Collection-Oriented Scientific Workflows for Integrating and Analyzing Biological Data*

Timothy McPhillips[1], Shawn Bowers[1], and Bertram Ludäscher[1,2]

[1] UC Davis Genome Center, University of California, Davis
[2] Department of Computer Science, University of California, Davis
{tmcphillips, sbowers, ludaesch}@ucdavis.edu

Abstract. Steps in scientific workflows often generate collections of results, causing the data flowing through workflows to become increasingly nested. Because conventional workflow components (or actors) typically operate on simple or application-specific data types, additional actors often are required to manage these nested data collections. As a result, conventional workflows become increasingly complex as data becomes more nested. This paper describes a new paradigm for developing scientific workflows that transparently manages nested data collections. Collection-oriented workflows have a number of advantages over conventional approaches including simpler workflow designs (*e.g.*, requiring fewer actors and control-flow constructs) that are invariant under changes in data nesting. Our implementation within the Kepler scientific workflow system enables the explicit representation of collections and collection schemas, concurrent operation over collection contents via multi-level pipeline parallelism, and allows collection-aware actors to be composed readily from conventional actors.

1 Introduction

Scientists today require access to data from diverse sources. Nowhere is this need more pressing than in the life sciences, where multiplying databases and rapidly growing data repositories promise to provide researchers with a wealth of information relevant to the systems they study. Effectively exploiting diverse sources of data requires a spectrum of data integration approaches.

In the database community, data integration traditionally means resolving different data structures that represent fundamentally the *same* kind of information [11]. This information may be stored using heterogeneous schemas, and may use different representations for data values (*e.g.*, for identifying objects). In such cases, data integration involves determining mappings between source schemas, and then transforming these schemas into a common schema and corresponding integrated data set that can be used for some other purpose. These mappings and transformations typically represent logically necessary relationships between different data sources.

In contrast, data integration in the life sciences often entails applying fundamentally *different* kinds of information to answer scientific questions, make discoveries,

* Work supported in part by SciDAC/SDM (DE-FC02-01ER25486), NSF/SEEK (DBI-0533368), and NSF/GEON (EAR-0225673).

U. Leser, F. Naumann, and B. Eckman (Eds.): DILS 2006, LNBI 4075, pp. 248–263, 2006.

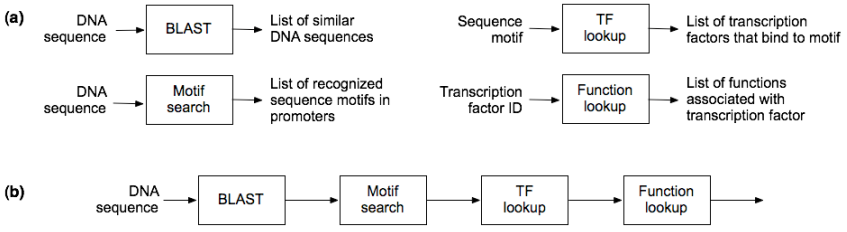

Fig. 1. Scientific workflow components frequently produce lists of results: (a) typical bioinformatics components; and (b) a hypothetical workflow composed from these components that leads to increasingly nested data collections

and test theories. Such scientific data integration procedures necessarily invoke scientific theories that cannot be inferred from schemas or data alone. For example, consider a systematist who wishes to use both genomic sequence data and morphological data in the process of inferring the evolutionary relationships among organisms. Instead of simply mapping DNA sequences and morphological data into a uniform data format, different processes may be applied to each data source to infer evolutionary (*i.e.*, phylogenetic) trees. The systematist then may use the assumption that the organisms have only one true set of evolutionary relationships, and that the phylogenetic trees inferred from genomic and morphological data approximate the true relationships. By employing this theory, the researcher may "integrate" these distinct data sources by computing a consensus tree that reflects commonalities in the distinct phylogenetic trees inferred from the different data sources. These consensus trees (*i.e.*, the resulting data product of integration) can then be analyzed further or applied in other studies.

The challenge of integrating life-science data from multiple sources becomes even more daunting as disciplines become increasingly specialized and as more diverse types of scientific data are desired. Scientific workflow systems [12,13,15,20,22,4] aim at facilitating these types of integration and analysis.[1] However, current scientific workflow systems still offer little or no support for effectively managing (and hiding) the inherent complexity of life-science data, leading to overly complex workflows that are hard to create, reuse, and optimize.

As shown in Figure 1, scientific workflow components (or actors) frequently generate lists of results. When carried out one after the other, such operations naturally yield increasingly nested collections of data that must be managed during workflow execution. This situation is further complicated by the fact that the steps in such workflows in general operate on different nesting levels. For example, a query of a database mapping sequence motifs to known transcription factors might take a single motif as an input, while the operation upstream of this step in the workflow might generate a list of motifs to operate upon. Similarly, the collection of all transcription factors associated with a number of different sequence motifs might be required as input to a downstream component. As these examples demonstrate, scientific workflows must be able to

[1] Figure 4 shows an implementation of a workflow for inferring and analyzing phylogenetic trees using the Kepler system.