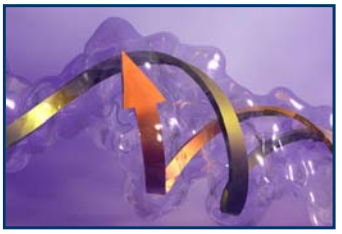
Bioinformatics Workflow using ASSIST on GRID

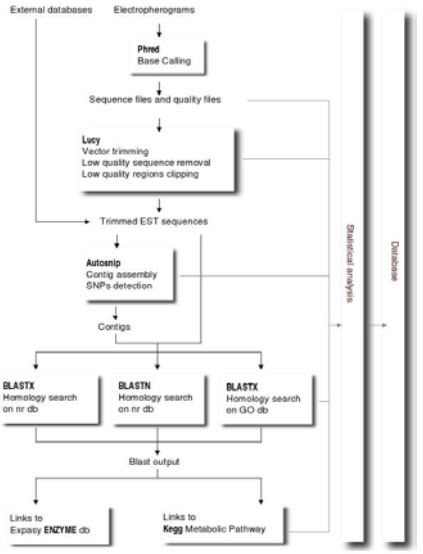
Ivan Merelli
ivan.merelli@itb.cnr.it



- Bioinformatics is a discipline which try to solve biological problems with the methodical approach typical of IT.
- The great challenge in biological research today is how to turn data into real knowledge.

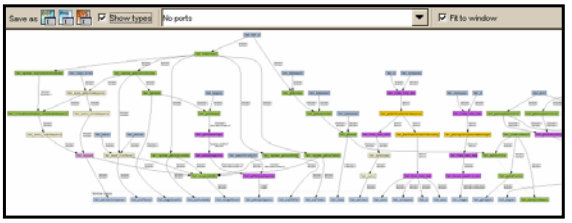
Workflow

- In order to understand complex biological processes it is consolidated for a long time in bioinformatics the workflow technology.
- Combining different software modules it is possibile to define very different data elaboration to integrate biological data.

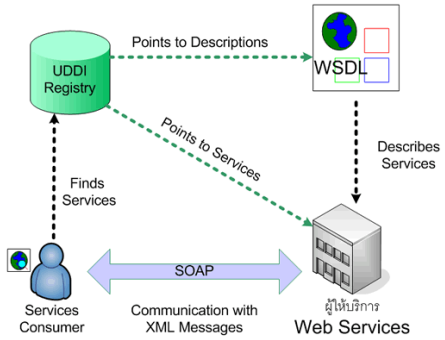


Workflow & Web Services

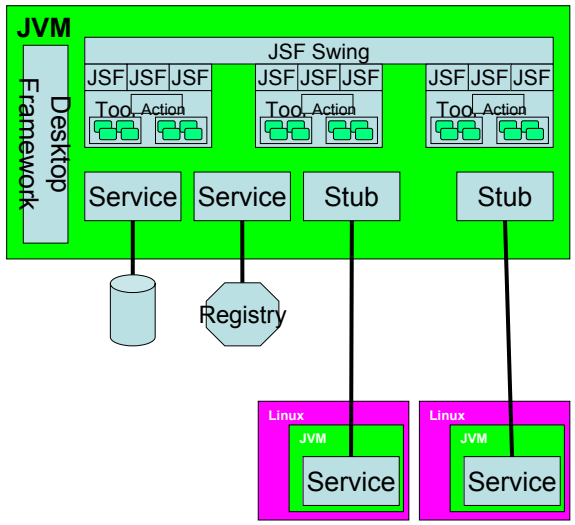
- The flexibility to define complex workflow is amplified by the possibility to integrate elaboration provided by remote server through the Web Services technology.



- The use of the WS technology allows the integration of very different systems through the definition of interoperability standards among the various modules in terms of communication protocols, but also from the semantic point of view.

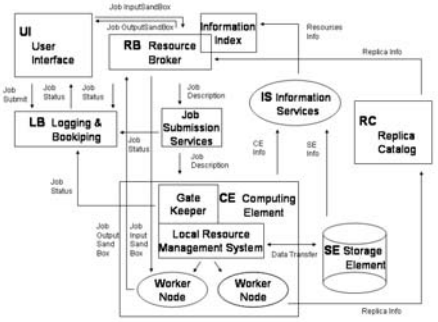


- The computational load for the most time-consuming services will take advantage from the redirection on a hidden calculation resource.
- A good solution is to use the grid platform to perform the jobs, while hidding the computation resource behind the Web Services interface.



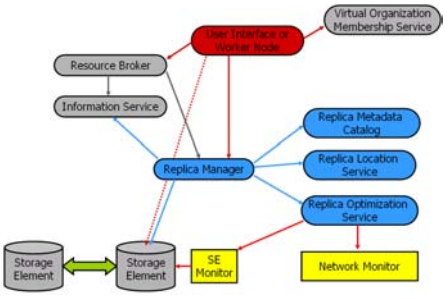
GRID – Computational resources

- Huge amount of nucleotide sequences demands increased necessity of computation power and a technology like grid, that allows high performance calculation, seems to be ideal for integrating such a vast quantity of heterogeneous data.
- Using a number of computing elements distributed in different grid sites it is possible to parallelize the most time-consuming steps of the computation, subdividing the elaboration in a series of small jobs.



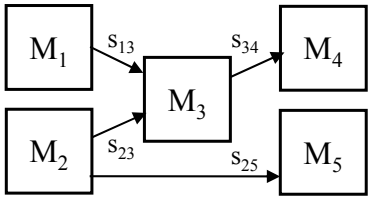
GRID – Data management

- Besides the problem of the calculation, once bioinformatics faces analysis at genomic scale, it is necessary to solve the problem of the data management.
- Also in this case the grid technology can be exploited to store data on a network of Storage Element maintaining the coherence among the replicas inside the whole infrastructure.



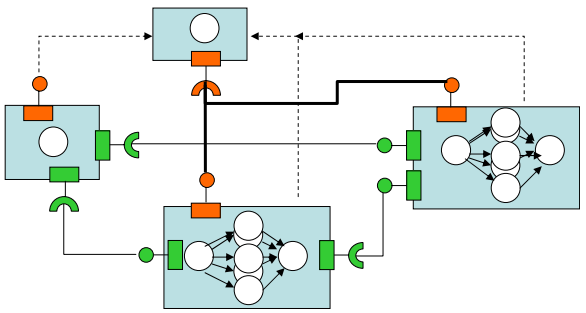
ASSIST – Program Environment

- The idea is to use a grid oriented programming language, like ASSIST, to perform high performance computations of software developed ad hoc for the analysis in matter, maintaining in the meantime the interoperability with the bioinformatics Web Services platform.
- ASSIST is a high level structured parallel programming system that integrates skeleton technology in a flexible and powerful environment in order to provide suitable support for the development of high performance portable applications.



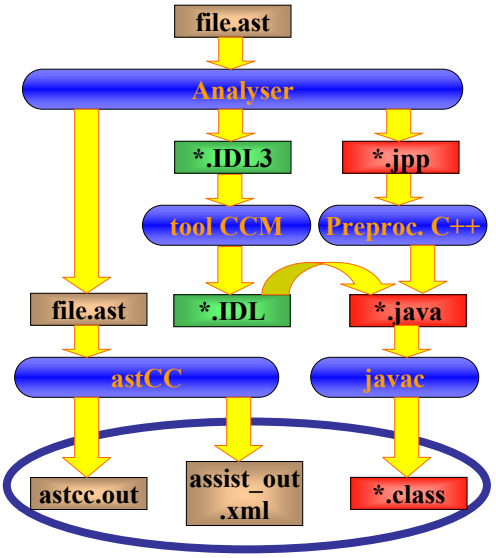
ASSIST – Program graph

- An ASSIST program is a graph in which nodes represent modules, or components, and arcs correspond to interfaces associated to a directional streaming of data.
- Streaming permits the composition of modules in a complex program in which the module can be executed in parallel or a sequential mode.



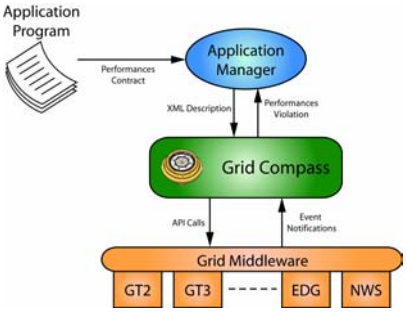
ASSIST - Compilation

- A powerfull integration is possible thanks to the multilanguage support of ASSIST that can be used to join both C++ stand alone application and JAVA portable implementation of Web Services client.



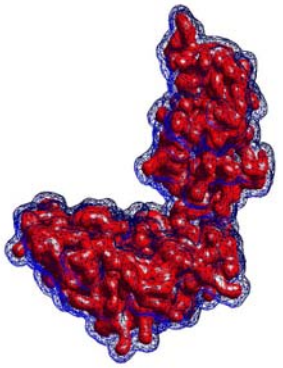
ASSIST - Grid loader

- Using a graphical loader it is possible to configure and execute an ASSIST application on a Globus based grids.
- It hides the programmer the structure of the grid and provides the interaction between the ASSIST Run Time Support and the Globus Middleware.



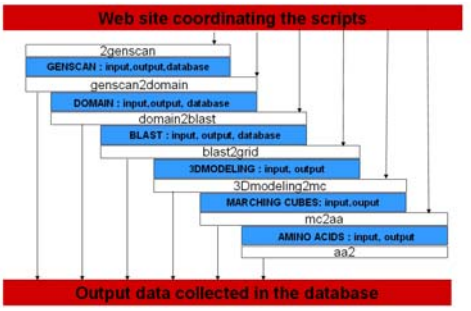
S2S – Case of study

- This study concerns the development of a high performance analysis to correlate different kind of Bioinformatics data.
- Starting from the Nucleotide Sequence it is possible to identify the exposed residue on the relative Protein Surface.



S2S - Structure

- The analysis consists of different steps of computation:
  - gene prediction
  - protein domain identification using different databases
  - blast search against the pdb sequences dataset
  - protein surface definition
    - 3D protein model generation
    - marching cubes algorithm
  - surface amino acids definition



S2S - Performance

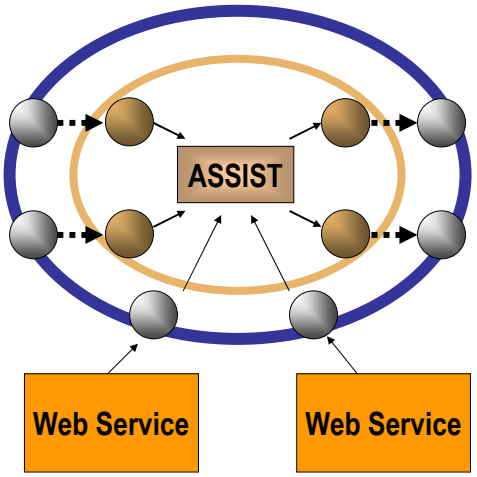
- The execution of the analysis tends to expand and a distributed implementation using the grid platform is justified even if there is an overhead for the network communications.
- Thanks to parallelism that has been implemented using different grid nodes this task has been carried out with particular success.

S2S

- This workflow analysis is already fully implemented on the grid platform and is accessible from a Web Site.
- The surface modeling and the amino acids analysis have been already re-implemented using ASSIST to use the grid platform resources in a more performing way.
- Several tests have been made to verify if a WS implementation of the sequence-based steps could be usefull.

Conclusion & Future Work

- Currently we are working to integrate JAVA Web Service Client in our workflow system for application that take effective advantages from this technology.
  - More portability
  - More interoperability
  - Less performance



Acknowledgement

This work has been supported by the Italian FIRB-MIUR projects "Laboratorio Italiano di Tecnologie Bioinformatiche, LITBIO" and "Enabling platforms for high-performance computational grids oriented scalable virtual organizations, GRID.IT".

- Luciano Milanesi, ITB-CNR
  - luciano.milanesi@itb.cnr.it
- Ivan Merelli, ITB-CNR
  - ivan.merelli@itb.cnr.it
- Giulia Morra, ITB-CNR
  - giulia.morra@itb.cnr.it