

BAAQ: A Grid-based Infrastructure for Integrating Bioinformatics Applications

Xiujun Gong¹
gong@apr.jaeri.go.jp

Kei Yura²
yura@apr.jaeri.go.jp

Nobuhiro Go^{1,2,3}
go@apr.jaeri.go.jp

¹ Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara, 630-0192 Japan

² Center for Promotion of Computational Science and Engineering, Japan Atomic Energy Research Institute, 8-1 Kizu, Souraku, Kyoto, 619-0215 Japan

³ Neutron Science Research Center, Japan Atomic Energy Research Institute, 8-1 Kizu, Souraku, Kyoto, 619-0215 Japan

Keywords: Grid, Active Service Provider, Wrapped Program Toolkits, Metadata

1 Introduction

Amount of available biological data is increasing at explosive speed. These huge data are heterogeneous, autonomous and dynamic. At the same time, many tools to analyze these data are produced by different research groups. Integration of these biological data and tools is becoming one of the major topics in bioinformatics community^{[1][2]}. One of the goals for the integration is to enable users to access the up-to-date biological data across multiple heterogeneous and distributed databases and analysis tools. The integration effort for bioinformatics application faces some requirements: uniform access to biological data sources, overviews for bioinformatics tools and their interface standardization, access to high performance and distributed computing resources, knowledge sharing and discovery.

To meet these requirements, we propose an integrated infrastructure for bioinformatics applications, which we named “Bioinformatics: Ask Any Questions (BAAQ)”. Via grid based computation (STA: Seamless Thinking Aid) and graphically-enabled editor environment (Task Mapping Editor)^[5], we emphasize how to compose a robust workflow to perform a complex bioinformatics analysis and to visualize results by Active Service Provider and Wrapped Program Toolkits. We also propose a metadata model to organize biological resources and to share and discover the knowledge in grid community.

2 Architecture overview

As a bioinformatics collaboration platform, the goal of BAAQ is to create and organize a repository of biological data, bioinformatics tools and analysis workflows, to recommend workflow-based solutions automatically or semi-automatically and to allocate grid-based computing resource to facilitate rapid bioinformatics analysis. From the conceptual view of the grid, BAAQ components fall into four-layer infrastructure, as shown in Fig 1.

- Computation resource layer provides grid-based computing environment that are responsible for authentication, authorization, communication and computing resource location and allocation.
- Information service layer consists of grid middleware such as TME, TextBrowser, and PluginTool. These middleware modules interact with user by GUI interface and communicate with STA by corresponding adaptors. TME is the core component for managing icons and building workflows.
- Application resource layer consists of Wrapped Program Toolkits (WPT) and Active Service Provider (ASP). WPT is a series of wrapped program tools to perform bioinformatics analysis. ASP is used to provide guidelines for choosing the biological data and bioinformatics tools hosted in grid environment and to aid composing workflows.
- Meta crawler layer collects and updates the data sources and maintains the consistency of the whole resources. The metadata management module is also responsible for the transportation of workflows between different grid communities.

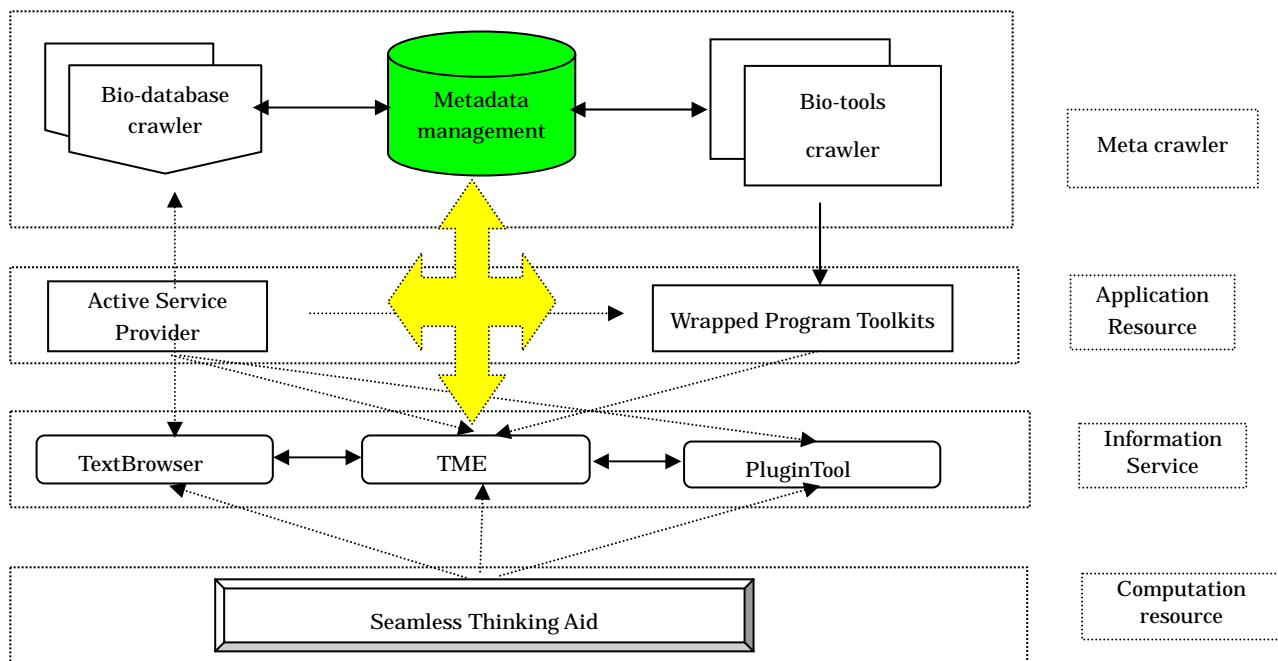


Fig 1 four-layer architecture

3 Discussions

Using knowledge rich metadata, BAAQ currently collects and manages amount of repositories of biological databases, bioinformatics tools and analysis workflow created by users. One of our targets for next research is to explore and discover knowledge from these repositories. For composing a workflow smoothly and visualizing the analysis results, we design a filter library, which is used to integrate biological data and wrap bioinformatics tools based on uniform XML format. To improve the model of our XML schema is another main task of the future work.

References

- [1] Achard, F., Vaysseix, and G., Barillot, E. "XML, Bioinformatics and Data Integration". *Bioinformatics*, Vol.17, No. 2, pp115-125, 2001.
- [2] Cannataro, M., Comito, C., Lo Schiavo F., and Veltri P. "Proteus, a Grid based Problem Solving Environment for Bioinformatics: Architecture and Experiments". *The IEEE Computational Intelligence Bulletin*, Vol. 3 No.1, pp7-18, 2004.
- [3] Gil, Y., Deelman, E., Blythe, J., Kesselman, C., and Tangmurarunkit, H. "Artificial Intelligence and Grids: Workflow Planning and Beyond". *IEEE Intelligent Systems*, Vol. 19, No.1, pp26-33, 2004.
- [4] Stevens, R.D., Robinson, A.J., and Goble, C.A. "myGrid: Personalised Bioinformatics on the Information Grid". *Bioinformatics*, Vol. 19, Suppl. 1, pp302-304, 2003.
- [5] Toshiyuki, I., Nobuhiro, Y., Hiroshi T., Yukihiro H., Kenji H., and Norihiro, N. "A Visual Resource Integration Environment for Distributed Applications on the ITBL System". *ISHPC 2003*, pp258-268, 2003.