



Proteomics applications in grid

Ivan Merelli

Institute for Biomedical Technology

National Research Council

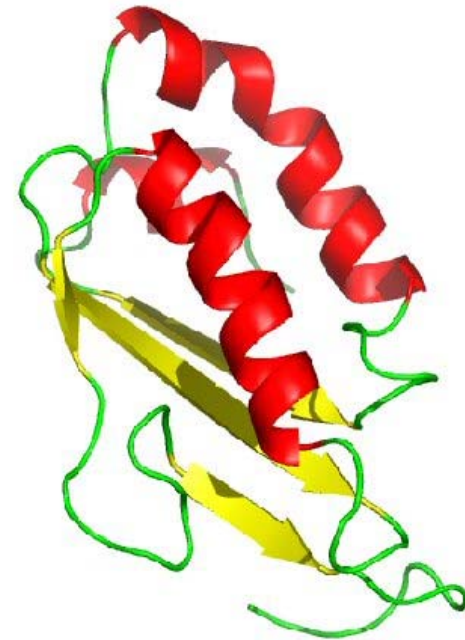


dkfz.



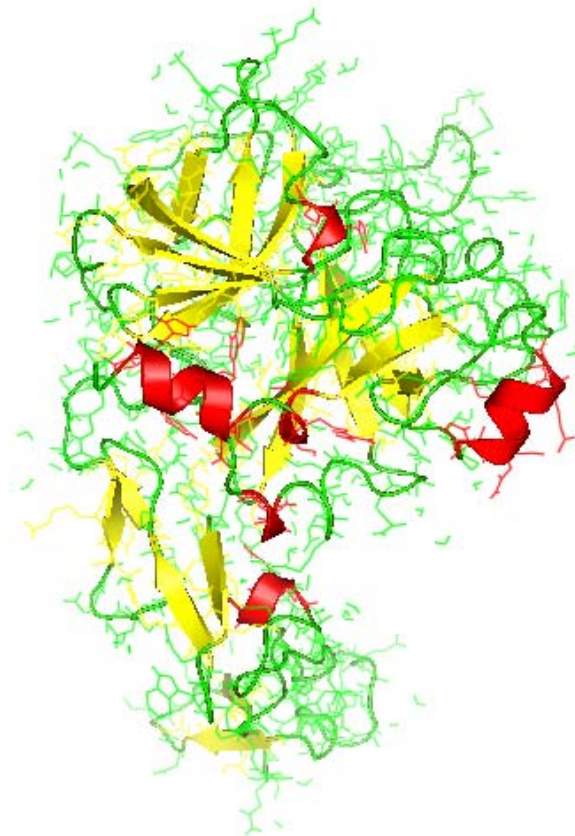


- Proteomics Work Package
- Sequence based analysis
 - Analysis software
 - Distribute approach
 - Relational job control
 - Database management
- Structural studies
 - Post Processing
- Web Interface
- Conclusions
- Acknowledgement





- The main task of the proteomics Work Package is the evaluation of different programs and databases to perform high throughput proteomics analysis in grid to face genome scale analysis.
- This Work Package is interested both in sequence based functional identification and in structural studies related to the surface atoms configuration.





- The objectives of the proteomics Work Package are:
 - an evaluation of the computational load needed by the most used proteomics software
 - a study on the possible strategy for porting on grid time consuming application
 - an analysis of the grid scalability for bioinformatics application.
- Thus, in order to achieve these objectives the creation of a suitable infrastructure to perform programs and databases in grid plays a crucial role.





- A crucial tasks of the sequence-based proteomics is to understand the protein functionality.
- The first approach when dealing with a new sequence is to perform a similarity search in order to compare the sequence against the available protein sequence database.

Sequences producing significant alignments: (bits) Score E Value

pdb 1F7S A Chain A, Crystal Structure Of Adf1 From Arabidopsis T...	51	2e-07
pdb 1AHQ Recombinant Actophorin	47	3e-06
pdb 1CNU A Chain A, Phosphorylated Actophorin From Acanthamoeba P...	47	5e-06

>pdb|1F7S|A Chain A, Crystal Structure Of Adf1 From Arabidopsis Thaliana
Length = 139

Score = 51.2 bits (121), Expect = 2e-07

Identities = 33/130 (25%), Positives = 65/130 (50%), Gaps = 5/130 (3%)

Query: 5 TGIQASEDVKEIFARA---RNGKYRLLKISIENEQLVIGSYSQPSDSWDKDYDSFVLPLL 61
+G+ +D K F R +++ KI ++Q+V+ QP ++++ + LP
Sbjct: 6 SGMVAVHDDCKLRFLELKAKRTHRFIVYKIEEKQKQVVVEKVGQPIQTYEEF--ACL PAD 63

Query: 62 EDKQPCYILFRLDSQNAQGYEWIFIAWSPDHSVSRQKMLYAATRATLKKEFGGGHIKDEV 121
E + Y +++N Q + FIAW PD + VR KM+YA+++ K+E G +++
Sbjct: 64 ECRYAIYDFDFVTAENCQKSKIFFIAWCPDIKVRSKMIYASSKDRFKRELDGIQVELQA 123

Query: 122 FGTVKEDVSL 131
+ D+ +

Sbjct: 124 TDPTEMDL DV 133

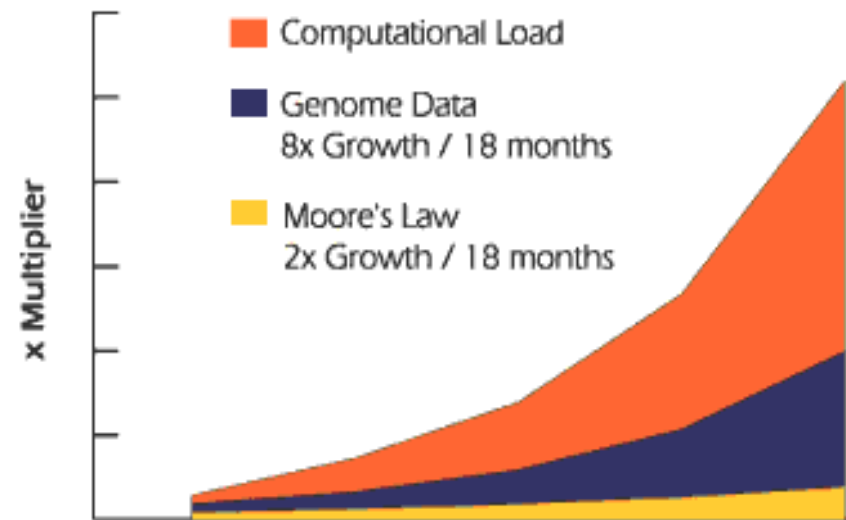


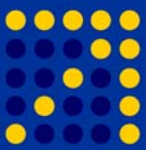
- Most genome-wide functional annotations are carried out by using sequence analysis tools such as BLAST, or motif and profile-based search tools like PROSITE and PFAM.
- Domain patterns are assuming high importance in the analysis of the macromolecular functionality in correlation with the protein sequence.





- The computational load needed for the protein analysis applications is quite important:
 - To perform an HMM analysis with hmmer against the Pfam database, for example, we need 0.3 sec for each amino acid
 - To perform a local alignment with blast against the NR database we need nearly 0.8 sec for each character.





- BLAST is the most used software for the sequence analysis in Bioinformatics and relies on plain sequence data that must be adequately indexed before starting the analysis.
- The type of analysis differ according to the type of input sequence that can be composed by either nucleotides or amino acids and the reference database that in the same way can present protein sequences or genetic sequences:
 - **blastp**: compares an amino acid query sequence against a protein sequence database
 - **blastn**: compares a nucleotide query sequence against a nucleotide sequence database
 - **blastx**: compares a nucleotide query sequence translated in all 6 reading frames (3 on each strand) against a protein sequence database
 - **tblastn**: compares an amino acid query sequence against a nucleotide sequence database translated in all 6 reading frames.



- The databases and tools considered for the protein domain analysis are freely available under GNU licence agreement from the EBI's ftp server:
 - **BlastProDom** is a wrapper script on top of a Blast package used to search against PRODOM database of protein domain families obtained by automated analysis of the SWISS-PROT and TrEMBL protein
 - **FingerPrintScan** is used to search against the PRINTS collection of protein signatures which is very useful to detect similarities in highly divergent protein super-families
 - **HMMPiR** is a script based on hmmer that performs a wide range analysis on PIR SuperFamily, a classification database based on evolutionary relationship
 - **HMMPfam** is a script based on hmmer used to search against the Pfam database that contains curated multiple sequence alignments for each family and the corresponding hidden Markov models (HMMs)
 - **HMMSmart** is a script based on hmmer for the identification of genetically mobile domains and for the analysis of their architectures against the SMART database
 - **TIGRfam** is a script based on hmmer that implements a full alignment against TIGRFAM, a collection of protein families curated multiple sequence alignments, Hidden Markov Models (HMMs) and associated information designed to support the automated functional identification of proteins by sequence homology.



- **ProfileScan** is used to search against the PROSITE profiles database, a set of position-specific table of amino acid weights and gap costs, to identify protein family with very divergent sequences
- **ScanRegExp** is used to search against the PROSITE patterns collection of regular expression and verify the matches by statistically significant confirm patterns
- **Superfamily** is a script based on hmmer used to search against the SUPERFAMILY library of Hidden Markov Models that represent all the proteins of known structure, based on SCOP
- **SignalPHMM** performs prediction of signal peptide cleavage sites, using HMM.
- **TMHMM** is used to predict the transmembrane helices in proteins using HMM.
- **PANTHER** is a software which queries a unique resource that classifies genes by their functions, using published scientific experimental evidence and evolutionary relationships to predict their function
- **Seg** is used to identify and mask some low compositional complexity segments in amino acid sequences
- **Coil** is used to predict coiled-coil regions, using the Lupas algorithm.



- In order to use this analysis software with grid technology a classic approach for the distributed computation has been used.
- This sequence based analysis can be addressed with a pure data parallel approach
 - This means that subdividing the main task in a set of small jobs it is possible to obtain a set of independent computations
 - In this way the performance of the grid platform can be fully exploited and the distributed approach becomes an efficient solution.



- Thus, the easiest and most effective way to use the protein analysis tools on the grid platform is to divide the query file into a series of small *multifasta* files containing a balanced number of sequences.
- The implementation of the protein domain analysis consists in creating an efficient system to coordinate the jobs submission, to check the computation status and to collect the results.

```
>gi|90110031|sp|Q15349|KS6A2_HUMAN Ribosomal protein S6 kinase alpha-2
MDLSMKKFVRRFFSVYLRRKSRSKSSLSRLEEVEGVKEIDISHHVKEGFEKADPSQFELLKVLGQGSY
GKVFLVRKVKGSDAGQLYAMKVLKATLKVDRVRSKMERDILAEVNHFFIVKLHYAFQTEGKLYLILDF
LRGGDLFTRLSKEVMFTEEDVKFYLAELALALDHLHSLGIIYRDLKPENILLDEEGHIKITDFGLSKEAI
DHDKRAYSFCGTIEYMAPEVNVNRRGHTQSADWWSFGVLMFEMLTGSLPFQGKDRKETMALILKAKLGMPQ
FLSGEAQSLLRALFKRNPENRNLGAGIDGVEEIKRHPFFVTIDWNTLYRKEIKPPFKPAVGRPEDTFHFDP
EFTARTPTDSPGVPPSANAHHLFRGFSFVASSLIQEPSQQDLHKVPVHPIVQQLHGNNIHFTDGYEIKED
IGVGSYSVCKRCVKATDTEYAVKIIDKSKRDPSEEIEILLRYGQHPNIIITLKDVIYDDGKFVYLMELMR
GGELLDRILRQRYFSEREASDVLCTITKTMDYLHSQGVVHRDLKPSNILYRDESGSPESIRVCDFGFAKQ
LRAGNLLMTPCYTANFVAPEVLKRQGYDAACDIWSLGIILYTMLAGFTPFANGPDDTPEEILARIGSGK
YALSGGNWDSISDAAKDVVSKMLHVDPHQRLTAMQVLKHPVVNREYLSPNQLSRQDVHLVKGAMAATYF
ALNRTPQAPRLEPVLSSNLAQRRGMKRLTSTRL
```

```
>gi|56749457|sp|Q15208|STK38_HUMAN Serine/threonine-protein kinase 38
MAMTGSTPCSSMSNHTKERVTMTKVTLNFYNSLIAQHEEREMRQKKLEKVMEEGLKDEEKRLRRSAHA
RKETEFLRLKRTLGLLEDFESLKVIGRGAFFVRLVQKQKDTGHVYAMKILRKADMLEKEQVGHIRARDI
LVEADSLWVVKMFYSFQDKLNLYLIMEFLPGGDMMTLLMKKDTLTHEETQFYIAETVLAIDSIHQLFH
RDIKPDNLLLDKSGHVKLSDFGLCTGLKKAHRTEFYRNLNHSLSDFTFQNMNSKRKAETWKRNRRLQAF
STVGTPTYIAPEVFMQTYGNKLCDWWSLGVIMYEMLIQYPPFCSETPQETQYKVMNWKETLTFPPEVPI
EKAKDLILRFCEWEHRIGAPGVVEIKNSFFEGVDWEHIRERPAASIEIKSIDDTSNFDEFPPESDILK
PTVATSNHPETDYKNKDWVFINYTYKRFEGLTARGAIPSYMKAAK
```



- For each split input file a JDL script is dynamically produced in order to describe the job and making the RB, which manages the policies for job allocation, able to route it.
- Moreover, the JDL script specifies the application reference database to be copied on the CE where the application will be performed.

```
[  
  Executable = "hmmpfam.sh";  
  Arguments = "test.seq exit.seq";  
  StdOutput = "stdout";  
  Stderr = "stderr";  
  InputData = "lfn:/grid/biomed/Pfam";  
  DataAccessProtocol = "gsiftp";  
  InputSandbox =  
    {"hmmpfam", "hmmpfam.pl",  
     "hmmpfam.sh", "input.seq"};  
  OutputSandbox =  
    {"stdout", "stderr", "output.seq"};  
]
```



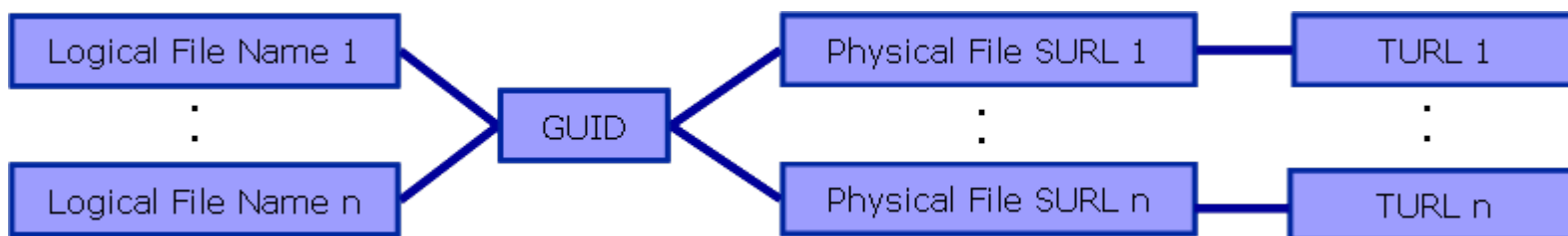
- All the JDL scripts are coordinated by an infrastructure that divides the input data and controls the advancement state of the jobs, retrieving the output when the jobs are successfully finished or submitting them again in the case they are incorrectly completed.
- A crucial problem is how to control the submission and how to integrate the output data
 - For each analysis tool and database, a different distributed implementation has been chosen in order to maximize the job efficiency
 - This aspect is highly influenced, for example, by the size of the I/O and of the reference database.



- The developed infrastructure relies on a Relational Database for monitoring the jobs' execution that is constantly updated with the computation status.
- It plays a crucial role for the distributed management, masking the grid complexity to the users.
- The key information stored in the Relation Database for each job are:
 - the application to perform
 - the related execution parameters
 - the user input data
 - the reference database.



- A key feature for this kind of analysis has been identified in storing and updating all the databases on the Grid platform.
- Working with the APIs of gLite for database handling, a solution has been developed to manage the database replicas and the updating.



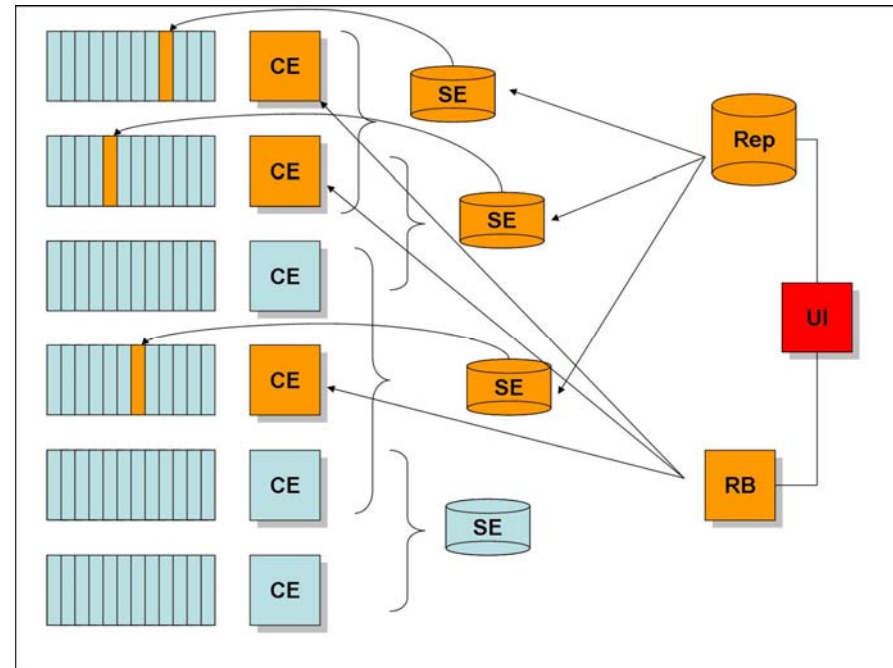


- The Automatic Updater (AU) constantly monitors FTP sites looking for newest versions of each databases
 - When a new timestamp on FTP sites is detected, the newest version is automatically downloaded and replaces the older version on the grid
 - Before clearing the older version, an xdelta patch is computed allowing to regenerate the old version starting from the new one.



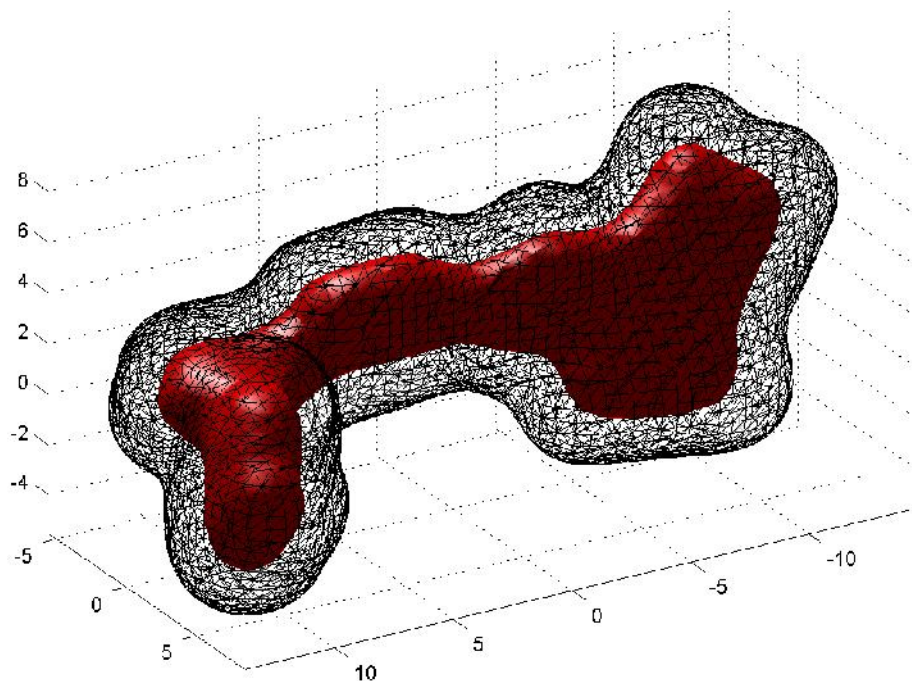


- Moreover, this software for the data management allows to replicate dynamically each database in relation with its usage in order to balance the number of replicas, and so the performance, taking into account the occupied disk space.
- It relies on the statistical analysis of the database usage by the grid jobs, working on data acquired after each job execution, regarding grid queue times, database set up times and overall job computation.



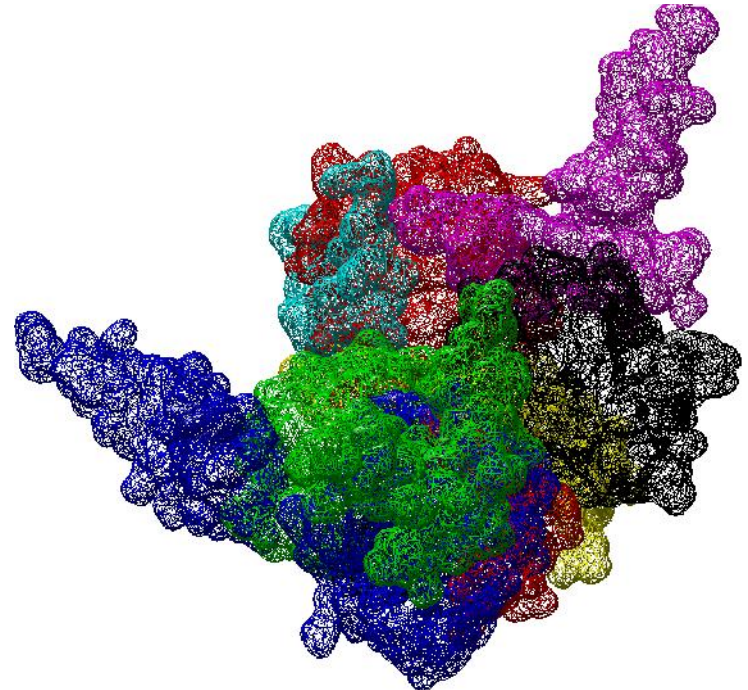


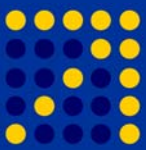
- The other main objective of the proteomics Wok Package is to address the protein functional analysis from a structural point of view.
- This kind of analysis is usually very time consuming because it relies on a pure geometrical approach.
- As for sequence based analysis, working at genome scale, this analysis takes great advantages from an HPC platform like grid.



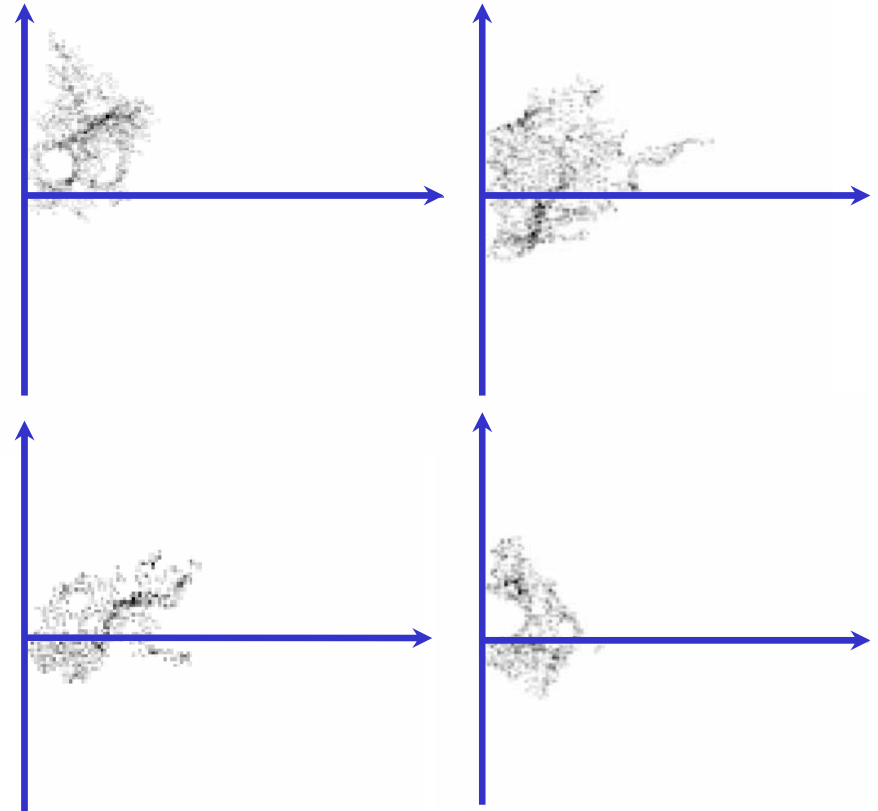


- In this context a novel approach consists in analysing the macromolecular surfaces, in order to establish possible functional correspondence among proteins.
- The problem of matching different surfaces is complex because the description of very similar surfaces can be performed in different way.



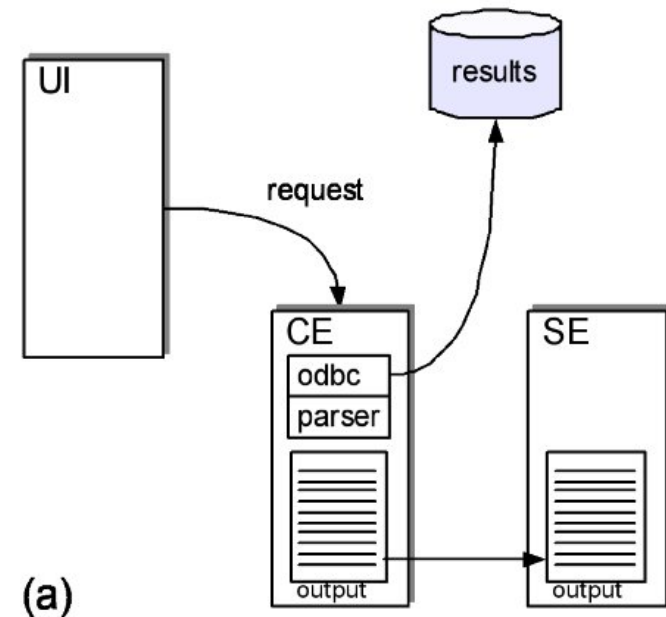


- To establish correspondence between different surfaces we will perform correlation relying on images of local description.
- Using this system we will identify 3D - transformation to align and compare the surface.
- The high number of correlations that have to be calculated represent a major issue: for this reason we will implement a grid version of this analysis system.



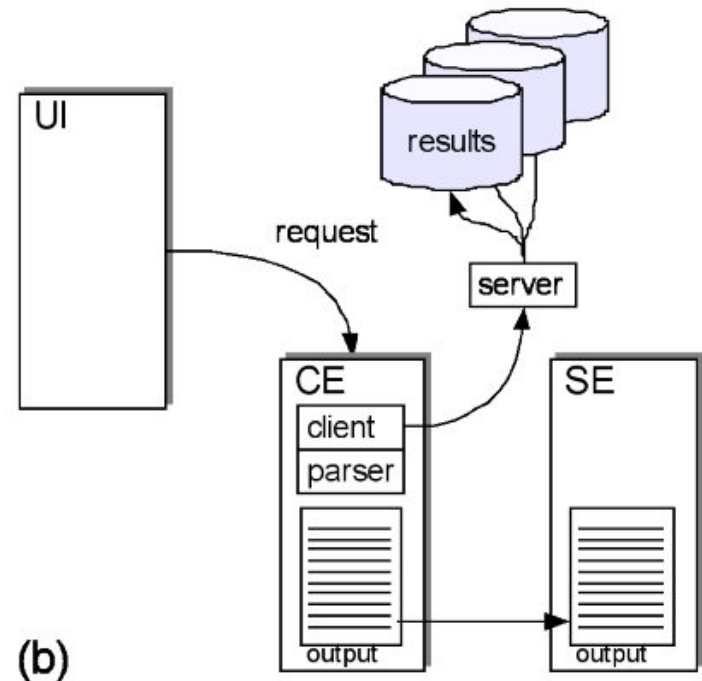


- The post process analysis of the results can be computationally very expensive, in some cases as much as the computation itself, if an adequate system to collect results has not been previously established.
- We face complex data challenges performing both the parsing of the output results and the storage of the data in the output database directly from the computational resource.



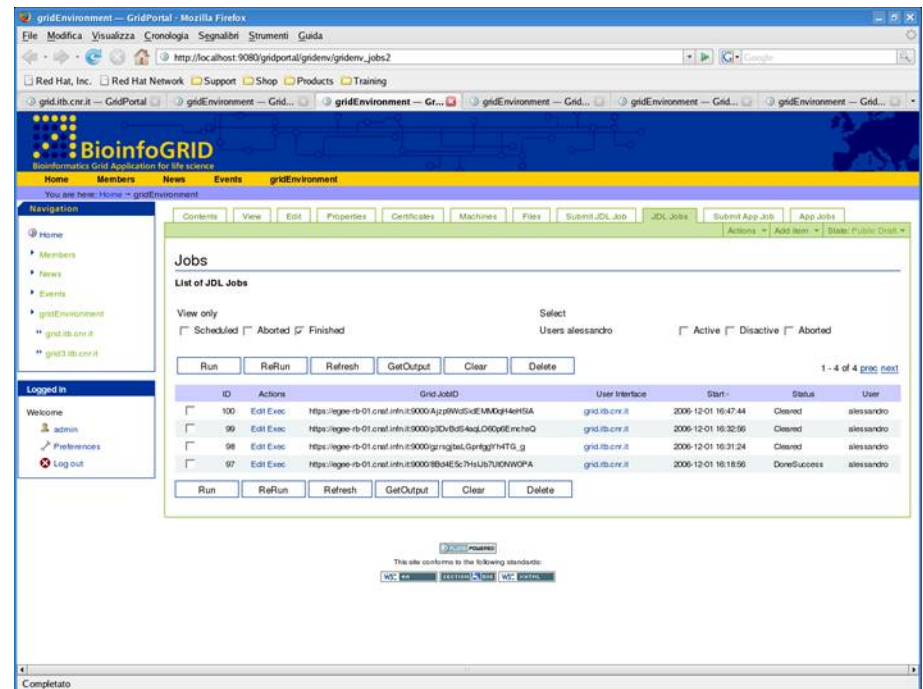


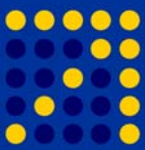
- By designing an analysis system that can be performed immediately after the computation to parse the results, the grid performance can be fully exploited and the post processing problem overcome.
- This is possible by implementing an output resource in which the results can be collected and using a client to contact this database in order to store relevant information.



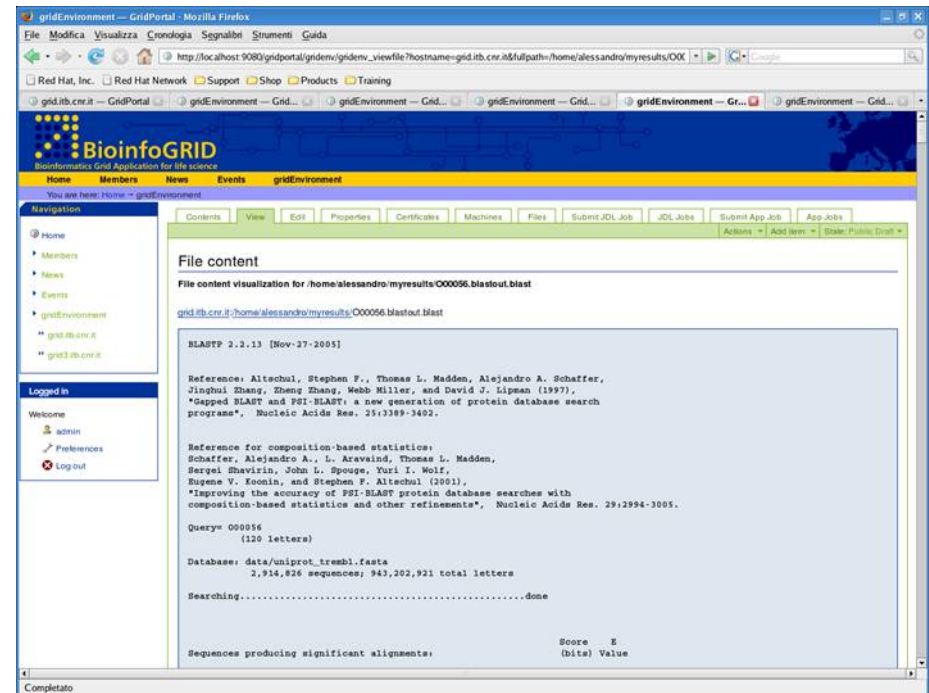


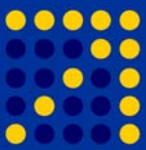
- In order to make this software rapidly accessible a user interface has been developed.
- It is used to submit jobs in the grid infrastructure, to visualize in a clear form the obtained results and to hide the complexity of the distributed platform.



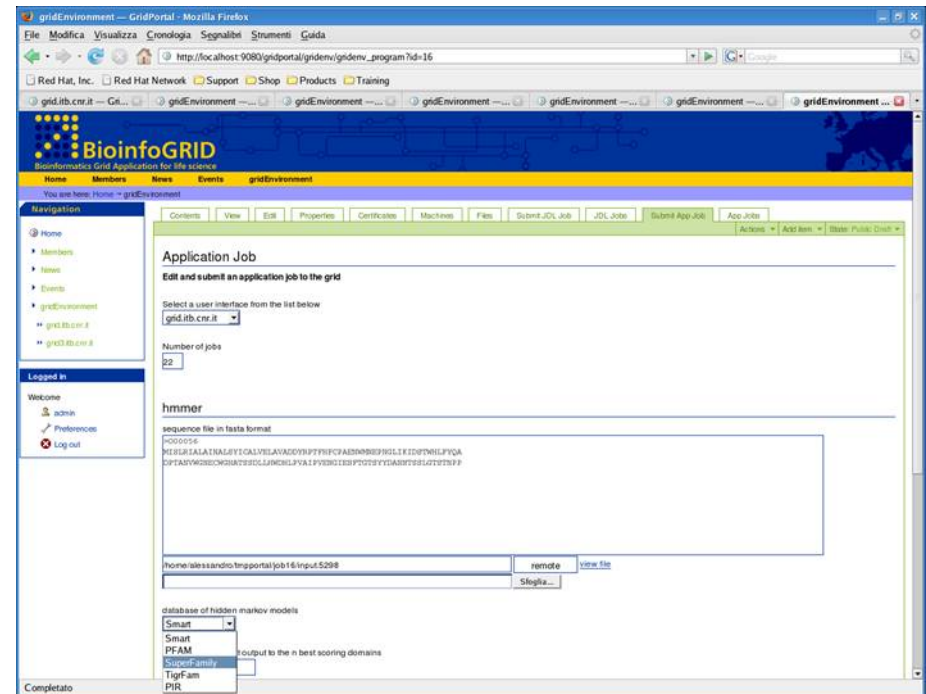


- The main feature of the portal is the possibility to hide completely the JDL scripts layer for the grid job submission.
- While it is still possible to submit simple job to grid writing it's own JDL script, the idea is to hide this process to make the grid use more user friendly for the bioinformatics community.



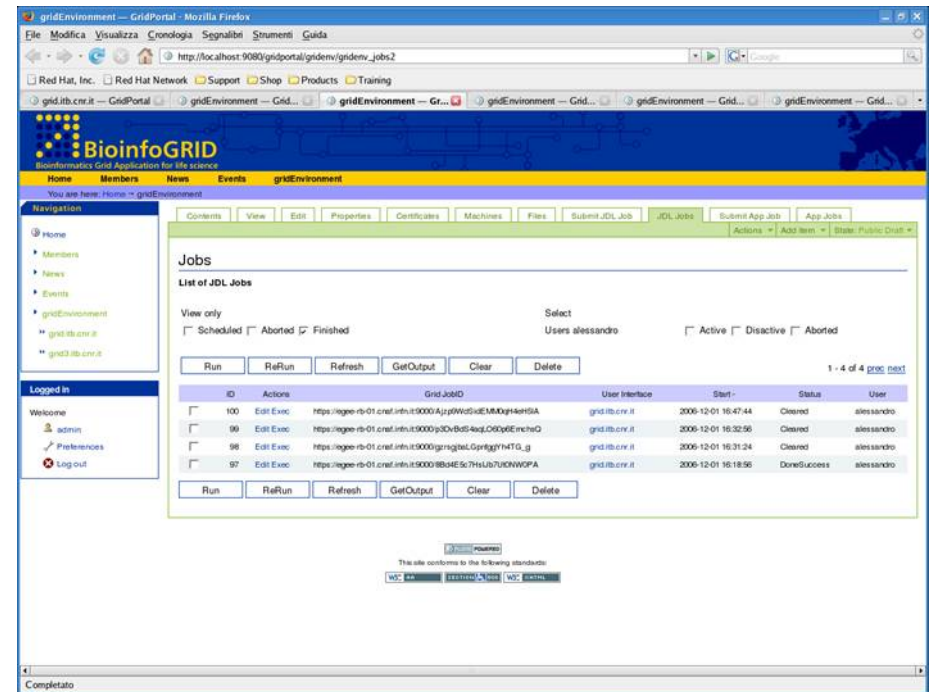


- The interfaces to application jobs are automatically generated by the conversion of XML files that describe both the end user parameters and the structure of the JDL scripts that have to be automatically generated to submit the jobs.



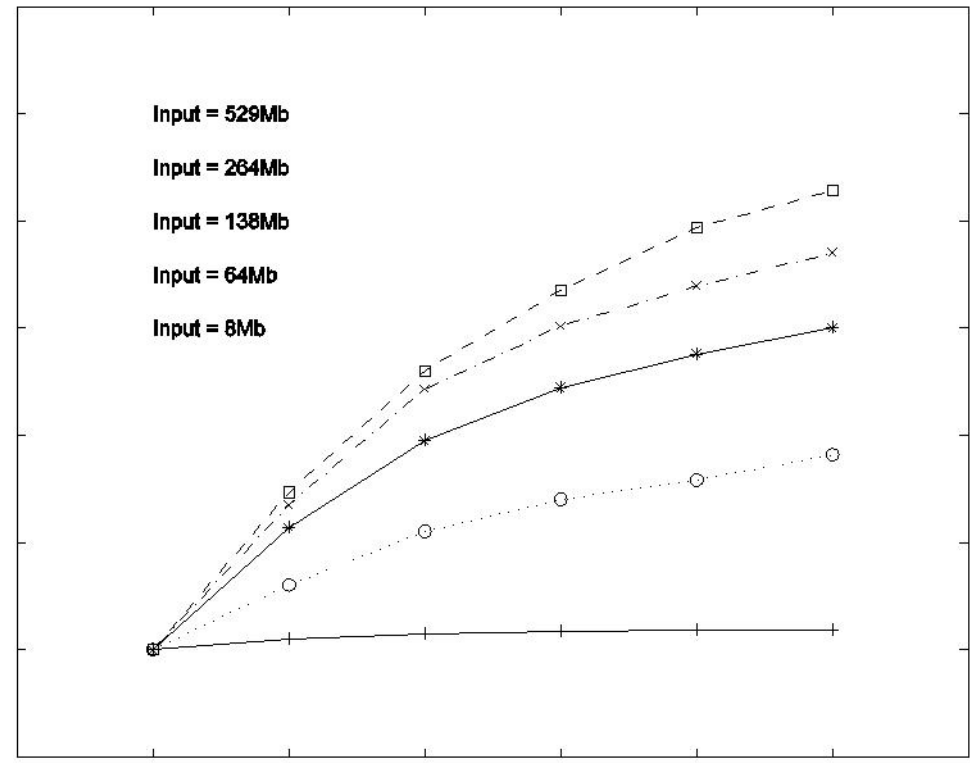


- A selection can be made among different databases against which to perform the analysis: all these databases are updated automatically.
- In figure is reported the summary of the submitted application jobs, with information about the analysis software, the global computation status and the user interface used for submission.





- The main task of the second year of the project is a complete report on the performance and the scalability of Bioinformatics software on the grid platform.
- All the tests performed until now have been studied for design solutions oriented to fully exploiting the grid's resources.





- In this study some grid based applications are presented to compute protein domain analysis in a distributed way.
- This approach has high throughput performance because the protein domains are investigated with different software computed in a distributed way using different grid sites.
- The implemented infrastructure can be used in the content of data intensive challenge of proteomics sequences analysis, according to the resource that the European EGEE platform makes available.



BioinfoGRID

Acknowledgments



- *BIOINFOGRID*
<http://www.bioinfoGRID.eu>

- Alessandro Orro
- Gabriele Trombetti
- Chiara Bishop
- John Hatton
- Luciano Milanese

