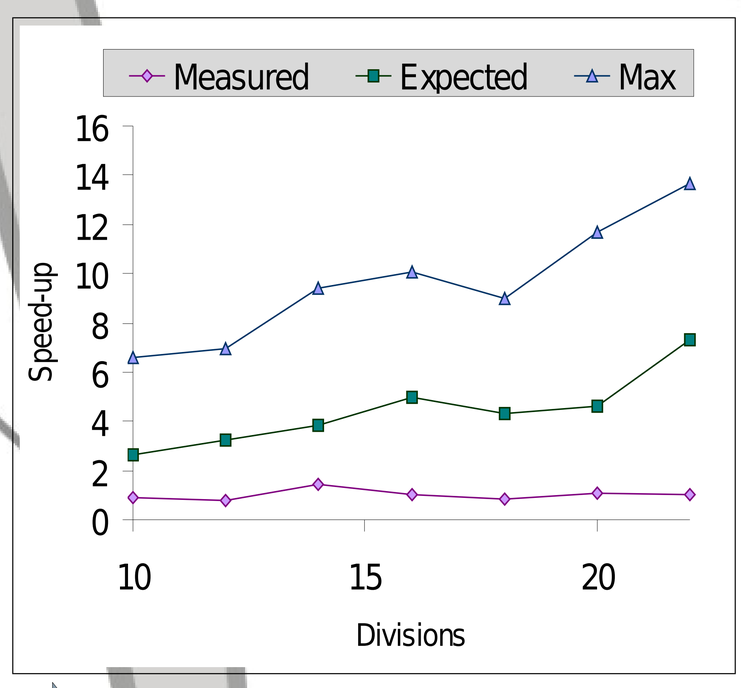
Input Database

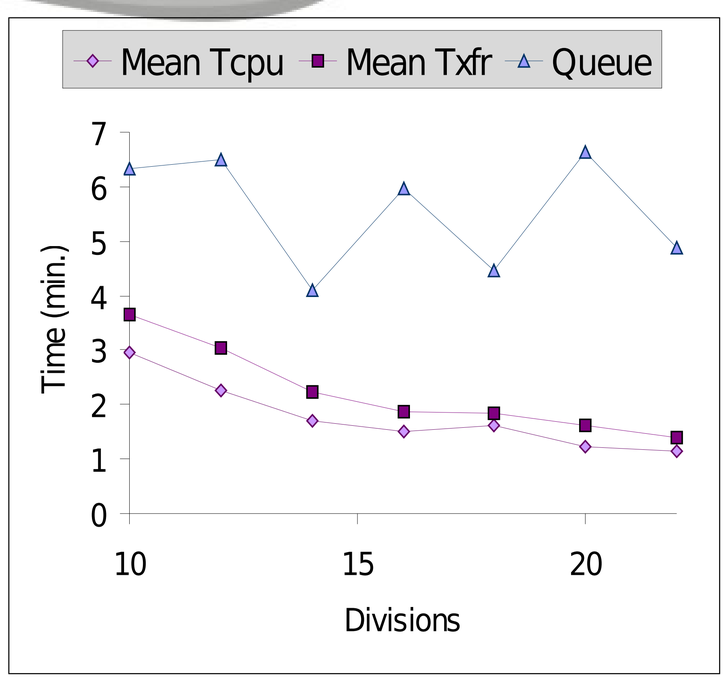
504,876 proteins (435MB) provided by National Center for Biotechnology Information



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