Bioinformatics Workflow using ASSIST on GRID

Ivan Merelli, Giulia Morra, Luciano Milanesi.
Istituto di Tecnologie Biomediche – Consiglio Nazionale delle Ricerche

Abstract

Workflows are very common in bioinformatics because they allow the elaboration of data by delegating the resources management to the information streaming. By the composition of different analysis tools, in fact, it is possible to follow the complex evolution of the biological process.

The Web Services technology is very important in this context because allows the exposition of tools that can be used by the bioinformatics community. The problem is the integration of these services because it presupposes the definition of an opportune semantic to be coordinated.

On the other hand, the enormous quantity of data to be elaborated demands the development of technologies for the management of the computational load. Grid technology allows the optimization of the different steps in the elaboration process, distributing the assignments.

A possible solution to define workflows, that exploit the power of the computational grid and allow the Web Services integration, is the use of a stream oriented programming language like ASSIST to design graph of tasks activated directly by the data flow.

Introduction

Bioinformatics studies complex biological processes in silico through both the analysis of the nucleotide or protein sequences and the study of the macromolecular structures interactions. The central aim is the data elaboration at all the biological levels of the informative streaming to turn the enormous quantity of data in our possession into real knowledge.

The data that undergo the biological mechanism are extremely difficult to interpret because the information stream suffers due to a series of complex transformations before being expressed in protein complexes. The data integration problem becomes of primary importance and always requires great calculation resources.

The use of the Web Services technology [1] allows the integration of very different systems through the definition of correct standards of interoperability both in terms of communication protocols and also from the semantic point of view. This is an enormous effort that will put in the hands of the researchers powerful tools for the analysis of huge amount of data, a problem that arises continually due to the enormous quantity of information that are daily produced.

The exponential growth of the sequence databases, due to the new sequencing technologies, creates an enormous flow of biological information to be elaborated. Such sequences have to be properly investigated and once the coding zones have been translated into protein sequences these have to be analyzed too. While the desire of understanding the biological process becomes deeper, the computational analyses that have to be implemented become more complex.

This huge amount of sequences demands increased necessity of computation power. For this reason a technology like grid [2], 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 distribute the most time-consuming steps of the computation, subdividing the elaboration in a series of small jobs.

Beside the calculation problem, once bioinformatics faces data 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 specific storage element maintaining the coherence among the database replicas [3].

The integration of these resources, the Web Services on one side and the computational grid from the other, is a critical problem for the bioinformatics of the years to come. The proposed solution is a high performance program environment, as ASSIST [4], to access these resources contemporarily.

Related Works

Many web sites offer services of sequence elaboration and a few of them nowadays offer services for structural analyses. Important web sites are certainly those of the big bioinformatics consortia like the NCBI, which developed BLAST [5], and the EBI, that developed INTERPRO [6] a famous analysis system for protein domains.

In recent years some services have been developed to be used directly through the SOAP protocol creating a platform of bioinformatics Web Services. Also the access to modern databases is now guaranteed by clients that can connect through the SOAP protocol to the international repositories.

Even if many bioinformatics Web Services are available, their attainment on the net and integration in complex workflows is not simple. For this reason projects like BioMoby [7] have been developed in order to build a repository of information for the Web Services through the definition of a specific bioinformatics semantic.

The BioMoby project has a workbench to exploit a semantic definition of the different Web Services present in its database. Through the progressive structure offered by S-Moby it is possible for the end user to elaborate step by step the bioinformatics data. For each object, in fact, only the services that can work on it are displayed to allow the user to focus only on the tools that can be useful for him.

This approach gives an interesting view on the bioinformatics workflow generation, but it focuses on small input datasets and can not be scaled on genome range analysis. This problem, instead, is tackled within the MyGrid project [8], that combines the flexibility of Web Services with the power of the computational grid. The services published on the Web, in fact, opportunely listed in a repository, can automatically be invoked on any data dimension thanks to the OGSA infrastructure that affords distributed access to the bioinformatics databases.

The milestone of the MyGRID project is the workflow manager, TAVERNA [9], that allows a graphical definition of the system starting from the information of the resource database. At the same time data can be elaborated in huge amount using the computational resources hidden by the Web interfaces.

The proposed system is an alternative for a workflow definition. The idea is to exploit a grid oriented programming language, like ASSIST, to perform high performance elaboration of software developed ad hoc, maintaining in the meantime the interoperability with the bioinformatics Web Services platform.

This approach has the necessity of informatics skills in the definition of the workflows, but allows the integration of different type of pre-existing code, both Web Services or stand alone, and permits the definition of high performance elaboration on the grid platform to work at genome scale.

The ASSIST Programming Environment

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 in multidisciplinary environments. It includes a skeleton based language and a set of compiling tools and run time libraries. The ensemble allows programs written using ASSIST to be seamlessly run on the top of workstation networks supporting POSIX and ACE [10] and computational grids.

An ASSIST program is a graph in which nodes represent modules or components, and arcs correspond to interfaces and are associated to a directional streaming of data. Streaming allows the composition of modules in a complex program. Modules can be written in different programming language like C/C++, FORTRAN and JAVA, that typically is used to interact with the Web Services platform. Each module can be a parallel module, parmod in ASSIST terminology, or a sequential module and it is possible to reuse a composition of modules as a component of a more complex program.

While the ASSIST graph expresses interaction among program components, the parmod expresses parallelism inside each component in a powerful and effective way. Parallel computation in the parmod is implemented by a set of virtual processors that interact using a topology, which provides a naming scheme for the virtual processors. The internal state of the parmod can be partitioned or replicated among virtual processors. The internal state can hold variables that allow the controlling of communications with input and output streams.

A parmod may have different input streams and through them interact with the rest of the program selecting input with a nondeterministic behaviour similar to that of CSP guarded commands. Moreover each input stream is associated to an independent distribution strategy like on demand, scatter, broadcast, multicast. Results of parmod computation are delivered to other components through parmod output streams.

The compiler works on three basic steps: first syntax form is produced. Then a task code is produced out of the abstract syntax tree. Last POSIX/ACE object code is generated out of the task code which is suitable to be run either on a cluster through the CLAM or on grid through a specific loader interface for the Globus toolkit 2.4 [11]. The object code is actually produced using standard C++ compilers. Along with this code, an XML configuration file is generated, holding all the information needed to map the specialized code to processing nodes.

Implementation

The elaboration of biological data, as apposed to the information present in the databases, is a typical bioinformatics operation. In the proposed case of study a workflow has been designed to integrate the typical aspects of the sequences analysis with the structural aspects of bioinformatics in order to identify what happens to some important informative patterns in the three-dimensional conformation. These three-dimensional analysis steps are certainly the most time-consuming in terms of calculation times, above all when considering aspects like the thermodynamic stability of the complexes or the dynamics of the protein interactions.

The first workflow step consists in the submission of a nucleotide sequence that is elaborated by a specific software called GENSCAN [12]. This software, that can be used remotely on a number of different resources, starts from a nucleotide sequence and checks if a gene is present, in positive case, translates it into the corresponding protein sequence. In particular this tool performs a search in the nucleotide sequence to find out the key components of gene expression: in fact GENSCAN looks for well determined biological elements like some gene promoters or the TATA box pattern. Once the gene in the nucleotide sequence is identified it is translated into the corresponding protein sequence, referring to the correct frame shift.

In a typical analysis it is then important to stress the domains that characterize the protein functionality using specific tools of domains prediction like HMMPFAM [13]. This bioinformatics tools of analysis allows the comparison of a Hidden Markov Models domains dataset, called Pfam, with a proper entry sequence in order to verify if a particular protein pattern is present. This service is exposed, for example, at the EBI and can be used through a simple JAVA client integrated in the workflow.

A protein possesses in general more than a functional domain that, once individualized, are looked for inside a three-dimensional structures dataset. This passage is crucial for this workflow because it allows the correlations of sequence and structural information. To identify the three-dimensional structures of a certain domain the workflow performs a BLAST against the Protein Data Bank [14] sequence database that contains all the proteins of which the atomic coordinates are known by crystallography or by magnetic nuclear resonance. BLAST is accessible in a number of websites through the SOAP protocol and perhaps the NCBI Web Services is the most famous resource. Using BLAST the workflow identifies a series of structures highly correlated with the submitted domain.

The atomic coordinates of all the sequences that introduce a strong correlation with the proposed domain are therefore downloaded from the RCSB site to create a three-dimensional model of the protein structure [15]. The protein model consists in a three-dimensional grid that represents the occupation volumes of all the atoms starting from the information of the nucleuses position.

The macromolecular surface is then extracted from the three-dimensional grid that models the protein. This step is of fundamental importance to understand in a protein domain what amino acids are effectively exposed to the surface and therefore if they are of key importance for the macromolecular functionality. The extraction of the protein surface is executed using a well known algorithm, called Marching Cubes [16], that allows the definition of a support mesh starting from the three-dimensional grid. ASSIST has allowed an high performance implementation of both the modelling phase and the isosurface extraction by the parallelisation on different nodes of each tasks.

According to the three-dimensional grid definition the extracted mesh represents the so called Lee & Richards protein surface. Analyzing each vertex of the surface and looking for the nearest atom it is possible to check which amino acids contribute to the external form of the protein. In this way the information about protein functionality collapses into the definition of a small number of key amino acids. Even in this case ASSIST provided an high performance environment to develop the searching algorithm of the key role amino acids of the surface.

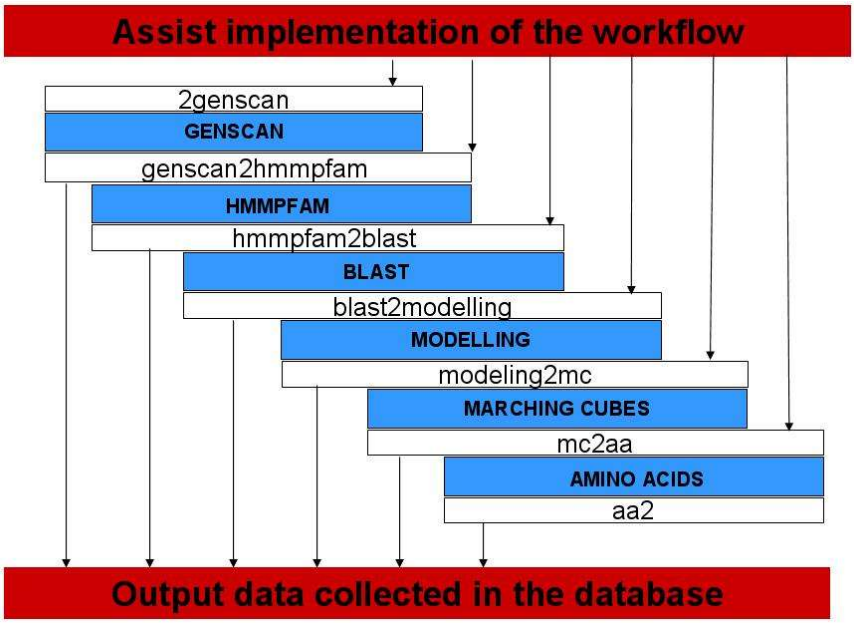

Figure 1. ASSIST graph of the presented case of study
Bioinformatics analysis workflow

Conclusion

ASSIST is a programming environment aimed at providing programmers with a user-friendly, efficient, portable, and fast way of implementing workflows. Using a graphical loader it is possible to configure and execute an ASSIST application on a Globus based grid. It hides to programmers the structure of the grid and provides the interaction between the ASSIST Run Time Support and the Globus Middleware.

Through the advanced programming feature of ASSIST is possible to exploit the software available for the bioinformatics community through Web Services and to integrate it with specific designed code for a workflow analysis. The integration is possible thanks to the multi-language support of ASSIST that can be used both with typical C stand alone applications and with JAVA distributed services. Moreover the possibility to create any kind of graph program allows a great flexibility in the workflow design. In this way it will be possible to produce high performance workflow to be performed on the computational grid.

This technology will allow the implementation of analysis systems at wide range and the integration of classical sequence based elaborations with structural analyses. This is very important in a discipline such as bioinformatics in which researchers typically work with datasets of huge dimensions rather than single input. Bioinformatics needs high performance elaboration of the data both on local clusters and on computational grid: ASSIST allows suitable implementation of powerful workflow on these different calculation solutions.

Aknowledgements

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”

References

[1] Neerincx PB, Leunissen JA: **Evolution of web services in bioinformatics.** *Brief Bioinform.*, 2005 **6**(2):178--188.

[2] Foster I, Kesselman C, Tuecke S: **The Anatomy of the Grid: Enabling Scalable Virtual Organizations.** *International J. Supercomputer Applications*, 2001 **15**(3).

[3] Foster I, Kesselman C, Nick J, Tuecke S: The **Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration, Open Grid Service Infrastructure WG.** *Global Grid Forum,* 2002.

[4] Aldinucci M, Campa S, Ciullo P, Coppola M, Danelutto M, Pesciullesi P, Ravazzolo R, Torquati M, Vanneschi M, ZoccoloC: **ASSISST demo: a high level, high performance, portable, structured parallel programming environment at work.** *Proceedings Euro-PAR*, 2003 712 –713.

[5] Altschul S: **Amino Acid Substitution Matrices from an Information Theoretic Perspective.** *J. Mol. Biol* 1991, **219**:555--565.

[6] Mulder N, Apweiler R, Attwood T, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P, Bucher P, Copley R, Courcelle E, Das U, Durbin R, Falquet L, Fleischmann W, Griffiths-Jones S, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lopez R, Letunic I, Lonsdale D, Silventoinen V, Orchard S, Pagni M, Peyruc D, Ponting C, Selengut J, Servant F, Sigrist C, Vaughan R, Zdobnov E: **The InterPro Database, 2003 brings increased coverage and new features.** *Nucleic Acids Res.,* 2003 **31**:315--318.

[7] Wilkinson MD, Links M: **BioMOBY: an open source biological web services proposal**. *Brief Bioinform.,* 2002 **3**(4):331--341.

[8] Stevens R, Robinson A, Goble CA: **myGrid: Personalised Bioinformatics on the Information Grid.** *Bioinformatics,* 2003 **19**:302--304.

[9] Oinn T, Addis M, Ferris J, Marvin D, Senger M, Greenwood M, Carver T, Glover K, Pocock MR, Wipat A, Li P: **Taverna: a tool for the composition and enactment of bioinformatics workflows.** *Bioinformatics,* 2004 **20**(17):3045--3054.

[10] Schmidt DC: **The ADAPTIVE Communication Environment: Object-Oriented Network Programming Components for Developing Client/Server Applications**. *Proceedings Sun Users Group Conference,* 1993.

[11] Foster I, Kesselman C: **Globus: A Metacomputing Infrastructure Toolkit**. *Intl J. SupercomputerApplications,* 1997 **11**(2):115--128.

[12] Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA**. *J. Mol. Biol.,* 1997 **268**:78--94.

[13] Hugley R, Krogh A: **Hidden Markov models for sequence analysis: Extension and analysis of the basic method.** *CABIOS,* 1996 **12**(2):95--107.

[14] Yang H, Guranovic V, Dutta S, Feng Z, Berman H, Westbrook J: **Automated and accurate deposition of structures solved by X-ray diffraction to the Protein Data Bank.** Biological Crystallography, 2004 **60**(10): 1833--1839.

[15] Ma B, Elkayam T, Wolfson H, Nussinov R: **Protein-protein interactions: Structurally conserved residues distinguish between binding sites and exposed protein surfaces.** *ComputProc Natl Acad Sci USA,* 2003 **100**(10):5772--5777.

[16] Merelli I, Milanesi L, D'Agostino D, Clematis A, Vanneschi M, Danelutto M: **Using Parallel Isosurface Extraction in Superficial Molecular Modeling.** Proceedings *DFMA - IEEE Computer Society*, 2005 79--84.