

A Methodology for Comparative Functional Genomics

Intikhab Alam¹, Mike Cornell^{1,2}, Darren M. Soanes³, Cornelia Hedeler¹, Han Min Wong³, Magnus Rattray¹, Simon J. Hubbard², Nicholas J. Talbot³, Stephen G. Oliver² and Norman W. Paton¹

¹ School of Computer Science & ² Faculty of Life Sciences, University of Manchester, Oxford Road, Manchester, M13 9PL, UK

³ Department of Biosciences, University of Exeter, Geoffrey Pope Building, Stocker Road Exeter, EX4 4QD, UK

Abstract

The continuing and rapid increase in the number of fully sequenced genomes is creating new opportunities for comparative studies. However, although many genomic databases store data from multiple organisms, for the most part they provide limited support for comparative genomics. We argue that refocusing genomic data management to provide more direct support for comparative studies enables systematic identification of important relationships between species, thereby increasing the value that can be obtained from sequenced genomes. The principal result of the paper is a methodology, in which comparative analyses are constructed over a foundation based on sequence clusters and evolutionary relationships. This methodology has been applied in a systematic study of the fungi, and we describe how comparative analyses have been implemented as an analysis library over the e-Fungi data warehouse.

1 Introduction

There is considerable and ongoing investment in genome sequencing. However, although low cost and systematic sequencing has led to rapidly growing sequence resources, experimental follow-up in the form of comprehensive functional analyses remains partial. As a result, it is important that analyses can be performed over sequenced genomes that both learn lessons in their own right and help to inform the design of functional studies. Although papers describing sequenced genomes invariably report on relationships between related organisms, such reports typically only scratch the surface in terms of depth and scope. Furthermore, although many valuable comparative studies have been carried out (e.g. [1,2,3,4]), there is as yet no well defined methodology for conducting comparative studies.

This paper presents a methodology for comparative genomics, in the form of a collection of analyses that, when used together, provide insights into the relationships between genomes that can be widely applied. Of course, the methodology is not complete, in that many additional analyses can usefully be performed, but we hope to encourage ongoing consideration as to the fundamental, or at least generally useful, constituents of comparative studies. In essence, we consider two principal stages:

1. *Foundational sequence analyses*: as sequences are the common resource in comparative genomics, they provide the foundation of the proposal. From sequences, clusters, families and evolutionary relationships provide the building blocks on which other analyses are built.
2. *Focused comparative studies*: given the foundational sequence analyses, it is then possible to explore relationships between pairs and groups of genomes; for example, genome redundancy, patterns in gene gain and loss, explorations of essential genes,

and preservation of pathways can all be explored using straightforward analyses over the same foundations.

These stages can both be implemented in different ways. The first stage principally takes sequence data as an input, but gives rise, through computationally demanding analyses, to substantial derived data resources on which the second stage builds. The second stage involves the integration of other data resources with the products of the first stage. Our implementation of the methodology in the e-Fungi project centres on a data warehouse in which the data used and produced by the first phase is integrated with that required by the second phase.

The remainder of the paper is structured as follows. Section 2 reviews techniques for relating genomes using sequence analyses, and thus describes the foundational sequence analyses phase. Section 3 identifies a collection of comparative studies that build directly on the results of the sequence analyses. Section 4 describes how the methodology has been implemented in the e-Fungi data warehouse (available online at <http://www.e-fungi.org.uk/>), and Section 5 concludes.

2 Relating Genomes by Sequence Comparison

Sequence analysis lies at the heart of comparative genomics. Comparative sequence analysis allows us to infer how the processes of mutation, deletion, duplication and selection have shaped the evolution of genomes [5, 6]. By making use of multiple genomes, we can identify the similarities and differences in the genes and regulatory regions, and thereby infer the selective pressure acting on these, which in turn provides useful evidence about their function [7]. Comparative sequence analysis is also useful for validation of predicted genes and gene models, and for the inference of protein function by homology [8]. Below we describe the sequence analysis methods used to provide the foundation on which other analyses are built.

2.1 Clustering of proteins into families and orthologous groups

There are several methods available for organising proteins into clusters or families [9]. Some methods cluster sequences based on some measure of sequence similarity such as BLAST scores [10]. Popular examples are BLASTClust and the Markov clustering (MCL) algorithm [11]. Alternatively, sequences can be scanned against motif or domain databases, such as Prosite [12] and Pfam [13]. Of the clustering methods, BLASTClust is a simple single-linkage clustering method that works reasonably well for identifying clusters of very similar sequences. However, BLASTClust is less effective for identifying more divergent clusters of sequences. MCL is an alternative approach, which uses graph theory and random walk simulation for clustering. It is computationally efficient, and is applicable to a huge number of sequences. It includes an inflation parameter that controls the granularity of clusters. As an example of the use of MCL, it was run over the 36 genomes contained in the current version of e-Fungi. These contain a total of about 350 thousand protein sequences, resulting in over 47 million similarities to be used in MCL clustering. We used 2.5 as a moderate inflation value and $1e-10$ as a comparatively strict E-value cut-off. With these parameters, the MCL algorithm produced 23724 clusters, containing 80% of all sequences, while the rest of the sequences were singletons.

OrthoMCL [14], a variant of the MCL algorithm that also uses BLAST similarity results, is able to construct putative orthologous groups, including recent paralogs, across multiple taxa. We use MCL to obtain comprehensive groups of similar fungal sequences and we use OrthoMCL to obtain smaller clusters containing putative orthologs and recent paralogs based on best-bidirectional BLAST hits from multiple fungal species.

An alternative way to organise protein sequences is to identify motifs or domains (independent functional units) contained within them. Prosite is a database of core functional motifs, while Pfam is a database containing a large number of known protein domain models. The models in Pfam are hidden Markov models (HMMs), which capture the statistical properties of the sequences in each family of protein domains. We have developed an alternative Pfam-like domain model database, which is specific to fungal genomes, called FPFam [15]. The fungal-specific HMM models in FPFam have been shown to detect more instances of known domains than the more general Pfam models. The coverage of domains increases even further when both databases are combined. Therefore, we recommend that both general and kingdom-specific models are used for identifying domains in predicted protein sequences from new genomes.

Domain-based methods provide a well-defined relationship between sequences, whereas clustering techniques like MCL provide a more *ad-hoc* organisation of sequences. However, even using a fairly liberal E-value cut-off of 0.1, 35% of all predicted protein sequences in E-Fungi contain no Pfam or FPFam domains. Therefore, clustering-based methods seem to obtain better coverage. However, one can also create new HMM models by fitting models to clusters of domains in sequences that do not contain Pfam domains using e.g. domainer/mkdom [16]. We are currently applying this approach in order to increase the coverage of FPFam models. The clustering and family analyses in sequences are illustrated in Figure 1. In the data flow diagram, data collections are drawn between horizontal lines, and processes applied to the data are depicted in ellipses.

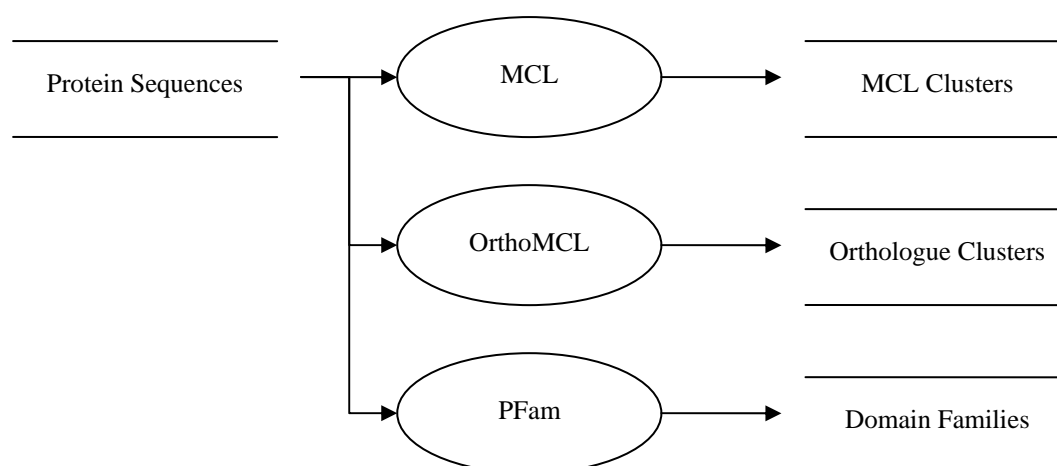


Figure 1: A data flow diagram for sequence clustering and family analyses.

2.2 Phylogenetic analysis

Phylogenetic trees allow results to be interpreted according to their evolutionary context. For example, when investigating the increase or decrease in the size or number of protein families, it is useful to consider the results in the context of a species tree. In the past, most phylogenetic analysis was carried out using single genes or small sets of concatenated genes. The availability of complete genome sequences offers the prospect of using large numbers of concatenated protein sequence alignments to reveal the evolutionary relationships between them (for a review on phylogenomic methods, see [17]). To obtain a species tree in e-Fungi, we selected several universal protein families present across all 36 genomes, aligned their sequences and concatenated these into one large alignment file. This concatenated alignment was submitted to PhyML [18], a maximum likelihood method to estimate phylogenies.

Once a good species tree has been obtained, other phylogenetic methods can be used to identify how changes have occurred over evolutionary time. For example, Dollo parsimony [19] or GeneTrace [20] can be used to determine the order of protein family gain or loss events occurring over evolutionary time. One can also compare the phylogenetic tree obtained by Dollo parsimony with the species tree obtained using a sequence alignment. This may expose trends in the distribution of protein families that do not correspond to the evolutionary relationship of species. The clustering and family analyses of sequences are illustrated in Figure 2.

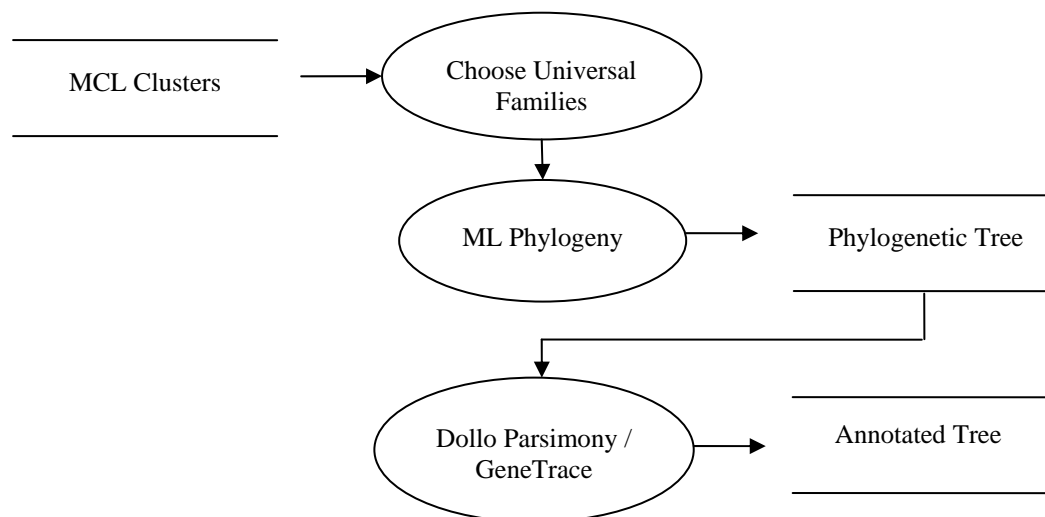


Figure 2: A data flow diagram for phylogenetic analyses.

3 Building on Sequence Comparisons for Systematic Analyses

This section describes how the foundation from Section 2 can be built on to systematically explore relationships between genomes.

3.1 Redundancy

Protein paralogs are the result of gene duplication events, which can be identified using OrthoMCL clusters. These duplication events can then be mapped to the phylogenetic tree, for instance, to identify duplications common to closely related species. Figure 3 shows a data flow diagram for an investigation of redundancy. By combining the OrthoMCL clusters with knowledge of the evolutionary relationship between the species, we can identify duplications that have occurred more frequently in one species than another. These patterns in the levels of redundancy can then be related to other information, such as the functional annotations of genes, to identify functions that are being enhanced between organisms or groups of organisms. For example, in the fungi we might wish to identify duplications common to the budding yeasts and the filamentous yeast *Sz. pombe* that were not identified in filamentous fungi.

Systematic gene deletion experiments using *S. cerevisiae* have demonstrated that the majority of its genes appear dispensable, their deletion having little effect on the survival of the organism. This dispensability has important consequences. For example, in the development of interacting protein networks, the robustness of the network will be dependent on its ability to withstand the deletion of protein nodes. Also, the rate at which a gene can accommodate mutations will be greatly affected by whether or not it is essential to the organism. One factor determining whether a protein might be dispensable is the presence of a closely related paralog, which might act as a “back up” in the event of gene deletion occurring.

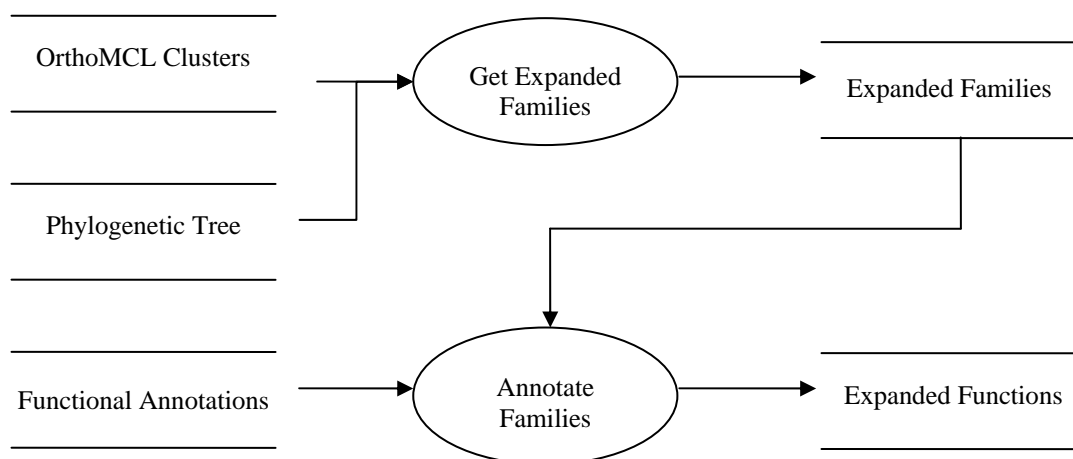


Figure 3: A data flow diagram for detecting and interpreting redundancy.

3.2 Conservation and diversification

Sequence clusters typically contain sequences from a subset of the genomes over which clustering was carried out. For example, in the fungi we can identify clusters specific to *Aspergillus* species or to the *Saccharomyces* “sensu stricto” species. This raises questions about how such clusters have originated. Are they the product of systematic gene deletions, or gene duplication followed by sequence divergence, or have they been acquired by horizontal gene transfer from other species. Figure 4 shows how the extent to which a proteome is conserved might be investigated. OrthoMCL clusters can be filtered to generate a set of clusters that contain proteins from an input species. For each of these clusters, we can then find the range of species that possess orthologs. This allows us to further divide our clusters into those that are specific to a genus, or that exhibit a specific phenotype (for example, they are pathogens).

By integrating these clusters with functional data we can look for features common to these sets of clusters. For example, do proteins associated with clusters limited to the *Saccharomyces sensu stricto* species share common GO terms or Pfam motifs? It also allows us to identify clusters with an unusual distribution of orthologs, such as a cluster containing a *S. cerevisiae* protein and its *Schizosaccharomyces pombe* ortholog, but lacking orthologs in more closely related Pezizomycotina and other Saccharomycotina species.

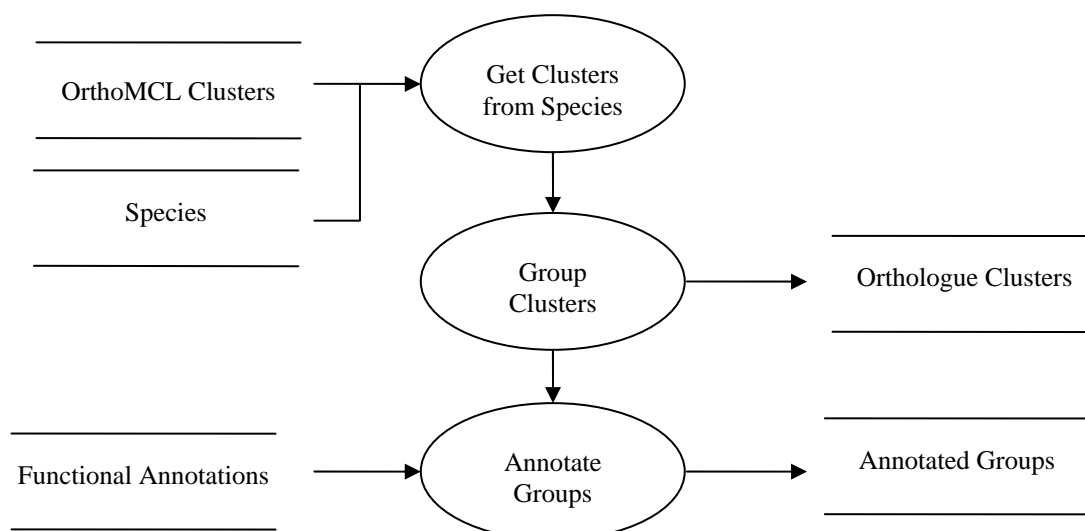


Figure 4: A data flow diagram for conservation and diversification.

3.3 Essential genes

Systematic gene deletion experiments have been widely used to relate genotype to phenotype, including the identification of essential genes. Figure 5 illustrates an approach to the analysis of essential genes. Firstly we might subdivide the OrthoMCL clusters into those which contain essential gene products and those containing non-essential gene products. We can then investigate species that have orthologs of these proteins and relate these to the phylogenetic tree.

We might expect that essential genes will tend to have orthologs over a greater range of species than non-essential genes, and this is indeed the case. However, in the fungi, there are many genes that are essential in *S. cerevisiae* but have no orthologs in any of the Pezizomycotina species, and some which are essential in the *Saccharomyces sensu stricto* species but appear to have no orthologs in other *Saccharomycotina* species.

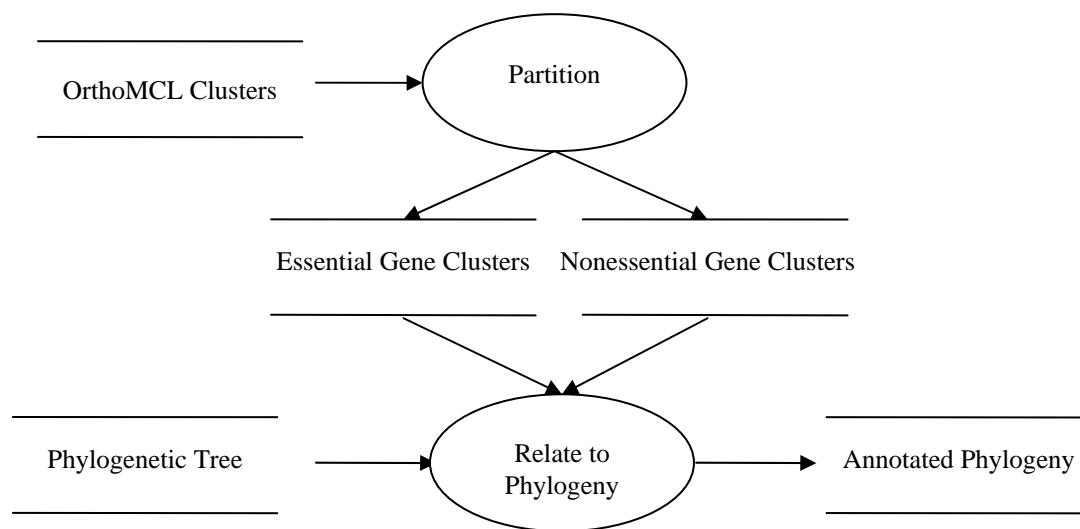


Figure 5: A data flow diagram for studying conservation of essential genes.

3.4 Pathway analyses

Proteins do not operate in isolation, but instead act as components in interacting networks. In many cases they directly interact with other proteins, to form a single unit, as in protein complexes. In other instances the interaction may not be physical but genetic. For example, a membrane bound receptor and a transcription factor that form part of the same signalling pathway. Figure 6 shows how conservation of metabolic pathways can be analysed by relating clusters to pathway data available in the LIGAND database [21]. LIGAND provides a mapping of enzyme classification to gene identifiers, allowing us to identify OrthoMCL clusters that contain enzymes for a given species. As with the examples above, we can determine the range of species that contain orthologs of these proteins. If an organism lacks an ortholog for an enzyme, we are able to search for the orthologs of other enzymes in the same pathway to identify multiple missing components. This approach can be used to study conservation of pathways across species.

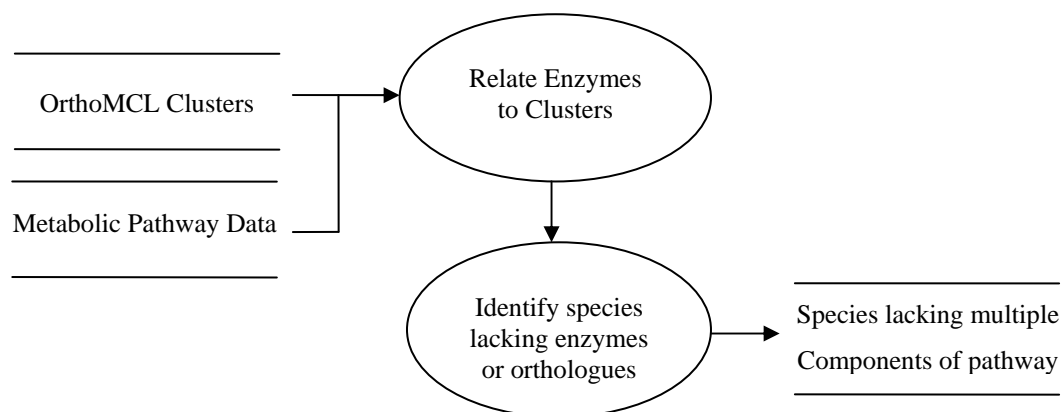


Figure 6: A data flow diagram for studying the conservation of pathways.

4 Implementing the methodology

To support analyses such as those mentioned above, primary genome data is integrated with derived data, generated using large-scale analyses of all the genome data, as well as publicly available functional annotation and pathway data. Primary genome data, obtained from a variety of data sources, in a variety of files, is parsed and mapped onto the part of the database schema representing genomic data. Result files generated by the analyses of the genome data, such as MCL or OrthoMCL clustering, are also parsed and stored in the part of the database representing derived data. The same process is followed for functional annotation and pathway data, which is obtained from publicly available data sources.

As part of the population process, the different kinds of data are linked in a way that supports comparative analyses. For example, MCL and OrthoMCL clusters are associated with all the proteins they contain, but also each genome is associated with all the clusters that contain its proteins, to improve the performance of analyses that retrieve all clusters with proteins of a particular genome. A similar approach has been followed for results on sub-cellular localisations of proteins and functional annotation: each localisation is associated with all the proteins for which that localisation has been predicted, whereas the scores associated with the predictions are stored separately and are only associated with a specific protein. Furthermore, genomes have been classified in terms of their growth form, pathogenicity and taxonomy. This classification has been used to group clusters with respect to the genomes they contain. This has been done to improve performance when retrieving all the clusters that contain proteins of genomes exhibiting a particular growth form or are pathogens. In summary, not only has primary genomic data, functional annotation, pathway and derived data been stored in the database, but additional data structures and associations between the data have been materialised to enable comparative analyses of the kind mentioned above to be carried out efficiently.

These analyses are supported by a number of available pre-determined analysis tasks that can be parameterised, also called canned queries. For example, to support the study of redundancy, introduced in Section 3.1, expanded families within a genome or a group of genomes can be identified using the query “Get the number of paralogs for all clusters containing proteins of a given genome”. This query can be used for all genomes of a particular group related to each other, for example, by their taxonomy, pathogenicity or growth form to identify duplications that have occurred more frequently in one species or a group of fungi than another. Particular clusters identified as being of interest due to their composition of proteins from certain genomes but not others can then be studied further using the query “Get the annotation for proteins in a given cluster”. This query returns all the

proteins with their associated annotations, such as, Pfam motifs identified in their sequence, their predicted sub-cellular location, and their functional annotation, if known. Thus, combining the analysis of paralogs with the study of functional annotation of proteins in clusters using the two queries mentioned can be used to identify expanded functions in certain genomes, as illustrated in Section 3.1 and the data flow diagram in Figure 1.

The analysis of conservation and diversification, introduced in Section 3.2, is supported by a number of queries. The filtering of all clusters to generate a set of clusters containing proteins of a given genome can be carried out using the query “Get clusters with proteins of a given genome”. The result of this query lists, for each cluster, the genomes to which the proteins in that particular cluster belong. This enables the identification of clusters that contain only proteins of a particular subset of genomes, such as the *Saccharomyces* “sensu stricto” group of genomes, or clusters with an unusual distribution of orthologs among genomes. The clusters of interest can then be further analysed using the above mentioned query to retrieve all the annotation for proteins in a given cluster. This can help to identify features common to the sets of clusters of interest, as motivated in Section 3.2.

Examining essential and non-essential *S. cerevisiae* genes and identifying their orthologs in other genomes, as introduced in Section 3.3, is supported by the query “Get all the clusters containing proteins of a given genome and proteins of only essential yeast genes”. A similar query is provided for clusters containing only non-essential yeast gene products. Both these queries can be used to subdivide OrthoMCL clusters into those containing essential gene products and those with non-essential gene products. Again, both queries return all the clusters matching the specified parameters along with a list of genomes which the proteins in those clusters belong to. This enables the analysis of the distribution of orthologs of essential yeast gene products among all the genomes and helps identify genes with an unexpected distribution of orthologs among related species.

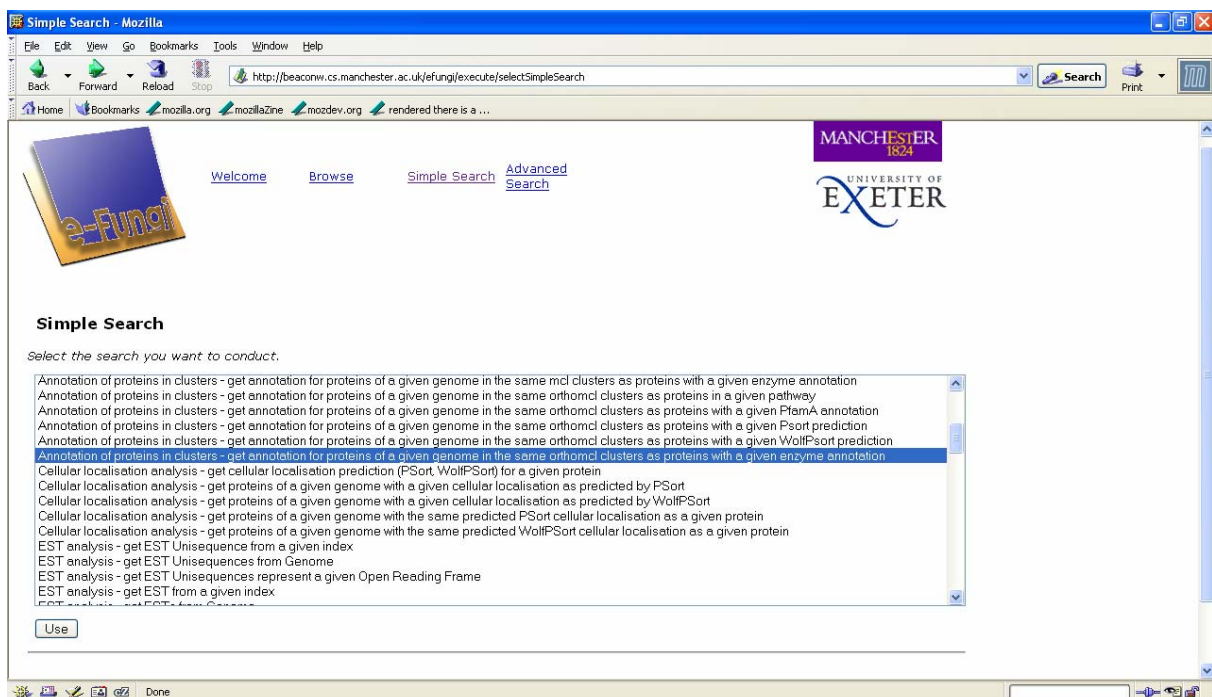


Figure 7: Selecting a canned query in e-Fundi.

The pathway analysis mentioned in Section 3.4 for studying pathways and the conservation of enzymes across species is supported by the query “Get annotation for proteins of a given genome in the same OrthoMCL clusters as proteins with a given enzyme annotation”. Figure 7 shows the selection of this query from the list of all available analysis tasks whereas Figure

8 shows the forms in which the input parameters required for the query can be provided. The counterpart of the query analysing MCL clusters is also provided. This analysis can be carried out for a number of genomes to identify those with and those that lack orthologs for a particular enzyme. This can be done for all enzymes in a particular pathway to determine whether some species are lacking multiple components of certain pathways and can thus be used to study conservation of pathways across species (see Section 3.4 and Figure 6).

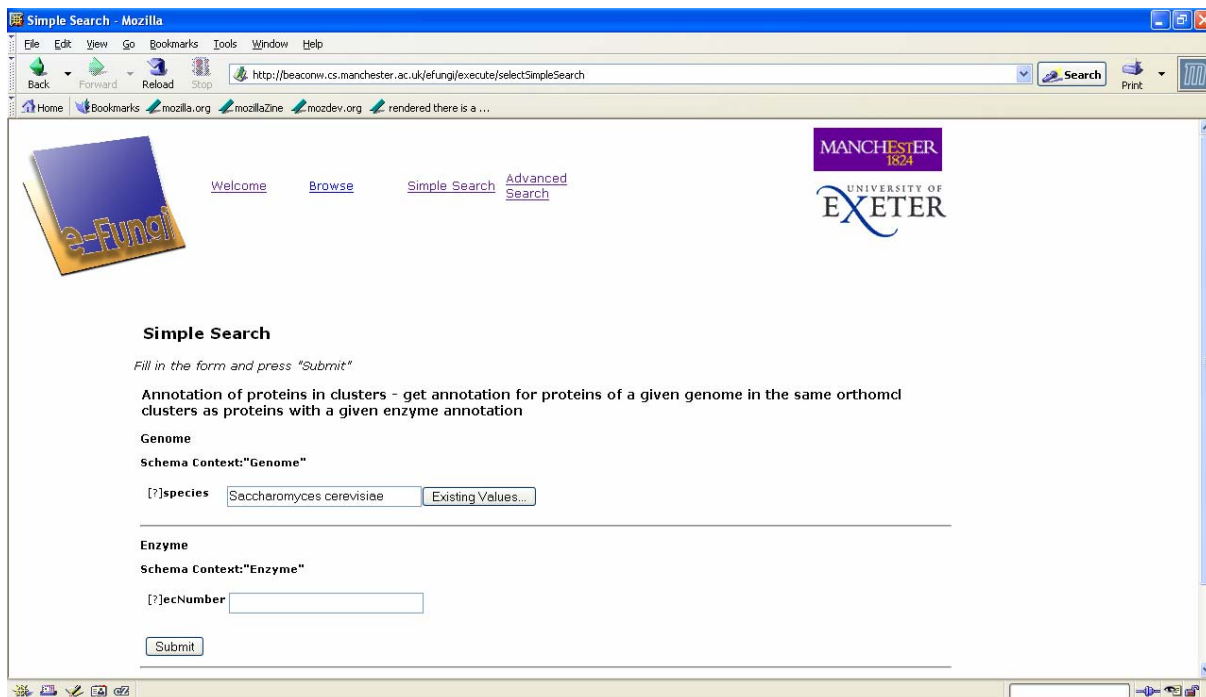
The image shows a screenshot of a Mozilla browser window displaying the 'Simple Search' page of the e-Fungi database. The browser's address bar shows the URL 'http://beaconw.cs.manchester.ac.uk/efungi/execute/selectSimpleSearch'. The page features a navigation menu with links for 'Welcome', 'Browse', 'Simple Search', and 'Advanced Search'. The University of Exeter logo is visible in the top right corner. The main content area is titled 'Simple Search' and includes the instruction 'Fill in the form and press "Submit"'. Below this, there is a descriptive sentence: 'Annotation of proteins in clusters - get annotation for proteins of a given genome in the same orthomcl clusters as proteins with a given enzyme annotation'. The form is divided into two sections: 'Genome' and 'Enzyme'. The 'Genome' section has a 'Schema Context: "Genome"' label and a dropdown menu for '[?]species' with 'Saccharomyces cerevisiae' selected and an 'Existing Values...' button. The 'Enzyme' section has a 'Schema Context: "Enzyme"' label and a text input field for '[?]ecNumber'. A 'Submit' button is located at the bottom of the form.

Figure 8: Providing parameters for a canned query in e-Fungi.

The canned queries mentioned here are by no means comprehensive lists of available analyses. Currently, more than 90 queries are provided, supporting a variety of different analyses such as secretome analysis, study of sub-cellular localisations of proteins, or transcript abundance analysis, to mention but a few additional types of analyses.

5 Conclusions

This paper has presented a methodology for comparative functional genomics, in which a collection of sequence-based analyses provide a foundation for the systematic exploration of relationships between genomes. Of course, other researchers have sought to support the systematic study of multiple genome sequences. Many genomic databases principally support search and gene-centred analyses on comprehensive genome data collections (e.g. [22]). The emphasis in this paper is not so much on archiving and annotating data collections, but rather is on genome-wide analyses. This has more in common with the Genome Atlas [23], which carries out and stores the results of many analyses on a large number of species. However, these analyses serve principally to provide a rich functional annotation of the genomes in the database, and thus are complementary to the cluster-centred approach proposed here. Others have sought to support comparative analyses directly; for example the Microbial Genome Database (MGDB) builds principally on orthologous groups [24]. However, a difference between MGDB and the approach proposed here is that in this paper we emphasise the development of higher-level analyses that build on top of different clustering schemes, rather than on querying the clusters directly.

Overall, current practice is extremely varied in the integrated management and analysis of genome-scale data sets. This paper seeks to characterise functional analyses independently of a specific database, thereby identifying analyses that can potentially be implemented as extensions to many existing platforms. The hope is that documenting a methodology for the analysis of genomic data can help to inform discussion on how best to manage and analyse genomic data, and also contribute to the design of future genome data resources.

References

- [1] Dujon B, *et al.* Genome evolution in yeasts. *Nature*, 430(6995): 35-44, 2004.
- [2] Kellis M, Patterson N, Endrizzi M, Birren B and Lander ES. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423(6937): 241-254, 2003.
- [3] Galagan JE, *et al.* Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*. *Nature*, 438(7071): 1105-1115, 2005.
- [4] El-Sayed NM, *et al.* Comparative Genomics of Trypanosomatid Parasitic Protozoa. *Science*, 309(5733): 404-409, 2005.
- [5] Frazer KA, Elnitski L, Church DM, Dubchak I and Hardison RC. Cross-species sequence comparisons: a review of methods and available resources. *Genome Research*, 13(1):1-12, 2003.
- [6] Chain P, Kurtz S, Ohlebusch E and Slezak T. An applications-focused review of comparative genomics tools: capabilities, limitations and future challenges. *Briefings in Bioinformatics*, 4(2):105-123, 2003.
- [7] Wei L, Liu Y, Dubchak I, Shon J and Park J. Comparative genomics approaches to study organism similarities and differences. *Journal of Biomedical Informatics*, 35(2):142-150, 2002.
- [8] Ouzounis CA and Karp PD. The past, present and future of genome-wide re-annotation. *Genome Biology*, 3(2):COMMENT2001, 2002.
- [9] Haifeng L and Loo-Nin T. Performance Evaluation of Protein Sequence Clustering Tools. *International Conference on Computational Science*, 2:877-885, 2005.
- [10] Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ. Basic local alignment search tool. *J Mol Biol*, 215:403-410, 1990.
- [11] Enright AJ, Van Dongen S and Ouzounis CA. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*, 30:1575-1584, 2002.
- [12] Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS, Pagni M and Sigrist CJA. The PROSITE database. *Nucl Acids Res*, 34:D227-230, 2006.
- [13] Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, *et al.* Pfam: clans, web tools and services. *Nucleic Acids Res*, 34:D247-251, 2006.
- [14] Li L, Stoeckert CJ, Jr. and Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*, 13:2178-2189, 2003.
- [15] Alam I, Hubbard SJ, Oliver SG and Rattray M. A kingdom-specific protein domain HMM library for improved annotation of fungal genomes. *BMC Genomics*, 8:97, 2007.

- [16] Gouzy J, Corpet F and Kahn D. Whole genome protein domain analysis using a new method for domain clustering. *Computational Chemistry*, 23:333-340, 1999.
- [17] Delsuc F, Brinkmann H and Philippe H. Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet*, 6:361-375, 2005.
- [18] Guindon S and Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*, 52:696-704, 2003.
- [19] Farris JS. Phylogenetic Analysis under Dollo's Law. *Systematic Zoology*, 26:77-88, 1977.
- [20] Kunin V and Ouzounis CA. GeneTRACE-reconstruction of gene content of ancestral species. *Bioinformatics*, 19:1412-1416, 2003.
- [21] Goto S, Nishioka T and Kanehisa, M. LIGAND database for enzymes, compounds and reactions, *Nucleic Acids Research*, 27:377-379, 1999.
- [22] Markowitz, VM, *et al.* The integrated microbial genomes (IMG) system, *Nucleic Acids Research*, 34(Database Issue):D344-D348, 2006.
- [23] Hallin, PF and Ussery DW, CBS Genome Atlas Database: a dynamic storage for bioinformatics results and data, *Bioinformatics*, 20:3682-3686, 2004.
- [24] Uchiyama, I. MBLD: a platform for microbial comparative genomics based on the automated construction of orthologous groups, *Nucleic Acids Research*, 35(Database Issue):D343-D346, 2007.