Exploring Microbial Genome Sequences to Identify Protein Families on the Grid

Yudong Sun, *Member, IEEE*, Anil Wipat, Matthew Pocock, Peter A. Lee, Keith Flanagan, and James T. Worthington

***Abstract*—The analysis of microbial genome sequences can identify protein families that provide potential drug targets for new antibiotics. With the rapid accumulation of newly sequenced genomes, this analysis has become a computationally intensive and data-intensive problem. This paper describes the development of a Web-service-enabled, component-based, architecture to support the large-scale comparative analysis of complete microbial genome sequences and the subsequent identification of orthologues and protein families (Microbase). The system is coordinated through the use of Web-service-based notifications and integrates distributed computing resources together with genomic databases to realize all-against-all comparisons for a large volume of genome sequences and to present the data in a computationally amenable format through a Web service interface. We demonstrate the use of the system in searching for orthologues and candidate protein families, which ultimately could lead to the identification of potential therapeutic targets.**

***Index Terms*—Genome analysis, grid, microbial genomes, protein families, Web services.**

I. Introduction

DEVELOPMENTS in comparative genomics have been helping to provide novel techniques for therapeutic antimicrobial drug discovery. The comparative analysis of complete microbial genome sequences can identify unique proteins and homologous protein families conserved in and between genomes, which can be screened in the search for new antibiotic targets [1]–[3]. The approach promises to enhance our capability to develop antibiotics to tackle the increasing risks of infectious diseases in humans that include the emergence of new bacterial pathogens, the spread of epidemic diseases, and the intensified resistance to existing antibiotics [2].

With the rapid increase of completed microbial genome sequences, the comparative analysis of whole microbial genomes has become a computationally intensive and data-intensive problem. For example, whole sequence alignment and homology search involve enormous computations over a huge volume of genomic datasets. Grid computing can federate distributed resources using open, general-purpose protocols to create a powerful computing system that meets end-user requirements of on-demand access to computing capabilities [4], [5] and promises to provide a solution to the highly increasing computational demand in biology, biomedicine, and bioinformatics [6]–[8]. Web services and service-oriented architecture are important principles and technologies in the implementation of the grid and the exposure of such resources as services to end-users [4].

The Microbase project[1] has developed a grid-based system to service the timely dynamic or on-demand comparative analysis of microbial genome sequences in biological and biomedical research. We employ Web-service-based technologies, in particular a Web-service-based notification system, to integrate distributed components and orchestrate their interoperability. Consequently, the system is able to perform large-scale genome comparison and analysis, using a variety of bioinformatics tools, and expose our precomputed dataset of comparison results to users across the Internet. The precomputed dataset provides a flexible data repository of genome sequence similarities that will support a number of subsequent analyses, both by a human user and by a computational client, such as a workflow. A Web-service-based client interface has been implemented to facilitate computational access to the data repository. Microbase enables biological and biomedical researchers to carry out customized analyses directly without having to repeat the computationally intensive genome comparisons. In order to demonstrate the utility of the system, we have used this precomputed dataset to discover and define protein families. The protein families we identify may aid in the discovery of new therapeutic agents and in the development of new antibiotics by highlighting proteins that are conserved in bacteria and may form suitable targets. A protein family that is conserved amongst a phylogenic group of bacteria can be viewed as a potential target for broad-spectrum antibiotics, whereas a protein unique to a specific pathogenic bacterium can be considered as the target of a narrow-spectrum drug.

Our preliminary version of the Microbase system—termed as Microbase*Lite*—has been implemented and integrates computing servers, a database server, and a campus grid. The system has been used to execute all-against-all comparisons, using different tools, for 250 microbial genomes, mainly bacteria and archaea. Genomic comparisons are automatically carried out in response to Web-service-based notification messages triggered as new genome sequences are deposited in remote public genome databases. Two algorithms have been implemented to search the 250 genomes for putative orthologues and clusters of orthologous groups (COGs). The system has been developed with

[1]Microbase project. [Online]. Available: http://www.microbase.org.uk

Manuscript received October 18, 2005; revised July 23, 2006. This work was supported in part by the U.K. Biotechnology and Biological Sciences Research Council (BBSRC) e-Science and Bioinformatics initiative and in part by the DTI under Grant 13/BEP17027.

Y. Sun was with Newcastle University, Newcastle Upon Tyne, NE1 7RU, U.K. He is now with the Oxford University Computing Laboratory, Oxford University, Oxford, OX1 3QD, U.K. (e-mail: yudong.sun@comlab.ox.ac.uk).

A. Wipat, M. Pocock, P. A. Lee, and K. Flanagan are with the School of Computing Science, Newcastle University, Newcastle Upon Tyne, NE1 7RU, U.K. (e-mail: anil.wipat@ncl.ac.uk; matthew.pocock@ncl.ac.uk; p.a.lee@ncl.ac.uk; keith.flanagan@ncl.ac.uk).

J. T. Worthington was with Newcastle University, Newcastle Upon Tyne, NE1 7RU, U.K. He is now with Convergys Corporation, London, SW1V 0HW, U.K. (e-mail: j.t.worthington@blueyonder.co.uk).

Digital Object Identifier 10.1109/TITB.2007.892913

Web-service-based user accessibility as a prime concern. Web-service-based interfaces including an application programming interface (API) and a graphical viewer have been developed to allow end-users to retrieve genome sequences, precomputed comparison results, and protein families via the Web.

A full overview of the project will be presented in the future. In Section II, we discuss previous work related to the project and the novelty of our approach. The Microbase*Lite* system is outlined in Section III. Section IV discusses the identification of protein families based on the system, and finally, our ideas for future work are covered in Section V.

II. Related Work

Grid computing is increasingly being employed in biological and biomedical research and in particular to support genome comparison and analysis. For example, Genome Analysis Research Environment (GNARE) [8] is a scalable grid-based system using Globus, Condor, and GriPhyN virtual data system and running on the grid systems as GRID2003, TeraGrid, and the Department of Energy's (DOE) Science Grid to automate genome analysis (including data acquisition from genome databases), mainly using basic local alignment search tool (BLAST). The Institute for Genomic Research (TIGR)'s distributed computing environment (DCE)[2] is an institutional grid system that connects the on-campus computers and database servers with Sun Grid Engine (SGE). Genome analysis is implemented on demand using BLAST, MUMmer, and HMMsearch, and a repository of protein and nucleotide sequence data and a protein database of all-versus-all searches to identify whether protein similarity is maintained or not. The grid protein sequence analysis (GPSA)[3] Web portal provides a user interface to run protein sequence analyses, including BLAST, FASTA, SSEARCH, and ClusterW, on the European Enabling Grids for E-sciencE (EGEE) Grid.[4] The PUMA2 system also provides a flexible system for grid-based, high-throughput analysis of genome sequences, based on the use of the GADU system, leveraging experience gained from the GriPhyN physics project [9].

While the aforementioned projects concentrate on grid-enabled gene and protein sequence comparison, there are far fewer projects providing data from analyses that employ sequence comparison data. One such application area is the identification of orthologous genes and their grouping into protein families. The identification of orthologues is an important application of comparative genomics that seeks to establish relationships between similar proteins and genes from different genomes for subsequent evolutionary and functional studies. The COG database,[5] [10] contains the clusters of orthologous proteins identified from different phylogenetic lineages and has become widely accepted for the annotation of proteins. The *coli*BASE,[6] [11] is a database of *Escherichia coli*, *Shigella*, and *Salmonella*, reflecting the full diversity of *E. coli* and its relatives, which includes the putative orthologues found in these genomes. The e-Fungi project [7] has performed homologue analysis for fungal genomes using protein BLAST (BLASTP) and has employed a Markov chain clustering (MCL) method to cluster protein families for phylogenetic and pathogenic analysis.

Drug discovery is also an emerging application area of grid computing. For example, myGrid [12] is a service-based grid middleware framework to manage the complex process of life-science research. myGrid supports data management, new discovery notification, and provenance management in the drug discovery process, in particular, through the use of e-Science workflows [13]. In addition, the EGEE Grid has a project for the virtual screening of a large amount of data to find potential drugs for infectious diseases such as malaria.[7]

In comparison, the Microbase project supports the computational analysis of microbial genomes, particularly bacteria. The project provides a grid-based system that possesses a distributed component infrastructure to integrate grid computing, remote public genome databases, with a precomputed, dynamically updated genome similarity dataset useful to biological and biomedical researches. We build on the approaches introduced by similar projects, extending them with a focus on the ease of computational access to the datasets and by introducing flexibility in the analyses available for these data. Unlike many existing systems, the individual components of the system are amenable to deployment in a distributed computing environment. Intercomponent communication and interoperability is facilitated via Web services, so theoretically, individual components may be located disparately, and in multiple instances, as long as network communication is maintained. The mechanism behind intercomponent communication is based on Web-service-based notifications, which are used to orchestrate the behavior of the system. In particular, notification facilitates the auto-update of the precomputed dataset to import and process new genomes from public genome databases as they are released. The system also provides a greater range of all-against-all genome comparison results at both nucleotide and protein levels than many existing systems. We provide tools including BLASTP, nucleotide BLAST (BLASTN), and the suffix-tree-based algorithms MUMmer and PROmer, but additional tools can easily be incorporated as the need arises.

The results generated provide a flexible dataset to support different, custom genomic analyses, in particular, those that underpin the successful identification of therapeutic targets such as protein family identification. Web-service-based client interfaces including a documented service API and a graphical genome viewer maximize the ease of access for end-users to the precomputed dataset. Users can directly undertake user-specified analyses without having to redo the time-consuming genome comparisons that usually exceed the computing capability of a single institute. A Web-service-based infrastructure also facilitates the enactment of workflows to manage subsequent

[2]TIGR grid computing. [Online]. Available: http://www.tigr.org/grid/

[3]GPSA: Grid protein sequence analysis. [Online]. Available: http://gpsa.ibcp.fr/

[4]EGEE. [Online]. Available: http://public.eu-egee.org/

[5]COGs. [Online]. Available: http://www.ncbi.nlm.nih.gov/COG/

[6]coliBASE. [Online]. Available: http://colibase.bham.ac.uk/

[7]EGEE battles malaria with grid wisdom. [Online]. Available: http://public.eu-egee.org/news/fullstory.php?news_id=53

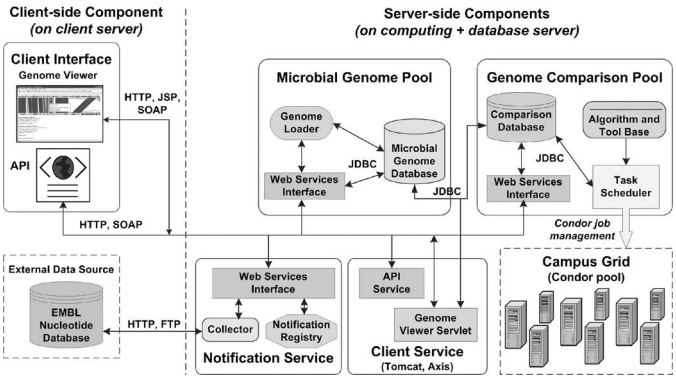

Fig. 1. Microbase*Lite* architecture.

sophisticated genome analysis processes. In this paper, we have used our precomputed genome comparison dataset, to establish orthologues and protein families from 250 proteomes, (to our knowledge, a greater range than the current versions of the related tools discussed earlier) in order to identify potential therapeutic protein targets.

III. MICROBASE*Lite*

Microbase*Lite* is the initial system implementation of the project. As shown in Fig. 1, Microbase*Lite* consists of distinct components for computation, data acquisition and database management, user access, and component orchestration. The components can be deployed on distributed servers and orchestrated via Web-service-based notification message-passing mechanisms. The main components include the microbial genome pool, the genome comparison pool, the notification service, and the client interface.

A. Microbial Genome Pool

The microbial genome pool provides an up-to-date database of complete microbial genome sequences. Genomes published in the public European Molecular Biology Laboratory (EMBL) nucleotide database[8] are imported into a local microbial genome database and subsequently compared to all other genomes in the repository. Automatic updates of the database are triggered by the Web-service-based *notification service*, where the collector component is deployed to regularly check for new microbial genomes in the EMBL database. When a new genome is available, the collector notifies the genome loader in this pool to download the new genome into the local database. More details of the notification service are presented in Section III-C.

To facilitate end-user access, the microbial genome pool uses BioJava[9] to parse the plain-text genome sequence records obtained from the EMBL database and enters the sequences and their annotations into the microbial genome MySQL database with a BioSQL schema.[10] At the time of writing, the microbial genome pool holds 250 microbial genome sequences.

Web-service-based client interfaces have been developed to allow end-users to flexibly access the genome sequences in the microbial genome database. A Java API to the database has been exposed as Web services and a graphical genome viewer application developed to visualize the data. The API has been implemented using Apache Tomcat and Codehaus XFire, a service-oriented simple object access protocol (SOAP) implementation.[11] Users can retrieve DNA and protein sequences, gene features (e.g., coding sequence (CDS), tRNA, and mRNA), and annotations (e.g., a sequence's ID, organism species, and references) using this interface. The genome viewer enables end-users to browse the genome sequences in a graphical format using a Web browser and has been developed using JavaServer Pages (JSP) and Java Servlet and deployed under Tomcat. The API and genome viewer are connected to the microbial genome database via Java Database Connectivity (JDBC). Details of the client graphical user interface are presented in Section III-D.

B. Genome Comparison Pool

The genome comparison pool is a central component responsible for conducting genome comparison and analysis within the system and for maintaining the precomputed comparison dataset.

The genome comparison pool performs pairwise sequence comparisons using existing tools to establish sequence similarities at nucleotide and protein sequence levels, both for whole genomes and individual genes. Currently, four sequence comparison tools are applied: BLASTP, BLASTN, MUMmer, and PROmer. BLASTP[12] is a protein–protein comparison tool that searches similar proteins. BLASTN[12] is a pairwise nucleotide alignment tool to find similar nucleotide fragments. MUMmer[13] is a suffix-tree-based fast tool for nucleotide alignment that gives a concise report on similar nucleotide sequence fragments. PROmer[13] is a variant of MUMmer that translates two nucleotide sequences into amino acids in all six frames, finds all matches in the amino acid sequences, and then maps the matches back to the positions in the nucleotide sequences.

Since our alignments are nonreciprocal statistical computations, we compare each member of a pair of genomes against each other (i.e., genome A is compared against B and genome B against A). In total, the all-against-all comparison of 250 microbial genomes requires 62 500 pairwise comparisons of genome sequences. The comparison of complete genome sequences and their encoded proteins is usually a computationally intensive task. For example, the BLASTP comparison of two species of *Bacillus*: *Bacillus anthracis* and *Bacillus cereus* (approximately 5500 proteins each) takes 12 min on a 2.8-GHz CPU and produces 95 MB of output data. The BLASTN comparison of two *Leptospira interrogans* genomes (approximately 4.3-M base pairs each) takes over 8 h and produces 193-MB data. Using the four comparison tools, the all-against-all comparison

[8]EMBL. [Online]. Available: http://www.ebi.ac.uk/embl/index.html
[9]BioJava. [Online]. Available: http://www.biojava.org/
[10]BioSQL. [Online]. Available: http://www.biosql.org/
[11]Codehaus XFire. [Online]. Available: http://xfire.codehaus.org/
[12]NCBI BLAST. [Online]. Available: http://www.ncbi.nlm.nih.gov/BLAST/
[13]MUMmer 3. [Online]. Available: http://mummer.sourceforge.net/

is an overloaded task that exceeds the capability of common computing systems.

To cope with this burden, the genome comparison pool utilizes a grid-based system to support all-against-all genome comparison. The system is based on a campus grid consisting of computing clusters distributed within different laboratories of our university. The clusters are federated into a powerful computing environment with Condor to handle large applications that overburden any single cluster. The access to the campus grid is authenticated by the user accounts in the Condor system. To manage the execution of the large number of pairwise comparisons, the genome comparison pool provides a *task scheduler* that calls the Condor job-management mechanism to realize parallel execution of the comparison jobs on the grid. Running on a computing server, the task scheduler creates an individual job for each comparison and submits the job to the Condor job queue. Condor, in turn, allocates the job to execute on an idle node. The task scheduler can consecutively submit a large number of jobs to run in parallel depending on the available computer nodes in the grid. Meanwhile, the task scheduler is also responsible for managing the pace of job submission to keep the overall workload at a reasonable level. The task scheduler uses a Condor command to check the status of each job. Once a running job has finished, a new job will be submitted to run. Therefore, the task scheduler along with the underlying Condor can progressively dispatch the large number of jobs for execution and yet maintain load balancing to prevent the system from being overloaded by excessive queuing jobs. The task scheduler is implemented in Java and can use different grid middleware. For example, the task scheduler can also be executed on SGE by calling the job submission and status-checking commands of SGE.

A comparison job also parses and loads its results into the comparison database when a comparison has finished. The comparison database is a MySQL database, which stores all pairwise comparison results. The Web-service-based API and genome viewer (discussed earlier) also support access to the comparison database. Since many techniques in genome analyses are based on sequence similarities, the comparison database provides an instantly accessible, base-level precomputed dataset that enables biological and bioinformatics researchers to directly implement in-depth genome analyses without having to redo time-consuming genome comparisons. In Section IV, we illustrate this concept by using these results to define protein families for all genomes within the database.

At present, 250-against-250 genome comparisons have been completed on the grid system, and the comparison database has reached 28 GB. Fig. 2 shows the execution time of all-against-all comparisons on a selected number of microbial genomes. The execution time includes the parallel execution time of all pairwise comparisons using the four tools, and the time for parsing and loading the results into the comparison database. The execution time is decreased by exploiting the CPUs available on the campus grid. However, the centralized database ultimately becomes a bottleneck to the overall performance when additional CPUs are used. We intend to address this limitation in future versions by deploying a decentralized database on multiple servers.

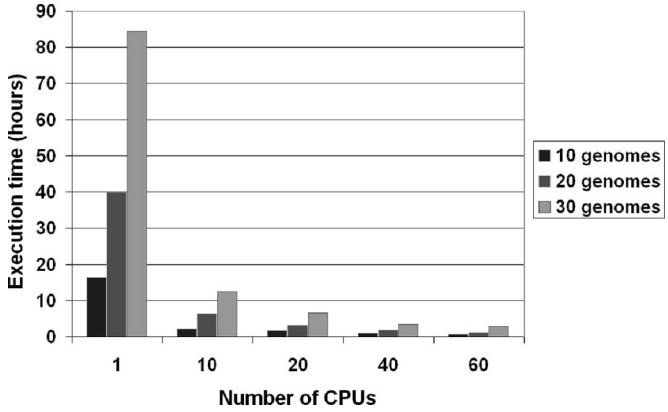

Fig. 2. Execution time of all-against-all genome comparison on the grid.

Once a pairwise comparison result has been generated, it is reusable in further analyses. Sixty-eight hours were required to run the comparisons of 165 genomes on 40 CPUs and to populate the results into the comparison database. Subsequently, the database was regularly updated to import the data from new genomes. When a new genome is imported, it is compared with each of the existing genomes in our database using the four tools, and the database is updated with the new comparison results. The update process is automatically invoked by the notification service, and the comparisons are also executed on the grid system with the support of the task scheduler.

C. Notification Service

The notification service is based on the myGrid notification system [14].[14] The myGrid notification system is a Web-service-based event notification that supports topic-based message publishing and subscribing. Subscribers can receive notification messages on a registered topic in push-and-pull models. A subscriber can be a real user or a software component. The push model delivers a notification by calling back to the client code deployed at a Web service endpoint. Microbase*Lite*'s notification service utilizes the push model to notify subscribed components about the arrival of new genomes. The microbial genome pool relies on this notification service to trigger the updates to the local microbial genome database. A collector is deployed in the notification service to monitor a remote genome repository (the EMBL database). When a new microbial genome is published, the collector will push a specific notification to the microbial genome pool, which in turn requests the new genome sequence from the collector. The collector then downloads the genome file from the remote repository via file transfer protocol (FTP), forwards it to the microbial genome pool, which parses the genome file, and then loads the sequence into the microbial genome database. Subsequently, the microbial genome pool sends a notification to the genome comparison pool that triggers the task scheduler to start the comparison of the new

[14]myGrid project. [Online]. Available: http://www.mygrid.org.uk/

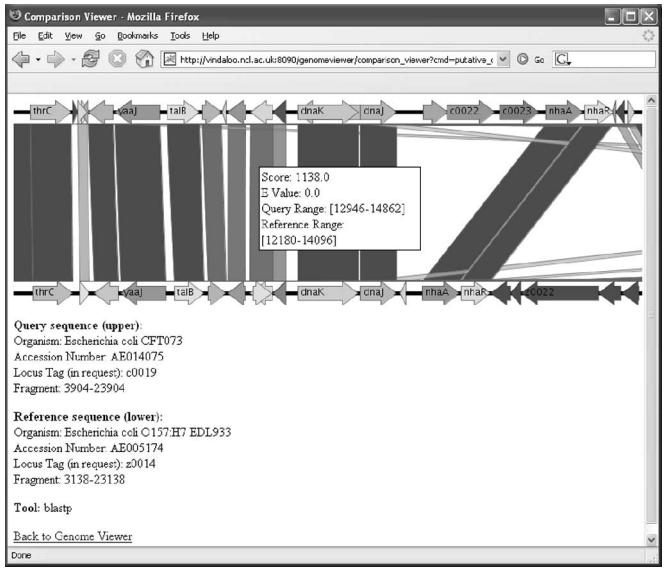

Fig. 3. Genome viewer showing BLASTP alignment between *Escherichia coli* CFT073 (upper) and *Escherichia coli* O157:H7 EDL933 (lower) fragments. The popup window shows the details of the hit between the dnaK genes of two sequences.

genome against all existing genomes to update the precomputed dataset. The notification service also allows human users to receive notification of new genomes. A notification message will be sent to the subscribers when a new genome has been added to the database. The notification service is implemented using Apache Tomcat and Codehaus XFire and uses a MySQL database to store the registered publishers, subscribers, and notification messages.

D. Client Interface

The client interface exposes the microbial genome database, the precomputed comparison database, and the protein families for external users to access over the Internet. The client interface includes a Java API and a genome viewer. The API provides various methods that can be called in user programs to retrieve the genome sequences and the comparison results, by connecting to the Web services deployed on the server side under Apache Tomcat and Codehaus XFire. The genome viewer uses JSP and servlet and is also deployed under Apache Tomcat. The genome viewer allows Web users to browse the genome sequences and comparison results, and to search the protein families via a Web browser such as Internet Explorer and Mozilla Firefox. In Fig. 3, the genome viewer shows the BLASTP alignment between *Escherichia coli* CFT073 (up) and *Escherichia coli* O157:H7 EDL933 (down). The strips in between highlight the hits of similar fragments between the two sequences. Hovering over a region using the mouse will pop up a window showing the detail of a hit or a gene.

IV. Protein Family Analysis

A protein family is a group of similar proteins that are related through evolution. Proteins directly related to each other through evolutionary processes are called homologues and can be further classified as orthologues and paralogues. Paralogues are homologous proteins in the same genome. Orthologues are homologous proteins in different genomes that evolved from a common ancestral gene. Orthologues often retain the same function in the process of evolution. Thus, orthologue search is an effective method to predict the evolutionary relationships and infer the functions of a group of genes or proteins [10], [15], [16]. Identifying orthologues can also aid in the drug discovery process. For example, when seeking to develop a new broad-spectrum antibiotic, it is useful to identify potential protein drug targets that are conserved in the target species, but not present in humans or higher eukaryotes. In this respect, groups of orthologues that are unique to bacteria can be considered as potential targets for new broad-spectrum antibiotics.

As a proof of principle for the Microbase system, we have used the precomputed BLASTP results from Microbase*Lite* to carry out orthologue searches providing a starting point for users wishing to identify conserved proteins as drug targets. We also extend this analysis to group orthologous genes into families, which may also provide useful information in this respect. In order to accelerate the identification of orthologues and protein families in such a large dataset, we parallelized the algorithms described as follows.

A. Putative Orthologues

Putative orthologues are defined as the proteins that have mutual best hits in the BLASTP comparison and satisfy specific requirements on the aligned portions. We use the same criteria of putative orthologues specified by *coli*BASE, [11]. Since orthologues reflect the evolutionary relationships of the genes that encode those proteins, for convenience, we use the terms "protein" and "gene" interchangeably when referring to orthologues in the following discussion. The search for putative orthologues begins by selecting the mutual best hits from the BLASTP results.

Definition 1: Given protein α from genome A and protein β from genome B (A and B are different genomes), α is the *best hit* to β, if the hit has the highest bit score and the lowest E-value in all BLASTP hits between α and any proteins of genome B. The hit between α and β is a *mutual best hit* if α is the best hit to β and β is also the best hit to α.

The mutual best hit means that α and β are the most similar proteins in all proteins between genome A and B, as defined by BLASTP. The evolutionary and functional relationships between the similar proteins, and therefore, the genes that encode the proteins can be inferred based on the mutual best hits that are also putative orthologues that are defined as follows.

Definition 2: If the mutual best hit between protein α and β satisfies two conditions on the aligned portion as follows, α and β are *putative orthologues*.

1) α and β have at least 80% amino acid identity.
2) The aligned portion covers at least 90% of the shorter sequence.

With the aforementioned definitions, the search for putative orthologues can be achieved in three steps.

1) Select the best hits in all BLASTP hits of each protein of a genome against the proteins from each of other genomes.
2) Identify mutual best hits among the best hits.
3) Check the amino acid identity and alignment coverage of the mutual best hits to determine the putative orthologues that satisfy the two conditions as in definition 2.

Microbase*Lite* has a dataset of 6 46 954 proteins from 250 genomes. Pairwise BLASTP comparisons have reported more than 400 million hits with a total size of 22 GB. A parallelized search was implemented to identify the putative orthologues among this huge number of hits. Running on eight 2.8-GHz CPUs, the search was completed in ten days (compared to in excess of two months if run on a single CPU). Additional CPUs have not been used because the search is data-intensive and restricted by the speed of the database server, and therefore, using more CPUs does not improve the performance. (This problem will be solved by deploying a decentralized database on distributed servers that can improve the parallel search.)

The number of putative orthologues found by the search depends on the specified values of cutoff conditions. Using the conditions in definition 2, the search found putative orthologues for 2 87 490 proteins. This represents 44.4% of the total proteins in our database. Among these proteins, most of them have more than one putative orthologues each. However, some genes are conserved in very limited numbers of organisms. There are 98 206 proteins that have only one putative orthologue. For example, the gene BH14430 (locus tag) of *Bartonella henselae* str. Houston-1 (an agent of cat scratch fever and bacillary angiomatosis) has only one putative orthologue, the gene BQ11380 (locus tag) of *Bartonella quintana* str. Toulouse (an agent of trench fever, bacillary angiomatosis, and bacteremia).

The comparison database in the genome comparison pool has also been populated with the putative orthologues defined using this approach. Users can search for orthologues of a given gene via the genome viewer.

B. COGs

COGs are a classification of homologous protein families [10], [15]. A COG is composed of orthologous proteins or orthologous groups of paralogous proteins from three or more genomes. A COGs search identifies both orthologous proteins from different genomes and paralogous proteins from the same genomes. The paralogues from a genome are collected into a group that is treated as a single candidate orthologue in the search for COGs. Putative orthologues only reflect one-to-many relationships of the proteins; nevertheless, COGs can reveal more comprehensive, many-to-many relationships amongst the proteins from the same and different genomes.

Our search for COGs is based on the same set of mutual best hits obtained in the putative orthologues search. However, the COGs search does not set any cutoff requirement on aligned portions. In addition, the COGs search needs to identify all paralogues that are the mutual best hits from the same genomes. Our COGs search includes the following steps based on the COG construction procedure from the COG database project[5], [10], [15].

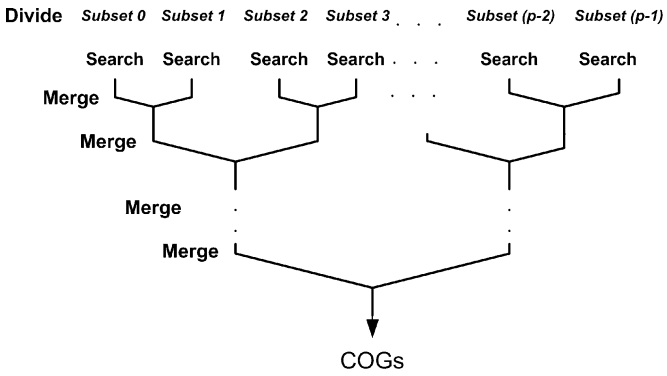

Fig. 4. Divide-and-conquer method for parallel COGs search. The set of proteins is split into p subsets. The search of three orthologue groups runs on each subset per CPU, followed by log p rounds of merge.

1) Identify best hits and mutual best hits from all BLASTP hits (available from the putative orthologues search).
2) Find paralogues in each genome and group them.
3) Search the groups of three orthologues in the mutual best hits. Given three proteins α, β, and γ, the proteins form a group of three orthologues if (α, β), (β, γ), and (α, γ) are mutual best hits. A group of paralogues is regarded as a single orthologue in the formation of the groups.
4) Merge the groups that have at least a common mutual best hit, if the merge will not put the proteins from the same genome (except those are paralogues) into a group.
5) COGs have finally been formed if the groups cannot be further merged.

The process involves an exhaustive search of the groups containing three orthologues (triples) and then an iterative merge of the groups that have common mutual best hits. This is an extremely compute- and data-intensive process. In order to establish a fast and parallel implementation of the COGs search, a *divide-and-conquer* method is used. As shown in Fig. 4, the divide-and-conquer method consists of three phases.

1) *Divide:* divide the whole protein set into p subsets.
2) *Search:* search the groups of three orthologues for the proteins in each subset based on the BLASTP hits, and perform an initial merge of the groups. This phase can be run in parallel on p CPUs.
3) *Merge:* merge the groups of orthologues from different subsets in log p rounds. Round i runs on $p/2^i$ CPUs $(i = 1, 2, \ldots, \log p)$ to merge the groups of orthologues between adjacent CPUs. The complete COGs are formed in the final round, which is run on one CPU.

In the search of three-orthologue groups (α, β, γ), only the starting point α is selected from the corresponding subset. Its orthologues β and γ can come from other protein subsets. Therefore, the parallel search can find the same groups of orthologues as a sequential search.

The COGs search over all proteins of the 250 microbial genomes took 30 days on eight 2.8-GHz CPUs excluding the time for filtering the mutual best hits, which are already

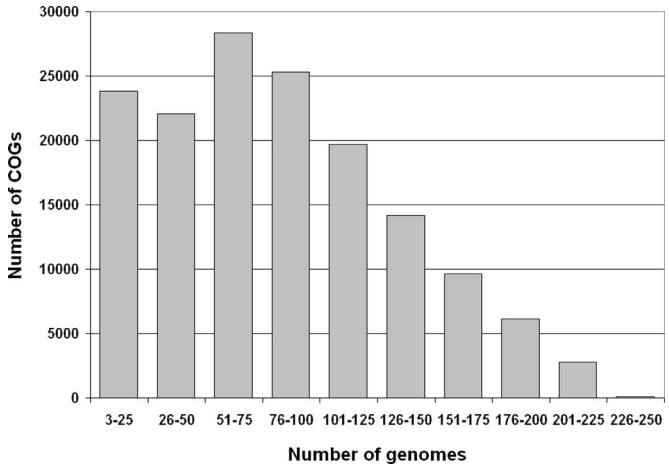

Fig. 5. Composition of the COGs in terms of the number of distinct genomes in each COG. For example, there are around 24 000 COGs each containing the orthologues contributed from 3 to 25 distinct genomes.

available. We estimate that the sequential search would take in excess of 200 days if running on a single CPU. Due to the more intensive search on the table of mutual best hits and the table of intermediate orthologous groups, the centralized database restricts the performance improvement of COGs search when using more than eight CPUs. This problem will also be solved when a decentralized database is deployed in the future.

Our COGs search identified 1 52 011 COGs containing 5 46 699 orthologues, of which 5 31 441 are single proteins and 15 258 are groups of paralogues. In total, 5 71 701 proteins were assigned to one or more COGs, representing 88.37% of all proteins from the 250 genomes. Also, 18 455 groups of paralogues have been found, which consist of 47 608 proteins. Fig. 5 shows the composition of the COGs in terms of the number of distinct genomes contributing to each COG, giving an indication of the degree of conservation of COGs across a range of genomes. The results demonstrate that a large number of COGs span between 50 and 75 genomes, and hence, appear to be well conserved. Around 3000 COGs contain members from 200 genomes, but far fewer span greater than 225 genomes.

As a COG is formed by merging the orthologous groups, the COGs search collects more orthologues together, reflecting the many-to-many relationships between proteins and between the genes that encode them. The results of our COGs identification are also incorporated into the comparison database within Microbase*Lite*, and users can search the COGs for a given gene via the genome viewer or access the results via the Web service exposed API. When a new genome is imported, it will be compared with existing genomes to find the mutual best hits between them. If a protein from the new genome has found two mutual best hits in a COG, it can be assigned to that COG.

V. Conclusion

Grid technologies enable a more rapid analysis of genome sequences facilitating a more intensive exploration of genomic data than can be achieved with traditional technology. In turn, this allows knowledge to be more quickly derived from our investment in sequencing programs and helps to address the problem of the analysis of rapidly accumulating genomic data. The Microbase project is developing a grid-based environment to support computationally intensive and data-intensive genome comparison and analysis, particularly for the analysis of microbial genomes. Microbase*Lite* is a system implementation that integrates distributed computing and data resources to perform the comparison and analysis of genome sequences. The system features the extensive use of Web service technologies for component orchestration, notification, database update, and user access. A large volume of precomputed comparison dataset has been generated on the system and exposed as a base level database to end-users for in-depth biological and biomedical research. We expose the results in both a computational amenable and user-friendly form through the use of Web services and graphical user interfaces.

One of the important applications implemented within Microbase*Lite* is the identification of protein families in a large number of proteomes. Such searches may aid the identification of potential targets for drug discovery and increase our understanding of protein evolution, and we hope that the system will prove useful in this respect.

In the future, we aim to develop a workflow framework to support the definition and enactment of custom applications based on which the system will be able to service user-defined, remotely conceived genome analyses. The system will provide the services to support user application submission and execution on the grid system. A decentralized database will be deployed. We will extend the protein family analysis to include the TribesMCL algorithm and implement other applications such as metabolic reconstruction and promoter searching. Finally, the system is not limited to the analysis of microbial genomes, and we intend to extend our approach to the analysis of eukaryotic genomes.

References

[1] M. T. Black and J. Hodgson, "Novel target sites in bacteria for overcoming antibiotic resistance" *Adv. Drug Del. Rev.*, vol. 57, no. 10, pp. 1528–1538, 2005.

[2] H. Loferer, "Mining bacterial genomes for antimicrobial targets," *Mol. Med. Today*, vol. 6, pp. 470–474, 2000.

[3] J. Rosamond and A. Allsop, "Harnessing the power of the genome in the search for new antibiotics," *Science*, vol. 287, no. 5460, pp. 1973–1976, 2000.

[4] I. Foster and S. Tuecke, "Describing the elephant: The different faces of IT as service," *ACM Queue*, vol. 3, no. 6, pp. 26–29, 2005.

[5] I. Foster, C. Kesselman, J. M. Nick, and S. Tuecke, "Grid services for distributed system integration," *Computer*, vol. 35, no. 6, pp. 37–46, 2002.

[6] N. Jacq, C. Blanchet, C. Combet, E. Cornillot, L. Duret, K. Kurata *et al.*, "Grid as a bioinformatic tool," *Parallel Comput.*, vol. 30, no. 9–10, pp. 999–1167, 2004.

[7] M. Cornell, I. Alam, D. Soanes, H. Wong, M. Rattray, S. Hubbard *et al.*, "e-Fungi: An e-science infrastructure for comparative functional genomics in fungal species," presented at the 4th U.K. e-Sci. All Hands Meet. (AHM 2005), Nottingham, U.K..

[8] D. Sulakhe, A. Rodriguez, M. D. Souza, M. Wilde, V. Nefedova, I. Foster, and N. Maltsev, "GNARE: An environment for grid-based high-throughput genome analysis," in *Proc. 5th IEEE Int. Symp. Cluster Comput. Grid (CCGrid 2005)*, Cardiff, U.K., May, pp. 455–462.

[9] N. Maltsev, E. Glass, D. Sulakhe, A. Rodriguez, M. H. Syed, T. Bompada *et al.*, "PUMA2-grid-based high-throughput analysis of genomes and metabolic pathways," *Nucl. Acids Res.*, vol. 34, no. 1, pp. D369–D372, 2006.

[10] R. L. Tatusov, M. Y. Galperin, D. A. Natale, and E. V. Koonin, "The COG database: a tool for genome-scale analysis of protein functions and evolution," *Nucl. Acids Res.*, vol. 28, no. 1, pp. 33–36, 2000.
[11] R. R. Chaudhuri, A. M. Khan, and M. J. Pallen, "coliBASE: An online database for Escherichia coli, Shigella and Salmonella comparative genomics," *Nucl. Acids Res.*, vol. 32, pp. D296–D299, 2004.
[12] R. Stevens, A. Robinson, and C. Goble, "myGrid: Personalised bioinformatics on the information grid," *Bioinformatics*, vol. 19, no. 1, pp. i302–i304, 2003.
[13] R. Stevens, R. McEntire, C. A. Goble, M. Greenwood, J. Zhao, A. Wipat *et al.*, "myGrid and the drug discovery process," *Drug Discovery Today: BIOSILICO*, vol. 2, no. 4, pp. 140–148, 2004.
[14] A. Krishna, V. Tan, R. Lawley, S. Miles, and L. Moreau, "myGrid notification service," in *Proc. U.K. e-Sci. All Hands Meet.*, Nottingham, U.K., 2003, pp. 475–482.
[15] R. L. Tatusov, E. V. Koonin, and D. J. Lipman, "A genomic perspective on protein families," *Science*, vol. 278, pp. 631–637, 1997.
[16] A. K. Bansal and T. E. Meyer, "Evolutionary analysis by whole-genome comparisons," *J. Bacteriol.*, vol. 184, no. 8, pp. 2260–2272, 2002.

Yudong Sun (M'05) received the B.Sc. and M.Sc. degrees in computer science from Shanghai Jiao Tong University, Shanghai, China, in 1985 and 1988, respectively, and the Ph.D. degree in computer science from the University of Hong Kong, Hong Kong, in 2002.

He is currently a Research Officer at the Oxford University Computing Laboratory, Oxford University, Oxford, U.K. He was a Research Associate in parallel and distributed computing at Hong Kong Polytechnic University, and a Research Associate in bioinformatics and e-Science at Newcastle University, Newcastle upon Tyne, U.K. He was a Lecturer, and then an Associate Professor in the Department of Computer Science and Engineering, Shanghai Jiao Tong University, from 1988 to 1996. His current research interests include parallel, distributed, and grid computing, scientific computing, e-Science, bioinformatics, computer architecture, system modeling, and simulation.

Anil Wipat received the B.Sc. degree in applied biology and the Ph.D. degree in molecular microbiology from John Moores University, Liverpool, U.K., in 1986 and 1990, respectively, and the M.Sc. degree with distinction in computing science from Newcastle University, Newcastle upon Tyne, U.K., in 2000.

He is a Reader and the Director of Degree Programme on Master of Research (MRes) in bioinformatics in the School of Computing Science, Newcastle University. He was a Postdoctoral Researcher in molecular microbiology at Newcastle University and the European Bioinformatics Institute, Cambridge, U.K. He was a Lecturer of microbial genomics in the Microbiology Department, Newcastle University. He was a Software Engineer and a Research and Development Scientist. His current research interests include microbiology, microbial genomics, e-Science grid computing, computational systems biology, and data integration.

Matthew Pocock was born in Oxford, U.K. He received the B.Sc. (Hons.) degree in genetics from Newcastle University, Newcastle Upon Tyne, U.K., in 1995, and the Ph.D. degree in bioinformatics from Darwing College, Cambridge, U.K., and the Sanger Centre, Cambridge, U.K., in 2003.

He is a member of the School of Computing Science, Newcastle University, where he is currently a Research Associate working on the Comparagrid Project. He was a Systems Administrator, Software Architect, and has worked with several genetics laboratories. His current research interests include interface between software architectures, ontological descriptions of knowledge,Web services, workflows (Taverna), semantic Web technologies [Web ontology language description logics (OWL-DL)], data integration (ComparaGRID), and community-based biological ontologies.

Peter A. Lee received the B.Sc., M.Sc., and Ph.D. degrees in computing science from the University of Manchester, Manchester, U.K., in 1972, 1973, and 1976, respectively.

He is a Professor and the Head of the School of Computing Science, Newcastle University, Newcastle Upon Tyne, U.K., where he started a research group addressing parallel processing tools, concentrating initially on shared-memory multiprocessors and visual, object-oriented programming languages. He was engaged in fault-tolerant systems. His current research interests include the support of bioinformatics research.

Keith Flanagan received the B.Sc. (Hons.) degree in software engineering in 2003, from Newcastle University, Newcastle Upon Tyne, U.K., where he is currently working toward the Ph.D. degree.

He is currently a Research Associate with Newcastle University. He had been with a Newcastle-based drug discovery company. His current research interests include comparative genomics and parallel/grid computing.

James T. Worthington received the B.Sc. degree in genetics and the M.Res. degree in bioinformatics from Newcastle University, Newcastle Upon Tyne, U.K., in 2003 and 2004, respectively.

He is currently a member of the Technical Support Team with Convergys Corporation, London, U.K. He was a Research Associate with Newcastle University. His current research interests include distributed computing, e-Science, and Web applications.