

e-Fungi: An e-Science Infrastructure for Comparative Functional Genomics in Fungal Species

Mike Cornell¹, Intikhab Alam¹, Darren Soanes³, Han Min Wong⁴, Magnus Rattray¹, Simon Hubbard², Nicholas J. Talbot³, Brian Lings⁴, David Hoyle⁴, Stephen G. Oliver² and Norman W. Paton¹.

¹ School of Computer Science, University of Manchester, Kilburn Building, Oxford Road, Manchester, M13 9PL.

² Faculty of Life Sciences, University of Manchester, Michael Smith Building, Oxford Road, Manchester, M13 9PT.

³ Department of Biological Sciences, University of Exeter, Hatherly Laboratories, Prince of Wales Road, Exeter, EX4 4PS.

⁴ School of Engineering, Computer Science and Mathematics, University of Exeter, Harrison Building, North Park Road, Exeter, EX4 4QF.

Abstract

e-Fungi aims to integrate sequence and functional data from multiple fungal species, to facilitate the systematic study of less well understood species with reference to model organisms. e-Fungi consists of a data warehouse and a library of bioinformatics queries and analyses, which can be combined in different ways to conduct studies of cellular processes, pathogenicity and evolution. Both the warehouse and analysis libraries will be made available within a service-oriented Grid.

1 Introduction

Classical comparative genomics identifies similar sequences in different organisms and infers functional equivalence from that sequence similarity. Its uses have included investigating orthologous genes (Tatusov *et al.*, 1997), inferring the presence of metabolic pathways, (Giles *et al.*, 2003), and defining transcription-factor binding sites and other regulatory sequences in DNA (Cliften *et al.*, 2001, Kellis *et al.*, 2003). Comparative functional genomics aims to extend such comparisons by including experimental data about function and phenotype. In order to facilitate these comparisons, it is essential that sequence and functional data are organised in a way that makes integrated analyses straightforward. At present this is not the case; biological data is stored in a variety of formats at different sites, making integration problematic. We have previously developed the Genome Information Management System (GIMS) (Cornell *et al.*, 2003), an integrated data warehouse storing sequence and functional data for the budding yeast *Saccharomyces cerevisiae*. The usefulness of this approach has been demonstrated by evaluating protein interaction data using microarray and annotation data (Cornell *et al.*, 2004). The e-Fungi project will

extend the functionality of GIMS to contain multiple fungal genomes, allowing cross-species analyses, as well as functional and annotation data.

In the development of the GIMS database, the focus has been on a single organism allowing the data selection, cleaning, loading and maintenance activities to be conducted manually. However, the size of the e-Fungi data warehouse requires a more systematic approach to maintenance. The cost/benefit trade-offs for replicating data in a warehouse, as against retrieving data on-demand and the systematic development of appropriate maintenance policies (Engström *et al.*, 2002, 2003) are being explored. In addition, the increased size of the warehouse will lead to database queries becoming more computationally intensive. In order to overcome this problem, the e-Fungi database and analysis tools will be made available as web services, allowing the efficient evaluation of computationally demanding requests over the data to be deployed as services on the Grid. In particular, OGSA-DQP (Alpdemir *et al.*, 2003) will be used to express queries over a combination of the e-Fungi warehouse, external databases, and bioinformatic analyses made available as web services.

In order to demonstrate the effectiveness of the e-Fungi approach to data management we are conducting a detailed exploration of fungal pathogenesis in *Magnaporthe grisea*, *Aspergillus fumigatus*, and *Candida albicans* by effective comparison with the model eukaryotes such as *Saccharomyces cerevisiae*.

2 The e-Fungi data warehouse.

The e-Fungi data warehouse has been implemented using the FastObjects t7 database (<http://www.versant.com/products/fastobjects/t7>). The object model is an extension of that used in the GIMS database, allowing sequence data to be stored for multiple genomes.

The data stored in the data warehouse can be split into two groups. The first is data that we have loaded from other data repositories, e.g. sequence data from the NCBI (Pruitt *et al.*, 2005) and pathway data from KEGG (Kanehisa and Goto, 2000), or from published data sets which have not been made available via online repositories, such as genes described as being involved in RNAi in a publication (Kim *et al.*, 2005). The second type of data is that which we obtain from analysing other data in the warehouse. For example the orthology relationships obtained from all vs. all Blast analysis (see Section 3).

The data warehouse currently contains genomic sequence data from 25 fungal species, including model organisms, non-pathogenic fungi and both human and plant pathogens (see Table 1). As further genomes are sequenced we intend to add them to the data warehouse. The extent to which genomes have been sequenced and annotated varies greatly. Some consist of fully sequenced chromosomes, with annotation including sequence similarities, intron positions, non-translated RNAs and repeat sequences. Others consist of contig sequences with annotation limited to predicted open reading frames (ORFs). The accuracy of the data also varies between genomes. In the case of *S.cerevisiae*, since the original publication of the genome there have been further studies which have led to re-annotation of the genome (e.g. Kellis *et al.*, 2003). This has not yet happened for more recently published genomes. In addition different sequencing projects may have used different protocols for annotating sequence data. For example, the two annotations of the *S.bayanus* and *S.mikatae* have very different numbers of ORFs and differing percentages of these encode

small (less than 100 amino acid residues) proteins. During our analysis of the fungal sequence data we aim to resolve some of the issues and provide more robust set of fungal coding sequences.

Large amounts of functional data are being generated for fungal species. In some cases, such as expression or protein interaction data, this data is restricted largely to *S.cerevisiae*. However, it is likely that functional data will become available for other species. For example, once an organism has been fully sequenced, microarray slides will often be generated for that organism. In other cases there is more data available for non-model organisms. For example, large amounts of EST data has been made available for *M.grisea* and *N.crassa*. It is intended that e-Fungi analyses will involve a diverse range of functional data including microarray expression data, protein interactions and metabolic pathways. In addition, there is annotation data from GO (Ashburner *et al.*, 2000) and MIPS (Mewes *et al.*, 2002). The size of the database as well as the diverse range of data and the large number of different data sources, will raise issues concerning the updating of the data warehouse and the requirements for running complex queries. These will be discussed more fully in Section 5.

3 Protein families analysis

One goal in comparative genomics is the functional assignment of predicted genes. Genes that are significantly similar (either paralogs resulting from gene duplication, or orthologs resulting from speciation) are frequently found to have a related biological function. In order to examine homologies through protein family analyses and to identify orthologs and paralogs we divide our analysis pipeline into two phases, Phase A and Phase B. In Phase A of our pipeline we follow a top level analyses approach, for example Blast based orthology assignments, while in Phase B we perform an in-depth analysis, such as phylogenetics based orthology assignments, to reveal relationships among proteins from the data set. Phases A and B are further divided into various steps as follows:

Genome	Data source	ORFS	%ORFS<100 amino acids	Genome size (Mb)
<i>Saccharomyces cerevisiae</i> *	Entrez (NCBI)	5866	5.3	12.2
<i>Schizosaccharomyces pombe</i> *	Entrez (NCBI)	5010	4	12.5
<i>Encephalitozoon cuniculi</i> *	Entrez (NCBI)	1996	1.95	2.5
<i>Candida glabrata</i> *	Entrez (NCBI)	5192	3.22	12.3
<i>Debaryomyces hansenii</i> *	Entrez (NCBI)	6317	7.92	12.2
<i>Eremothecium gossypii</i> *	Entrez (NCBI)	4726	2.98	8.8
<i>Kluyveromyces lactis</i> *	Entrez (NCBI)	5336	3.75	10.7
<i>Yarrowia lipolytica</i> *	Entrez (NCBI)	6544	3.1	20.6
<i>Aspergillus nidulans</i>	AACD01000001	21181	1.76	47.2
<i>Candida albicans</i>	AACQ01000001	14217	0.82	27.6
<i>Gibberella zeae</i>	AACM01000001	11640	1.81	36.1
<i>Magnaporthe grisea</i>	AACU01000001	11109	3.03	38.8
<i>Neurospora crassa</i>	AABX01000001	10079	10.72	38.0
<i>Ustilago maydis</i>	AACP01000001	6522	1.06	19.7
<i>Saccharomyces bayanus</i>	MIT (via SGD)	9424	23.84	11.5
<i>Saccharomyces bayanus</i>	WashU (via SGD)	4966	3.3	11.9
<i>Saccharomyces castellii</i>	WashU (via SGD)	4677	2.67	11.4
<i>Saccharomyces kluyveri</i>	WashU (via SGD)	2968	3.64	11.0
<i>Saccharomyces kudriavzevii</i>	WashU (via SGD)	3768	4.62	11.2
<i>Saccharomyces mikatae</i>	MIT (via SGD)	9057	23.46	11.5
<i>Saccharomyces mikatae</i>	WashU (via SGD)	3100	5.35	10.8
<i>Saccharomyces paradoxus</i>	MIT (via SGD)	8955	23.66	11.9
<i>Trichoderma reesei</i>	JGI	8592	0.07	34.5
<i>Phytophthora sojae</i>	JGI	19071	3.12	86.0
<i>Phytophthora ramorum</i>	JGI	15890	2.91	66.7
<i>Phanerochaete chrysosporium</i>	JGI	11532	19.18	28.8
<i>Aspergillus fumigatus</i>	CADRE	9945	3.95	29.7

Table 1. Genomes currently in e-Fungi data warehouse. * indicates that this genome has been fully sequenced, i.e. there are complete chromosome sequences. Other genomes are available as a set of contig sequences and there may be DNA sequences that have not been sequenced for these organisms. For these genomes, the genome size is an estimate based upon the total sizes of all the contigs.

Phase A:	A3)	Categorization of protein families based on presence or absence of genes, across 24 fungal genomes.
A1) All against all Blast (Altschul <i>et al.</i> , 1997) analysis with Evalue threshold of 1e-5.	A4)	Best Bi-directional Blast Hit (BBH) analysis to assign orthologies.
A2) Clustering Blast similarities into non-overlapping protein families using MCL.	A5)	InterProScan analyses to get information about sequences by scanning signature

databases e.g. GO terms, Prosite motifs, Pfam domains etc.

Phase B:

- B1) More sensitive similarity searches based on results from A2, using CHASE (Alam *et al.*, 2004) to get maximum possible members of a protein family and their alignments.
- B2) Repeat A3,
- B3) Finding missing genes using GenCHASE.
- B4) Categorization of protein families based on presence or absence of genes from genomes. For example, fungi specific, filamentous fungi specific, ascomycete specific and eukaryotic specific.
- B5) Phylogenetic analysis based on A1 and B4.

In Phase A, we generated a dataset of 214,777 amino acid sequences, corresponding to all predicted protein sequences with length greater than 40 amino acids from 25 fungal genomes. Step A1 produced 30,715,566 similarities between protein sequences. Based on these similarities the Markov Chain Clustering (MCL) (Van Dongen, 2000) was carried out (A2), resulting in 178,821 proteins being clustered into 19,203 families of homologues and the rest (35,956) as singletons.

Categorization of protein families shows that there are about 176 families, of sizes between 27 and 1732 which are conserved throughout pathogenic and non pathogenic genomes. We are currently generating phylogenetic trees based on these conserved families. Among the unevenly distributed families, 19,027 have sizes between 2 and 1401. Of these 18,367 families have less than 30 members.

Analyses in Phase B will result in a more detailed picture of the relationships among the participating pathogenic and non-pathogenic genomes. In general, a complete phylogenetic analysis of all groups of homologous genes is required to decipher true orthologous relationships (Koonin and Galperin, 2002). To define robust protein families, containing maximum possible members, we used CHASE (Alam *et al.*, 2004) which uses multiple sequence-based homology

search methods. Further, to recover possible missing genes, we apply GenCHASE (Alam, unpublished) a tool that employs multiple methods, taking protein sequences as an input and finding homologues in DNA sequences using six reading frame translations. Robust protein families obtained as a result of this phase of analysis will improve the phylogenetic analyses, helping identify orthologs and paralogs which in turn aids the functional assignment of predicted genes. The results from these analyses will be made available to users via the e-Fungi data warehouse, allowing them to be combined in queries with appropriate functional data.

4 Study of pathogenicity factors

Molecular phylogeny has shown that pathogenic fungi are found in many taxonomic groups, suggesting that these lifestyles have evolved repeatedly within the fungal kingdom. It can be speculated that there are three possible mechanisms that account for the evolution of pathogens (Tunlid and Talbot, 2002). Firstly, the genomes of pathogens may have acquired novel genes, enabling them to infect and colonise plants. This might occur by horizontal gene transfer (Rosewich and Kistler, 2000), or, more likely, as a result of gene duplication followed by sequence divergence (Ohno, 1970). Secondly, pathogenicity might be associated with gene loss. Thirdly, there may be genes in pathogens that are also present in non-pathogens but take on a different role in the former, possibly due to changes in gene regulation.

By searching the literature we have identified 292 experimentally determined pathogenicity factors from both plant and animal pathogens. These are involved in a wide range of cellular processes. We hope that comparison of genomic and functional data between pathogenic and non-pathogenic species of fungi, combined with our knowledge of these pathogenicity factors, will enable us to start to answer the question “What makes a pathogen different from a non-pathogen?” Comparison of gene inventories between fungal genomes, as described in Section 3, will allow us to identify instances of gene duplication or loss, as well as genes which are unique to pathogens or non-pathogens. By integrating functional data, for example on metabolic pathways or gene expression, we may be able to investigate whether altered gene regulation is responsible for pathogenicity.

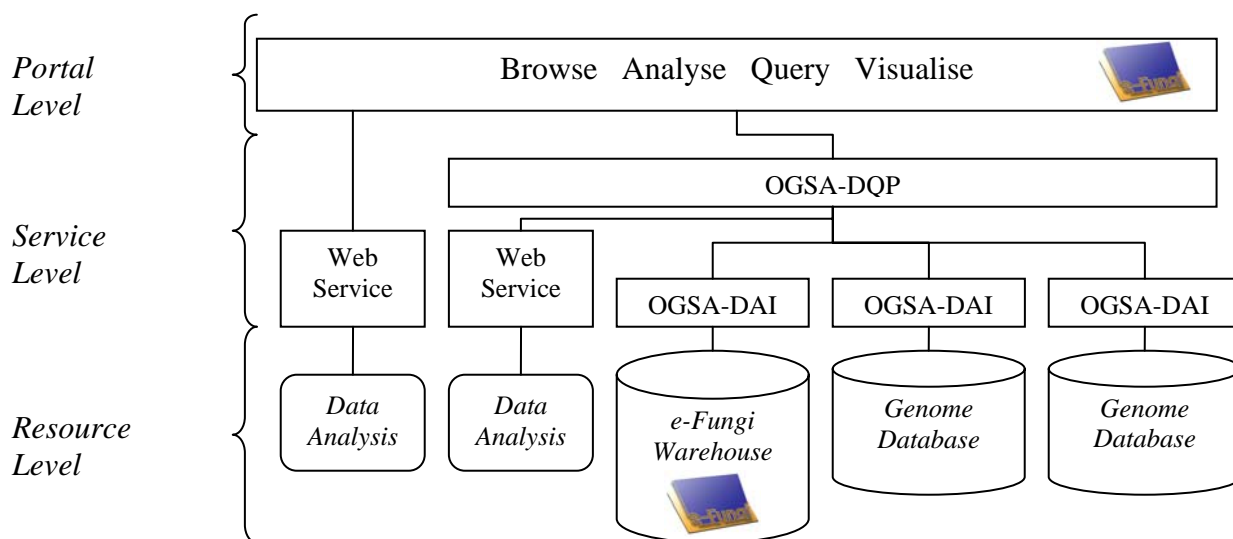


Figure 1. The e-Fungi architecture.

As a result of our analyses described in Section 3, we have identified 71 protein families, in which proteins from pathogenic fungi appear to be over-represented. These are now being subjected to more detailed analysis.

5 The e-Fungi architecture.

As stated earlier, the development of e-Fungi from the GIMS database has greatly increased the size of the data warehouse and the requirements for running complex queries. The three tiers of the e-Fungi architecture are shown in Figure 1. At the resource level, there is the e-Fungi data warehouse, other on-line databases which might be used for updating the e-Fungi warehouse or for running queries in association with e-Fungi data, plus a library of queries for performing data analyses. At the service level, OGSA-DAI (Antonioletti 2005) middleware will be used to integrate data, while OGSA-DQP (Alpdemir *et al.*, 2003) will be used to express queries over a combination of the e-Fungi warehouse, external databases, and bioinformatic analyses made available as web services. At the portal level, the e-Fungi interface will be implemented using Pierre (Garwood, unpublished), a model driven application for creating front ends for data repositories.

5.1 Data Warehouse Updates

Data integration currently requires intensive maintenance activities by a bioinformatician, as

new data sources are integrated and old sources updated. There is a need to make use of aspects of Data Warehouse technologies in order to reduce the need for ongoing human interactions in maintenance activities. The cost/benefit trade-offs for replicating data in a warehouse, as against retrieving data on-demand and the systematic development of appropriate maintenance policies have been explored (Engström *et al.*, 2002, 2003).

For example, the current GIMS system incorporates multiple external data sources and is in need of mechanisms that provide some means of automated updating to the local databases. While maintaining the local databases in the best possible manner, these mechanisms should not overload the warehouse or the source by updating the database excessively. Careful considerations on the updating policy are required to optimize the performance without sacrificing data accuracy. Such policies may be based on information such as the type of source, analysis, as well as data users' behavior. Passed query results may also be cached to further enhance performance.

When new data sets are integrated it would be desirable if the amount of human interaction required were minimized, although the lack of standards within biological data sources will make some interaction inevitable. However we will endeavor to make e-Fungi data conform to the emerging standards, such as SBML (Hucka *et al.*,

2003), to maximize the extent that e-Fungi can be integrated with other online resources.

5.2 Data Integration

In data integration, it is important for any tool to be able to present itself as a service such that it can be used/integrated into other higher-level services. For example, as an analysis tool itself, the potential of GIMS can be significantly expanded by having the ability to integrate itself to other tools/services, i.e. by latching on standardized protocols such as Grid and Web-services. As a resource, the results from the local analyses are also important. They need to be structured in a standardized, consistent and meaningful manner and, where possible, accompanied by meta data, such that the results can be compatible as input data to other services.

5.3 Computational infrastructure

There is a need to satisfy QoS requirements within the context of the possibilities offered by computational Grids. A request for modest resources within a lightly constrained timeframe may allow local (and cheap) computation. However, a request for high precision over large data sets may well trigger a Grid-based solution constrained only by specified cost. All of this, including the selection of the appropriate algorithms to deliver the correct level of precision, should be transparent to a user.

6 Conclusions

The development of high throughput DNA sequencing combined with advances in gene expression profiling, proteomics and metabolomics, have resulted in an explosion in the amount of publicly available sequence and functional data. The development of systems biology necessitates the integration of this data in order to help us answer complex biological problems. However, automated data integration has proved problematic, often due to the lack of compatibility between data formats and the inability of databases to communicate. The development of web services and agreed data structures will provide some solutions to these problems.

However, these approaches have only been applied to a subset of biological data sources, a problem that was evident when integrating the genomic data in e-Fungi. The e-Fungi approach, building an integrated data repository and analysis

software, and making them available as web services, may provide some solutions. Data integration is still problematic but having been done; the data can be made available in a structured format. In addition, the problem of functional annotation in fungal genomes, with the vast majority of data applying to a couple of model organisms, might be eased by making the results of comparative genomics available in an integrated format via the e-Fungi data warehouse.

References

- Alam I *et al.* (2004) Comparative homology agreement search: an effective combination of homology-search methods. *Proceedings of the National Academy of Sciences (PNAS)*, USA. Volume 101 (38) 13814-19.
- Alpdemir, M.N *et al.*, (2003), Service-Based Distributed Querying on the Grid, *1st International Conference on Service Oriented Computing (ISOC)*. 467-482, Springer-Verlag.
- Antonioletti, M *et al.*, (2005) The design and implementation of Grid database services in OGSA-DAI. *Concurrency and Computation: Practice and Experience*. 17, 357-376.
- Ashburner M *et al.* (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics* 25:25-29.
- Cliften, P.F *et al.*, (2001) Surveying Saccharomyces genomes to identify functional elements by comparative DNA sequence analysis, *Genome Research* 11:1175-1186.
- Cornell M *et al.* (2003) GIMS: an integrated data storage and analysis environment for genomic and functional data. *Yeast* 20:1291-1306.
- Cornell M *et al.* (2004) A critical and integrated view of the yeast interactome. *Comparative and Functional Genomics* 5:382-402.
- Engström H, *et al.*, (2003), A Heuristic for Refresh Policy Selection in Heterogeneous Environments, *19th Intl. Conference on Data Engineering (ICDE 2003)*, Bangalore, IEEE Press.
- Engström H, *et al.*, (2002), A Systematic Approach to Selecting Maintenance Policies in a Data Warehouse Environment, *8th Intl. Conference on Extending Database Technology (EDBT)*, Prague, March 24-28, Springer, 317-335.

Giles PM *et al.*, (2003) A relational database for the discovery of genes encoding amino acid biosynthetic enzymes in pathogenic fungi. *Comparative and Functional Genomics*, 4:4-15.

M. Kanehisa and S. Goto (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28:27-30.

Kim JK, *et al.*, (2005) Functional genomic analysis of RNA interference in *C. elegans*. *Science*.308:1164-7.

Kellis M, *et al.*, (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*. 423:241-54.

Koonin EV and Galperin MY, (2002) Sequence - Evolution – Function. Computational Approaches in Comparative Genomics Kluwer Academic Publishers. ISBN 1-40207-274-0

M. Hucka (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*. 19(4):524-31.

Mewes HW *et al.* (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* 30(1):31-4.

Ohno, S. (1970) Evolution by Gene Duplication (Springer: New York)

Pruitt KD, *et al.*, (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 33:D501-4.

Rosewich UL and Kistler HC, (2000) Role of horizontal gene transfer in the evolution of fungi. *Annu. Rev. Phytopathol.* 38, 325-363.

Tatusov RL, *et al.*, (1997) A genomic perspective on protein families. *Science*. 278: 631-637.

Tunlid A and Talbot NJ, (2002), Genomics of parasitic and symbiotic fungi. *Current Opinion in Microbiology* 5, 513-519.

Van Dongen S. (2000) PhD Thesis, University of Utrecht, The Netherlands. (<http://micans.org/mcl/>)