

Microbase

Abstract

A microbial genome sequence can reveal a great deal about the biology of an organism, but the comparative analysis of microbial genomes allows this rich source of biological data to be most efficiently exploited. To date (July 2003) there are approximately 130 complete genome sequences and the rate at which microbial genomes are being sequenced is increasing rapidly. Soon the volume of data will put comparative analyses beyond the capability of the computing resources of most individual laboratories. Microbase aims to establish an object based information repository that will support a variety of microbial genome comparison algorithms. Microbase will develop a novel architecture that is both scalable and able to support user defined, remotely conceived, computationally intensive algorithms through the use of distributed computing technology and integration with The Grid.

Aims and Background

Tools that are effective for the comparison of microbial genomes are being continually refined. However, as more genome sequences are published, there is an increasing need to develop effective solutions for the storage and querying of the increasing number of genomes that will be available in the near future. Methods are also required that will allow genome comparison results to be interpreted in the context of a range of other biologically relevant data (e.g. gene expression, protein function, protein interactions, metabolic pathways, virulence, environmental niche and taxonomy). All-against-all genome comparisons, in particular, are computationally intensive. Even at the current rate at which new sequences are entering the databases, all-against-all genome comparisons are becoming increasingly difficult to perform locally using the resources available to the majority of researchers. Moreover, rapid developments in DNA sequencing technology promise to speed up whole genome sequencing to a rate far exceeding our current capabilities, threatening to release a flood of new genomes into the public domain. It is essential that new systems for genome comparison are scalable to cope with the influx of new genomes.

Faced with the prospect of such large datasets, it is

important that a well-structured but flexible data model is employed for storing the outcomes of genome comparisons. Currently, most comparison databases are based on relational systems. However, genome sequence features are well described by an object-oriented (OO) model and the results of protein comparisons lend themselves very naturally to being modelled as object graphs. An object database management system (ODBMS) promises to provide **a rich information model** that is better able to capture and query the many-to-many relationships between compared objects and incorporate them into the model in a dynamic fashion. Inevitably, such a database could be expected to contain a huge number of objects and could not be effectively implemented using a single site ODBMS. Simply making an object dataset available over the Internet is not sufficient to allow users to perform their own computations. Computations running on users' local machines are restricted by the low bandwidth connection of the Internet and the size of the dataset.

Recently, work at Newcastle has addressed the design of architecture to support computation on very large datasets stored in object-oriented databases.

The design has been termed an Active Information Repository (AIR) (Fig.1). The AIR architecture provides a means by which OO data may be queried, processed or mined using computationally intensive analyses without the necessity to download huge datasets to the remote user's machine. Computations to be performed on the persistent object data are sent in the form of mobile code from a remote client and are executed in a high performance environment, close to the data. Thus, an AIR comprises two major components: an object database, and an execution server. Both the execution server and the object database server exploit parallelism to achieve scalability: each runs on a set of nodes interconnected by a high performance network.

The AIR architecture is being developed specifically as a scalable node for The Grid.

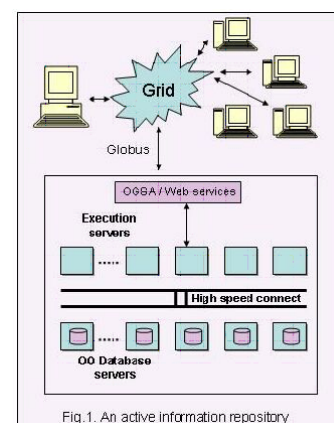


Fig.1. An active information repository

The AIR architecture is intended to act as a storage and processing node, specifically designed to address the problem of processing of complex queries against large datasets, which higher-speed networking alone cannot solve efficiently.

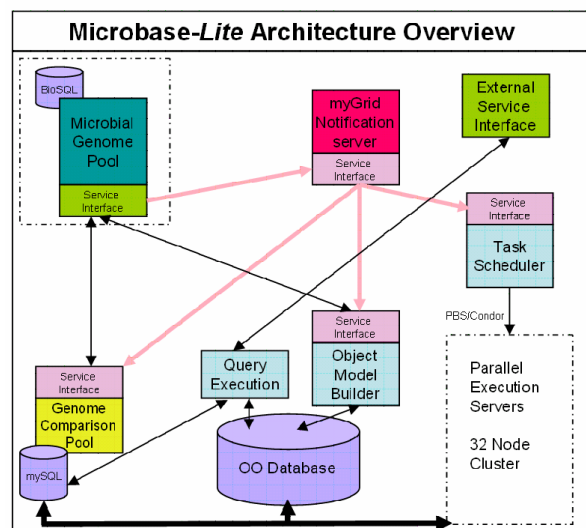
Microbase-Lite

Microbase-Lite is a microbial genome comparison system being developed using mostly conventional technology that will provide a tool for biologists to browse and analyse pre-compared microbial genomes. Microbase-Lite will provide an intermediate solution for service based system for microbial genome comparison that will help to establish user requirements, build up use cases for the project, develop specifications for the AIR architecture and develop the schema for the object oriented database. The preliminary architecture of the Microbase-Lite system is shown in figure 2:

The Microbase-Lite system comprises a number of distinct components that communicate through the use of web-service interfaces. These will be upgraded to Grid-service interfaces as the system evolves.

A standalone 'Microbial Genome Pool' has been designed and implemented. It provides a local source of 120 complete microbial genome sequences and has a web service interface that provides access to complete or subsections of microbial genome sequences and their features. The Microbial genome pool is now available for use by external users that require notified, computational access to microbial genome sequences. A description of the functionality and service interface for the genome pool can be found at <http://vindaloo.ncl.ac.uk:8090/genomepool/index.html>.

An implementation of the myGrid notification system is used to coordinate the processing of new genomes by the Microbase-Lite system as they arrive in the Genome Pool. New genomes are compared to each other at the nucleotide and protein sequence level using a variety of algorithms including Blast and suffix tree based sequence comparisons. In addition, a number of novel algorithms are employed to provide orthologue and paralogue determination, protein family classification and an object based description of genome rearrangements (e.g. Inversions, insertions, deletions and translocations etc.) between genomes. A task scheduler is used to farm processor intensive tasks to



a parallel cluster using a combination of Condor and PBS to manage the jobs. Comparison results are stored in both relational and object databases.

A graphical genome comparison viewer in the form of a Java application is under development and will act as a client to allow a biologist to access the Microbase-Lite system. This client will be made freely available for academic use and will provide a means by which a user can remotely browse genomes, view genome and gene comparison data and formulate their own queries for execution within the system.

Summary

The Microbase project started in April 2003 and will finish in August 2006. We hope to release Microbase-Lite for use by the biological community in autumn 2003. Overall, the Microbase project aims to combine the biological and computational expertise at Newcastle to provide systems that will **support the biological and bioinformatics community**, by allowing them to perform their own microbial genome comparison studies in a remote high performance environment. Such systems will not only store and return their experimental results but will ultimately provide seamless access to additional analytical services and data sources over the Grid.



Dr. Anil Wipat



Prof. Pete Lee



Prof. Paul Watson



Dr. Yudong Sun

Mr. Keith Flanagan

Mr. Jake Wu