Microbase2.0: A Generic Framework for Computationally Intensive Bioinformatics Workflows in the Cloud

Keith Flanagan[1], Sirintra Nakjang[1,2], Jennifer Hallinan[1], Colin Harwood[2], Robert P. Hirt[2], Matthew R. Pocock[1], Anil Wipat[1]

[1] School of Computing Science and [2]Institute for Cell and Molecular Biosciences, Newcastle University, UK

Introduction

As bioinformatics datasets grow ever larger, and analyses become increasingly complex, there is a need for data handling infrastructures to keep pace with developing technology. One solution is to apply Grid and Cloud technologies to address the computational requirements of analysing high throughput datasets. We present an approach for writing new, or wrapping existing applications, and a reference implementation of a framework, Microbase2.0, for executing those applications using Grid and Cloud technologies.

We used Microbase2.0 to develop an automated Cloud-based bioinformatics workflow executing simultaneously on two different Amazon EC2 data centres and the Newcastle University Condor Grid. Several CPU years' worth of computational work was performed by this system in less than two months. The workflow produced a detailed dataset characterising the cellular localisation of 3,021,490 proteins from 867 taxa, including bacteria, archaea and unicellular eukaryotes.

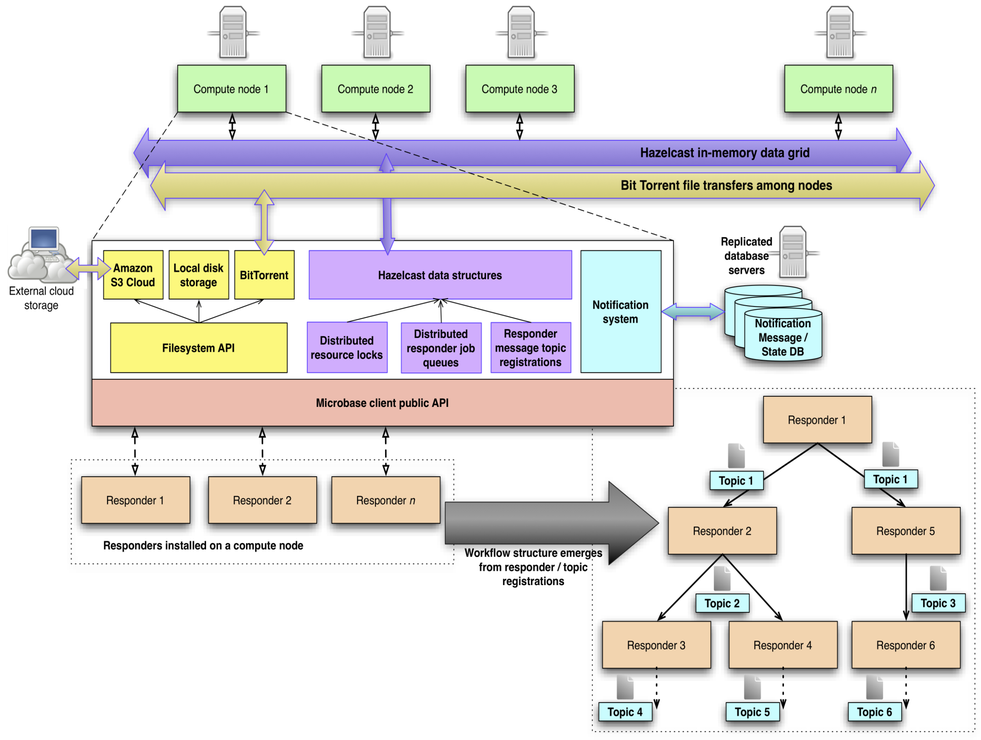
A distributed computing approach

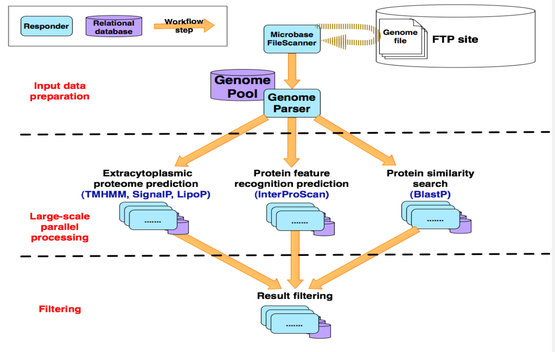
Microbase2.0 is a distributed computing bioinformatics framework, consisting of a set of independent, loosely coupled services that co-operate to provide the infrastructure required by Grid- or Cloud-based analysis workflows. These components are:

- A notification system facilitating de-coupled communication between workflow components;
- The filesystem: a scalable, distributed file store;
- A distributed process manager;
- Domain-specific application components (termed responders): user-written components that either perform an analysis, or delegate a task to an existing analysis program.

Responders can be written *ab initio*, or built around existing bioinformatics tools. New responders can be added on-the-fly, and have full access to all previous results, meaning that no re-computation is required.

Microbase architecture





A set of Microbase2.0 responders was developed to encapsulate bioinformatics tools for the identification of putative targeting signals, cell surface anchoring regions and homology.

The workflow:

- Automatically retrieves sequence data from an FTP site;
- Generates appropriately formatted input data for the downstream analysis processes;
- Processes the input sequence via various bioinformatics tools; and
- Extracts a list of candidate extracytoplasmic proteins based on the prediction results.

Predicting extracytoplasmic proteins

The extracytoplasmic proteome is likely to be important to the phenotype of an organism, as it mediates many primary aspects of its environment. We developed an analysis workflow using Microbase2.0, incorporating multiple targeting-signal prediction tools to identify extracellular proteins and domains.

Organism	Total proteins	Predicted Extracytoplasmic
Gram +	693,402	214,955
Gram -	1,922,673	665,194
Microbial eukaryote	272,389	67,121
Archaea	133,026	34,499
Total	3,021,490	981,769

The performance of the workflow was assessed using an experimentally-verified dataset.

Organism	PPV (%)	Sensitivity (%)
Gram+	90.84	85.56
Gram-	95.24	88.73
Archaea	100.00	90.67

Conclusions

Microbase2.0 executed the extracellular protein prediction pipeline over 3,021,490 protein sequences using six different bioinformatics tools in parallel. This task, which would have taken four years on a two-CPU desktop machine, was completed in one month using between 10 and 100 computers, depending upon availability, on a Condor grid and the Amazon cloud.

Microbase2.0 is a powerful and flexible tool for the analysis of large amounts of biological data.