Integration of Bioinformatics Data Sources Using Proteus[1]

Parikshit Pol, Mihail Bota, Shahram Ghandeharizadeh, Larry Swanson
Departments of Computer Science and Biological Sciences
University of Southern California
{pol,mbota,shahram,swanson}@usc.edu

Key Words: distributed execution, data integration, what-oriented framework.

Purpose and Motivation

MANY different information sources expose both their functionalities and their descriptions. Web Services (WSs) provide a standard XML-based specification for this purpose. They are building blocks of a "what"-oriented framework such as Proteus to empower scientists to specify "what" is needed using a query, freeing them from the details of "how" to process the query. This empowers scientists to focus on their scientific research instead of integrating information from discovering different data sources on the Internet and integrating them. Once a scientist submits a query, Proteus identifies the relevant WSs, composes a plan that describes how the WSs should be invoked, and executes the plan to obtain results. This paper demonstrates Proteus for a bio-informatic application.

Methods

Two primary components of the Proteus framework are query plan composition and execution. During composition, Proteus discovers the relevant WSs and integrates them into a plan. Proteus executes this plan in either a centralized or distributed manner. With the distributed execution paradigm different physical data and computational services are used. All the services are assumed to be autonomous components. This paradigm provides greater flexibility on privacy of data. Proteus employs the relational data model to include each WS as a set-oriented operator that consumes either an element or a set to produce a set as its output.

Implementation

We demonstrate Proteus for data integration using the following query: "What are the input and output connections for Caudoputamen, and what are definitions of the Caudoputamen and closely related terms?" We targeted this query for the "rat" species, and in the "swanson-1998" nomenclature. A plan is generated for this query, employing two bioinformatics data sources to process our target query – BAMS and UMLS. The Brain Architecture Management System (http://brancusi.usc.edu/bkms) is an online domain specific system that handles biological data from different species and across several levels of organization of Central Nervous System: cell types, brain regions, networks of brain regions and gene representational data. UMLS is the medical standard developed by the National Library of Medicine and contains a repository of biomedical information. We use UMLS to get the synonyms for the given structure specified as input to the plan.

[1] Supported in part by an unrestricted cash gift from Microsoft research and NSF grant IIS-0307908.