Applications of Combinatorial Pattern Discovery in Computational Genomics

Nikos Darzentas

Wolfson College

A dissertation submitted to the University of Cambridge for the degree of Doctor of Philosophy

European Molecular Biology Laboratory
European Bioinformatics Institute
Wellcome Trust Genome Campus
Hinxton, Cambridge, CB10 1SD
United Kingdom

Email: nikos@ebi.ac.uk

26 November 2005

To my mother and father and sister, for reasons which would easily cover this Thesis,
and another, and another...

To all those who have put up with me, especially Melanie,
and to all those who have shown me respect, my life’s fuel.

This Thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

This Thesis does not exceed the specified length limit of 300 pages as defined by the Biology Degree Committee.

This Thesis has been typeset in 12pt font according to the specifications defined by the Board of Graduate Studies and the Biology Degree Committee.

Applications of Combinatorial Pattern Discovery in Computational Genomics: Summary

This Thesis describes various applications of combinatorial pattern discovery in protein sequence analysis and computational comparative genomics. The work is the result of more than three years of personal and collaborative research towards the completion of my doctoral research in computational biology and bioinformatics.

In this Thesis, Chapters 1 to 4 correspond to an introduction that describes aspects of computational genomics, resources used and specific components developed by the candidate. **Chapter 1** forms an introduction to Bioinformatics, and discusses key issues of data availability and integration. **Chapter 2** introduces joint work involving the monitoring of genome sequences and their re-distribution, as well as the availability of scientific literature for the corresponding species. **Chapter 3** describes key resources jointly developed and further used in the following Chapters, including the COGENT genome sequence database and its various extensions, the OFAM database of reciprocal best hits corresponding to putative orthologs, and the integration of genome information into metabolic pathways using the BioCyc collection. **Chapter 4** discusses other methods that were used in the analyses presented in this Thesis, including TRIBE-MCL, GeneTRACE and Cluster Annotator.

Chapter 5 is an interlude, describing the representation of the tree of life as a network and the reconstruction of the last universal common ancestor. The phylogeny of the microbial world with vertically and horizontally inherited gene flows results in the net of life, a new representation of phylogenetic relationships. The subsequent analysis uses ancestral reconstruction for gene content to infer a minimal estimate for the genome of the last universal common ancestor.

The remaining Chapters, namely Chapters 6 to 9, address issues of sequence analysis and in particular the detection of distantly related protein families. **Chapter 6** addresses sequence comparison, sequence alignment, and methods for the detection of remote sequence similarity. **Chapter 7** advocates combinatorial pattern discovery as an attractive alternative to widely used methods, in terms of computational efficiency, biological accuracy and general applicability, where a number of other collaborative projects are also described. **Chapter 8** presents a well-defined control experiment for the detection of the blue-copper binding domain across distantly related protein families. Finally, **Chapter 9** extends the previous work across the entire range of the protein fold hierarchy and is reported as work in progress.

Preface

This Thesis is the end result of an adventurous and challenging but also immensely rewarding Ph.D. career at the European Bioinformatics Institute (EBI), an outstation of the European Molecular Biology Laboratory (EMBL). It started in October 2001 with an excellent course in Heidelberg (the EMBL headquarters), and ends late 2005. I was awarded an EMBL predoctoral fellowship to work with the Computational Genomics Group under the supervision of Dr. Christos Ouzounis.

Firstly, I would like to thank Dr. Christos Ouzounis, for his supervision in general, and in particular for those long late night meetings at the EBI that have taught me a lot about science and life in equal measure.

Then, a very loud thank you to every member of the Computational Genomics Group, past and present, long-term and visiting, everyone, for making this worthwhile and being there in the good times and the bad.

Also, I would like to acknowledge familiar faces at the EBI and EMBL who shared science, sports, lunches and dinners and coffee breaks with me. To my fellow PhD students, I would like to say that it has been a privilege to be around so talented individuals, and of course wish everyone best of luck!

Finally, to my Cambridge University Automobile Club (CUAC) friends, I would like to send my appreciation for not throwing me in tyre walls too often during kart races, and for making it possible for me to have so much fun racing with and against them.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The following few paragraphs constitute a brief preface to Bioinformatics and areas of the discipline that have been of interest to us during the research described elsewhere in this Thesis.

Additional introductory information, relevant to **Chapters 5** and **7-9**, can be found in **Chapters 4** and **6** respectively.

1.1 Bioinformatics

It has been approximately 40 years since the first biological sequence was tediously completed, a 75-base long tRNA molecule (Holley, Everett *et al.* 1965). It would take another decade until various laboratories were able to sequence entire genes (Sanger and Coulson 1975) (Sanger, Coulson *et al.* 1982), and a further two decades in the case of whole genomes (Fleischmann, Adams *et al.* 1995).

Naturally, it was not long before mathematicians and computer scientists took a real interest in deciphering this digital code, a linear string of characters encoding the blueprint for the assembly and operation of cells and organisms.

As DNA sequencing became a fundamental molecular biology tool, the large volumes of DNA data warranted the establishment of Bioinformatics, a separate field of study in which biology, computer science, and information technology merge to form an intriguing discovery playground. Subsequently, a number of concepts emerged that remain central to the discipline today.

It is known that biological systems are extremely complex, dynamic systems, and DNA and protein molecules are no exception. Yet, a remarkably simple abstraction, i.e. the representation of DNA and protein molecules as linear strings, has made it possible for current computer resources to be able to manage the billions of bytes of information, and has consequently led to the wealth of information the scientific community shares and enjoys today. This information is exploited in areas like molecular medicine (drug targets, personalized medicine, preventative medicine, and gene therapy), microbial genome applications (food technology, waste cleanup, climate change, alternative energy sources, antibiotic resistance, evolutionary studies, even biological warfare), and agriculture (enhanced crops with traits like insect and drought resistance and improved nutritional qualities).

Selection has put its mark on every sequence we see today, arguably the main force being the conservation across organisms of important sequence segments. Simplistically, DNA sequence determines protein sequence, protein sequence determines protein structure, and protein structure determines protein function. Proteins with significant sequence and structural similarity tend to possess functional similarity. Of course, evolutionary relationship, or divergent evolution, is not the only path to functional similarity: in numerous occasions, the physics of molecular interactions alone have pushed proteins to adopt similar structural conformations resulting in similar functions – the concept of convergent evolution. A considerable number of algorithms in Bioinformatics attempt to detect similarity (see **Chapter 6**).

Finally, there is the realization that Bioinformatics is far from being a theoretical science. The sheer volume of information in need of gathering, storing, and understanding, has many scientists, but also vast computational resources, busy. The signs are good: central processing unit manufacturers are managing to pack more and more components on their silicon chips, and there are predictions of imminent ground-breaking technologies in the field that will finally overturn the infamous Moore’s law (Moore 1965) (**Figure 1**).

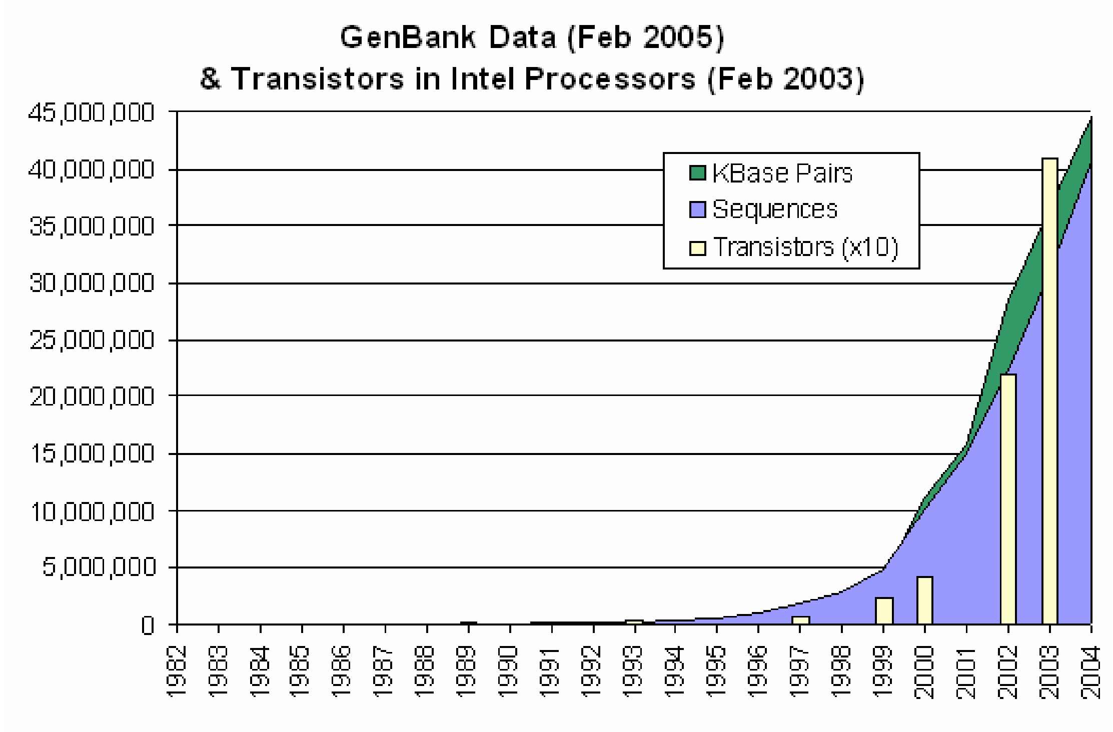


Figure 1. GenBank data in KBase pairs and sequences, and the number of transistors in Intel processors, over time (values during the first decade of the graph are barely visible, highlighting the extraordinary progress made during the second decade).

1.2 Comparative Genomics and Phylogenetic Studies

The Human Genome Project, a huge undertaking in many respects, has made comparative genomics a very high profile area of research. Now that the human genome has been made available, the international sequencing laboratories have moved on deciphering many others (Bernal, Ear *et al*. 2001) to add to those of higher organisms like mouse, rat, fruit fly, zebrafish, chicken, mosquito, and chimp, few eukaryotes, and numerous prokaryotes.

The reason behind this is the power of comparative genomics. The underlying theory is that common features of two organisms will often be encoded within the DNA that is conserved between them, whereas sequences that encode (or control the expression of) molecules responsible for differences between them will themselves be different.

Different phylogenetic distances between genomes in comparison answer different questions. Very short distances (e.g. between humans and chimpanzees) highlight the key sequence differences developed under positive selection that may account for the differences in the organisms. Moderate distances lead to both functional and non-functional DNA being found within the conserved regions, with the functional sequences showing a signature of negative selection; this approach is also contributing to identifying the general function of certain DNA segments, such as coding exons, non-coding RNAs, and gene regulatory regions. Finally, very long phylogenetic distances can help tackle questions about the core set of genes of the last common ancestor of two species, or the controversial subject of the Last Universal Common Ancestor (see **Section 5.2**).

1.3 Structural and Functional Genomics

Structural genomics involve the systematic determination of all macromolecular structures represented in a genome. X-ray crystallography and Nuclear Magnetic Resonance (NMR) have produced the tens of thousands of protein 3-Dimensional (3D) structures currently residing in the Protein Data Bank (PDB) (Deshpande, Addess *et al*. 2005). Subsequently, projects and databases like the Structural Classification Of Proteins (SCOP) (Andreeva, Howorth *et al*. 2004) and the ASTRAL Compendium for Sequence and Structure Analysis (Chandonia, Hon *et al*. 2004), handle this data and provide invaluable datasets of proteins domains (see **Section 1.4** for an overview) in a hierarchical scheme, highlighting evolutionary, structural and functional relationships.

As a matter of fact, protein structures cover only a very small percentage of the publicly available protein universe (currently ~0.075% according to the PDB and GenBank (Benson, Karsch-Mizrachi *et al*. 2005)), and although they truly are indispensable in studying and understanding protein evolution and function, there are a number of other large projects looking at protein domains directly on the primary sequence level. Examples include Pfam (Bateman, Coin *et al*. 2004), and InterPro (Mulder, Apweiler *et al*. 2005).

1.4 Protein Domains and Sequence Patterns

A protein domain is a discrete portion of a protein assumed to fold independently of the rest of the protein, and to possess evolutionary, structural, and functional significance. The importance of domains is that they cannot be divided into smaller units, thus they represent fundamental building blocks that can be used to understand the evolution and function of proteins.

Most proteins are multi-domain - the interactions between protein domains can be complex and instrumental for protein function. This concept is also integral in creating new proteins by bringing domains together, i.e. domain shuffling. It is thought that this is a major way that new proteins have arisen during evolution. Consequently, mining of databases for homology by domains, rather than by whole proteins (which are not as evolutionarily conserved), is important in obtaining clues to functionality. Protein sequence patterns frequently represent highly conserved parts of domains, thus can be considered as the true core signal of evolution, structure, and function.

1.5 Data Availability, Integration and Knowledge Discovery

Cutting edge applications in the life sciences, e.g. molecular biology, biodiversity research, drug discovery and personalized medical research, increasingly depend on bioinformatics methods to manage and analyze vast amounts of considerably diverse and complex data.

Consequently, there has been an unprecedented explosion in the number and size of public data resources, and a rapid growth in the variety and volume of laboratory data, from microarrays measuring gene transcription levels and genome sequence variations, to protein interaction screens measuring components of protein complexes. This has been fuelled by both world-wide research activity and the emergence of technologies such as mass spectrometry for protein identification. The modelling, management, and analysis of this data often requires a comprehensive integration of heterogeneous and typically semi-structured data, distributed across many data sources. Relatively recent interoperability standards, such as the Extensible

Markup Language (XML - http://www.w3.org/XML/), alleviate some problems, but data and process integration remain time-consuming and error-prone (usually manual) tasks. The difficulty of these tasks is compounded by the high degree of semantic heterogeneity across data sources, varying data quality, as well as domain specific application requirements. Bioinformatics is an inherently integrative discipline, requiring access to data from a wide range of sources. Without the underlying data, and the ability to combine these data in new and interesting ways, the field of bioinformatics would be much limited in scope.

These hardships offer unprecedented challenges that researchers seem happy to face. The *Nucleic Acids Research* Molecular Biology Database Collection is a public online resource that lists the databases described in issues of *Nucleic Acids Research*. All databases included in this Collection are freely available to the public. The 2005 update includes 719 databases, 171 more than 2004 (Galperin 2005). As already mentioned, the immense challenge we have is to integrate these towards novel levels of understanding.

1.6 Conclusion

Genomic research makes it possible to look at biological phenomena on a scale not previously possible. Our research exploits this possibility in diverse ways towards a better understanding of mechanisms of evolution and function.

1.7 Disclaimer

Several World Wide Web sources have been used to compile this **Chapter**.

Chapter 2
Genome explosion

This **Chapter** introduces joint work involving the monitoring of genome sequences and their re-distribution, as well as the availability of scientific literature for the corresponding species.

The information contained in this **Chapter**, and especially in **Section 2.1**, has not been updated to reflect current status, in order to maintain discussion points and conclusions.

2.1 Beyond 100 genomes

Since the publication of the first entire genome sequence seven years ago (Fleischmann, Adams *et al*. 1995), a multitude of other genomes have been – or are in the process of being – sequenced (Nelson, Paulsen *et al*. 2000). By the end of 2002, we have witnessed the landmark submission of the 100th complete genome sequence in the databases (Akman, Yamashita *et al*. 2002). There are now 106 complete genomes in the public domain, thanks to advances in sequencing technology and sustained funding. An overview, and in particular the rank ordering, of these genomes reveals certain interesting trends and provides valuable insights into possible future developments.

First, the contribution of genome sequencing projects in terms of actual protein sequence entries has been staggering. There are 433,238 protein sequences derived exclusively from entire genomes[1] (**Figure 2**), out of a total of a million protein sequences known to date. In contrast, there are only 101,602 entries in SwissProt (release 40), underlining the significant effort that is required for high-quality annotation (Bairoch and Apweiler 2000). The growth of protein sequence data coming from entire genomes is expected to reach over one million entries in two years' time (**Figure 2**). Given that approximately 40% of genes of any organism

[1] http://cgg.ebi.ac.uk/services/cogent/

cannot be assigned to a specific functional role (Iliopoulos, Tsoka *et al*. 2001), this suggests that in just a few years hundreds of thousands of sequences will be uncharacterized. While the large-scale characterization of protein function obtained from high-throughput experimental techniques (Martzen, McCraith *et al*. 1999) will alleviate some of the above problems, it is clear that more research should also be devoted to the development of intelligent automated genome annotation systems that are able to predict functional properties of protein sequences (Andrade, Brown *et al*. 1999), to capitalize on the information explosion of genome biology.

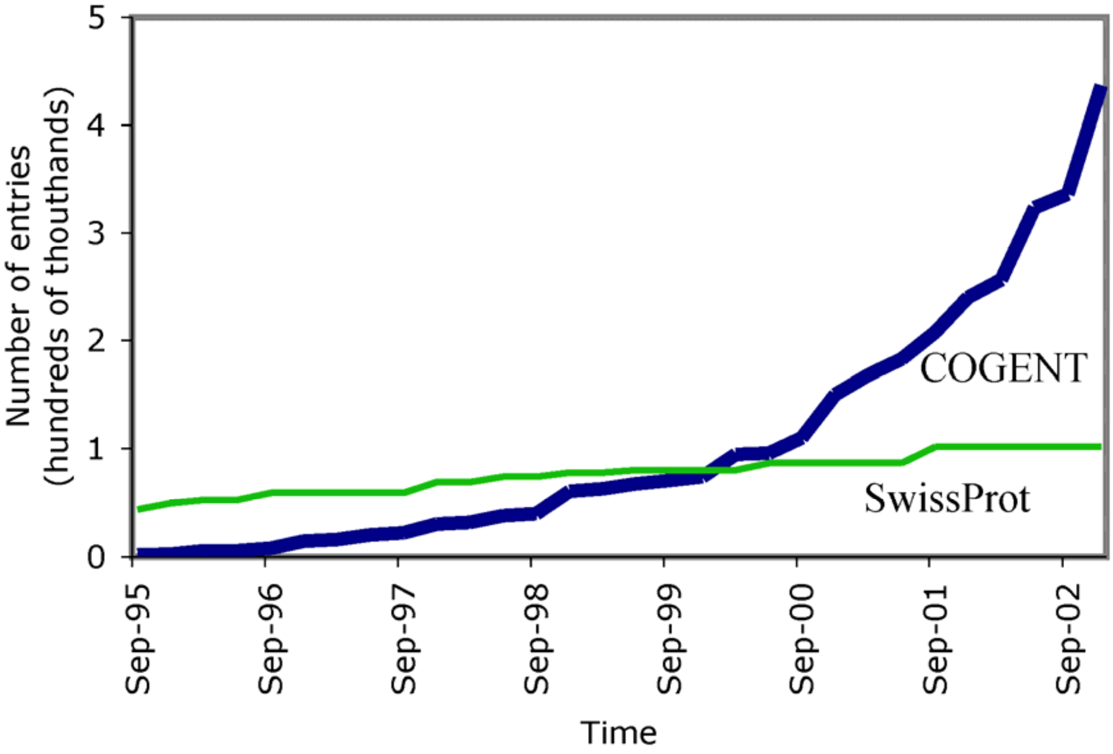


Figure 2. Cumulative number of protein sequence entries (y-axis) in completed genomes (CoGenT, in blue) and SwissProt (in green) as a function of time (x-axis).

Second, in addition to the well-defined collection of 106 completed and published genomes, there are another 544 ongoing projects, covering a large number of taxa. Yet, the known taxa of Bacteria and Archaea are far better represented amongst the completed genome projects, compared to Eukarya (**Figure 3**). Using comparative genomics, we have already obtained a glimpse of the bewildering biological diversity of the prokaryotic world (Torsvik, Ovreas *et al*. 2002). Very soon, a similar trend might emerge for the Eukarya: 208 out of the 544 ongoing genome projects are dedicated to eukaryotic species. However, many eukaryotic taxa are still not represented (**Figure 3**). A better sampling of phylogenetic diversity might be required, to fully explore the genomes of eukaryotic cells.

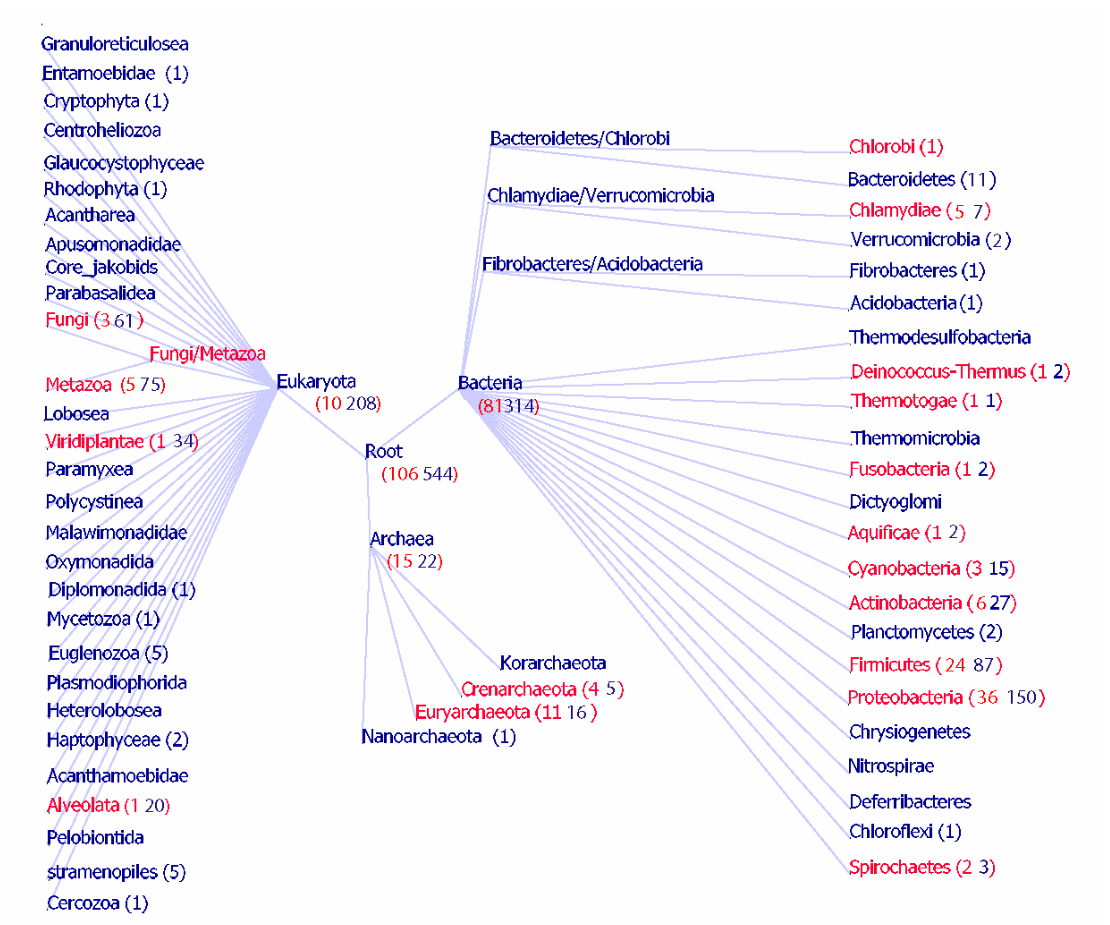


Figure 3. Phylogenetic distribution of genome sequencing projects. Archaea and Bacteria are shown to the Phyla level and Eukaryotes are shown to their first taxonomic branching, with the exception of Metazoa and Fungi. The numbers in parentheses represent the number of completed, published (red) and ongoing (blue) genome projects. The tree is based on the Taxonomy database from the National Center for Biotechnology Information (NCBI). Information about ongoing genome projects has been obtained from the Genomes OnLine Database (GOLD), as of 22 January 2003.

Third, over time, both the range of sequenced genome sizes and the selection of species on the basis of their social impact has expanded (Doolittle 2002) (**Figure 4**). Sequenced genome sizes range from 0.5 to 300 Megabases, with the exception of the human and mouse genomes, which span 2,900 and 2,500 Megabases respectively (and together constitute almost 90% of the 106 available DNA sequences data). Although species of medical and academic interest were initially the main targets of genome projects, there has been a recent trend to sequence genomes from species with impact on agriculture, environmental sciences or industrial processes. In addition, a growing number of genomes are being sequenced in order to provide a better perspective for the structure and function of evolutionarily related genes and genomes through comparative analysis.

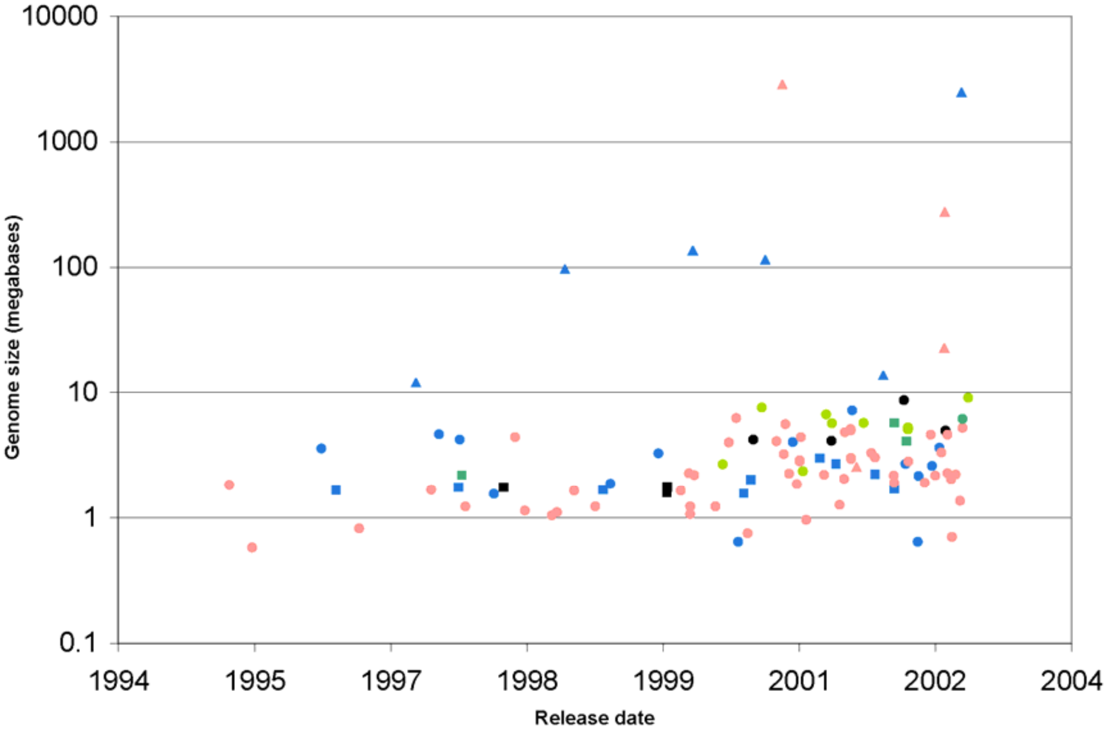


Figure 4. Representation of completed genome sequences over time (x-axis) and size (in Megabases, y-axis - logarithmic) labelled according to their social impact. Genomes from Archaea (squares), Bacteria (circles) and Eukarya (triangles) are collared according to their academic (blue), medical (red), agricultural (green), ecological (cypress green) and industrial (black) relevance.

Thus, ten years after the computational analysis of the first eukaryotic chromosome (Bork, Ouzounis *et al.* 1992) and seven years after an exhaustive analysis of the first complete genome (Casari, Andrade *et al.* 1995; Fleischmann, Adams *et al.* 1995), genomic science has become a stand-alone discipline, and genome sequencing and computational analysis have become mutually dependent, intertwined in a fascinating interplay. Not so long ago, it would have been unthinkable that from a set of DNA fragments, it would be possible to assemble them into a single genome, find the genes, translate the proteins, identify their potential functional roles and ultimately integrate all this structural and functional information into complex biochemical networks (Eisenberg, Marcotte *et al.* 2000). Although there are still significant challenges, these technologies, along with scientific advances, have now come of age and are expected to have a growing impact on various aspects of human welfare.

2.1.1 Disclaimer

This work has been a collaborative effort with the Computational Genomics Group, and has been published in (Janssen, Audit *et al.* 2003).

2.2 Genome coverage, literally speaking - The challenge of annotating 200 genomes with 4 million publications

In late 2004, 200 complete genomes had been sequenced and made available to the research community. At the time of writing, that number had further risen to 221 and will have undoubtedly increased again before publication. These genomes, which represent a wide range of species from archaea to human, are a highly valuable knowledge resource for the scientific community. However, the sequencing of a full genome is just the first step in research; it must be followed by the functional characterization of genes and proteins. In this context, it is interesting to see how well represented these sequenced species are in terms of publications on their genes and proteins. We have thus obtained the number of abstracts per species and normalized that count by the number of genes in that species to obtain a comparable measure for the number of publications per gene for all completed and published genomes. This simple measure highlights the current knowledge gap between various organisms and could further serve as a guideline for selecting genomes for sequencing projects, high-throughput functional genomics and database annotation efforts.

The 200 complete genome sequences published by December 2004 include 118 genera, 166 species and 34 additional strains for 21 species. This rate translates to a doubling time for genome sequence availability of less than two years (Janssen, Audit *et al*. 2003). And it remains steady: in 2003, an average of one complete genome was released per week; 47 genomes were made available in the first 44 weeks of 2004 (rank numbers 154-200). This trend will accelerate further, as more than 1,000 genome projects are currently underway (Bernal, Ear *et al*. 2001).

For the 221 genomes currently available, the total number of predicted proteins is 822,114, according to the CoGenT database (Janssen, Enright *et al*. 2003). One of the great challenges for computational and experimental genomics is the functional characterization of the genes and proteins encoded in these genomes (Eisenberg, Marcotte *et al*. 2000), a process that should be considered as continuous (Ouzounis and Karp 2002). To achieve this goal, it is important to rely on existing knowledge and draw from previous studies conducted on these organisms. We have analysed how well each of the currently available genomes is characterized in terms of the number of publications pertinent to the corresponding species. To achieve a reliable measure for the available knowledge per genome, we obtained the number of abstracts per species–but not strain–from Medline and divided this number by the number of predicted protein-encoding genes. The corresponding ratio, which we term the ‘Species Knowledge Index’ or SKI, thus reflects our current understanding of each of these species. More detailed information on the literature coverage of sequenced organisms, tracked by CoGenT, is available through our GenMed server (http://cgg.ebi.ac.uk/cgi-bin/genmed/genmed.pl).

This analysis encompasses 3,806,293 Medline abstracts corresponding to 200 genomes. On average, there are 5 abstracts per gene for the first 200 genomes in the CoGenT database, ranging from 1 abstract per 1000 genes for poorly characterized species to 55.1 abstracts per gene for *Escherichia coli* (**Figure 5**). This arithmetic mean value is grossly distorted by a few outliers (**Figure 6**), namely *E. coli*, human

(48.5), *Staphylococcus aureus* strains (ca. 16-17), mouse (15.6), *Helicobacter pylori* strains (ca. 13) and *Saccharomyces cerevisiae* (10.6). If these outliers are excluded, then the average SKI value drops to 0.9 (580,016 abstracts for 651,183 genes). Yet, this value is still dominated either by important pathogens, such as *Chlamydia trachomatis* (9.5) and *Haemophilus influenzae* (8.9), or by model organisms such as *Pseudomonas aeruginosa* (5.7) and *Bacillus subtilis* (4.8).

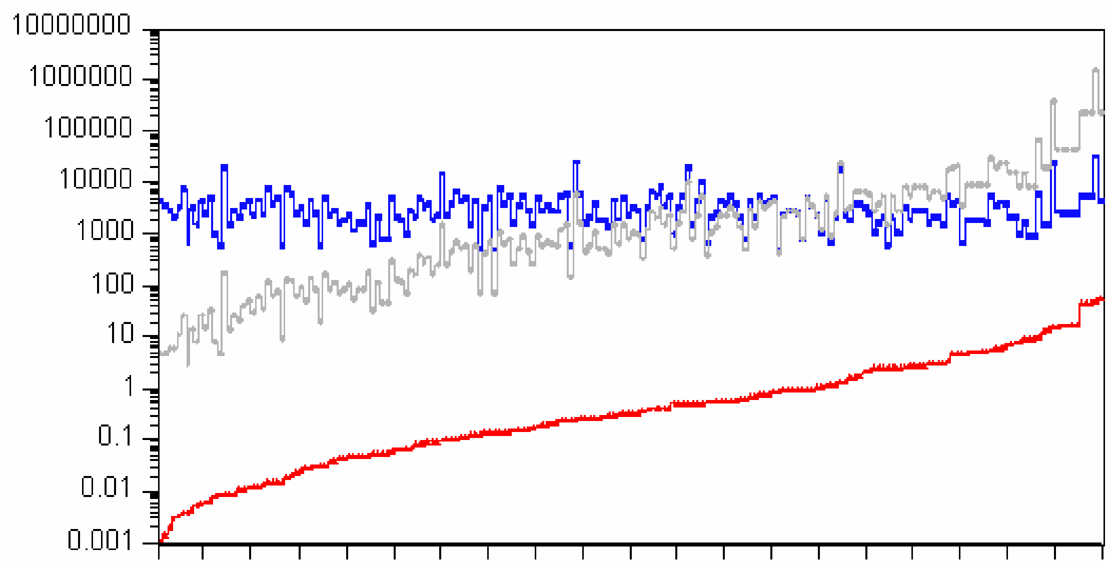


Figure 5. Relationship between the number of genes and number of abstracts for the first 200 genomes made available in CoGenT (v. 207, 29 Oct 2004; abstracts collected from Medline on 5 Nov 2004). The x-axis represents the genome rank, sorted by SKI value corresponding to the order of appearance (release and publication). The y-axis corresponds to the number of genes (blue), abstracts (grey) and the SKI value (red), on a logarithmic scale. It is evident that there is a range of two orders of magnitude for gene numbers and a range of six orders of magnitude for abstract numbers, which results in a range of four orders of magnitude for the SKI value, between 0.001 and <100.

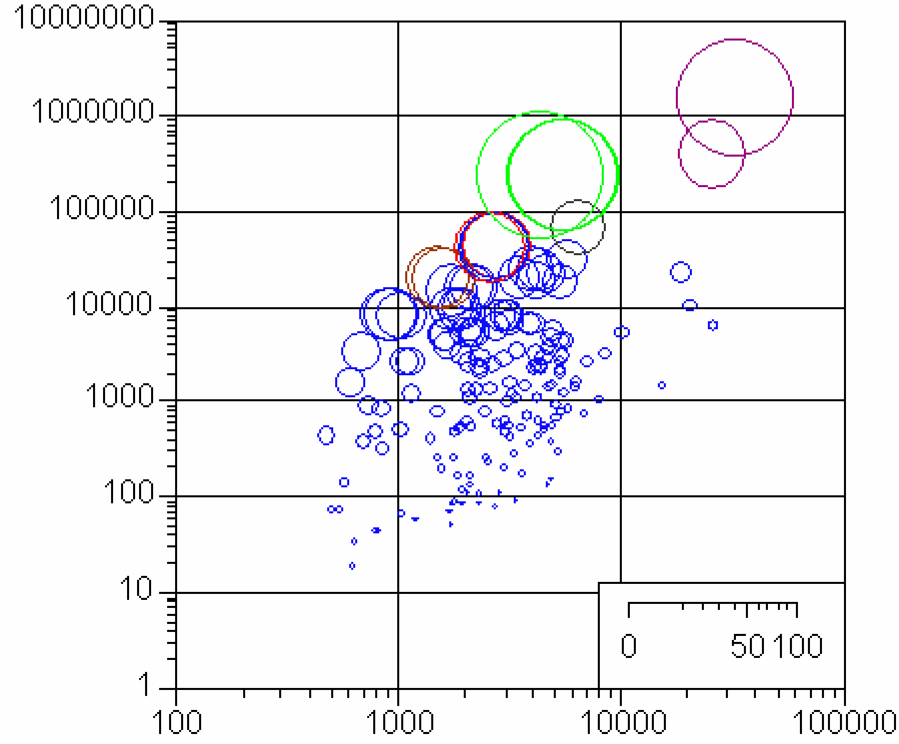


Figure 6. Relationship between the number of genes and number of abstracts for the first 200 genomes made available in CoGenT (v. 207, 29 Oct 2004; abstracts collected from Medline on 5 Nov 2004). The x-axis represents the number of genes, the y-axis the number of abstracts and the radius of circles the SKI value (scale shown). The five principal outliers mentioned in the text are coloured as follows: *Escherichia coli* strains (green), human-mouse pair (purple), *Staphylococcus aureus* strains (red), *Helicobacter pylori* strains (brown) and *Saccharomyces cerevisiae* (grey). It is evident that there is a range of two orders of magnitude for gene numbers and a range of six orders of magnitude for abstract numbers, which results in a range of four orders of magnitude for the SKI value, between 0.001 and <100.

Other model species follow at lesser ranks, for instance *Drosophila melanogaster* (1.3), *Schizosaccharomyces pombe* (1.1) and *Caenorhabditis elegans* (0.5). The most studied archaeal species are *Methanobacterium thermoautotrophicum* and *Methanococcus jannaschii,* both with SKI values of 0.3. At the very end of the scale are species that have been recently characterized or of environmental interest (Janssen, Audit *et al.* 2003). Such examples are *Gloeobacter violaceus* and

Oceanobacillus iheyensis with only a few abstracts, which corresponds to an average of 1 abstract per 1,000 genes. It is worth noting that there is a slightly decreasing tendency for the SKI value with respect to the order in which a genome was completed (not shown).

These results demonstrate a number of limitations to our approach. First, the time of characterization of the corresponding species is not taken into account; obviously, recently characterized species have lower SKI values. A rate value that takes into account the number of abstracts per gene per year would produce a more accurate estimate for the increase in scientific interest for each organism. Second, naming conventions are not strictly followed in the literature, thus making it difficult to accurately retrieve the number of abstracts for *Homo sapiens*, for example. Not all abstracts that contain the word ‘human’ necessarily relate to molecular biology and genes, but they might contain relevant information; as a filter, we used the keyword ‘protein’ for human and mouse to focus on molecular information. In addition, species that are used in biotechnology, such as *E. coli*, are unquestionably over-represented. Similarly, there is no sufficient resolution for strain names; however, it is safe to assume that closely related strains have similar properties and the body of existing knowledge might therefore be easily transferable across strains in most cases. Third, Medline indices primarily cover medical literature and related biological sciences. It is therefore conceivable that journals that publish studies on organisms of environmental or industrial interest are not included.

The accuracy of the SKI measure for organisms could further improve if the dynamics of their nomenclature was considered. For instance, *Ralstonia solanacearum* had 159 hits in Medline, but this organism was first classified as *Burkholderia solanacearum* (50 hits) and thereafter as *Pseudomonas solanacearum* (167 hits). The keyword '*solanacearum*' produced 320 hits, close to the sum for all three species (376 hits). The slight discrepancy results from the use of synonyms in abstracts, for instance “…in *Ralstonia solanacearum* (formerly *Pseudomonas solanacearum*), a phytopathogenic bacterium…”. In this particular case, the term '*solanacearum*' would thus result in a more accurate count of scientific papers for the organism *R. solanacearum*, assuming that no other species name contains this term. However, it would be unrealistic to consider the dynamics in nomenclature for every single species. Nevertheless, it may be possible to obtain improved SKI values in the future, assuming that the literature keeps pace with taxonomic modifications. In this respect, it may be worthwhile to use the yearly average of SKI values over a suitably long period and compare those values only with each other. In essence, global and average SKI values will never be static but will change over time, thus reflecting renewed or dwindling scientific interest in a particular organism.

Despite the above limitations–namely the time of species characterization, taxonomic conventions and potential biases–the SKI measure reflects our current status of knowledge for each species or strain at the molecular level and essentially delimits the number of published studies according to the number of genes. Nevertheless, to achieve this improvement, better descriptions of strains and species in the literature are necessary. We have recently shown that current definitions of species and genera in particular are not entirely satisfactory, according to a measure of genome sequence similarity called genome conservation (Kunin, Ahren *et al.* 2005). It is thus imperative that sufficient literature coverage, which reflects the active experimental interest of research communities, is available for a given organism before a genome sequencing project is initiated. Certain communities might not successfully advocate the sequencing of their favourite species, despite its relative importance in terms of available abstracts in Medline, whereas closely related species with less literature support may be sequenced.

To exemplify this point, we have analyzed the group of *Streptomyces* species that have been or are in the process of being sequenced, according to CoGenT and the Genomes OnLine Database (GOLD; (Bernal, Ear *et al.* 2001) respectively. The *Streptomyces* genus is a complex and important group of Actinobacteria, with many unresolved branches and a variety of phenotypic attributes (Anderson and Wellington 2001). Two species whose genomes have been sequenced, *S. coelicolor* and *S. avermitilis*, are represented in Medline by 1,014 and 119 abstracts, respectively, corresponding to SKI values of 0.13 and 0.014. According to GOLD, five other *Streptomyces* species are currently being sequenced: *S. ambofaciens* (116 abstracts), *S. diversa* (no abstracts), *S. noursei* (81 abstracts), *S. peucetius* (87 abstracts) and *S. scabies* (60 abstracts). It is therefore surprising to find that *S. aureofaciens* (365 abstracts), *S. antibioticus* (277 abstracts) and *S. griseus* (1,223 abstracts) are not listed in GOLD as ongoing projects, although all three strains are representative species of key *Streptomyces* groups (Anderson and Wellington 2001).

Obviously, the criterion of published abstracts alone is not sufficient to prioritize genome sequencing targets, yet it provides a rational measure of current knowledge and interest by taking into account the number of published studies. It is conceivable that a more elaborate listing of all species and/or strains could be obtained and ranked by the corresponding number of published abstracts. These corpora could serve as focal points for the experimental communities and would facilitate the identification of 'neglected' organisms that might be considered for genome sequencing in the future. Other uses could obviously benefit computational analysis, including database annotation and text mining.

We believe that the SKI measure demonstrates the significant variation in the number of publications for each genome and the huge challenge of using this literature to accurately annotate these genome sequences. Simply put, we would need to achieve more than 26,000 publications for *Mycoplasma genitalium*–which has 479 genes and is currently covered by about 400 abstracts–to reach the current SKI value for *E. coli*. We have a long way to go.

2.2.1 Disclaimer

My contribution to this work has been data collection and analysis, along with co-authoring the manuscript. The work has been published in (Janssen, Goldovsky *et al*. 2005).

Chapter 3
Developed Databases

This **Chapter** describes key resources jointly developed and further used in the following **Chapters**, including the CoGenT genome sequence database and its various extensions, the OFAM database of reciprocal best hits corresponding to putative orthologs, and the integration of genome information into metabolic pathways using the BioCyc collection.

3.1 Complete Genome Tracking (CoGenT): a flexible data environment for computational genomics

We present a database of fully sequenced and published genomes to facilitate the re-distribution of data and ensure reproducibility of results in the field of computational genomics. For its design we have implemented an extremely simple yet powerful schema to allow linking of genome sequence data to other resources.

3.1.1 Introduction

The number of completely sequenced genomes is constantly increasing, with over 80 entire genome sequences published to date. In principle, this unprecedented resource of genomic data opens up new opportunities for computational biology, by expanding the possibilities of genome-wide investigations. Yet, this data avalanche also challenges the current practices with respect to accessibility, reproducibility and usability. The capacity of linking the genomic sequence data with other information, for example functional classes, cellular localization, chromosomal position and expression profiles clearly leads to new knowledge. However, the combination of diverse data resources for genome analysis can be cumbersome, because of the wide variability of syntactic and semantic conventions deployed by different data repositories. Although some operations should be easy to perform, in practice a significant amount of time is spent to achieve them.

First, although there exist resources that allow full access to genome sequence information, there is a clear lack of a single point of reference that allows flexible and direct access to full-genome information with a few mouse clicks. Most of the web resources are not designed for genome-scale analysis, as they were generally created for a ‘one gene at a time’ mode of browsing and, as a consequence, the raw data are not readily available. This information is necessary for large-scale computational analyses and should thus be conveniently accessible and re-distributable. It is frequently the case that computational biologists may be working on the same data set using different identifiers and may not be able to easily compare their results.

Furthermore, databases are tremendously heterogeneous so that it is in practice impossible to link their contents in an automated way. The primary reason is that the naming conventions for genes are not adequate and might differ from one resource to another. For example, the COGs database uses gene names as they can be found in EMBL or GenBank entries, e.g. AF1241 and BS gsaB, found in COG0001 cluster. In this example, gsaB had to be prefixed by ‘BS’ to uniquely point to the *Bacillus subtilis* gsaB gene, since gsaB is not a unique identifier. In order to map this COG cluster to the InterPro domain database, an additional piece of information, such as the SwissProt identifier or accession number, is also necessary. Therefore, simply linking two commonly used databases for protein sequence annotation can be a challenging task.

Finally, even if the data, naming schemes and their mappings are fully available, file parsing and format conversion can still be a bottleneck in bioinformatics research. For example, the mapping of identifiers and other terms to keywords, various classification schemes or species names can be a challenge. In practice, every single user will have to write their own parser to effectively utilize this information. In our view, the difficulties mentioned above raise a major issue about consistency and reproducibility for computational biology. To overcome some of the obstacles that hinder consistency and reproducibility, we have developed a new data environment for complete genome analysis, called CoGenT.

3.1.2 Implementation

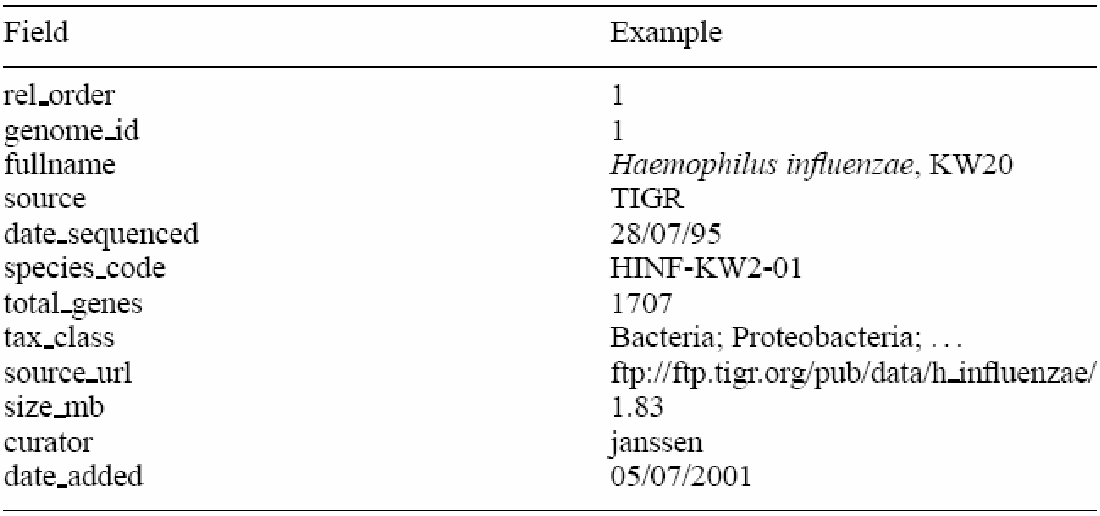
Some ongoing projects already offer functional frameworks to share software development (e.g. http://www.bioperl.org) or high-throughput data analysis. For example, the Ensembl project (http://www.ensembl.org) provides full access by distributing the software (Perl modules) to query its complex MySQL database either on a local installation or on their server. The aim is to provide a simple and practical tool on which to develop computational genomics projects, to facilitate sharing of data and results and, hopefully, to enable synergy in the field. We are successfully using CoGenT in our research group, so that we now advocate its use by the wider community.

CoGenT can be considered a realistic solution, because it is simple, highly flexible with a small usage overhead. Working with CoGenT entails the use of: (i) the core CoGenT database of complete genome sequences, which defines a common naming convention for genomes and proteins (identifiers); and (ii) SQL tables as an exchange format, as well as a potential working environment. These few and simple guidelines allow the convenient exchange of results pertaining to complete genome sequences, necessary for high-throughput computation. Table joining enables the linking of results from different origins and the indexing mechanisms available in relational databases such as MySQL ensure efficient query and retrieval.

3.1.3 Design

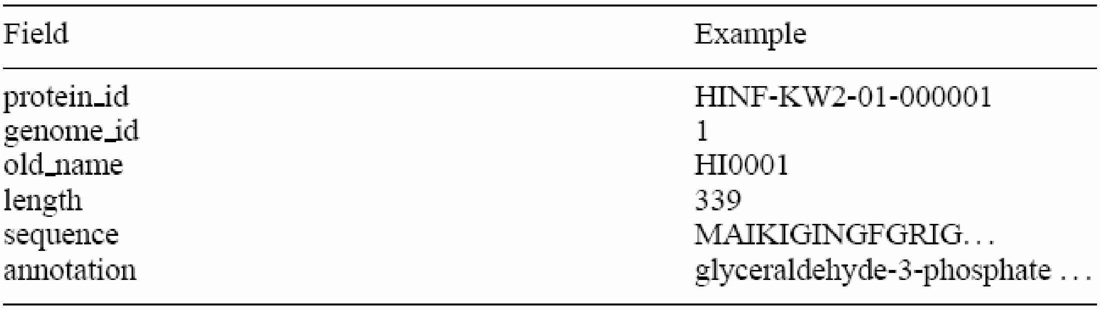
The complete design details are available on the CoGenT web pages, along with Perl DBI-based scripts and modules to perform basic operations, such as retrieving a sequence or a complete genome peptide file in FASTA format. Nevertheless, it is noteworthy that the core database is composed of just two tables. The *genomes* table contains genome-related information (**Table 1**). A design decision was made to keep track of the relative timing of genome sequences in order of their appearance in the literature (*rel_order*). Along with the *genome_id*, a mnemonic species code is generated and used as a prefix to construct unique protein identifiers. This mnemonic notation is based on the genus, species and strain name as well as version number for this genome, in order to enable updates e.g. HINF-KW2-01 corresponds to *Haemophilus influenzae* strain KW2, version 01. The *proteins* table holds the amino acid sequence data (**Table 2**). The unique *protein_id*s are constructed by the *species_code* followed by a dash and a number. The use of mnemonic identifiers enable to clearly distinguish genomes and proteins from a given genome in a simple manner. As new genomes are made available, they are added to the *genomes* and *proteins* tables by the database curators upon release. Whenever possible, genome sequences are downloaded directly from the sequencing centre web sites—otherwise GenBank is used as a source.

Table 1. CoGenT *genomes* MySQL table.



Field	Example
rel_order	1
genome_id	1
fullname	*Haemophilus influenzae*, KW20
source	TIGR
date_sequenced	28/07/95
species_code	HINF-KW2-01
total_genes	1707
tax_class	Bacteria; Proteobacteria; . . .
source_url	ftp://ftp.tigr.org/pub/data/h_influenzae/
size_mb	1.83
curator	janssen
date_added	05/07/2001

Table 2. CoGenT *proteins* MySQL table.



Field	Example
protein_id	HINF-KW2-01-000001
genome_id	1
old_name	HI0001
length	339
sequence	MAIKIGINGFGRIG. . .
annotation	glyceraldehyde-3-phosphate . . .

We endeavour to maintain consistency between the content of the database and published complete genomes.

3.1.4 Usage and future plans

The database has already been used effectively in a number of ongoing projects, including genome sequence clustering, genome annotation and phylogenetic analysis.

GeneQuiz annotation in CoGenT

As an example of the capability of CoGenT, we have mapped the data obtained by the automatic genome annotation system GeneQuiz (Andrade, Brown *et al*. 1999) for 60 complete genomes to the CoGenT working environment. By making these corresponding tables available, we will enable CoGenT users to benefit from an expanded query capability for this vast amount of data on a genome-wide context.

Future uses of CoGenT

We anticipate that the computational biology community will benefit from the advantages offered by this open-source environment. All data and SQL tables can be easily downloaded from the web site. In the future, we hope that new developments might be openly exchanged by other groups using the CoGenT data environment.

3.1.5 Disclaimer

This work was in full progress when I joined the Computational Genomics Group. Nevertheless, it is presented because it forms the basis for much of the research conducted in the Group in general and by myself in particular. Since its publication in (Janssen, Enright *et al*. 2003), I have been involved in this project and its major developments (in particular CoGenT++ discussed in the next **Chapter**).

3.2 CoGenT++: An extensive and extensible data environment for computational genomics

3.2.1 Abstract

Motivation: CoGenT++ is a data environment for computational research in comparative and functional genomics, designed to address issues of consistency, reproducibility, scalability and accessibility. *Description:* CoGenT++ facilitates the re-distribution of all fully sequenced and published genomes, storing information about species, gene names and protein sequences. We describe our scalable implementation of ProXSim, a continually updated all-against-all similarity database, which stores pairwise relationships between all genome sequences. Based on these similarities, derived databases are generated for gene fusions – AllFuse, putative orthologs – OFAM, protein families – TRIBES, phylogenetic profiles – ProfUse and phylogenetic trees. Extensions based on the CoGenT++ environment include disease gene prediction, pattern discovery, automated domain detection, genome annotation and ancestral reconstruction. *Conclusion:* CoGenT++ provides a comprehensive environment for computational genomics accessible primarily for large-scale analyses as well as manual browsing. *Access*: The database and component downloads are accessible from http://cgg.ebi.ac.uk/cogentpp.html

3.2.2 Introduction

Since the publication of the first entire genome sequence for *Haemophilus influenzae* Rd in 1995 (Fleischmann, Adams *et al*. 1995), more than 220 genomes comprising over 800,000 genes/proteins have been sequenced and published (**Figure 7**) (Janssen, Enright *et al*. 2003). This wealth of sequence information provides immense opportunities for genome-wide mining and exploration, including aspects of comparative and evolutionary studies, functional genomics, protein interactions and protein family discovery. Yet, most of the existing databases are designed to provide access to genomic information on a gene-by-gene basis (Maglott, Ostell *et al*. 2005), and are not fully suitable for large-scale studies. To address this problem in our own research, we first developed the Complete Genome Tracking (CoGenT) database (Janssen, Enright *et al*. 2003), enabling both large-scale analyses and unified linking of various projects. We have now extended the original sequence database to include other facets of computational genomics with precomputed entities, such as phylogenetic profiles and protein families.

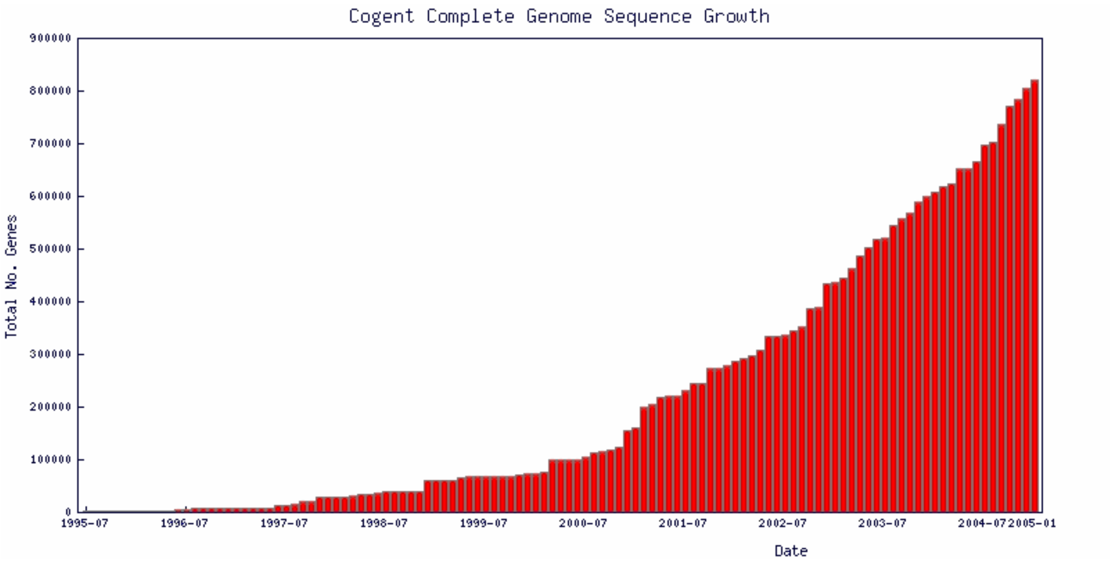


Figure 7. Exponential growth of the CoGenT genome collection as number of protein sequence entries. The x-axis corresponds to the date of publication and the y-axis to the number of protein entries in CoGenT. Despite the fact that the CoGenT project was initiated in 2000, we record the original date of publication for each genome since 1995 (Janssen, Enright *et al*. 2003).

3.2.3 The CoGenT++ environment

The CoGenT++ system is designed to provide a comprehensive, robust, flexible and useful environment guided by research issues in computational genomics. The original design principle of the CoGenT++ environment is to capture various aspects of genomic information, including a multitude of pre-computed entities, in a uniform manner. We opted for a relational database schema, currently implemented in MySQL (mysql.org), and made both the schema and contents of the databases available. **Figure 8** shows the conceptual schema of the system, replicated from the web site, where it is available as a clickable site map. The CoGenT++ environment has been designed as a three-tier system: (i) the database tier: this includes the primary input data and results from a self-comparison; (ii) the application tier: this includes all results computed for specific purposes via a range of applications and further maintained in secondary databases, either directly derived (cyan in **Figure 8**) or linked (grey in **Figure 8**) to other imported resources (blue in **Figure 8**); (iii) the client tier: this includes external access for both precomputed data and interactive computation available to users through the World Wide Web (WWW).

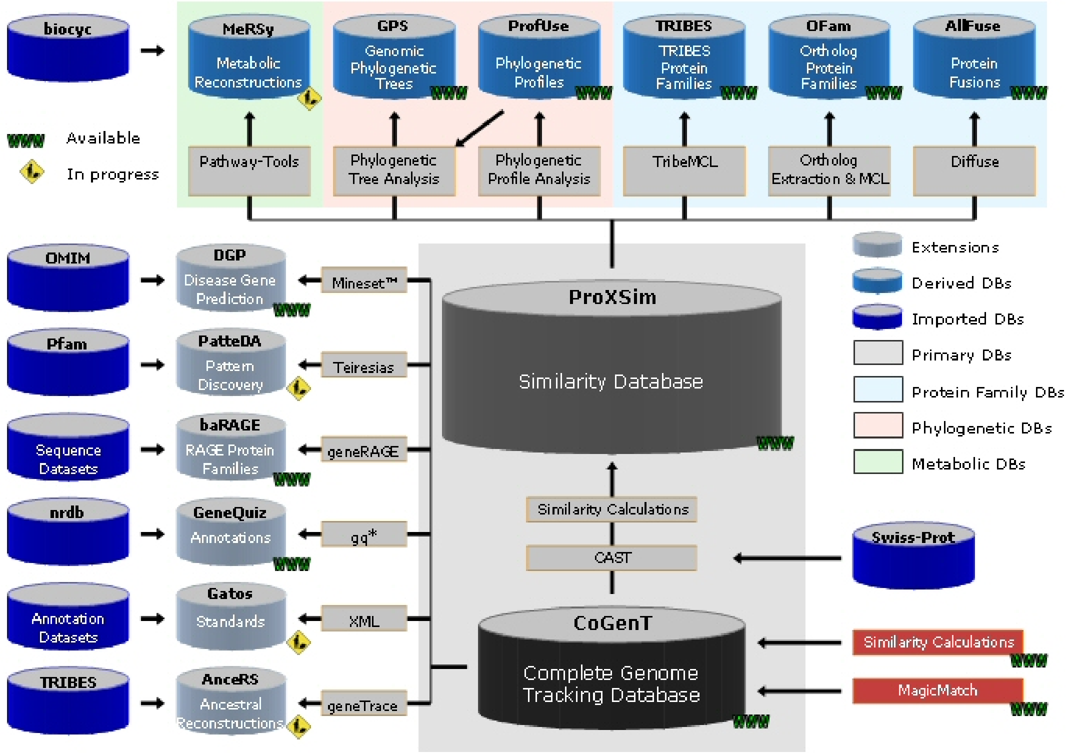


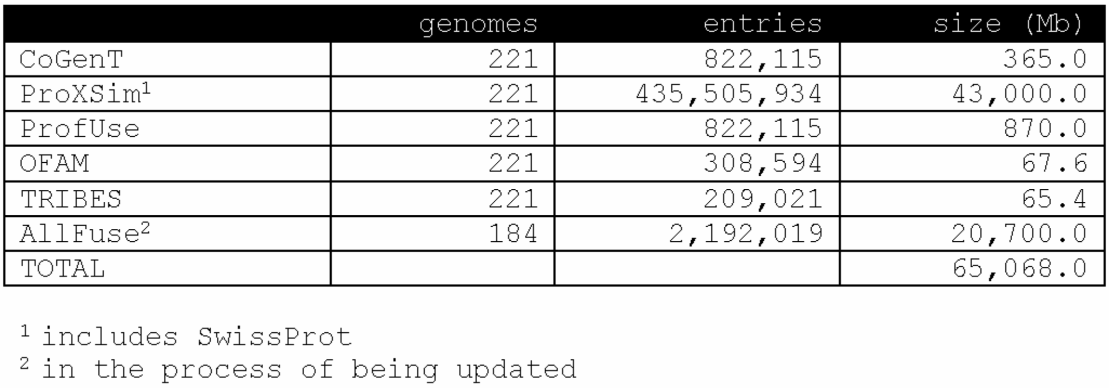
Figure 8. Conceptual design of the CoGenT++ environment. This representation is the entry point of the web site as a clickable map. The database tier is composed of the CoGenT and ProXSim databases, the application tier is composed of derived (cyan) or linked (grey) databases to other imported resources (blue). The client tier includes external access to data and interactive servers via the WWW. Users can query the system either by identifier using precomputed database cross-references (via MagicMatch) or sequence (Similarity calculations, via BLAST).

The first tier consists of two layers. The first layer corresponds to the original CoGenT database, storing information about genomes and the corresponding proteins (Janssen, Enright *et al*. 2003). The second layer corresponds to a pairwise similarity database for protein sequences, called ProXSim, which provides the basis for the production of derived databases for the next tier (**Figure 8**).

The second tier consists of directly derived databases containing protein families, groups of orthologs, phylogenetic profiles and fusion events (**Figure 8**). These databases are computed by a range of algorithms that interact with the database, for instance GeneRAGE (Enright and Ouzounis 2000), TRIBE-MCL (Enright, Van Dongen *et al*. 2002) and GeneTRACE (Kunin and Ouzounis 2003b). Other applications may require the importing of other resources for the creation of secondary databases, for example the Disease Gene Prediction (DGP) (Lopez-Bigas and Ouzounis 2004) or the Genome Phylogeny Server (GPS) (Kunin, Ahren *et al*. 2005) (**Figure 8**).

The third tier facilitates access to a substantial amount of precomputed data via WWW access as well as interactive querying and processing, for example CAST masking (Promponas, Enright *et al*. 2000) and BLAST searching (Altschul, Madden *et al*. 1997) (**Figure 8**). The amount of data made available for research purposes is substantial: the current CoGenT++ environment provides access to more than 65 Gigabytes (GB) of data (**Table 3**), e.g. almost twelve times larger than the current UniProt release (Bairoch, Apweiler *et al*. 2005). The details of database components are described below.

Table 3. Contents of the principal databases in the CoGenT++ environment. Entries: protein sequences in CoGenT, pairwise similarities in ProXSim, phylogenetic profiles in ProfUse, putative ortholog clusters in OFAM, protein families in TRIBES, fusion events in AllFuse.



	genomes	entries	size (Mb)
CoGenT	221	822,115	365.0
ProXSim[1]	221	435,505,934	43,000.0
ProfUse	221	822,115	870.0
OFAM	221	308,594	67.6
TRIBES	221	209,021	65.4
AllFuse[2]	184	2,192,019	20,700.0
TOTAL			65,068.0

[1] includes SwissProt
[2] in the process of being updated

3.2.4 System backbone

The core of the system is the CoGenT database (Janssen, Enright *et al*. 2003), which provides information about genomes and proteins from all published and fully sequenced genomes, the two criteria that define admittance of a genome into the database. We admit only genomes for which sufficient coverage has been achieved so that protein sequence information has been made available either at the original sequencing centre site or a major molecular biology database (Janssen, Enright *et al*. 2003). The recent rate of genome inclusion has reached one genome per week, on average (Janssen, Goldovsky *et al*. 2005).

The CoGenT database consists of two tables (database tier): the *genomes* table and the *proteins* table (also see **Section 3.1.3**). The *genomes* table stores information regarding each fully sequenced genome: full name of the species and strain, its taxonomic classification, genome size, number of genes, date of publication, additional curator information and a simple versioning mechanism (Janssen, Enright *et al*. 2003). The *proteins* table stores information about proteins from these genomes including the full sequence, original annotation, the originally submitted protein identifier and further information (Janssen, Enright *et al*. 2003). Additional CoGenT identifiers are generated for both genomes and protein sequences in a consistent fashion (Janssen, Enright *et al.* 2003), to facilitate easy recognition of sequences by users and programs alike.

One of the most important recent developments is the linking of the CoGenT identifier space with 'official' identifier conventions. Thus, the CoGenT++ environment is now fully cross-linked with 5 million links to major molecular biology databases, namely UniProt (Bairoch, Apweiler *et al.* 2005), EMBL (Kanz, Aldebert *et al.* 2005), GenBank (Benson, Karsch-Mizrachi *et al.* 2005), RefSeq (Pruitt, Tatusova *et al.* 2005) and PDB (Deshpande, Addess *et al.* 2005). This is achieved through the MagicMatch algorithm (Smith, Kunin *et al.* 2005), accessible at http://cgg.ebi.ac.uk/services/magicmatch/

The CoGenT++ schema allows the user to include additional protein databases and link them to the system. In the current set-up, we have included the Swiss-Prot database (Boeckmann, Bairoch *et al.* 2003), in order to link to a high-quality annotation resource.

3.2.5 ProXSim: a similarity database

The similarity database contains all-against-all pairwise similarities of proteins computed using BLASTP (Altschul, Madden *et al.* 1997), filtered for compositionally biased regions using CAST (Promponas, Enright *et al.* 2000). This database is built from all protein sequence data in CoGenT, plus other imported sequence collections, currently Swiss-Prot, release 42.11 (Boeckmann, Bairoch *et al.* 2003). The principal use of this component (database tier) is the storage of pre-computed similarities and their subsequent use by different applications (application tier), depending on the question at hand, for example phylogenetic profiles or protein families (see below). In this way, a single large-scale computation provides the basis upon which other systems or resources are automatically built: for example, the entry of each new genome sequence triggers its comparison against CoGenT sequences (see below). Thus, we reduce the need for recurrent similarity searches for specific genome subsets within a finite protein universe, which are performed daily consuming significant computational resources. It is hoped that other colleagues will find the pre-computed set of similarities useful in their own research. Due to the large size of the database (**Table 3**), users can interactively access the similarity information only by providing a protein identifier and the 'Phylogen' server (**Figure 8**) returns the phylogenetic profile of the query protein. Interactive sequence searches against genomes are also supported at this level via BLASTP (Altschul, Madden *et al.* 1997). Both Phylogen and BLASTP support *ad hoc* analyses.

3.2.6 Incremental update mechanism

All similarities are computed by the BLASTP program (Altschul, Madden *et al.* 1997). However, the main estimation of similarity significance by BLASTP is the *E-value*, which is dependent on the database size. With the natural growth of the database, the old E-value estimation needs to be recomputed. We aimed to reduce the computational load on the system and avoid recalculating the complete database every time a new genome is being included. We use the BLASTP bit-score as an estimate of sequence similarity, that is independent of the database size which is set to a constant value (see below). The E-value might be computed on the basis of bit-scores and query-dependent database sizes (Altschul, Madden *et al.* 1997).

Therefore, we use the BLASTP bit-score b as an estimate of sequence similarity because it only depends on the alignment and the substitution matrix. Then, the E-value is calculated in a simplified yet uniform manner as follows: $E_{simpl}=L_{eff}*S_{eff}*2^{(-b)}$ where the effective database size S_{eff} is set to 10^8 residues and L_{eff} is the effective protein length (number of amino acid residues not masked by CAST). In order to perform a consistent and incremental update of the similarity database each time a new genome is released and processed, while keeping the previously computed values in the database, we use an E-value cut-off of 10^{-5} on E_{simpl}. Note that for very small peptides this cut-off is too stringent to accept even full length alignments. Hence, we calculate the cut-off that would accept alignments covering 40% of the query protein length and we actually use the most permissive cut-off between this alternative cut-off and 10^{-5}.

This setup allows us to perform a consistent and incremental update of the database each time a new genome is released, while keeping the previously computed values in the database. By designing the system (database tier) so that it makes incremental updates of genome similarities, CoGenT++ provides a scalable and automatic update mechanism for the pairwise comparison of all genomes, at least for the foreseeable future. With the growth of genome sequence information (**Figure 7**), the size of the similarity database increases quadratically. This will eventually lead to a data size explosion that could challenge methods of storage and distribution for end users, and other solutions must be sought, e.g. distributed databases across the GRID (Stevens, Robinson *et al*. 2003; Teo, Wang *et al*. 2004).

3.2.7 TRIBEs and OFAM: Protein families and putative orthologs

Protein family classification is a key step in many computational genomics projects. CoGenT++ provides protein family information in two forms: the TRIBES protein family database and the OFAM ortholog family database. The TRIBES database (Enright, Kunin *et al*. 2003) is derived from the complete set of pairwise similarities using the TRIBE-MCL algorithm (Enright, Van Dongen *et al*. 2002), a method suited to the rapid and accurate detection of protein families on a large scale. OFAM is a database of protein ortholog families derived using best bidirectional hits – instead of pairwise similarities – and clustered as in TRIBES. Reciprocal best hits have been proposed as an operational definition of orthologs (Overbeek, Fonstein *et al*. 1999). The OFAM database provides higher granularity, i.e. specific clusters, while TRIBES provides wider protein families.

3.2.8 AllFuse: protein fusions

The detection of gene fusion events across genomes can be used for predicting functional associations of proteins (Marcotte, Pellegrini *et al*. 1999), including physical interaction or complex formation (Enright, Iliopoulos *et al*. 1999). CoGenT++ incorporates data on gene fusion events computed with an updated automatic protocol called Diffuse-2 (Enright, Iliopoulos *et al*. 1999) and is expected to supersede the previous AllFuse collection (**Table 3**). Each genome is used as a query against the remaining complete genomes to detect gene fusions. Pairs of proteins in the query genome that are not similar to each other and are found to be similar to a single, composite protein from the reference genomes, across non-overlapping segments, are predicted to be functionally associated (Enright, Iliopoulos *et al*. 1999).

3.2.9 ProfUse: phylogenetic profiles

A phylogenetic profile is defined as a string that encodes the presence or absence of a protein in every known genome (Pellegrini, Marcotte *et al.* 1999). Proteins with similar phylogenetic profiles have been shown to be involved in related cellular processes. Phylogenetic profiles in CoGenT++ are generated from the ProXSim similarity database data and are represented as binary vectors, where each bit represents an individual protein's hit to a genome. Given a query protein, the 'Phylogen' server returns its pre-computed phylogenetic profile (see above). The reverse operation is achieved through the 'ProfUse' server: given an arbitrary phylogenetic profile from a list of species, it returns all proteins consistent with the specified profile. Note that this set of proteins might contain homologous sequences exhibiting identical profiles due to evolutionary relationships or non-homologous sequences exhibiting identical profiles due to functional relationships.

3.2.10 CoGenT++ extensions

CoGenT++ can easily be linked to other systems (**Figure 8**). Currently it is loosely coupled to GeneQuiz – an expert system for automated genome annotation (Andrade, Brown *et al.* 1999), GeneRAGE – a sequence clustering algorithm (Enright and Ouzounis 2000), Disease Gene Prediction (DGP) – a database of human genes with their probability of being involved in a hereditary disease (Lopez-Bigas and Ouzounis 2004) and the Genome Phylogeny Server (GPS) – an approach for the computation of phylogenetic trees using estimation of species distances from genome-wide sequence similarity (Kunin, Ahren *et al*. 2005). Other systems are in the process of being linked to CoGenT++, in the near future. Current work towards this direction includes the reconstruction of ancestral states using the GeneTRACE algorithm (Kunin and Ouzounis 2003b), pattern discovery methods applied to protein families using TEIRESIAS (Rigoutsos and Floratos 1998) and automated metabolic reconstructions with the Pathway Tools software suite (Karp, Paley *et al*. 2002) ('in progress', **Figure 8**). We hope that other groups who find this resource useful in their research projects will both link and provide pertinent open-access data following the CoGenT++ architecture.

The potential adoption of the CoGenT schema by other groups and the provision of additional datasets that could easily be linked to the original CoGenT tables, for instance gene expression or protein interaction information, is expected to contribute towards highly consistent results that are readily accessible and reproducible.

3.2.11 Comparison to other systems

Similar systems that deliver genome information have been developed before. Notable cases are the NCBI genomes[2] and EBI genomes[3] databases, serving a varied community of end-users, ranging from novice to sophisticated. No other provided information is integrated in the form of protein families or interactions, or any tightly coupled methods that generate secondary information. The COG database maintains a list of orthologs for various species but requires manual intervention and no database dumps are readily available (Tatusov, Fedorova *et al*. 2003). The STRING database contains information about both genomes and gene context (such as gene fusions) and derives orthology directly from the COG database (von Mering, Jensen *et al*. 2005). Analogous databases providing information for gene context are ProLinks (Bowers, Pellegrini *et al*. 2004) and Predictome (Mellor, Yanai *et al*. 2002). Both of these databases provide flat-file downloads and links to other databases, but no database dumps or link mechanisms. Finally, the Comprehensive Microbial Resource provides detailed information about gene and protein function and various computed characteristics of sequences and genomes (Peterson, Umayam *et al*. 2001), yet it does not provide tightly coupled systems for analysis and inference.

[2] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome
[3] http://www.ebi.ac.uk/integr8/

Unique features of CoGenT/CoGenT++ include its transparency and the tracking of genome information in precise chronological order (Janssen, Enright *et al.* 2003). In this manner, we are able to trace the time of discovery of novel protein families and the sampling of sequence or phylogenetic space (Kunin, Cases *et al.* 2003). Moreover, the two strict criteria of both publication and availability of genomes allow us to incorporate genome data in an objective manner. Other unique features include the naming scheme that facilitate both computer and human interaction, the full availability of the entire resource in both flat-file and MySQL dumps, and finally the simplicity of design and implementation. All these assets should make the CoGenT++ environment useful for a wider community and facilitate research in computational genomics.

3.2.12 Data access and platform requirements

CoGenT++ is available via an interactive website, MySQL or flat files. The URL http://cgg.ebi.ac.uk/cogentpp.html is the main entry point to the CoGenT++ environment.

The MySQL database and the Apache HTTP server run on a Sun Microsystems Enterprise E450 server with 4 CPUs and 4 Gigabytes of shared memory with access to a 140 Gigabyte disk partition. All data are downloadable as MySQL dumps and flat files, where applicable. MySQL is available for multiple platforms.

Similarity searches are performed on a 200-CPU cluster kindly provided by IBM to the Research Programme of the European Bioinformatics Institute.

3.2.13 Disclaimer

My contribution to this project and article has been very significant. I have been involved in the design, implementation, and availability of CoGenT++ and its databases (especially OFAM), and the co-authoring of the article (Goldovsky, Janssen *et al.* 2005).

3.3 Expansion of the BioCyc Collection of Pathway/Genome Databases to 160 Genomes

3.3.1 Abstract

The BioCyc Database Collection is a set of 160 Pathway/Genome Databases (PGDBs) for all eukaryotic and prokaryotic species whose genomes have been completely sequenced to date. Each PGDB in the BioCyc collection describes the genome and predicted metabolic network of a single organism, inferred from the MetaCyc database, which is a reference source on metabolic pathways from multiple organisms. In addition, each bacterial PGDB includes predicted operons for the corresponding species. The BioCyc collection provides a unique resource for computational systems biology, namely global and comparative analyses of genomes and metabolic networks, and a supplement to the BioCyc resource of curated PGDBs. The Omics viewer available through the BioCyc Web site allows scientists to visualize combinations of gene expression, proteomics, and metabolomics data on the metabolic maps of these organisms. We seek scientists to adopt and curate individual PGDBs within the BioCyc collection. Only by harnessing the expertise of many scientists can we hope to produce biological databases that accurately reflect the depth and breadth of knowledge that the biomedical research community is producing.

3.3.2 Introduction

How should biological knowledge repositories be created and updated in the post-genome era to maximize the accuracy of our rapidly evolving knowledge about the genome and biochemical network of each organism? Without a clear roadmap for this vast information space, chaos will reign among the hundreds of competing and overlapping databases that are arising. We propose a strategy based on the following tenets.

It is critical that for every organism whose genome has been sequenced and which has a significant experimental community, that an organism-specific database be created, and curated on an ongoing basis, to provide an up-to-date, authoritative, central resource on the evolving knowledge about the genome and the biochemical network of that organism. Often called model organism databases, or organism-specific databases, examples include EcoCyc for *Escherichia coli* (Keseler, Collado-Vides *et al*. 2005), PseudoCyc for *Pseudomonas aeruginosa* (Romero and Karp 2003), PlasmoCyc for *Plasmodium falciparum* (Yeh, Hanekamp *et al*. 2004), SGD for *Saccharomyces cerevisiae* (Christie, Weng *et al*. 2004), TAIR for *Arabidopsis thaliana* (Rhee, Beavis *et al*. 2003), MGD for *Mus musculus* (mouse) (Eppig, Bult *et al*. 2005) and FlyBase for *Drosophila melanogaster* (Drysdale, Crosby *et al*. 2005).

No one group can curate all the world's genomes. Therefore, it is imperative to involve many scientists in the updating of each database, ideally centered around scientific communities sharing an interest in a specific organism, or a related group of organisms (Ouzounis and Karp 2002). Pathway/Genome Databases provide a mechanism to integrate genome information into higher-order biochemical or gene networks, such as metabolic pathways and transcription units (Karp 1998; Karp, Krummenacker *et al*. 1999). They also express a scientific theory within a formal ontology and make it available for computational analysis (Karp 2001).

The BioCyc collection described herein is partly a mechanism for bootstrapping this process. We originally demonstrated the value of this approach by constructing a PGDB for *Haemophilus influenzae* (Karp, Ouzounis *et al*. 1996), the first species whose genome was deciphered, ten years ago (Fleischmann, Adams *et al*. 1995). By creating draft versions of many organism-specific PGDBs that can then be adopted by outside groups for curation, we are lowering the barrier to initiating such curation efforts. Because most of the BioCyc PGDBs are openly available, outside groups can use them, update them, and redistribute them, without intellectual property restrictions. We request (although we do not require) that adopters make the modified versions of their PGDBs openly available as well. Adopters are free to publish PGDBs on their own Web sites using the same Pathway Tools software that powers the BioCyc Web site (Karp, Paley *et al*. 2002). We further propose that when adopters insert new experimentally determined metabolic pathways in an adopted PGDB, that they also kindly submit those pathways for inclusion in the MetaCyc DB (Krieger, Zhang *et al*. 2004). The set of reference metabolic pathways in MetaCyc will thus continue to expand by including community contributions, enabling more accurate pathway prediction in the future.

As well as bootstrapping curation efforts by other groups, the BioCyc database collection is a useful resource in its own right. It provides a collection of metabolic pathway reconstructions and (where applicable) operon predictions for many organisms, available as a reference source through the BioCyc Web site. Each database can serve as a resource for the analysis of gene expression, proteomics, and metabolomics data (alone or in combination) using the Pathway Tools Omics Viewer, a Web-based tool for painting overlaying datasets onto the Cellular Overview diagram – a wiring diagram for the metabolic network of the cell [4].

The BioCyc collection is also an important tool for both scientific research and technology development on a genome-wide scale (Tsoka and Ouzounis 2003). For instance, EcoCyc has been used to profile properties of metabolic enzymes and pathways (Ouzounis and Karp 2000) and their relationships to protein families (Tsoka and Ouzounis 2001). EcoCyc has also been used to explore the conservation of a well-defined set of metabolic enzymes across all domains of life (Peregrin-Alvarez, Tsoka *et al.* 2003). Examples of method development include the recognition of enzymes from sequence (des Jardins, Karp *et al.* 1997) or their genome and pathway context (Green and Karp 2004). Other uses include the benchmarking of automatic annotation projects (Karp, Paley *et al.* 2004) and metabolic reconstructions (Paley and Karp 2002). For instance, we are in the process of comparing the imported annotations for BioCyc against automatically-derived annotations using a consistent sequence comparison strategy across all species, to investigate whether imported annotations reflect the most up-to-date information from public databases of functional annotation (Goldovsky *et al.*, unpublished).

The 160 genome-specific databases within the BioCyc collection are organized into three different tiers according to the quality of their database content, which is a function of the degree of manual curation they have undergone:

[4] See URL http://BioCyc.org/expr-examples/animation.html for an example.

- Tier 1 BioCyc databases are of the highest quality. They have undergone multiple person-years of curation and are updated on an ongoing basis, with regular releases. The Tier 1 PGDBs are EcoCyc, MetaCyc, and the BioCyc Open Chemical Database (BOCD) [5]. Note that MetaCyc and BOCD are not organism-specific DBs, thus yielding the 162 total DBs in the BioCyc collection.
- Tier 2 BioCyc databases were computationally generated by the PathoLogic program, and have undergone less than one person-year of manual curation to review and polish their contents. BioCyc contains 17 PGDBs in Tier 2, including HumanCyc (Romero, Wagg *et al.* 2005), AgroCyc, and FrantCyc (Larsson, Oyston *et al.* 2005).
- Tier 3 BioCyc databases are those emphasized in this article. They were computationally generated by the PathoLogic program, and have undergone no manual curation nor review. For example, we have not manually reviewed the PathoLogic pathway predictions, nor have we refined the contents of Tier 3 PGDBs, such as by manually adding additional experimentally known metabolic pathways for that organism. BioCyc contains 142 PGDBs in Tier 3.

The remainder of this article describes the methods used to create the Tier 3 BioCyc databases.

3.3.3 Methods

Data Sources

Creation of the Tier 3 BioCyc PGDBs began with annotated genomes for each organism. Our approach does not re-annotate each genome, but instead builds new layers of knowledge on the basis of existing genome annotation. Annotations for most Tier 3 prokaryotic genomes were obtained from the Comprehensive Microbial Resource (CMR) (Peterson, Umayam *et al.* 2001), version as of Nov-13-2004. For those genomes not included in CMR, annotations were imported from the UniProt

[5] See URL http://biocyc.org/open-compounds.shtml

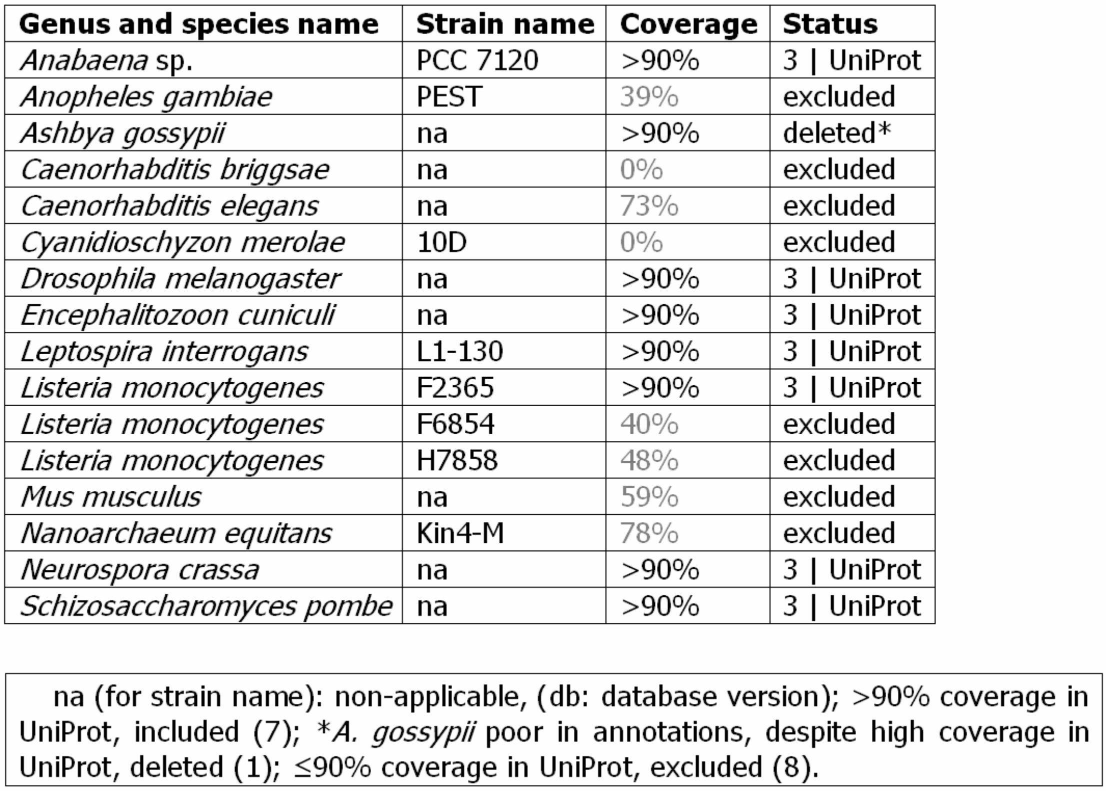
database (which includes the curated SwissProt database and the automatically generated TrEMBL supplement) (Bairoch, Apweiler *et al*. 2005). Most of the latter annotations were indeed imported from TrEMBL and directly reflect the original function assignments deposited in the corresponding GenBank files by the original genome sequencing projects (Benson, Karsch-Mizrachi *et al*. 2005).

Inclusion Criteria and Input Format

Annotations for species not available in CMR were obtained from the UniProt database (Bairoch, Apweiler *et al*. 2005). Only those species with high coverage in UniProt have been considered (more than 90% of total number of proteins). To determine this level of coverage, all protein sequences from the CoGenT database were cross-referenced to the corresponding UniProt entries using the MagicMatch algorithm (Smith, Kunin *et al.* 2005). MagicMatch is an efficient algorithm that matches identical sequences across databases. The corresponding UniProt files were checked for species names to ensure that the same species is considered.

Despite the fact that metabolic map reconstruction is robust and can be achieved even with partial genome information (Ahren and Ouzounis 2004), we decided not to include any organisms not available in CMR and with ≤90% coverage of their genome by protein-coding genes in UniProt (**Table 4**), as mentioned above.

Table 4. A list of the 16 species not available in CMR. Columns: genus/species name, strain name, coverage in UniProt and status in BioCyc (tier level 3 or exclusion).



Genus and species name	Strain name	Coverage	Status
Anabaena sp.	PCC 7120	>90%	3 | UniProt
Anopheles gambiae	PEST	39%	excluded
Ashbya gossypii	na	>90%	deleted*
Caenorhabditis briggsae	na	0%	excluded
Caenorhabditis elegans	na	73%	excluded
Cyanidioschyzon merolae	10D	0%	excluded
Drosophila melanogaster	na	>90%	3 | UniProt
Encephalitozoon cuniculi	na	>90%	3 | UniProt
Leptospira interrogans	L1-130	>90%	3 | UniProt
Listeria monocytogenes	F2365	>90%	3 | UniProt
Listeria monocytogenes	F6854	40%	excluded
Listeria monocytogenes	H7858	48%	excluded
Mus musculus	na	59%	excluded
Nanoarchaeum equitans	Kin4-M	78%	excluded
Neurospora crassa	na	>90%	3 | UniProt
Schizosaccharomyces pombe	na	>90%	3 | UniProt

na (for strain name): non-applicable, (db: database version); >90% coverage in UniProt, included (7); **A. gossypii* poor in annotations, despite high coverage in UniProt, deleted (1); ≤90% coverage in UniProt, excluded (8).

Of the 16 species not in CMR but available in the CoGenT tracking database (Janssen, Enright *et al*. 2003), the breakdown is as follows. Only seven species have a >90% coverage in UniProt and are not in any other Tier, thus included in Tier 3 (see **Table 4**, for details). One species (*Ashbya gossypii*) contains very few annotations in the form of EC numbers and was excluded from the current release, despite >90% coverage in UniProt. Finally, the remaining eight species whose coverage is ≤90% and are not supported by other resources were not included (see **Table 4**). It is conceivable that, depending on community requests, PGDBs for some of these species can be readily created, if our criteria are relaxed, due to the significance of these organisms, such as the mosquito *Anopheles gambiae*, the two *Caenorhabditis* species, or mouse. Thus, of all genomes available at the time of this project (September 2004), only nine genomes were not ultimately incorporated into the BioCyc collection of PGDBs (**Table 4**).

The Pathway Tools engine for generating new PGDBs, called PathoLogic, accepts as its input a set of files describing the annotated genome of an organism. It accepts a file format called PathoLogic format, in which each file describes all genes within one replicon. For each gene the file specifies its name, base-pair position, product type (e.g., protein, rRNA), functional description of the gene product, and EC number.

CMR Data Transformation

We transformed each CMR genome to PathoLogic format using a two-step process. In step 1 we loaded the entire CMR into BioWarehouse, an Oracle-based bioinformatics DB warehouse system [6]. CMR provides two alternative annotations of the genome: one produced by the sequencing centre that sequenced the genome, and one produced by TIGR using an automated annotation pipeline; we always loaded the former annotation into BioWarehouse. In step 2 we developed a Lisp-based program that issued SQL queries to BioWarehouse to retrieve the CMR data for each genome, and wrote that data out in PathoLogic format.

In the course of validating the preceding programs, we discovered that CMR does not contain most RNA-coding genes for many genomes – therefore those genes are lacking from many CMR-derived genomes. For scientists wishing to adopt a CMR-derived genome, SRI will generate anew a PGDB from the original GenBank entry for the genome, thus providing a complete PGDB.

In the future, because of the complexity of the CMR database and its omission of RNA-coding genes, other databases such as CoGenT (Janssen, Enright *et al.* 2003) or RefSeq (Pruitt, Tatusova *et al.* 2005) will have to be considered.

UniProt Data Transformation

UniProt annotations in the form of description line and EC number were imported into PathoLogic-format files for further processing. The following fields were extracted from UniProt: description line and EC number for function assignment, as well as gene and species name for the derivation of the corresponding entities in PGDBs.

[6] See URL http://bioinformatics.ai.sri.com/biowarehouse/

PathoLogic Execution

The Tier 3 BioCyc PGDBs were created by the PathoLogic program (Karp, Paley *et al*. 2002). PathoLogic computationally creates a new PGDB from the annotated genome of an organism. Its first step is to transform the input genome (in the form of a GenBank file or a PathoLogic-format file) into a set of objects within the Ocelot object database system (Karp, Paley *et al*. 2002). Each replicon, gene, and protein within the input file is converted to an Ocelot object that represents those replicons, genes, and proteins.

Its second step is to infer the metabolic pathway complement of the organism by reference to the MetaCyc DB, copying appropriate metabolic pathway, reaction, and small-molecule objects from MetaCyc to the new PGDB. PathoLogic matches enzymes in the annotated genome against metabolic pathways in the MetaCyc DB. The matching is based on EC number, and on the enzyme name assigned in the annotated genome. The PathoLogic operon predictor (Romero and Karp 2004) was executed to populate the microbial PGDBs with objects describing predicted operons. PathoLogic execution time on a SunBlade-1500 (1.06GHz/2GB RAM) workstation is approximately 15 minutes per microbial genome.

MetaCyc

The MetaCyc database (see the URL http://MetaCyc.org/) is a collection of metabolic pathways and enzymes from a wide variety of organisms, primarily microorganisms and plants (Krieger, Zhang *et al*. 2004). The goal of MetaCyc is to contain a representative sample of each experimentally elucidated pathway, and thereby to catalogue the universe of metabolism. MetaCyc also describes reactions, chemical compounds, and genes. As of the most recent release (version 9.0), MetaCyc contains 547 pathways elucidated in more than 340 organisms, 5,000 reactions, 2,062 enzymes, 3,900 compounds, and 5,400 literature citations. In addition, 55% of the pathways and 90% of the enzymes contain comments. More than 120 species have two or more pathways represented in MetaCyc, with *Escherichia coli* and *Arabidopsis thaliana* represented by the greatest number of pathways, 169 and 59, respectively.

Database Implementation of BioCyc

The entire collection of BioCyc DBs is stored within the same database management system, the Ocelot object database system (Karp, Paley *et al.* 2002). Ocelot employs an object-oriented data model that includes a taxonomic hierarchy of classes that define the DB schema, a set of instances of those classes that represent biological entities, and a set of slots that define attributes of (and relationships among) those entities.

All BioCyc DBs share the same DB schema, namely the Pathway Tools schema (Karp 2000). By sharing the same schema among all DBs, we ensure that the same software environment can be used to manipulate the DBs, and that comparisons can be computed consistently across all DBs. The schema consists of 1350 class definitions that define data types (such as biochemical reactions, small molecules, genes, promoters, operons, and metabolic pathways), and taxonomic classification systems (Karp 2000). For example, the Pathway Tools schema includes a classification system for pathways, for small molecules, for biochemical reactions (the Enzyme Commission system), and for genes (Serres and Riley 2000).

3.3.4 Results

Shared Aspects of BioCyc Databases

The entire collection of PGDBs encompasses 21,187 pathways distributed across 160 species, thus resulting in a mean value of 132.4 (standard deviation 52) and a median value of 137.5 pathways per species (**Figure 9**). Thus, most metabolic reconstructions generate a substantial amount of pathway information from the imported annotations. There are 3 species with 25 or less pathways, namely *Mycobacterium avium paratuberculosis* (5 pathways), *Ralstonia solanacearum* GMI1000 (14 pathways), and *Pyrococcus horikoshii* shinkaj OT3 (15 pathways) and 3 species with more than 230 pathways, namely *Bradyrhizobium japonicum* USDA 110 (231 pathways), *Streptomyces avermitilis* MA-4680 (231 pathways), and *Mesorhizobium loti* MAFF303099 (234 pathways). The organisms with very small numbers of pathways seem to be artefacts caused by very sparse genome annotations, or annotations in which the gene functions are provided in unusual formats.

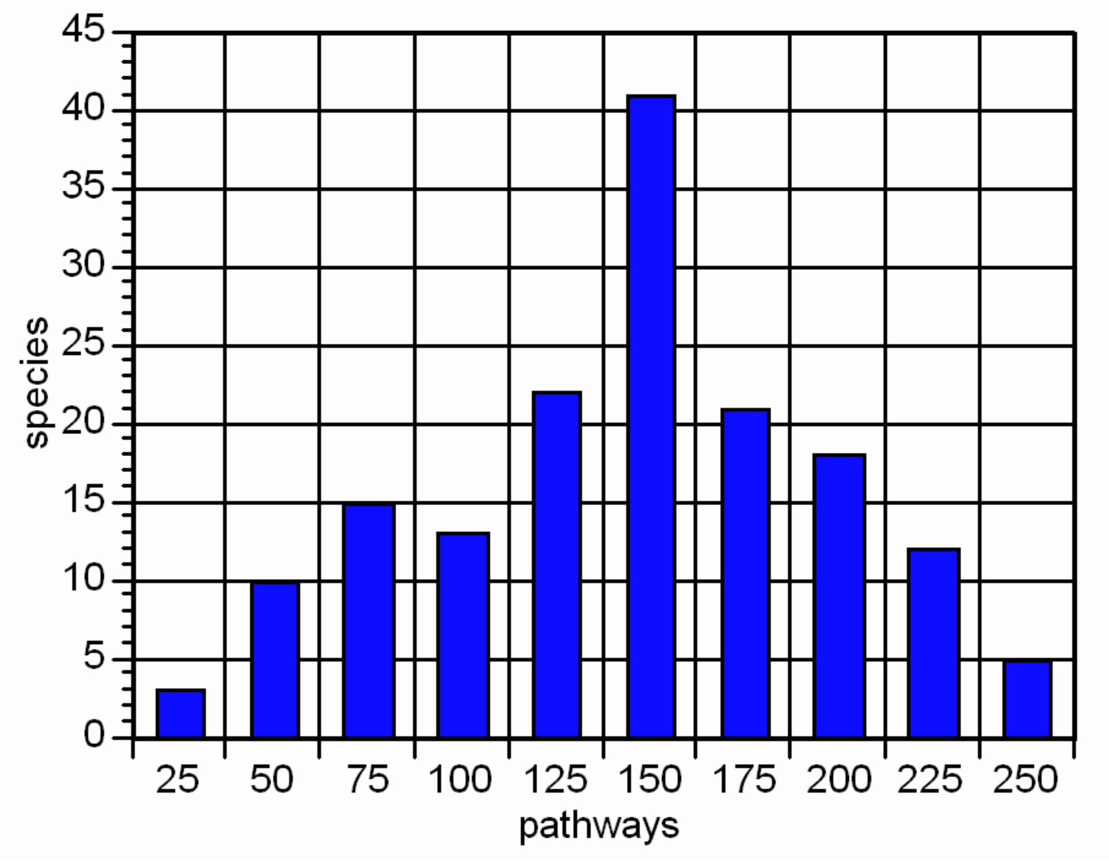
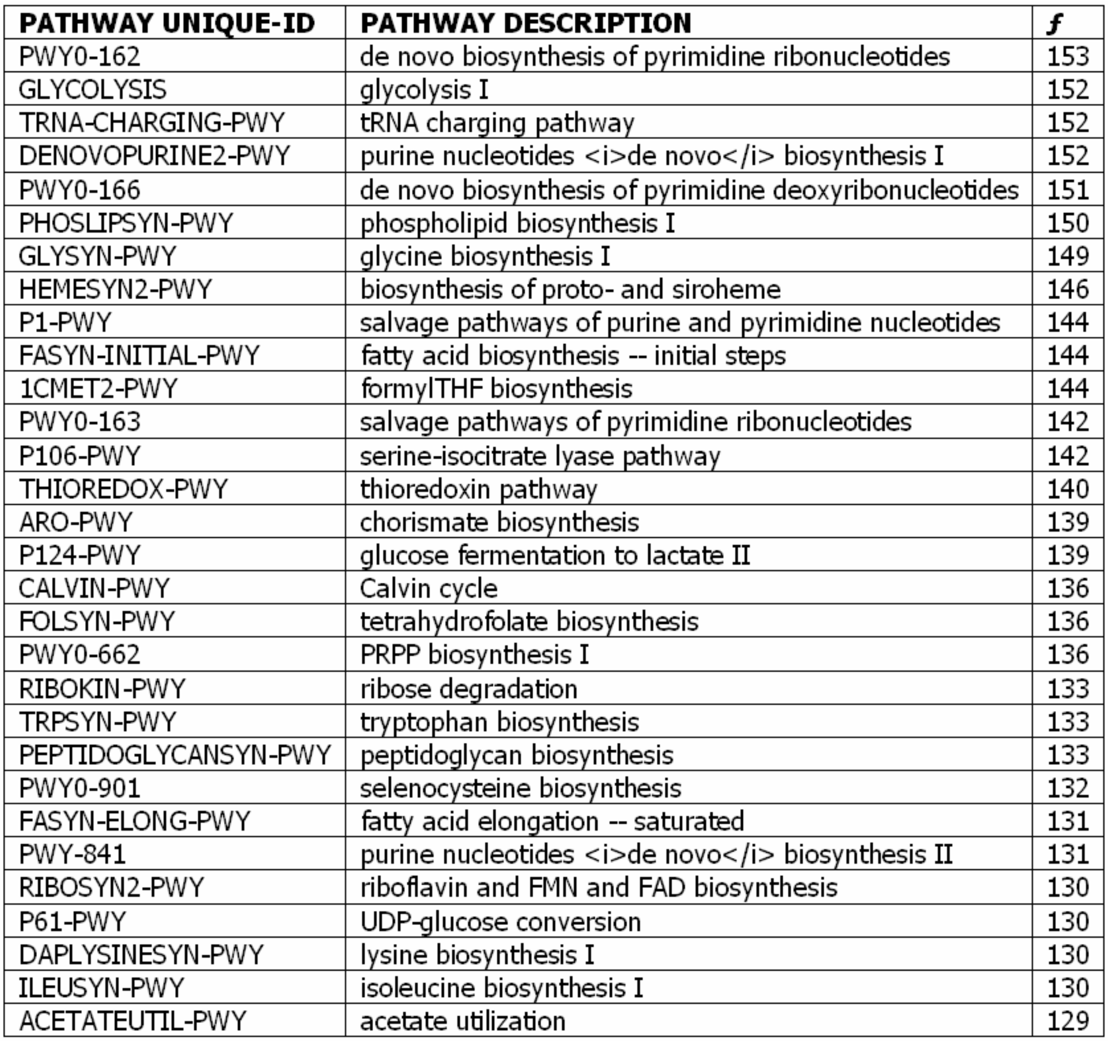


Figure 9. Frequency distribution of BioCyc pathways across species. The x-axis shows the number of detected pathways and the y-axis the number of species containing those pathways.

Interestingly, some of the most conserved pathways at the level of enzymes (Peregrin-Alvarez, Tsoka *et al*. 2003) are recovered in the BioCyc PGDB collection, such as glycolysis, nucleotide biosynthesis and amino acid (e.g. glycine, tryptophan, lysine) biosynthesis (**Table 5**). This observation might be biased by the sample of organisms under consideration, yet it uncovers some of the most conserved (or, possibly, well-annotated) segments of metabolism.

Table 5. The 30 pathways that occur most frequently across the BioCyc DBs, and their frequency (f) of occurrence.



PATHWAY UNIQUE-ID	PATHWAY DESCRIPTION	f
PWY0-162	de novo biosynthesis of pyrimidine ribonucleotides	153
GLYCOLYSIS	glycolysis I	152
TRNA-CHARGING-PWY	tRNA charging pathway	152
DENOVOPURINE2-PWY	purine nucleotides <i>de novo</i> biosynthesis I	152
PWY0-166	de novo biosynthesis of pyrimidine deoxyribonucleotides	151
PHOSLIPSYN-PWY	phospholipid biosynthesis I	150
GLYSYN-PWY	glycine biosynthesis I	149
HEMESYN2-PWY	biosynthesis of proto- and siroheme	146
P1-PWY	salvage pathways of purine and pyrimidine nucleotides	144
FASYN-INITIAL-PWY	fatty acid biosynthesis -- initial steps	144
1CMET2-PWY	formylTHF biosynthesis	144
PWY0-163	salvage pathways of pyrimidine ribonucleotides	142
P106-PWY	serine-isocitrate lyase pathway	142
THIOREDOX-PWY	thioredoxin pathway	140
ARO-PWY	chorismate biosynthesis	139
P124-PWY	glucose fermentation to lactate II	139
CALVIN-PWY	Calvin cycle	136
FOLSYN-PWY	tetrahydrofolate biosynthesis	136
PWY0-662	PRPP biosynthesis I	136
RIBOKIN-PWY	ribose degradation	133
TRPSYN-PWY	tryptophan biosynthesis	133
PEPTIDOGLYCANSYN-PWY	peptidoglycan biosynthesis	133
PWY0-901	selenocysteine biosynthesis	132
FASYN-ELONG-PWY	fatty acid elongation -- saturated	131
PWY-841	purine nucleotides <i>de novo</i> biosynthesis II	131
RIBOSYN2-PWY	riboflavin and FMN and FAD biosynthesis	130
P61-PWY	UDP-glucose conversion	130
DAPLYSINESYN-PWY	lysine biosynthesis I	130
ILEUSYN-PWY	isoleucine biosynthesis I	130
ACETATEUTIL-PWY	acetate utilization	129

It is worth noting that degradation pathways (with the exception of ribose degradation) appear to be much less conserved (data not shown).

3.3.5 Discussion

Scientists interested in adopting one or more of the BioCyc PGDBs should contact the authors. We will supply data files for the PGDBs, as well as the Pathway Tools software. Pathway Tools supports updating, querying, analysis, and web publishing of PGDBs in the following manner.

One can envisage a model where a central repository points to availability of PGDBs, similar in spirit to peer-to-peer file-sharing. In this mode, any new PGDB that is created registers with a central repository for inclusion and flagging for potential downloads from interested parties. A registry of shared PGDBs is available[7].

The Pathway/Genome Editors within Pathway Tools are graphical interactive tools for updating information within a PGDB. For example, given new findings in the literature, a scientist could add a new metabolic pathway to a PGDB, or alter an existing pathway. They can alter the annotation of a gene, and add commentary and literature citations. They can annotate features on proteins, such as enzyme active sites or phosphorylation sites. They can also update the description of the genetic network of an organism by defining new operons, promoters, and transcription-factor binding sites, and by defining regulatory interactions between transcription factors and their binding sites. These are the same tools used by the curators who update the EcoCyc and MetaCyc DBs.

Pathway Tools is also the software used to power the BioCyc.org Web site. We encourage PGDB adopters to publish their PGDBs on their Web sites using Pathway Tools and thus make them available to the scientific community.

3.3.6 Related Work

Because the WIT database (Overbeek, Larsen *et al*. 2000) has not been available on the Web for more than one year, the KEGG database (Kanehisa, Goto *et al*. 2004) is the closest related existing resource to BioCyc. Both KEGG and the BioCyc collection predict pathways by comparing the enzymes within a given genome against a known set of reference pathways. But many differences exist in how this is done.

One difference in pathway prediction methodology is that the two resources use different reference pathway databases as the basis for prediction – MetaCyc and the KEGG reference maps, respectively. MetaCyc contains extensive comments that describe individual pathways and enzymes; KEGG has fewer such comments. MetaCyc cites the primary literature sources from which pathway and enzyme data

[7] See URL http://BioCyc.org/registry.html

were obtained. KEGG contains very few literature citations. KEGG pathways are very much larger than those in MetaCyc, and we showed that KEGG pathways are less biologically meaningful (Green and Karp, submitted). This is due to the fact that proteins within a KEGG pathway show a lower degree of functional association than do those in EcoCyc pathways (and therefore, than MetaCyc pathways, because MetaCyc contains all pathways in EcoCyc, and because MetaCyc pathways are on average about the same size as those in EcoCyc).

The KEGG pathway prediction algorithm is also different. It begins by computing new functions for all gene products within a genome, thus replacing a genome annotation that is often derived from significant manual work by scientists with one that is purely computationally derived, and thus potentially of lower quality. Our work builds from the originally submitted genome annotation. In addition, the KEGG algorithm does not actually predict pathways – it simply colours a set of static KEGG map diagrams to indicate which enzymes within a pathway are present within a given genome. This approach avoids actually predicting whether a pathway is present or absent, whereas our PathoLogic algorithm does call each pathway in MetaCyc as present or absent in the organism it is analyzing.

KEGG also does not make its pathway databases available for adoption and curation by experts. Furthermore, the software environment underlying KEGG does not support the rich level of curation and annotation as does the Pathway Tools software underlying BioCyc. For example, the KEGG software does not allow the editing of pathway information, to remove erroneous reactions that are not present in a given organism, or to add organism-specific reactions to a pathway.

3.3.7 BioCyc Availability

All BioCyc PGDBs are accessible online through the Web at the URL http://BioCyc.org/ for interactive querying. The Web site is freely available to all users.

The Pathway Tools software is freely available to academics. See the URL http://biocyc.org/download.shtml for more information on downloading the software and databases.

We provide several mechanisms by which computational or experimental biologists can compute with the pathway data within the BioCyc PGDBs. BioCyc PGDBs in Tiers 1 and 2 are queryable via Application Program Interfaces (APIs), and are available by data file download. We plan to make Tier 3 PGDBs available in those manners in summer 2005. Nonetheless, we will make any Tier 3 PGDBs available for download for adoption based on email requests to the authors.

Pathway Tools APIs allow users to query BioCyc DBs in the Java, Perl, and Lisp languages [8]. All three APIs provide extremely comprehensive and easy-to-use query facilities. The APIs access the data through a binary executable program that bundles together the Pathway Tools software with all the BioCyc DBs and runs under Linux, Windows, and Sun workstations. This executable runs as a desktop application, and also supports local installation of the BioCyc Web site on an intranet. Flatfile versions of BioCyc PGDBs can be downloaded in four formats including SBML (see sbml.org) and BioPAX (see biopax.org).

Most BioCyc PGDBs are freely available and may be redistributed freely. A fee applies to commercial installations of some BioCyc PGDBs, and to the Pathway Tools software.

3.3.8 Disclaimer

My contribution to this project has been the participation in design and method development with members of the Computational Genomics Group and our collaborators of the Bioinformatics Research Group at SRI International. This work has now been published (Karp, Ouzounis *et al.* 2005).

[8] described at URL http://bioinformatics.ai.sri.com/ptools/ptools-resources.html

Chapter 4
Methods

The first two **Sections** (**4.1** and **4.2**) of this **Chapter** offer a glimpse into two brilliant computational methods developed within the Computational Genomics Group by two of my former colleagues, Dr. Anton Enright and Dr. Victor Kunin respectively. Both methods have been fundamental in research described in this Thesis.

Section 4.3 describes a computational method developed solely by myself, for producing consensus annotation of groups of annotated entities, primarily, in our case, clusters of proteins.

4.1 TRIBE-MCL

Despite significant progress in the field of sequence clustering, new challenges have emerged due to the availability of large eukaryotic genomes, in terms of their size and complexity. In particular, eukaryotic protein families constitute a bottleneck for most methods. Many eukaryotic proteins contain large numbers of protein domains (Apic, Gough *et al.* 2001a; Hegyi and Bork 1997), each of which needs to be detected and resolved by an efficient clustering algorithm. Iterative automatic domain detection algorithms such as GeneRAGE (Enright and Ouzounis 2000) suffer from an excessive and unpredictable number of additional sequence comparison steps, which renders them somewhat impractical when using modest computational resources. Another approach is to detect proteins with very similar domain architectures (Apic, Gough *et al.* 2001b), rather than attempt to detect each domain individually. The assumption is that proteins with near-identical sets of domains may have very similar biochemical roles (Hegyi and Gerstein 1999; Ouzounis and Karp 2000).

Given the difficulty of detecting domains accurately and the ever increasing amount of eukaryotic data available, this problem has been approached using an elegant mathematical approach based on probability and graph flow theory. An ideal method would require sequence similarity relationships as input and be able to rapidly detect clusters solely using this information, without being led astray by the complex modular domain structure of eukaryotic proteins. Traditionally, most methods deal with similarity relationships in a pairwise manner, while graph theory allows the classification of proteins into families based on a global treatment of all relationships in similarity space simultaneously. To this end, the TRIBE-MCL algorithm was developed as an efficient and reliable method for protein sequence clustering (Enright, Van Dongen *et al*. 2002). TRIBE-MCL is based on the Markov Cluster (MCL) algorithm, previously developed for graph clustering using flow simulation (Van Dongen 2000). This approach for protein sequence clustering is extremely fast and appears to be highly accurate. It avoids most of the problems mentioned above and has already been successfully utilised for the clustering of large datasets and the classification and annotation of proteins from the draft human genome (Hubbard, Barker *et al.* 2002; Lander, Linton *et al.* 2001).

The algorithm's input is a protein similarity graph. In order to build such a graph, a FASTA file containing all sequences that are to be clustered into families is assembled. Peptides within this file are filtered using the CAST (Promponas, Enright *et al*. 2000) algorithm, then compared against each other using BLAST (Altschul, Madden *et al*. 1997). All similarities and associated scores are used to build the protein-protein similarity graph. A Markov matrix is constructed, representing transition probabilities from any protein in this graph to any other connected protein. Each column of the matrix represents a given protein, and each entry in a column represents a transition probability between this protein and another protein. Diagonal elements are set arbitrarily to a *neutral* value. The entries in the Markov matrix are probabilities generated from weighted sequence similarity scores (e.g. from BLAST). A weight is assigned to each edge of a protein similarity graph by taking the average pairwise $-log_{10}$(E-Value) (Altschul, Madden *et al*. 1997), resulting in a symmetric matrix. This simple weighting scheme produces reliable results but other more complex schemes can also be used. This Markov matrix is supplied to the MCL algorithm. Initial expansion of the Markov matrix simulates random walks, which allow one to measure 'flow' in the graph. Areas of high flow indicate that a large number of random walks go through this area. The MCL algorithm iteratively promotes flow through the graph where it is strong, and removes flow where it is weak. This process terminates when equilibrium has been reached.

In a biological sense, members of a protein family are expected to be more similar to each other than to proteins in another family. Experiments using the Bio-Layout graph visualisation algorithm (Enright and Ouzounis 2001) have shown this to be true for most protein similarity graphs. Because of this property of biological graphs, flow within protein families is strong, i.e. a random walk starting at any given protein in a family is more likely to remain within this family than to cross to another family. Flow between protein families will be weaker than flow within a family as there are relatively few (if any) paths that cross two distinct protein families. Intra-family paths represent either sequence similarity relationships due to multi-domain proteins or mere false positive similarity detections. These properties of biological similarity graphs make them ideally suited to the MCL algorithm, which removes this weak flow across protein families, and promotes the stronger flow within protein families. This boot-strapping procedure allows protein families hidden in the graph to become visible by gradually stripping the graph down to its basic components as detected by stochastic flow. Many of the problems that normally hinder protein sequence clustering are eliminated by the Markov Clustering approach. Proteins possessing a promiscuous domain, that is present in many functionally unrelated proteins, are normally very difficult to cluster correctly. Promiscuous domains will connect a member of a given protein family to members of that family and possibly to a large number of other (possibly unrelated) protein families. Because these inter-family connections are still far fewer than intra-family connections, the algorithm gradually eliminates these inter-family similarities and detects protein families accurately. The algorithm requires no *a priori* knowledge of protein domains, and clusters proteins into families purely based on observed relationships through the entire similarity graph. However, proteins containing different domains or sets of domains will have very different sequence similarity patterns, and hence the MCL algorithm is expected to cluster proteins with different domain structures into distinct families.

MCL and its modules, like TRIBE-MCL, is under constant development by Dr. Stijn van Dongen, currently in the research group of Dr. Anton Enright at the Sanger Centre (see http://micans.org/mcl/).

4.1.1 Disclaimer

I was not involved in the development of TRIBE-MCL, nevertheless it was deemed necessary to include a short introduction to the system because of its significance in the research described elsewhere in this Thesis. This work has been published in (Enright, Van Dongen *et al*. 2002).

4.2 GeneTRACE

The availability of protein families in a multitude of genomes enables researchers for the first time to take a glimpse at the evolution of genomes. While methods were developed to computationally reconstruct ancestral DNA and protein sequences (Pupko, Pe'er *et al*. 2000), these studies generally focus on single sequences, rather than complete genomes. GeneTRACE (Kunin and Ouzounis 2003b) is a method aimed to reconstruct the gene content of ancestral prokaryotes.

To reconstruct the gene content of ancestral species, a framework for the inference of presence or absence of individual protein families at any node on a phylogenetic tree has been suggested (Kunin and Ouzounis 2003a). The approach is based on the following assumptions:

1. when most of the clade members contain a representative of a protein family, the observed distribution pattern would normally result from vertical gene descent. The common ancestor of the clade is thus assumed to contain the corresponding family;
2. if a protein family is present in most of the descendants of a particular ancestor, but is not found in some subclade, the observed gene absence would normally result from gene loss;
3. protein family distribution interspersed in distantly related clades would be indicative of horizontal gene transfer.

Previously, these gene distribution patterns, also called phylogenetic profiles, have been used to predict protein function (Pellegrini, Marcotte *et al*. 1999), to build gene content based phylogenetic trees (Snel, Bork *et al*. 1999) and to deduce the cell localization of gene products (Marcotte, Xenarios *et al*. 2000). To avoid the issues of ortholog definition, it has been decided to use phylogenetic profiles that contain information about the presence or absence of an entire protein family. Family information was obtained from the TRIBES database (Enright, Kunin *et al*. 2003).

GeneTRACE allows the inference of the most likely evolutionary scenario that led to the observed present-day distribution of protein families (**Figure 10**). The GeneTRACE input consists of phylogenetic profiles of protein families and an evolutionary tree including all organisms involved. Inner nodes on this tree represent ancestral organisms (**Figure 10a**). Two types of events are considered: protein family gain and loss. The algorithm consists of the following stages:

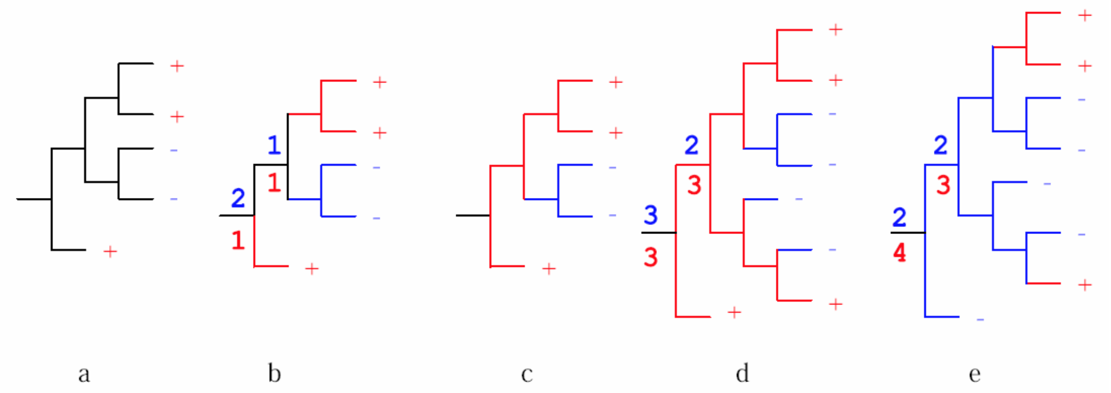


Figure 10. The flow of the GeneTRACE algorithm. Gene presence is marked by red colour, and gene absence is marked by blue. The input consists of a trusted species tree and phylogenetic profile. Plus (+) shows gene presence and minus (-) absence in an extant species (terminal nodes on the tree). a. The input data. b. Unambiguous cases are resolved, and the number of independent changes required to obtain the given data is calculated for both gene presence (red numbers) and absence (blue numbers) for each internal node. c. A putative scenario for evolutionary history of the gene is suggested, based on the Gain threshold (see text). d,e. When the difference of potential gains and losses is between the Gain and Loss thresholds, the final assignment is dependent on the subtree neighbourhood.

1. for each inner node, the minimal number of potential changes that are required to obtain the observed family distribution is calculated for both possible cases: gene family presence and absence at the node (**Figure 10b**). Both gene acquisition and loss are penalized by a single point. The calculation proceeds from terminal nodes of the tree towards the root. For each node down the tree, the penalty is equal to the sum of the penalties of its daughter nodes. These penalties are transformed into assignments of family presence or absence at the node in any of the following cases (**Figure 10c**):
    - if the descendants of the node exhibit a uniform pattern, either family presence or absence, the corresponding pattern is assigned to the node;
    - if the difference between the number of potential gains and losses is larger than a threshold value called the Gain threshold, and the family presence is observed on at least two daughter subtrees, family presence is assigned to the node;
    - if the difference between number of potential gains and losses is smaller than a threshold value called the Loss threshold, family absence is assigned to the node
2. starting from the root of the tree, unassigned nodes inherit the parental assignment (**Figure 10d,e**). The parent of the root is assumed not to contain any genes, thus delaying the first assignment to the first evidence of family presence.

The algorithm infers the presence and absence of the family on the nodes of the evolutionary tree, and generates a list of nodes where family gain and loss is predicted to occur. Horizontal gene transfer is inferred if more than one family gain is reported.

Gain and Loss thresholds are different, as they stand for family gain and loss at the node of interest. The Gain threshold is conceptually analogous to the HGT penalty described earlier (Snel, Bork *et al.* 2002). This threshold stands for the assessment of the probability for multiple gene loss events versus HGT events. Family gain is assumed if the cost of all losses is smaller than the allowed penalty for horizontal transfer. The number of suspected horizontal transfers would be the number of family gains minus one (accounting for family genesis). When family presence is observed on the parental tree node, assigning family absence to a node would imply gene loss. If some descendants of this node appear to have the protein family, the loss would be followed by regaining. In such a scenario, the introduction of the Loss threshold brings an additional requirement for higher amounts of gene loss for assigning gene absence and allows a more parsimonious version of events. The described system comprised of two walks on the tree and two thresholds allows considering subtree neighbourhoods (**Figure 10d,e**), and thus is allows potentially higher level of resolution than the previous model (Snel, Bork *et al*. 2002).

4.2.1 Disclaimer

I was not involved in the development of GeneTRACE, nevertheless it was deemed necessary to include a short introduction to the system because of its significance in the research described elsewhere in this Thesis. This work has been published in (Kunin and Ouzounis 2003b).

4.3 Cluster Annotator

As part of our efforts towards an integrated computational genomics data environment described in the previous **Chapter**, we developed a simple in concept, yet powerful, cluster annotation algorithm, i.e. a means to summarize down to a single line the annotation information of two or more entries (e.g. proteins) considered to be related by some measure. The approach is very general and can be applied to any dataset. However, since its main usage is in the manipulation of protein families, the algorithm has elements specifically suited for the management of biological information.

Previous efforts consisted of the simple implementation of the Longest Common Subsequence (LCS) algorithm (Gusfield 1997) - given two sequences of symbols, X and Y, determine the longest subsequence of symbols that appears in both X and Y. Despite the theoretical elegance of this approach, the reality was that the extreme variability of the protein annotation vocabulary and semantics created significant problems. Our current approach is based simply on the frequencies of words in the annotation information of each cluster.

By design, all operations are per cluster. The number of members in the cluster is calculated but this number is not directly used for the word frequency calculations; instead, the number of members with meaningful annotation information is calculated and used, termed henceforth the meaningful size of the cluster. By meaningful we mean annotation that delivers specific information of the cluster members sequence, structure, function, history, or else that contains other words than the ones in the following list: conserved, hypothetical, unknown, known, putative, uncharacterized, probable, possible, predicted, unnamed, gene, protein. Blank annotation (i.e. no information) is also not considered as meaningful. Thus, for example, a cluster could contain 100 members, but only 60 members would contain meaningful annotation information – the second number would be used for our calculations. Each annotation is also treated for syntactic inconsistencies like multiple spaces.

Each word is converted to lower case to avoid letter case variability, and stored in memory along with its original form. The characters “-“ “,” and “‘” are considered important linkers and are also stored in memory, connected to the words they follow in the annotation. The position of the word in the annotation is also stored. This is important because words are treated independently at this stage but in the final output they must be placed in a meaningful order. It is also a point that needs to be addressed in later versions of the algorithm since the variability in the annotation syntax means that the same word could be found in a totally different order and position, while essentially carrying the same meaning, therefore creating problems in the final assembly of the cluster’s consensus annotation. Another feature of the algorithm is the storage in memory of pairs of words. This is used at the cluster consensus annotation assembly stage to include words that are below the threshold but follow a word that is above the threshold a significant number of times. Finally, short words (3-6 characters) that are not included in a comprehensive user-defined English dictionary form the input for the pattern discovery algorithm TEIRESIAS (Rigoutsos and Floratos 1998). The expectation is that short words not found in the dictionary are likely be gene names, therefore pattern discovery would recover the gene family prefix, e.g. **atp*** from **atpB**, **atpC**, **atpE**, **atpG** etc.

During this process, the number of occurrences of each word in the cluster is stored in memory – it is important to note that only one occurrence per annotation is counted, i.e. the maximum number of word occurrences can only be the number of meaningful annotations in the cluster.

After all the members of a cluster have been processed, the frequency of each word is calculated by dividing the number of its occurrences in the meaningful annotations by the number of meaningful annotations, leading to a number between zero and one [0 – 1]. Depending on the meaningful size of the cluster, an arbitrary word frequency threshold is set:

Meaningful size of cluster		**Threshold**
from…	**to…**	
1	1	1
2	2	1
3	10	0.4
11	50	0.4
51	500	0.3
501	2500	0.2
2500	∞	0.1

The words that are over the appropriate threshold are assembled to form the cluster's consensus annotation. The original form of the words is recovered from memory, the average word order is calculated and used for the assembly, and any stored linkers are put in place. If no words are above the threshold, then the cluster is labelled as containing 'NO CONSENSUS ANNOTATION'; if the cluster contains no meaningful annotation, it is labelled as containing 'NO ANNOTATION'. Finally, a consensus annotation score is given by calculating the average percent frequency of the words in the consensus annotation.

A recent development has been to create a three-tier annotation of each cluster, by lowering the threshold in two steps and using the resulting two extra threshold ranges to represent less abundant annotation information in the cluster – this was deemed as an important step, especially in the case of multi-functional, multi-domain protein families. It was also observed that the pre-defined thresholds were leading to exclusion of words by a very small margin, therefore we aimed to include these in the lower levels as can be seen below (ORTHO- is the cluster ID from OFAM, followed by the cluster size and the cluster's meaningful size; the rest of the lines show the three tiers of consensus cluster annotation (A/B/C), with the consensus annotation scores (e.g. A@78) and the consensus annotations themselves(if any)):

```
ORTHO-000000    153     146
        |A@78| glucose inhibited division A
        |B@27| gidA
        |C@0|
ORTHO-000015    116     67
        |A@0|
        |B@18| TIGR00278 alpha-hemolysin
        |C@0|
ORTHO-000027    58      55
        |A@85| RNA polymerase sigma-32 factor
        |B@25| rpoH
        |C@9| heat shock
ORTHO-000031    48      33
        |A@58| hydrolase [HAD]
        |B@27| HAD superfamily
        |C@18| haloacid dehalogenase-like family
ORTHO-000047    143     138
        |A@87| UDP-N-acetylenolpyruvoylglucosamine reductase
        |B@20| murB
        |C@12| acetylmuramate dehydrogenase
        EC 1.1.1.158 [1]
ORTHO-000076    160     154
        |A@86| glutamyl-tRNA synthetase
        |B@16| gltX
        |C@10| glutamate-ligase
        EC 6.1.1.17 [1]
```

Note that Enzyme Classification (E.C.) numbers are also reported along with their absolute frequency in the cluster given in square brackets.

The computational efficiency of the method is remarkable: ~92,000 clusters with 500,000 entries were annotated in approximately 5min (without the use of TEIRESIAS, which adds a relatively significant overhead). This constitutes a vast improvement over the LCS approach previously used.

Overall, we believe that despite its simplicity, our method is an elegant, easily-customisable, and efficient solution for annotating clusters. It is currently an integral part of our CoGenT++ system, and has been implemented in the TRIBES and OFAM databases (see **Section 3.2**).

4.3.1 Disclaimer

I was the sole contributor to this work. Useful feedback was and is being provided by members of the Computational Genomics Group.

Chapter 5

Evolutionary studies of families

This **Chapter** describes the representation of the tree of life as a network and the reconstruction of the last universal common ancestor. The phylogeny of the microbial world with vertically and horizontally inherited gene flows results in the net of life, a new representation of phylogenetic relationships. The subsequent analysis uses ancestral reconstruction for gene content to infer a minimal estimate for the genome of the last universal common ancestor.

5.1 The net of life – Reconstructing the microbial phylogenetic network

5.1.1 Abstract

It has previously been suggested that the phylogeny of microbial species might be better described as a network containing vertical and Horizontal Gene Transfer (HGT) events. Yet, all phylogenetic reconstructions so far have presented microbial trees rather than networks. Here, we present a first attempt to reconstruct such an evolutionary network, which we term the ‘net of life’. We use available tree reconstruction methods to infer vertical inheritance, and use an ancestral state inference algorithm to map HGT events on the tree. We also describe a weighting scheme used to estimate the number of genes exchanged between pairs of organisms. We demonstrate that vertical inheritance constitutes the bulk of gene transfer on the tree of life. We term the bulk of horizontal gene flow between tree nodes as ‘vines’, and demonstrate that multiple, but mostly tiny vines interconnect the tree. Our results strongly suggest that the HGT network is a scale-free graph, a finding with important implications for genome evolution. We propose that genes might propagate extremely rapidly across microbial species through the HGT network, using certain organisms as hubs.

5.1.2 Introduction

Following the legacy of Darwin's "Origin of species" (Darwin 1859), most current methods for phylogenetic reconstruction depict evolutionary history of organisms as a tree. Phylogenetic trees have been derived from compositional signatures (Fox, Stackebrandt *et al*. 1980), sequence alignments (Doolittle 1981) or alignments of artificially concatenated conserved orthologs (Brown, Douady *et al*. 2001; Rokas, Williams *et al*. 2003). With genome sequencing technology, methods based on complete genome sequences appeared, including trees based on gene content (Fitz-Gibbon and House 1999; Korbel, Snel *et al.* 2002; Lin and Gerstein 2000; Snel, Bork *et al.* 1999; Tekaia, Lazcano *et al.* 1999), gene order (Korbel, Snel *et al*. 2002), average ortholog similarity (Clarke, Beiko *et al*. 2002) and genome conservation – a novel genome-based method combining gene content and sequence similarity (Kunin, Ahren *et al*. 2005).

All these tree-like representations of evolution have an inherent drawback, dealing solely with vertical inheritance (Bapteste, Boucher *et al*. 2004). Yet, a well-established consensus between evolutionary biologists is that the genomic history of most microbial species is mosaic, with a significant amount of Horizontal Gene Transfer (HGT) present (Boucher, Douady *et al*. 2003). Though the quantification of the evolutionary effect of the HGT is still a subject of an ongoing debate (Kunin and Ouzounis 2003a; Snel, Bork *et al.* 2002), its existence is not questioned. The strong influence of HGT led to a proposal that presentation of microbial phylogeny as a tree is inaccurate as instances of HGT are not recorded in this presentation (Doolittle 1999; Martin 1999), and a correct representation should reflect horizontal gene transfer events.

Attempts to deal with this issue include algorithmic solutions for network-like tree reconstruction, mostly addressing recombination (but not HGT) as a form of non-vertical inheritance (Gusfield, Eddhu *et al.* 2004; Wang, Zhang *et al.* 2001), and topological analyses of tree structure (Makarenkov and Legendre 2004; Piel, Sanderson *et al.* 2003). Thus, the widely accepted view that the phylogenetic history of genomes should be represented as a network rather than a tree has not been realized yet.

Here we present a first attempt to reconstruct the history of the microbial world, recording both horizontal and vertical gene transfer. As a scaffold depicting vertical gene transfer we use established tree reconstruction methods, on which we document the instances of horizontal transfer that intertwine the tree. We discuss the major properties of this complex phylogenetic network based on a multitude of genome comparisons, demonstrate its scale-free, small-world nature and discuss the patterns of gene propagation through the network.

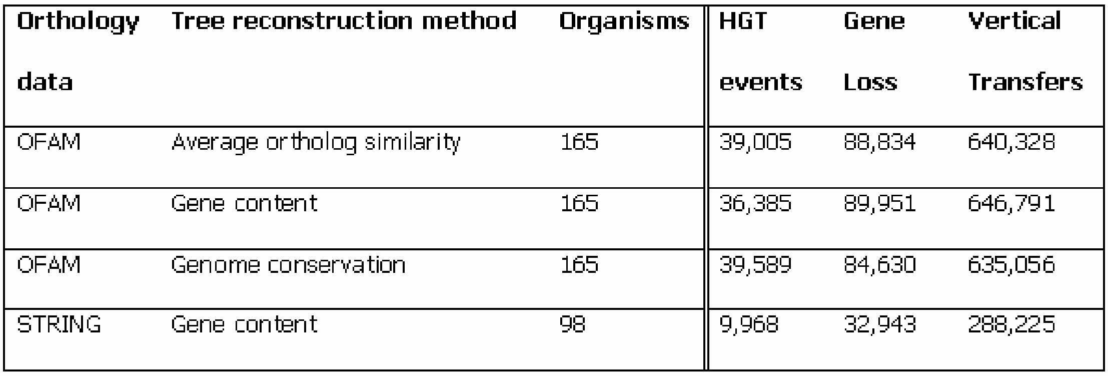
5.1.3 Results

Data

To ensure that our results are not affected solely by the orthology data (see **Methods Section 5.1.5**), we used two data sets: OFAM (see **Methods Section 5.1.5**) and groups of orthologs defined by STRING (von Mering, Huynen *et al*. 2003). Similarly, to avoid possible bias from a single tree reconstruction method, we derived genomic trees with three independent methods: gene content, average ortholog similarity and genome conservation (see **Methods Section 5.1.5**) for OFAM data and gene content for STRING data.

The summary of the evolutionary events reconstructed with each method is presented in **Table 6**. It is evident that though HGT is readily detectable, the bulk of the genes are still transferred by vertical gene transfer which is the most prevailing mode of inheritance (Kunin and Ouzounis 2003a). In analogy, the net of life is not a grid where all edges are of a similar strength, but more like a tree with robust branching stems connected by thin climbing vines.

Table 6. Summary of settings and results from various experimental designs.

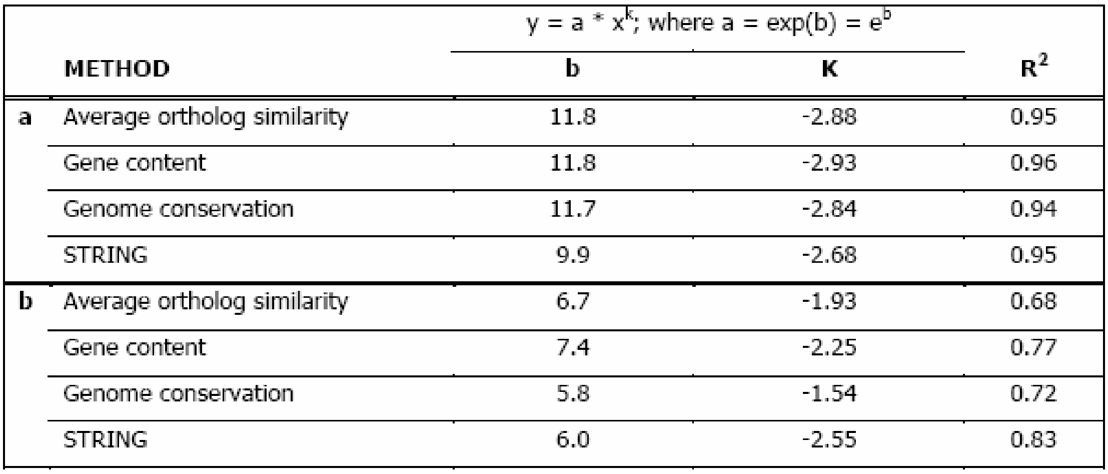


Orthology data	Tree reconstruction method	Organisms	HGT events	Gene Loss	Vertical Transfers
OFAM	Average ortholog similarity	165	39,005	88,834	640,328
OFAM	Gene content	165	36,385	89,951	646,791
OFAM	Genome conservation	165	39,589	84,630	635,056
STRING	Gene content	98	9,968	32,943	288,225

HGT vine width distribution

We define the HGT vine width as a summary of all horizontal transfer events between two nodes on the tree, subsequently fixated within the genome. The distribution of HGT vine widths, or number of genes transferred between any two nodes on the tree is shown in **Figure 11**. All data sets and trees produce virtually identical frequency distribution (**Figure 11**), following a power law (**Table 7a**), with the STRING data shifted by an order of magnitude, due to lower coverage of genomes (**Table 6**).

Table 7. Parameters for the power-law distribution (b, k) for (a) HGT vine widths (Figure 11) and (b) the connectivity of the network (Figure 12), according to the four methods used. Goodness-of-fit is expressed as the coefficient of determination (R^2) defined as R^2=1-SSE/SSM; where SSE is the sum of squared errors and SSM is the sum of squares around the mean.



		$y = a * x^k$; where $a = \exp(b) = e^b$		
	METHOD	b	K	R^2
a	Average ortholog similarity	11.8	-2.88	0.95
	Gene content	11.8	-2.93	0.96
	Genome conservation	11.7	-2.84	0.94
	STRING	9.9	-2.68	0.95
b	Average ortholog similarity	6.7	-1.93	0.68
	Gene content	7.4	-2.25	0.77
	Genome conservation	5.8	-1.54	0.72
	STRING	6.0	-2.55	0.83

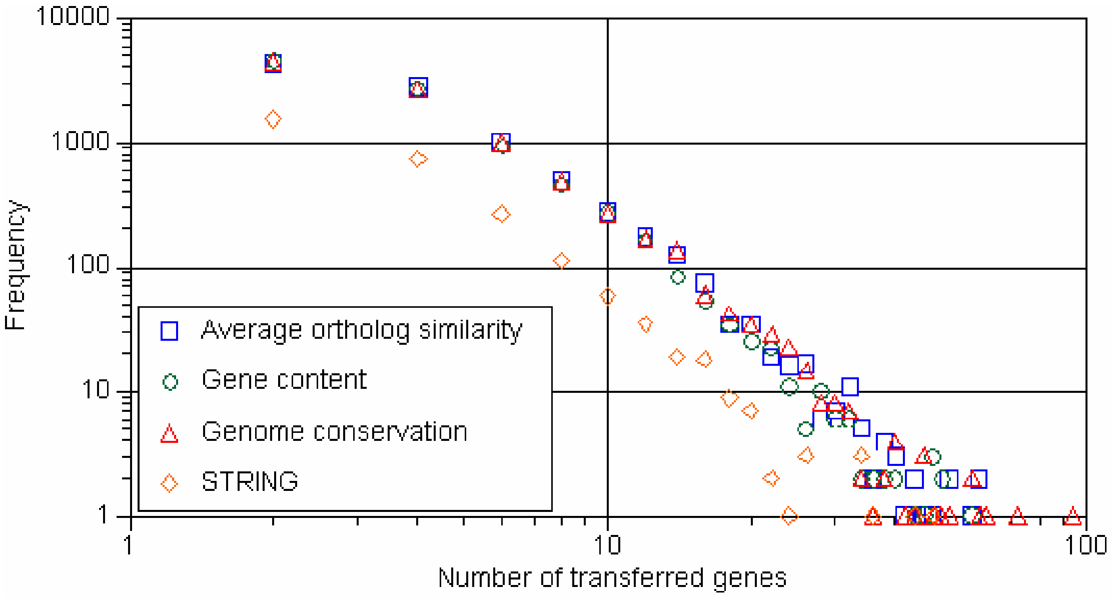


Figure 11. Distribution of HGT vine widths.

Connectivity of the network

To investigate the properties of the HGT network, we removed the underlying (vertical inheritance) tree from the net of life. Since our inference of HGT vine widths is probabilistic (see **Methods Section 5.1.5**), we had to select a meaningful threshold to depict the inferred events. Thus, to investigate the connectivity of the HGT network, we experimented with several thresholds, namely one (a single HGT), five and ten. Irrespectively of the tree used and dataset, the HGT network displays small-world behaviour, with the diameter of the network fluctuating between 5 and 6.

When higher thresholds are chosen for the analysis, the network also demonstrates power law distribution of connectivity of nodes (**Table 7b**), once again irrespectively of the data set or the tree used (**Figure 12**). This power-law signal is obscured at the lowest thresholds, where many nodes appear to have high connectivity. We suggest that this deviation from the power law is a result of noise inevitable when a probability model is examined at low thresholds, i.e. possibly containing more false positive instances. Our usage of thresholds higher than one for evidence of HGT is indeed reinforced by biological observations that genes often travel between organisms as groups rather than singletons (Boucher, Douady *et al.* 2003). We thus conclude that the HGT network is likely to have a power-law distribution of connectivity, and thus be scale-free.

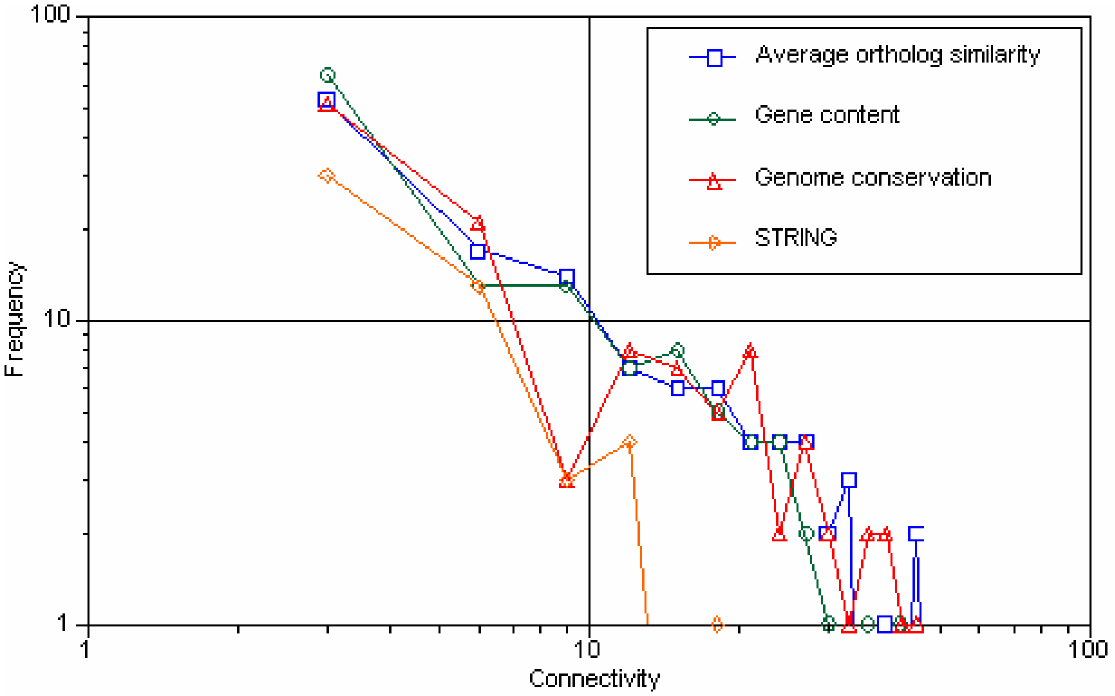


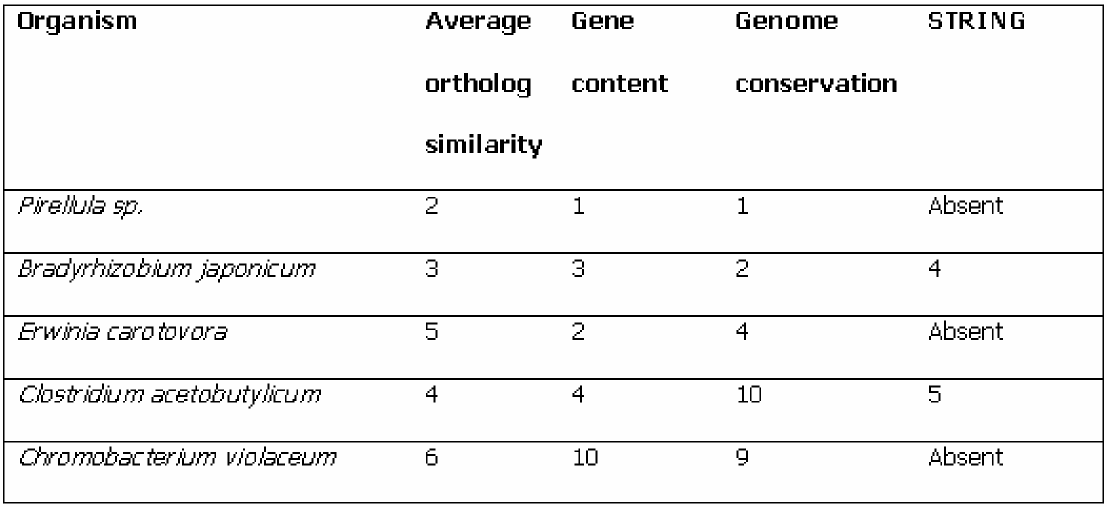
Figure 12. Connectivity of the HGT network.

HGT champions

We aimed at investigating the HGT network in search of hubs and the widest HGT vines. Unlike the global properties of the network which are virtually identical and independent on the dataset, the exact number of predicted gene transfers between two nodes is highly dependent on the tree structure. Incorrect tree architecture can cause the mistaken inference of high amounts of HGT, particularly when two related organisms are positioned distantly on a tree. We thus aimed to exclude tree architecture bias from our analysis and examined results consistent between different tree architectures. Also, since the tree architectures are different, inner nodes (i.e. ancestral states) are often incomparable, and thus we limited the analysis to the leaves (terminal nodes) of the tree, i.e. the sequenced genomes from contemporary species.

When examining 165 microbial genomes for the network hubs, certain species came out on the top of the connectivity list with a remarkable consistency between the results obtained from different trees and datasets (**Table 8**). We found *Pirellula sp., Bradyrhizobium japonicum* and *Erwinia carotovora* always at the top of the list of (terminal) nodes with the largest number of HGT partners. Interestingly, the original genome report for *Pirellula* sp. provides certain hints for HGT events in this species (Glockner, Kube *et al.* 2003). Furthermore, there is evidence for HGT between *B. japonicum* and *E. carotovora* in the literature (Streit, Schmitz *et al.* 2004). In conclusion, these hubs can serve as bacterial 'gene banks', providing a medium to acquire and redistribute genes in the microbial communities, either due to specific genetic mechanisms or by virtue of their close proximity to and interaction with other species in their environmental niches.

Table 8. The list of species representing the major hubs in the HGT network and their connectivity ranking in the three trees considered. (HGT vine width threshold is set to 10, see Methods Section 5.1.5). Inner nodes of the tree are ignored during the ranking. 'Absent' signifies absence of the organism in the input data.



Organism	Average ortholog similarity	Gene content	Genome conservation	STRING
Pirellula sp.	2	1	1	Absent
Bradyrhizobium japonicum	3	3	2	4
Erwinia carotovora	5	2	4	Absent
Clostridium acetobutylicum	4	4	10	5
Chromobacterium violaceum	6	10	9	Absent

We have also examined HGT vines that are reported to be wide and consistent across datasets and trees. One of the widest HGT vines is observed between the *Bradyrhizobium* genus (or sometimes the broader Rhizobiales group) of Alpha Proteobacteria and the Beta Proteobacterium *Ralstonia solanacearum*. Phylogenetically distant, both these species are soil bacteria, penetrating plant roots and forming – symbiotic in case of *Bradyrhizobium* (Kiers, Rousseau *et al.* 2003) and parasitic in case of *Ralstonia* (Alfano and Collmer 2004; Genin and Boucher 2004) – relationships with plants. Both cause tumour-like structures, and possess complex molecular mechanisms to interact with the host plants (Sawada, Kuykendall *et al.* 2003). Both bacteria are reported to have acquired large number of genes horizontally

(Kaneko, Nakamura *et al.* 2002; Salanoubat, Genin *et al.* 2002). Careful analysis of the genes that are transferred between the two bacteria can help to understand the mechanisms of pathogen-host interactions in these species, as well in other cases of HGT detected between species with similar life styles.

5.1.4 Discussion

The strongest limitation of the types of the network reconstruction presented here is the inability of the ancestral state inference methods to precisely establish the donor organism for a HGT event. Often a HGT event is inferred across nodes of the tree that existed at different time periods. In this case, GeneTRACE determines the donor *group* of organisms rather than a particular donor species and the prediction should be read as ‘the donor is a progeny of the node’. Though this effect might influence the character of the inferred network, the consistency between the results of medium and high confidence HGT vine width thresholds as well any input data used in this study indicate that the properties of the phylogenetic network reported here are genuine and realistic. A method to correctly infer HGT donors should greatly improve reconstruction of the network.

The GeneTRACE method applied here uses phylogenetic distribution as a marker of HGT events. However, HGT more often occurs between related organisms, followed by homologous recombination (Vulic, Dionisio *et al.* 1997). Rather than introducing new protein families into a genome, this type of HGT causes orthologous gene replacement. In this study, we did not address this mechanism, we focus instead on events that introduce novel protein families into genomes. We are currently working on incorporation of detecting homologous HGT events in the phylogenetic network.

Another limitation is our inability to determine the correct path across organisms when multiple HGT events happened. Though the probabilistic schema described in the **Methods Section 5.1.5** was designed to reduce the impact of this phenomenon, identification of the exact order and direction of HGT events would drastically improve reconstruction of the network.

The hubs of the HGT network presented here might partially result from the phylogenetic coverage of the sequenced species. When the coverage is low, multiple HGT events accumulate on long branches, and an artificial ‘hub’ might appear. Thus, the reconstruction and understanding of the net of life will improve with better phylogenetic representation of sequenced organisms.

The currently acceptable representation of phylogenetic data is in the form of a tree-like structure in a two-dimensional space, often referred to as a ‘dendrogram’ (meaning tree-graph in Greek). This presentation has the limitation of an inherent inability to depict HGT events. We propose to represent the phylogenetic data in the form of a three-dimensional tree, where beyond a tree drawn in the conventional two-dimensional space, HGT vines require a third dimension. When convergence of gene content is particularly high, participating nodes can be drawn closer in the third dimension. An example of such drawing is shown in **Figure 13**, with real data from this study.

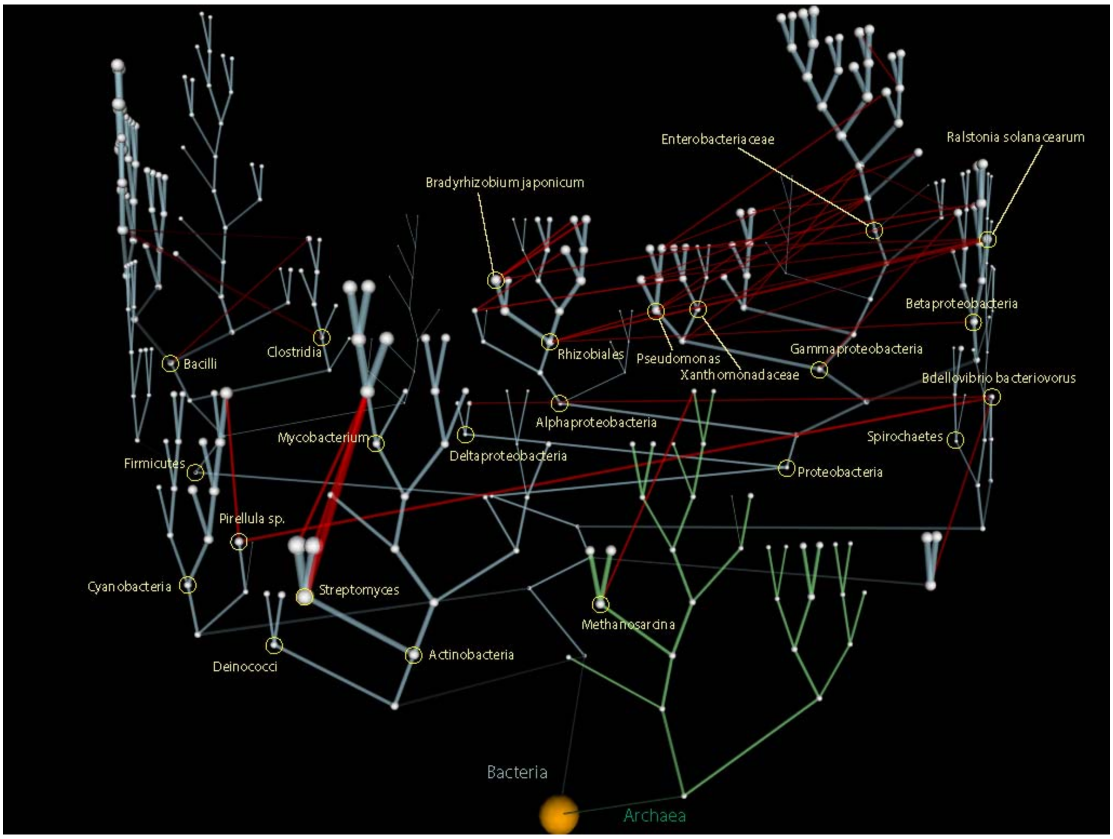


Figure 13. Three-dimensional representation of the net of life. The tree backbone was generated using the average gene similarity approach (see Methods Section 5.1.5). The root is represented as a yellow sphere. Bacteria are shown as nodes on cyan branches and Archaea as nodes on green branches. Red lines correspond to the vines representing HGT. The radius of the nodes is proportional to the estimated gene content size (in terms of number of gene families). Also, the widths of both the vertical inheritance branches and the horizontal inheritance vines correspond to the numbers of gene families transferred by either mechanism. For visualization purposes, only values for HGT vine width above 30 are shown. Certain key species and taxa are labelled.

Our results suggest that the connectivity of microbial HGT network has a power-law behaviour, i.e. the connectivity distribution appears as a decreasing straight line on a log-log scale (**Figure 12**). A network in which connectivity of nodes distributes as a power-law has also scale-free and small-world properties. Scale-free networks display identical properties when any random subset of the complete network is sampled, suggesting that our conclusions should not be strongly affected by an ever-increasing number of genomes.

In a small-world network, the average shortest path between any two of its nodes (termed 'network diameter') involves traversing only relatively few nodes. This has a profound ecological meaning and strong implications for genome evolution. In the context of the HGT network, a small-world structure means that a substantially beneficial gene appearing in any organism can swing across species barriers and reach any other organism via a very small number of HGT events. In fact, this prediction of our hypothesis has an independent verification from the 'experiment' of antibiotics-resistance genes that are known to spread extremely rapidly across species (Jacoby 1996), or the preferential involvement of specific functional classes (Nakamura, Itoh *et al.* 2004). Though most of the reported instances of drug resistance involve pathogenic bacteria, based on the scale-free model we predict that the initial donor and final acceptor organisms might have nothing in common in terms of phylogenetic origin, ecological niche or geographical distribution, and communicate indirectly through the 'hubs' in the network of life.

5.1.5 Methods

In order to reconstruct the phylogenetic network of microbial species, we required a dataset of orthologs across all currently sequenced species. We used BLASTP (Altschul, Madden *et al.* 1997) to find best bidirectional hits across 165 microbial genomes in CoGenT database release 184 (Janssen, Enright *et al.* 2003). To eliminate paralogy, we used only bidirectional best hits across genomes. We then clustered these hits using Markov clustering algorithm (MCL) (Enright, Van Dongen *et al.* 2002). The exhaustive nature of this schema ensures that all genes that had at least one bidirectional best hit in another organism are represented (Goldovsky, Janssen *et al.* 2005). We call the resulting protein families used for the analysis described herein as the 'OFAM' dataset. This dataset is accessible at http://cgg.ebi.ac.uk/services/ortho-fam/

To ensure that the results are not an artefact of the orthology definition, we used orthology information for 110 species from the STRING database (von Mering, Huynen *et al*. 2003), from which we cross-linked 106 species to CoGenT, resulting in 98 prokaryotic species, after excluding Eukaryotes. STRING adopts the definition of orthologs as groups of homologs built from at least one triplet of best-matching pairs of sequences, also known as Clusters of Orthologous Genes (COGs) (Tatusov, Koonin *et al*. 1997).

To reconstruct the microbial phylogenetic network, we required a phylogenetic tree. There are many methods for the reconstruction of phylogenetic trees (see **Introduction Section 5.1.2**), however neither guarantees 100% accuracy. To avoid biases generated by any single tree, we used three methods of genome-based phylogenetic reconstruction, namely gene content (Korbel, Snel *et al*. 2002), average gene similarity (Clarke, Beiko *et al*. 2002) and genome conservation (Kunin, Ahren *et al*. 2005). The first method derives phylogenetic distances from conservation of gene content, the second uses only sequence similarity between genomes, while the third combines the two measures to achieve maximum precision and contrast (Kunin, Ahren *et al*. 2005). While being based on complete genomes, all these methods produce phylogenies that are remarkably similar to the classical 16S rRNA trees. All methods are implemented as it appears on the Genome Phylogeny Server (http://cgg.ebi.ac.uk/cgi-bin/gps/GPS.pl) and described elsewhere (Kunin, Ahren *et al*. 2005). Only results consistent across different trees and with consistently high jacknife scores (Kunin and Ouzounis 2003b) are considered robust. For STRING data, we used a gene content tree constructed according to (Korbel, Snel *et al*. 2002).

Just as there are many methods to reconstruct phylogenetic trees, there are several available methods to identify HGT events. We could not use methods that are based on identification of biased GC content or codon usage, as these can only identify recently acquired genes and are not designed to reconstruct early events. We thus used GeneTRACE – a method that identifies HGT from the phylogenetic distribution of protein families on the tree of life (Kunin and Ouzounis 2003b). GeneTRACE assumes that presence of a gene family in multiple members of a clade reveals its ancestral nature, absence of a gene in some members of a clade indicates gene loss, and patchy presence of the gene family in distantly related clades implies HGT. This method was shown to have at least 90% accuracy on simulated data (Kunin and Ouzounis 2003b) and at least 81% accuracy on biological data (Kunin and Ouzounis 2003a), being capable of reconstructing HGT events on most levels on the tree of life.

A limitation of the GeneTRACE approach to reconstructing HGT events is its inability to distinguish between the donor and the acceptor genomes (Kunin and Ouzounis 2003b). Thus, a gene that was extensively transferred horizontally creates links between all lineages that possess the gene, regardless whether they were involved in the particular transfer or not. We thus adopted a schema for normalization of the number of transferred genes, to avoid multiple counts of a single HGT event, as below.

Consider a situation when a protein family appears twice in distant sections of a tree. In this case, at least one HGT event may be necessary to explain the phylogenetic distribution of the family. Consider now a protein family that has three dispersed roots within a tree. Then, at least 2 horizontal transfer events are necessary to explain the distribution. However, simple linking of all nodes creates 3 possible edges for horizontal transfer. Assuming equal probability for all possible scenarios, we then assign the value of 2/3 as a probability for each possible event to be depicted correctly (and 1/3 for an incorrect detection). Thus, while the minimal number of edges required to connect all nodes (*n*) by HGT is *n-1*, the number of all possible connections is *n(n-1)/2*. This gives us the probability that each of the edges describes a valid HGT event as *(n-1)/(n(n-1)/2)*, or *2/n*. Thus, to each edge that connects independent origins of a protein family, previously labelled by GeneTRACE as arising from HGT, we assign a probability of *2/n*.

To describe the inferred sum of all HGT events between two nodes within an evolutionary net, we sum up all probabilities of transfer for each gene family transferred between the two nodes and term the resulting edge as ‘vine’ and the weight of the edge as ‘vine width’.

5.1.6 Acknowledgements

We thank members of the Computational Genomics Group for useful discussions. CAO acknowledges support from the UK Medical Research Council and IBM Research.

5.1.7 Disclaimer

My contribution to this work has been in data analysis and data visualization, as well as discussions with the rest of the authors. In particular, I have developed a PERL script for the creation of a VRML representation of the net of life (see **Figure 13** above) – useful advice from the rest of the authors have assisted greatly towards the final result. This work has now been published (Kunin, Goldovsky *et al.* 2005).

5.2 A minimal estimate for the gene content of the last universal common ancestor – exobiology from a terrestrial perspective

Paper presented to the "3rd International Workshop on Space Microbiology" organized by the European Space Agency at Mol, Belgium, 22-25 May 2005

5.2.1 Abstract

Using an algorithm for ancestral state inference of gene content, given a large number of extant genome sequences and a phylogenetic tree, we aim to reconstruct the gene content of the Last Universal Common Ancestor (LUCA), a hypothetical life form that presumably was the progenitor of the three domains of life. The method allows for gene loss, previously found to be a major factor in shaping gene content, and thus the estimate of LUCA's gene content appears to be substantially higher than that proposed previously, with a typical number of over 1,000 gene families, of which more than 90% are also functionally characterized. More precisely, when only prokaryotes are considered, the number varies between 1,006 and 1,189 gene families, while when eukaryotes are also included, this number increases between 1,344 and 1,529 families, depending on the underlying phylogenetic tree. Therefore, the common belief that the hypothetical genome of LUCA should resemble those of the smallest extant genomes of obligate parasites is not supported by recent advances in computational genomics. Instead, a fairly complex genome similar to those of free-

living prokaryotes, with a variety of functional capabilities including metabolic transformation, information processing, membrane/transport proteins and complex regulation, shared between the three domains of life, emerges as the most likely progenitor of life on Earth, with profound repercussions for planetary exploration and exobiology.

5.2.2 Introduction

The quest for extraterrestrial life (Des Marais and Walter 1999; Horneck 1995) in this early phase of space exploration within our solar system will necessarily focus on microbial life forms, perhaps resilient to extreme environmental conditions such as temperature, pressure and chemical composition (Cavicchioli 2002; Cleaves and Chalmers 2004). Recent exciting advances from planetary exploration of the three so-called Jupiter's Galilean satellites (Showman and Malhotra 1999) include the discovery of an ocean under Europa's surface (Carr, Belton *et al.* 1998), organic molecules in Callisto and Ganymede (McCord, Carlson *et al.* 1997), even solid oxygen (Vidal, Bahr *et al.* 1997). Thus, if there is any hope of identifying non-terrestrial microorganisms via relevant biomarkers (Chyba and Phillips 2001), efforts must focus on the delineation of life's extreme conditions for existence (Cavicchioli 2002), only recently unravelled by the exploration of planetary biodiversity on Earth and the spectacular progress in environmental genomics (Venter, Remington *et al.* 2004). The ability of certain microbial forms to endure extremes of pressure (>20MPa) (Prieur, Erauso *et al.* 1995; Yayanos 1995), temperature (≥121^{o} C) (Kashefi and Lovley 2003), pH/salinity (e.g. pH 5 and NaCl concentrations near 4M) (Kamekura 1998) and other parameters (Rothschild and Mancinelli 2001), is primarily defined by their genotypic properties, in other words by their gene content. For instance, through genome sequencing, it was found that the UV-resistant *Deinococcus radiodurans* contains multiple components, including DNA repair systems, that render it capable of withstanding high doses of UV radiation (White, Eisen *et al.* 1999). Other microbes have been found to thrive in ice, boiling water, acid, water cores of nuclear reactors, salt crystals and toxic waste (Cavicchioli 2002).

As Thomas Gold commented on subsurface microbial ecosystems, “[…] life exists in all the locations where microbes can survive” (Gold 1992).

More than 100 molecules have been found in the interstellar medium (ISM), including glycolaldehyde and ethanol. ISM organic molecules have been detected (Snyder 1997), including the controversial, alleged spectroscopic discovery of glycine (Kuan, Charnley *et al*. 2003), later rebutted by others (Snyder, Lovas *et al*. 2005). Cometary (Chyba, Thomas *et al*. 1990) and meteoritic (Bernstein, Dworkin *et al*. 2002) bombardment of the early planets is thought to have played a major role in the delivery of various organic compounds. Given that interplanetary – l*et al*one interstellar – transfer of life may be unlikely due to the vast distances between planetary systems (e.g. Earth and Jupiter) (Cleaves and Chalmers 2004), the chance that life can be found on jovian moons, for instance, will rely on the premise that these life forms might have been formed independently at the early stages of the formation of our solar system (Chyba and Sagan 1992), following similar self-organization processes as those on Earth (de Duve 2003; Nisbet and Sleep 2001). Despite the fact that survival in extreme conditions does not necessarily imply that life actually originated in such environments (Cleaves and Chalmers 2004; Miller and Lazcano 1995), it is possible that primordial features might include certain aspects of extremophiles. It is not inconceivable, therefore, to assume that certain organisms, found in the most extreme conditions on our planetary biosphere, might potentially tolerate (even thrive in) conditions elsewhere in our solar system, provided that adequate sources of energy – via, e.g., chemolithotrophy – can be utilized (Chyba and Phillips 2002). Such known and observable energy sources include the strong magnetic field of Ganymede and the high-temperature volcanism of Io (Showman and Malhotra 1999). Theoretical studies that take into account chemical cycling, thermal and osmotic gradients as well as magnetic or gravitational fields indicate that availability of free energy to harbour life on Europa is indeed possible (Schulze-Makuch and Irwin 2002). Thus, the search for life in the solar system must rely on the identification of common, possibly primordial, features and genetic composition between known terrestrial life forms and their postulated extra-terrestrial counterparts.

Consequently, the genome reconstruction of the Last Universal Common Ancestor (LUCA) (Lazcano and Forterre 1999b), baptized as such at a workshop in France in 1996 (Lazcano and Forterre 1999a) by Kyrpides and Ouzounis (Kyrpides, Overbeek *et al.* 1999), is of paramount importance in our understanding of the most primitive life forms that once emerged on Earth, 3.8 Million years ago (Nisbet and Sleep 2001). Previous attempts include the intersection of common protein families (Ouzounis and Kyrpides 1996a), the identification of common functional properties on a genome-wide scale (Kyrpides, Overbeek *et al.* 1999), as well as theoretical approaches with respect to the nature of this hypothetical entity (Woese 1998) and horizontal gene transfer (HGT) (Woese 2002). This discussion has been confounded in the past by the role of HGT in molecular evolution (Doolittle and Brown 1994), and in particular the difficulties in quantifying HGT in an objective way across taxa (Doolittle 1999). Moreover, there have been various controversies as to the nature of LUCA with regard to its mesophily (Penny and Poole 1999) or thermophily (Glansdorff 2000), simplicity or complexity (Forterre and Philippe 1999), monophyly or polyphyly (Brown, Douady *et al.* 2001) and the conceptual difficulties in rooting the tree of life (Bapteste and Brochier 2004). Note that most of these analyses were based either on cytological information, e.g. RNA processing (Penny and Poole 1999), or sets of proteins, e.g. 23 orthologous families (Brown, Douady *et al.* 2001), with certain exceptions (Kyrpides, Overbeek *et al.* 1999; Ouzounis and Kyrpides 1996a).

More recently, two major advances in comparative and evolutionary genomics allow us to approach the problem of the nature of LUCA from a fresh perspective, by taking into account entire genome sequences, their evolutionary relationships and the reconstructed gene content of ancestral forms. First, a host of methods have been devised that take into account whole-genome comparisons for phylogeny construction and inference, which include gene content (CT) (Fitz-Gibbon and House 1999; Snel, Bork *et al.* 1999; Tekaia, Lazcano *et al.* 1999), average sequence similarity (AS) (Clarke, Beiko *et al.* 2002) and genome conservation (GC) (Kunin, Ahren *et al.* 2005). Whole-genome phylogenies are far more robust with respect to HGT, undetected paralogy and variable evolutionary rates (Kunin, Ahren *et al.* 2005; Snel, Bork *et al.* 1999) and it has been shown that genome conservation is by far the most robust method to detect similarities from the strain up to the domain level (Kunin, Ahren *et al*. 2005). Furthermore, a number of parsimony-based algorithms have been developed that allow the reconstruction of ancestral states using as a starting point the gene content (as represented by protein families) of extant organisms and a given phylogenetic tree, allowing both for HGT and gene loss. The first such method was proposed in 2002 (Snel, Bork *et al*. 2002), using a limited number of genomes and lacking a rigorous estimation for the HGT penalty parameter, followed by a precisely formulated algorithm, called GeneTRACE (Kunin and Ouzounis 2003b), which attempted to calibrate the HGT penalty parameter on the basis of two independent assumptions for genome evolution and a multitude of genomes (Kunin and Ouzounis 2003a). Other similar algorithms also based on parsimony have been proposed later (Mirkin, Fenner *et al*. 2003).

5.2.3 Materials and methods

The first 184 genome sequences that were both published and released in the public domain were obtained from CoGenT, the complete genome tracking database (Janssen, Enright *et al*. 2003).

An all-vs-all comparison of all proteins was previously performed using BLAST (Altschul, Madden *et al*. 1997) and all reciprocal best hits (Overbeek, Fonstein *et al*. 1999) were obtained and further clustered with the Markov Clustering algorithm (Enright, Van Dongen *et al*. 2002). We name this dataset OFAM, for putative Ortholog Family clustering (Goldovsky, Janssen *et al*. 2005), since reciprocal best hits have been previously proposed as an operational definition of orthologs (Overbeek, Fonstein *et al*. 1999). The families (clusters) obtained at this stage roughly correspond to protein families (Enright, Kunin *et al*. 2003), with a much higher granularity due to the stricter criterion of including only reciprocal best hits. Their annotation records are thus expected to reliably describe their functional properties, when available – these are automatically extracted from the description lines of family members (Cluster Annotator, **Section 4.3**). In total, 37,402 putative ortholog families have been analyzed across 184 entire genomes.

Having defined the gene content of extant species by their corresponding OFAM cluster representatives, we then constructed species trees for gene content, average sequence similarity and genome conservation as previously described (Kunin, Ahren *et al*. 2005). We have also opted for two variants, by either excluding or including Eukaryotes.

Ancestral state reconstructions were performed using GeneTRACE (Kunin and Ouzounis 2003b) with default parameters, for three different trees and with/without Eukaryotes, thus resulting in six different estimates for the gene content of LUCA, with regard to OFAM protein families and their corresponding annotations. A short discussion of some limitations of the method is pertinent here: first, a HGT event between deeply branching Bacteria and Archaea could place that particular family at the root, and these cases would result in very low bootstrap values (Kunin and Ouzounis 2003b) (which have not been treated further in this analysis but are available on the web site, see below); second, multiple HGT events can obliterate any evolutionary trace, so that the corresponding families could be incorrectly attributed to an ancestral, instead of a derived, state (Kunin and Ouzounis 2003b); third, scenarios regarding the fusion of Eukaryotes with Archaea (Margulis 1996) are not taken into further consideration, as they are not yet firmly established.

5.2.4 Results

The estimated gene content of LUCA on the basis of 184 entire genomes and three different methods of whole-genome phylogeny construction and three taxonomic combinations is surprisingly robust. The gene content, average sequence similarity and genome conservation methods yield a core of 1,042 (CT), 946 (AS) and 891 (GC) ortholog families, respectively. Interestingly, when excluding Eukaryotes, the corresponding Archaea/Bacteria-specific families are 147 (CT), 160 (AS) and 115 (GC) only, less than 10% of the total estimates. The inclusion of Eukaryotes, despite their limited number of representative species, adds 487 (CT), 398 (AS) and 505 (GC) families. In total, the approach results in 1,676 (CT), 1,504 (AS) and 1,511 (GC) families.

This analysis can be depicted as a Venn diagram, however, the three variant tree construction approaches are represented independently (**Figure 14**). The gene content (CT) tree in both cases produces the largest estimate (without or with Eukaryotes at 1,189 or 1,529 respectively) possibly suffering from over-detection of gene loss (Kunin, Ahren *et al*. 2005), while the smallest estimates are produced by the other two tree construction methods, without Eukaryotes at 1,006 by genome conservation (GC) and with Eukaryotes at 1,344 by average sequence similarity (AS). While these estimates are robust, there clearly exist sizeable differences, and multiple intersection of these sets results in variable annotation sets. We make these results available for further investigation by the scientific community[9].

[9] http://www.ebi.ac.uk/research/cgg/projects/phylogeny/luca/

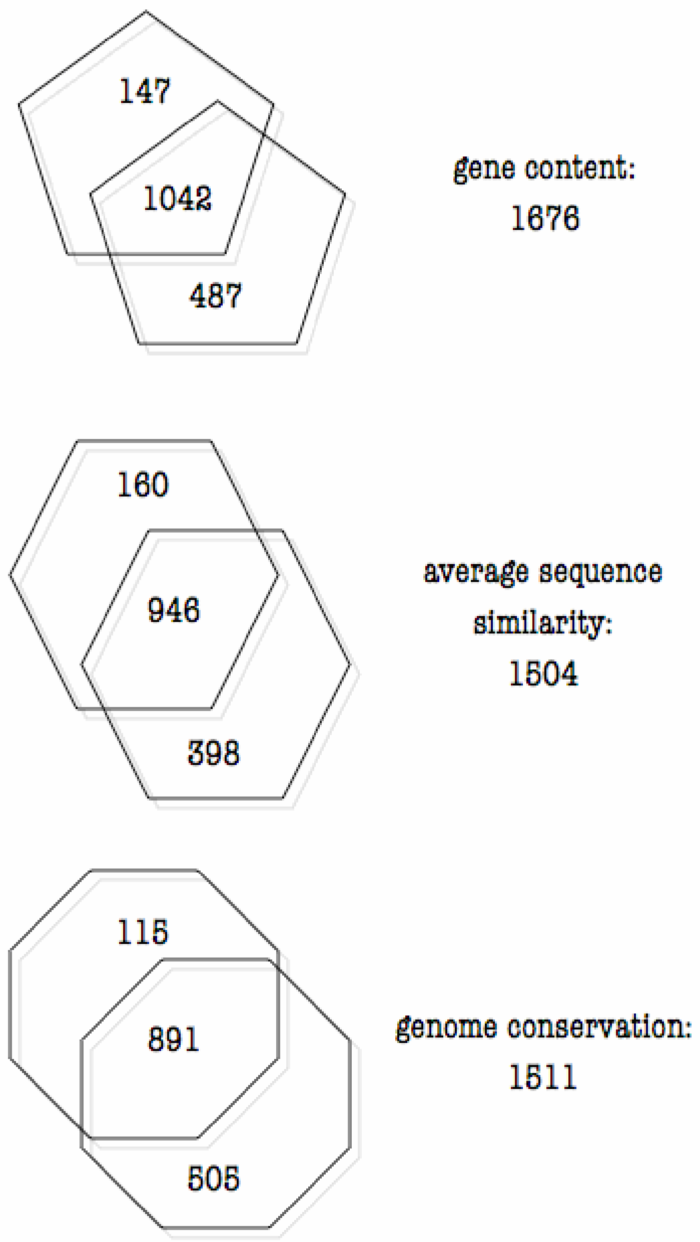


Figure 14. A representation of the minimal gene content for LUCA. Upper diagrams represent analyses without Eukaryotes, lower diagrams represent analyses with Eukaryotes; pentagons represent gene content (CT), hexagons represent average sequence similarity (AS), octagons represent genome conservation (GC) - see Materials and methods Section 5.2.3. The number of unique (outside the intersection) and common (inside the intersection) gene families per category are given in the diagrams; the number of total unique families is also provided (listed below the corresponding method).

We have previously shown that trees based on genome conservation are substantially more robust than those derived from the other two genome-wide tree construction methods (Kunin, Ahren *et al*. 2005). In this sense, while we believe that the gene content of LUCA is probably most accurately reflected by the inclusion of Eukaryotes and the genome conservation tree, corresponding to 891 (core) + 505 (with Eukaryotes) = 1,396 ortholog families, this estimate is within a considerably narrow range compared to those obtained from the other two methods, namely gene content (1,529 or 10% above) or average sequence similarity (1,344 or 4% below) (**Figure 14**). These gene content estimates come as a challenge to the widely held view of a 'minimal' genome, allegedly supported by functional genomics experiments of essential genes (Koonin 2003). While it would be pertinent to analyze the results obtained by genome conservation alone, we opt for the estimate obtained from the common intersection of the three methods, which generates 669 ortholog families, which are entirely covered by 561 functional description classes (**Appendix B**). This set of 669/561 sequence/function categories can be considered as the truly minimal estimate for the gene content of LUCA. Depending on particular questions, further analysis using a particular tree construction method or the inclusion (or not) of Eukaryotes can be justified in the future.

The interpretation of LUCA's truly minimal gene content based on the intersection of the three methods necessarily requires comparison with previous results. We will attempt to provide a description of the main functional categories (**Appendix B**), by referring to the case number (first column) as necessary. Reference to **Appendix B** is included in brackets for all functional descriptions; citations are given only for either recent functional characterizations or instances that are not evident why they are classified under the corresponding categories. Inevitably, the following paragraphs list the various functional properties of the ortholog clusters detected by all methods in a concise manner. Full results and pointers to sequence identifiers are available on the web site.

REPLICATION/RECOMBINATION/REPAIR/MODIFICATION

The gene content of LUCA with respect to DNA processing (replication, recombination, modification and repair) contains a wide range of functions. The following families/functions are identified: DNA polymerase (76-77), excinuclease ABC (252-254), DNA gyrase (69-70) and topoisomerase (79), NAD-dependent DNA ligase (72), DNA helicases (57, 59, 71, 301-302), DNA mismatch repair MutS (73-74) and MutT (97, 353), endonucleases (248-249), RecA (440), chromosome segregation SMC (204), methyltransferase (342, 540), methyladenine glycosylase (9) and adenine glycosylase (49), adenine phosphoribosyltransferase (144), deoxyribodipyrimidine photolyase (231), integrase (319), HAM1 (88), Sir2 (122)- involved in various aspects of genomic stability (Lombard, Chua *et al*. 2005), TatD (125)- a recently discovered DNase (Wexler, Sargent *et al*. 2000), histone deacetylase (308), and restriction modification (515). Thus, one can reason that most aspects of DNA metabolism and information processing are well-represented in the minimal reconstruction of LUCA (DiRuggiero, Brown *et al*. 1999).

TRANSCRIPTION/REGULATION

It does not come as a surprise that reconstructed transcription and its regulation contains just a handful of protein families/functions, since this process across the three domains of life is so divergent. We have previously shown that transcription exhibits a remarkable degree of taxon specificity (Coulson, Enright *et al*. 2001). Apart from RNA polymerase (531), only four regulator families are detected: AsnC (501), ArsR (546), iron-dependent repressor (322) and ferric uptake regulator (262). Two processing families are represented by transcription-repair coupling factor mfd (500) and RNA helicase (61). Finally the bi-functional transcriptional regulator-GntR-aminotransferase class I (502) is also included here, along with wRBA, a trp-repressor binding protein family (511). The deep phylogeny of transcription has been discussed elsewhere, and our findings are consistent with those studies (Kyrpides and Ouzounis 1999; Ouzounis and Kyrpides 1996b).

TRANSLATION/RIBOSOME

Undoubtedly, the analysis yields a number of translation-related protein families, known to be highly conserved during evolution (Ouzounis and Kyrpides 1996a). This particular group serves as an internal control for the validity of the entire approach as well, and it is encouraging that most of the findings in this category are confirmatory of previous analyses. This set includes 17 aminoacyl-tRNA synthetases – including bi-functional Gln/Glu-tRNA synthetase (284) and the two subunits of Phe-tRNA synthetase (374-375) – covering 18 (or 19 if Asn-tRNA is considered to be covered by a dual-specificity enzyme or Gln amidotransferases (Raczniak, Becker *et al*. 2001), 285-287), with the sole exception of Tyr-tRNA synthetase. Certain – but not all – ribosomal proteins (12 small and 9 large subunit) are also identified, along with ribosome modification enzymes (454-456). Furthermore, a set of key translation initiation factors (506-508) as well as elongation factors EF-G (504) and EF-Tu (505) are found. Finally, the following families can also be assigned to this class: translation-associated protein SUA5 (123) (Teplova, Tereshko *et al*. 2000), rRNA methyltransferase sun family (481) and the modification enzymes queuine tRNA-guanine ribosyltransferase transglycosylase (435) and tRNA pseudouridine synthase (483-484). As previously observed, translation is conserved (Kyrpides, Overbeek *et al*. 1999; Kyrpides and Woese 1998b), thus ultimately found to be part of LUCA.

RNA PROCESSING

Three families/functions can only be classified under this category: ribonucleases (446-448), RNA methyltransferase (120) and HIT pyrophosphatases (89), with a role in RNA processing (Liu, Rodgers *et al*. 2002). This is a remarkable finding, given the importance of RNA processing, yet LUCA does not appear to have contained many representatives from this process. Our findings confirm previous comparative genomics analyses partly addressing the evolution of RNA processing (Anantharaman, Koonin *et al.* 2002; Delaye and Lazcano 2000).

CELLULAR PROCESSES

A number of families/functions involved in various aspects of cell division, thermoprotection, signalling and proteolysis are detected. Namely, cell division is represented by FtsH/Z/Y (192-195). Related to thermoprotection are chaperones DnaJ (196), DnaK/HSP70 (82), chaperonin GroES (2) and GroEL (48), heat shock (299-300), hsp20 (467) and cold-shock (214). Some domains involved in signalling include chemotaxis CheW/A/R (197-199), two-component systems (555-556), GTP-binding (85-86, 550), Ser/Thr kinase (462), tyrosine kinase TrkA (127), GGDEF domain (83) and sensor histidine kinase (457). Finally, groups involved in proteolysis or modification include proteases (60, 63, 421-422, 460-461), terminal (485) and amino-peptidases (161-162, 339-340), oligoendopeptidase F (357), peptidases (369), peptidyl-cis-trans isomerases (370-371) and inhibitors (316).

TRANSPORT/MEMBRANE

Most importantly, LUCA appears to have been a complete cell with well-established membrane systems. Families/functions recovered in this analysis include the following most characteristic groups: (i) ABC transporters: cobalt (50, 211), iron (320-321), molybdenum (345), glycine (293-294), spermidine ABC (470), sugar (477-478), oligopeptide ABC (358-362, 541), phosphate (376-381), amino acid (157-158) and dipeptide (52, 243), other non-specific ABC transporters (e.g. 51, 529-530); (ii) ammonium (163); (iii) heavy-metal ATPases: copper (190) and other P-ATPase (191), magnesium (332) and/or cobalt (333); (iv) multidrug resistance (352); (v) ion ATPases: potassium ATPase A/B/C chains (411-413), sodium (469); (vi) permeases: transport system kinases (509), L-lactate permease (95), glutathione-Na antiporter (289), sodium symporter (424, 468), non-specific antiporters (87, 532), non-specific efflux systems (535-536); (vii) ion channels: chloride channel (200), mechanosensitive channel (336), Trk (126) and other potassium channel (409) and uptake (410); (viii) protein translocases: export SecD/F (257-258) and SecY (414), translocase TatC (121), general secretion pathway components (439) [automatic annotation generates an incomplete string: “ral”]; (ix) other: bacterioferritin comigratory protein (179), non-specific membrane protein families (558), SRP54 (465), arsenical pump membrane (171), Mrp subfamily of ABC transporters (96) and the rhomboid family (337)- previously suggested that it was not present in LUCA (Koonin, Makarova *et al.* 2003). This is the first time that the transport/membrane protein complement of LUCA is described in such detail, thanks to the availability of new data and novel methodologies.

UNCLASSIFIED FUNCTION

A few other classes could not be assigned to any of the above categories, and they are listed here. These included the following cases: CrcB camphor resistance

(Sand, Gingras *et al*. 2003) (215), inorganic pyrophosphatase (317) (Kornberg, Rao *et al*. 1999), TPR-containing proteins (124)- shown to be present in Archaea (Kyrpides and Woese 1998a), and ankyrin repeat proteins (164).

ELECTRON TRANSPORT

A number of key electron transport systems also appear to be part of LUCA's genomic signature. These include ferredoxin oxidoreductase components (7, 261, 431), ferredoxin (260), flavoproteins (245-247), NADH dehydrogenase components (99-112), iron-sulfur proteins (323), thioredoxin reductase (492) and thioredoxin (545), ferrochelatase (263), HesB (90), alkyl hydroperoxide reductase (153), arsenate reductase (170) and finally, superoxide dismutase (482) (Fe/Mn type; correctly, Cu/Zn superoxide dismutase not found in Archaea is not detected). The latter finding lends support to the unorthodox hypothesis on the possible antiquity of aerobic respiration (Castresana and Saraste 1995).

METABOLISM

The most impressive and complex functional complement of LUCA can only be the set of metabolic enzymes. We have previously shown that metabolism is highly conserved (Peregrin-Alvarez, Tsoka *et al*. 2003), so this fact is not entirely unexpected. Indeed, this pattern had already been suggested in the pre-genomics era (Ouzounis and Kyrpides 1996a), and later confirmed with a limited set of genomes (Kyrpides, Overbeek *et al*. 1999). To describe in detail the metabolic complement of LUCA is beyond the scope of the present work; the metabolic sophistication of LUCA with regard to bioenergetics has been discussed elsewhere (Castresana 2001). However, it is important to emphasize that most major pathways are represented, including amino acid, nucleotide, sugar and lipid biosynthesis, and limited degradation. We also identify more specialized reactions that are sometimes confined to close relatives but are also sporadically found across higher taxa. We highlight all families/functions considered to belong to metabolism by green colour (**Appendix B**) for further analysis.

UNKNOWN FUNCTION

There are only 69 ortholog families which were not classified in any of the above categories and are marked in red (**Appendix B**), corresponding to 17 functions (52 families contain no annotation). What is therefore quite remarkable is the fact that in terms of families the set of unknowns represents only 69/669=10% and in terms of functions it represents a mere 17/561=3% of total. Whether this is a veritable property or a historical accident – with research having focused on the most phylogenetically abundant proteins – is not easy to resolve, and it is possible that both factors play some role. Whatever the reason might be, it is a fortunate fact that the hypothetical gene content and functional complement of LUCA is a set of characterized protein families and cellular functions. Under diametrically different circumstances, LUCA could have been just a bag of unknown genes.

5.2.5 Discussion and conclusions

Having automatically deduced the gene content of LUCA with recent methods of phylogeny construction and inference, along with a large number of genome sequences not available previously, it is actually surprising how robust earlier estimates have been (Kyrpides, Overbeek *et al.* 1999; Ouzounis and Kyrpides 1996a). Moreover, it is remarkable that proposals on the nature of LUCA, and its position in the DNA world (Forterre, Benachenhou-Lahfa *et al*. 1992), are now vindicated by the discovery of well-developed cellular functions across a wide range, resembling contemporary organisms. At the same time, the root of the tree of life is still an open question and not, as is commonly believed, decidedly close to Bacteria (Forterre, Benachenhou-Lahfa *et al*. 1992).

In fact, it is rather surprising that many investigators support the idea that Bacteria are closest to the root, given that the analyses offering this view were based on scarce information and suffered methodology problems, in particular the well-known effect of long-branch attraction arising from Bacteria (Gribaldo and Philippe 2002). It has been shown that genome-wide tree construction methods (Snel, Bork *et al*. 1999), and in particular, genome conservation (Kunin, Ahren *et al*. 2005), provide a much more robust way to construct deep phylogenies, and tolerate issues such as horizontal gene transfer (HGT) (Kunin, Ahren *et al.* 2005; Snel, Bork *et al.* 1999), yet having the obvious drawback that entire genomes are required. It has also been shown that ancestral state inference, by taking into account gene loss and HGT (Kunin and Ouzounis 2003a; Kunin and Ouzounis 2003b; Snel, Bork *et al.* 2002), offers subtler estimates of ancient gene content, especially when deep phylogenies and large evolutionary distances are considered. In contrast, typical analyses using intersections of protein families suffer from their inability to identify gene loss, so that the common genes detected are usually few and highly conserved (Harris, Kelley *et al.* 2003; Ouzounis and Kyrpides 1996a).

Most importantly, methodological issues aside, a fundamental issue with regard to the nature of LUCA concerns its first initial, which stands for “Last”. Insightful discussions about the nature of a universal ancestor have been offered (Woese 1998), addressing the origins of life and the emergence of genetic information transfer through heredity (Woese 2002). It should be realized, however, that when the gene content of the *last* (i.e. most recent) ancestor is examined, the question is whether extant domains of life share a sufficient number of gene/protein families that indicate what functional properties of primeval life have been conserved across the aeons. This type of ancestral reconstructions as the one presented here essentially delineate in a sophisticated, yet conceptually elegant, way the common protein families and corresponding functions of the three domains of life. While we will probably never fully understand how the first ancestor emerged, we can glimpse into the nature of LUCA, which does not appear dramatically different from extant life, after all. We hope that our present analysis will provide insights towards the detection of possible life forms elsewhere and more specifically the development of useful biomarkers with respect to the functional properties of a ‘universal cell’.

5.2.6 Disclaimer

My contribution to this work has been in data generation, collection and analysis, as well as discussions with the authors. In particular, my Cluster Annotator PERL script (**Section 4.3**) was used to create the annotation descriptions seen in **Appendix B**. This work is in press (Ouzounis, Kunin *et al.* 2005).

Chapter 6
Similarity Detection

6.1 Sequence Alignment

Sequence comparison is one of the most fundamental operations in computational biology. Procedures relying on sequence comparison are diverse and range from database searches to secondary structure prediction. Sequences can be compared two by two to search databases for homologs, or they can be multiply aligned to examine the evolutionary pattern of a protein family.

Optimal sequence alignment algorithms are implemented using dynamic programming (Holmes and Durbin 1998), ultimately a technique that identifies optimal alignment by maximizing the score of the path that produces it. Although the algorithmic solutions appear satisfactory, the computational load escalates as a power function of the length of the sequences making its use for searching large databases infeasible. Subsequently, a few heuristic approaches were proposed, mostly based on the recognition of alignment 'seeds', with BLAST (Altschul, Gish *et al.* 1990; Altschul, Madden *et al.* 1997) and FASTA (Pearson 1990; Pearson and Lipman 1988) being the most ubiquitous applications.

Multiple alignments constitute a powerful means of revealing the constraints imposed by structure and function on the evolution of a protein family, making it possible to tackle exciting problems; examples (Notredame 2002) relevant to the scope of this Thesis follow.

6.1.1 Identification of conserved motifs and domains

Multiple sequence alignments make it possible to identify motifs preserved by evolution that play an important role in the structure and function of a group of related proteins. Within a multiple alignment, these elements often appear as columns with a lower level of variation than their surroundings. When coupled with experimental data, these motifs constitute a very powerful means of characterizing sequences of unknown function. When a motif is too subtle, one may use another type of descriptor known as a profile or a hidden Markov model (HMM). These are meant to exhaustively summarize (column by column) the properties of a protein family or a domain. Profiles and HMMs make it possible to identify very distant members of a protein family when searching a database. Although it is thought that their sensitivity and specificity is much higher than that provided by a single sequence or a pattern, in **Chapter 8** we show that patterns can arguably provide better results in both sensitivity and specificity if used correctly. In practice, one can derive their own profile from multiple alignments, pre-established collections like Pfam, or compute the profiles on the fly with PSI-BLAST, the position specific version of BLAST. The specificity and sensitivity of a profile are tightly correlated to the biological quality of the multiple alignment it was derived from – this constitutes the breaking point of all the aforementioned applications, an issue that will be referred to again in later **Sections**.

6.1.2 Structure prediction

Structure prediction is another important use of multiple alignments. Secondary and tertiary structure prediction aim at predicting the role a residue plays in a protein structure (buried or exposed, helix or strand etc.). The rationale behind predictions based on a multiple sequence alignment is that the pattern of substitutions observed at a position directly reflects the type of constraints imposed on that position in the course of evolution. In the context of tertiary structure determination or when predicting non-local contacts, multiple alignments can also help to identify correlated mutations (Casari, Sander *et al.* 1995).

6.2 BLAST and PSI-BLAST

Basic Local Alignment Search Tool (BLAST) is one of the most heavily used sequence analysis tools available in the public domain (Altschul, Madden *et al.* 1997). There are several types of BLAST to compare all combinations of nucleotide or protein queries with nucleotide or protein databases. BLAST is a heuristic that finds short matches between two sequences and attempts to start alignments from these 'hot spots'. In addition to performing alignments, BLAST provides statistical information to help decipher the biological significance of the alignment; this is the 'Expect' value or E-value, or false-positive rate (McGinnis and Madden 2004).

Position-Specific Iterated (PSI)-BLAST is the most sensitive BLAST program, making it useful for finding very distantly related proteins. The first round of PSI-BLAST is a standard protein-protein BLAST search. The program builds a position-specific scoring matrix (PSSM or profile) from an alignment of the sequences returned with Expect values better (lower) than the inclusion threshold. The PSSM will be used to evaluate the alignment in the next iteration of search. Any new database hits below the inclusion threshold are included in the construction of the new PSSM. A PSI-BLAST search is said to have converged when no more new database sequences are added in subsequent iterations (http://www.ncbi.nlm.nih.gov/BLAST/).

6.3 Profiles and HMMs

Profile HMMs are statistical models of multiple sequence alignments. They capture position-specific information about how conserved each column of the alignment is, and which residues are likely to be observed. Anders Krogh and co-workers introduced profile HMMs to computational biology (Krogh, Brown *et al.* 1994), adopting HMM techniques which have been used for years in speech recognition. HMMs had been used in biology before, notably by Gary Churchill (Churchill 1989), but the Krogh paper had a dramatic impact, because HMM technology was so well-suited to the popular "profile" methods for searching databases using multiple sequence alignments instead of single query sequences. "Profiles" were introduced by Gribskov and colleagues (Gribskov, Luthy *et al.* 1990; Gribskov, McLachlan *et al.* 1987), and several other groups introduced similar approaches at about the same time. All of the profile methods are statistical descriptions of the consensus of a multiple sequence alignment. They use *position-specific* scores for amino acids (or nucleotides) and position specific penalties for opening and extending an insertion or deletion. Traditional pairwise alignment (for example, BLAST (Altschul, Gish *et al.* 1990), FASTA (Pearson 1990), or the Smith/Waterman algorithm (Smith and Waterman 1981)) uses position-*independent* scoring parameters. This property of profiles captures important information about the degree of conservation at various positions in the multiple alignment, and the varying degree to which gaps and insertions are permitted. The advantage of using HMMs is that HMMs have a formal probabilistic basis to guide how all the scoring parameters should be set. This probabilistic basis allows HMMer to do what most heuristic methods cannot easily. For example, a profile HMM can be trained from unaligned sequences, if a trusted alignment is not yet known. Another consequence is that HMMs have a consistent theory behind gap and insertion scores. Profile HMMs are a slight improvement over a carefully constructed profile – but less skill and manual intervention are necessary to use profile HMMs. This allows to make libraries of multiple profile HMMs and apply them on a large scale to whole genome analysis. One such database of protein domain models is Pfam (Bateman, Coin *et al.* 2004) (The HMMER User's Guide, http://hmmer.wustl.edu/).

6.4 Limitations

In the majority of cases, because of computational limitations, available tools are only heuristics providing an approximate solution to a problem that remains largely open.

A limitation to using PSI-BLAST for large-scale automated protein analysis is that on a small, but certainly not negligible percentage of queries, false positives could enter the list of matches at one iteration with an *E*-value low enough to corrupt the PSSMs constructed for searching in subsequent iterations. In some cases the corruption could result in non-convergence with the accumulation of false positives, or produce biologically incorrect output. But it has to be noted that the procedure is nevertheless powerful, and constantly being improved, e.g. (Schaffer, Aravind *et al.* 2001).

HMMs do have important limitations. One is that HMMs do not capture any higher-order correlations. An HMM assumes that the identity of a particular position is independent of the identity of all other positions. HMMs make poor models of

RNAs, for instance, because an HMM cannot describe base pairs, unless they are hand-crafted.

6.5 Sequence comparison without alignment

Although initially sequence analysis algorithms were mostly borrowed from string processing computer science methodologies (Gusfield 1997), in a second stage biological sequence analysis quickly incorporated additional concepts and algorithms from computational statistics, such as stochastic modelling of sequences. These approaches carry a bias, very clear in present days, that views biological molecules as being linear sequences of discrete units similar to linguistic representations, in spite of their physical nature as a 3D structure and the dynamic nature of molecular evolution. The alignment approach overlooks well-documented long-range interactions and general fluidity resulting from recombination with shuffling of conserved regions without loss of function (Lynch 2002; Zhang, Perry *et al*. 2002). On the other hand, assuming conservation of contiguity allows the employment of a large set of well-developed effective computational procedures. A review of alignment-free methodologies can be found in (Vinga and Almeida 2003).

6.6 Conclusions

Nobody can deny the elegant power and statistical foundations of alignment-based methods like the ones discussed in this **Chapter**. We however want to believe that other methodologies that do not rely on error-prone sequence alignment are a clear alternative. We discuss our work with combinatorial pattern discovery in **Chapters 7**, **8**, and **9**.

Section 8.2 offers further insight into similarity detection methodologies.

Chapter 7

Combinatorial Pattern Discovery

This **Chapter** advocates combinatorial pattern discovery as an attractive alternative to widely used methods, in terms of computational efficiency, biological accuracy and general applicability, where a number of other collaborative projects are also described.

Readers should appreciate this **Chapter** as a repository of ideas, both theoretical and practical. We describe work in progress, and provide preliminary results, for projects that we consider promising and worth pursuing in the future. For various practical reasons we have given priority to projects described elsewhere in this Thesis, and therefore much of the material presented in this **Chapter** has not been published.

7.1 TEIRESIAS

TEIRESIAS is an algorithm for the discovery of rigid patterns in unaligned biological sequences. It has been demonstrated that the algorithm, in the absence of any context information, is able to derive results of proven biological significance (Floratos 1999). What distinguishes TEIRESIAS from the existing methods of discovering local similarities in biological sequences, is the combined effect of the following two features:

- it is very fast (and scales well with the number and the composition of the patterns in the input);
- it finds all the maximal patterns with (a user specified) minimum support.

The main reason for the enhanced performance achieved by the algorithm is the utilization of the convolution operation. Most of the existing pattern enumeration algorithms start with a seed pattern and extend it by a single position at a time, checking at every step of the process whether the new pattern has the required support and pruning the search if it does not. This process guarantees completeness of the results but non-maximal patterns are generated and can be very time consuming, especially if the input contains long patterns. Non-maximal patterns show up almost invariably when the minimum support becomes small enough, making the performance of algorithms in this class prohibitive at such parameter settings. Consequently, and in order to achieve reasonable running times, either the maximum length of a pattern has to be bounded and/or the minimum support must be set very close to the total number of sequences in the input set.

The convolution operation, on the other hand, permits the extension of a pattern by more than one position at a time, allowing for considerable speed up. Furthermore, the ordering of the intermediate patterns when performing the convolutions gives another performance boost by avoiding the generation of redundant patterns. The achieved speed gains afford one the ability to look for patterns with very small supports. This is particularly useful when the composition of the input is not uniform, i.e. when it is comprised of sequences that do not necessarily all belong to one group.

Another property that differentiates TEIRESIAS from existing work, is the kind of structural restriction the user is allowed to impose on the sought patterns. Typically, the speed of the pattern discovery process can be controlled by bounding the length of the reported patterns. This, however, has the drawback that long patterns either escape attention or are broken into multiple redundant and overlapping pieces. With TEIRESIAS, only the parameter W which indicates basically the maximum number of don't care characters (wildcards) between two successive residues in a pattern needs to be set. It thus becomes possible to discover patterns of arbitrary length as long as preserved positions are not more than W residues away.

Finally, TEIRESIAS is guaranteed to report all the maximal patterns meeting the structural restrictions set by the user. Other approaches restrict the search space by incorporating a probabilistic model of importance that is used to decide what patterns to seek. The authors of TEIRESIAS were of the opinion that the assignment of a measure of importance on the patterns should be disjoint from the discovery process. This way all the existing patterns are indeed reported. The task of choosing which of them to keep ought to be a post-discovery problem-specific consideration. In its current implementation the algorithm does not handle flexible gaps.

7.2 Functional Site Discovery and Clustering

7.2.1 Introduction

There exists a vast public knowledgebase regarding biological sequence, structure, and function, a product of immense scientific effort. It is therefore tempting and unquestionably wise to exploit this information in an effort to create better and faster methods in Bioinformatics.

A comparative genomics approach to the discovery of functional sites in protein sequences through the use of combinatorial pattern discovery was previously attempted, however, the amount of input data (hundreds of genomes, hundreds of thousands of protein sequences, millions of amino acids) led to serious problems regarding producing and analyzing results. It became obvious that we needed to filter out, or mask, genomic information that would not contribute to biologically significant results.

7.2.2 Data and Methods

Recently, we built a computational model based on the PDBSITE database (Ivanisenko, Pintus *et al*. 2005) which would capture and describe protein functional features such as enzyme catalytic centres, sites of protein post-translational modification, ligand binding sites, and protein-protein/protein-DNA/protein-RNA interaction sites (i.e. information contained in PDBSITE). In theory, the model should then be able to highlight amino acids with a high probability of being involved in functional sites, based, by design (see next paragraph), on the amino acid itself and its sequence neighbourhood. PDBSITE is based on PDB, and therefore also contains information about secondary and tertiary structure; however, to allow the model to be generally applicable (i.e. where no protein structures are available), we refrained from using this kind of information.

The model was built using the J48 decision tree implementation of the excellent Weka package, a collection of machine learning algorithms for data mining tasks (http://www.cs.waikato.ac.nz/ml/weka/). A decision tree constitutes a supervised approach to classification; it is a simple structure where non-terminal nodes represent tests on one or more attributes and terminal nodes reflect decision outcomes.

The input to Weka included all the amino acids from PDB sequence files involved in functional sites according to PDBSITE. We also constructed the model to take into account the immediate PDB primary sequence neighbourhood of the functionally important amino acids, currently plus/minus five amino acids. The model was then used to mask amino acids in 200 CoGenT genome sequences if they were deemed probabilistically unlikely to be involved in functional sites. The genome sequences were previously filtered for >80% identity using CD-HIT, a program for clustering large protein database at high sequence identity threshold (Li, Jaroszewski *et al*. 2001) – this resulted in a reduction of the protein sequences from 751,742 down to 551,798, i.e. by approximately 25%. The amino acid masking through the use of the Weka model led to a reduction of the number of amino acids TEIRESIAS had to handle by a further 85% (**Figure 15**).

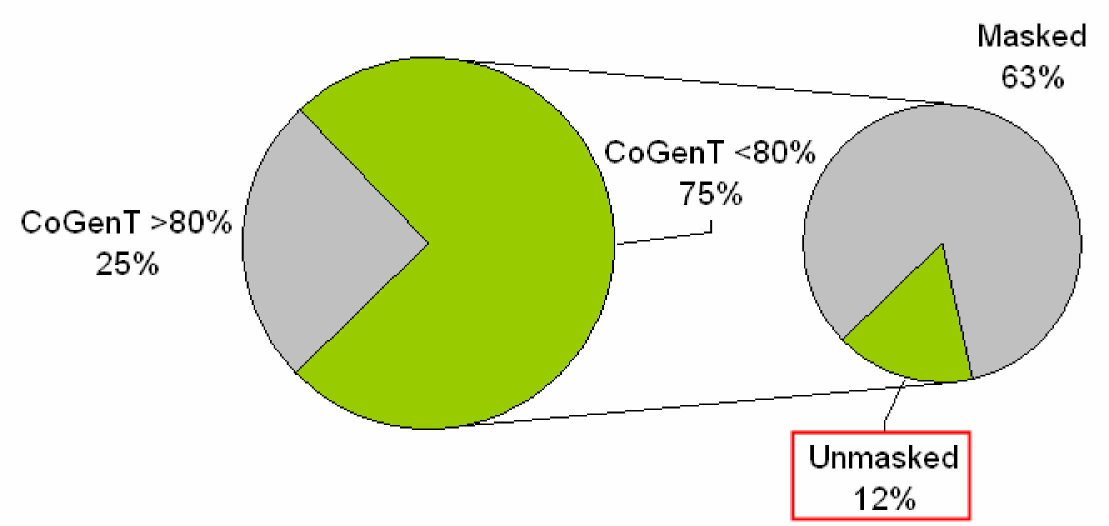


Figure 15. The effect of sequence identity filtering and amino acid masking by the Weka model on the data input for TEIRESIAS (red box).

Looking in detail on the masking of amino acids by the Weka model, it is interesting to note that histidine (H) is only masked by ~40%, by far the lowest percentage of all twenty amino acids (**Figure 16**) – this is not only expected, as histidine is a major contributor to functional sites in proteins and in particular to metal-binding sites (see **Chapter 8** and our study of blue copper-binding domains), but also strongly indicates that our model works. On the same side of the range we find cysteine (C) and arginine (R) with a moderate 66% and 54% masking. On the other side we find alanine (A) (95% masking), leucine (L) (95% masking), and valine (V) (96% masking).

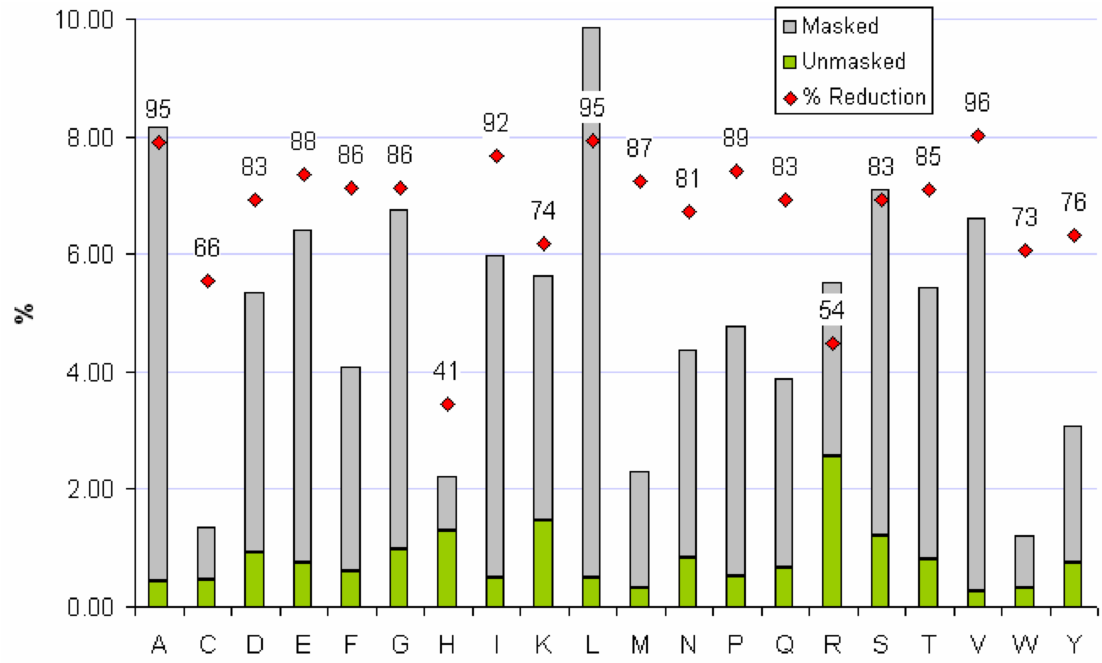


Figure 16. The breakdown of amino acid content in the 80% sequence-identity-filtered CoGenT, and the effect of masking by the Weka model on each one. On the x-axis we list the twenty amino acids. On the y-axis are the percentages of each amino acid in CoGenT represented by the green (unmasked) and grey (masked) bars. The percentage masking (or 'reduction' - see legend) for each amino acid is represented by the red rotated rectangle underneath the value itself.

Pattern discovery was performed on the filtered and masked CoGenT genome sequences using a multitude of TEIRESIAS parameters. Finally, a compromise between sensitivity and computational cost was reached using the following parameters: the minimum number *L* of literals in a reported pattern was set to 3, the maximum distance *W* between any *L* consecutive literals was set to 10, the minimum allowed support *K* for all reported patterns was set to 2, and the maximum allowed support *Q* for all reported patterns was set to 10. Furthermore, pattern overlaps were not allowed (option *–r*).

7.2.3 Results

More than 8,6 million patterns were discovered in a matter of a few hours, a vast improvement over previous attempts. Patterns of length less than 12 were filtered out, as were patterns with more than one appearances in any single protein sequence. The final set contained just over 4,5 million patterns.

Analysis of the data provided the following figures: 469,078, or 85%, of the 551,798 proteins were covered by the 4,5 million patterns. PROSITE (01/03/05 - no profiles, no promiscuous patterns) hit 122,542 (or 22% of) proteins with 1,290 patterns. Fractions of our patterns (73,718 out of 4,5 million) and PROSITE patterns (823 out of 1,290 discovered out of 1,802 in PROSITE) overlapped in 16,003 proteins. Although the proteins in the overlap represent a relatively small subset of the almost half a million proteins in the filtered 200 CoGenT genome sequences, it is important to note that the overlap involves almost 65% of all PROSITE patterns.

7.2.4 Conclusion

We strongly believe that with the tens of millions of amino acids currently available, and the order of magnitude increase expected soon, our approach will allow us to perform pattern discovery and functional analysis in a comparative genomics context on a vast scale.

7.3 Phylogenetic profile patterns

7.3.1 Introduction

Phylogenetic profiling, a relatively recent development in Bioinformatics and a consequence of increasing genome sequence information availability, is a method of assigning functional clues to proteins based on their patterns of inheritance across multiple organisms. What clearly distinguishes the approach from other widely used functional assignment algorithms, is that it is independent of amino acid sequence similarity to proteins of known function.

Phylogenetic profile analysis facilitates the study of protein function by assigning functional clues to uncharacterised proteins, or novel functional clues to proteins of known function. By design, the coverage and accuracy of phylogenetic profile analysis increases with increasing numbers of whole genome sequences, it should therefore benefit immensely from the rapidly increasing genome sequence information available. However, it has been observed that this increased genome availability also reduces the size of clusters of proteins with identical phylogenetic profiles (and in theory similar functions) and eventually results in most clusters having only one protein member in which no functional linkages can be inferred (Wu, Kasif *et al*. 2003).

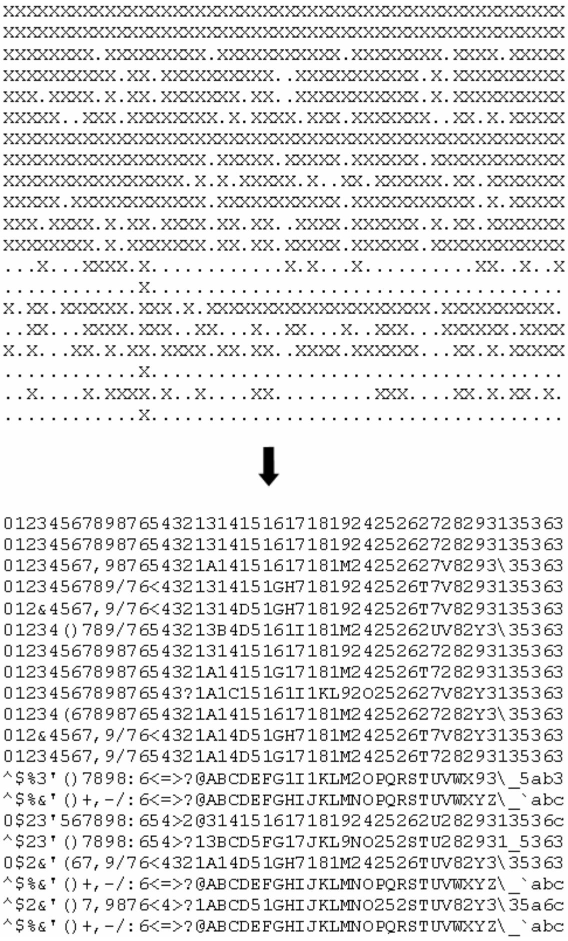
7.3.2 Methods, Results and Observations

Part of our CoGenT++ system (see **Section 3.2**) involves the calculation and storage of phylogenetic profiles of all the proteins stored in CoGenT. Currently, there are more than 250 genomes stored in our database, making the phylogenetic profile binary vector as many bits long. As mentioned in the previous paragraph, this has the drawback of increased cluster granularity because of the decreased probability of two proteins having the same profile.

We attempted to alleviate this problem by performing combinatorial pattern discovery in the phylogenetic profiles set, therefore capturing similar phylogenetic profiles under a phylogenetic profile pattern. Hamming distance, a widely used metric, ignores the position of the mismatches between the vectors, which in the case of phylogenetic profiling can be misleading.

A major obstacle for the use of combinatorial pattern discovery is the sheer number of possible forms that a ~250-character-long binary vector can take, thus making a naive approach computationally very expensive. Also, patterns less than the length of the vector (which in the case of TEIRESIAS cannot be avoided due to restrictions in parameter setting), need to be position specific in order to be meaningful, i.e. a **1011..00.11** pattern can represent different sets of genomes in different proteins in a ~250-long bit vector – avoiding the discovery and report of such patterns can save huge amounts of time and effort.

Ideally, we would use a different ASCII character for each genome and the zero (0) and one (1) state of its bit on the vector (i.e. 500 characters for 250 genomes), thus making the columns genome specific and denying TEIRESIAS discovering the same pattern in different offsets. However, TEIRESIAS cannot use the whole ASCII character set; and even if it could, we are currently very rapidly reaching the point when the number of genomes will be greater than the ASCII character set. So a compromise was reached where ASCII characters would be used more than once but in such a way that would minimise the aforementioned problem. An example is given in **Figure 17**. The current implementation of this ASCII character usage has not yet been optimised by any theoretical model for such a task.



```
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXX.XXXXXXXXX.XXXXXXXXXXX.XXXXXXXX.XXXX.XXXXX
XXXXXXXXXX.XX.XXXXXXXXXX..XXXXXXXXXXX.X.XXXXXXXXXX
XXX.XXXX.X.XX.XXXXXXX.XX..XXXXXXXXXXX.X.XXXXXXXXXX
XXXXX..XXX.XXXXXXXX.X.XXXX.XXX.XXXXXXX..XX.X.XXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXX.XXXXX.XXXXX.XXXXXX.XXXXXXXXXXXX
XXXXXXXXXXXXXXXX.X.X.XXXXX.X..XX.XXXXXX.XX.XXXXXXX
XXXXX.XXXXXXXXXXXX.XXXXXXXXXXX.XXXXXXXXXXX.X.XXXXX
XXX.XXXX.X.XX.XXXX.XX.XX..XXXX.XXXXXX.X.XX.XXXXXXX
XXXXXXXX.X.XXXXXXX.XX.XX.XXXXX.XXXXXX.XXXXXXXXXXXX
...X...XXXX.X............X.X...X..........XX..X..X
............X.....................................
X.XX.XXXXXX.XXX.X.XXXXXXXXXXXXXXXXXXXX.XXXXXXXXXX.
..XX...XXXX.XXX..XX...X..XX...X..XXX...XXXXXX.XXXX
X.X...XX.X.XX.XXXX.XX.XX..XXXX.XXXXXX...XX.X.XXXXX
............X.....................................
..X....X.XXXX.X..X....XX.........XXX....XX.X.XX.X.
............X.....................................
```

```
01234567898765432131415161718192425262728293135363
01234567898765432131415161718192425262728293135363
01234567,987654321A14151617181M24252627V8293\35363
0123456789/76<4321314151GH71819242526T7V8293135363
012&4567,9/76<4321314D51GH71819242526T7V8293135363
01234()789/76543213B4D5161I181M2425262UV82Y3\35363
01234567898765432131415161718192425262728293135363
012345678987654321A14151G17181M242526T728293135363
0123456789876543?1A1C15161I1KL92O252627V82Y3135363
01234(678987654321A14151617181M24252627282Y3\35363
012&4567,9/76<4321A14D51GH7181M242526T7V82Y3135363
01234567,9/7654321A14D51G17181M242526T728293135363
^$%3'()7898:6<=>?@ABCDEFG1I1KLM2OPQRSTUVWX93\_5ab3
^$%&'()+,-/:6<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ\_`abc
0$23'567898:654>2@31415161718192425262U2829313536c
^$23'()7898:654>?13BCD5FG17JKL9NO252STU282931_5363
0$2&'(67,9/76<4321A14D51GH7181M242526TUV82Y3\35363
^$%&'()+,-/:6<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ\_`abc
^$2&'()7,9876<4>?1ABCD51GHIJKLMNO252STUV82Y3\35a6c
^$%&'()+,-/:6<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ\_`abc
```

Figure 17. From binary (for better visualisation, '1' has been substituted by 'X', and '0' by '.'), to ASCII (sample).

Our tests have indicated a significant performance increase: using 674 profiles we achieved a 40% gain, finally making the problem computationally tractable.

Before our approach can be used towards the functional annotation of proteins, a series of issues have to be resolved, including the masking of the evolutionary signal inherent in phylogenetic profiles, when two proteins may travel together across organisms and therefore have the same or very similar phylogenetic profiles just because the organisms are (very) closely related. Furthermore, sequence similarity thresholds for homology assignment can be tweaked to allow only meaningful homologies to appear in the profiles. Finally, the ideal phylogenetic profile should represent the presence or absence of an ortholog for a given protein within the subject genome, as true orthologs are more likely to share the same function than homologs (Wu, Kasif *et al*. 2003).

Leaning on this last point, we have used orthologs (see **Section 3.2.7** for a definition) stored in our CoGenT++ system to construct phylogenetic profiles. Innovatively, we have also introduced a filtering step to increase the robustness of the profiles, by requiring transitivity in orthology: transitivity indicates that if, for example, human gene A is a reciprocal best BLAST hit of worm gene B and fly gene C, then worm gene B and fly gene C also need to be reciprocal best BLAST hits in order for genes A, B, and C to be grouped as orthologs. The idea was that the functional signal contained in the phylogenetic profiles would increase if we required the orthologs of the protein the profile is being built for, to also be orthologs of each other. Finally, we incorporated a sequence conservation threshold between a protein and its potential ortholog. This approach is currently being tested. An illustration of the effects of the modifications is given in **Figure 18**.

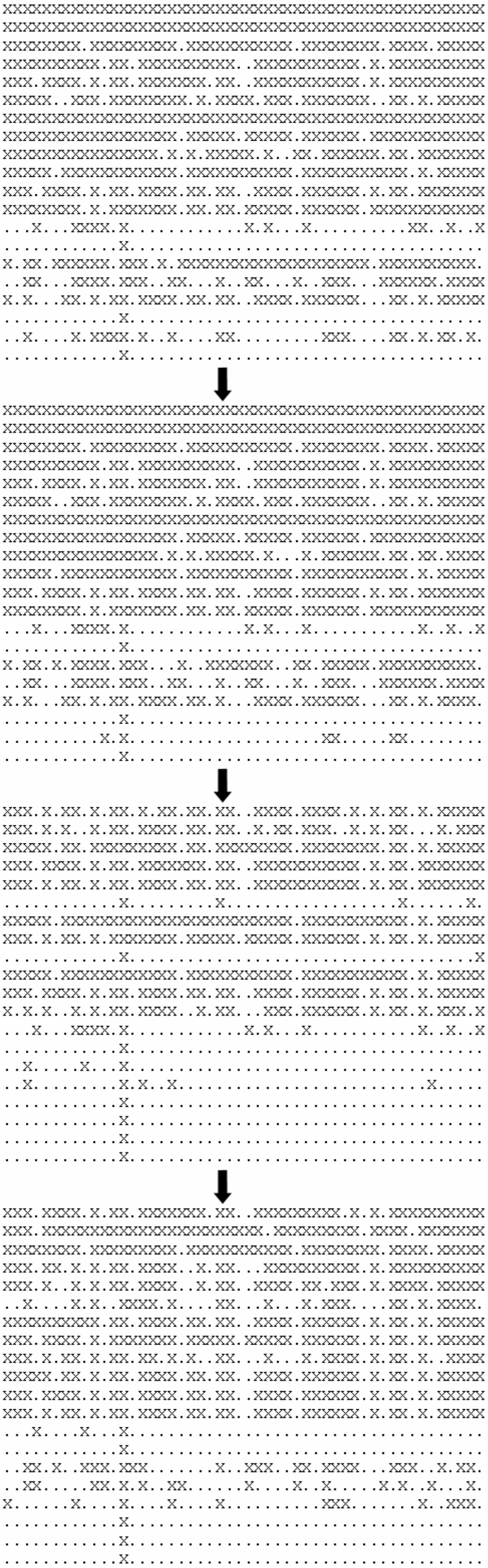


Figure 18. Different phylogenetic profiles: Homolog phylogenetic profiles (671-329) Ortholog phylogenetic profiles (646-354) Above + transitivity filtering (413-587) Above + conservation filtering (501-499) (in parentheses: the number of ‘X’s and ‘.’s, or ‘1’s and ‘0’s, in each set).

7.4 Clusters of orthologs

Recently, the ‘transitivity in orthology’ idea outlined in the previous **Section** was extended and used in the construction of clusters of orthologs (an alternative to OFAM, featuring in **Section 3.2.7**). Calculating the common orthologs of a pair of orthologs provided a scoring function. By example: protein A has 100 orthologs, one of them is B; B, out of A's 100 orthologs, has 80 of them as orthologs; so the A-B ratio is 80/100, or 0.80. We perform this step for all pairs of orthologs showing more that 20% sequence conservation. The arbitrary scoring function we devised incorporates both the number of common orthologs of the orthologs pair and the ratio raised to the power of two (to penalise for low ratios) - using the previous example that is 80 * 0.80^2. Starting with the highest-scoring pair of orthologs, we store them as members of a cluster, denying them appearing again (albeit with a lower score).

After the clustering, we check the clusters for paralogs: if there are two proteins from the same genome in a cluster, we remove the ortholog with the lowest pairwise score; if the scores with which the paralogs were included in the cluster are the same, we take into account the conservation of the paralogs choosing the most conserved. We understand that this is not a complete solution to the paralogy problem – ideally one or more phylogenetic trees would be constructed for each cluster of proteins and the sequence of speciation and duplication events would be defined (Storm and Sonnhammer 2002). However, this was considered beyond the scope of this work.

We used the algorithm to look at 200 CoGenT genomes. In approximate numbers, OFAM produced 230,000 singletons (single member clusters) out of 740,000 proteins, whereas the new approach produced 200,000 singletons. OFAM produces 75,000 clusters, whereas the new approach produced 145,000 clusters, obviously generating more granularity. Preliminary analysis has indicated that there is considerable overlap between the two databases, but no further analysis of the data has been performed.

7.5 Phosphorylation sites prediction on targets of protein kinase Par-1

7.5.1 Introduction

Phosphorylation is the addition of a phosphate (PO^4) group to a protein or a small molecule - in eukaryotes, protein phosphorylation is arguably the most important regulatory event. Many enzymes and receptors are switched "on" by phosphorylation by various specific protein kinases, and "off" by dephosphorylation by phosphatases.

A prime example of the important role that phosphorylation plays is the p53 tumour suppressor gene, whose activation stimulates transcription of genes that suppress the cell cycle, even to the extent that the cell undergoes apoptosis. The p53 protein is extensively regulated, since its activity should be limited to situations where the cell is damaged or physiology is disturbed. Consequently, human p53 contains more than 18 different phosphorylation sites (Appella and Anderson 2000).

As exemplified above, phosphorylation can occur on several amino acids within a protein. Serine (S) phosphorylation is the most common, followed by threonine (T), whereas tyrosine (Y) phosphorylation is relatively rare. However, since tyrosine phosphorylated proteins are relatively easy to purify using antibodies, tyrosine phosphorylation sites are relatively well understood (Chong and Daar 2001). Histidine and aspartate phosphorylation occurs in prokaryotes as part of two-component signalling (Saito 2001).

Protein kinases are enzymes that can transfer a phosphate group from a donor molecule (usually ATP) to an amino acid residue of a protein. Although most protein kinases are specialized for a single kind of amino acid residue, some exhibit dual kinase activity, i.e. they can phosphorylate two different kinds of amino acids.

7.5.2 Background of collaboration – Data

The laboratory of Dr. Anne Ephrussi at the European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany, asked us to investigate the possibility of using combinatorial pattern discovery to help them identify potential phosphorylation sites on experimentally predicted *in vitro* substrates, or targets, of the protein kinase Par-1 in *Drosophila melanogaster* (Tomancak, Piano *et al.* 2000). The physiological relevance of these targets could not be determined by the screen that identified them (mobility shifts on SDS-PAGE). The assumption was that the possible phosphorylation sites would be the same *in vivo* and *in vitro*, regardless of physiological relevance. This remains to be proved, but it seems to be the case for known targets of Par-1. The predicted phosphorylation sites would then be selectively mutated seeking experimental confirmation.

In personal communications with Dr. Anne Ephrussi's laboratory, there were a few key points that would form part of our approach.

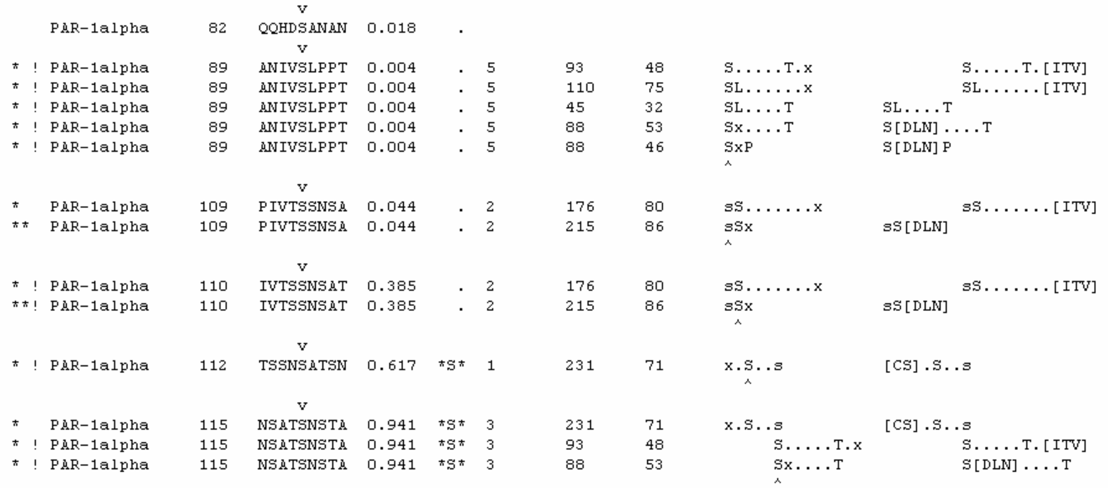
Firstly, it was not clear how many target sites were phosphorylated. Ideally one universal Par-1-phosphorylated site was to be found, but there was no data to back this expectation up, e.g. Bazooka (Benton and St Johnston 2003) had two phosphorylation sites, whereas for Oskar (Riechmann, Gutierrez *et al.* 2002) and LKB1 (Martin and St Johnston 2003) (two other known Par-1 targets) the number was not known. Par-1 was also autophosphorylated, but the sites were not known. These proteins would form part of our target set.

In the case where a 14-3-3 protein (Dougherty and Morrison 2004) binding site existed on target proteins, that seemed to be the phosphorylation site (Benton, Palacios *et al.* 2002); however, a lack of 14-3-3 binding sites was not proof of absence of phosphorylation sites. Also, mammalian Par-1 targets, such as Tau (Nishimura, Yang *et al.* 2004), were likely to have a KxGS motif (where 'x' any amino acid), the serine of which is phosphorylated by Par-1 (Biernat, Wu *et al.* 2002), although in *Drosophila* this motif was not considered as abundant. There was one more target known for Par-1 which had not been unambiguously proven to be a direct target, the Dishevelled protein (Sun, Lu *et al.* 2001); the area of it that was phosphorylated had been mapped down to 35 amino acids containing two 14-3-3 binding sites, although it was not known which of the ten S/T residues in those 35 amino acids were the actual phosphorylation sites. This information would be used as a post-pattern-discovery filter to add value to the patterns and their occurrences on the targets.

7.5.3 Methods and Results

A set of 135 potential and four experimentally verified Par-1 targets were the input of TEIRESIAS. The potential targets represented approximately 2% of the proteins collection of the *Drosophila* Genomics Resource Center, which itself represented 40% (approximately 6,000 proteins) of the *Drosophila melanogaster* proteome. There was a significant probability that many targets were missed by the assay performed at the Ephrussi laboratory, either because the proteins were too large, or because they were not correctly folded, or because the mobility shift was not large enough to be visible on the SDS-PAGE. TEIRESIAS was launched with very sensitive parameter settings, including allowing for amino acid equivalences. The discovered patterns were computationally filtered for inclusion of a S/T amino acid and participation of three experimentally verified phosphorylation sites (which were also part of 14-3-3 binding motifs): S457 on Exuperantia (Riechmann and Ephrussi 2004) and S151 and S1085 on Bazooka. In numbers: 377,310 patterns were discovered; after the filtering described above, 54 patterns remained; after manual checking for pattern overlaps, 35 patterns remained.

The same 139 targets were run through the NetPhos phosphorylation site prediction server (Blom, Sicheritz-Ponten *et al.* 2004). The final set of 35 patterns were overlaid on the results of the server, and a comprehensive output with evidence from both methods was generated (**Figure 19**).



```
                             v
    PAR-1alpha      82   QQHDSANAN  0.018     .
                             v
* ! PAR-1alpha      89   ANIVSLPPT  0.004     .  5       93      48      S.....T.x                S.....T.[ITV]
* ! PAR-1alpha      89   ANIVSLPPT  0.004     .  5       110     75      SL......x                SL......[ITV]
* ! PAR-1alpha      89   ANIVSLPPT  0.004     .  5       45      32      SL....T         SL....T
* ! PAR-1alpha      89   ANIVSLPPT  0.004     .  5       88      53      Sx....T         S[DLN]....T
* ! PAR-1alpha      89   ANIVSLPPT  0.004     .  5       88      46      SxP             S[DLN]P
                                                                         ^
                             v
*   PAR-1alpha     109   PIVTSSNSA  0.044     .  2       176     80      sS.......x               sS.......[ITV]
**  PAR-1alpha     109   PIVTSSNSA  0.044     .  2       215     86      sSx             sS[DLN]
                                                                         ^
                             v
* ! PAR-1alpha     110   IVTSSNSAT  0.385     .  2       176     80      sS.......x               sS.......[ITV]
**! PAR-1alpha     110   IVTSSNSAT  0.385     .  2       215     86      sSx             sS[DLN]
                                                                          ^
                             v
* ! PAR-1alpha     112   TSSNSATSN  0.617  *S*   1       231     71      x.S..s          [CS].S..s
                                                                           ^
                             v
*   PAR-1alpha     115   NSATSNSTA  0.941  *S*   3       231     71      x.S..s          [CS].S..s
* ! PAR-1alpha     115   NSATSNSTA  0.941  *S*   3       93      48           S.....T.x          S.....T.[ITV]
* ! PAR-1alpha     115   NSATSNSTA  0.941  *S*   3       88      53           Sx....T            S[DLN]....T
                                                                              ^
```

Figure 19. A sample of overlaying the 35 patterns on the results of the NetPhos phosphorylation site prediction server for the 139 Par-1 targets.

<u>Columns (space separated)</u>

target name

offset of candidate phosphorylation site on target

context of candidate phosphorylation site on target

prediction score [0-1]

prediction or not (using a cut-off)

number of patterns containing the candidate phosphorylation site

number of pattern instances in the 139 targets (can be more than 1 per target)

number of targets containing the pattern (up to 139)

the pattern itself

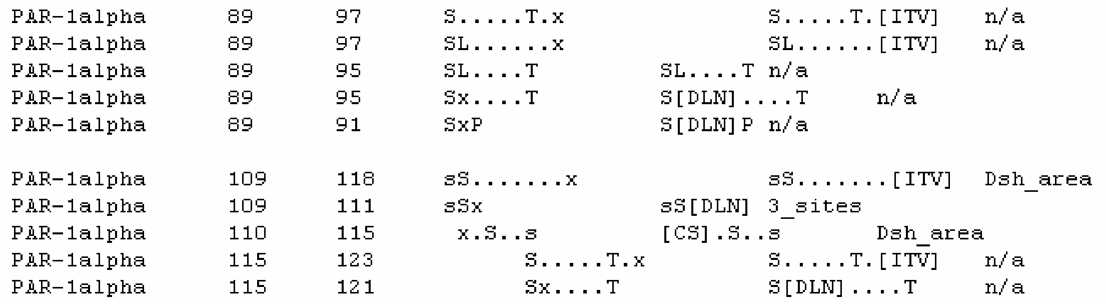
<u>Symbols</u>

'*', a pattern (or more) covering that position

'', the pattern describes all three experimentally verified sites**

'!', the candidate phosphorylation site and the corresponding amino-acid on the pattern overlap, i.e. the amino acid on the pattern was actually the experimentally described phosphorylation site.

The final step of the procedure was to check whether the patterns overlapped with the 14-3-3 binding site motifs, the KxGS motif, and the area on the Dishevelled protein (**Figure 20**).



```
PAR-1alpha      89      97      S.....T.x               S.....T.[ITV]   n/a
PAR-1alpha      89      97      SL......x               SL......[ITV]   n/a
PAR-1alpha      89      95      SL....T         SL....T n/a
PAR-1alpha      89      95      Sx....T         S[DLN]....T     n/a
PAR-1alpha      89      91      SxP             S[DLN]P n/a

PAR-1alpha      109     118     sS.......x              sS.......[ITV]  Dsh_area
PAR-1alpha      109     111     sSx             sS[DLN] 3_sites
PAR-1alpha      110     115      x.S..s         [CS].S..s       Dsh_area
PAR-1alpha      115     123           S.....T.x         S.....T.[ITV]   n/a
PAR-1alpha      115     121           Sx....T           S[DLN]....T     n/a
```

Figure 20. Sample of the 35 patterns overlap with the 14-3-3 binding site motifs, the KxGS motif, and the area on the Dishevelled protein.

7.5.4 Discussion

Since all 35 patterns were required to include the three experimentally-defined phosphorylation sites, they would necessarily overlap, but they would not all cover each of the three phosphorylation sites; therefore one should not expect or require all of them to appear on a target (for example, one of the experimentally-defined sites was covered by only six patterns). In fact, although the patterns did overlap, as a complete set they would cover a large number of positions around the phosphorylation site, therefore targets with many patterns shared would also share considerable amino acid identity/similarity around the phosphorylation site, which was obviously not a requirement – in relevant literature and databases, kinase binding sites in very different proteins were not defined by more than two to four amino acids.

Our approach has probably led to an intentional over-prediction, since our aim was to aid the expert eyes of our collaborators in identifying highly probable phosphorylation sites on the predicted Par-1 targets. Therefore, instead of calculating probabilities for our predictions or adopting any other scoring function that could potentially hide part of our evidence, we opted for a phosphorylation site prediction prioritization scheme based on the following (in suggested order and based on **Figure 20**):

- whether the pattern(s) on the predicted site was (/were) discovered on experimentally-defined phosphorylation sites;
- number of patterns per predicted site (up to 35; however 35 was not to be expected – see above);
- number of targets (out of 139) each pattern was discovered in;
- co-occurrence of 14-3-3 binding motifs;
- co-occurrence of the KxGS motif;
- whether the patterns are in the phosphorylated area of the Dishevelled protein;
- overlap with NetPhos predictions.

Our work is currently being used to assist Dr. Anne Ephrussi's research, and definitive results are not yet available.

7.6 Functional network patterns

7.6.1 Introduction

Network motifs were defined recently as patterns of interconnections occurring in complex networks at numbers that are significantly higher than those in randomized networks (Shen-Orr, Milo *et al*. 2002). They can be used to uncover the design principles of these networks, and can be considered as basic building blocks, or, in biological terms, as structural and functional modules. The same authors applied their algorithms to the one of the characterized gene regulation networks, that of direct transcriptional interactions in *Escherichia coli*. They excitingly found that much of that network was composed of three highly significant motifs, which led to an easily interpretable view of the network. Their approach involved detection of subgraphs with a user-defined number of nodes through the use of the corresponding matrix, but computational limitations restricted this user-defined number to be rather low.

Istvan Albert and Reka Albert (Albert and Albert 2004) used conserved network motifs to predict protein-protein interactions with relative success. They claimed that the method's success was based on the presence of conserved interaction motifs within the network.

7.6.2 Methods

We attempted to tackle this issue using combinatorial pattern discovery. We strongly believed that an appropriate (possibly problem-specific) representation of the network for combinatorial pattern discovery with TEIRESIAS, and a similarly knowledgeable approach in analyzing the output could yield a powerful method. Another feature we aimed for was the unlimited size and complexity of the patterns.

However, it quickly became obvious that there were a number of issues that needed to be addressed; an obvious and major limitation is that TEIRESIAS is by design only able to discover linear patterns in character strings – networks are much more complex structures, therefore there would be a need for a considerable post-discovery manipulation of the linear patterns to create higher-structure network patterns. This led to the conclusion that we had to devise a consistent way to comprehensively decompose a network to linear segments for use with TEIRESIAS. Consequently, it was decided that only directed graphs could be used efficiently, and that the sources (nodes with only outgoing edges) and sinks (nodes with only incoming edges) of the network would be used as starting and finishing points of the linear segments respectively. Additionally, it was decided that the method would massively benefit from labelling the nodes of the network (e.g. with a functional classification scheme), therefore requiring the patterns to describe both structure and e.g. function.

In summary, the methodology involved the following steps:

- obtaining a directed cyclic/acyclic graph;
- labelling the nodes of the graph with a user-defined classification scheme of low granularity (e.g. the 15 functional classes of GeneQuiz);
- starting from all graph sources, following and exporting all the possible linear paths to all graph sinks;

- gathering all the linear paths and creating the input for TEIRESIAS;
- running TEIRESIAS to discover linear patterns;
- creating higher-structure patterns (termed "super-patterns") using linear patterns as building blocks;
- using an abstract representation of graph structural elements to represent the super-patterns while ignoring node labels and the corresponding classification scheme, thus only describing structure, leading to clustering together of structurally-similar super-patterns.

7.6.3 Data

During methodological development, we used the *Escherichia coli* regulatory and metabolic networks; the former was provided from Dr. Ildenfonso Cases, the latter from Dr. Anton Enright, both former members of the Computational Genomics Group. The structural characteristics of the regulatory network, i.e. a small number of highly connected nodes and short average path length, were better suited to our approach; the metabolic network on the other hand would generate a huge amount of linear paths, making the process computationally inefficient. We decided to proceed with the *E. coli* regulatory network.

7.6.4 Results

A working example: two linear patterns discovered by TEIRESIAS (G = regulatory functions; C = cell envelope; W = unknown):

(1) GC

(2) WGC

Three super-patterns in three columns built using the above patterns (**Figure 21**):

(1) crp fadL
(1) fadR fadL
(1) ompR fadL
(2) ihfB ompR fadL
(2) ihfA ompR fadL

(1) lrp ompC
(1) envY ompC
(1) ompR ompC
(2) ihfB ompR ompC
(2) ihfA ompR ompC

(1) lrp ompF
(1) envY ompF
(1) ompR ompF
(2) ihfB ompR ompF
(2) ihfA ompR ompF

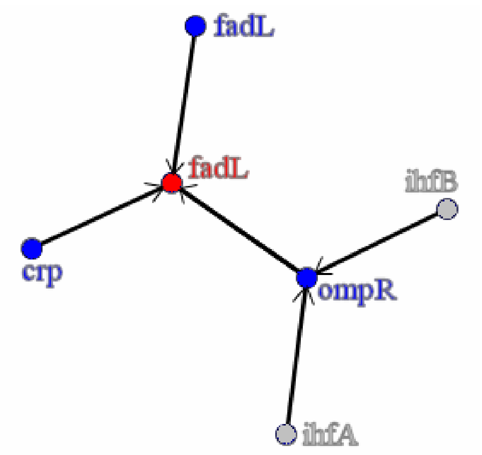


Figure 21. Section of the *Escherichia coli* gene regulatory network on which the super-patterns presented in the text were discovered (blue = regulatory functions; red = cell envelope; grey = unknown).

The most common super-pattern by a large margin was the 1-on-2 structure (**Figure 22a**); the feed-forward loop also featured numerous times (**Figure 22b**):

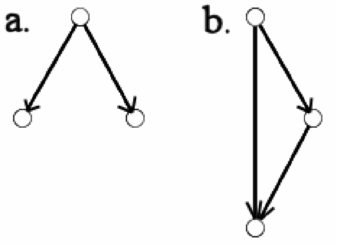


Figure 22. Graphical representation of the 1-on-2 pattern (a), and the feed-forward loop (b).

7.6.5 Conclusion

This project has been a major challenge methodologically, and in many respects rewarding. Future uses and development directions are considered.

7.7 Functional class prediction

7.7.1 Introduction

Arguably, the single most important and exciting goal of Genomics (laboratory-based and computational alike) is the assignment of function to proteins. It is the prerequisite for the understanding of biological systems of any scale and therefore the field of Systems Biology.

Sequence analysis and the concept of homology have been the dominant way of transferring function from a studied protein to a new unknown one (Cai and Doig 2004) (King, Karwath *et al.* 2000). However, they are not the only ones (Shin, Tsuda *et al.* 2004). We have already seen the importance of phylogenetic profiles in **Section 7.3**. Another is gene neighbourhoods, a concept based on the assumption that functionally related genes are likely to be closer to each other than do unrelated ones.

This high degree of clustering of genes belonging to the same functional class makes the classification using conserved gene clusters biologically meaningful. The genes in one class may be conserved under natural selection (Tamames J. *et al.*, 1997). Furthermore, early studies showed that conserved gene order could be correlated with physical interactions between the encoded proteins (Dandekar T. *et al.*, 1998). Therefore, conservation of order of genes could be used as a tool for predicting both protein-protein interaction and protein function. For example, if a certain gene cluster was identified in one or more genomes and a new protein sequence in another organism was located within a gene cluster similar to the previous cluster, the new protein could be considered as functionally coupled to the others in the second cluster, and more specifically be assigned the function of the corresponding protein of the first cluster. Whether this approach can be used to assign functions to the uncharacterized genes has already been tested with positive results (Overbeek *et al.*, 1999).

7.7.2 Data and Methods

Based on our resources and expertise on comparative genomics, we decided to use CoGenT, GeneQuiz, and combinatorial pattern discovery with TEIRESIAS to develop a system for functional class prediction based on conserved gene clusters.

GeneQuiz is an expert annotation system designed to functionally characterize protein sequences. It uses an internal functional classification scheme – we used this scheme and a single character code for our purposes:

A	Amino acid biosynthesis
S	Biosynthesis of cofactors, prosthetic groups, and carriers
C	Cell envelope
D	Cellular processes
I	Central intermediary metabolism
E	Energy metabolism
F	Fatty acid and phospholipid metabolism
Q	Other categories
N	Purines, pyrimidines, nucleosides, and nucleotides
G	Regulatory functions
R	Replication
T	Transcription
L	Translation
M	Transport and binding proteins
W	Unknown

For the CoGenT genomes analysed by GeneQuiz, we substituted each whole protein sequence with its functional class, creating a FASTA-like file with each genome being a single entry:

> genome X
ANAwwMESwwDCDDMMwMGQwLwNLDQwIAN […]

The organisms used in the preliminary analysis are shown in **Appendix C** along with information about their functionally unclassified proteins, i.e. proteins assigned functional class W. The remaining functional classes appeared in the genomes as illustrated in **Figure 23** (grey bars).

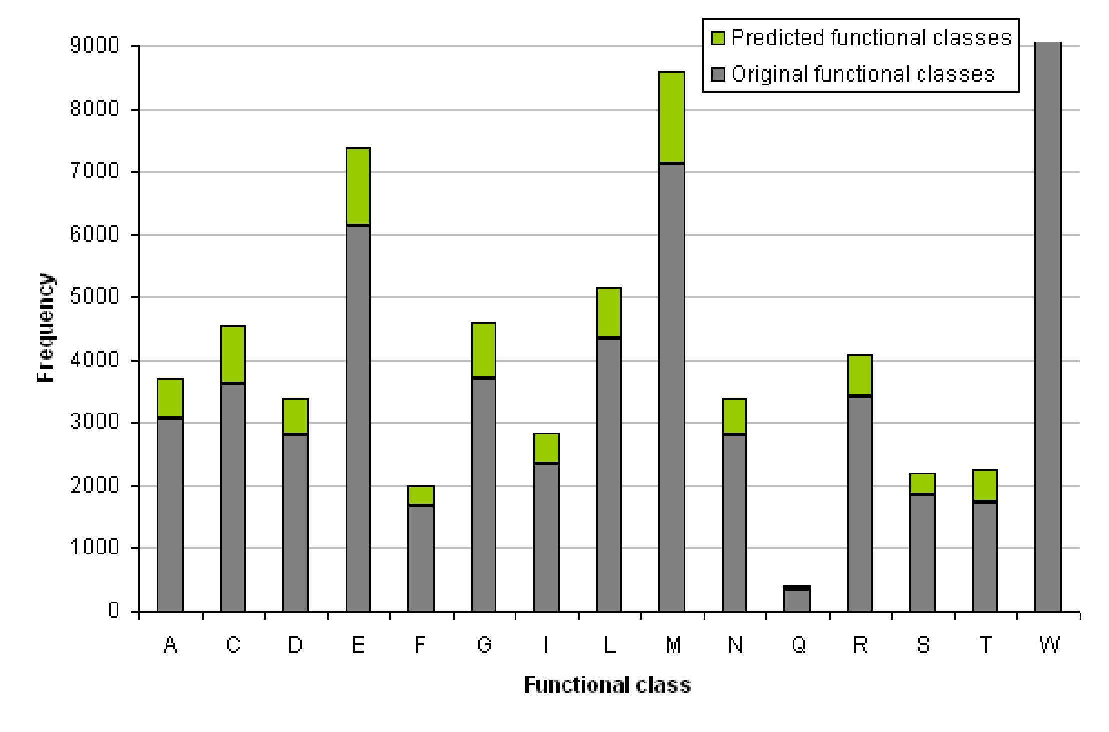


Figure 23. The distribution of GeneQuiz functional classes in the genomes used in this experiment, original (grey bars) and predicted (green bars).

TEIRESIAS was limited to consider uppercase characters only. By converting the 'Unknown'/'W' functional class of GeneQuiz to lower-case ('w' – see example above) we chose to ignore it during pattern discovery, since capturing this class in patterns would be uninformative as well as computationally expensive. The minimum number L of literals in a reported pattern was set to 3, the maximum distance W between any L consecutive literals was set to 5, and the minimum allowed support K for all reported patterns was set to 2.

Only patterns with wildcards would be considered, the idea being that a wildcard position would ideally recover an 'Unknown' class and a significant number of instances of another class, e.g. 'Amino acid biosynthesis', therefore allowing us to assign the latter class to the former. A scoring function for each assignment was devised, that took into account the ratio of known and unknown class instances in each pattern position, and penalised if the pattern recovered more than one known classes per position:

score = (S/C)*S - (T/ℓ(pattern)) - (M*2/ ℓ (pattern)) where:

S : number of known functional classes supporting the prediction

C : number of pattern instances

T : number of times a pattern position has two known functional classes

M : number of times a pattern position has more than two known functional classes

ℓ : length function

7.7.3 Results

Overall 14,290 patterns were discovered, of which 12,582 contained wildcards and therefore further considered.

Below we provide a working example of functional class assignment:
the pattern **C.C.A.M** was discovered in four different organisms:
in ***Escherichia coli*** **O157:H7** as **CMCNAMM** starting at protein with offset **124**
in ***Agrobacterium tumefaciens*** as **CwCwAwM** starting at protein with offset **999**
in ***Salmonella enterica*** as **CMCNAMM** starting at protein with offset **168**
in ***Escherichia coli*** **K12** as **CMCNAMM** starting at protein with offset **120**

For each of the seven pattern positions (0 to 6), we gathered functional information:
0: **C**
1: **MW**
2: **C**
3: **NW**
4: **A**
5: **MW**
6: **M**

Thus:

C.C.A.M → **C[MW]C[NW]A[MW]M**

This information was used to assign:

- the functional class M with a score of 9.00 to the protein of *Agrobacterium tumefaciens* at offset 1000 with CoGenT ID ATUM-C58-01-1-1004 and original ID AGR_C_1879 previously classified as Unknown;
- the functional class N with a score of 9.00 to the protein of *Agrobacterium tumefaciens* at offset 1002 with CoGenT ID ATUM-C58-01-1-1006 and original ID AGR_C_1881 previously classified as Unknown and annotated as putative transglycosylase;
- the functional class M with a score of 9.00 to the protein of *Agrobacterium tumefaciens* at offset 1004 with CoGenT ID ATUM-C58-01-1-1008 and original ID AGR_C_1885 previously classified as Unknown and annotated as SSRA-binding protein.

Predictions were made for 6,811 proteins. The majority of proteins were assigned a single functional class (**Figure 24**). The distribution of predicted functional classes predicted are given in **Figure 23** (green bars).

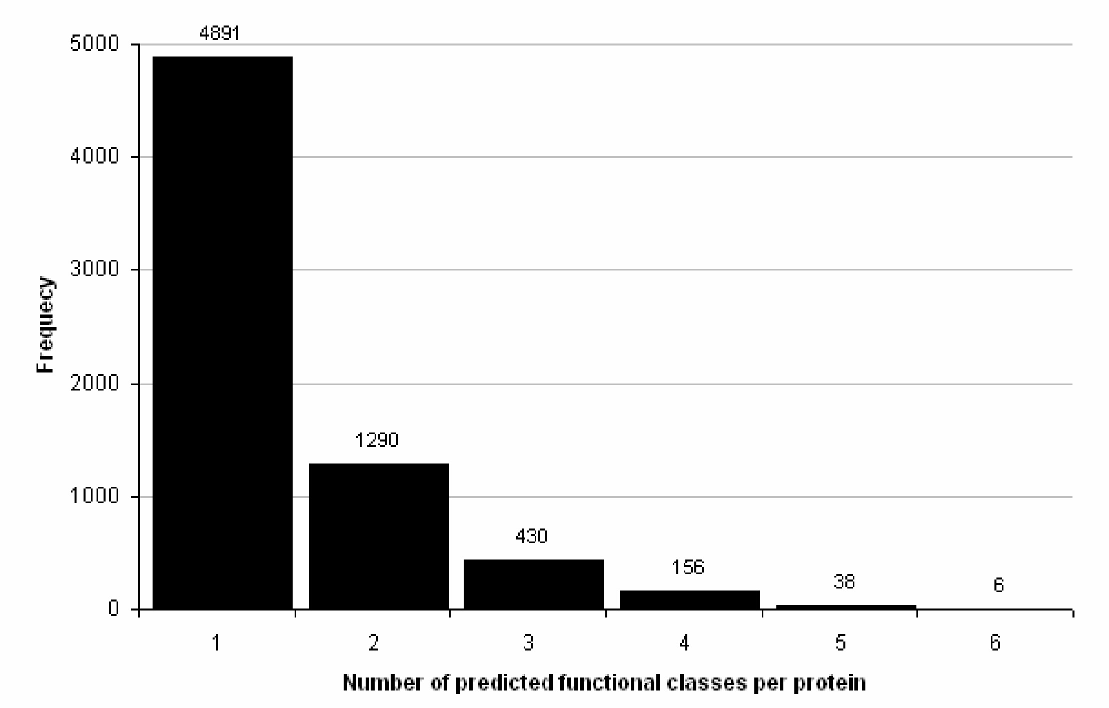


Figure 24. Number of proteins assigned one to six GeneQuiz functional classes.

For proteins for which multiple functional classes were predicted, there were 443 different functional class combinations, e.g. G/M with 56 appearances. Most notably, E and M were not only the most frequent featured and predicted functional classes, but they also only combined with each other in cases of multiple functional class prediction (56 times).

7.7.4 Discussion

Clearly, the fact that the GeneQuiz functionally annotated content of the organisms used in this experiment was significantly low was a drawback. The ‘Unknown’ class did not contribute to the discovery of patterns (since it is invisible to TEIRESIAS) and long stretches of it did not only hamper pattern discovery but also limited pattern support used for prediction. It is envisaged that an iterative procedure where predictions would be incorporated in the input of TEIRESIAS to increase the functionally annotated content and aid in the next round of predictions could be a solution to this problem.

7.8 Disclaimer

I was the sole contributor to this work, with acknowledgements going to the Computational Genomics Group for invaluable discussions, and to Dr. Anne Ephrussi and Dr. Piyi Papadaki of the Ephrussi Group at the European Molecular Biology Laboratory in Heidelberg.

Chapter 8

Sensitive Detection of Sequence Similarity Using Combinatorial Pattern Discovery: a Challenging Study of Two Distantly Related Protein Families

This **Chapter** presents a well-defined control experiment for the detection of the blue-copper binding domain across distantly related protein families.

8.1 Abstract

We investigate the performance of combinatorial pattern discovery to detect remote sequence similarities in terms of both biological accuracy and computational efficiency for a pair of distantly related families, as a case study. The two families represent the cupredoxins and multicopper oxidases, both containing blue copper-binding domains. These families present a challenging case due to low sequence similarity, different local structure and variable sequence conservation at their copper-binding active sites. In this study, we investigate a new approach for automatically identifying weak sequence similarities that is based on combinatorial pattern discovery. We compare its performance with a traditional, HMM-based scheme and obtain estimates for sensitivity and specificity of the two approaches. Our analysis suggests that pattern discovery methods can be substantially more sensitive in detecting remote protein relationships while at the same time guaranteeing high specificity.

8.2 Introduction

Sensitive detection of sequence similarity is a key element of protein structure and function analysis in computational and experimental biology. Through sensitive detection methods, it is possible to identify remote sequence relationships, linked through divergent evolution, and unite protein families into superfamilies with a common origin (Holm 1998). Examples of such detection studies abound in the literature and are systematically deposited in databases of protein families and common motifs (Andreeva, Howorth *et al.* 2004; Bateman, Coin *et al.* 2004). Thus, the sensitive detection of sequence similarity is a fundamental problem (Murzin and Bateman 1997) and multiple solutions have been developed to this end.

The use of multiple sequence alignments as a more sensitive approach of detection was first proposed in the form of sequence profiles (Gribskov, McLachlan *et al.* 1987), followed by a multitude of methods, including BLOCKS (Henikoff, Pietrokovski *et al.* 1998), MAST (Bailey and Gribskov 1998) and hidden Markov models (HMMs) (Eddy 1998; Karplus, Barrett *et al.* 1998). Various tests were devised, usually relying on structural similarity, to resolve specificity (precision) versus sensitivity (recall) of sequence searches. Another approach is sequence threading (Bowie, Luthy *et al.* 1991; Ouzounis, Sander *et al.* 1993), which allows the detection of structural similarities in the absence of evident sequence homology, e.g. between colicins and globins (Holm and Sander 1993), or tenascin and immunoglobulin (Leahy, Hendrickson *et al.* 1992). Yet, at the structural level, with no apparent sequence similarity, the issue of divergence versus convergence raised a generation ago (Zuckerkandl and Pauling 1965) is still with us, despite attempts to delineate homology on the basis of structural similarity (Holm, Ouzounis *et al.* 1992; Holm and Sander 1996). In fact, conflicting arguments are often made to resolve this issue (Murzin 1998).

The argument that sensitive detection of sequence similarity can be utilized to reason about a common structure and thus a common origin of functional properties was established by classical studies of protein sequence/structure relationships (Chothia and Lesk 1986). This early work demonstrated that sequence similarity (presumably reflecting adaptation to a functional role) drifts faster than structure similarity. It should be noted that measures of function similarity are not readily quantifiable, although simple measures based on the Enzyme Commission (EC) (des Jardins, Karp *et al.* 1997; Pawlowski, Jaroszewski *et al.* 2000; Shah and Hunter 1997) and the Gene Ontology (GO) (Jensen, Gupta *et al.* 2003; Lord, Stevens *et al.* 2003) classification schemes have been used to correlate sequence with function.

8.2.1 Multiple sequence alignment

In order to assess divergent relationships between distantly related protein families, a number of approaches have been developed over the past few decades. Most of these relied (explicitly or implicitly) on multiple sequence alignment, whereas other methods leveraged the detection of conserved regions discovered during database searches.

Early methods sought pairwise similarities between proteins. Among such algorithms, the Smith-Waterman dynamic programming alignment algorithm (Smith and Waterman 1981) is arguably among the most accurate. Subsequently, heuristic algorithms were introduced which struck a compromise between improved search performance and achieved sensitivity, with BLAST (Altschul, Gish *et al.* 1990) and FASTA (Pearson 1990) being among the more notable efforts. Further increases in accuracy were achieved by collecting aggregate statistics from a set of aligned sequences and comparing them to a single, unlabeled protein of interest (i.e. in single-sequence mode database searching). Sequence profiles (Gribskov, McLachlan *et al.* 1987) and hidden Markov models (HMMs) (Baldi, Chauvin *et al.* 1994) mentioned above are two such methods for representing these aggregate alignment statistics.

Increases in both sensitivity and specificity can be achieved by leveraging the information in large databases of unlabeled protein sequences. Methods such as PSI-BLAST (Altschul, Madden *et al.* 1997) improve upon profile-based methods by iteratively collecting homologous sequences from the target database and incorporating the resulting statistics into a multiple-sequence alignment model. Furthermore, because this task requires discriminating between related (positive set) and unrelated sequences (negative set), explicitly modelling the difference between these two sets can lead to very powerful methods (Jaakkola, Diekhans *et al.* 1999; Jaakkola, Diekhans *et al.* 2000).

A wealth of information has been published on how the various methods compare to each other: Probabilistic Smith-Waterman (PSW – which is based on HMMs for a single sequence using a standard scoring matrix) (Agarwal and States 1998) and WU-BLAST2 (which uses sum statistics for gapped alignments) (Altschul and Gish 1996) against Smith-Waterman SSEARCH (Smith and Waterman 1981), FASTA and BLASTP (Agarwal and States 1998); BLAST, WU-BLAST2, FASTA and SSEARCH (Brenner, Chothia *et al.* 1998); the SAM-T98 implementation of a HMM procedure, WU-BLASTP and DOUBLE-BLAST (a two-step method similar to ISS – see below, using BLAST instead of FASTA) (Karplus, Barrett *et al.* 1998); SAM-T98, PSI-BLAST and the intermediate sequence search (ISS) procedure (Park, Karplus *et al.* 1998); HMMER, SAM and PSI-BLAST (Madera and Gough 2002); a simple nearest neighbour approach (BLAST), methods based on multiple alignments generated by a statistical profile HMM, and methods – including Support Vector Machines – that transform protein sequences into fixed-length feature vectors (Karchin, Karplus *et al.* 2002).

All methods mentioned above rely on multiple sequence alignment, and one needs to be generated that involves all the sequences in the input. In the case where the sequences are distantly related to one another, generating the necessary alignment can prove to be a very challenging task that frequently necessitates manual intervention. It is worth pointing out here that one can always build a multiple sequence alignment out of *unrelated* sequences. The alignment will of course be of poor quality, but the implication is that if at this stage one makes use of a contaminated set of sequences, the resulting alignment and the performance of all downstream steps will likely be adversely affected. Thus, it is no surprise that the generation of accurate of multiple sequence alignments has developed into a separate subject (Raghava, Searle *et al.* 2003).

To partly alleviate the problems that are associated with methods based on an alignment, pattern discovery methodologies have been proposed in the literature in recent years. These methodologies rely on algorithms that determine patterns that correspond to conserved regions of related sequences. One great example in this category is the BLOCKS database (Pietrokovski, Henikoff *et al.* 1996). These computed 'blocks' can be used to generate profiles (Bucher and Bairoch 1994), HMMs (Krogh, Brown *et al.* 1994), or to derive regular expressions describing the sequence regions of interest. These ideas also formed the intellectual underpinnings for the PRINTS (Attwood, Bradley *et al.* 2003) and PROSITE (Hulo, Sigrist *et al.* 2004) databases. Other applications of pattern discovery include sensitive multiple sequence alignment (Parida, Floratos *et al.* 1998; Smith and Smith 1992), protein annotation (Rigoutsos, Floratos *et al.* 1999; Rigoutsos, Huynh *et al.* 2002), gene discovery (Shibuya and Rigoutsos 2002) or gene expression analysis (Rigoutsos, Floratos *et al.* 2000).

8.2.2 Combinatorial Pattern Discovery

TEIRESIAS (Rigoutsos and Floratos 1998) is an unsupervised pattern discovery algorithm developed to overcome some of the above-mentioned problems while obviating the requirement to align the input sequences. The algorithm is combinatorial in nature, and reports all maximal patterns that are present in a given *unaligned* sequence dataset that have a minimum, user-specified support. Importantly, the algorithm does not need to enumerate the underlying search space in order to report the patterns-answers. TEIRESIAS is output-sensitive, i.e. its running time is quasi-linear to the size of the generated output, and can handle patterns of arbitrary length (Rigoutsos and Floratos 1998).

Using TEIRESIAS, the entire sequence space that was known at the time was explored and a dictionary of all sequence patterns with two or more instances was thus obtained (Rigoutsos, Floratos *et al.* 1999). The entries of this dictionary covered 98.12% of all amino acid positions in the input database and effectively provided a comprehensive and finite set of descriptors for protein sequence space. Thus the patterns derived in this manner could be effectively used to describe virtually every naturally occurring protein. Indeed, one can think of these patterns as building blocks of protein molecules that are a necessary (but not sufficient) condition for function or family equivalence memberships. These patterns either define conserved family signatures or, more interestingly, cut across protein families capturing previously undetected sequence signals. This last premise is consistent with parallel work showing that motifs are often able to detect meaningful sequence similarity across divergent sequences (Nevill-Manning, Wu *et al*. 1998).

Given the highly desirable properties of the TEIRESIAS algorithm, the present study seeks to examine and evaluate the specificity and sensitivity of a pattern-discovery based approach to the problem of detecting members of a given protein family. Using as our vehicle a well-defined experiment that focuses on two distinct protein families with hard-to-detect sequence similarities (but known structural similarities), we set out to answer two questions. First, we want to know how sensitive pattern-based approaches are when compared to hidden Markov model (HMM) schemes. The latter have traditionally been considered to be among the most sensitive search methods. And second, we wish to determine whether sequence-based patterns have the power to bridge the apparent gap between distantly-related families.

8.3 A case study of two distantly related protein families

As noted in earlier work, a test has to be devised which is independent from the method itself (Spang, Rehmsmeier *et al*. 2002). This aims to avoid a circular argument in the evaluation procedure. In that regard, the SCOP database (Murzin, Brenner *et al*. 1995) where proteins have been manually classified to reflect structural, functional, and evolutionary relationships, provides an ideal testing ground for our purposes.

We decided to focus on two protein *families*, namely the cupredoxins and multicopper oxidases. According to the SCOP classification, these two families share the same *fold* and belong to the same *superfamily* but exhibit very low sequence similarity (approximately 10%). The Pfam database (Bateman, Birney *et al*. 2002) contains two distinct entries for these two families.

This superfamily arguably represents one of the most difficult cases for methods that depend on an accurate multiple sequence alignment: in fact, there are only a few conserved sequence patterns around the respective active sites that are indicative of divergent evolution. This extremely challenging, yet well-studied case, was first delineated on the basis of the tertiary structure (Ouzounis and Sander 1991).

8.3.1 The cupredoxin family

The first family is the plastocyanin/azurin family of copper-binding proteins (cupredoxins), also known as ‘blue’ or 'type-1' copper proteins. These small proteins bind a single copper atom and their spectroscopic properties account for their name (De Rienzo, Gabdoulline *et al*. 2000; De Rienzo, Gabdoulline *et al*. 2004). The best known family members include plant chloroplastic plastocyanins and the distantly related bacterial azurins. This family also includes amicyanin from bacteria such as *Methylobacterium extorquens* or *Thiobacillus versutus*, auracyanins A and B from *Chloroflexus aurantiacus*, the blue copper protein from *Alcaligenes faecalis*, cupredoxin (CPC) from cucumber peelings, cusacyanin (basic blue protein; plantacyanin, CBP) from cucumber, halocyanin from *Natrobacterium pharaonis* (a membrane-associated cupredoxin), pseudoazurin from *Pseudomonas* sp., rusticyanin from *Thiobacillus ferrooxidans*, stellacyanin from the Japanese lacquer tree, umecyanin from horseradish roots, and allergen Ra3 from ragweed (this pollen protein seems to have lost the ability to bind copper).

Blue copper proteins are relatively small (10-15 kDa) stable polypeptides built out of beta-strands that are arranged in two beta-sheets, forming a beta-sandwich (**Figure 25**). The alpha-helical content is low and is occasionally restricted to a single alpha-helix and turn-like structures in the loops connecting the beta-strands (Chothia and Lesk 1982). They contain a single copper atom located at the interface of the hydrophobic core between the two beta-sheets (**Figure 25**). The copper ion resides almost in the plane of three strong ligands, two histidines and a cysteine. A fourth axial ligand is provided by a methionine that weakly binds to the copper ion (Canters and Gilardi 1993).

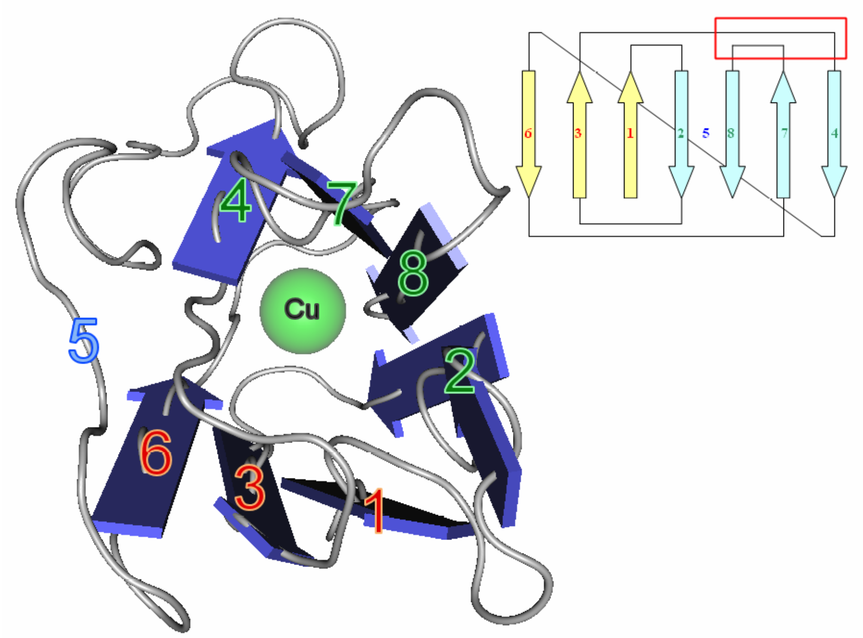


Figure 25. Three-dimensional structure of 1PLC (plastocyanin - PF00127) and a topology diagram. The numbers on both representations indicate the sequential arrangement of beta-strands in three dimensions. The structure is viewed from above the Type I copper-binding site (red box on the topology diagram). The Type I copper-binding site is formed by four ligands - one residue (H) from beta-strand 4 and three residues (C, H, and M) from beta-strands 7 and 8, not shown. The three colours for the strand numbers signify the membership of each beta-strand in one of the two beta-sheets (except beta-strand 5 which belongs to neither). The two beta-sheets of the topology diagram are in yellow and cyan, forming a beta-sandwich in the structure.

The copper ligands are highly conserved, but located at variable distances along the sequences, as the proteins differ substantially in length. The first ligand, always a histidine (His), lies just before a beta-strand, away in sequence from the other three ligands. The other ligands, cysteine (Cys), histidine (His), and methionine (Met), are located within a loop between the last two beta-strands in all known structures. All strands adjacent to the copper-binding site belong to a single beta-sheet (**Figure 25**). The interaction of the residues in the antiparallel beta-strands 4 and 7 appears to be important for the construction of the copper site (Ouzounis and Sander 1991).

8.3.2 The multicopper oxidase family

The second family is that of multicopper oxidases. These enzymes possess three spectroscopically different copper centres: type 1 (or blue), type 2 (or normal) and type 3 (or coupled binuclear). Typical family members include laccase, an enzyme found in fungi and plants, L-ascorbate oxidase, a higher plant enzyme, and ceruloplasmin (ferroxidase), a protein found in the serum of mammals and birds (Ryden and Hunt 1993). On the basis of sequence similarities, the family includes copper resistance protein A (copA) from a plasmid in *Pseudomonas syringae*, human blood coagulation factor V (Fa V) and VIII (Fa VIII), the yeast FET3 and hypothetical protein YFL041w, as well as SpAC1F7.08, the fission yeast homolog of the latter.

Multicopper oxidases were suspected to contain cupredoxin domains on the basis of sequence similarities (Messerschmidt and Huber 1990), later confirmed by the structure determination of ascorbate oxidase from zucchini (Messerschmidt, Ladenstein *et al.* 1992). Although the size/length range among the members of this family varies significantly, multicopper oxidases invariantly contain a type-1 copper-binding site located at the C-terminal domain with a topology similar to that of the cupredoxin family. In ascorbate oxidase, there are three tandem domains with this topology that contribute other ligands to the type-2 and type-3 copper-binding sites (Messerschmidt, Ladenstein *et al.* 1992). More complex structures, domain insertions and extensive length variation are observed in other members of this family (Ryden and Hunt 1993).

8.4 Methods

Our methodology consists of the following steps: (i), data collection and definition of the gold-standard sets; (ii), pattern discovery in the learning sets; (iii), pattern matching and validation to assess specificity and sensitivity; (iv), generation of various HMM profiles; (v), validation of HMM searches; (vi) assessment of specificity and finally (vii) visualization. These seven steps are reflected by the corresponding headings in this section.

8.4.1 Data collection

The SCOP database (Release 1.61 – November 2002) was used to identify the relationship between the two protein families; they are both classified in the same *Class* ("All beta proteins"), *Fold* ("Cupredoxin-like: *sandwich; 7 strands in 2 sheets, greek-key*"), *Superfamily* ("Cupredoxins: *contain copper-binding site*"), and named Plastocyanin/azurin-like (cupredoxins) – with 8 protein types and 31 sequences in total – and Multidomain cupredoxins (multicopper oxidases) – with 5 protein types and 11 sequences in total.

Information from SCOP was used to determine the accession numbers of the families in the Pfam database (Release 7.7 – October 2002). Pfam PF00127 and PF00394 point to "Copper binding proteins, plastocyanin/azurin family" and "Multicopper oxidases", respectively. PF00127 has 31 seed sequences and PF00394 has 127 seed sequences, corresponding to individual domains from the multicopper oxidases; the source of the seed sequences is the corresponding entries from the PROSITE database (Hulo, Sigrist *et al*. 2004).

To enrich the two datasets, BLASTP (Altschul, Gish *et al.* 1990) (version 2.1.3) was used to search the SwissProt database (Boeckmann, Bairoch *et al.* 2003) (Release 40.31 – October 2002), with an E-value cut-off of 1.0e-10; the members of the two families were used as queries in turn. This step was meant to augment the families by including closely related sequences that were absent from the original Pfam alignments, and not to validate the performance of the BLAST algorithm. Note that PSI-BLAST was not used for this step.

Prior to inclusion in the respective families, the search results were examined and verified, whereas sequences with erroneous annotation were excluded. Subsequently, the datasets comprised 73 sequence members for PF00127 and 53 for PF00394. Further examination of the SwissProt annotation records (see below) indicated that PF00127 has 80 potential members (including one plastocyanin fragment) and that PF00394 has 53; these two groups of 80 and 53 sequences are considered to be the gold-standard datasets for the two families. Consequently, the maximum expected recall level for both methods used in this study is 80 and 53 respectively. As a final step, all compositionally biased regions of the two datasets were masked using the CAST algorithm (Promponas, Enright *et al.* 2000) with default settings.

8.4.2 Pattern Discovery – TEIRESIAS

TEIRESIAS requires a set of parameters that express the features of discovered patterns in terms of specified characters, maximum length of elementary patterns, minimum support (frequency), maximum number of amino acid equivalences and various other formatting options (Rigoutsos and Floratos 1998).

TEIRESIAS was used with the following parameter choices: the minimum number *L* of literals in a reported pattern was set to 4, the maximum distance *W* between any *L* consecutive literals was set to 8, and the minimum allowed support *K* for all reported patterns was set to 2 (Rigoutsos and Floratos 1998). During the pattern discovery step, the following amino acid equivalences were permitted: [AG], [DE], [FY], [KR], [ILMV], [QN], [ST], [DN], [EQ]. These equivalences signify the conservative replacement of chemically equivalent amino acids by one another (Taylor 1986). We allowed up to two equivalences per pattern, expressed by the parameter n that was set to 2. The TEIRESIAS software package includes a utility that allows the calculation of log-probability estimates for the discovered patterns that are based on a second order Markov chain (Rigoutsos, Floratos *et al.* 1999). For both PF00127 and PF00394, we allowed only for patterns whose log-probability estimates were in the range [-27, -24], effectively filtering out all patterns that were either too sensitive (high values of log-probability, i.e. common) or too specific (low values of log-probability, i.e. rare). As we were interested in capturing only inter-sequence information, any patterns that appeared more than once in a given sequence were also discarded from our collection (referred to as ‘xpNc’ in **Table 9**).

Table 9. Tabulated summary of results for each Pfam family examined (a. cupredoxins represented by PF00127, b. multicopper oxidases represented by PF00394)

Column Description

1 family identifier in Pfam

2 family members grouped by annotation

(in parentheses, phylogenetic origin: B=Bacteria, Vp=Viridiplantae, A=Archaea, F=Fungi, M=Metazoa)

3 number of proteins in SwissProt (40.31) for each annotation group

4 number of proteins in dataset for each annotation group

5 HMMer PF00127 hits in SwissProt (a. True+, b. cross-hits to the other family)

6 TEIRESIAS PF00127 hits in SwissProt (a. True+, b. cross-hits to the other family)

7 HMMer PF00394 hits in SwissProt (a. cross-hits for the other family, b. True+)

8 TEIRESIAS PF00394 hits in SwissProt (a. cross-hits for the other family, b. True+)

9 HMMer False+

10 TEIRESIAS False+

11 composite precision & recall for HMMer, see formulas below *

12 composite precision & recall for TEIRESIAS, see formulas below *

13 percentage difference in composite precision & recall between TEIRESIAS & HMMer*

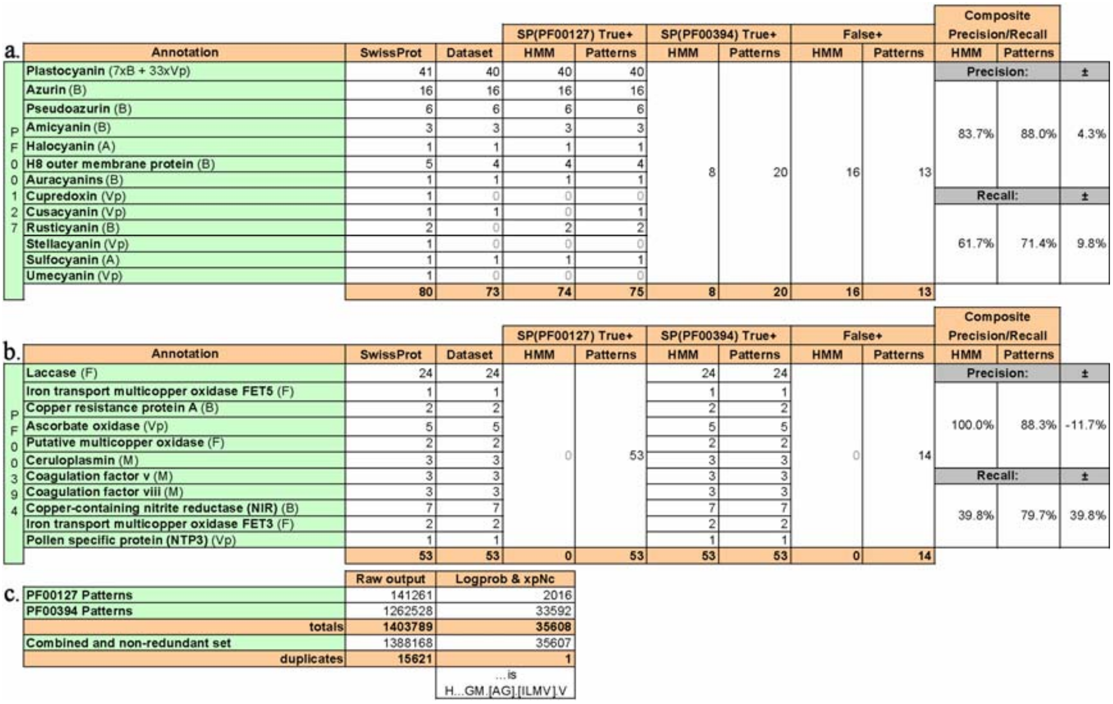
*** Precision of method in family i =**

(True+i + True+j) / (True+i + True+j + False+i)

*** Recall of method in family i =**

(True+i + True+j) / (True+i + True+j + False-i)

c. Number of patterns discovered in the two families, and the effect of filtering in their reduction. 'Original output' is the raw TEIRESIAS output; 'Logprob & xpNc' refers to the effect of pattern score and occurrence filtering (see Methods). The row representing the combined, non-redundant pattern set corresponds to the intersection of the distinct sets from each family. Note that after filtering, the only common pattern between the two families is the cross-hitting pattern H...GM.[AG].[ILMV].V (see Text for details).



a.

				SP(PF00127) True+		SP(PF00394) True+		False+		Composite Precision/Recall		
	Annotation	SwissProt	Dataset	HMM	Patterns	HMM	Patterns	HMM	Patterns	HMM	Patterns	
PF00127	Plastocyanin (7xB + 33xVp)	41	40	40	40	8	20	16	13	Precision:		±
	Azurin (B)	16	16	16	16					83.7%	88.0%	4.3%
	Pseudoazurin (B)	6	6	6	6							
	Amicyanin (B)	3	3	3	3							
	Halocyanin (A)	1	1	1	1							
	H8 outer membrane protein (B)	5	4	4	4							
	Auracyanins (B)	1	1	1	1							
	Cupredoxin (Vp)	1	0	0	0					Recall:		±
	Cusacyanin (Vp)	1	1	0	1					61.7%	71.4%	9.8%
	Rusticyanin (B)	2	0	2	2							
	Stellacyanin (Vp)	1	0	0	0							
	Sulfocyanin (A)	1	1	1	1							
	Umecyanin (Vp)	1	0	0	0							
		80	73	74	75	8	20	16	13			

b.

				SP(PF00127) True+		SP(PF00394) True+		False+		Composite Precision/Recall		
	Annotation	SwissProt	Dataset	HMM	Patterns	HMM	Patterns	HMM	Patterns	HMM	Patterns	
PF00394	Laccase (F)	24	24	0	53	24	24	0	14	Precision:		±
	Iron transport multicopper oxidase FET5 (F)	1	1			1	1			100.0%	88.3%	-11.7%
	Copper resistance protein A (B)	2	2			2	2					
	Ascorbate oxidase (Vp)	5	5			5	5					
	Putative multicopper oxidase (F)	2	2			2	2					
	Ceruloplasmin (M)	3	3			3	3					
	Coagulation factor v (M)	3	3			3	3			Recall:		±
	Coagulation factor viii (M)	3	3			3	3			39.8%	79.7%	39.8%
	Copper-containing nitrite reductase (NIR) (B)	7	7			7	7					
	Iron transport multicopper oxidase FET3 (F)	2	2			2	2					
	Pollen specific protein (NTP3) (Vp)	1	1			1	1					
		53	53	0	53	53	53	0	14			

c.

	Raw output	Logprob & xpNc
PF00127 Patterns	141261	2016
PF00394 Patterns	1262528	33592
totals	1403789	35608
Combined and non-redundant set	1388168	35607
duplicates	15621	1
		...is H...GM.[AG].[ILMV].V

8.4.3 Validation of patterns

With the two collections of patterns describing the two families under consideration, we turned our attention to the identification of instances of these patterns in a candidate sequence. For this step, we decided to use a widely and freely available tool, NR-grep that is, on the average, a significantly faster alternative to GNU-grep (Navarro 2001).

Each pattern was used to retrieve SwissProt sequences, whose annotation records were examined. The two families and the gold-standard sets were not excluded from SwissProt. If these records matched the annotations from the gold-standard sets, previously classified into annotation groups (e.g. plastocyanin, etc.), then the recovered SwissProt entry was considered as a true positive (True+). The recovery rate for both gold-standard sets was used to assess the rate of recall for all patterns. Labels in none of the above categories were arranged into an “other/hypothetical” category and were considered as false positives (False+), in order to assess accuracy. Those patterns that only recovered members of the gold-standard sets were collected and made non-redundant through clustering that examined their location along the input sequences (also called offset lists). We also carried out the clustering process on those patterns of one family that recovered members of the other (henceforth called cross-hitting patterns).

The keywords that dictated the formation of the annotation groups for the gold-standard sets were as follows: “laccase, iron transport multicopper oxidase, copper resistance protein a, ascorbate oxidase, putative multicopper oxidase, ceruloplasmin, coagulation factor v, coagulation factor viii, copper-containing nitrite reductase, fet3, ntp3, plastocyanin, azurin, amicyanin, halocyanin, h.8, auracyanin, cupredoxin, cusacyanin, rusticyanin, stellacyanin, sulfocyanin, umecyanin, major outer membrane protein, multicopper oxidase, blue copper, copper resistance, copa, copper-containing, copper-binding, copper tolerance, blue-copper-, nitrite reductase, copper oxidase”. This augmented dataset based on annotation allows the possibility of detecting further genuine members without relying on BLAST and thus validate the two methods independently.

8.4.4 Hidden Markov models – HMMer

Starting with the seed datasets provided by Pfam, we augmented the multiple alignment by incorporating the sequences that were retrieved using BLASTP (see ‘Data collection’ above). We explored several alternatives at this point: (i) automatic alignment using ClustalW (Thompson, Higgins *et al*. 1994); (ii) manual editing of the ClustalW results; (iii) direct use of the provided seed alignment; and (iv) the automatic alignment of the seed Pfam alignment with the retrieved sequences using BLASTP, using hmmalign. The first three alternative options were deemed not applicable either because of poor performance of the multiple alignment procedure, or incomplete datasets. Thus, for our analysis, we adopted option (iv) above.[10]

For each of the two augmented input sets, we used the newly constructed alignment to build an HMM (hmmbuild), calibrate it (hmmcalibrate) and search the SwissProt database (hmmsearch) for instances of the corresponding family using an E-value cut-off for hmmsearch of 100. All three hmm* utilities are part of the HMMER package (Eddy 1998) – [see also: S.R. Eddy, 2001, HMMER: Profile hidden Markov models for biological sequence analysis, http://hmmer.wustl.edu/, version 2.1.1].

8.4.5 Validation of HMM searches

Those sequences that did not belong to the annotation groups mentioned above, and had an equal or lower E-value than the one from the lowest-ranking, known True+ hit (Jaakkola, Diekhans *et al.* 1999) were considered to be false positives. This adaptive cut-off strategy ensures that the resulting search maximises recall (by recovering all known True+ hits) at the expense of precision (by including certain False+ hits).

8.4.6 Assessment of specificity

To further evaluate the False+ entries of both pattern and HMM searches, we used BLASTP with each False+ entry as a query against the SwissProt database. The E-value cut-off was set at 100 and all hits were checked for potential membership in the Pfam families or any other positive indication of relatedness to the gold-standard sets. Additionally, we checked whether the local similarity regions highlighted by BLAST had a corresponding pattern (analysis not shown). These procedures were used to ensure that all False+ hits were indeed unrelated to the two families under consideration and do not result in an evident detection of sequence similarity.

[10] Clearly, we ensured that the HMM-based search procedure was directly comparable with the pattern-discovery-based one in that they used the same input dataset, the latter first having been augmented with the help of BLASTP as described earlier.

8.4.7 Data visualization

To generate and manipulate multiple sequence alignments, we used ClustalW (Thompson, Higgins *et al.* 1994) (version 1.82) and Seaview for Windows (Galtier, Gouy *et al.* 1996). For structural studies we used Cn3D (Wang, Geer *et al.* 2000) (version 4.1) with a selection of structure files from the Molecular Modelling Database (MMDB) (Chen, Anderson *et al.* 2003).

8.4.8 Data availability

To ensure reproducibility, all our results and links to relevant resources are available at the following URL: http://www.ebi.ac.uk/research/cgg/patterns/type_I/

8.5 Results

8.5.1 PF00127

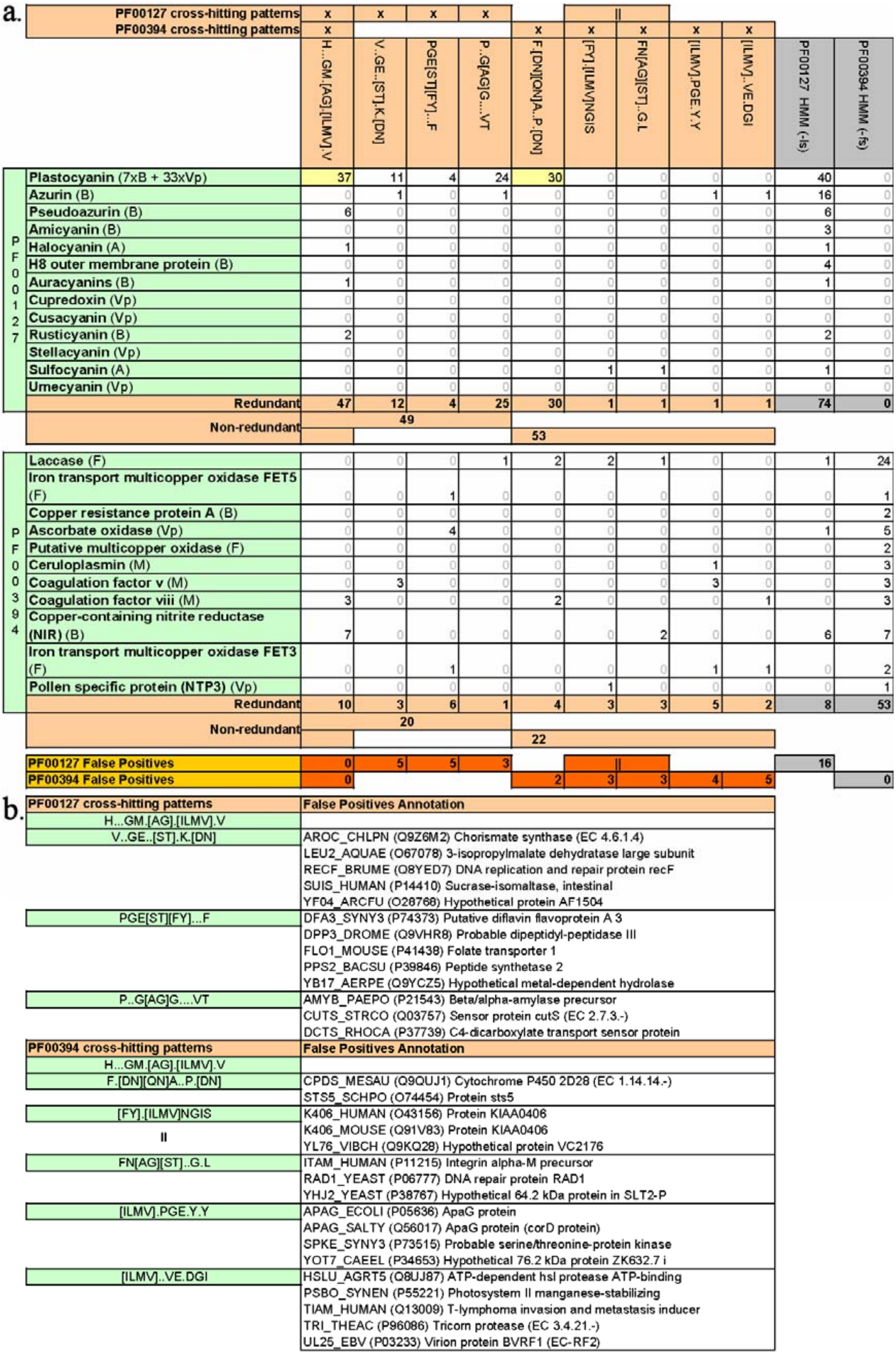
Out of the 80 of the PF00127-related gold-standard set, the BLAST-expanded dataset contained 73 entries. Seven sequences, namely 1 plastocyanin fragment, 1 H8 outer membrane protein, 1 'cupredoxin', 2 rusticyanins, 1 stellacyanin, and 1 umecyanin were not detected by BLAST and the selected parameter settings.

Pattern Discovery

Using TEIRESIAS, we discovered 141,261 patterns in ~2min CPU time in the 73 sequences of PF00127. Log-probability-based filtering and exclusion of patterns with multiple instances in a given sequence reduced this set to 2016 patterns. These 2016 patterns were checked against SwissProt, and the results are summarised in **Table 9a** and illustrated in **Figure 26a**. The patterns recovered all PF00127 family members. In particular, 160 patterns managed this without any error (False+). An extra 4 patterns succeeded in recovering 20 sequences from the PF00394 family (part of the PF00394 gold-standard set, plus ANIA_NEIGO with accession number Q02219, a major outer membrane protein Pan 1 precursor annotated in Pfam as a positive case), plus the 2 rusticyanins, while generating 13 False+ hits (**Table 9a** and **Table 10**; **Figure 26a**).

Table 10. a. A detailed analysis of the cross-hitting patterns from TEIRESIAS and the two HMMs. Columns 1 and 2 as in Table 9 (a, b). Four cross-hitting patterns were identified in PF00127 and five in PF00394, with one in common, listed in the third column. Columns 4-6 correspond to patterns identified in PF00127 and columns 7-11 to those identified in PF00394. For all columns 3-11, the family identifiers where the patterns were discovered are listed in the first two rows and linked to the corresponding pattern by the symbol 'x'. Columns 8 and 9 list two patterns which are mutually exclusive for family PF00394, marked by the symbol '||'. Columns 12 and 13 list the corresponding hits for each HMM. Additionally, the rows for redundant and non-redundant hits are displayed: redundant hits correspond to the counts in the individual cells, and non-redundant hits correspond to the sum of unique hits within each family and thus are not derivable from the individual cells. These patterns correspond to known structural motifs (see Text for details). The highlighted cells in PF00127 correspond to 37 hits covering all 7 bacterial and 30 plant plastocyanins, while 30 additional hits are plant-specific; their union covers all 40 plastocyanins, indicating the value of cross-hitting patterns to attain high coverage. The last two rows provide the False+ counts for each pattern and for each HMM.

b. The list of putative False+ hits in SwissProt for each of the cross-hitting patterns, sorted by the same column order in a. The corresponding entries as listed with SwissProt identifier (accession number) and description line.



a.

	PF00127 cross-hitting patterns	x	x	x	x		||					
	PF00394 cross-hitting patterns	x				x	x	x	x	x		
		H...GM.[AG].[ILMV].V	V..GE..[ST].K.[DN]	PGE[ST][FY]...F	P..G[AG]G....VT	F.[DN][QN]A..P.[DN]	[FY].[ILMV]NGIS	FN[AG][ST]..G.L	[ILMV].PGE.Y.Y	[ILMV]..VE.DGI	PF00127 HMM (-ls)	PF00394 HMM (-fs)
PF00127	Plastocyanin (7xB + 33xVp)	37	11	4	24	30	0	0	0	0	40	0
	Azurin (B)	0	1	0	1	0	0	0	1	1	16	0
	Pseudoazurin (B)	6	0	0	0	0	0	0	0	0	6	0
	Amicyanin (B)	0	0	0	0	0	0	0	0	0	3	0
	Halocyanin (A)	1	0	0	0	0	0	0	0	0	1	0
	H8 outer membrane protein (B)	0	0	0	0	0	0	0	0	0	4	0
	Auracyanins (B)	1	0	0	0	0	0	0	0	0	1	0
	Cupredoxin (Vp)	0	0	0	0	0	0	0	0	0	0	0
	Cusacyanin (Vp)	0	0	0	0	0	0	0	0	0	0	0
	Rusticyanin (B)	2	0	0	0	0	0	0	0	0	2	0
	Stellacyanin (Vp)	0	0	0	0	0	0	0	0	0	0	0
	Sulfocyanin (A)	0	0	0	0	0	1	1	0	0	1	0
	Umecyanin (Vp)	0	0	0	0	0	0	0	0	0	0	0
	Redundant	47	12	4	25	30	1	1	1	1	74	0
	Non-redundant	49										
						53						
PF00394	Laccase (F)	0	0	0	1	2	2	1	0	0	1	24
	Iron transport multicopper oxidase FET5 (F)	0	0	1	0	0	0	0	0	0	0	1
	Copper resistance protein A (B)	0	0	0	0	0	0	0	0	0	0	2
	Ascorbate oxidase (Vp)	0	0	4	0	0	0	0	0	0	1	5
	Putative multicopper oxidase (F)	0	0	0	0	0	0	0	0	0	0	2
	Ceruloplasmin (M)	0	0	0	0	0	0	0	1	0	0	3
	Coagulation factor v (M)	0	3	0	0	0	0	0	3	0	0	3
	Coagulation factor viii (M)	3	0	0	0	2	0	0	0	1	0	3
	Copper-containing nitrite reductase (NIR) (B)	7	0	0	0	0	0	2	0	0	6	7
	Iron transport multicopper oxidase FET3 (F)	0	0	1	0	0	0	0	1	1	0	2
	Pollen specific protein (NTP3) (Vp)	0	0	0	0	0	1	0	0	0	0	1
	Redundant	10	3	6	1	4	3	3	5	2	8	53
	Non-redundant	20										
						22						
	PF00127 False Positives	0	5	5	3		||				16	
	PF00394 False Positives	0				2	3	3	4	5		0

b.

PF00127 cross-hitting patterns	False Positives Annotation
H...GM.[AG].[ILMV].V	
V..GE..[ST].K.[DN]	AROC_CHLPN (Q9Z6M2) Chorismate synthase (EC 4.6.1.4) LEU2_AQUAE (O67078) 3-isopropylmalate dehydratase large subunit RECF_BRUME (Q8YED7) DNA replication and repair protein recF SUIS_HUMAN (P14410) Sucrase-isomaltase, intestinal YF04_ARCFU (O28768) Hypothetical protein AF1504
PGE[ST][FY]...F	DFA3_SYNY3 (P74373) Putative diflavin flavoprotein A 3 DPP3_DROME (Q9VHR8) Probable dipeptidyl-peptidase III FLO1_MOUSE (P41438) Folate transporter 1 PPS2_BACSU (P39846) Peptide synthetase 2 YB17_AERPE (Q9YCZ5) Hypothetical metal-dependent hydrolase
P..G[AG]G....VT	AMYB_PAEPO (P21543) Beta/alpha-amylase precursor CUTS_STRCO (Q03757) Sensor protein cutS (EC 2.7.3.-) DCTS_RHOCA (P37739) C4-dicarboxylate transport sensor protein
PF00394 cross-hitting patterns	**False Positives Annotation**
H...GM.[AG].[ILMV].V	
F.[DN][QN]A..P.[DN]	CPDS_MESAU (Q9QUJ1) Cytochrome P450 2D28 (EC 1.14.14.-) STS5_SCHPO (O74454) Protein sts5
[FY].[ILMV]NGIS ||	K406_HUMAN (O43156) Protein KIAA0406 K406_MOUSE (Q91V83) Protein KIAA0406 YL76_VIBCH (Q9KQ28) Hypothetical protein VC2176
FN[AG][ST]..G.L	ITAM_HUMAN (P11215) Integrin alpha-M precursor RAD1_YEAST (P06777) DNA repair protein RAD1 YHJ2_YEAST (P38767) Hypothetical 64.2 kDa protein in SLT2-P
[ILMV].PGE.Y.Y	APAG_ECOLI (P05636) ApaG protein APAG_SALTY (Q56017) ApaG protein (corD protein) SPKE_SYNY3 (P73515) Probable serine/threonine-protein kinase YOT7_CAEEL (P34653) Hypothetical 76.2 kDa protein ZK632.7 i
[ILMV]..VE.DGI	HSLU_AGRT5 (Q8UJ87) ATP-dependent hsl protease ATP-binding PSBO_SYNEN (P55221) Photosystem II manganese-stabilizing TIAM_HUMAN (Q13009) T-lymphoma invasion and metastasis inducer TRI_THEAC (P96086) Tricorn protease (EC 3.4.21.-) UL25_EBV (P03233) Virion protein BVRF1 (EC-RF2)

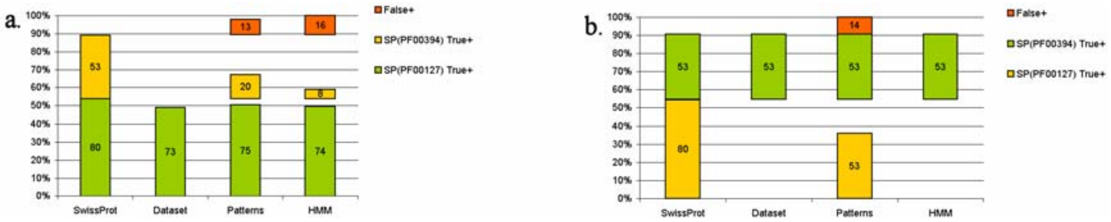


Figure 26. Graphical representation of total counts from columns 3-10 of Table 9 for families PF00127 (a, as in Table 9a) and PF00394 (b, as in Table 9b). The hits corresponding to within-family members are shown as green bars and cross-family hits are shown as yellow bars. False+ hits are shown as orange bars. The scale is up to 100% that includes all hits, including False+ ones.

Hidden Markov model Searching

The SwissProt search with the global HMM provided by augmented multiple alignments for PF00127 took ~25min CPU time and recovered all family members plus the 2 rusticyanins (**Table 9a**). At an E-value cut-off of 54 (which was selected in order to accommodate sulfocyanin, see ‘Validation of HMM searches’), HMMer also recovered 8 PF00394 members (part of the PF00394 gold-standard set, plus ANIA_NEIGO – see above) and generated 16 False+ hits (**Table 9a** and **Table 10**; **Figure 26a**). The Smith/Waterman model did not perform equally well (results not shown).

8.5.2 PF00394

For this family, the BLAST-expanded dataset contained all 53 of the SwissProt sequences that were also appropriately annotated and formed our gold-standard set.

Pattern Discovery

Using TEIRESIAS, we discovered 1,262,528 patterns in ~45min CPU time in the 53 sequences of PF00394. Log-probability-based filtering and exclusion of patterns with multiple instances in a given sequence reduced this set to 33,592 patterns. All family members were recovered with the help of 2,652 patterns without generating any False+ hits. The patterns also recovered the auracyanin and the 2 rusticyanins from PF00127 (not shown). An additional 6 patterns (2 of them recovering sulfocyanin but with different PF00394 hits and the same False+ cost, thus considered as alternative solutions, see **Table 10**) crossed the family boundaries and recovered 53 members from PF00127 while generating only 14 False+ hits (**Table 9b** and **Table 10**; **Figure 26b**).

It is important to point out that the cross-hitting pattern H...GM.[AG].[ILMV].V was discovered and included in both pattern collections, something that emphatically supports the ability to recover the existence of a functional relationship between the two protein families at the sequence level alone.

Hidden Markov model Searching

The local HMM provided by augmented multiple alignments performed better than the global model for this family: it took ~35min CPU time to search through SwissProt and all 53 PF00394 members were recovered. The HMM also managed to recover 8 PF00127 members at a cost of 59 False+ at an *adaptive* E-value cut-off of 78 (**Table 9b** and **Table 10**).

8.5.3 Sensitivity analysis of the two approaches

As **Figure 27** illustrates, combinatorial pattern discovery within the present analysis framework succeeded in being at least as specific and at least as sensitive as HMMer and Pfam HMMs in both experiments. In fact, as it can be seen from the results in the case of the cupredoxin family PF00127 (**Figure 27a**), combinatorial pattern discovery performed significantly better. In the case of the multicopper oxidase family PF00394 results were mixed (**Figure 27b**).

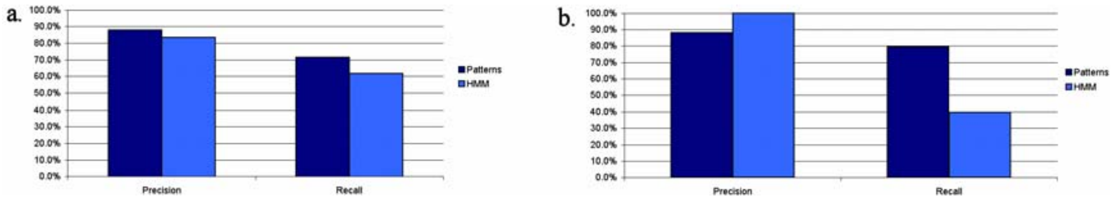


Figure 27. Graphical representation of columns 11-12 of a: Table 9a and b: Table 9b. Dark blue corresponds to TEIRESIAS pattern discovery performance measures and light blue to HMMer, for the two families. It is evident that, with the exception of precision for PF00394, the overall performance of TEIRESIAS is slightly better.

It is also worth noting that the global model for PF00127 outperformed the fragment-tolerant model for the same family. Interestingly, the opposite was true for PF00394. This behaviour most likely results from the fact that PF00394 members are longer, variable and more complex, compared to the single domain case of the PF00127 family.

8.5.4 The importance of parameter settings

Masking compositionally biased regions prior to applying a pattern discovery approach is an advisable step. This masking has a dual benefit of a decrease in the quantity and an increase in the quality of the reported patterns. Even if this step were omitted, the vast majority of patterns based on compositionally biased areas would not pass our tests at the subsequent stages.

Other inherent parameters influence the pattern discovery analysis directly. Most importantly, the ratio L/W (see above for a definition) controls the minimum amount of local homology that a discovered pattern captures. A large value of L will likely result in patterns that are too specific whereas a small L will result in uninformative patterns that fail to capture what is important about the dataset at hand.

Clearly these parameters can affect the quality of the generated results and typically both the biological significance and the validation phase should be considered prior to deciding values for L and W. In that regard, the parameter selection in this study is rather generic and no attempt was made to optimize the values for L and W. In fact, any attempt to optimize would defeat the purpose as it would unnecessarily diminish the general applicability of the proposed framework to other protein families.

Analogously, the logarithmic probability boundaries that we have employed represent rather generic choices as well. The only threshold that matters in this case is the maximum threshold which is currently set to -24. Raising it to larger values is not recommended as doing so would permit the inclusion of patterns that are likely to appear by pure chance. On the other hand, it should be clear that further lowering the minimum log-probability (i.e. to values lower than -27) would not affect neither the sensitivity nor the specificity of the framework; however, it would affect the speed of searching for instances of the resulting patterns simply due to the increased cardinality, i.e. large number of highly specific patterns.

8.5.5 Bridging families with cross-hitting patterns

We now concentrate on the two sets of cross-hitting patterns and examine their potential to detect weak similarities across two distantly related families.

Pattern H...GM.[AG].[ILMV].V is the only common significant pattern between the two sets. Thus, it is the only one surviving the filtering procedure (**Table 9c**), out of 15,621 common patterns deduced from both families. **Figure 28** shows the pattern highlighted on the 3D structures of a plastocyanin (**Figure 28a**) and a copper-containing nitrite reductase (**Figure 28b**) – in both structures, the pattern covers the section of the turn leading to beta-strand 8 of the plastocyanin(-like) domain and the beta-strand itself (Ouzounis and Sander 1991). This region, and especially the turn, is part of the structure of the copper-binding site (**Figure 25**). The only additional hit of the pattern, ANIA_NEIGO, is actually a genuine homolog (see above), very similar to copper-containing nitrite reductases.

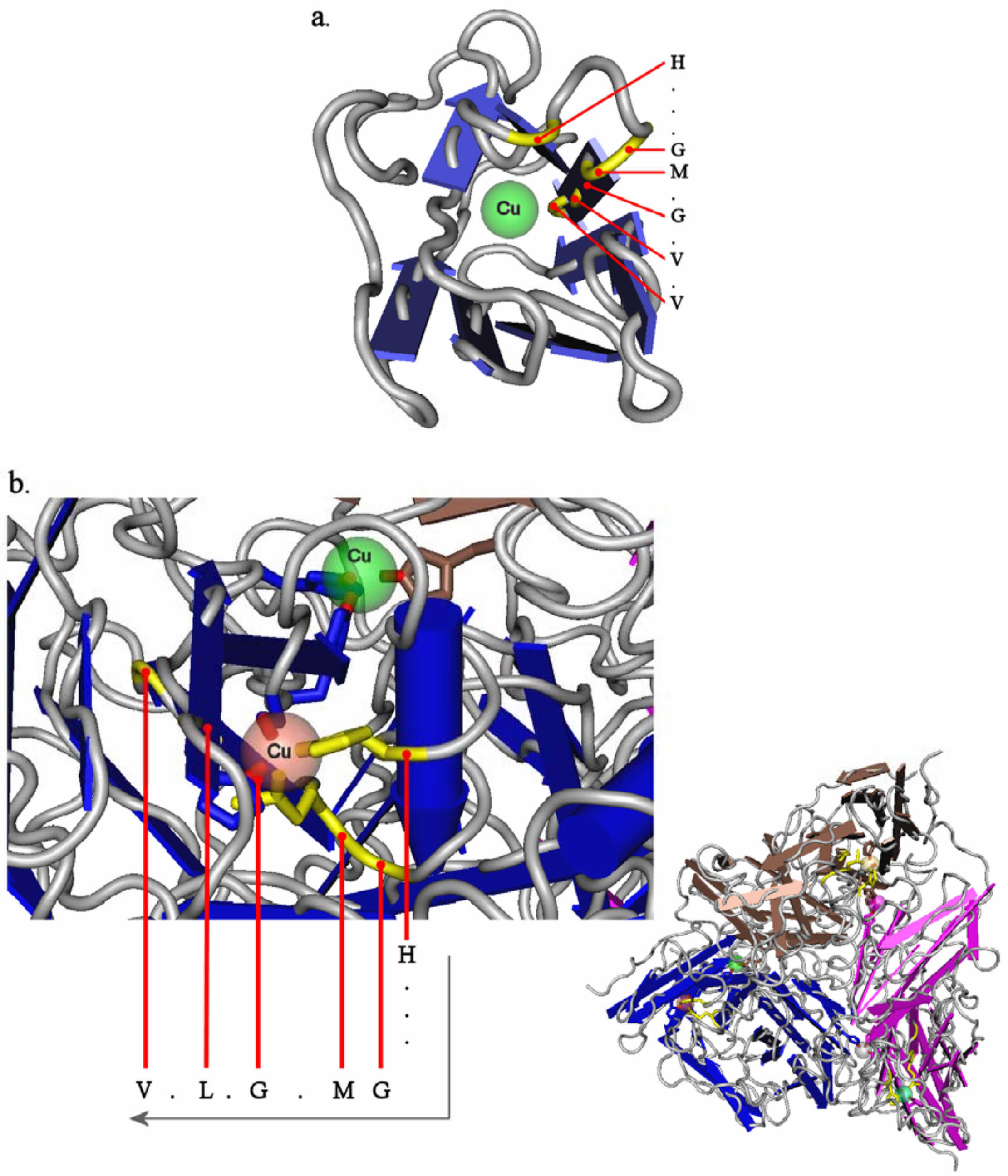


Figure 28. a. View of 1PLC (plastocyanin – PF00127) with the H...GM.[AG].[ILMV].V cross-hitting pattern highlighted in yellow. The layout of the structure is exactly the same as in Figure 1. b. View of 1NDR (nitrite reductase – PF00394) with the H...GM.[AG].[ILMV].V cross-hitting pattern highlighted in yellow. Nitrite reductase consists of a trimer of monomers each containing two cupredoxin domains (bottom right). The zoomed-in and rotated (for clarity) region displays two copper-binding sites, one intra-domain Type I (red Cu atom –accepting electrons from the donor protein azurin) and one inter-domain Type II (green Cu atom – where nitrite reduction takes place).

The consensus pattern for Type-1 copper (blue) proteins from PROSITE (entry PDOC00174; C and H are copper ligands) is aligned with H...GM.[AG].[ILMV].V (**Table 10a**) as follows:

```
[…]C-x(1,2)-[PG]-x(0,1)-H-x(2,4)-[MQ]
                        H . . . G M.[AG].[ILMV].V
```

The same for the consensus pattern for the copper-binding domain of multicopper oxidases (entry PDOC00076; the first two H symbols are the copper Type-3 ligands; the C, the third H, and [LM] are Type-1 ligands) (**Table 10a**):

```
H-C-H-x(3)-H-x(3)-[AG]-[LM]
           H . . . G    M.[AG].[ILMV].V
```

Pattern P..G[AG]G....VT (**Table 10a**) overlaps with the aforementioned pattern:

```
   H... GM.[AG].[ILMV].V
P..G[AG]G...   .    V T
```

and thus describes the same structural motif. Pattern PGE[ST][FY]...F (**Table 10a**) is situated on beta-strand 6 in some plastocyanins and ascorbate oxidases, and V..GE..[ST].K.[DN] (**Table 10a**) at the end of beta-strand 2 and the complete beta-strand 3 of certain plastocyanins (**Figure 25**). Pattern F.[DN][QN]A..P.[DN] (**Table 10a**) is positioned at the end of beta-strand 3 and the loop before beta-strand 4 of certain plastocyanins (not containing the H copper ligand). Another pattern mapped to structural elements is pattern [ILMV].PGE.Y.Y found in some azurins (**Table 10a**) – it occupies the turn before beta-strand 7 and two positions (Y.Y) of the beta-strand itself. The two mutually exclusive patterns (**Table 10a**) are found in the first beta-sheet of different multicopper oxidases as well as sulfocyanin. Finally, pattern [ILMV]..VE.DGI maps to regions found in helix 5 (**Figure 25**) in homologous azurins (but not found in currently known structures).

Thus, the above-mentioned patterns are able to detect members from both families, via the identification of sequence regions that correspond to conserved structural features in this fold. Further examination of the putative false positive sequences containing instances of the above cross-hitting patterns indicated that in fact ANIA_NEIGO is a true positive (as mentioned above). However, no connection could be established between any of the remaining putative false positives and the two families at hand (**Table 10b**). Thus, the cross-hitting patterns are remarkably specific in this case and are not truly able to identify remote homologs, beyond the learning set of known cases.

8.5.6 A brief note on phylogenetic distribution

Using a coarse-grained phylogenetic profiling which partitions sequence space into seven major taxa (Peregrin-Alvarez, Tsoka *et al.* 2003), we classified the cross-hitting patterns according to their phylogenetic profiles and explored their taxon specificity. It is evident that particular groups are located in specific taxonomic categories, and this could be used in the future to filter out unlikely cases of identification in large-scale experiments. More specifically, PF00127 is evenly distributed among Bacteria (40 members, plus 1 fragment) and Viridiplantae (37 members), and also contains 2 Archaeal proteins. PF00394 is mostly found in Fungi (29 members), and less so in Bacteria (9 members), Metazoa (9 members) and Viridiplantae (6 members) (see **Table 9** and **Table 10** for the corresponding taxonomic labels).

From this perspective, an interesting result is the discovery of a True+ from a partition that is not supported by any member of either family – it was located in Protists and annotated as multicopper oxidase (accession number Q9NKK0, L3223.1 from *Leishmania major*, a eukaryote).

8.6 Discussion

With the help of modern sequence analysis methods, we have examined another facet of this challenging superfamily. Through a new approach, we have managed to translate the families' subtle evolutionary relationship to sequence patterns (Ouzounis and Sander 1991). More than a decade ago, the means to perform such an analysis were rather limited; sequence and structural information was sparse, and computational tools were less advanced.

We have thus compared two methods, combinatorial pattern discovery versus one of the most sensitive approaches based on hidden Markov models. Our results suggest that, at least in the case of this controlled experiment, combinatorial pattern discovery represents an important alternative to HMMs, in that it can detect remote sequence similarities without sacrificing specificity. One of the issues that will need to be addressed is the handling of multiple, sometimes redundant, patterns, and their reduction into sets of biologically meaningful structural descriptors. The lack of gold-standard sets for this type of experiments is another limiting factor. This specific study focused on a challenging case and, through a highly detailed validation step, demonstrated the type of issues that need to be addressed. We are in the process of extending this analysis to the entire set of known protein structures in order to assess the general applicability of combinatorial pattern discovery in remote similarity detection.

8.7 Disclaimer

I was the sole contributor to this work. I would like to acknowledge the advice of Dr. Isidore Rigoutsos, manager of the Bioinformatics and Pattern Discovery group at IBM's Computational Biology Center and co-author of the manuscript (Darzentas, Rigoutsos *et al.* 2005).

Chapter 9
Bridging SCOP Folds and Superfamilies with structurally equivalent sequence patterns

This **Chapter** extends the work described in **Chapter 8** across the entire range of the protein fold hierarchy.

9.1 Abstract

Motivation: Structural and functional annotation by association through homology detection is an essential tool in Bioinformatics. We wanted to investigate the limits of the approach by looking at cases of extreme sequence distance. *Methods*: We employed combinatorial pattern discovery to establish evolutionary links between SCOP Superfamilies of the same or different Fold of the same Class. Importantly, we overlaid structural alignment information from the Dali server to validate and strengthen the sequence-based connections. *Results*: We show by example that sequence patterns are an efficient and foremost robust way for the detection of very distant protein relationships. We provide all data for further investigation by experts.

9.2 Background

Detection of homology has been in the forefront of biological sequence analysis for decades (Agarwal and States 1998; de Haen, Swanson *et al.* 1976; Nishikawa and Ooi 1986; Ponomarenko, Bourne *et al.* 2005). The need for robust, efficient, and sensitive tools for this task is more pressing now than ever, considering the flood of sequence information from genome sequencing centres around the world, and its very low coverage with structural and functional information (Janssen, Enright *et al*. 2003). Computational approaches vary significantly from pairwise alignment to Support Vector Machines and neural networks (see **Chapter 6** and **Section 8.2.1**).

For every such method, two major evaluation points are sensitivity and specificity. There is inevitably a trade-off between the two, such that it is very difficult to ensure high recall with high precision, i.e. a large number of true positives with a relatively low number of false positives. Other desirable traits include computational speed, and general and easy applicability.

Research by the authors has demonstrated that combinatorial pattern discovery can successfully handle challenging cases of sensitive detection of evolutionary relationships, like in the case of two distantly related copper-binding protein families where we showed the methodology to be sensitive yet specific, and efficient (Darzentas, Rigoutsos *et al.* 2005)/(**Chapter 8**).

In this **Chapter**, mainly in an attempt to evaluate the robustness of our approach, we are describing our work towards a more systematic study of the capabilities of combinatorial pattern discovery to link (divergent) proteins with no apparent sequence similarity.

For such an undertaking, a primary requirement is a well-defined dataset providing robust learning and testing material. A widely used resource, structured in such a way that the user can obtain data of considerable structural but low sequence similarity is the Structural Classification Of Proteins (SCOP) database (Andreeva, Howorth *et al.* 2004). SCOP aims to provide a description of the structural and evolutionary relationships between all proteins whose structure is known, including all entries in the Protein Data Bank (Deshpande, Addess *et al.* 2005). It has been constructed manually by visual inspection and comparison of structures, but with the assistance of tools to make the task manageable and help provide generality. Proteins are classified to reflect both structural and evolutionary relatedness (http://scop.mrc-lmb.cam.ac.uk/scop/intro.html).

According to SCOP, a Superfamily consists of proteins that have low sequence identities, but whose structural and functional features suggest that a common evolutionary origin is probable. A Fold contains proteins that have the same major secondary structures in the same arrangement and with the same topological connections. Proteins placed together in the same Fold category may not have a common evolutionary origin: the structural similarities could arise just from the physics and chemistry of proteins (http://scop.mrc-lmb.cam.ac.uk/scop/intro.html).

Most of the folds are assigned to one of the five structural classes on the basis of the secondary structures of which they are composed: (a) *all alpha* (for proteins whose structure is essentially formed by alpha-helices), (b) *all beta* (for those whose structure is essentially formed by beta-sheets), (c) *alpha and beta* (for proteins with alpha-helices and beta-strands that are largely interspersed), (d) *alpha plus beta* (for those in which alpha-helices and beta-strands are largely segregated) and (e) *multi-domain* (for those with domains of different fold and for which no homologs are known at present) (Murzin, Brenner *et al*. 1995). *Membrane and cell surface proteins and peptides* (f), and *small proteins* (g) are two more Classes we have included in this work. We believe that this evolutionarily grey area of SCOP is the ideal playground for our purposes.

The question put forward was whether pattern discovery could connect different Superfamilies within the same Fold in a meaningful way on the protein sequence level. The previous paragraph highlights the challenges any such attempt faces: not only Superfamily members have low sequence identities, but also the evolutionary distance between Superfamilies in a Fold can be severe (if not computationally insurmountable).

Importantly, we included another layer of evidence above sequence similarity, namely Dali (http://www.bioinfo.biocenter.helsinki.fi:8080/dali/index.html), a database based on exhaustive all-against-all 3D structure comparison of protein structures currently in the PDB (Holm and Sander 1996). We thus rephrased the problem as the discovery of structurally similar sequence patterns.

Investigators led by Dr. Rob Russell (Lupas, Ponting *et al*. 2001) have already presented evidence hinting at the evolutionary origins of domains by considering the occurrence of structurally similar elements in seemingly different folds. They argued in favour of the evolution of modern single polypeptide domains from ancient short peptide ancestors (antecedent domain segments, or ADSs, as the authors named them). Our work supports their theory through a vastly larger dataset and totally different methodology.

9.3 Materials – Methods

Step 1 - Data Collection and Preparation

We obtained the one-protein-per-species sequence file from the ASTRAL Compendium (Chandonia, Hon *et al.* 2004) of the SCOP database version 1.65, which consisted of 7 Classes, 798 Folds, 1,290 Superfamilies, 2,319 Families, and 5,126 PDB Domains. We decided to use Classes a-g, and the ASTRAL data was divided into six corresponding files, one for each Class.

Each file was constructed in such a way that each Superfamily formed a single pseudo-sequence, thus with the appropriate parameters we could ensure that only inter-Superfamily patterns could be discovered. This was deemed essential towards making the data generation and analysis computationally efficient.

To avoid obtaining results of low biological significance, compositionally biased regions in the data were masked with CAST (Promponas, Enright *et al.* 2000) and with default settings.

Step 2 - Combinatorial Pattern Discovery

TEIRESIAS (Rigoutsos and Floratos 1998) was run on each Class separately. TEIRESIAS parameters used were: amino acids in the pattern (-l), number of overlapping characters in the convolved pattern (-c), maximum length of an elementary pattern (-w), minimum number of appearances of the pattern (-k), maximum number of appearances of the pattern (-q), maximum number of brackets (indicating equivalent amino acids) allowed in the pattern (-n), flag for the support k to be the minimum number of sequences in which a pattern should appear (compared to the minimum number of instances of the pattern) (-v), flag for the algorithm to output the list of positions (offset list) for each discovered pattern (-p), flag for the algorithm to use amino acid equivalences (-b<equivalences file>), and flag for the algorithm to consider only the uppercase characters during pattern discovery (-u). The parameters were set to [l=3] [c=2] [w=6] [k=2] [q=2] [n=2] [-v] [-p] [-b<equivalences>] [-u] for all Classes, except for Class 'c', where [w=4], for purely computational reasons. The equivalences were [CS] [DLN] [EQ] [FHWY] [ITV] [KMR], all based on the structural properties of amino acids.

Step 3 - Preliminary pattern filtering

Only patterns longer than three characters connecting Domains in different Superfamilies of the same Class were considered further.

Step 4 - Determination of Structural Similarity of patterns

To assess whether the patterns were similar in structure as well as sequence, we overlaid the patterns on the structurally aligned sequence regions from Dali. To achieve this, we had to calculate the coordinates of the patterns PDB sequence file used and provided by Dali, and not on the ASTRAL SCOP data.

A pattern was considered further only if it was perfectly aligned in, and either 'completely included in' or 'completely including', the pair of Dali-aligned regions.

Step 5 - Determination of Specificity and Sensitivity of patterns

The complete (multiple-species) file from the ASTRAL Compendium was used to assess whether a pattern was able to represent an adequate number of Domains per Family while remaining Class specific. An average Family coverage of equal or greater than 50% was required for a pattern to be considered further.

9.4 Results

Overall, 23,528,918 patterns (23,268,001 unique) were discovered in Step 2. Steps 3 and 4 reduced this number down to 3,031 patterns participating in 3,109 Domain connections, and Step 5 down to 231 patterns in 240 connections. Out of these 240 connections, 87 were between Domains of different Superfamilies in the same Fold and Class, and 153 were between Domains of different Superfamilies in different Folds in the same Class. The coverage of SCOP from this subset was ~13% at Fold level, going down to ~4% at Domain level; Classes e and g were not represented (**Table 11**). All the connections and patterns can be viewed in **Appendix D**. These same data were used in **Figure 29**, which emphasizes the complexity of the connections made by the 231 patterns at Fold level. The distribution of Z-scores of DALI structural alignments participating in the SCOP connections made by the 231 patterns in the two levels of connections is plotted in **Figure 30**.

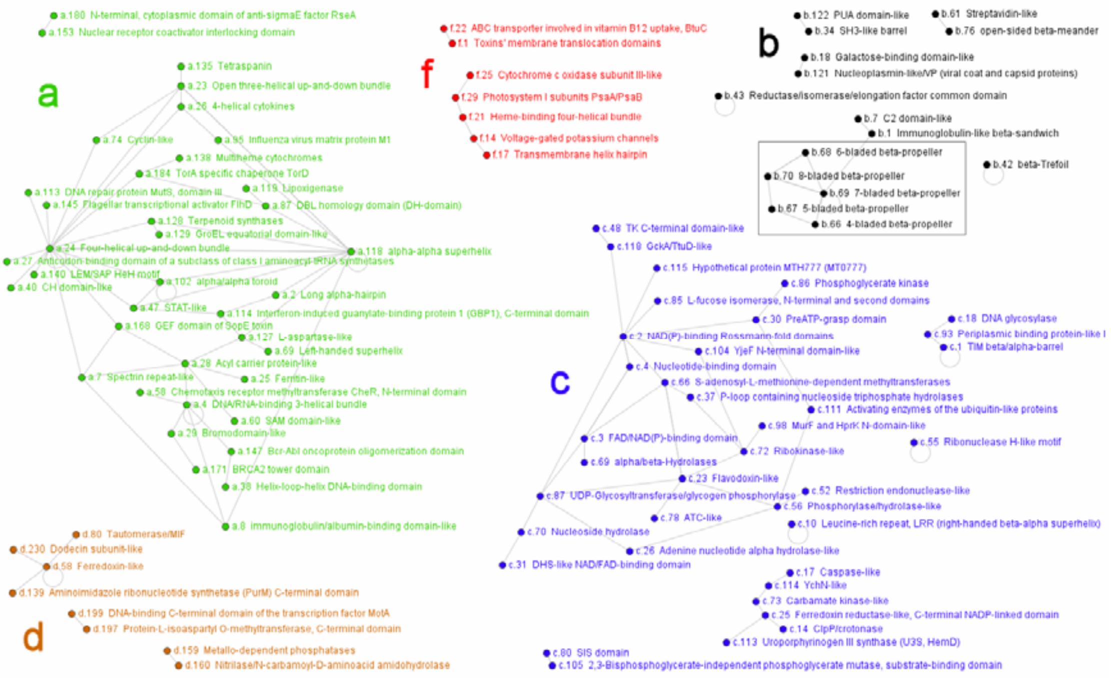

a
a.180 N-terminal, cytoplasmic domain of anti-sigmaE factor RseA
a.153 Nuclear receptor coactivator interlocking domain
a.135 Tetraspanin
a.23 Open three-helical up-and-down bundle
a.26 4-helical cytokines
a.74 Cyclin-like
a.95 Influenza virus matrix protein M1
a.138 Multiheme cytochromes
a.184 TorA specific chaperone TorD
a.113 DNA repair protein MutS, domain III
a.119 Lipoxigenase
a.145 Flagellar transcriptional activator FlhD
a.87 DBL homology domain (DH-domain)
a.128 Terpenoid synthases
a.129 GroEL equatorial domain-like
a.24 Four-helical up-and-down bundle
a.118 alpha-alpha superhelix
a.27 Anticodon-binding domain of a subclass of class I aminoacyl-tRNA synthetases
a.140 LEM/SAP HeH motif
a.102 alpha/alpha toroid
a.40 CH domain-like
a.2 Long alpha-hairpin
a.47 STAT-like
a.114 Interferon-induced guanylate-binding protein 1 (GBP1), C-terminal domain
a.168 GEF domain of SopE toxin
a.127 L-aspartase-like
a.69 Left-handed superhelix
a.28 Acyl carrier protein-like
a.7 Spectrin repeat-like
a.25 Ferritin-like
a.58 Chemotaxis receptor methyltransferase CheR, N-terminal domain
a.4 DNA/RNA-binding 3-helical bundle
a.60 SAM domain-like
a.29 Bromodomain-like
a.147 Bcr-Abl oncoprotein oligomerization domain
a.171 BRCA2 tower domain
a.38 Helix-loop-helix DNA-binding domain
a.8 immunoglobulin/albumin-binding domain-like
f
f.22 ABC transporter involved in vitamin B12 uptake, BtuC
f.1 Toxins' membrane translocation domains
f.25 Cytochrome c oxidase subunit III-like
f.29 Photosystem I subunits PsaA/PsaB
f.21 Heme-binding four-helical bundle
f.14 Voltage-gated potassium channels
f.17 Transmembrane helix hairpin
b
b.122 PUA domain-like
b.61 Streptavidin-like
b.34 SH3-like barrel
b.76 open-sided beta-meander
b.18 Galactose-binding domain-like
b.121 Nucleoplasmin-like/VP (viral coat and capsid proteins)
b.43 Reductase/isomerase/elongation factor common domain
b.7 C2 domain-like
b.1 Immunoglobulin-like beta-sandwich
b.68 6-bladed beta-propeller
b.42 beta-Trefoil
b.70 8-bladed beta-propeller
b.69 7-bladed beta-propeller
b.67 5-bladed beta-propeller
b.66 4-bladed beta-propeller
c
c.48 TK C-terminal domain-like
c.118 GckA/TtuD-like
c.115 Hypothetical protein MTH777 (MT0777)
c.86 Phosphoglycerate kinase
c.85 L-fucose isomerase, N-terminal and second domains
c.30 PreATP-grasp domain
c.18 DNA glycosylase
c.2 NAD(P)-binding Rossmann-fold domains
c.93 Periplasmic binding protein-like I
c.104 YjeF N-terminal domain-like
c.1 TIM beta/alpha-barrel
c.4 Nucleotide-binding domain
c.66 S-adenosyl-L-methionine-dependent methyltransferases
c.37 P-loop containing nucleoside triphosphate hydrolases
c.111 Activating enzymes of the ubiquitin-like proteins
c.98 MurF and HprK N-domain-like
c.3 FAD/NAD(P)-binding domain
c.55 Ribonuclease H-like motif
c.72 Ribokinase-like
c.69 alpha/beta-Hydrolases
c.23 Flavodoxin-like
c.87 UDP-Glycosyltransferase/glycogen phosphorylase
c.52 Restriction endonuclease-like
c.56 Phosphorylase/hydrolase-like
c.78 ATC-like
c.10 Leucine-rich repeat, LRR (right-handed beta-alpha superhelix)
c.70 Nucleoside hydrolase
c.26 Adenine nucleotide alpha hydrolase-like
c.31 DHS-like NAD/FAD-binding domain
c.17 Caspase-like
c.114 YchN-like
c.73 Carbamate kinase-like
c.25 Ferredoxin reductase-like, C-terminal NADP-linked domain
c.14 ClpP/crotonase
c.113 Uroporphyrinogen III synthase (U3S, HemD)
c.80 SIS domain
c.105 2,3-Bisphosphoglycerate-independent phosphoglycerate mutase, substrate-binding domain
d
d.80 Tautomerase/MIF
d.230 Dodecin subunit-like
d.58 Ferredoxin-like
d.139 Aminoimidazole ribonucleotide synthetase (PurM) C-terminal domain
d.199 DNA-binding C-terminal domain of the transcription factor MotA
d.197 Protein-L-isoaspartyl O-methyltransferase, C-terminal domain
d.159 Metallo-dependent phosphatases
d.160 Nitrilase/N-carbamoyl-D-aminoacid amidohydrolase

Figure 29. Visualisation of the SCOP connections made by the 231 patterns and listed in Appendix D, at the Fold level per Class (colour-coded; Green:Class a, Black:Class b, Blue:Class c, Brown:Class d, and Red:Class f). The beta-propellers Folds connected by our patterns and discussed in the text is highlighted by a black box in the Class b network.

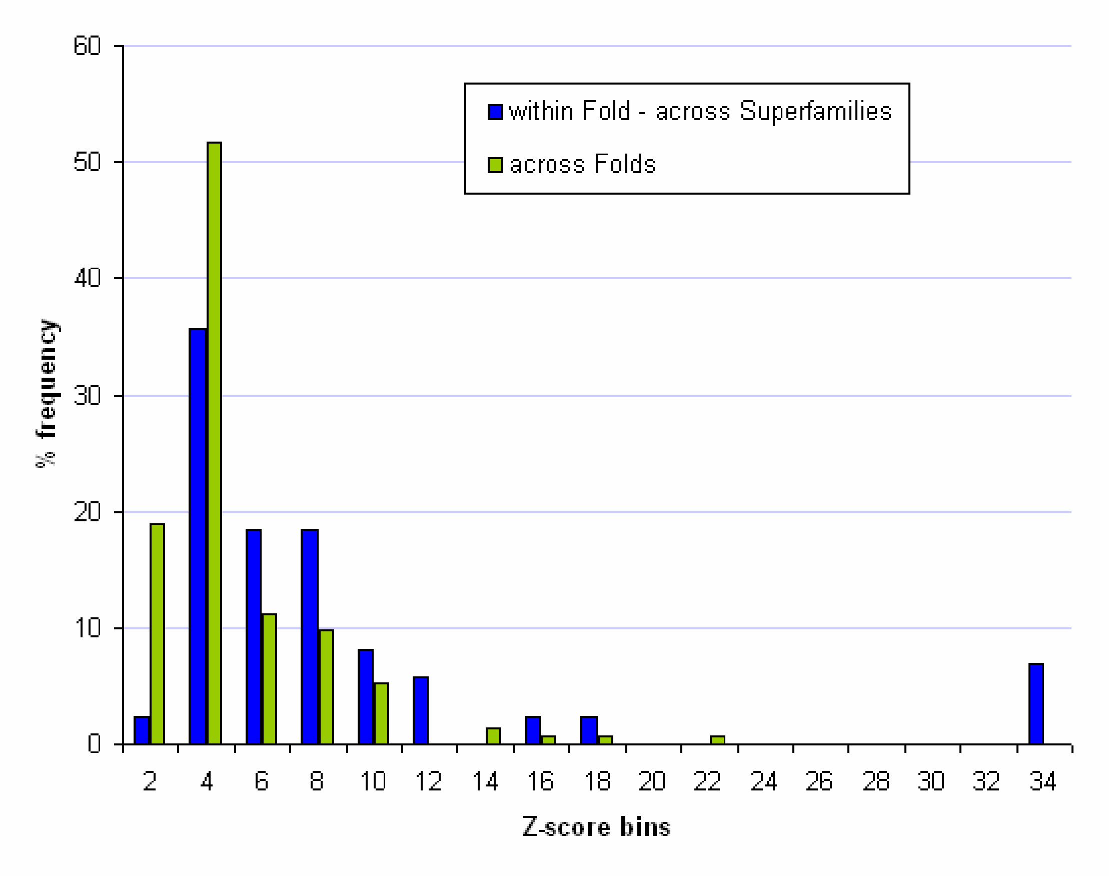
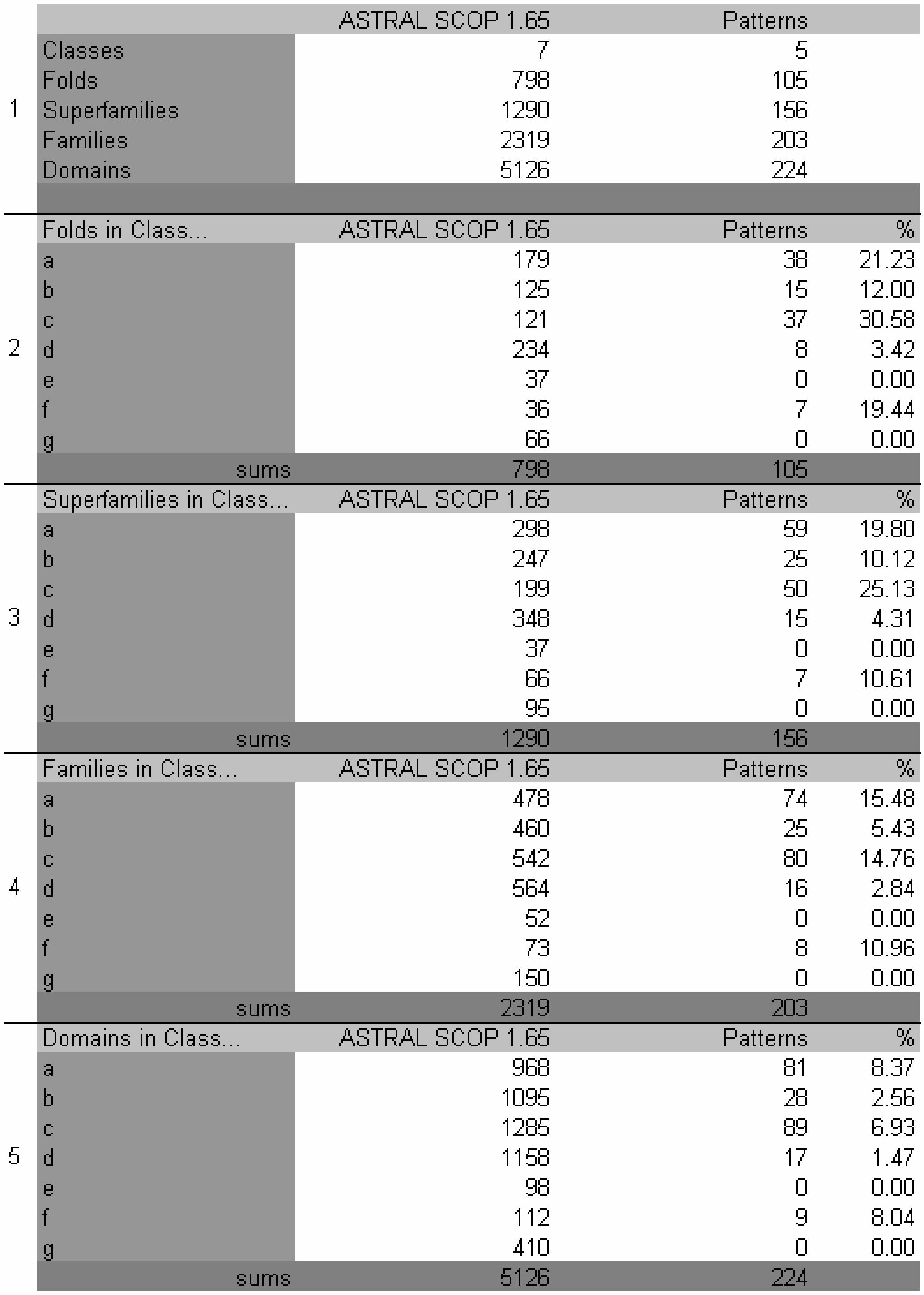


Figure 30. Distribution of Z-scores of DALI structural alignments participating in the SCOP connections made by the 231 patterns in: Blue:within Fold – across Superfamilies, Green:across Folds. Bins contain values greater than x-2 and less than or equal to x, where x is the category value (2 to 34).

Table 11. Tabulated counts featuring combinations of all different sublevels of SCOP in the ASTRAL SCOP dataset version 1.65 (reference) and the subset recovered by our patterns. Sub-table 1 is an overview of hierarchy in the two sets. Sub-tables 2,3,4, and 5 are the counts of Folds, Superfamilies, Families and Domains respectively per Class in each set. The last column is the percentage of the reference ASTRAL SCOP set recovered by patterns.



		ASTRAL SCOP 1.65	Patterns	
1	Classes	7	5	
	Folds	798	105	
	Superfamilies	1290	156	
	Families	2319	203	
	Domains	5126	224	
	Folds in Class...	ASTRAL SCOP 1.65	Patterns	%
2	a	179	38	21.23
	b	125	15	12.00
	c	121	37	30.58
	d	234	8	3.42
	e	37	0	0.00
	f	36	7	19.44
	g	66	0	0.00
	sums	798	105	
	Superfamilies in Class...	ASTRAL SCOP 1.65	Patterns	%
3	a	298	59	19.80
	b	247	25	10.12
	c	199	50	25.13
	d	348	15	4.31
	e	37	0	0.00
	f	66	7	10.61
	g	95	0	0.00
	sums	1290	156	
	Families in Class...	ASTRAL SCOP 1.65	Patterns	%
4	a	478	74	15.48
	b	460	25	5.43
	c	542	80	14.76
	d	564	16	2.84
	e	52	0	0.00
	f	73	8	10.96
	g	150	0	0.00
	sums	2319	203	
	Domains in Class...	ASTRAL SCOP 1.65	Patterns	%
5	a	968	81	8.37
	b	1095	28	2.56
	c	1285	89	6.93
	d	1158	17	1.47
	e	98	0	0.00
	f	112	9	8.04
	g	410	0	0.00
	sums	5126	224	

Beta-Propellers

A significant result of our experiment was the appearance of all five beta-propeller (four-bladed to eight-bladed) folds in our results.

The beta-propeller fold appears as a very fascinating architecture based on four-stranded antiparallel and twisted beta-sheets, radially arranged around a central tunnel. Similar to the alpha/beta-barrel (TIM-barrel) fold, the beta-propeller has a wide range of different functions, and is gaining substantial attention. Proteins containing beta-propeller domains have been implicated in the pathogenesis of a variety of diseases such as cancer, Alzheimer's, Huntington's Disease, arthritis, familial hypercholesterolemia, retinitis pigmentosa, osteogenesis, hypertension, and microbial and viral infections (Pons, Gomez *et al.* 2003).

As can be seen in the boxed section of **Figure 29**, these folds participate in several connections. The most-connected is the seven-bladed beta-propeller fold, which actually is linked to the very large Immunoglobulin-like beta-sandwich Fold (b.1). Murzin suggested that there is a preference for a seven-bladed beta-propeller over six- or eight-bladed ones, based on a model to describe both geometrical parameters and residue packing (Murzin 1992).

In more detail, we looked into the pattern GR[FHWY]LF...[KMR] linking Domains d1qksa2 and d1fwxa2 of Folds b.70.2.1 and b.69.3.1 respectively – these codes translate to the C-terminal (heme d1) domain of cytochrome cd1-nitrite reductase from *Paracoccus denitrificans* and the N-terminal domain of Nitrous oxide reductase from *Paracoccus denitrificans* of the eight- and seven-bladed beta-propeller Folds respectively.

The two proteins are enzymes (EC 1.7.2.1 and EC 1.7.99.6 respectively) and part of the denitrification pathway, one of the major branches of the global nitrogen cycle. In this five-step pathway, nitrate is reduced to dinitrogen through a series of reactions by specific enzymes, named after the substrate they utilise. Nitrite reductase performs step two, and nitrous oxide reductase step four – please see (Moura and Moura 2001) for a full review.

The pattern occupies very similar 3D-shapes in the two Domains, as shown in bright yellow in **Figure 31** (**central** and **A & B)** – Dali had aligned the two regions with a relatively high Z-score of 22. Another pattern that caught our attention was

[KMR].[DLN]..VA.[ITV]..D (grey in **Figure 31 A & B**) – this pattern was methodologically deemed unspecific because it linked two Classes (see Step 5 in the **Materials – Methods Section 9.3**) and therefore did not qualify for out final set of 231 patterns. However, not only is it more popular in the two beta-propeller folds we are discussing, it also follows pattern GR[FHWY]LF...[KMR] in sequence (they overlap by one amino acid position, captured by [KMR]) on Domain d1fwxa2 - on Domain d1qksa2 the two patterns are approximately 200 amino acid positions apart. Both patterns seem to occupy regions that tend to be close in structure (and in sequence in the majority of cases) to Fe-/Cu-binding residues (**Figure 31 C & D**).

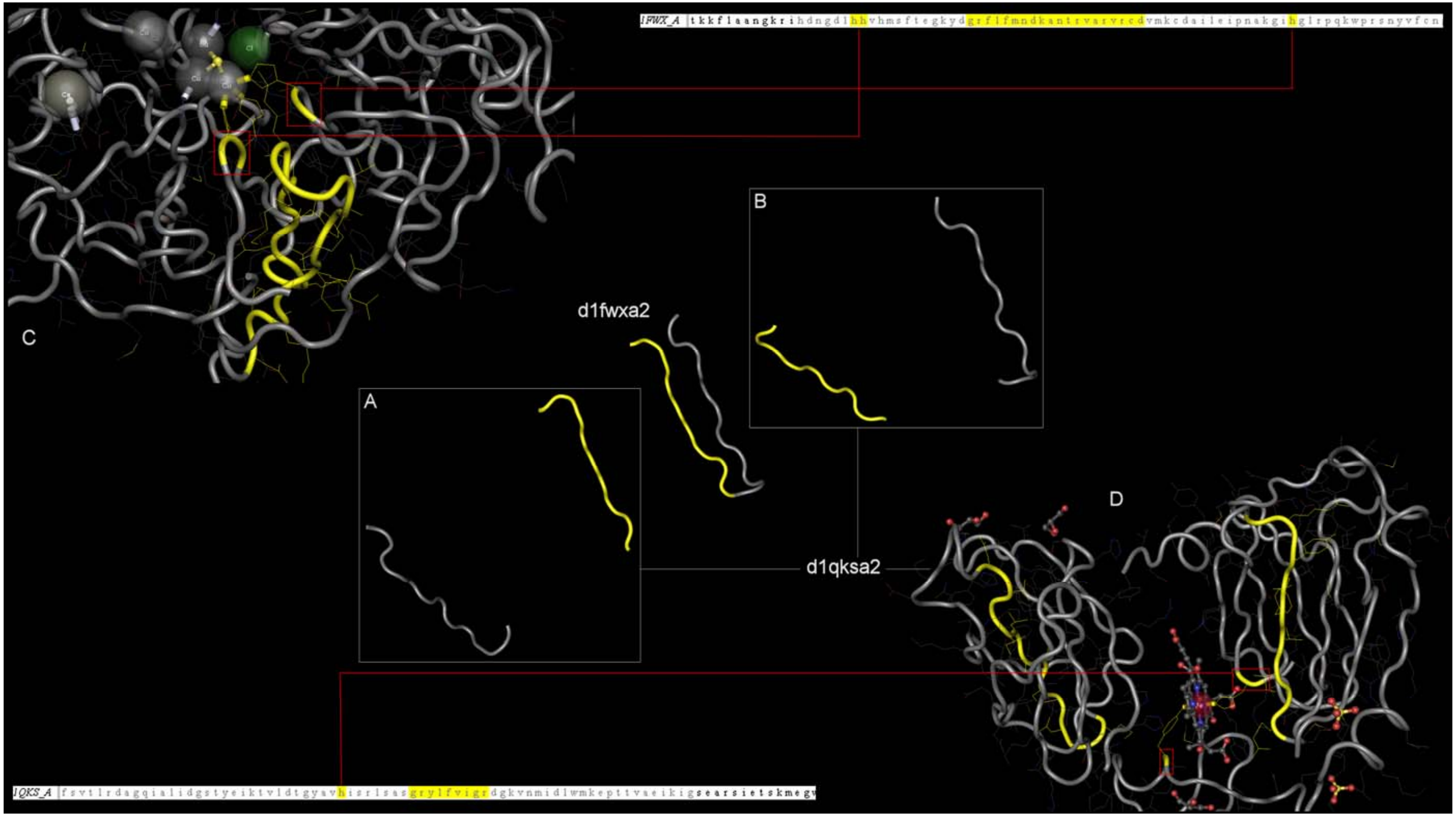

1FWX_A
tkkflaangkrihdngdlhhvhmsftegkydgrflfmndkantrvarvrcdvmkcdaileipnakgihglrpqkwprsnyvfcn
C
d1fwxa2
B
A
d1qksa2
D
1QKS_A
fsvtlrdagqialidgstyeiktvldtgyavhisrlsasgrylfvigrdgkvnmidlwmkepttvaeikigsearsietskmegv

Figure 31 (previous page). Patterns GR[FHWY]LF...[KMR] and [KMR].[DLN]..VA.[ITV]..D linking Domains d1qksa2 and d1fwxa2 of Folds b.70.2.1 and b.69.3.1 respectively. At the centre of the figure: the two patterns in tandem on Domain d1fwxa2 - GR[FHWY]LF...[KMR] in bright yellow. In box A: the patterns on Domain d1qksa2, rotated in such a way that GR[FHWY]LF...[KMR] is aligned between the two Domains. The reverse is performed in box B, where pattern [KMR].[DLN]..VA.[ITV]..D (in grey) is aligned between the two Domains. Box C depicts the two patterns (both in bright yellow) in context on Domain d1fwxa2; the two red boxes contain the three Histidines involved in binding copper (Cu) atoms – all three residues are close in sequence (shown separately in the sequence box) and structure in relation to the two patterns. Box D is equivalent to Box C but looking at the d1qksa2 Domain; again the two patterns are close to two residues binding an iron (Fe) atom – both are close in structure but only one in sequence (see corresponding sequence box) in relation to the two patterns.

9.5 *Discussion*

Previous work in remote sequence similarity detection (Darzentas, Rigoutsos *et al.* 2005)/(**Chapter 8**) has shown both the difficulty of the experiment and the positive potential of our pattern discovery approach. However, we were aware that a more systematic study would be required to reach a safer conclusion about its general applicability. Here we present such a study.

We made great effort not to undermine the generality of our approach by adapting the pattern discovery parameters to each Class. The parameters are based mainly on the requirement for increased sensitivity balanced by computational constraints (i.e. more sensitive parameters would have led to increased pattern discovery time, larger pattern numbers, and more costly data analysis). Furthermore, our approach is sequence guided and structures are only used as an external control. The opposite strategy, i.e. pattern discovery within structural alignments, is limiting on top of being less challenging.

Although we do understand that it is difficult to claim that the structurally similar sequence patterns reported here describe valid evolutionary relationships (in other words cases of divergent evolution), we believe that their intrinsic properties suggest that the evolutionary signal is strong. To showcase this belief, we investigate one example in depth; we show that two enzymes of the fundamental denitrification pathway of the global nitrogen cycle featuring two different SCOP folds (albeit of the multi-bladed beta-propeller kind) can be linked with two patterns with significant characteristics. We provide to the expert readership the full data for further investigation.

9.6 *Disclaimer*

I was the sole contributor to this work, which has been submitted (Darzentas and Ouzounis 2005).

Bibliography

Agarwal, P. and D. J. States (1998). "Comparative accuracy of methods for protein sequence similarity search." Bioinformatics **14**(1): 40-7.

Ahren, D. G. and C. A. Ouzounis (2004). "Robustness of metabolic map reconstruction." J Bioinform Comput Biol **2**(3): 589-93.

Akman, L., A. Yamashita, H. Watanabe, K. Oshima, T. Shiba, M. Hattori and S. Aksoy (2002). "Genome sequence of the endocellular obligate symbiont of tsetse flies, Wigglesworthia glossinidia." Nat Genet **32**(3): 402-7.

Albert, I. and R. Albert (2004). "Conserved network motifs allow protein-protein interaction prediction." Bioinformatics **20**(18): 3346-52.

Alfano, J. R. and A. Collmer (2004). "Type III secretion system effector proteins: double agents in bacterial disease and plant defense." Annu Rev Phytopathol **42**: 385-414.

Altschul, S. F. and W. Gish (1996). "Local alignment statistics." Methods Enzymol **266**: 460-80.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-10.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-402.

Anantharaman, V., E. V. Koonin and L. Aravind (2002). "Comparative genomics and evolution of proteins involved in RNA metabolism." Nucleic Acids Res **30**(7): 1427-64.

Anderson, A. S. and E. M. Wellington (2001). "The taxonomy of Streptomyces and related genera." Int J Syst Evol Microbiol **51**(Pt 3): 797-814.

Andrade, M. A., N. P. Brown, C. Leroy, S. Hoersch, A. de Daruvar, C. Reich, A. Franchini, J. Tamames, A. Valencia, C. Ouzounis*, et al.* (1999). "Automated genome sequence analysis and annotation." Bioinformatics **15**(5): 391-412.

Andreeva, A., D. Howorth, S. E. Brenner, T. J. Hubbard, C. Chothia and A. G. Murzin (2004). "SCOP database in 2004: refinements integrate structure and sequence family data." Nucleic Acids Res **32 Database issue**: D226-9.

Apic, G., J. Gough and S. A. Teichmann (2001a). "Domain combinations in archaeal, eubacterial and eukaryotic proteomes." J Mol Biol **310**(2): 311-25.

Apic, G., J. Gough and S. A. Teichmann (2001b). "An insight into domain combinations." Bioinformatics **17 Suppl 1**: S83-9.

Appella, E. and C. W. Anderson (2000). "Signaling to p53: breaking the posttranslational modification code." Pathol Biol (Paris) **48**(3): 227-45.

Attwood, T. K., P. Bradley, D. R. Flower, A. Gaulton, N. Maudling, A. L. Mitchell, G. Moulton, A. Nordle, K. Paine, P. Taylor*, et al.* (2003). "PRINTS and its automatic supplement, prePRINTS." Nucleic Acids Res **31**(1): 400-2.

Bailey, T. L. and M. Gribskov (1998). "Methods and statistics for combining motif match scores." J Comput Biol **5**(2): 211-21.

Bairoch, A. and R. Apweiler (2000). "The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000." Nucleic Acids Res **28**(1): 45-8.

Bairoch, A., R. Apweiler, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane*, et al.* (2005). "The Universal Protein Resource (UniProt)." Nucleic Acids Res **33 Database Issue**: D154-9.

Baldi, P., Y. Chauvin, T. Hunkapiller and M. A. McClure (1994). "Hidden Markov models of biological primary sequence information." Proc Natl Acad Sci U S A **91**(3): 1059-63.

Bapteste, E., Y. Boucher, J. Leigh and W. F. Doolittle (2004). "Phylogenetic reconstruction and lateral gene transfer." Trends Microbiol **12**(9): 406-11.

Bapteste, E. and C. Brochier (2004). "On the conceptual difficulties in rooting the tree of life." Trends Microbiol **12**(1): 9-13.

Bateman, A., E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S. R. Eddy, S. Griffiths-Jones, K. L. Howe, M. Marshall and E. L. Sonnhammer (2002). "The Pfam protein families database." Nucleic Acids Res **30**(1): 276-80.

Bateman, A., L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. Sonnhammer, *et al.* (2004). "The Pfam protein families database." Nucleic Acids Res **32 Database issue**: D138-41.

Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell and D. L. Wheeler (2005). "GenBank." Nucleic Acids Res **33 Database Issue**: D34-8.

Benton, R., I. M. Palacios and D. St Johnston (2002). "Drosophila 14-3-3/PAR-5 is an essential mediator of PAR-1 function in axis formation." Dev Cell **3**(5): 659-71.

Benton, R. and D. St Johnston (2003). "Drosophila PAR-1 and 14-3-3 inhibit Bazooka/PAR-3 to establish complementary cortical domains in polarized cells." Cell **115**(6): 691-704.

Bernal, A., U. Ear and N. Kyrpides (2001). "Genomes OnLine Database (GOLD): a monitor of genome projects world-wide." Nucleic Acids Res **29**(1): 126-7.

Bernstein, M. P., J. P. Dworkin, S. A. Sandford, G. W. Cooper and L. J. Allamandola (2002). "Racemic amino acids from the ultraviolet photolysis of interstellar ice analogues." Nature **416**(6879): 401-3.

Biernat, J., Y. Z. Wu, T. Timm, Q. Zheng-Fischhofer, E. Mandelkow, L. Meijer and E. M. Mandelkow (2002). "Protein kinase MARK/PAR-1 is required for neurite outgrowth and establishment of neuronal polarity." Mol Biol Cell **13**(11): 4013-28.

Blom, N., T. Sicheritz-Ponten, R. Gupta, S. Gammeltoft and S. Brunak (2004). "Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence." Proteomics **4**(6): 1633-49.

Boeckmann, B., A. Bairoch, R. Apweiler, M. C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, *et al.* (2003). "The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003." Nucleic Acids Res **31**(1): 365-70.

Bork, P., C. Ouzounis, C. Sander, M. Scharf, R. Schneider and E. Sonnhammer (1992). "What's in a genome?" Nature **358**(6384): 287.

Boucher, Y., C. J. Douady, R. T. Papke, D. A. Walsh, M. E. Boudreau, C. L. Nesbo, R. J. Case and W. F. Doolittle (2003). "Lateral gene transfer and the origins of prokaryotic groups." Annu Rev Genet **37**: 283-328.

Bowers, P. M., M. Pellegrini, M. J. Thompson, J. Fierro, T. O. Yeates and D. Eisenberg (2004). "Prolinks: a database of protein functional linkages derived from coevolution." Genome Biol **5**(5): R35.

Bowie, J. U., R. Luthy and D. Eisenberg (1991). "A method to identify protein sequences that fold into a known three-dimensional structure." Science **253**(5016): 164-70.

Brenner, S. E., C. Chothia and T. J. Hubbard (1998). "Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships." Proc Natl Acad Sci U S A **95**(11): 6073-8.

Brown, J. R., C. J. Douady, M. J. Italia, W. E. Marshall and M. J. Stanhope (2001). "Universal trees based on large combined protein sequence data sets." Nat Genet **28**(3): 281-5.

Bucher, P. and A. Bairoch (1994). "A generalized profile syntax for biomolecular sequence motifs and its function in automatic sequence interpretation." Proc Int Conf Intell Syst Mol Biol **2**: 53-61.

Cai, Y. D. and A. J. Doig (2004). "Prediction of Saccharomyces cerevisiae protein functional class from functional domain composition." Bioinformatics **20**(8): 1292-300.

Canters, G. W. and G. Gilardi (1993). "Engineering type 1 copper sites in proteins." FEBS Lett **325**(1-2): 39-48.

Carr, M. H., M. J. Belton, C. R. Chapman, M. E. Davies, P. Geissler, R. Greenberg, A. S. McEwen, B. R. Tufts, R. Greeley, R. Sullivan, *et al.* (1998). "Evidence for a subsurface ocean on Europa." Nature **391**(6665): 363-5.

Casari, G., M. A. Andrade, P. Bork, J. Boyle, A. Daruvar, C. Ouzounis, R. Schneider, J. Tamames, A. Valencia and C. Sander (1995). "Challenging times for bioinformatics." Nature **376**(6542): 647-8.

Casari, G., C. Sander and A. Valencia (1995). "A method to predict functional residues in proteins." Nat Struct Biol **2**(2): 171-8.

Castresana, J. (2001). "Comparative genomics and bioenergetics." Biochim Biophys Acta **1506**(3): 147-62.

Castresana, J. and M. Saraste (1995). "Evolution of energetic metabolism: the respiration-early hypothesis." Trends Biochem Sci **20**(11): 443-8.

Cavicchioli, R. (2002). "Extremophiles and the search for extraterrestrial life." Astrobiology **2**: 281-292.

Chandonia, J. M., G. Hon, N. S. Walker, L. Lo Conte, P. Koehl, M. Levitt and S. E. Brenner (2004). "The ASTRAL Compendium in 2004." Nucleic Acids Res **32**(Database issue): D189-92.

Chen, J., J. B. Anderson, C. DeWeese-Scott, N. D. Fedorova, L. Y. Geer, S. He, D. I. Hurwitz, J. D. Jackson, A. R. Jacobs, C. J. Lanczycki*, et al.* (2003). "MMDB: Entrez's 3D-structure database." Nucleic Acids Res **31**(1): 474-7.

Chong, L. D. and I. O. Daar (2001). "Cloning protein tyrosine kinases by screening cDNA libraries with antiphosphotyrosine antibodies." Methods Mol Biol **124**: 21-37.

Chothia, C. and A. M. Lesk (1982). "Evolution of proteins formed by beta-sheets. I. Plastocyanin and azurin." J Mol Biol **160**(2): 309-23.

Chothia, C. and A. M. Lesk (1986). "The relation between the divergence of sequence and structure in proteins." Embo J **5**(4): 823-6.

Christie, K. R., S. Weng, R. Balakrishnan, M. C. Costanzo, K. Dolinski, S. S. Dwight, S. R. Engel, B. Feierbach, D. G. Fisk, J. E. Hirschman*, et al.* (2004). "Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms." Nucleic Acids Res **32**(Database issue): D311-4.

Churchill, G. A. (1989). "Stochastic models for heterogeneous DNA sequences." Bull Math Biol **51**(1): 79-94.

Chyba, C. and C. Phillips (2001). "Possible ecosystems and the search for life on Europa." Proc Natl Acad Sci U S A **98**(3): 801-4.

Chyba, C. and C. Sagan (1992). "Endogenous production, exogenous delivery and impact-shock synthesis of organic molecules: an inventory for the origins of life." Nature **355**: 125-32.

Chyba, C. F. and C. B. Phillips (2002). "Europa as an abode of life." Orig Life Evol Biosph **32**(1): 47-68.

Chyba, C. F., P. J. Thomas, L. Brookshaw and C. Sagan (1990). "Cometary delivery of organic molecules to the early Earth." Science **249**: 366-73.

Clarke, G. D., R. G. Beiko, M. A. Ragan and R. L. Charlebois (2002). "Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized BLASTP scores." J Bacteriol **184**(8): 2072-80.

Cleaves, H. J. and J. H. Chalmers (2004). "Extremophiles may be irrelevant to the origin of life." Astrobiology **4**: 1-9.

Coulson, R. M., A. J. Enright and C. A. Ouzounis (2001). "Transcription-associated protein families are primarily taxon-specific." Bioinformatics **17**(1): 95-7.

Darwin, C. (1859). On the origin of species by means of natural selection. London,, J. Murray.

Darzentas, N. and C. Ouzounis (2005). "Bridging SCOP Folds and Superfamilies with structurally equivalent sequence patterns." **[submitted]**.

Darzentas, N., I. Rigoutsos and C. A. Ouzounis (2005). "Sensitive detection of sequence similarity using combinatorial pattern discovery: a challenging study of two distantly related protein families." Proteins **61**(4): 926-37.

de Duve, C. (2003). "A research proposal on the origin of life." Orig Life Evol Biosph **33**(6): 559-74.

de Haen, C., E. Swanson and D. C. Teller (1976). "The evolutionary origin of proinsulin. Amino acid sequence homology with the trypsin-related serine proteases detected and evaluated by new statistical methods." J Mol Biol **106**(3): 639-61.

De Rienzo, F., R. R. Gabdoulline, M. C. Menziani and R. C. Wade (2000). "Blue copper proteins: a comparative analysis of their molecular interaction properties." Protein Sci **9**(8): 1439-54.

De Rienzo, F., R. R. Gabdoulline, R. C. Wade, M. Sola and M. C. Menziani (2004). "Computational approaches to structural and functional analysis of plastocyanin and other blue copper proteins." Cell Mol Life Sci **61**(10): 1123-42.

Delaye, L. and A. Lazcano (2000). RNA-binding peptides as molecular fossils. Origins from the Big-Bang to Biology: Proceedings of the First Ibero-American School of Astrobiology. J. Chela-Flores, G. A. Lemerchand and J. Oró. Dordrecht, Kluwer Academic Publishers**:** 285-288.

des Jardins, M., P. D. Karp, M. Krummenacker, T. J. Lee and C. A. Ouzounis (1997). "Prediction of enzyme classification from protein sequence without the use of sequence similarity." Proc Int Conf Intell Syst Mol Biol **5**: 92-9.

Des Marais, D. J. and M. R. Walter (1999). "Astrobiology: exploring the origins, evolution, and distribution of life in the Universe." Annu Rev Ecol Syst **30**: 397-420.

Deshpande, N., K. J. Addess, W. F. Bluhm, J. C. Merino-Ott, W. Townsend-Merino, Q. Zhang, C. Knezevich, L. Xie, L. Chen, Z. Feng, *et al.* (2005). "The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema." Nucleic Acids Res **33 Database Issue**: D233-7.

DiRuggiero, J., J. R. Brown, A. P. Bogert and F. T. Robb (1999). "DNA repair systems in archaea: mementos from the last universal common ancestor?" J Mol Evol **49**(4): 474-84.

Doolittle, R. F. (1981). "Similar amino acid sequences: chance or common ancestry?" Science **214**(4517): 149-59.

Doolittle, R. F. (2002). "Biodiversity: microbial genomes multiply." Nature **416**(6882): 697-700.

Doolittle, W. F. (1999). "Phylogenetic classification and the universal tree." Science **284**(5423): 2124-9.

Doolittle, W. F. and J. R. Brown (1994). "Tempo, mode, the progenote, and the universal root." Proc Natl Acad Sci U S A **91**(15): 6721-8.

Dougherty, M. K. and D. K. Morrison (2004). "Unlocking the code of 14-3-3." J Cell Sci **117**(Pt 10): 1875-84.

Drysdale, R. A., M. A. Crosby, W. Gelbart, K. Campbell, D. Emmert, B. Matthews, S. Russo, A. Schroeder, F. Smutniak, P. Zhang*, et al.* (2005). "FlyBase: genes and gene models." Nucleic Acids Res **33 Database Issue**: D390-5.

Eddy, S. R. (1998). "Profile hidden Markov models." Bioinformatics **14**(9): 755-63.

Eisenberg, D., E. M. Marcotte, I. Xenarios and T. O. Yeates (2000). "Protein function in the post-genomic era." Nature **405**(6788): 823-6.

Enright, A. J., I. Iliopoulos, N. C. Kyrpides and C. A. Ouzounis (1999). "Protein interaction maps for complete genomes based on gene fusion events." Nature **402**(6757): 86-90.

Enright, A. J., V. Kunin and C. A. Ouzounis (2003). "Protein families and TRIBES in genome sequence space." Nucleic Acids Res **31**(15): 4632-8.

Enright, A. J. and C. A. Ouzounis (2000). "GeneRAGE: a robust algorithm for sequence clustering and domain detection." Bioinformatics **16**(5): 451-7.

Enright, A. J. and C. A. Ouzounis (2001). "BioLayout--an automatic graph layout algorithm for similarity visualization." Bioinformatics **17**(9): 853-4.

Enright, A. J., S. Van Dongen and C. A. Ouzounis (2002). "An efficient algorithm for large-scale detection of protein families." Nucleic Acids Res **30**(7): 1575-84.

Eppig, J. T., C. J. Bult, J. A. Kadin, J. E. Richardson, J. A. Blake, A. Anagnostopoulos, R. M. Baldarelli, M. Baya, J. S. Beal, S. M. Bello*, et al.* (2005). "The Mouse Genome Database (MGD): from genes to mice--a community resource for mouse biology." Nucleic Acids Res **33 Database Issue**: D471-5.

Fitz-Gibbon, S. T. and C. H. House (1999). "Whole genome-based phylogenetic analysis of free-living microorganisms." Nucleic Acids Res **27**(21): 4218-22.

Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick*, et al.* (1995). "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd." Science **269**(5223): 496-512.

Floratos, A. (1999). Pattern Discovery in Biology: Theory and Applications. Department of Computer Science, New York University**:** 250.

Forterre, P., N. Benachenhou-Lahfa, F. Confalonieri, M. Duguet, C. Elie and B. Labedan (1992). "The nature of the last universal ancestor and the root of the tree of life, still open questions." Biosystems **28**(1-3): 15-32.

Forterre, P. and H. Philippe (1999). "The last universal common ancestor (LUCA), simple or complex?" Biol Bull **196**(3): 373-5; discussion 375-7.

Fox, G. E., E. Stackebrandt, R. B. Hespell, J. Gibson, J. Maniloff, T. A. Dyer, R. S. Wolfe, W. E. Balch, R. S. Tanner, L. J. Magrum*, et al.* (1980). "The phylogeny of prokaryotes." Science **209**(4455): 457-63.

Galperin, M. Y. (2005). "The Molecular Biology Database Collection: 2005 update." Nucleic Acids Res **33 Database Issue**: D5-24.

Galtier, N., M. Gouy and C. Gautier (1996). "SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny." Comput Appl Biosci **12**(6): 543-8.

Genin, S. and C. Boucher (2004). "Lessons learned from the genome analysis of *Ralstonia solanacearum*." Annu Rev Phytopathol **42**: 107-34.

Glansdorff, N. (2000). "About the last common ancestor, the universal life-tree and lateral gene transfer: a reappraisal." Mol Microbiol **38**(2): 177-85.

Glockner, F. O., M. Kube, M. Bauer, H. Teeling, T. Lombardot, W. Ludwig, D. Gade, A. Beck, K. Borzym, K. Heitmann*, et al.* (2003). "Complete genome sequence of the marine planctomycete *Pirellula* sp. strain 1." Proc Natl Acad Sci U S A **100**(14): 8298-303.

Gold, T. (1992). "The deep, hot biosphere." Proc Natl Acad Sci U S A **89**(13): 6045-9.

Goldovsky, L., P. Janssen, D. Ahren, B. Audit, I. Cases, N. Darzentas, A. J. Enright, N. Lopez-Bigas, J. M. Peregrin-Alvarez, M. Smith*, et al.* (2005). "CoGenT++: an extensive and extensible data environment for computational genomics." Bioinformatics **21**(19): 3806-10.

Green, M. L. and P. D. Karp (2004). "A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases." BMC Bioinformatics **5**(1): 76.

Gribaldo, S. and H. Philippe (2002). "Ancient phylogenetic relationships." Theor Popul Biol **61**(4): 391-408.

Gribskov, M., R. Luthy and D. Eisenberg (1990). "Profile analysis." Methods Enzymol **183**: 146-59.

Gribskov, M., A. D. McLachlan and D. Eisenberg (1987). "Profile analysis: detection of distantly related proteins." Proc Natl Acad Sci U S A **84**(13): 4355-8.

Gusfield, D. (1997). Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology, Cambridge University Press.

Gusfield, D., S. Eddhu and C. Langley (2004). "Optimal, efficient reconstruction of phylogenetic networks with constrained recombination." J Bioinform Comput Biol **2**(1): 173-213.

Harris, J. K., S. T. Kelley, G. B. Spiegelman and N. R. Pace (2003). "The genetic core of the universal ancestor." Genome Res **13**(3): 407-12.

Hegyi, H. and P. Bork (1997). "On the classification and evolution of protein modules." J Protein Chem **16**(5): 545-51.

Hegyi, H. and M. Gerstein (1999). "The relationship between protein structure and function: a comprehensive survey with application to the yeast genome." J Mol Biol **288**(1): 147-64.

Henikoff, S., S. Pietrokovski and J. G. Henikoff (1998). "Superior performance in protein homology detection with the Blocks Database servers." Nucleic Acids Res **26**(1): 309-12.

Holley, R. W., G. A. Everett, J. T. Madison and A. Zamir (1965). "Nucleotide Sequences in the Yeast Alanine Transfer Ribonucleic Acid." J Biol Chem **240**: 2122-8.

Holm, L. (1998). "Unification of protein families." Curr Opin Struct Biol **8**(3): 372-9.

Holm, L., C. Ouzounis, C. Sander, G. Tuparev and G. Vriend (1992). "A database of protein structure families with common folding motifs." Protein Sci **1**(12): 1691-8.

Holm, L. and C. Sander (1993). "Structural alignment of globins, phycocyanins and colicin A." FEBS Lett **315**(3): 301-6.

Holm, L. and C. Sander (1996). "Mapping the protein universe." Science **273**(5275): 595-603.

Holmes, I. and R. Durbin (1998). "Dynamic programming alignment accuracy." J Comput Biol **5**(3): 493-504.

Horneck, G. (1995). "Exobiology, the study of the origin, evolution and distribution of life within the context of cosmic evolution: a review." Planet Space Sci. **43**: 189-217.

Hubbard, T., D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down*, et al.* (2002). "The Ensembl genome database project." Nucleic Acids Res **30**(1): 38-41.

Hulo, N., C. J. Sigrist, V. Le Saux, P. S. Langendijk-Genevaux, L. Bordoli, A. Gattiker, E. De Castro, P. Bucher and A. Bairoch (2004). "Recent improvements to the PROSITE database." Nucleic Acids Res **32 Database issue**: D134-7.

Iliopoulos, I., S. Tsoka, M. A. Andrade, P. Janssen, B. Audit, A. Tramontano, A. Valencia, C. Leroy, C. Sander and C. A. Ouzounis (2001). "Genome sequences and great expectations." Genome Biol **2**(1): INTERACTIONS0001.

Ivanisenko, V. A., S. S. Pintus, D. A. Grigorovich and N. A. Kolchanov (2005). "PDBSite: a database of the 3D structure of protein functional sites." Nucleic Acids Res **33 Database Issue**: D183-7.

Jaakkola, T., M. Diekhans and D. Haussler (1999). "Using the Fisher kernel method to detect remote protein homologies." Proc Int Conf Intell Syst Mol Biol: 149-58.

Jaakkola, T., M. Diekhans and D. Haussler (2000). "A discriminative framework for detecting remote protein homologies." J Comput Biol **7**(1-2): 95-114.

Jacoby, G. A. (1996). "Antimicrobial-resistant pathogens in the 1990s." Annu Rev Med **47**: 169-79.

Janssen, P., B. Audit, I. Cases, N. Darzentas, L. Goldovsky, V. Kunin, N. Lopez-Bigas, J. M. Peregrin-Alvarez, J. B. Pereira-Leal, S. Tsoka*, et al.* (2003). "Beyond 100 genomes." Genome Biol **4**(5): 402.

Janssen, P., A. J. Enright, B. Audit, I. Cases, L. Goldovsky, N. Harte, V. Kunin and C. A. Ouzounis (2003). "COmplete GENome Tracking (COGENT): a flexible data environment for computational genomics." Bioinformatics **19**(11): 1451-2.

Janssen, P., L. Goldovsky, V. Kunin, N. Darzentas and C. A. Ouzounis (2005). "Genome coverage, literally speaking. The challenge of annotating 200 genomes with 4 million publications." EMBO Rep **6**(5): 397-9.

Jensen, L. J., R. Gupta, H. H. Staerfeldt and S. Brunak (2003). "Prediction of human protein function according to Gene Ontology categories." Bioinformatics **19**(5): 635-42.

Kamekura, M. (1998). "Diversity of extremely halophilic bacteria." Extremophiles **2**(3): 289-95.

Kanehisa, M., S. Goto, S. Kawashima, Y. Okuno and M. Hattori (2004). "The KEGG resource for deciphering the genome." Nucleic Acids Res **32**(Database issue): D277-80.

Kaneko, T., Y. Nakamura, S. Sato, K. Minamisawa, T. Uchiumi, S. Sasamoto, A. Watanabe, K. Idesawa, M. Iriguchi, K. Kawashima*, et al.* (2002). "Complete genomic sequence of nitrogen-fixing symbiotic bacterium *Bradyrhizobium japonicum* USDA110." DNA Res **9**(6): 189-97.

Kanz, C., P. Aldebert, N. Althorpe, W. Baker, A. Baldwin, K. Bates, P. Browne, A. van den Broek, M. Castro, G. Cochrane*, et al.* (2005). "The EMBL Nucleotide Sequence Database." Nucleic Acids Res **33 Database Issue**: D29-33.

Karchin, R., K. Karplus and D. Haussler (2002). "Classifying G-protein coupled receptors with support vector machines." Bioinformatics **18**(1): 147-59.

Karp, P. D. (1998). "Metabolic databases." Trends Biochem Sci **23**(3): 114-6.

Karp, P. D. (2000). "An ontology for biological function based on molecular interactions." Bioinformatics **16**(3): 269-85.

Karp, P. D. (2001). "Pathway databases: a case study in computational symbolic theories." Science **293**(5537): 2040-4.

Karp, P. D., M. Krummenacker, S. Paley and J. Wagg (1999). "Integrated pathway-genome databases and their role in drug discovery." Trends Biotechnol **17**(7): 275-81.

Karp, P. D., C. Ouzounis and S. Paley (1996). "HinCyc: a knowledge base of the complete genome and metabolic pathways of *H. influenzae*." Proc Int Conf Intell Syst Mol Biol **4**: 116-24.

Karp, P. D., C. A. Ouzounis, C. Moore-Kochlacs, L. Goldovsky, P. Kaipa, D. Ahren, S. Tsoka, N. Darzentas, V. Kunin and N. Lopez-Bigas (2005). "Expansion of the BioCyc collection of pathway/genome databases to 160 genomes." Nucleic Acids Res **33**(19): 6083-9.

Karp, P. D., S. Paley, C. J. Krieger and P. Zhang (2004). "An evidence ontology for use in pathway/genome databases." Pac Symp Biocomput: 190-201.

Karp, P. D., S. Paley and P. Romero (2002). "The Pathway Tools software." Bioinformatics **18 Suppl 1**: S225-32.

Karplus, K., C. Barrett and R. Hughey (1998). "Hidden Markov models for detecting remote protein homologies." Bioinformatics **14**(10): 846-56.

Kashefi, K. and D. R. Lovley (2003). "Extending the upper temperature limit for life." Science **301**(5635): 934.

Keseler, I. M., J. Collado-Vides, S. Gama-Castro, J. Ingraham, S. Paley, I. T. Paulsen, M. Peralta-Gil and P. D. Karp (2005). "EcoCyc: a comprehensive database resource for *Escherichia coli*." Nucleic Acids Res **33 Database Issue**: D334-7.

Kiers, E. T., R. A. Rousseau, S. A. West and R. F. Denison (2003). "Host sanctions and the legume-rhizobium mutualism." Nature **425**(6953): 78-81.

King, R. D., A. Karwath, A. Clare and L. Dehaspe (2000). "Accurate prediction of protein functional class from sequence in the Mycobacterium tuberculosis and Escherichia coli genomes using data mining." Yeast **17**(4): 283-93.

Koonin, E. V. (2003). "Comparative genomics, minimal gene-sets and the last universal common ancestor." Nat Rev Microbiol **1**(2): 127-36.

Koonin, E. V., K. S. Makarova, I. B. Rogozin, L. Davidovic, M. C. Letellier and L. Pellegrini (2003). "The rhomboids: a nearly ubiquitous family of intramembrane serine proteases that probably evolved by multiple ancient horizontal gene transfers." Genome Biol **4**(3): R19.

Korbel, J. O., B. Snel, M. A. Huynen and P. Bork (2002). "SHOT: a web server for the construction of genome phylogenies." Trends Genet **18**(3): 158-62.

Kornberg, A., N. N. Rao and D. Ault-Riche (1999). "Inorganic polyphosphate: a molecule of many functions." Annu Rev Biochem **68**: 89-125.

Krieger, C. J., P. Zhang, L. A. Mueller, A. Wang, S. Paley, M. Arnaud, J. Pick, S. Y. Rhee and P. D. Karp (2004). "MetaCyc: a multiorganism database of metabolic pathways and enzymes." Nucleic Acids Res **32**(Database issue): D438-42.

Krogh, A., M. Brown, I. S. Mian, K. Sjolander and D. Haussler (1994). "Hidden Markov models in computational biology. Applications to protein modeling." J Mol Biol **235**(5): 1501-31.

Kuan, Y.-J., S. Charnley, H.-C. Huang, W.-L. Tseng and Z. Kisiel (2003). "Interstellar glycine." Astroph. J. **593**: 848-867.

Kunin, V., D. Ahren, L. Goldovsky, P. Janssen and C. A. Ouzounis (2005). "Measuring genome conservation across taxa: divided strains and united kingdoms." Nucleic Acids Res **33**(2): 616-21.

Kunin, V., I. Cases, A. J. Enright, V. de Lorenzo and C. A. Ouzounis (2003). "Myriads of protein families, and still counting." Genome Biol **4**(2): 401.

Kunin, V., L. Goldovsky, N. Darzentas and C. A. Ouzounis (2005). "The net of life: reconstructing the microbial phylogenetic network." Genome Res **15**(7): 954-9.

Kunin, V. and C. A. Ouzounis (2003a). "The balance of driving forces during genome evolution in prokaryotes." Genome Res **13**(7): 1589-94.

Kunin, V. and C. A. Ouzounis (2003b). "GeneTRACE-reconstruction of gene content of ancestral species." Bioinformatics **19**(11): 1412-6.

Kyrpides, N., R. Overbeek and C. Ouzounis (1999). "Universal protein families and the functional content of the last universal common ancestor." J Mol Evol **49**(4): 413-23.

Kyrpides, N. C. and C. A. Ouzounis (1999). "Transcription in archaea." Proc Natl Acad Sci U S A **96**(15): 8545-50.

Kyrpides, N. C. and C. R. Woese (1998a). "Tetratrico-peptide-repeat proteins in the archaeon *Methanococcus jannaschii*." Trends Biochem Sci **23**(7): 245-7.

Kyrpides, N. C. and C. R. Woese (1998b). "Universally conserved translation initiation factors." Proc Natl Acad Sci U S A **95**(1): 224-8.

Lander, E. S.L. M. LintonB. BirrenC. NusbaumM. C. ZodyJ. BaldwinK. DevonK. DewarM. DoyleW. FitzHugh*, et al.* (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.

Larsson, P., P. C. Oyston, P. Chain, M. C. Chu, M. Duffield, H. H. Fuxelius, E. Garcia, G. Halltorp, D. Johansson, K. E. Isherwood*, et al.* (2005). "The complete genome sequence of *Francisella tularensis*, the causative agent of tularemia." Nat Genet **37**(2): 153-9.

Lazcano, A. and P. Forterre (1999a). The Last Universal Common Ancestor and Beyond. Proceedings of a workshop. France, July, 1996. J Mol Evol.

Lazcano, A. and P. Forterre (1999b). "The molecular search for the last common ancestor." J Mol Evol **49**(4): 411-2.

Leahy, D. J., W. A. Hendrickson, I. Aukhil and H. P. Erickson (1992). "Structure of a fibronectin type III domain from tenascin phased by MAD analysis of the selenomethionyl protein." Science **258**(5084): 987-91.

Li, W., L. Jaroszewski and A. Godzik (2001). "Clustering of highly homologous sequences to reduce the size of large protein databases." Bioinformatics **17**(3): 282-3.

Lin, J. and M. Gerstein (2000). "Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels." Genome Res **10**(6): 808-18.

Liu, H., N. D. Rodgers, X. Jiao and M. Kiledjian (2002). "The scavenger mRNA decapping enzyme DcpS is a member of the HIT family of pyrophosphatases." Embo J **21**(17): 4699-708.

Lombard, D. B., K. F. Chua, R. Mostoslavsky, S. Franco, M. Gostissa and F. W. Alt (2005). "DNA Repair, Genome Stability, and Aging." Cell **120**(4): 497-512.

Lopez-Bigas, N. and C. A. Ouzounis (2004). "Genome-wide identification of genes likely to be involved in human genetic disease." Nucleic Acids Res **32**(10): 3108-14.

Lord, P. W., R. D. Stevens, A. Brass and C. A. Goble (2003). "Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation." Bioinformatics **19**(10): 1275-83.

Lupas, A. N., C. P. Ponting and R. B. Russell (2001). "On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world?" J Struct Biol **134**(2-3): 191-203.

Lynch, M. (2002). "Intron evolution as a population-genetic process." Proc Natl Acad Sci U S A **99**(9): 6118-23.

Madera, M. and J. Gough (2002). "A comparison of profile hidden Markov model procedures for remote homology detection." Nucleic Acids Res **30**(19): 4321-8.

Maglott, D., J. Ostell, K. D. Pruitt and T. Tatusova (2005). "Entrez Gene: gene-centered information at NCBI." Nucleic Acids Res **33 Database Issue**: D54-8.

Makarenkov, V. and P. Legendre (2004). "From a phylogenetic tree to a reticulated network." J Comput Biol **11**(1): 195-212.

Marcotte, E. M., M. Pellegrini, H. L. Ng, D. W. Rice, T. O. Yeates and D. Eisenberg (1999). "Detecting protein function and protein-protein interactions from genome sequences." Science **285**(5428): 751-3.

Marcotte, E. M., I. Xenarios, A. M. van Der Bliek and D. Eisenberg (2000). "Localizing proteins in the cell from their phylogenetic profiles." Proc Natl Acad Sci U S A **97**(22): 12115-20.

Margulis, L. (1996). "Archaeal-eubacterial mergers in the origin of Eukarya: phylogenetic classification of life." Proc Natl Acad Sci U S A **93**(3): 1071-6.

Martin, S. G. and D. St Johnston (2003). "A role for Drosophila LKB1 in anterior-posterior axis formation and epithelial polarity." Nature **421**(6921): 379-84.

Martin, W. (1999). "Mosaic bacterial chromosomes: a challenge en route to a tree of genomes." Bioessays **21**(2): 99-104.

Martzen, M. R., S. M. McCraith, S. L. Spinelli, F. M. Torres, S. Fields, E. J. Grayhack and E. M. Phizicky (1999). "A biochemical genomics approach for identifying genes by the activity of their products." Science **286**(5442): 1153-5.

McCord, T. B., R. W. Carlson, W. D. Smythe, G. B. Hansen, R. N. Clark, C. A. Hibbitts, F. P. Fanale, J. C. Granahan, M. Segura, D. L. Matson*, et al.* (1997). "Organics and other molecules in the surfaces of Callisto and Ganymede." Science **278**(5336): 271-5.

McGinnis, S. and T. L. Madden (2004). "BLAST: at the core of a powerful and diverse set of sequence analysis tools." Nucleic Acids Res **32**(Web Server issue): W20-5.

Mellor, J. C., I. Yanai, K. H. Clodfelter, J. Mintseris and C. DeLisi (2002). "Predictome: a database of putative functional links between proteins." Nucleic Acids Res **30**(1): 306-9.

Messerschmidt, A. and R. Huber (1990). "The blue oxidases, ascorbate oxidase, laccase and ceruloplasmin. Modelling and structural relationships." Eur J Biochem **187**(2): 341-52.

Messerschmidt, A., R. Ladenstein, R. Huber, M. Bolognesi, L. Avigliano, R. Petruzzelli, A. Rossi and A. Finazzi-Agro (1992). "Refined crystal structure of ascorbate oxidase at 1.9 A resolution." J Mol Biol **224**(1): 179-205.

Miller, S. L. and A. Lazcano (1995). "The origin of life--did it occur at high temperatures?" J Mol Evol **41**: 689-92.

Mirkin, B. G., T. I. Fenner, M. Y. Galperin and E. V. Koonin (2003). "Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes." BMC Evol Biol **3**(1): 2.

Moore, G. E. (1965). "Cramming more components onto integrated circuits." Electronics **38**(8).

Moura, I. and J. J. Moura (2001). "Structural aspects of denitrifying enzymes." Curr Opin Chem Biol **5**(2): 168-75.

Mulder, N. J., R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bradley, P. Bork, P. Bucher, L. Cerutti*, et al.* (2005). "InterPro, progress and status in 2005." Nucleic Acids Res **33 Database Issue**: D201-5.

Murzin, A. G. (1992). "Structural principles for the propeller assembly of beta-sheets: the preference for seven-fold symmetry." Proteins **14**(2): 191-201.

Murzin, A. G. (1998). "How far divergent evolution goes in proteins." Curr Opin Struct Biol **8**(3): 380-7.

Murzin, A. G. and A. Bateman (1997). "Distant homology recognition using structural classification of proteins." Proteins **Suppl 1**: 105-12.

Murzin, A. G., S. E. Brenner, T. Hubbard and C. Chothia (1995). "SCOP: a structural classification of proteins database for the investigation of sequences and structures." J Mol Biol **247**(4): 536-40.

Nakamura, Y., T. Itoh, H. Matsuda and T. Gojobori (2004). "Biased biological functions of horizontally transferred genes in prokaryotic genomes." Nat Genet **36**(7): 760-6.

Navarro, G. (2001). "NR-grep: a Fast and Flexible Pattern Matching Tool." Software Practice and Experience (SPE) **31**: 1265-1312.

Nelson, K. E., I. T. Paulsen, J. F. Heidelberg and C. M. Fraser (2000). "Status of genome projects for nonpathogenic bacteria and archaea." Nat Biotechnol **18**(10): 1049-54.

Nevill-Manning, C. G., T. D. Wu and D. L. Brutlag (1998). "Highly specific protein sequence motifs for genome analysis." Proc Natl Acad Sci U S A **95**(11): 5865-71.

Nisbet, E. G. and N. H. Sleep (2001). "The habitat and nature of early life." Nature **409**(6823): 1083-91.

Nishikawa, K. and T. Ooi (1986). "Amino acid sequence homology applied to the prediction of protein secondary structures, and joint prediction with existing methods." Biochim Biophys Acta **871**(1): 45-54.

Nishimura, I., Y. Yang and B. Lu (2004). "PAR-1 kinase plays an initiator role in a temporally ordered phosphorylation process that confers tau toxicity in Drosophila." Cell **116**(5): 671-82.

Notredame, C. (2002). "Recent progress in multiple sequence alignment: a survey." Pharmacogenomics **3**(1): 131-44.

Ouzounis, C., V. Kunin, N. Darzentas and L. Goldovsky (2005). "A minimal estimate for the gene content of the last universal common ancestor - exobiology from a terrestrial perspective." Res. Microbiol. **[in press]**.

Ouzounis, C. and N. Kyrpides (1996a). "The emergence of major cellular processes in evolution." FEBS Lett **390**(2): 119-23.

Ouzounis, C. and C. Sander (1991). "A structure-derived sequence pattern for the detection of type I copper binding domains in distantly related proteins." FEBS Lett **279**(1): 73-8.

Ouzounis, C., C. Sander, M. Scharf and R. Schneider (1993). "Prediction of protein structure by evaluation of sequence-structure fitness. Aligning sequences to contact profiles derived from three-dimensional structures." J Mol Biol **232**(3): 805-25.

Ouzounis, C. A. and P. D. Karp (2000). "Global properties of the metabolic map of *Escherichia coli*." Genome Res **10**(4): 568-76.

Ouzounis, C. A. and P. D. Karp (2002). "The past, present and future of genome-wide re-annotation." Genome Biol **3**(2): COMMENT2001.

Ouzounis, C. A. and N. C. Kyrpides (1996b). "Parallel origins of the nucleosome core and eukaryotic transcription from Archaea." J Mol Evol **42**(2): 234-9.

Overbeek, R., M. Fonstein, M. D'Souza, G. D. Pusch and N. Maltsev (1999). "The use of gene clusters to infer functional coupling." Proc Natl Acad Sci U S A **96**(6): 2896-901.

Overbeek, R., N. Larsen, G. D. Pusch, M. D'Souza, E. Selkov, Jr., N. Kyrpides, M. Fonstein, N. Maltsev and E. Selkov (2000). "WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction." Nucleic Acids Res **28**(1): 123-5.

Paley, S. M. and P. D. Karp (2002). "Evaluation of computational metabolic-pathway predictions for *Helicobacter pylori*." Bioinformatics **18**(5): 715-24.

Parida, L., A. Floratos and I. I. Rigoutsos (1998). "MUSCA: An Algorithm for Constrained Alignment of Multiple Data Sequences." Genome Inform Ser Workshop Genome Inform **9**: 112-119.

Park, J., K. Karplus, C. Barrett, R. Hughey, D. Haussler, T. Hubbard and C. Chothia (1998). "Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods." J Mol Biol **284**(4): 1201-10.

Pawlowski, K., L. Jaroszewski, L. Rychlewski and A. Godzik (2000). "Sensitive sequence comparison as protein function predictor." Pac Symp Biocomput: 42-53.

Pearson, W. R. (1990). "Rapid and sensitive sequence comparison with FASTP and FASTA." Methods Enzymol **183**: 63-98.

Pearson, W. R. and D. J. Lipman (1988). "Improved tools for biological sequence comparison." Proc Natl Acad Sci U S A **85**(8): 2444-8.

Pellegrini, M., E. M. Marcotte, M. J. Thompson, D. Eisenberg and T. O. Yeates (1999). "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles." Proc Natl Acad Sci U S A **96**(8): 4285-8.

Penny, D. and A. Poole (1999). "The nature of the last universal common ancestor." Curr Opin Genet Dev **9**(6): 672-7.

Peregrin-Alvarez, J. M., S. Tsoka and C. A. Ouzounis (2003). "The phylogenetic extent of metabolic enzymes and pathways." Genome Res **13**(3): 422-7.

Peterson, J. D., L. A. Umayam, T. Dickinson, E. K. Hickey and O. White (2001). "The Comprehensive Microbial Resource." Nucleic Acids Res **29**(1): 123-5.

Piel, W. H., M. J. Sanderson and M. J. Donoghue (2003). "The small-world dynamics of tree networks and data mining in phyloinformatics." Bioinformatics **19**(9): 1162-8.

Pietrokovski, S., J. G. Henikoff and S. Henikoff (1996). "The Blocks database--a system for protein classification." Nucleic Acids Res **24**(1): 197-200.

Ponomarenko, J. V., P. E. Bourne and I. N. Shindyalov (2005). "Assigning new GO annotations to protein data bank sequences by combining structure and sequence homology." Proteins **58**(4): 855-65.

Pons, T., R. Gomez, G. Chinea and A. Valencia (2003). "Beta-propellers: associated functions and their role in human diseases." Curr Med Chem **10**(6): 505-24.

Prieur, D., G. Erauso and C. Jeanthon (1995). "Hyperthermophilic life at deep-sea hydrothermal vents." Planet Space Sci **43**(1-2): 115-22.

Promponas, V. J., A. J. Enright, S. Tsoka, D. P. Kreil, C. Leroy, S. Hamodrakas, C. Sander and C. A. Ouzounis (2000). "CAST: an iterative algorithm for the complexity analysis of sequence tracts. Complexity analysis of sequence tracts." Bioinformatics **16**(10): 915-22.

Pruitt, K. D., T. Tatusova and D. R. Maglott (2005). "NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins." Nucleic Acids Res **33 Database Issue**: D501-4.

Pupko, T., I. Pe'er, R. Shamir and D. Graur (2000). "A fast algorithm for joint reconstruction of ancestral amino acid sequences." Mol Biol Evol **17**(6): 890-6.

Raczniak, G., H. D. Becker, B. Min and D. Soll (2001). "A single amidotransferase forms asparaginyl-tRNA and glutaminyl-tRNA in *Chlamydia trachomatis*." J Biol Chem **276**(49): 45862-7.

Raghava, G. P., S. M. Searle, P. C. Audley, J. D. Barber and G. J. Barton (2003). "OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy." BMC Bioinformatics **4**(1): 47.

Rhee, S. Y., W. Beavis, T. Z. Berardini, G. Chen, D. Dixon, A. Doyle, M. Garcia-Hernandez, E. Huala, G. Lander, M. Montoya*, et al.* (2003). "The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community." Nucleic Acids Res **31**(1): 224-8.

Riechmann, V. and A. Ephrussi (2004). "Par-1 regulates bicoid mRNA localisation by phosphorylating Exuperantia." Development **131**(23): 5897-907.

Riechmann, V., G. J. Gutierrez, P. Filardo, A. R. Nebreda and A. Ephrussi (2002). "Par-1 regulates stability of the posterior determinant Oskar by phosphorylation." Nat Cell Biol **4**(5): 337-42.

Rigoutsos, I. and A. Floratos (1998). "Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm." Bioinformatics **14**(1): 55-67.

Rigoutsos, I., A. Floratos, C. Ouzounis, Y. Gao and L. Parida (1999). "Dictionary building via unsupervised hierarchical motif discovery in the sequence space of natural proteins." Proteins **37**(2): 264-77.

Rigoutsos, I., A. Floratos, L. Parida, Y. Gao and D. Platt (2000). "The emergence of pattern discovery techniques in computational biology." Metab Eng **2**(3): 159-77.

Rigoutsos, I., T. Huynh, A. Floratos, L. Parida and D. Platt (2002). "Dictionary-driven protein annotation." Nucleic Acids Res **30**(17): 3901-16.

Rokas, A., B. L. Williams, N. King and S. B. Carroll (2003). "Genome-scale approaches to resolving incongruence in molecular phylogenies." Nature **425**(6960): 798-804.

Romero, P. and P. Karp (2003). "PseudoCyc, a pathway-genome database for *Pseudomonas aeruginosa*." J Mol Microbiol Biotechnol **5**(4): 230-9.

Romero, P., J. Wagg, M. L. Green, D. Kaiser, M. Krummenacker and P. D. Karp (2005). "Computational prediction of human metabolic pathways from the complete human genome." Genome Biol **6**(1): R2.

Romero, P. R. and P. D. Karp (2004). "Using functional and organizational information to improve genome-wide computational prediction of transcription units on pathway-genome databases." Bioinformatics **20**(5): 709-17.

Rothschild, L. J. and R. L. Mancinelli (2001). "Life in extreme environments." Nature **409**(6823): 1092-101.

Ryden, L. G. and L. T. Hunt (1993). "Evolution of protein complexity: the blue copper-containing oxidases and related proteins." J Mol Evol **36**(1): 41-66.

Saito, H. (2001). "Histidine phosphorylation and two-component signaling in eukaryotic cells." Chem Rev **101**(8): 2497-509.

Salanoubat, M., S. Genin, F. Artiguenave, J. Gouzy, S. Mangenot, M. Arlat, A. Billault, P. Brottier, J. C. Camus, L. Cattolico*, et al.* (2002). "Genome sequence of the plant pathogen *Ralstonia solanacearum*." Nature **415**(6871): 497-502.

Sand, O., M. Gingras, N. Beck, C. Hall and N. Trun (2003). "Phenotypic characterization of overexpression or deletion of the *Escherichia coli* crcA, cspE and crcB genes." Microbiology **149**(Pt 8): 2107-17.

Sanger, F. and A. R. Coulson (1975). "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase." J Mol Biol **94**(3): 441-8.

Sanger, F., A. R. Coulson, G. F. Hong, D. F. Hill and G. B. Petersen (1982). "Nucleotide sequence of bacteriophage lambda DNA." J Mol Biol **162**(4): 729-73.

Sawada, H., L. D. Kuykendall and J. M. Young (2003). "Changing concepts in the systematics of bacterial nitrogen-fixing legume symbionts." J Gen Appl Microbiol **49**(3): 155-79.

Schaffer, A. A., L. Aravind, T. L. Madden, S. Shavirin, J. L. Spouge, Y. I. Wolf, E. V. Koonin and S. F. Altschul (2001). "Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements." Nucleic Acids Res **29**(14): 2994-3005.

Schulze-Makuch, D. and L. Irwin (2002). "Energy cycling and hypothetical organisms in Europa's ocean." Astrobiology **2**: 105-121.

Serres, M. H. and M. Riley (2000). "MultiFun, a multifunctional classification scheme for Escherichia coli K-12 gene products." Microb Comp Genomics **5**(4): 205-22.

Shah, I. and L. Hunter (1997). "Predicting enzyme function from sequence: a systematic appraisal." Proc Int Conf Intell Syst Mol Biol **5**: 276-83.

Shen-Orr, S. S., R. Milo, S. Mangan and U. Alon (2002). "Network motifs in the transcriptional regulation network of Escherichia coli." Nat Genet **31**(1): 64-8.

Shibuya, T. and I. Rigoutsos (2002). "Dictionary-driven prokaryotic gene finding." Nucleic Acids Res **30**(12): 2710-25.

Shin, H., K. Tsuda and B. Scholkopf (2004). "Protein Functional Class Prediction with a Combined Graph." Proceedings of the Korean Data Mining Conference: 200-219.

Showman, A. P. and R. Malhotra (1999). "The Galilean satellites." Science **286**(5437): 77-84.

Smith, M., V. Kunin, L. Goldovsky, A. J. Enright and C. A. Ouzounis (2005). "MagicMatch--cross-referencing sequence identifiers across databases." Bioinformatics **21**(16): 3429-30.

Smith, R. F. and T. F. Smith (1992). "Pattern-induced multi-sequence alignment (PIMA) algorithm employing secondary structure-dependent gap penalties for use in comparative protein modelling." Protein Eng **5**(1): 35-41.

Smith, T. F. and M. S. Waterman (1981). "Identification of common molecular subsequences." J Mol Biol **147**(1): 195-7.

Snel, B., P. Bork and M. A. Huynen (1999). "Genome phylogeny based on gene content." Nat Genet **21**(1): 108-10.

Snel, B., P. Bork and M. A. Huynen (2002). "Genomes in flux: the evolution of archaeal and proteobacterial gene content." Genome Res **12**(1): 17-25.

Snyder, L., F. Lovas, J. Hollis, D. Friedel, P. Jewell, A. Remijan, V. Ilyushin, E. Alekseev and S. Dyubko (2005). "A rigorous attempt to verify interstellar glycine." Astroph. J. **619**: 914-930.

Snyder, L. E. (1997). "The search for interstellar glycine." Orig Life Evol Biosph **27**(1-3): 115-33.

Spang, R., M. Rehmsmeier and J. Stoye (2002). "A novel approach to remote homology detection: jumping alignments." J Comput Biol **9**(5): 747-60.

Stevens, R. D., A. J. Robinson and C. A. Goble (2003). "myGrid: personalised bioinformatics on the information grid." Bioinformatics **19 Suppl 1**: i302-4.

Storm, C. E. and E. L. Sonnhammer (2002). "Automated ortholog inference from phylogenetic trees and calculation of orthology reliability." Bioinformatics **18**(1): 92-9.

Streit, W. R., R. A. Schmitz, X. Perret, C. Staehelin, W. J. Deakin, C. Raasch, H. Liesegang and W. J. Broughton (2004). "An evolutionary hot spot: the pNGR234b replicon of *Rhizobium* sp. strain NGR234." J Bacteriol **186**(2): 535-42.

Sun, T. Q., B. Lu, J. J. Feng, C. Reinhard, Y. N. Jan, W. J. Fantl and L. T. Williams (2001). "PAR-1 is a Dishevelled-associated kinase and a positive regulator of Wnt signalling." Nat Cell Biol **3**(7): 628-36.

Tatusov, R. L., N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya*, et al.* (2003). "The COG database: an updated version includes eukaryotes." BMC Bioinformatics **4**(1): 41.

Tatusov, R. L., E. V. Koonin and D. J. Lipman (1997). "A genomic perspective on protein families." Science **278**(5338): 631-7.

Taylor, W. R. (1986). "The classification of amino acid conservation." J Theor Biol **119**(2): 205-18.

Tekaia, F., A. Lazcano and B. Dujon (1999). "The genomic tree as revealed from whole proteome comparisons." Genome Res **9**(6): 550-7.

Teo, Y. M., X. Wang and Y. K. Ng (2004). "GLAD: a system for developing and deploying large-scale bioinformatics grid." Bioinformatics.

Teplova, M., V. Tereshko, R. Sanishvili, A. Joachimiak, T. Bushueva, W. F. Anderson and M. Egli (2000). "The structure of the yrdC gene product from *Escherichia coli* reveals a new fold and suggests a role in RNA binding." Protein Sci **9**(12): 2557-66.

Thompson, J. D., D. G. Higgins and T. J. Gibson (1994). "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." Nucleic Acids Res **22**(22): 4673-80.

Tomancak, P., F. Piano, V. Riechmann, K. C. Gunsalus, K. J. Kemphues and A. Ephrussi (2000). "A Drosophila melanogaster homologue of Caenorhabditis elegans par-1 acts at an early step in embryonic-axis formation." Nat Cell Biol **2**(7): 458-60.

Torsvik, V., L. Ovreas and T. F. Thingstad (2002). "Prokaryotic diversity--magnitude, dynamics, and controlling factors." Science **296**(5570): 1064-6.

Tsoka, S. and C. A. Ouzounis (2001). "Functional versatility and molecular diversity of the metabolic map of *Escherichia coli*." Genome Res **11**(9): 1503-10.

Tsoka, S. and C. A. Ouzounis (2003). "Metabolic database systems for the analysis of genome-wide function." Biotechnol Bioeng **84**(7): 750-5.

Van Dongen, S. (2000). Graph Clustering by Flow Simulation, University of Utrecht.

Venter, J. C., K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson*, et al.* (2004). "Environmental genome shotgun sequencing of the Sargasso Sea." Science **304**(5667): 66-74.

Vidal, R. A., D. Bahr, R. A. Baragiola and M. Peters (1997). "Oxygen on Ganymede: laboratory studies." Science **276**(5320): 1839-42.

Vinga, S. and J. Almeida (2003). "Alignment-free sequence comparison-a review." Bioinformatics **19**(4): 513-23.

von Mering, C., M. Huynen, D. Jaeggi, S. Schmidt, P. Bork and B. Snel (2003). "STRING: a database of predicted functional associations between proteins." Nucleic Acids Res **31**(1): 258-61.

von Mering, C., L. J. Jensen, B. Snel, S. D. Hooper, M. Krupp, M. Foglierini, N. Jouffre, M. A. Huynen and P. Bork (2005). "STRING: known and predicted protein-protein associations, integrated and transferred across organisms." Nucleic Acids Res **33 Database Issue**: D433-7.

Vulic, M., F. Dionisio, F. Taddei and M. Radman (1997). "Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria." Proc Natl Acad Sci U S A **94**(18): 9763-7.

Wang, L., K. Zhang and L. Zhang (2001). "Perfect phylogenetic networks with recombination." J Comput Biol **8**(1): 69-78.

Wang, Y., L. Y. Geer, C. Chappey, J. A. Kans and S. H. Bryant (2000). "Cn3D: sequence and structure views for Entrez." Trends Biochem Sci **25**(6): 300-2.

Wexler, M., F. Sargent, R. L. Jack, N. R. Stanley, E. G. Bogsch, C. Robinson, B. C. Berks and T. Palmer (2000). "TatD is a cytoplasmic protein with DNase activity. No requirement for TatD family proteins in sec-independent protein export." J Biol Chem **275**(22): 16717-22.

White, O., J. A. Eisen, J. F. Heidelberg, E. K. Hickey, J. D. Peterson, R. J. Dodson, D. H. Haft, M. L. Gwinn, W. C. Nelson, D. L. Richardson*, et al.* (1999). "Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1." Science **286**(5444): 1571-7.

Woese, C. (1998). "The universal ancestor." Proc Natl Acad Sci U S A **95**(12): 6854-9.

Woese, C. R. (2002). "On the evolution of cells." Proc Natl Acad Sci U S A **99**(13): 8742-7.

Wu, J., S. Kasif and C. DeLisi (2003). "Identification of functional links between genes using phylogenetic profiles." Bioinformatics **19**(12): 1524-30.

Yayanos, A. A. (1995). "Microbiology to 10,500 meters in the deep sea." Annu Rev Microbiol **49**: 777-805.

Yeh, I., T. Hanekamp, S. Tsoka, P. D. Karp and R. B. Altman (2004). "Computational analysis of *Plasmodium falciparum* metabolism: organizing genomic information to facilitate drug discovery." Genome Res **14**(5): 917-24.

Zhang, Y. X., K. Perry, V. A. Vinci, K. Powell, W. P. Stemmer and S. B. del Cardayre (2002). "Genome shuffling leads to rapid phenotypic improvement in bacteria." Nature **415**(6872): 644-6.

Zuckerkandl, E. and L. Pauling (1965). "Molecules as documents of evolutionary history." J Theor Biol **8**(2): 357-66.

APPENDIX A

Papers published / in press / submitted

Darzentas, N. and C. Ouzounis (2005). "Bridging SCOP Folds and Superfamilies with structurally equivalent sequence patterns." **[submitted]**.

Darzentas, N., I. Rigoutsos and C. A. Ouzounis (2005). "Sensitive detection of sequence similarity using combinatorial pattern discovery: a challenging study of two distantly related protein families." Proteins **61**(4): 926-37.

Goldovsky, L., P. Janssen, D. Ahren, B. Audit, I. Cases, N. **Darzentas**, A. J. Enright, N. Lopez-Bigas, J. M. Peregrin-Alvarez, M. Smith*, et al.* (2005). "CoGenT++: an extensive and extensible data environment for computational genomics." Bioinformatics **21**(19): 3806-10.

Janssen, P., B. Audit, I. Cases, N. **Darzentas**, L. Goldovsky, V. Kunin, N. Lopez-Bigas, J. M. Peregrin-Alvarez, J. B. Pereira-Leal, S. Tsoka*, et al.* (2003). "Beyond 100 genomes." Genome Biol **4**(5): 402.

Janssen, P., L. Goldovsky, V. Kunin, N. **Darzentas** and C. A. Ouzounis (2005). "Genome coverage, literally speaking. The challenge of annotating 200 genomes with 4 million publications." EMBO Rep **6**(5): 397-9.

Karp, P. D., C. A. Ouzounis, C. Moore-Kochlacs, L. Goldovsky, P. Kaipa, D. Ahren, S. Tsoka, N. **Darzentas**, V. Kunin and N. Lopez-Bigas (2005). "Expansion of the BioCyc collection of pathway/genome databases to 160 genomes." Nucleic Acids Res **33**(19): 6083-9.

Kunin, V., L. Goldovsky, N. **Darzentas** and C. A. Ouzounis (2005). "The net of life: reconstructing the microbial phylogenetic network." Genome Res **15**(7): 954-9.

Ouzounis, C., V. Kunin, N. **Darzentas** and L. Goldovsky (2005). "A minimal estimate for the gene content of the last universal common ancestor - exobiology from a terrestrial perspective." Res. Microbiol. **[in press]**.

APPENDIX B

Table from Section 5.2.4

A truly minimal estimate for the gene content of the last universal common ancestor, obtained by three different tree construction methods and the inclusion or not of Eukaryotes. In total, there are 669 ortholog families distributed in 561 functional annotation descriptions, including 52 which remain uncharacterized.

case	count	consensus annotation
1	1	1-deoxyxylulose-5-phosphate synthase
2	1	10 chaperonin GroES kDa
3	1	2,3-phosphoglycerate bisphosphoglycerate-mutase independent
4	1	2-4-1,isomerase 7
5	1	2-dehydro-3-deoxyphosphooctonate aldolase [chorismate]
6	1	2-isopropylmalate synthase
7	1	2-oxoacid ferredoxin oxidoreductase, beta subunit
8	1	3,GTP 4-cyclohydrolase dihydroxy-II 2-butanone phosphate synthase
9	1	3-DNA-methyladenine [glycosidase] glycosylase
10	1	3-demethylubiquinone-9 methyltransferase
11	1	3-hydroxyacyl-CoA dehydrogenase
12	1	3-hydroxyisobutyrate dehydrogenase
13	1	3-isocitrate isopropylmalate dehydrogenase
14	1	3-isopropylmalate dehydratase large subunit
15	1	3-isopropylmalate dehydratase small subunit
16	1	3-methyl-2-oxobutanoate hydroxymethyltransferase
17	1	3-octaprenyl-4-decarboxylase hydroxybenzoate carboxy-lyase
18	1	3-oxoacyl-acyl-carrier reductase
19	1	3-oxoacyl-acyl-carrier-synthase III
20	1	3-phosphoshikimate 1-carboxyvinyltransferase
21	1	30S ribosomal S10
22	1	30S ribosomal S11
23	1	30S ribosomal S12

24	1	30S ribosomal S13
25	1	30S ribosomal S17
26	1	30S ribosomal S19
27	1	30S ribosomal S3
28	1	30S ribosomal S4
29	1	30S ribosomal S5
30	1	30S ribosomal S7
31	1	30S ribosomal S8
32	1	30S ribosomal S9
33	1	4-aminobutyrate aminotransferase
34	1	4-hydroxybenzoate
35	1	4-hydroxybenzoate octaprenyltransferase
36	1	5-methyltetrahydrofolate-homocysteine methyltransferase
37	1	50S ribosomal L1
38	1	50S ribosomal L11
39	1	50S ribosomal L13
40	1	50S ribosomal L14
41	1	50S ribosomal L2
42	1	50S ribosomal L22
43	1	50S ribosomal L23
44	1	50S ribosomal L5
45	1	50S ribosomal L6
46	1	6-phosphogluconate dehydrogenase decarboxylating
47	1	6-pyruvoyl [tetrahydropterin] tetrahydrobiopterin synthase
48	1	60 kDa chaperonin GroEL
49	1	A G-specific adenine glycosylase
50	1	ABC cobalt transport permease
51	1	ABC periplasmic transporter, binding
52	1	ABC transporter, periplasmic [dipeptide] binding
53	1	ADP-heptose synthase
54	1	ATP phosphoribosyltransferase
55	1	ATP synthase [F1] alpha subunit
56	1	ATP synthase [F1] beta chain subunit
57	1	ATP-DNA dependent II helicase

58	1	ATP-NAD kinase
59	1	ATP-dependent DNA helicase RecQ
60	1	ATP-dependent Lon protease La
61	1	ATP-dependent RNA helicase
62	1	ATP-dependent helicase
63	1	ATP-dependent protease La [lon]
64	1	ATPase
65	1	CTP synthase
66	1	D-2-dehydrogenase
67	1	D-3-phosphoglycerate dehydrogenase
68	1	DHH
69	1	DNA [topoisomerase] gyrase subunit B
70	1	DNA gyrase subunit A
71	1	DNA helicase
72	1	DNA ligase NAD
73	1	DNA mismatch MutS repair
74	1	DNA mismatch repair MutL
75	1	DNA pantothenate flavo [dfp] metabolism
76	1	DNA polymerase III gamma subunit and tau subunits
77	1	DNA polymerase [bacteriophage]
78	1	DNA repair RadA
79	1	DNA topoisomerase I
80	1	DNA-damage-inducible P
81	1	DNA-stress binding
82	1	DnaK chaperone heat shock [HSP70]
83	1	GGDEF [domain]
84	1	GMP synthase glutamine
85	1	GTP-binding HflX
86	1	GTP-binding [GTP1] Obg
87	1	H Na antiporter
88	1	HAM1
89	1	HIT family
90	1	HesB
91	1	KpsF

92	1	L-fuculose-aldolase phosphate
93	1	L-isoaspartate O-methyltransferase [pcm]
94	1	L-lactate malate dehydrogenase
95	1	L-lactate permease
96	1	Mrp binding
97	1	MutT nudix family
98	1	N-acetyl-gamma-glutamyl-phosphate reductase
99	1	NADH [flavin] oxidoreductase
100	1	NADH [ubiquinone] dehydrogenase I chain B subunit
101	1	NADH [ubiquinone] dehydrogenase I chain subunit
102	1	NADH [ubiquinone] dehydrogenase I chain subunit A
103	1	NADH [ubiquinone] dehydrogenase I chain subunit L
104	1	NADH dehydrogenase
105	1	NADH dehydrogenase I chain C subunit
106	1	NADH dehydrogenase I chain subunit D
107	1	NADH dehydrogenase I chain subunit H
108	1	NADH dehydrogenase I chain subunit K
109	1	NADH dehydrogenase I chain subunit M
110	1	NADH dehydrogenase I chain subunit N
111	1	NADH dehydrogenase I oxidoreductase chain J subunit
112	1	NADH reductase
113	1	NifU
114	1	O-acetylhomoserine sulfhydrylase
115	1	O-acetyltransferase
116	1	O-methyltransferase
117	1	O-sialoglyco endopeptidase
118	1	O-succinylbenzoic acid-CoA ligase
119	1	PmbA
120	1	RNA methyltransferase
121	1	Sec-independent translocase TatC
122	1	Sir2 family
123	1	Sua5
124	1	TPR
125	1	TatD

126	1	Trk potassium uptake
127	1	TrkA domain
128	1	UDP-N-acetylglucosamine 2-epimerase
129	1	UDP-N-acetylglucosamine pyrophosphorylase
130	1	UDP-glucose 4-epimerase
131	1	UDP-glucose 6-dehydrogenase
132	1	UTP-glucose-1-phosphate thymidylyltransferase uridylyltransferase
133	1	V-type ATP ATPase synthase subunit [vacuolar] D
134	1	acetolactate synthase large subunit
135	1	acetolactate synthase small subunit
136	1	acetyl-CoA carboxylase carboxyl transferase beta subunit
137	1	acetyl-biotin CoA carboxylase
138	1	acetyl-coenzyme CoA A synthetase
139	1	acetylglutamate kinase
140	1	acetylornithine aminotransferase
141	1	aconitate hydratase
142	1	acyl-CoA dehydrogenase
143	1	acylphosphatase
144	1	adenine phosphoribosyltransferase [apt]
145	1	adenosylhomocysteinase S-hydrolase
146	1	adenylate kinase [adk]
147	1	adenylosuccinate lyase
148	1	adenylosuccinate synthetase
149	1	agmatinase
150	1	alanyl-tRNA synthetase
151	1	alcohol dehydrogenase
152	1	aldehyde semialdehyde dehydrogenase
153	1	alkyl hydroperoxide reductase
154	1	alpha-glucosidase
155	1	amidophosphoribosyltransferase
156	1	amidotransferase glutamine hisH
157	1	amino acid ABC transporter permease
158	1	amino acid ABC transporter, ATP-binding
159	1	amino acid permease

160	1	aminomethyltransferase glycine cleavage system T
161	1	aminopeptidase N
162	1	aminopeptidase Pro [dipeptidase]
163	1	ammonium transporter
164	1	ankyrin [repeat]
165	1	anthranilate para-aminobenzoate synthase component I II
166	1	anthranilate phosphoribosyltransferase
167	1	argininosuccinate lyase
168	1	argininosuccinate synthase
169	1	arginyl-tRNA synthetase
170	1	arsenate reductase
171	1	arsenical pump membrane
172	1	asparaginase L
173	1	asparagine synthetase glutamine
174	1	aspartate aminotransferase
175	1	aspartate carbamoyltransferase catalytic
176	1	aspartate-semialdehyde dehydrogenase [asd]
177	1	aspartokinase aspartate kinase
178	1	aspartyl-tRNA synthetase
179	1	bacterioferritin comigratory
180	1	band 7
181	1	beta-lactamase
182	1	biotin BioY family
183	1	biotin [operon] birA acetyl-ligase CoA-carboxylase synthetase
184	1	biotin synthase synthetase
185	1	branched-chain amino acid aminotransferase
186	1	carbamoyl-phosphate synthase large subunit chain
187	1	carbamoyl-phosphate synthase small subunit chain
188	1	carboxymethylenebutenolidase hydrolase
189	1	catalase peroxidase
190	1	cation-copper-transporting P-ATPase type
191	1	cation-transporting P-ATPase type
192	1	cell division FtsH
193	1	cell division FtsZ

194	1	cell division inhibitor
195	1	cell signal division recognition particle [receptor/docking] FtsY
196	1	chaperone DnaJ
197	1	chemotaxis binding CheW
198	1	chemotaxis histidine CheA kinase
199	1	chemotaxis methyltransferase CheR
200	1	chloride channel
201	1	chlorohydrolase
202	1	chorismate mutase prephenate dehydratase
203	1	chorismate synthase
204	1	chromosome segregation SMC
205	1	citrate synthase
206	1	cob I alamin adenosyltransferase
207	1	cobalamin 5-phosphate synthase
208	1	cobalamin biosynthesis CobD CbiB
209	1	cobalamin magnesium CobN chelatase
210	1	cobalamin synthesis [P47K]
211	1	cobalt ABC transport transporter, ATP-binding
212	1	cobyric acid synthase
213	1	cobyrinic acid a, c-diamide synthase
214	1	cold shock [domain]
215	1	crcB
216	1	cystathionine gamma-synthase lyase
217	1	cysteine desulfurase
218	1	cysteine synthase
219	1	cysteinyl-tRNA synthetase
220	1	cytidylate kinase [cmk]
221	1	cytochrome P450
222	1	cytochrome d ubiquinol oxidase subunit I
223	1	cytochrome d ubiquinol oxidase subunit II
224	1	dTDP-4-dehydrorhamnose 3,5-epimerase
225	1	dTDP-4-dehydrorhamnose reductase
226	1	dTDP-glucose 4,6-dehydratase
227	1	deaminase

228	1	dehydrogenase reductase
229	1	delta-aminolevulinic acid dehydratase [porphobilinogen]
230	1	deoxycytidine triphosphate deaminase [dcd]
231	1	deoxyribodipyrimidine photolyase
232	1	dependent NAD synthetase
233	1	diaminopimelate decarboxylase
234	1	dihydrodipicolinate reductase
235	1	dihydrodipicolinate synthase
236	1	dihydrolipoamide [S] 2-acetyltransferase succinyltransferase component [of] dehydrogenase [complex] E2
237	1	dihydrolipoamide dehydrogenase [E3]
238	1	dihydroorotase
239	1	dihydroorotate dehydrogenase
240	1	dihydropteroate synthase
241	1	dihydroxy-acid dehydratase
242	1	dimethyladenosine transferase
243	1	dipeptide oligopeptide ABC transport transporter, permease
244	1	diphosphate synthase [octaprenyl]
245	1	electron transfer flavo alpha subunit
246	1	electron transfer flavo beta subunit
247	1	electron transfer flavo-ubiquinone oxidoreductase
248	1	endonuclease III [nth]
249	1	endonuclease IV
250	1	enolase [eno]
251	1	enoyl-CoA hydratase
252	1	excinuclease ABC subunit A
253	1	excinuclease ABC subunit B
254	1	excinuclease ABC subunit C
255	1	exodeoxyribonuclease III
256	1	exopolyphosphatase
257	1	export membrane SecD
258	1	export membrane SecF
259	1	exsB
260	1	ferredoxin

261	1	ferredoxin dioxygenase [Rieske] subunit
262	1	ferric uptake regulator
263	1	ferrochelatase
264	1	fibronectin fibrinogen-binding
265	1	folylpolyglutamate synthase dihydrofolate
266	1	formate dehydrogenase reductase subunit
267	1	fructokinase kinase
268	1	fructose-6-1,bisphosphatase [fbp]
269	1	fructose-bisphosphate aldolase class I
270	1	fumarate hydratase class II
271	1	galactokinase
272	1	gamma-glutamyl phosphate reductase
273	1	glucokinase [ROK] kinase
274	1	glucosamine-fructose-6-phosphate aminotransferase isomerizing
275	1	glucose-1-phosphate [adenylyltransferase/guanyltransferase]
276	1	glucose-6-phosphate isomerase [pgi]
277	1	glutamate 5-kinase
278	1	glutamate N-acetyltransferase
279	1	glutamate specific dehydrogenase
280	1	glutamate synthase large subunit
281	1	glutamate-1-semialdehyde 2,aminomutase
282	1	glutamine amidotransferase
283	1	glutamine amidotransferase [SNO]
284	1	glutamyl-glutaminyl-tRNA synthetase
285	1	glutamyl-tRNA Gln amidotransferase subunit A
286	1	glutamyl-tRNA Gln amidotransferase subunit B
287	1	glutamyl-tRNA Gln amidotransferase subunit C
288	1	glutamyl-tRNA reductase
289	1	glutathione-Na regulated potassium-H efflux antiporter system
290	1	glycerol kinase
291	1	glycerol-3-phosphate dehydrogenase
292	1	glycerophosphoryl diester phosphodiesterase
293	1	glycine ABC betaine transporter transport permease
294	1	glycine betaine [carnitine] ABC transport transporter permease

295	1	glycine cleavage system H
296	1	glycine dehydrogenase [decarboxylating] cleavage system P
297	1	glycogen phosphorylase
298	1	glycyl-tRNA synthetase
299	1	heat GrpE shock
300	1	heat shock HtpX
301	1	helicase
302	1	helicase [SNF2]
303	1	hemolysin
304	1	histidine ammonia-lyase
305	1	histidinol dehydrogenase
306	1	histidinol-phosphate aminotransferase
307	1	histidyl-tRNA synthetase
308	1	histone deacetylase
309	1	homoserine dehydrogenase
310	1	homoserine kinase
311	1	hydrolase phosphatase
312	1	hydrolase, dehalogenase-like
313	1	hydroxyacylglutathione hydrolase
314	1	imidazoleglycerol-phosphate dehydratase [histidinol]
315	1	indole-3-glycerol phosphate synthase
316	1	inhibitor
317	1	inorganic pyrophosphatase [ppa]
318	1	inosine-5-monophosphate dehydrogenase
319	1	integrase recombinase
320	1	iron ABC transporter, ATP-transport binding
321	1	iron [compound] ABC transport transporter, permease
322	1	iron-dependent repressor
323	1	iron-sulfur
324	1	isoleucyl-tRNA synthetase
325	1	ketol-acid reductoisomerase
326	1	leucyl-tRNA synthetase
327	1	lipase esterase
328	1	lipoic acid synthetase

329	1	lysophospholipase
330	1	lysyl-tRNA synthetase
331	1	maf
332	1	magnesium Mg2 transporter mgtE
333	1	magnesium and cobalt transport
334	1	malate [oxidoreductase] dependent malic enzyme
335	1	mannose-1-phosphate guanylyltransferase isomerase
336	1	mechanosensitive [ion] channel
337	1	membrane rhomboid
338	1	metallo-beta-lactamase
339	1	methionine aminopeptidase
340	1	methionine aminopeptidase [map]
341	1	methionyl-tRNA synthetase
342	1	methylated-DNA-cysteine methyltransferase
343	1	methylenetetrahydrofolate folD [bifunctional] dehydrogenase methenyltetrahydrofolate cyclohydrolase
344	1	methylmalonyl-CoA mutase subunit
345	1	molybdenum ABC transport transporter, permease
346	1	molybdenum cofactor biosynthesis A
347	1	molybdenum cofactor biosynthesis C
348	1	molybdenum molybdopterin cofactor biosynthesis [mog]
349	1	molybdopterin biosynthesis MoeA
350	1	molybdopterin biosynthesis MoeB
351	1	molybdopterin converting factor, subunit 2
352	1	multidrug [efflux] resistance transporter
353	1	mutator MutT [nudix]
354	1	nicotinate-nucleotide pyrophosphorylase [carboxylating]
355	1	nitrogen regulatory [PII] P-II
356	1	nucleoside diphosphate kinase [ndk]
357	1	oligoendopeptidase F
358	1	oligopeptide ABC transport [system] transporter, permease
359	1	oligopeptide ABC transport transporter permease
360	1	oligopeptide ABC transport transporter, permease
361	1	oligopeptide ABC transporter, binding

362	1	oligopeptide [dipeptide] ABC transport transporter, ATP-binding
363	1	ornithine carbamoyltransferase
364	1	orotate phosphoribosyltransferase
365	1	oxidoreductase [Gfo]
366	1	oxidoreductase aldo keto reductase
367	1	oxidoreductase oxidase
368	1	oxidoreductase short chain dehydrogenase
369	1	peptidase
370	1	peptidyl-FKBP-type prolyl cis-trans isomerase
371	1	peptidyl-prolyl cis-trans isomerase [cyclophilin]
372	1	periplasmic divalent cation tolerance cutA
373	1	phenazine biosynthesis PhzF
374	1	phenylalanyl-tRNA synthetase alpha subunit chain
375	1	phenylalanyl-tRNA synthetase beta chain subunit
376	1	phosphate ABC periplasmic transporter, binding
377	1	phosphate ABC transport transporter system permease
378	1	phosphate ABC transporter transport permease
379	1	phosphate ABC transporter, periplasmic binding
380	1	phosphate ABC transporter, transport ATP-binding
381	1	phosphate [permease] transporter
382	1	phosphate hisF [imidazoleglycerol] cyclase synthase subunit
383	1	phosphate phosphotransacetylase acetyltransferase [pta]
384	1	phosphate transport system regulator regulatory PhoU
385	1	phosphatidylserine decarboxylase
386	1	phosphoadenosine phosphosulfate [sulfotransferase] reductase
387	1	phosphoenolpyruvate pyruvate, phosphate dikinase
388	1	phosphoglucomutase phosphomannomutase mutase
389	1	phosphoglycerate kinase [pgk]
390	1	phosphomannomutase phosphoglucomutase
391	1	phosphomethylpyrimidine kinase
392	1	phosphoribosyl-AMP cyclohydrolase ATP [pyrophosphohydrolase]
393	1	phosphoribosylamine-glycine ligase
394	1	phosphoribosylaminoimidazole carboxylase ATPase subunit
395	1	phosphoribosylaminoimidazole carboxylase catalytic subunit

396	1	phosphoribosylaminoimidazole-succinocarboxamide synthase [SAICAR]
397	1	phosphoribosylaminoimidazolecarboxamide formyltransferase IMP cyclohydrolase [bifunctional]
398	1	phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase
399	1	phosphoribosylformylglycinamidine cyclo-ligase
400	1	phosphoribosylformylglycinamidine synthase I
401	1	phosphoribosylformylglycinamidine synthase II
402	1	phosphoribosylglycinamide formyltransferase
403	1	phosphoserine aminotransferase
404	1	phosphoserine phosphatase
405	1	phytoene dehydrogenase
406	1	pirin
407	1	polysaccharide biosynthesis
408	1	porphobilinogen deaminase
409	1	potassium channel
410	1	potassium uptake
411	1	potassium-transporting ATPase A chain
412	1	potassium-transporting ATPase B chain
413	1	potassium-transporting ATPase C chain
414	1	pre translocase SecY subunit
415	1	precorrin-2 C20-methyltransferase
416	1	precorrin-3B C17-3 methyltransferase methylase
417	1	precorrin-4 C11-methyltransferase methylase
418	1	precorrin-6Y [C5] decarboxylase methyltransferase
419	1	precorrin-8X methylmutase isomerase
420	1	prolyl-tRNA synthetase
421	1	protease
422	1	protease IV
423	1	protoheme IX farnesyltransferase
424	1	proton sodium-dicarboxylate glutamate symporter
425	1	protoporphyrinogen HemK oxidase
426	1	protoporphyrinogen oxidase
427	1	pterin-4-alpha-carbinolamine dehydratase

428	1	purine nucleoside phosphorylase
429	1	pyridoxine biosynthesis
430	1	pyrroline-5-carboxylate reductase
431	1	pyruvate 2-ferredoxin oxidoreductase, alpha subunit
432	1	pyruvate dehydrogenase E1 component, alpha subunit
433	1	pyruvate dehydrogenase E1 component, beta subunit
434	1	pyruvate kinase
435	1	queuine tRNA-guanine ribosyltransferase transglycosylase [tgt]
436	1	quinolinate synthetase A
437	1	quinone oxidoreductase
438	1	radical activating enzyme
439	1	ral secretion pathway E
440	1	recA
441	1	reductase
442	1	resistance
443	1	riboflavin 6,7-dimethyl-synthase beta 8-ribityllumazine
444	1	riboflavin biosynthesis RibD deaminase reductase
445	1	riboflavin synthase alpha chain
446	1	ribonuclease HII
447	1	ribonuclease PH
448	1	ribonuclease R
449	1	ribonucleoside-diphosphate reductase alpha subunit chain
450	1	ribonucleoside-diphosphate reductase beta chain subunit
451	1	ribonucleoside-diphosphate ribonucleotide reductase alpha chain
452	1	ribose 5-phosphate isomerase
453	1	ribose-phosphate pyrophosphokinase
454	1	ribosomal L11 methyltransferase
455	1	ribosomal-acetyltransferase alanine
456	1	ribosomal-alanine acetyltransferase N
457	1	sensor histidine kinase
458	1	serine [O] acetyltransferase
459	1	serine hydroxymethyltransferase
460	1	serine protease
461	1	serine protease [DO]

462	1	serine threonine kinase
463	1	seryl-tRNA synthetase
464	1	shikimate 5-dehydrogenase
465	1	signal recognition particle [SRP54/ffh]
466	1	siroheme synthase
467	1	small heat shock [Hsp20]
468	1	sodium [solute] symporter
469	1	sodium-dependent transporter
470	1	spermidine putrescine ABC transport transporter permease
471	1	spermidine synthase
472	1	succinate dehydrogenase [fumarate] flavo subunit
473	1	succinate dehydrogenase [fumarate] iron-sulfur
474	1	succinyl-CoA synthetase alpha chain subunit
475	1	succinyl-CoA synthetase beta chain subunit
476	1	succinyl-diaminopimelate desuccinylase
477	1	sugar ABC transporter binding
478	1	sugar ABC transporter, ATP-binding
479	1	sulfate adenylyltransferase
480	1	sulfite
481	1	sun
482	1	superoxide dismutase [Mn/Fe]
483	1	tRNA pseudouridine [55] synthase B
484	1	tRNA pseudouridylate pseudouridine synthase A
485	1	terminal protease
486	1	thermostable carboxypeptidase
487	1	thiamine biosynthesis ThiC
488	1	thiamine biosynthesis ThiI
489	1	thiamine-monophosphate kinase
490	1	thiamine-thiamin-phosphate pyrophosphorylase
491	1	thioesterase
492	1	thioredoxin reductase
493	1	thiosulfate sulfurtransferase
494	1	threonine dehydratase
495	1	threonine synthase dehydratase

496	1	threonyl-tRNA synthetase
497	1	thymidylate kinase [tmk]
498	1	thymidylate synthase
499	1	tldD
500	1	transcription-repair coupling factor [mfd]
501	1	transcriptional leucine-regulator [lrp] regulatory AsnC family
502	1	transcriptional regulator aminotransferase family
503	1	transketolase
504	1	translation elongation factor G
505	1	translation elongation factor [EF] Tu
506	1	translation initiation factor IF-2
507	1	translation initiation factor SUI1
508	1	translation initiation factor [aIF] eIF-2B alpha subunit
509	1	transport system kinase
510	1	triosephosphate isomerase
511	1	trp repressor binding
512	1	tryptophan synthase alpha subunit chain
513	1	tryptophan synthase beta chain subunit
514	1	tryptophanyl-tRNA synthetase
515	1	type I restriction enzyme modification system R subunit
516	1	type V-ATP synthase ATPase subunit I
517	1	ubiquinone biosynthesis
518	1	ubiquinone menaquinone biosynthesis methyltransferase
519	1	undecaprenyl pyrophosphate synthetase synthase
520	1	undecaprenyl synthase pyrophosphate
521	1	uracil phosphoribosyltransferase [upp]
522	1	urease accessory UreG hydrogenase [expression]
523	1	uridine kinase [udk]
524	1	uridylate kinase
525	1	uroporphyrin-III C-synthase methyltransferase [uroporphyrinogen]
526	1	uroporphyrinogen decarboxylase
527	1	valyl-tRNA synthetase [valine]
528	1	xanthine-guanine phosphoribosyltransferase
529	2	ABC transporter

530	2	ABC transporter, ATP-binding
531	2	DNA-directed RNA polymerase beta subunit
532	2	Na H antiporter
533	2	acetyl-CoA acetyltransferase thiolase
534	2	acetyltransferase
535	2	cation efflux
536	2	efflux
537	2	glutamine synthetase
538	2	inositol monophosphatase
539	2	long-chain-fatty-acid-CoA ligase
540	2	methyltransferase
541	2	oligopeptide ABC transporter transport ATP-binding
542	2	peptide methionine sulfoxide reductase
543	2	phosphoglycerate mutase
544	2	polysaccharide
545	2	thioredoxin
546	2	transcriptional regulator ArsR family
547	2	transporter
548	2	universal stress
549	3	ABC transporter ATP-transport binding
550	3	GTP-binding
551	3	aminotransferase
552	3	hydrolase
553	3	oxidoreductase
554	3	transferase
555	3	two-component response regulator
556	3	two-component sensor histidine kinase
557	5	ABC transporter permease
558	5	membrane
559	6	glycosyl transferase
560	9	ABC transporter ATP-binding
561	52	NO CONSENSUS ANNOTATION

Columns:

- **case: index for the 561 annotation classes to facilitate discussion in text;**
- **count: number of instances in all six derived gene content datasets;**
- **consensus annotation: annotation derived by consensus pattern matching of all putative orthologs and their database homologs.**

Notes:

- **sorted by number of putative orthologs and annotation;**
- **red strings represent unknown function;**
- **green strings represent metabolic enzymes.**

APPENDIX C

Table from Section 7.3.5

The 64 genomes used in the functional class prediction experiment. Column 1 is the name of the organism, column 2 is the number of proteins classified as ‘Unknown’/’W’, column 3 is the corresponding percentage.

Organism (and chromosome where applicable)	**Proteins classified as Unknown**	**%**
Haemophilus influenzae Rd	674	39.6%
Mycoplasma genitalium	184	38.5%
Synechocystis sp. PCC	2334	73.8%
Methanococcus jannaschii	1249	73.0%
Mycoplasma pneumoniae	296	44.5%
Helicobacter pylori 26695	1105	70.9%
Escherichia coli K12	1798	42.0%
Methanothermobacter thermautotrophicus	1489	80.0%
Bacillus subtilis	2547	62.3%
Archaeoglobus fulgidus	1998	83.0%
Borrelia burgdorferi	629	74.9%
Aquifex aeolicus	1089	72.2%
Pyrococcus horikoshii	1756	85.3%
Mycobacterium tuberculosis H37Rv	1481	72.4%
Treponema pallidum	803	78.1%
Chlamydia trachomatis	620	71.6%
Rickettsia prowazekii	541	65.3%
Helicobacter pylori J99	1051	70.6%
Chlamydophila pneumoniae CWL029	794	76.0%
Aeropyrum pernix	2464	91.5%
Thermotoga maritima	1357	73.5%
Deinococcus radiodurans chromosome 1	1676	65.2%
Deinococcus radiodurans chromosome 2	221	62.1%
Thermoplasma volcanium	934	82.9%
Pyrococcus abyssi	1429	81.0%
Campylobacter jejuni	950	58.2%
Neisseria meningitidis MC58	1098	54.2%

Chlamydophila pneumoniae AR39	843	76.8%
Chlamydia muridarum	651	71.9%
Neisseria meningitidis Z2491	1130	54.7%
Bacillus halodurans	2410	59.5%
Chlamydophila pneumoniae J138	810	75.9%
Xylella fastidiosa	1787	64.7%
Vibrio cholerae chromosome 1	1400	51.3%
Vibrio cholerae chromosome 2	726	67.2%
Pseudomonas aeruginosa	3812	68.5%
Buchnera sp. APS	142	25.2%
Thermoplasma acidophilum	931	63.0%
Ureaplasma urealyticum	343	56.2%
Halobacterium sp. NRC 1	1443	70.6%
Mesorhizobium loti	5268	78.2%
Mycobacterium leprae	1091	68.1%
Escherichia coli O157:H7	2831	53.1%
Pasteurella multocida	977	48.5%
Caulobacter vibrioides	2905	77.7%
Streptococcus pyogenes M1	1174	69.2%
Staphylococcus aureus subsp.	1780	69.9%
Lactococcus lactis subsp.	1635	72.2%
Mycoplasma pulmonis	580	74.3%
Sulfolobus solfataricus	2585	86.9%
Streptococcus pneumoniae	1523	72.8%
Sinorhizobium meliloti	2248	68.0%
Clostridium acetobutylicum	2766	76.3%
Agrobacterium tumefaciens str. chromosome 1	2021	74.5%
Agrobacterium tumefaciens str. chromosome 2	1311	71.8%
Rickettsia conorii	1038	75.6%
Sulfolobus tokodaii	2418	85.6%
Yersinia pestis	2064	55.6%
Salmonella enterica subsp.	2315	52.8%
Listeria monocytogenes	1945	68.3%
Listeria innocua	2107	71.0%
Schizosaccharomyces pombe chromosome 1	1331	64.1%
Schizosaccharomyces pombe chromosome 2	1123	66.4%
Schizosaccharomyces pombe chromosome 3	586	69.8%

APPENDIX D

Table from Section 9.4

The complete set of significant connections made by patterns. Columns are (numbers in parentheses refer to other Columns):

1 Light grey: within Fold across Superfamilies – Dark grey: across Folds;

2 ASTRAL SCOP Domain;

3 PDB chain and ASTRAL SCOP sequence coordinates information for Domain (2);

4 ASTRAL SCOP Family of Domain (2);

5 DALI PDB sequence coordinates of pattern (14) on Domain (2);

6 DALI PDB sequence coordinates from DALI structural alignment (7) on Domain (2);

7 DALI structural alignment ID;

8 DALI structural alignment Z-score for (7) – White: 0-5, Light grey: 5-10, Dark grey: >10;

9 DALI PDB sequence coordinates of DALI structural alignment (7) on Domain (11);

10 DALI PDB sequence coordinates for pattern (14) on Domain (11);

11 ASTRAL SCOP Domain;

12 PDB chain and ASTRAL SCOP sequence coordinates information for Domain (11);

13 ASTRAL SCOP Family of Domain (11);

14 Structurally similar sequence pattern responsible for the connection.

Columns 2&11, 3&12, 4&13, 5&9, 6&10 contain equivalent information for the two connected Domains.

1	2	3	4	5	6	7	8	9	10	11	12	13	14
░	d1gai__	(-)	a.102.1.1	188-198	182-199	718646	9	549-559	543-560	d2sqca1	(A:8-36,A:308-630)	a.102.4.2	[ITV]A...[KMR]AL[ITV].G
░	d1fp3a_	(A:)	a.102.1.3	2-9	2-8	504281	2.8	26-33	26-32	d1cb8a1	(A:26-335)	a.102.3.2	EK...TLQ
░	d1nc5a_	(A:)	a.102.1.6	338-350	339-350	2391984	6.1	484-496	485-496	d2sqca1	(A:8-36,A:308-630)	a.102.4.2	[ITV]..[DLN].G.GA.[ITV].A
░	d1dl2a_	(A:)	a.102.2.1	126-135	123-142	504324	14.3	143-152	140-159	d1fp3a_	(A:)	a.102.1.3	Y...A[ITV]D..D
▓	d2sqca1	(A:8-36,A:308-630)	a.102.4.2	6-17	2-19	259614	2.8	74-85	70-87	d1k04a_	(A:)	a.24.14.1	R...[KMR]A...L[DLN]S
░	d1dceb_	(B:)	a.102.4.3	281-290	281-300	476486	6.7	241-250	241-260	d1fp3a_	(A:)	a.102.1.3	G...[DLN]P.H.L
░	d1dceb_	(B:)	a.102.4.3	224-234	231-234	114854	3.9	41-51	48-51	d1fp3a_	(A:)	a.102.1.3	[CS].G.[DLN]GR...[DLN]
▓	d1f5na1	(A:284-583)	a.114.1.1	345-351	336-353	866770	2	206-212	197-214	d1j3ua_	(A:)	a.127.1.1	RD.ER.A
░	d1iala_	(A:)	a.118.1.1	423-434	421-438	1106012	2.6	52-63	50-67	d1kpsb_	(B:)	a.118.12.1	[DLN].SV..A.L[DLN].I
░	d1b3ua_	(A:)	a.118.1.2	187-196	182-197	3485775	3.1	228-237	223-238	d1hf8a_	(A:)	a.118.10.1	[KMR]..EF.KV.E
▓	d1b3ua_	(A:)	a.118.1.2	495-505	491-516	2244325	3.1	52-62	48-73	d1k04a_	(A:)	a.24.14.1	L..[KMR]T.L..[ITV][DLN]
░	d1hx8a_	(A:)	a.118.10.1	223-232	211-244	711854	2.7	352-361	340-373	d1kt1a1	(A:254-421)	a.118.8.1	K.[FHWY][DLN].R..R[ITV]
░	d1kpsb_	(B:)	a.118.12.1	59-68	53-70	225368	3.7	253-262	247-264	d1gw5b_	(B:)	a.118.1.10	[ITV]L.A[ITV]...M[KMR]
░	d1l5ja1	(A:1-160)	a.118.15.1	82-95	81-96	3109338	2	120-133	119-134	d1iala_	(A:)	a.118.1.1	[ITV].E.A[ITV]..LG...G
░	d1ldja2	(A:17-410)	a.118.17.1	342-351	337-354	910835	2.8	86-95	81-98	d1dcea1	(A:1-241,A:351-443)	a.118.6.1	AAL.KA...F
▓	d1ldja2	(A:17-410)	a.118.17.1	136-149	145-148	504105	2	44-57	53-56	d1c3ca_	(A:)	a.127.1.1	[ITV]E[KMR].RN...I[DLN][ITV].L
▓	d1ldja2	(A:17-410)	a.118.17.1	325-336	326-335	910085	3.4	162-173	163-172	d1foea1	(A:1034-1239)	a.87.1.1	KY..L.[KMR]..F.[DLN]
▓	d1qsaa1	(A:1-450)	a.118.5.1	251-262	250-262	591755	3.1	291-302	290-302	d1cb8a1	(A:26-335)	a.102.3.2	[ITV][EQ]E..D.[ITV]A..[DLN]
▓	d1qsaa1	(A:1-450)	a.118.5.1	154-165	153-166	591758	2.7	109-120	108-121	d1cb8a1	(A:26-335)	a.102.3.2	P.A..E[KMR]..L.[KMR]
░	d1qsaa1	(A:1-450)	a.118.5.1	318-328	317-333	591606	3.4	162-172	161-177	d1b3ua_	(A:)	a.118.1.2	K.E.R.[FHWY]..[DLN]L
░	d1qsaa1	(A:1-450)	a.118.5.1	214-222	213-228	591638	2.5	258-266	257-272	d1b3ua_	(A:)	a.118.1.2	[ITV]R.M.A..F
░	d1qsaa1	(A:1-450)	a.118.5.1	253-262	250-263	329994	2.8	177-186	174-187	d1lrv__	(-)	a.118.1.5	E...[ITV]VA.RL
░	d1qsaa1	(A:1-450)	a.118.5.1	253-262	250-263	329994	2.8	177-186	174-187	d1lrv__	(-)	a.118.1.5	E.R..VA.RL
░	d1qsaa1	(A:1-450)	a.118.5.1	253-262	250-263	329994	2.8	177-186	174-187	d1lrv__	(-)	a.118.1.5	E.R..[ITV]A.RL
░	d1qsaa1	(A:1-450)	a.118.5.1	253-262	250-262	329995	2.6	105-114	102-114	d1lrv__	(-)	a.118.1.5	E...[ITV]VA.RL
░	d1qsaa1	(A:1-450)	a.118.5.1	246-262	250-262	329995	2.6	98-114	102-114	d1lrv__	(-)	a.118.1.5	L..D[EQ]..E.R..VA.RL
░	d1qsaa1	(A:1-450)	a.118.5.1	253-262	250-262	329995	2.6	105-114	102-114	d1lrv__	(-)	a.118.1.5	E.R..VA.RL
░	d1qsaa1	(A:1-450)	a.118.5.1	249-262	250-262	329995	2.6	101-114	102-114	d1lrv__	(-)	a.118.1.5	D[EQ]..[EQ].R..VA.RL
░	d1qsaa1	(A:1-450)	a.118.5.1	253-262	250-262	329995	2.6	105-114	102-114	d1lrv__	(-)	a.118.1.5	E.R..[ITV]A.RL
▓	d1dcea1	(A:1-241,A:351-443)	a.118.6.1	85-99	63-105	356244	4	147-161	125-167	d1jfaa_	(A:)	a.128.1.5	S..L[ITV][KMR]..L.F.E.C

d1dcea1	(A:1-241,A:351-443)	a.118.6.1	417-426	417-439	355654	3.4	149-158	149-171	d1fpoa2	(A:77-171)	a.23.1.1	[KMR]...[FHWY]LD.LR
d1dcea1	(A:1-241,A:351-443)	a.118.6.1	417-427	417-439	355654	3.4	149-159	149-171	d1fpoa2	(A:77-171)	a.23.1.1	[KMR]...[FHWY]LD.[DLN][KMR]S
d1dcea1	(A:1-241,A:351-443)	a.118.6.1	417-427	417-439	355654	3.4	149-159	149-171	d1fpoa2	(A:77-171)	a.23.1.1	[KMR][KMR]...LD..RS
d1dcea1	(A:1-241,A:351-443)	a.118.6.1	402-411	399-411	3014871	1.8	57-66	54-66	d1aa7a_	(A:)	a.95.1.1	K.[ITV]L.[FHWY]..TL
d1kt1a1	(A:254-421)	a.118.8.1	276-294	262-301	887629	3	25-43	11-50	d1e52a_	(A:)	a.2.9.1	L.M.[FHWY]..[DLN].E[FHWY]..A.[EQ]..D
d1hz4a_	(A:)	a.118.8.2	292-303	285-306	204225	6.3	184-195	177-198	d1fp3a_	(A:)	a.102.1.3	[KMR].L.L..QL..[EQ]
d1yge_1	(150-839)	a.119.1.1	780-791	777-798	843215	2	158-169	155-176	d1foea1	(A:1034-1239)	a.87.1.1	Q...K[FHWY].[DLN].L[KMR]E
d1yge_1	(150-839)	a.119.1.1	780-791	777-798	843215	2	158-169	155-176	d1foea1	(A:1034-1239)	a.87.1.1	Q...K[FHWY].[DLN].L.E
d1j3ua_	(A:)	a.127.1.1	54-63	59-63	4276066	1	28-37	33-37	d1dv5a_	(A:)	a.28.1.3	L.[DLN].E[ITV]GLLD
d1jfaa_	(A:)	a.128.1.5	179-189	184-187	259166	3.2	108-118	113-116	d1k04a_	(A:)	a.24.14.1	[EQ][FHWY]...M[DLN]...H
d1kp8a1	(A:2-136,A:410-526)	a.129.1.1	58-70	61-65	3035218	1.1	111-123	114-118	d1di1a_	(A:)	a.128.1.4	E..L.D.[FHWY]E.M.A
d1g8qa_	(A:)	a.135.1.1	9-20	2-26	354186	2.1	119-130	112-136	d1fpoa2	(A:77-171)	a.23.1.1	K.VK..[FHWY]D...Q
d1jeqa1	(A:559-609)	a.140.2.1	536-546	532-548	499676	2.1	221-231	217-233	d1knya1	(A:126-253)	a.24.16.1	[EQ].LLE.L..[FHWY][FHWY]
d1g8ea_	(A:)	a.145.1.1	12-27	1-29	258896	2.1	84-99	73-101	d1k04a_	(A:)	a.24.14.1	[DLN].[DLN]L..L.[DLN]..[KMR]L..Q
d1k1fa_	(A:)	a.147.1.1	35-46	30-67	2711621	2.6	42-53	37-74	d1nlwa_	(A:)	a.38.1.1	L.[KMR]AK..I[KMR][KMR]LE
d1k1fa_	(A:)	a.147.1.1	35-46	30-67	2711621	2.6	42-53	37-74	d1nlwa_	(A:)	a.38.1.1	L.[KMR]A...I[KMR].L[EQ]
d1k1fa_	(A:)	a.147.1.1	39-45	30-49	2662452	0.2	46-52	37-56	d1hsta_	(A:)	a.4.5.13	K.SIRRL
d1gzsb_	(B:)	a.168.1.1	8-20	14-20	266836	2.5	14-26	20-26	d1h6ga1	(A:377-507)	a.24.9.1	V.D..L[EQ]T[DLN]...[DLN]
d1gzsb_	(B:)	a.168.1.1	8-18	4-23	4275818	1.1	5-15	1-20	d1dv5a_	(A:)	a.28.1.3	[ITV]K[DLN]..L.[ITV]L.D
d1gzsb_	(B:)	a.168.1.1	8-18	4-20	343188	2.2	466-476	462-478	d1qlaa1	(A:458-655)	a.7.3.1	[ITV]K[DLN].M..[ITV].[DLN]D
d1miua2	(A:2752-2887)	a.171.1.1	432-441	429-453	810232	2.5	94-103	91-115	d1jm6a1	(A:1003-1169)	a.29.5.1	R[ITV].S[EQ]F..AL
d1or7c_	(C:)	a.180.1.1	18-28	13-44	3186461	1.8	28-38	23-54	d1kbhb_	(B:)	a.153.1.1	[EQ].LN.L..NP[EQ]
d1n1ca_	(A:)	a.184.1.1	142-153	148-152	469076	3.1	437-448	443-447	d1fgja_	(A:)	a.138.1.3	S.[EQ]L..L[EQ]...N
d1n1ca_	(A:)	a.184.1.1	141-152	141-146	267185	3.3	12-23	12-17	d1h6ga1	(A:377-507)	a.24.9.1	[DLN].V.[DLN]SFL[EQ]T.V
d1n1ca_	(A:)	a.184.1.1	140-152	148-151	390595	2.1	22-34	30-33	d1ki1b1	(B:1229-1438)	a.87.1.1	EN.V.[DLN]..L.T.[ITV]
d1fpoa1	(A:1-76)	a.2.3.1	18-30	15-39	931916	3.7	338-350	335-359	d1f5na1	(A:284-583)	a.114.1.1	Q.L.[DLN].[FHWY].D.[EQ]R[EQ]
d1fpoa2	(A:77-171)	a.23.1.1	110-124	110-115	441462	2.1	551-565	551-556	d1b3ua_	(A:)	a.118.1.2	D...L[EQ]S.[ITV]K.[ITV]..[KMR]
d1fpoa2	(A:77-171)	a.23.1.1	112-126	112-139	258839	3.5	38-52	38-65	d1k04a_	(A:)	a.24.14.1	A..E.[FHWY]..[KMR]VK...[DLN]
d1fpoa2	(A:77-171)	a.23.1.1	112-126	112-139	258839	3.5	38-52	38-65	d1k04a_	(A:)	a.24.14.1	A..E.[FHWY][ITV]..VK...[DLN]
d1fpoa2	(A:77-171)	a.23.1.1	115-122	113-137	442181	2.7	34-41	32-56	d1guxa_	(A:)	a.74.1.3	ES..KRVK
d1k04a_	(A:)	a.24.14.1	43-57	44-47	259785	2.2	121-135	122-125	d1ldja2	(A:17-410)	a.118.17.1	[FHWY].P..K[EQ]V.[DLN]A...L

░	d1k04a_	(A:)	a.24.14.1	88-98	88-91	259210	3.3	215-225	215-218	d1knya1	(A:126-253)	a.24.16.1	G[EQ]L.[DLN]..KL.[EQ]
▓	d1k04a_	(A:)	a.24.14.1	46-55	43-64	258947	2.6	39-48	36-57	d1guxa_	(A:)	a.74.1.3	[KMR]VK.[ITV]G...[KMR]
▓	d1knya1	(A:126-253)	a.24.16.1	143-154	142-158	499632	2.1	11-22	10-26	d2spca_	(A:)	a.7.1.1	[DLN].E.A..W[KMR]..R
░	d1nlxa_	(A:)	a.24.17.1	44-50	28-51	3391339	2.1	125-131	109-132	d1gs9a_	(A:)	a.24.1.1	KR[DLN]L.DA
▓	d1nlxa_	(A:)	a.24.17.1	69-81	72-75	304035	3	780-792	783-786	d1h3na1	(A:687-814)	a.27.1.1	[FHWY]..A.H.A.E[DLN].[FHWY]
▓	d1nlxa_	(A:)	a.24.17.1	7-21	1-22	449480	2.1	38-52	32-53	d1guxa_	(A:)	a.74.1.3	K.[ITV].D[ITV]...F[KMR]...A
▓	d1h6ga1	(A:377-507)	a.24.9.1	22-35	22-35	265057	2	405-418	405-418	d1ewqa1	(A:267-541)	a.113.1.1	[DLN]..L[DLN].L..A.[KMR].G
▓	d1jkva_	(A:)	a.25.1.2	53-62	51-65	4276131	0.1	3-12	1-15	d1dv5a_	(A:)	a.28.1.3	E..K[DLN]..LD[ITV]
▓	d1cnt1_	(1:)	a.26.1.1	40-48	38-66	996325	3.5	75-83	73-101	d1k04a_	(A:)	a.24.14.1	E.E[KMR].Q.[DLN]L
▓	d2ilk__	(-)	a.26.1.3	27-37	10-37	311317	2.8	124-134	107-134	d1fpoa2	(A:77-171)	a.23.1.1	[KMR]..T.[FHWY]Q.[KMR].[EQ]
▓	d1is2a1	(A:272-460)	a.29.3.2	425-435	420-446	856579	2.1	6-16	1-27	d1fx7a1	(A:1-64)	a.4.5.24	[DLN].T..[FHWY]L[KMR].IY
▓	d1is2a1	(A:272-460)	a.29.3.2	425-435	420-446	856579	2.1	6-16	1-27	d1fx7a1	(A:1-64)	a.4.5.24	[DLN].T.[KMR][FHWY]L..IY
▓	d1is2a1	(A:272-460)	a.29.3.2	425-435	420-446	856579	2.1	6-16	1-27	d1fx7a1	(A:1-64)	a.4.5.24	[DLN].T.[KMR].L[KMR].IY
░	d1iw7f1	(F:258-318)	a.4.13.1	192-208	189-209	891324	3.5	5-21	2-22	d1fx7a1	(A:1-64)	a.4.5.24	V.T[ITV].[KMR]..RT...L[EQ][EQ]E
▓	d1or7a1	(A:120-187)	a.4.13.2	141-151	145-148	4276805	0.7	22-32	26-29	d1dv5a_	(A:)	a.28.1.3	[ITV].[KMR].LD.[DLN].[FHWY]E
░	d1l0oc_	(C:)	a.4.13.2	35-43	31-57	497217	4	21-29	17-43	d1i1ga1	(A:2-61)	a.4.5.32	[ITV]EIA..L.[ITV]
░	d1hsta_	(A:)	a.4.5.13	33-39	35-38	436685	0.8	33-39	35-38	d1igna1	(A:360-445)	a.4.1.6	SHY...H
▓	d1ka8a_	(A:)	a.4.5.20	61-73	55-74	449899	2.2	42-54	36-55	d1af7_1	(11-91)	a.58.1.1	[DLN].L..[KMR]L[KMR]..G.[DLN]
░	d1i1ga1	(A:2-61)	a.4.5.32	1-15	1-18	436717	2.8	5-19	5-22	d1igna1	(A:360-445)	a.4.1.6	[ITV]DE.D..IL.[ITV]..K[DLN]
░	d1i1ga1	(A:2-61)	a.4.5.32	2-11	1-18	436717	2.8	6-15	5-22	d1igna1	(A:360-445)	a.4.1.6	DE.D..IL.[ITV]
░	d1i1ga1	(A:2-61)	a.4.5.32	23-38	18-44	537957	5.9	165-180	160-186	d1a04a1	(A:150-216)	a.4.6.2	IA..L.I.E..V[KMR]..[ITV]
░	d1i1ga1	(A:2-61)	a.4.5.32	19-39	18-44	537957	5.9	161-181	160-186	d1a04a1	(A:150-216)	a.4.6.2	P...IA[KMR][KMR]L.I.E..V[KMR]..VK
░	d1p4wa_	(A:)	a.4.6.2	41-47	38-65	2120853	4.6	21-27	18-45	d1i1ga1	(A:2-61)	a.4.5.32	[ITV].IAKK[DLN]
▓	d1bkra_	(A:)	a.40.1.1	34-43	33-44	466221	2.9	171-180	170-181	d1knya1	(A:126-253)	a.24.16.1	MA...LI..H
▓	d1bg1a1	(A:136-321)	a.47.1.1	135-144	129-145	1467901	2.5	369-378	363-379	d1dcea1	(A:1-241,A:351-443)	a.118.6.1	[EQ]..K.L[EQ]EL[EQ]
▓	d1lvfa_	(A:)	a.47.2.1	48-61	41-65	1934606	2.3	385-398	378-402	d1fp3a_	(A:)	a.102.1.3	R...[KMR][CS].E.[DLN]..[DLN]L
▓	d1lvfa_	(A:)	a.47.2.1	11-19	1-26	792181	4.1	315-323	305-330	d1f5na1	(A:284-583)	a.114.1.1	VQKA[ITV]...[EQ]
▓	d1lvfa_	(A:)	a.47.2.1	20-28	22-28	792923	2.8	152-160	154-160	d1ldja2	(A:17-410)	a.118.17.1	G..Q.[FHWY][ITV]EL
▓	d1lvfa_	(A:)	a.47.2.1	57-65	39-68	2229651	3.7	57-65	39-68	d1k04a_	(A:)	a.24.14.1	[DLN]L...DETI
▓	d1jmsa1	(A:148-242)	a.60.6.1	25-31	17-33	4276142	1.4	10-16	2-18	d1dv5a_	(A:)	a.28.1.3	LDILA.[DLN]
▓	d1jvsa1	(A:301-399)	a.69.3.1	380-388	369-399	886365	3	98-106	87-117	d1j3ua_	(A:)	a.127.1.1	S[ITV][DLN].NA.EV

d1qlaa1	(A:458-655)	a.7.3.1	466-479	462-480	4277061	2.7	5-18	1-19	d1dv5a_	(A:)	a.28.1.3	IKN...D[ITV]..D[DLN][ITV]G
d1gvna_	(A:)	a.8.2.1	66-80	65-81	2883937	0.4	70-84	69-85	d1hh8a_	(A:)	a.118.8.1	AV..[FHWY].R.M[DLN].Y..[EQ]
d1gvna_	(A:)	a.8.2.1	46-55	49-53	3588777	0.7	26-35	29-33	d1cf7a_	(A:)	a.4.5.17	[DLN].LK.A..[ITV][DLN]
d1gvna_	(A:)	a.8.2.1	28-41	29-32	602324	2.6	3-16	4-7	d2spca_	(A:)	a.7.1.1	L[DLN].[EQ]L...D[CS][EQ]L.E
d1gvna_	(A:)	a.8.2.1	28-41	29-32	602324	2.6	3-16	4-7	d2spca_	(A:)	a.7.1.1	L[DLN].[EQ]L.[KMR].D[CS].L.E
d1gvna_	(A:)	a.8.2.1	29-41	29-32	602324	2.6	4-16	4-7	d2spca_	(A:)	a.7.1.1	[DLN].[EQ]L.[KMR].D.[EQ]L.E
d1aa7a_	(A:)	a.95.1.1	2-14	5-12	485944	1.7	4-16	7-14	d2gmfa_	(A:)	a.26.1.2	S..T[EQ]...[FHWY]V..I
d1aa7a_	(A:)	a.95.1.1	139-148	139-158	462827	1.4	60-69	60-79	d1hula_	(A:)	a.26.1.2	T[ITV]E..F.[DLN].[CS]
d1iray1	(Y:1-101)	b.1.1.4	69-81	71-80	763168	3.2	313-325	315-324	d1dcea2	(A:242-350)	b.7.4.1	[DLN]..H.[FHWY].V[ITV]...S
d1lqsr1	(R:2-100)	b.1.2.1	56-68	57-63	191388	6.9	299-311	300-306	d1dcea2	(A:242-350)	b.7.4.1	[ITV]...DL.A..L[DLN][DLN]
d1czsa_	(A:)	b.18.1.2	69-76	66-78	1652172	4.2	41-48	38-50	d1stma_	(A:)	b.121.7.1	[ITV][ITV]A.[ITV][ITV]QG
d1mhna_	(A:)	b.34.9.1	20-25	15-30	821986	2.6	48-53	43-58	d1nxza1	(A:2-73)	b.122.1.2	IYPA.I
d1ihka_	(A:)	b.42.1.1	52-58	53-58	189027	6	81-87	82-87	d1hcd__	(-)	b.42.5.2	G.VSI[KMR]G
d1krha1	(A:106-205)	b.43.4.2	168-178	166-179	580817	4.8	502-512	500-513	d1fuia1	(A:356-591)	b.43.2.1	T.F..R[DLN][ITV]..G
d1jmxa5	(A:163-281)	b.61.4.1	256-263	256-259	775444	2.7	85-92	85-88	d1h3ia1	(A:52-193)	b.76.2.1	DGEM.G.[KMR]
d1itva_	(A:)	b.66.1.1	16-25	18-25	199205	5.2	16-25	18-25	d1gyha_	(A:)	b.67.2.1	G[DLN]..YLF..G
d1itva_	(A:)	b.66.1.1	109-121	110-116	199249	2.3	405-417	406-412	d1k3ia3	(A:151-537)	b.69.1.1	[KMR]GK.L.F.G..[DLN][FHWY]
d1itva_	(A:)	b.66.1.1	23-33	24-33	1024886	2.7	342-352	343-352	d1a12a_	(A:)	b.69.5.1	KDG..[FHWY].[FHWY]..G
d1h6la_	(A:)	b.68.3.1	145-153	143-161	3804904	4.1	202-210	200-218	d1qksa2	(A:136-567)	b.70.2.1	Y..V[ITV]G[KMR].G
d1k3ia3	(A:151-537)	b.69.1.1	227-236	224-236	407725	0.8	87-96	84-96	d1lqsr1	(R:2-100)	b.1.2.1	S...VT.T[KMR][FHWY]
d1k3ia3	(A:151-537)	b.69.1.1	227-236	224-236	407725	0.8	87-96	84-96	d1lqsr1	(R:2-100)	b.1.2.1	S..T.T.T[KMR][FHWY]
d1k3ia3	(A:151-537)	b.69.1.1	227-236	224-236	407725	0.8	87-96	84-96	d1lqsr1	(R:2-100)	b.1.2.1	S..TV..T[KMR][FHWY]
d1k3ia3	(A:151-537)	b.69.1.1	246-257	250-256	3823891	8.4	152-163	156-162	d1tbga_	(A:)	b.69.4.1	[DLN].[DLN].QIV[ITV]..G[DLN]
d1k3ia3	(A:151-537)	b.69.1.1	248-257	250-256	3823891	8.4	154-163	156-162	d1tbga_	(A:)	b.69.4.1	[DLN]..IV[ITV]..G[DLN]
d1k3ia3	(A:151-537)	b.69.1.1	507-519	510-514	3823898	2.3	145-157	148-152	d1tbga_	(A:)	b.69.4.1	Y.S.[CS]..L.D..[ITV]
d1k3ia3	(A:151-537)	b.69.1.1	258-270	261-266	2897941	4.6	116-128	119-124	d1a12a_	(A:)	b.69.5.1	D..[KMR][ITV].L[FHWY].S..D
d1k3ia3	(A:151-537)	b.69.1.1	225-239	234-238	3805203	12.4	178-192	187-191	d1qksa2	(A:136-567)	b.70.2.1	[ITV]...[KMR]TV..T.[FHWY]..[FHWY]
d1jofa_	(A:)	b.69.10.1	281-289	280-290	491690	14.1	328-336	327-337	d1crza1	(A:141-409)	b.68.4.1	Q.[FHWY]IA...L
d1jofa_	(A:)	b.69.10.1	282-291	282-290	2703906	5	161-170	161-169	d1gxra_	(A:)	b.69.4.1	G.IA.[FHWY].L.[DLN]
d1jofa_	(A:)	b.69.10.1	281-291	281-288	2619598	7.7	238-248	238-245	d1a12a_	(A:)	b.69.5.1	[EQ]G[FHWY]..GF.L.[DLN]
d1jofa_	(A:)	b.69.10.1	281-291	281-288	2619598	7.7	238-248	238-245	d1a12a_	(A:)	b.69.5.1	[EQ]..[ITV].GF.L.[DLN]

d1jofa_	(A:)	b.69.10.1	281-291	281-288	2619598	7.7	238-248	238-245	d1a12a_	(A:)	b.69.5.1	[EQ]G.[ITV].GF.[DLN].[DLN]
d1jofa_	(A:)	b.69.10.1	281-291	281-288	2619598	7.7	238-248	238-245	d1a12a_	(A:)	b.69.5.1	[EQ]G.[ITV].G[FHWY].L.[DLN]
d1jofa_	(A:)	b.69.10.1	281-291	281-288	2619598	7.7	238-248	238-245	d1a12a_	(A:)	b.69.5.1	[EQ]G.[ITV]..F.L.[DLN]
d1l0qa2	(A:1-301)	b.69.2.3	121-131	115-135	75033	32.2	145-155	139-159	d1gxra_	(A:)	b.69.4.1	LA.SPD.K..[FHWY]
d1l0qa2	(A:1-301)	b.69.2.3	122-131	115-135	75033	32.2	146-155	139-159	d1gxra_	(A:)	b.69.4.1	A.SPD.K..[FHWY]
d1fwxa2	(A:8-451)	b.69.3.1	397-403	397-406	2897766	4.2	161-167	161-170	d1a12a_	(A:)	b.69.5.1	ND[FHWY]LV.L
d1k8kc_	(C:)	b.69.4.1	149-160	141-161	491665	13.3	268-279	260-280	d1crza1	(A:141-409)	b.68.4.1	W..[DLN]S.[DLN]L...S
d1tbga_	(A:)	b.69.4.1	231-242	223-246	491605	7.5	221-232	213-236	d1crza1	(A:141-409)	b.68.4.1	A..F.P[DLN]G...A
d1k8kc_	(C:)	b.69.4.1	146-157	139-161	491667	3.1	177-188	170-192	d1crza1	(A:141-409)	b.68.4.1	S..[FHWY].P[DLN]...LA
d1gxra_	(A:)	b.69.4.1	190-198	182-201	211270	32.1	124-132	116-135	d1l0qa2	(A:1-301)	b.69.2.3	S.DG.K.[FHWY][ITV]
d1gxra_	(A:)	b.69.4.1	146-155	139-159	211270	32.1	80-89	73-93	d1l0qa2	(A:1-301)	b.69.2.3	A[ITV]SPD.K..[FHWY]
d1gxra_	(A:)	b.69.4.1	146-155	139-159	211270	32.1	80-89	73-93	d1l0qa2	(A:1-301)	b.69.2.3	A.SPD.K..[FHWY]
d1gxra_	(A:)	b.69.4.1	190-198	182-201	211270	32.1	124-132	116-135	d1l0qa2	(A:1-301)	b.69.2.3	S.DG.KL[FHWY][ITV]
d1a12a_	(A:)	b.69.5.1	161-171	151-178	3775909	5.6	342-352	332-359	d1nexb2	(B:370-744)	b.69.4.1	[DLN].[FHWY]LV...ADG
d1g72a_	(A:)	b.70.1.1	145-160	151-160	2895896	3	183-198	189-198	d1gyha_	(A:)	b.67.2.1	D...GS..[ITV][EQ]APF[ITV].[KMR]
d1qksa2	(A:136-567)	b.70.2.1	440-446	432-451	491579	8.9	269-275	261-280	d1crza1	(A:141-409)	b.68.4.1	[FHWY]P[DLN]SQ.L
d1qksa2	(A:136-567)	b.70.2.1	200-208	198-210	3804814	22	99-107	97-109	d1fwxa2	(A:8-451)	b.69.3.1	GR[FHWY]LF...[KMR]
d1qksa2	(A:136-567)	b.70.2.1	362-371	355-382	2064778	17.9	212-221	205-232	d1tbga_	(A:)	b.69.4.1	D[ITV][KMR]EG[KMR]...[ITV]
d1gzga_	(A:)	c.1.10.3	247-253	233-253	811184	4.6	311-317	297-317	d1iexa1	(A:1-388)	c.1.8.7	G.DM[ITV]MV
d1nvma2	(A:2-290)	c.1.10.5	106-113	107-113	1754969	7.5	68-75	69-75	d1dxea_	(A:)	c.1.12.5	A.VVRV.T
d1qtwa_	(A:)	c.1.15.1	159-166	156-175	1083746	7.4	291-298	288-307	d1epxa_	(A:)	c.1.10.1	E[FHWY]L.AI.[DLN]
d1i60a_	(A:)	c.1.15.4	50-58	51-55	1018322	7.1	73-81	74-78	d1epxa_	(A:)	c.1.10.1	[DLN].AE.F[EQ].[FHWY]
d1i60a_	(A:)	c.1.15.4	123-131	120-131	751948	6.4	513-521	510-521	d1jqna_	(A:)	c.1.12.3	DV.T[EQ]L.[DLN]I
d1i60a_	(A:)	c.1.15.4	20-26	15-27	544808	2.8	163-169	158-170	d1a4ma_	(A:)	c.1.9.1	LELC.K[FHWY]
d1i60a_	(A:)	c.1.15.4	94-102	76-105	1018552	3.1	43-51	25-54	d1bf6a_	(A:)	c.1.9.3	[KMR]T.GV[KMR].V[ITV]
d1b5ta_	(A:)	c.1.23.1	88-93	74-98	1587574	2	67-72	53-77	d2liv__	(-)	c.93.1.1	N[DLN]GI[KMR][FHWY]
d1b5ta_	(A:)	c.1.23.1	88-94	74-98	1587574	2	67-73	53-77	d2liv__	(-)	c.93.1.1	N[DLN]GI[KMR].[ITV]
d1hkva2	(A:46-310)	c.1.6.1	129-136	118-136	518979	7.8	90-97	79-97	d1b5ta_	(A:)	c.1.23.1	G[ITV].HIV.[DLN]
d1a4ma_	(A:)	c.1.9.1	95-102	93-103	3674469	3	227-234	225-235	d1icpa_	(A:)	c.1.4.1	V.[ITV]R.SP[FHWY]
d1gkra2	(A:55-379)	c.1.9.6	137-143	131-151	1005824	4.7	151-157	145-165	d1b5ta_	(A:)	c.1.23.1	[KMR][KMR].DAGA
d1j5sa_	(A:)	c.1.9.8	416-422	401-423	892345	5	234-240	219-241	d1hkva2	(A:46-310)	c.1.6.1	L.[DLN]VVGE

d1fqva2	(A:146-431)	c.10.1.3	117-124	105-126	211135	17.2	223-230	211-232	d1o6va2	(A:33-416)	c.10.2.1	LSL.G[DLN].L
d1fqva2	(A:146-431)	c.10.1.3	137-144	129-146	211137	16.2	132-139	124-141	d1o6va2	(A:33-416)	c.10.2.1	NL.RL.LS
d1fqva2	(A:146-431)	c.10.1.3	163-170	159-172	355676	3.5	509-516	505-518	d1dcea3	(A:444-567)	c.10.2.2	RL.EL[DLN]L[CS]
d1h6ua2	(A:36-262)	c.10.2.1	131-139	130-142	28121	14.6	114-122	113-125	d1a4ya_	(A:)	c.10.1.1	L[EQ].L[FHWY]L.[DLN]N
d1o6va2	(A:33-416)	c.10.2.1	223-230	219-231	1173795	11.2	288-295	284-296	d1a4ya_	(A:)	c.10.1.1	LSL.G[DLN].L
d1o6va2	(A:33-416)	c.10.2.1	264-272	263-275	2627052	10.8	228-236	227-239	d1a4ya_	(A:)	c.10.1.1	L.EL.LG.N
d1h6ua2	(A:36-262)	c.10.2.1	169-178	166-185	28137	12	132-141	129-148	d1fqva2	(A:146-431)	c.10.1.3	LA.[DLN]S.L[ITV].L
d1h6ua2	(A:36-262)	c.10.2.1	169-176	166-185	28137	12	132-139	129-148	d1fqva2	(A:146-431)	c.10.1.3	LA.[DLN]S.L[ITV]
d1o6va2	(A:33-416)	c.10.2.1	189-198	189-206	1937917	12	129-138	129-146	d1fqva2	(A:146-431)	c.10.1.3	[ITV].[ITV]LAK[DLN].NL
d1dcea3	(A:444-567)	c.10.2.2	532-540	530-546	355669	8.9	215-223	213-229	d1fqva2	(A:146-431)	c.10.1.3	CP.LV.L[DLN]L
d1dcea3	(A:444-567)	c.10.2.2	535-542	522-545	355671	8.4	138-145	125-148	d1fqva2	(A:146-431)	c.10.1.3	LV.L[DLN][DLN].G
d1dcea3	(A:444-567)	c.10.2.2	535-542	522-545	355671	8.4	138-145	125-148	d1fqva2	(A:146-431)	c.10.1.3	LV.LNL.G
d1jzta_	(A:)	c.104.1.1	130-137	127-138	728626	5.3	254-261	251-262	d1kola2	(A:161-355)	c.2.1.1	[ITV][DLN]C.VDA[ITV]
d1jzta_	(A:)	c.104.1.1	129-134	129-139	2229278	5.7	134-139	134-144	d1o14a_	(A:)	c.72.1.1	K[ITV][DLN]C[ITV]V
d1jw9b_	(B:)	c.111.1.1	32-40	30-53	570119	6.6	4-12	2-25	d1gsoa2	(A:-2-103)	c.30.1.1	[KMR]VL[ITV][ITV]G.GG
d1ngva_	(A:)	c.111.1.2	132-138	128-142	325766	4.1	95-101	91-105	d1cfza_	(A:)	c.56.1.1	L.LADVL
d1jr2a_	(A:)	c.113.1.1	154-160	150-160	431766	4.8	380-386	376-386	d1cqxa3	(A:262-403)	c.25.1.5	ALK[DLN].GI
d1jx7a_	(A:)	c.114.1.1	3-8	1-10	1644449	2.5	12-17	10-19	d1cvra2	(A:1-350)	c.17.1.2	[KMR]IVIVA
d1tyfa_	(A:)	c.14.1.1	42-49	29-55	3382220	0.7	120-127	107-133	d1qfja2	(A:98-232)	c.25.1.1	A.NP.[KMR]DI
d1oe4a_	(A:)	c.18.1.3	181-188	167-188	2434713	2.5	68-75	54-75	d2liv__	(-)	c.93.1.1	[DLN]G[ITV]K.VIG
d1pqwa_	(A:)	c.2.1.1	86-91	79-93	2427707	5.5	101-106	94-108	d1o0ua_	(A:)	c.118.1.1	[ITV]LEL[ITV]D
d1pqwa_	(A:)	c.2.1.1	113-121	103-127	1808624	8.7	125-133	115-139	d1ej0a_	(A:)	c.66.1.2	[ITV]LAPGG.F[ITV]
d1jqba2	(A:1140-1313)	c.2.1.1	257-261	245-264	511287	7.5	135-139	123-142	d1g38a_	(A:)	c.66.1.27	KPGG[ITV]
d1jqba2	(A:1140-1313)	c.2.1.1	255-260	247-264	509698	6.4	163-168	155-172	d1xvaa_	(A:)	c.66.1.5	MV[KMR]PGG
d1uaya_	(A:)	c.2.1.2	21-30	11-32	1089307	8.5	27-36	17-38	d1kjna_	(A:)	c.115.1.1	LK.[KMR]G[FHWY]RV[ITV]V
d1iy8a_	(A:)	c.2.1.2	34-41	29-49	2983709	9.8	53-60	48-68	d1m6ya2	(A:2-114,A:216-294)	c.66.1.23	[ITV]DV.SE.L
d1hdca_	(A:)	c.2.1.2	130-137	128-138	3719207	9.1	235-242	233-243	d1m6ya2	(A:2-114,A:216-294)	c.66.1.23	GG.IV.IS
d1h5qa_	(A:)	c.2.1.2	140-146	138-147	71430	8.5	236-242	234-243	d1m6ya2	(A:2-114,A:216-294)	c.66.1.23	G.IVV[ITV]S
d1nvmb1	(B:1-131,B:287-312)	c.2.1.3	9-16	6-26	2128053	6.4	6-13	3-23	d1i8ta1	(A:1-244,A:314-367)	c.4.1.3	I[ITV]GSG[DLN].G
d1i36a2	(A:1-152)	c.2.1.6	24-30	23-31	520433	5.6	53-59	52-60	d1fuia2	(A:1-355)	c.85.1.1	VE.V[ITV]S[DLN]
d1a04a2	(A:5-142)	c.23.1.1	49-56	46-56	3960724	3.5	168-175	165-175	d1o9ga_	(A:)	c.66.1.29	PD.[ITV]L.DL

d1a04a2	(A:5-142)	c.23.1.1	49-56	46-56	3960724	3.5	168-175	165-175	d1o9ga_	(A:)	c.66.1.29	PD.[ITV]L.[DLN]L
d1a04a2	(A:5-142)	c.23.1.1	49-56	46-56	3960724	3.5	168-175	165-175	d1o9ga_	(A:)	c.66.1.29	P[DLN].[ITV]L.DL
d1i3ca_	(A:)	c.23.1.1	85-91	84-95	758386	4.8	191-197	190-201	d1jbwa2	(A:1-296)	c.72.2.2	[KMR].IPVV[ITV]
d1g2ia_	(A:)	c.23.16.2	21-30	13-36	240394	7.9	22-31	14-37	d1f0ka_	(A:)	c.87.1.2	H.L[KMR].[EQ]G.[EQ]V
d1cqxa3	(A:262-403)	c.25.1.5	295-301	293-303	859847	5.5	45-51	43-53	d1e19a_	(A:)	c.73.1.1	[EQ]VV.[ITV]HG
d1h3na3	(A:1-225,A:418-686)	c.26.1.1	67-75	52-78	329714	2.5	25-33	10-36	d1f0ka_	(A:)	c.87.1.2	[KMR].QG[FHWY][EQ]V.[FHWY]
d1coza_	(A:)	c.26.1.2	27-34	27-39	965798	2.1	54-61	54-66	d1cfza_	(A:)	c.56.1.1	[DLN].D[FHWY]L[ITV][ITV]A
d1ju2a1	(A:1-293,A:464-521)	c.3.1.2	27-35	26-48	2128036	9.1	2-10	1-23	d1i8ta1	(A:1-244,A:314-367)	c.4.1.3	YDY[ITV]IVG.G
d1ju2a1	(A:1-293,A:464-521)	c.3.1.2	27-33	26-48	2128036	9.1	2-8	1-23	d1i8ta1	(A:1-244,A:314-367)	c.4.1.3	YDY[ITV]I.G
d1ng4a1	(A:1-218,A:307-364)	c.3.1.2	194-199	194-201	1837013	3.3	248-253	248-255	d1f0ka_	(A:)	c.87.1.2	WA[DLN].VV
d1gsa_1	(1-122)	c.30.1.3	77-83	75-89	166030	4.4	67-73	65-79	d1hyha1	(A:21-166)	c.2.1.5	LAD.DV[ITV]
d1gsa_1	(1-122)	c.30.1.3	103-110	89-114	2047288	1.2	251-258	237-262	d1php__	(-)	c.86.1.1	E[KMR]A.EKG[ITV]
d1d4oa_	(A:)	c.31.1.4	128-135	128-138	588100	6.2	144-151	144-154	d2masa_	(A:)	c.70.1.1	[KMR].K[EQ]V[ITV].M
d1gvnb_	(B:)	c.37.1.21	43-49	43-58	445022	3.9	109-115	109-124	d1gg4a4	(A:99-312)	c.72.2.1	SGKTS.[KMR]
d1i8ta1	(A:1-244,A:314-367)	c.4.1.3	6-14	1-26	651403	4.2	6-14	1-26	d1hyha1	(A:21-166)	c.2.1.5	I[ITV]G.G[DLN].GA
d1i8ta1	(A:1-244,A:314-367)	c.4.1.3	2-8	1-24	651395	7.6	25-31	24-47	d1gpea1	(A:1-328,A:525-587)	c.3.1.2	YDYII.G
d1i8ta1	(A:1-244,A:314-367)	c.4.1.3	26-32	25-34	651395	7.6	50-56	49-58	d1gpea1	(A:1-328,A:525-587)	c.3.1.2	[KMR][ITV].V[ITV]EK
d1i8ta1	(A:1-244,A:314-367)	c.4.1.3	26-32	25-34	651395	7.6	50-56	49-58	d1gpea1	(A:1-328,A:525-587)	c.3.1.2	KVLVIEK
d1i8ta1	(A:1-244,A:314-367)	c.4.1.3	2-8	1-24	651395	7.6	25-31	24-47	d1gpea1	(A:1-328,A:525-587)	c.3.1.2	YDY[ITV]I.G
d1qgda3	(A:528-663)	c.48.1.1	563-570	558-571	3164880	1.7	50-57	45-58	d1o0ua_	(A:)	c.118.1.1	A.AAYE.L
d1gefa_	(A:)	c.52.1.18	20-26	1-28	487906	1.8	53-59	34-61	d1cg2a1	(A:26-213,A:327-414)	c.56.5.4	GF.V[ITV]RS
d1g99a1	(A:1-156)	c.55.1.2	3-10	1-21	533451	8.9	3-10	1-21	d1hjra_	(A:)	c.55.3.6	[ITV]L.I[DLN].GS
d1hjra_	(A:)	c.55.3.6	3-10	1-20	140422	7.9	4-11	2-21	d1huxa_	(A:)	c.55.1.5	[ITV]L.I[DLN].GS
d1hjra_	(A:)	c.55.3.6	3-10	1-20	140422	7.9	4-11	2-21	d1huxa_	(A:)	c.55.1.5	[ITV]LGID.GS
d1hjra_	(A:)	c.55.3.6	3-9	1-20	140422	7.9	4-10	2-21	d1huxa_	(A:)	c.55.1.5	[ITV]LGID.G
d1hjra_	(A:)	c.55.3.6	5-10	1-20	140422	7.9	6-11	2-21	d1huxa_	(A:)	c.55.1.5	GID.GS
d1kcfa2	(A:39-256)	c.55.3.7	42-49	38-55	323351	5.6	7-14	3-20	d1jcfa1	(A:1-140)	c.55.1.1	GIDLG[ITV].N
d1kcfa2	(A:39-256)	c.55.3.7	40-46	38-57	452794	4.7	4-10	2-21	d1huxa_	(A:)	c.55.1.5	[ITV]LGID.G
d1cfza_	(A:)	c.56.1.1	144-151	134-153	3460615	4.9	117-124	107-126	d1m2fa_	(A:)	c.23.1.5	QV[DLN]AAL.E
d1ej0a_	(A:)	c.66.1.2	26-31	25-33	424613	3.5	6-11	5-13	d1d7ya1	(A:5-115,A:237-308)	c.3.1.5	VV.LGA
d1ne2a_	(A:)	c.66.1.32	74-79	67-84	2109208	3.8	84-89	77-94	d1hv8a1	(A:3-210)	c.37.1.19	[DLN]AI[EQ][ITV]A

▓	d1xvaa_	(A:)	c.66.1.5	73-81	68-91	240584	3.7	22-30	17-40	d1g2ia_	(A:)	c.23.16.2	[KMR]L.EEG[FHWY].V
▓	d1qlwa_	(A:)	c.69.1.15	303-311	301-316	470817	2	573-581	571-586	d1gpea1	(A:1-328,A:525-587)	c.3.1.2	L.VAD.ILD
▓	d1e19a_	(A:)	c.73.1.1	2-9	1-9	3098861	1.4	3-10	2-10	d1jx7a_	(A:)	c.114.1.1	K.V[ITV][ITV]A[DLN]G
▓	d1b74a1	(A:1-105)	c.78.2.1	85-92	87-92	1580227	4.8	78-85	80-85	d1qcza_	(A:)	c.23.8.1	[ITV][DLN]VPV.GV
▓	d1jeoa_	(A:)	c.80.1.3	67-74	71-74	690549	4.3	173-180	177-180	d1o98a1	(A:77-310)	c.105.1.1	[FHWY].VGE[ITV].T
▓	d1f0ka_	(A:)	c.87.1.2	17-24	14-30	583706	6.5	13-20	10-26	d1hyha1	(A:21-166)	c.2.1.5	G.AVAH.L
▓	d1f0ka_	(A:)	c.87.1.2	247-254	242-255	583707	6.3	68-75	63-76	d1hyha1	(A:21-166)	c.2.1.5	A.AD.V[ITV][CS]
▓	d1gz5a_	(A:)	c.87.1.6	23-30	21-33	374293	3.4	14-21	12-24	d2masa_	(A:)	c.70.1.1	[DLN]AV.IL.A
▓	d1ko7a1	(A:1-129)	c.98.2.1	85-92	89-92	2944	2.3	195-202	199-202	d1jbwa2	(A:1-296)	c.72.2.2	[ITV]VT.[DLN]L.P
▓	d1emsa2	(A:10-280)	d.160.1.1	25-35	21-35	26857	3.5	18-28	14-28	d1nnwa_	(A:)	d.159.1.5	AA..[KMR]IE...E
▓	d1dl5a2	(A:214-317)	d.197.1.1	270-278	272-278	463522	2.1	53-61	55-61	d1kafa_	(A:)	d.199.1.1	[DLN][DLN]..MRI[FHWY]G
▓	d1go3e2	(E:1-78)	d.230.1.1	46-56	49-52	243463	2.7	71-81	74-77	d1ekra_	(A:)	d.58.21.1	LS.V.V.[DLN]..E
▓	d1azsa_	(A:)	d.58.29.1	41-49	40-55	374681	2.7	17-25	16-31	d1otga_	(A:)	d.80.1.2	L..LFA[KMR].[DLN]
░	d1i19a1	(A:274-613)	d.58.32.3	399-410	401-405	876432	3.9	44-55	46-50	d1phza1	(A:19-115)	d.58.18.3	[DLN]..[ITV]E.R.[CS].L[DLN]
░	d1i19a1	(A:274-613)	d.58.32.3	398-410	401-405	876432	3.9	43-55	46-50	d1phza1	(A:19-115)	d.58.18.3	[DLN]...[ITV]E.R.[CS].L[DLN]
░	d1dj0a1	(A:7-114)	d.58.35.1	84-98	88-92	487816	3.8	286-300	290-294	d1cg2a2	(A:214-326)	d.58.19.1	[DLN]...D[ITV].V..[ITV][KMR]..P
░	d1iwga1	(A:38-134)	d.58.44.1	93-102	92-103	447531	5.1	380-389	379-390	d1psda3	(A:327-410)	d.58.18.1	D.D[ITV]A[EQ]...Q
▓	d1b7fa1	(A:123-204)	d.58.7.1	72-81	77-79	1579541	4	195-204	200-202	d1clia2	(A:171-345)	d.139.1.1	K.L.VS...P
░	d1cvja1	(A:11-90)	d.58.7.1	55-62	54-66	993246	5.7	384-391	383-395	d1psda3	(A:327-410)	d.58.18.1	AE[KMR]AL..M
▓	d1orsc_	(C:)	f.14.1.1	35-44	32-55	2441231	2.4	172-181	169-192	d1qlac_	(C:)	f.21.2.1	YLV.L[FHWY].V.L
▓	d1c0va_	(A:)	f.17.1.1	17-38	32-38	4023944	3.1	4-25	19-25	d1lnqa2	(A:19-98)	f.14.1.1	[KMR].L..[ITV].A.I..G[ITV].G..F.EG
▓	d1l7va_	(A:)	f.22.1.1	150-159	147-169	10884	2.6	145-154	142-164	d1cola_	(A:)	f.1.1.1	VALGI.[CS]..L
▓	d1jb0a_	(A:)	f.29.1.1	424-433	422-454	2770084	2.4	24-33	22-54	d1nekc_	(C:)	f.21.2.2	AI.S.L..V[CS]
▓	d1jb0a_	(A:)	f.29.1.1	422-432	422-457	646918	2.4	77-87	77-112	d1fftc_	(C:)	f.25.1.1	[KMR]..[ITV]IS[FHWY]L..[ITV]