Web Services Interface to Run Protein Sequence Tools on Grid, Testcase of Protein Sequence Alignment

Christophe Blanchet, Christophe Combet, Vladimir Daric, and Gilbert Deléage

Institut de Biologie et Chimie des Protéines (IBCP UMR 5086); CNRS; Univ. Lyon 1;
IFR128 BioSciences Lyon-Gerland; 7, passage du Vercors, 69007 Lyon, France
{Christophe.Blanchet,C.Combet,V.Daric, G.Deleage}@ibcp.fr

Abstract. Bioinformatics analysis of data produced by high-throughput biology, for instance genome projects, is one of the major challenges for the next years. Some of the requirements of this analysis are to access up-to-date databanks (of sequences, patterns, 3D structures, etc.) and relevant algorithms (for sequence similarity, multiple alignment, pattern scanning, etc.). GPS@ is a Web portal devoted to bioinformatics applications on the grid (Grid Protein Sequence Analysis, http://gpsa-pbil.ibcp.fr). GPS@ is the grid release of the NPS@ bioinformatics portal, and is wrapping the mechanisms required for submitting bioinformatics analyses on the grid infrastructure. For example, we have put online two multiple alignment Web Services that are submitting the computing job on a remote grid environment. One is accessible through a classical Web interface by using a simple Web browser; the other one can be used through a SOAP and workflow client such as Taverna or Triana. These Web services can process the submitted alignment on two different computing environments: a local and classical one which is a cluster of 30 CPUs, but we are also providing biologists with a large-scale distributed one: the grid platform of the EU-EGEE project (more than 20,000 CPUs available at the European scale).

Keywords: Bioinformatics, Grid computing, Web Services, Protein Sequence Analysis.

1 Introduction

Bioinformatics analysis of data produced by high-throughput biology, for instance genome projects [1], is one of the major challenges for the next years. Some of the requirements of this analysis are to access up-to-date databanks (of sequences, patterns, 3D structures, etc.) and relevant algorithms (for sequence similarity, multiple alignment, pattern scanning, etc.) [2]. There are more and more tools that are put online for molecular biology [3]. But most of these portal are proposing isolated bioinformatics methods, some times several analysis methods, but only few of these Web servers are proposing the integration of several program devoted to molecular Biology. Since 1998, we are developing the Network Protein Sequence Analysis (NPS@) Web server [4], that provides the biologist with many of the most common resources for protein sequence analysis, integrated into several pre-defined and connected workflows.

N. Maglaveras et al. (Eds.): ISBMDA 2006, LNBI 4345, pp. 240 – 249, 2006.

1.1 Related Work: Multiple Sequence Alignment Resources Available Online

Starting in 2003, the Bioinformatics Links Directory [3], [5] is referencing all the online resources, tools and databases devoted to molecular Bioinformatics. This list is curated according to expert recommendations. More than two thousands of links are listed on this molecular bioinformatics Web repository. Among all of them, almost two hundreds are classified as “Sequence comparison” resources, and among these ones, only 39 are furnishing a service for multiple sequence alignments. But all of them are computing the user queries on classical computing resources, local machine or batch cluster, according to their description available on the Bioinformatics Link Directory.

1.2 NPS@, Bioinformatics Web Portal

NPS@ [4] is providing biologist with a Web form to input their data (like protein sequences) in order to run, for example, a BLAST similarity scan against a given protein sequence database, or a multiple alignment of a subset of sequences. In NPS@, user inputs his protein sequences by pasting them in the corresponding field. Then, in case of BLAST, he chooses the database that will be scan with the query sequence. All the protein databases available on NPS@ can be selected through a multi-valued list of the form. These methods and data can be accessed through a classical web browsing and HTTP connection, or through a specialized interface like MPSA [6] or AntheProt [7] programs.

Today, the computing resources available behind the NPS@ Web portal may limit the capabilities put available to the research community. And it is the case also for other genomics and proteomics Web portals. Indeed some methods are very computing-time and memory consuming. Our NPS@ portal is facing an increasing demand of CPU and disk resources and the management of numerous bioinformatics resources (algorithms, databases).

1.3 Testcase: Build Hepatitis C Virus Sequence Alignment

Hepatitis C virus (HCV) causes chronic liver disease in humans, including cirrhosis and hepatocellular carcinoma. The HCV genome shows remarkable sequence variation. Analysis of this variability is essential not only to investigate the correlation between HCV molecular components and diseases expression or antiviral resistance, but also to study the structure-function relationships of these components. To date, more than 40000 HCV sequences are deposited in the generalist databases DDBJ, EMBL, and Genbank [8].

In this testcase, we will consider a common task for bioinformaticians working on Hepatitis C Virus: doing a multiple alignment of sequences issued from different strains. User will upload its own sequence databank (in Pearson/FASTA format) or will extract the sequences from annotated HCV sequence database using a retrieval system (SRS, SQL through a web site). The upload or the query will be done thanks to an integrated Web interface. From the subset of sequences, the user will launch a multiple sequence alignment tools, for example ClustalW [9] or Muscle [10]. A set of tools should be proposed to the user to analyze the computed alignment.

2 Distributed Computing: Web Services and Grid

2.1 Web Services, Weak Connected Concept of Distributed Computing

Web Services (WS) is describing programmatic interfaces that allow different resources, different by location or implementation, to collaborate in a distributed environment base on the Web. Web Services are most of time using three components: WSDL, SOAP and HTTP. Users that will use Web Services will do it through a SOAP client able to create SOAP messages, and able of understanding WSDL file in order to import available WS processors from a remote site.

WSDL. The Web Service Description Language (WSDL) is a language based on XML and aiming to describe Web Services. A WSDL file should define both the resources available in the Web Service, their interfaces and their location (Appendix A). A WSDL file is a directory of several Web Services that could be completely independent and located in different Web places.

SOAP. The Simple Object Access Protocol (SOAP) is a framework that allows describing objects and that they talk together in a distributed and weakly connected Web environment.

HTTP. The Hypertext Transfer Protocol is defining the language and the rules used to exchange messages from clients and servers that are connected to the World-Wide Web (W3). Client are contacting servers, that answer with the same protocol based on messages mainly composed of two parts, a header and a body, containing different command and valued defined with tags.

SOAP client. Taverna [11] and Triana [12] are both combining capabilities of workflow editor, workflow enactor and Web services client, compliant with SOAP and other WS protocols. They provide a visual interface to build simple and complex workflows, providing simple way to link processors available locally or remotely, to datasets or large range of tests. They are both using their own workflow description language, but both based on XML. They are written in Java, and so able to run on most operating systems.

2.2 Grid Computing, Integrated Concept of Distributed Computing

Grid computing concept defines a set of information resources (computers, databases, networks, instruments, etc.) that are integrated to provide users with tools and applications that treat those resources as components within a « virtual » system [13][14][15]. Grid middleware provides the underlying mechanisms necessary to create such systems, including authentication and authorization, resource discovery, network connections, and other kind of components.

The Enabling Grids for E-sciencE project (EGEE [16]), funded by the European Commission, aims to build on recent advances in grid technology and to develop a service grid infrastructure. The EGEE consortium involves 70 leading institutions in 27 countries, federated in regional Grids, with currently a combined capacity of 20,000 CPUs and 5 petabytes of storage. The platform is built on the LCG-2 middleware, inherited from the EDG middleware developed by the European

DataGrid Project [17] (EDG, FP5 2001-2003). The middleware LCG-2 is based upon the Globus toolkit release 2 (GT2) and the Condor middleware [14]. The new middleware gLite [16], that is being developed, have the goals to improve the performances and the services provided by the future EGEE platform.

There are several important components into the EGEE grid: first on the user point of view is the user interface (UI) where the user log in and submit their jobs. These jobs need to be described by JDL files (Job Description Language) with the Condor "ClassAd" formalism. The "workload management system" (WMS) is responsible of the job scheduling on the platform. The scheduler (or "resource broker", RB) analyzes the JDL file and determines where and when to compute a job: (i) using one "computing element" (CE) near one "storage element" (SE) containing the data in case of simple jobs, or (ii) several CEs and SEs in case of larger jobs. A computing element is a gatekeeper to a cluster of several CPUs, the worker nodes (WN) managed by a batch scheduler system. The "information system" that centralize all parameters raised by the grid components (CPUs, storage, network, …).

3 Web Resources for Protein Sequence Analysis on the Grid

3.1 GPS@, Web Server for Grid Protein Sequence Analysis

GPS@ grid Web portal (Grid Protein Sequence Analysis, http://gpsa-pbil.ibcp.fr) is the grid release of the NPS@ bioinformatics portal. GPS@ portal hides the required mechanisms for submitting bioinformatics analyses on the grid infrastructure. Selecting the "EGEE" check-box will schedule the submission of the ClustalW on the EGEE grid when clicking on the "submit" button. The bioinformatics programs and databases available on GPS@ have been distributed and registered on the grid [18], and GPS@ runs its own EGEE interface to the grid [19].

3.2 gBIO-WS, Interfacing WSDL-Compliant Web Services with the GRID

The Grid capability is added to our Web services by the use of the *bio_launcher* tool (Figure 1). We have developed this *bio_launcher* tool to be able to submit remotely a job to the EGEE Grid.

Indeed, in the normal job submission process on EGEE, user has to connect on a special host of the grid: the user interface. Then, he authenticates itself by creating a proxy certificate signed by his own, valid and recognized by EGEE, electronic certificate. Afterwards, he submits and manages his job by using the appropriate command line interface. These are the commands *edg-job-submit*, *edg-job-status*, *edg-job-get-output*, …

Bio_launcher gets, as input, an XML file describing the bioinformatics task to compute: which program to use (ClustalW in this case), the data to process, the user values of program options. This XML description file is written by the Web service according to our gBIO DTD, and filled-in with the data provided by the user. Finally, *bio_launcher* connects on our EGEE user interface through a secure connection (SSH), opens an authenticated connection to the EGEE Grid, submits the job on the Grid, and once the job is finished, gets the results back to the Web service, which forwards them to the user.

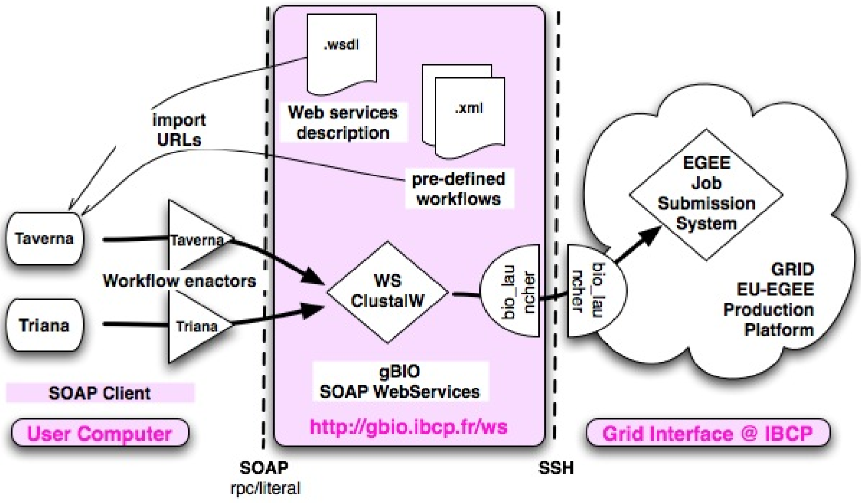


Fig. 1. Architecture of the Bioinformatics Web services at gBIO-WS server, interfaced to the GRID

4 User-Friendly Access to Multiple Alignment Web Services on Grid

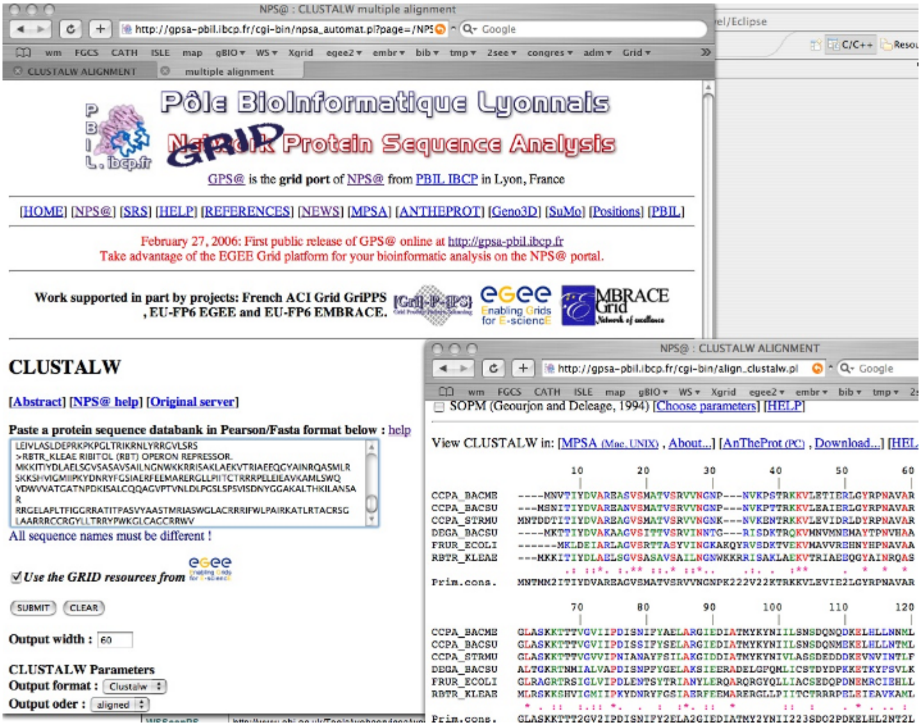
We have put online two multiple alignments Web Services on the CNRS IBCP servers. One is accessible through a classical Web interface, the other one can be used through a SOAP client such as Taverna or Triana, but also a user one built with gSOAP, perl SOAP::Lite or Java.

These Web services can process the submitted alignment on two different computing environments: a local and classical one which is a cluster of 30 CPUs, but we are also providing biologist with an original distributed one: the grid platform of the EU-EGEE project (more than 20,000 CPUs available at the European scale)

4.1 Via a Web Browser

Biologist can access this grid service of multiple sequence alignment through a classical Web page on our Grid Protein Sequence Analysis server (http://gpsa-pbil.ibcp.fr). The protein sequences of HCV can be pasted from the euHCVdb server to the submission form on the GPS@ Web portal (Figure 2). There, the user can chose to process the alignment on our cluster or on the Grid. When the “Grid” checkbox is checked, the multiple alignment is then processed on the EGEE grid platform.

First, the job description in the Web form is converted to a JDL file that can then be submitted to the workload management system of EGEE. The GPS@ sub-process that have submitted the job, is also checking periodically the status of this job by querying the resource broker with the good commands. All steps are notified to the user through the Web page of the submission, indicating the time and the duration of the current step. When achieved, i.e. reaching the “Done” step, the GPS@ automat downloads the result file containing the multiple alignment computed by ClustalW. Then this raw result file in ClustalW format is processed and converted into a HTML page showing, in a colored and graphical way, the list of aligned protein sequences, (Figure 2). This formatting process is directly inherited from the original NPS@ portal, providing biologists with a well-known interface and way of displaying results. After biologist has analyzed the alignment thanks the graphical display in colored shape, he can submit new queries to obtain a modified and better alignment.


NPS@ : CLUSTALW multiple alignment
http://gpsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPS
Google
wm FGCS CATH ISLE map gBIO WS Xgrid egee2 embr bib tmp 2see congres adm Grid
CLUSTALW ALIGNMENT
multiple alignment
el/Eclipse
C/C++
Resou
P B I L. ibcp.fr
Pôle BioInformatique Lyonnais
GRID
Network Protein Sequence Analysis

GPS@ is the grid port of NPS@ from PBIL IBCP in Lyon, France

[HOME] [NPS@] [SRS] [HELP] [REFERENCES] [NEWS] [MPSA] [ANTHEPROT] [Geno3D] [SuMo] [Positions] [PBIL]

February 27, 2006: First public release of GPS@ online at http://gpsa-pbil.ibcp.fr
Take advantage of the EGEE Grid platform for your bioinformatic analysis on the NPS@ portal.

Work supported in part by projects: French ACI Grid GriPPS , EU-FP6 EGEE and EU-FP6 EMBRACE.

CLUSTALW

[Abstract] [NPS@ help] [Original server]

Paste a protein sequence databank in Pearson/Fasta format below : help

```
LEIVLASLDEPRKPKPGLTRIKRNLYRRGVLSRS
>RBTR_KLEAE RIBITOL (RBT) OPERON REPRESSOR
MKKITIYDLAELSGVSASAVSAILNGNWKKRRISAKLAEKVTRIAEEQGYAINRQASMLR
SKKSHVIGMIIPKYDNRYFGSIAERFEEMARERGLLPIITCTRRRPELEIEAVKAMLSWQ
VDWVVATGATNPDKISALCQQAGVPTVNLDLPGSLSPSVISDNYGGAKALTHKILANSA
R
RRGELAPLTFIGGRRATITPASVYAASTMRIASWGLACRRRIFWLPAIRKATLRTACRSG
LAARRRCCRGYLLTRRYPWKGLCAGCRRWV
```

All sequence names must be different !

☑ Use the GRID resources from egee

SUBMIT CLEAR

Output width : 60

CLUSTALW Parameters
Output format : Clustalw
Output oder : aligned

NPS@ : CLUSTALW ALIGNMENT
http://gpsa-pbil.ibcp.fr/cgi-bin/align_clustalw.pl
Google
wm FGCS CATH ISLE map gBIO WS Xgrid egee2 embr bib tmp 2s

□ SOPM (Geourjon and Deleage, 1994) [Choose parameters] [HELP]

View CLUSTALW in: [MPSA (Mac, UNIX) , About...] [AnTheProt (PC) , Download...] [HEL

```
                     10        20        30        40        50        60
                      |         |         |         |         |         |
CCPA_BACME   ----MNVTIYDVAREASVSMATVSRVVNGNP---NVKPSTRKKVLETIERLGYRPNAVAR
CCPA_BACSU   ---MSNITIYDVAREANVSMATVSRVVNGNP---NVKPTTRKKVLEAIERLGYRPNAVAR
CCPA_STRMU   MNTDDTITIYDVAREAGVSMATVSRVVNGNK---NVKENTRKKVLEVIDRLDYRPNAVAR
DEGA_BACSU   ----MKTTIYDVAKAAGVSITTVSRVINNTG---RISDKTRQKVMNVMNEMAYTPNVHAA
FRUR_ECOLI   ------MKLDEIARLAGVSRTTASYVINGKAKQYRVSDKTVEKVMAVVREHNYHPNAVAA
RBTR_KLEAE   ---MKKITIYDLAELSGVSASAVSAILNGNWKKRRISAKLAEKVTRIAEEQGYAINRQAS
                   .: ::*. :.** ::.* ::*..    .:. .  :**      .  *  *  *
Prim.cons.   MNTMM2ITIYDVAREAGVSMATVSRVVNGNPK222V22KTRKKVLEVIE2LGYRPNAVAR

                     70        80        90       100       110       120
                      |         |         |         |         |         |
CCPA_BACME   GLASKKTTTVGVIIPDISNIFYAELARGIEDIATMYKYNIILSNSDQNQDKELHLLNNML
CCPA_BACSU   GLASKKTTTVGVIIPDISSIFYSELARGIEDIATMYKYNIILSNSDQNMEKELHLLNTML
CCPA_STRMU   GLASKKTTTVGVVIPNIANAYFSILAKGIDDIATMYKYNIVLASSDEDDDKEVNVINTLF
DEGA_BACSU   ALTGKRTNMIALVAPDISNPFYGELAKSIEERADELGFQMLICSTDYDPKKETKYFSVLK
FRUR_ECOLI   GLRAGRTRSIGLVIPDLENTSYTRIANYLERQARQRGYQLLIACSEDQPDNEMRCIEHLL
RBTR_KLEAE   MLRSKKSHVIGMIIPKYDNRYFGSIAERFEEMARERGLLPIITCTRRRPELEIEAVKAML
             * .  ::  :.:: *.   .   :   :*. ::   *         ::  :    . * .  .. :
Prim.cons.   GLASKKTTT2GV2IPDISNIFY2ELA2GIEDIATMY2YNII223SDQ2PDKELHL2NT2L
```

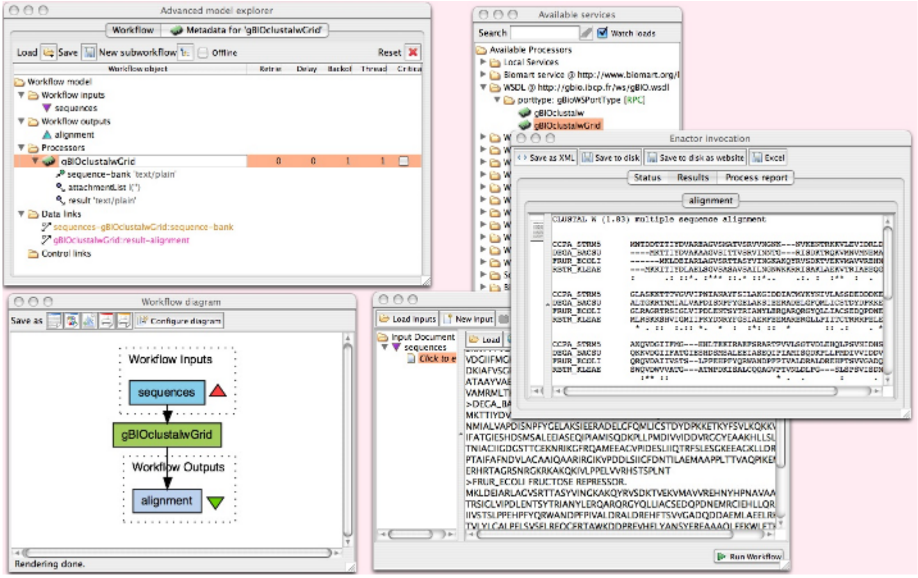
Fig. 2. Multiple alignment of protein sequences obtained through a submission form on the GPS@ Web portal and processed on the EGEE grid platform

4.2 Via a SOAP Client and Workflow Enactor

Biologists and Bioinformaticians can also access this grid service of multiple sequence alignment through Web Services on our Grid Bioinformatics server (http://gbio.ibcp.fr/ws). Our Web services are using standard protocols (SOAP, WSDL and HTTP) and have been built with the gSOAP toolkit, and hosted on Apache HTTP server.

Before to be able to use it, user needs obviously to get the SOAP client which may be Taverna [11] or Triana [12]. We have tested only with these two ones, but other client compatible with WSDL and SOAP standard will certainly be able to connect to our Web services.

User then needs to import our Web service within the Taverna tool (Figure 3). He can do it by importing the WSDL file available at http://gbio.ibcp.fr/ws/gBIO.wsdl (Figure 1). He also has to build a workflow with the HCV sequences as input of the Web service processor, and the multiple alignment as the output downloaded after computing. We are providing two pre-defined workflows to submit an alignment query on our Web service. These pre-defined workflows can be also imported in Taverna directly from the following URLs (Figure 1): http://gbio.ibcp.fr/wf/Clustalw.xml (to import a workflow that process a multiple alignment on our own computing resources), or http://gbio.ibcp.fr/wf/ClustalwGrid.xml (to import a workflow that submit the sequence set on the EGEE grid resources).



```
CLUSTAL W (1.83) multiple sequence alignment


CCPA_STRMS      MNTDDTTIIYDVAREAGVSMATVSRVVNGNK---NVKESTRKKVLEVIDRLD
DEGA_BACSU      ----MKTTIYDVAKAAGVSITTVSRVINNTG---RISDKTRQKVMNVMDEMA
FRUR_ECOLI      ------MKLDEIARLAGVSRTTASYVINGKAKQYRVSDKTVEKVMAVVREHN
RBTR_KLEAE      ---MKKITIYDLAELSGVSASAVSAILNGNWKKRRISAKLAEKVTRIAEEQG
                :        .: ::*. :*** ::.* ::*..     .:. :  :** :  .

CCPA_STRMS      GLASKKTTTVGVVIPNIANAYFSILAKGIDDIAKKYKYNIVLASSDEDDOKE
DEGA_BACSU      ALTGKRTNMIALVAPDISNPFYGELAKSIEERADELGFQMLICSTDYDPKKE
FRUR_ECOLI      GLRAGRTRSIGLVIPDLENTSYTRIANYLERQARQRGYQLLIACSEDQPDNE
RBTR_KLEAE      MLMSKKSHVIGMIIPKYDDRYFGSIAERFEEMAREHGLLPIITCTRKKPELE
                 * . ::  :.:: *.  *  :  :*: ::  *         :: .:   . *

CCPA_STRMS      AKQVDGIIFMG---NHLTEEIRAEFSRARTPVVLSGTVDLENQLPSVNIDNS
DEGA_BACSU      QKKVDGIIFATGIESHDSMSALEEIASEQIPIAMISQDKPLLPMDIVVIDDV
FRUR_ECOLI      QRQVDAIIVSTS--LPPEHPFYQRWANDPFPIVALDRALDREHFTSVVGADQ
RBTR_KLEAE      ENQVDNVVATG---ATNPDKISALCQQAGVPIVSLDLPG---SLSPSVISDN
                  :** ::                          *  .  .        :     .
```

```
VDGIIFMG
DKIAFVSG
ATAAYVAE
VAMRMLT
>DECA_BA
MKTTIYDV
NMIALVAPDISNPFYGELAKSIEERADELGFQMLICSTDYDPKKETKYFSVLKQKKV
IFATGIESHDSMSALEDIASEQIPIAMISQDKPLLPMDIVVIDDVRGGYEAAKHLLSL
TNIACIIGDGSTTGEKNRIKGFRQAMEEAGVPIDESLIIQTRFSLESGKEEAGKLLDR
PTAIFAFNDVLACAAIQAARIRGIKVPDDLSIIGFDNTILAEMAAPPLTTVAQPIKE
ERHRTAGRSNRGKRKAKQKIVLPPELVVRHSTSPLNT
>FRUR_ECOLI FRUCTOSE REPRESSOR.
MKLDEIARLAGVSRTTASYVINGKAKQYRVSDKTVEKVMAVVREHNYHPNAVAA
TRSICLVIPDLENTSYTRIANYLERQARQRGYQLLIACSEDQPDNEMRCIEHLLQR
IIVSTSLPPEHPFYQRWANDPFPIVALDRALDREHFTSVVGADQDDAEMLAEELR
TVLYLCALPELSVSFLREQCERTAWKDDPREVHFLYANSYEREAAAQLFEKWLET
```

Fig. 3. Multiple alignment of protein sequences obtained through a workflow launched in the Taverna tool, submitted to the IBCP Web Services and processed on the EGEE grid platform.

The protein sequences of HCV can be pasted from the euHCVdb server to the submission field on the Taverna tool (Figure 3). There, the user can chose to process the alignment on our cluster or on the Grid. When the "Grid" processor is used, i.e. the "Grid" workflow ClustalwGrid.xml has been imported, the multiple alignment is then processed on the EGEE grid platform. Afterwards, Biologist analyzes the alignment, and submits new queries to obtain a modified and better alignment.

5 Conclusion

GPS@ Web portal and gBIO-WS make the remote access and bioinformatics job submission easier on the grid. We have used, as testcase, the ClustalW multiple alignment tool run on a remote Grid platform, to analyze the variability of a subset of

sequences. The GPS@ portal and gBIO Web services are compliant with standard protocol, guaranty of a good access with common Web browsers and SOAP clients. Biologists can then submit bioinformatics jobs on the Grid by using their usual Web client, but also integrate these grid services within complex workflow combining different databases and tools. They will then benefit from the large-scale computing resources of the Grid, from their usual and local working environment. Grid computing and storage facilities will also permit GPS@ and gBIO services to scale to thousands of daily user as much as aligning complete genomes or proteomes.
Future works will be done about applying this WebServices-to-Grid interface to other programs. We will, for example, work to put online, as Web Services, a selected panel of other protein alignment methods, but also similarity searching programs, like BLAST or SSEARCH, raising the issues of large and numerous databases management in Grid environment [20].

Acknowledgments. This work was supported in part by projects: French ACI Grid GriPPS (ACI GRID PPL02-05), European projects EGEE (EU FP6, INFSO-508833) and EMBRACE (EU FP6, LHSG-CT-2004-512092).

References

1. Bernal, A., Ear, U., Kyrpides, N. : Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. NAR 29 (2001) 126-127
2. G. Perrière, C. Combet, S. Penel, C. Blanchet, J. Thioulouse, C. Geourjon, J. Grassot, C. Charavay, M. Gouy, L. Duret and G. Deléage, Integrated databanks access and sequence/structure analysis services at the PBIL. Nucleic Acids Res., 31:3393-3399, 2003.
3. Fox JA, McMillan S, Ouellette BF. A compilation of molecular biology web servers: 2006 update on the Bioinformatics Links Directory. Nucleic Acids Res 34(Web Server Issue) W3-5. (2006).
4. Combet, C., Blanchet, C., Geourjon, C. et Deléage, G. : NPS@: Network Protein Sequence Analysis. Tibs, 25 (2000) 147-150.
5. Bioinformatics Links Directory. Online at bioinformatics.ubc.ca/resources/links_directory
6. Blanchet, C., Combet, C., Geourjon, C. et Deléage, G. : MPSA: Integrated System for Multiple Protein Sequence Analysis with client/server capabilities. Bioinformatics, 16 (2000) 286-287.
7. Deleage, G, Combet, C, Blanchet, C, Geourjon, C. : ANTHEPROT: an integrated protein sequence analysis software with client/server capabilities. Comput Biol Med., 31 (2001) 259-267
8. Combet C., Penin F., Geourjon C. and Deleage G. HCVDB: Hepatitis C Virus Sequences Database. Appl. Bioinformatics, 2004, 3(4):237-240
9. Thompson, JD, Higgins, DG, Gibson, TJ : CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22 (1994) 4673-4680.
10. Robert C Edgar . MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 2004, 5:113
11. D. Hull, K. Wolstencroft, R. Stevens, C. Goble, M. R. Pocock, P. Li, and T. Oinn. Taverna: a tool for building and running workflows of services. Nucleic Acids Res., July 1, 2006; 34(suppl_2): W729 - W732.

12. I. Taylor, M. Shields, I. Wang, and A. Harrison. Visual Grid Workflow in Triana. In Journal of Grid Computing, 3(3-4):153-169, September 2005.
13. Foster, I. And Kesselman, C. (eds.) : The Grid 2 : Blueprint for a New Computing Infrastructure, (2004).
14. Thain, D., Tannenbaum, T. Livny, M.: Distributed computing in practice: the Condor experience. Concurrency and Computation 17 (2005) 323-356.
15. Vicat-Blanc Primet, P., d'Anfray, P., Blanchet, C., Chanussot, F. : e-Toile : High Performance Grid Middleware. Proceedings of Cluster'2003 (2003).
16. Enabling Grid for E-sciencE (EGEE). Online at www.eu-egee.org
17. European DataGrid project (EDG). Online at www.eu-datagrid.org
18. Blanchet, C., Combet, C. and Deléage, G., Integrating Bioinformatics Resources on the EGEE Grid Platform. ccgrid, p. 48, Sixth IEEE International Symposium on Cluster Computing and the Grid Workshops (CCGRIDW'06), 2006.
19. Blanchet, C., Lefort, V., Combet, C., Deléage, G., GPS@ Bioinformatics Portal: from Network to EGEE Grid. Stud Health Technol Inform. 2006;120:187-93.
20. Desprez, F., Vernois, A., Blanchet, C., Simultaneous Scheduling of Replication and Computation for Bioinformatic Applications on the Grid. ISBMDA 2005: 262-273

Appendix A: gBIO WebServices Description (WSDL)

Description of Web Services for Multiple Sequence Alignment, available on the gBIO-WS server (http://gbio.ibcp.fr/ws), written according to WSDL standard.

```
<?xml version="1.0" encoding="UTF-8"?>
<definitions name="gBioWS"
 targetNamespace="http://gbio.ibcp.fr:8090/gBioWS.wsdl"
 xmlns:tns="http://gbio.ibcp.fr:8090/gBioWS.wsdl"
 xmlns:SOAP-ENV="http://schemas.xmlsoap.org/soap/envelope/"
 xmlns:SOAP-ENC="http://schemas.xmlsoap.org/soap/encoding/"
 xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 xmlns:xsd="http://www.w3.org/2001/XMLSchema"
 xmlns:ns="urn:gbioWS"
 xmlns:SOAP="http://schemas.xmlsoap.org/wsdl/soap/"
 xmlns:MIME="http://schemas.xmlsoap.org/wsdl/mime/"
 xmlns:DIME="http://schemas.xmlsoap.org/ws/2002/04/dime/wsdl/"
 xmlns:WSDL="http://schemas.xmlsoap.org/wsdl/"
 xmlns="http://schemas.xmlsoap.org/wsdl/">

<types>

 <schema targetNamespace="urn:gbioWS"
  xmlns:SOAP-ENV="http://schemas.xmlsoap.org/soap/envelope/"
  xmlns:SOAP-ENC="http://schemas.xmlsoap.org/soap/encoding/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema"
  xmlns:ns="urn:gbioWS"
  xmlns="http://www.w3.org/2001/XMLSchema"
  elementFormDefault="unqualified"
  attributeFormDefault="unqualified">
  <import namespace="http://schemas.xmlsoap.org/soap/encoding/"/>
 </schema>

</types>

<message name="gBIOclustalwRequest">
 <part name="sequence-bank" type="xsd:string"/>
</message>
```

```
<message name="gBIOclustalwResponse">
 <part name="result" type="xsd:string"/>
</message>

<message name="gBIOclustalwGridRequest">
 <part name="sequence-bank" type="xsd:string"/>
</message>

<message name="gBIOclustalwGridResponse">
 <part name="result" type="xsd:string"/>
</message>

<portType name="gBioWSPortType">
 <operation name="gBIOclustalw">
  <documentation>Service definition of function
ns__gBIOclustalw</documentation>
  <input message="tns:gBIOclustalwRequest"/>
  <output message="tns:gBIOclustalwResponse"/>
 </operation>
 <operation name="gBIOclustalwGrid">
  <documentation>Service definition of function
ns__gBIOclustalwGrid</documentation>
  <input message="tns:gBIOclustalwGridRequest"/>
  <output message="tns:gBIOclustalwGridResponse"/>
 </operation>
</portType>

<binding name="gBioWS" type="tns:gBioWSPortType">
 <SOAP:binding style="rpc"
transport="http://schemas.xmlsoap.org/soap/http"/>
 <operation name="gBIOclustalw">
  <SOAP:operation style="rpc" soapAction=""/>
  <input>
     <SOAP:body use="encoded" namespace="urn:gbioWS"
encodingStyle="http://schemas.xmlsoap.org/soap/encoding/"/>
  </input>
  <output>
     <SOAP:body use="encoded" namespace="urn:gbioWS"
encodingStyle="http://schemas.xmlsoap.org/soap/encoding/"/>
  </output>
 </operation>
 <operation name="gBIOclustalwGrid">
  <SOAP:operation style="rpc" soapAction=""/>
  <input>
     <SOAP:body use="encoded" namespace="urn:gbioWS"
encodingStyle="http://schemas.xmlsoap.org/soap/encoding/"/>
  </input>
  <output>
     <SOAP:body use="encoded" namespace="urn:gbioWS"
encodingStyle="http://schemas.xmlsoap.org/soap/encoding/"/>
  </output>
 </operation>
</binding>
<service name="gBioWS">
 <documentation>gSOAP 2.7.8c generated service
definition</documentation>
 <port name="gBioWS" binding="tns:gBioWS">
  <SOAP:address location="http://gbio.ibcp.fr:8090"/>
 </port>
</service>
</definitions>
```