Exploring the Non-vertical Component of Bacterial Evolution Using Network Structures

A thesis submitted to the National University of Ireland for the Degree of
Doctor of Philosophy

Presented by:
Leanne S. Haggerty
Department of Biology,
NUI Maynooth,
Co. Kildare, Ireland.

NUI MAYNOOTH

Ollscoil na hÉireann Má Nuad

October 2012

Supervisor: Professor James O. McInerney B.Sc., Ph.D. (Galway)
Head of Dept.: Professor Paul Moynagh, BA(mod), PhD (Dublin)

Table of Contents

Acknowledgements

Firstly I'd like to acknowledge my supervisor, James. I have the utmost appreciation for all the opportunities that you have provided me with, during both my degree and PhD. I can honestly say that without your knowledge and guidance I would not have had the confidence to achieve all that I have under your supervision.

My research would not have been possible without funding from Science Foundation Ireland and the excellent computing facilities provided by the NUIM HPC facility and the Irish Centre for High-End Computing (ICHEC). In particular I'd like to thank Dr. Simon Wong for his unwavering patience and invaluable computing advice. In addition I have to say that this would have been a much more arduous journey without the technical support provided by Brian Daly and Dr. Vanush Paturyan. For both it seemed that no problem was unsolvable, I am extremely grateful for the time they gave me.

I cannot deny that this entire process would not have been nearly as enjoyable without all of my friends, both inside and outside of the lab. Carla and Lahcen, you are by far the two coolest people I have ever met. You were never short of time to help with my problems in work, to teach me extreme sports or to relax and talk nonsense with me. Sinead and Aoife, you have been my support system inside of the lab. I don't think I could have survived the conferences, the presentations, the thesis writing and all the nasty surprises of the PhD without your tea, chocolate, giggles and gossip. I'd like to thank my housemates in particular Kev and Ciaran. Your encouragement and motivation over the last two years have been more important than you know. I must reserve a special mention for Ms. Ruth Brennan, one of the loveliest ladies I have ever met. You may not realize it but the pain of the final two months of writing was often alleviated by a chat and a cup of tea with you, for which I am so grateful. My brother Charlie also deserves my thanks, for though he does not understand what I do or even why I do it he never fails to express his pride in me.

Traditionally, I have saved the best for last, for Graham you have been my rock through all of this. You have been by my side whenever I've needed you, to provide encouragement or perspective or even just a hug. You never fail to put a smile on my face, I couldn't ask for a better friend or partner.

Declaration

This thesis has not been submitted in whole, or in part, to this, or any other University for any other degree and is, except where otherwise stated, the original work of the author.

Signed:_____

Leanne S. Haggerty

Abbreviations Used

DNA	DeoxyriboNucleic Acid
BLAST	Basic Local Alignment Search Tool
RNA	RiboNucleic Acid
16S rRNA	16S ribosomal RNA
ToL	Tree of Life
HGT	Horizontal Gene Transfer
MGE	Mobile Genetic Element
GTA	Gene Transfer Agent
YESS	*Yersinia*, *Escherichia*, *Salmonella*, *Shigella*
COG	Clusters of Orthologous Genes

Index of Figures

Chapter 1 - Introduction

Chapter 2 –FUSION: A Network-based Approach to Finding Fusion Genes

Chapter 3 - Using FUSION: Homology Network Properties Reveal Fusions in *Salmonella* enterica

Chapter 4 – Phylogenetic and Non-phylogenetic signals During the Separation of the Enteric Bacterial Species Cloud

Index of Tables

Chapter 3 - Using FUSION: Homology Network Properties Reveal Fusions in *Salmonella* enterica

Chapter 4 – Phylogenetic and Non-phylogenetic signals During the Separation of the Enteric Bacterial Species Cloud

Abstract

The central tree metaphor has been challenged over the last couple of decades with the observation of incongruent trees derived largely from protein-coding genes in prokaryotic genomes. There are an increasing number of evolutionary processes and entities that confuse and confound the traditional understanding of evolution. As a result, these processes and entities are very often omitted from phylogenetic studies altogether.

In this thesis I attempt to uncover the importance of non-tree like evolution. I discuss the types of genes that do not adhere to vertical patterns of inheritance such as fusion genes and mobile genetic elements. Furthermore I explore the alternative of using network structures in describing the evolutionary history of bacteria.

This thesis recounts two key uses of networks for revealing the less commonly noted aspects of bacterial evolution. Firstly I present each stage in the development of a new method for identifying fusions of unrelated genes from conception of the idea, through the implementation to its application to data. Secondly I use networks of gene sharing to elucidate patterns of divergence among a group of closely related bacteria that would have once formed a single species cloud.

These studies reveal an abundance of the types of genes that contradict traditional tree-thinking and support the notion that a strictly vertical view of evolution is inadequate when describing bacterial relationships.

Chapter 1 – Introduction

1.1: Studying Bacteria

"Any good biologist finds it intellectually distressing to devote his life to the study of a group that cannot be readily and satisfactorily defined in biological terms; and the abiding intellectual scandal of bacteriology has been the absence of a clear concept of a bacterium." - Stanier and van Niel (1962)

1.1.1: Discovery

Antonie van Leeuwenhoek first observed bacteria under a microscope in 1676. In his letters to the Royal Society van Leeuwenhoek wrote of "animalcules": microorganisms that he had observed in water and in scrapings from teeth (Gest 2004). It was not until 1828, however, that Christian Gottfried Ehrenberg coined the name bacterium (Ehrenberg 1828). Ehrenberg defined bacteria as the non-spore-forming, rod-shaped microorganisms, and later named the spore-forming kind bacillus. A drawn out debate on the definition of bacteria followed.

Until 1859 it was not known that bacteria caused disease. Louis Pasteur and Robert Koch were early advocates of the germ theory of disease, otherwise known as the pathogenic theory of medicine (Worboys 2000). In his experiments with Tuberculosis, Koch proved that bacteria caused disease through infection and

reproduction. The bacterium was suddenly a topic for research. Following this discovery, in 1910 Paul Ehrlich invented the first antibiotic. By combining arsenic with other chemicals he found a compound that would kill bacteria without harming the animal or human. Ehrlich named the drug Salvarsan and it was used to cure syphilis.

The discovery of bacteria and, as a result, antibiotics, has revolutionized medicine. However the bacterial species concept and defining the genera within this kingdom is ongoing.

1.1.2: Classification

Since very early studies bacteriologists have struggled to define the relationships among bacteria. Linnaeus placed all microbes in one species called the "Chaos Infusoria", and the chaos has persisted (Linnaeus 1774). Many have claimed that a phylogenetic classification is "impossible to apply to bacteria" (Winogradsky 1952). A universally accepted concept remains unstated to this day. Even at the highest levels of organization there has been much debate surrounding the nomenclature.

Early classification relied on phenotypic traits; bacteria were grouped based on a whole repertoire of physical characteristics and their usefulness in industry or medicine. Characteristics included cell shape, plane of cell division, possession of flagella, ability to form spores or colonies, staining reactions, pathogenicity and many more (Sapp 2009).

Carl Nageli thought of microbes as fission fungi or Schizomycetes (Mazumdar 2002) and using morphological traits, Ferdinand Cohn classified 6 genera of bacteria

as members of the Plantae (Smith and Gordon 1957). Bergey's manual even classified the genera of bacteria as "typically unicellular plants" (Bergey *et al.* 1923) and later "primitive plants" (Smith and Gordon 1957).

Ernst Haeckel did not agree that bacteria belonged within the Plantae and in 1866 he designated a third living kingdom within which the bacteria could fall. He named this kingdom Protista: the first living creatures. It included the Protozoa; unicellular organisms with animal-like behaviour, the Protophyta; those with plant-like behaviour and the Protista Neutralia; those ancestral to neither plant nor animal. Haekel named the lowest level of the protest kingdom the Moneres (later the Monera), he assigned bacteria to this level, claiming they were unique because they possessed no nucleus.

Herbert Faulkner Copeland adopted Haeckel's way of thinking and argued that the Monera should be a kingdom of its own due to the sharp distinction from protists by the absence of nuclei.

Edouard Chatton is famed for his prescient generalization of taxa at the highest level. By recognizing two general patterns of organization in the cell he coined the prokaryote-eukaryote distinction, it appears that he first used the terms in 1925 (Sapp, 2005). Although Chatton is credited for inventing the names it was his student Andre Lwoff that recommended the use of these names to Roger Stanier (Sapp 2005). Stanier and van Niel went on to distinguish prokaryotes (Greek for before karyon or nucleus) in negative terms in relation to eukaryotes (Greek for true nucleus) (Stanier and Niel 1962). The definition of prokaryotes relied on three common features: absence of true nuclei, absence of sexual reproduction and absence of plastids. In 1963, Stanier, Douderoff and Adelberg stated that "this basic divergence in cellular structure, which separates the bacteria and blue-green algae

from all other cellular organisms, probably represents the greatest single evolutionary discontinuity to be found in the present-day world".

Finally in 1974 Bergey's manual described bacteria as a Prokaryote, their relationships still defined by such phenotypic traits as morphology, gram staining and oxygen requirements.

1.1.3: SSU rRNA Phylogeny

In 1965 Zuckerkandl and Pauling suggested that molecular data might be used to understand evolutionary processes. They showed that the relative recentness of common ancestry of a group of animals (as judged against the fossil record) was in good agreement with the relative similarities of some proteins found in those animals (Zuckerkandl and Pauling 1965). When molecular evolutionary studies took off, it meant that the comparison of conserved sequences might give us an insight into bacterial phylogeny (Stackebrandt and Goebel 1994). In particular, the small subunit of rRNA (SSU rRNA) made it possible to infer evolutionary relationships between different bacteria. The SSU rRNA was ideal because it is universally distributed across microbial organisms and the highly conserved nature of these sequences makes it easy to obtain (Bruijn 2011). Almost the full length of an SSU rRNA sequence can be obtained using "universal" primers and without having to culture the organism. This 1.5kb sequence has often been used exclusively to classify microbes. Although for a long time it was considered the "Gold Standard" in bacterial phylogeny construction, the SSU rRNA has its drawbacks as a molecular marker. There is very little support for phylogeny using only one gene. Often SSU rRNA genes from evolutionarily distant organisms are similar in nucleotide composition and have very

often, though incorrectly, been placed together on trees (Doolittle 1999). In some cases other phylogenetic markers have been used such as the protein-coding genes, recA (Thompson *et al.* 2004), HSP70 and others but there have problems with cloning protein-encoding genes from diverse organisms.

1.1.4: Early Genome Sequencing

The very first methods of DNA sequencing were developed in the 1970s and were somewhat laborious. Sanger and Coulson's procedure, published in 1975, involved generating short oligonucleotides. These were then fractioned by electrophoresis on a polyacrylamide gel and visualized using autoradiography (Sanger and Coulson 1975).This technology provided the first fully sequenced DNA-based genome, belonging to the single-stranded bacteriophage φX174 (Sanger *et al.* 1977). At the time it would take a year to sequence a thousand base pairs (bp), the *Escherichia coli K-12* genome would have taken more than a thousand years to sequence (Binnewies *et al.* 2006). The chain-termination method, which essentially mimics DNA replication in-vivo, has always proven to be more efficient (Sanger *et al.* 1977).

Large-scale sequencing was made possible with the invention of polymerase chain reaction (PCR) and the shotgun cloning procedure. On July 28, 1995 the first complete sequence for a cellular life form was published. The sequence belonged to the 1.8 Megabase (Mb) genome of *Haemophilus influenza* (Fleischmann *et al.* 1995). Only 3 months later the 0.58 Mb sequence of *Mycoplasma genitalium* was released (Fraser *et al.* 1995) and suddenly there was excitement surrounding the future for genome sequencing. Scientists predicted that in the two years following these

breakthrough there would be at least another eight fully sequenced bacterial genomes available (Koonin *et al.* 1996). A division of the GenBank database was opened, dedicated solely to storing the many genome sequences to come.

1.1.5: Second generation sequencing and its impact

The trajectory of a new technology can be summarized as a race to a commoditization phase, in which competitive forces drive the price down while performance and reliability approach the ideal (web link 1).

Between 2005 and 2007, the demand for fast, accurate and, most of all low-cost sequencing drove the development of high-throughput or second-generation sequencing. Three new methods were commercialized that were based on amplification strategies as alternatives to the standard cloning system: 454 (Margulies *et al.* 2005), Illumina (Bentley 2006) and SOLiD.

In October 2006 the Archon X prize was announced. $10 million would be awarded to the team that could build a device to sequence 100 human genomes within 30 days or less. There must be no more than one error per 1,000,000 bases, an accuracy rate of 98% of the genome and a cost of, at most, $1,000 per genome (Kedes 2011). The competition starts 3rd January 2013 and already a number of companies putting in massive effort including Complete Genomics, Illumina, ION Torrent Systems, GE Global and many more (Mukhopadhyay 2009). The technology that has emerged as a result of this competition includes the conversion of chemical information to digital, human-readable, information (Glenn 2011). Novel nanopore strand sequencing techniques are achievable on a device as small as a USB memory stick that achieves power and computer analysis from a normal laptop computer (web

link 2). It was even announced early this year that an entire human genome sequence could be obtained in one day for just $1,000 (web link 3).

According to the most up to date report from the Genomes Online Database or GOLD (web link 4), there are 18,306 genome projects completed or currently in progress. Of 3,705 completed genome projects, 3,363 are of bacterial species and of the 14,601 still in progress, 11,831 represent bacterial genome projects. From the time the first genome was sequenced in 1995 until 2008, almost 1,000 genomes were added to the GenBank database, according to GOLD statistics. Between 2008 and 2011 this number almost doubled. In 2011, of the 8,448 bacterial genome projects, either completed or in progress, 46% were focused on strains from the Proteobacteria. Many disease-causing bacteria belong to this group. The nature of many human pathogens is to evolve continually by mutation and by exchanging sequences with one another, so sequencing clinical isolates is of interest, especially if rapid data about antibiotic susceptibility and/or resistance and other virulence markers can be obtained (Mardis 2008). The benefit of next-generation sequencing platforms in strain-to-reference sequencing is that each DNA sequence in a library is obtained from a single genomic fragment, such that if there are rare variants in the clinical population, these can be detected by virtue of the depth of sampling obtained.

For bacteria, improvements in sequencing technology, has lead to greater insight into gene content and diversity between strains. Obtaining the pan genome of a species and as a result the core genome has revealed some interesting results. *E. coli* is a widely used model organism. Databases have been created with the purpose of tracking genomes that are available (web link 5) or characterizing the gene pool of horizontally transferred genes and virulence determinants (web link 6). In a study of 61 publically available *E. coli* and *Shigella* spp. sequenced genomes, Lukjancenko et

al found that of the predicted pan genome only 6% represented the core genome of *E. coli*.

1.2: A subset of Bacteria

1.2.1: The Proteobacteria

One of the largest and most diverse divisions of the prokaryotes was previously known as the "Purple and Photosynthetic bacteria and their relatives (Woese 1987). This was not appropriate as most of the organisms comprising this group are neither purple nor photosynthetic. In 1988 a new name was proposed after the Greek God Proteus, who can assume many forms, the "Proteobacteria" were born (Stackebrandt *et al.* 1988). The name gives credit to the vast assemblage of phenotypic and physiological traits represented within this group. Though they share the characteristic of being gram-negative the common trend does not extend much further. The Proteobacteria include many human, animal and plant pathogens as well as free-living bacteria, they can move *via* flagella or bacterial gliding and some even aggregate to form fruiting bodies. Most members are facultatively or obligately anaerobic, chemoautotrophs, and heterotrophic, but exceptions include the photosynthetic organisms. They come in a vast range of shapes and sizes and have a wide variety of metabolic systems (Madigan and Brock 2011).

Different species have been placed in this extensive group based on a number of analyses including 16S oligonucleotide catalogs, phylogenetic analysis based on full and partial sequences, rRNA cistron similarities and the results of DNA-RNA

hybridization. The main basis of classifying the proteobacterial group, however, has been their placement in a distinct clade on phylogenetic trees. Woese et al first circumscribed the group using 16S rRNA/rDNA analyses. As no overall 16S rRNA signature can be assigned to the group, it has been divided into 5 subclasses designated α, β, γ, δ and ε.

1.2.2: The Gammaproteobacteria

The *Gammaproteobacteria* as a subclass is richer in genera than any bacterial phyla bar the *Firmicutes* (Williams *et al.* 2010). Although its members exhibit a broad range of aerobicity, tropisms, morphologies and phenotypes the Gammaproteobacteria are defined solely by their 16S sequence relationships (Williams *et al.* 2010).

Phylogenetic studies of this group show that the deep branching makes it difficult to construct a well-resolved phylogeny. A single gene is insufficient for recovering deep relationships and even multigene studies have revealed instabilities in areas of the tree (Williams *et al.* 2010).

1.2.3: The YESS Group

The closely related enteric bacterial genera *Yersinia*, *Escherichia*, *Shigella* and *Salmonella* are known collectively as the YESS group. This group is medically and scientifically important as many of its members are human pathogens. For instance, *Y. pestis* is the causative agent of plague, a bacterial infection that is spread from rodents

to humans via fleas and can become airborne in a human population. In the 1300s plague killed almost one third of the population of Europe (~20 million people) (Scott and Duncan 2001). In 2011 it was shown that, due to horizontal acquisition of genes, *Y. pestis* has evolved rapidly resulting in cases of multi-antibiotic resistant strains (Bland *et al.* 2011). The pathogenic strains of *E. coli* can cause pneumonia, bacteremia, neonatal meningitis and gastroenteritis. Uropathogenic *E. coli* causes more than 90% of uncomplicated urinary tract infections (UTIs) and chance of recurrence after the first infection is 44% over 12 months (Rosen *et al.* 2007). Enteroinvasive *E. coli* (EIEC) and *Shigella* cause bacillary dysentery or shigellosis. There are an estimated 160 million cases worldwide each year, approximately 1.1 million result in deaths and of these, most involve children under the age of five (Kotloff *et al.* 1999). *Samonella* is responsible for as many as 1.3 billion cases of disease each year (Coburn *et al.* 2006) including enteric fever, enterocolitis and bacteremia.

As well as its medical significance, this group proves interesting in its complicated evolutionary history. The phylogenetic relationships of different *Shigella* strains have been the subject of intense debate. Joshua Lederberg famously said that Enterohaemorrhagic *E. coli* were '*Shigella* in a little cloak of *E. coli* antigens' (Lederberg 1998). *Shigella* are essentially *E. coli* that have acquired a virulence plasmid (VP) (Sansonetti *et al.* 1981). There are two conflicting theories on the origin of *Shigella*. The multiple independent origin theory (Pupo *et al.* 2000) suggested that *Shigella* strains formed through multiple acquisitions of the VP. The analysis of Pupo et al. found three clusters of *Shigella* strains occurring within *E. coli* and concluded that *Shigella* strains, much like EIEC, do not have a single evolutionary origin. Later it was argued that there was a single origin of *Shigella* (Escobar-Paramo *et al.* 2003).

The argument was based upon similarities between the phylogenies of genes on the VP with phylogenies for chromosomal genes. Because the phylogenies did not conflict significantly, Escobar-Paramo et al. (2003) suggested that there was a single ancestral VP that accounted for the emergence of *Shigella* and that the VP has not been horizontally transferred (as the multiple origins theory would imply). It was said that any conflicts in the trees were accounted for by transfer of fragments of the VP as opposed to the transfer of an entire VP. More recently, the two hypotheses were revisited using more robust data and support for the multiple origin hypothesis was found (Yang *et al.* 2007). Like Pupo et al., they found three major clusters of *Shigella*. They concluded that ancestral VPs entered various strains of *E. coli* and that convergent evolution explains why we see diverse *Shigella* genomes with similar phenotypic properties.

It has been previously argued that the deeper branches in prokaryote phylogeny are the source of conflict and a tree-like phylogeny may exist only at the tips (Creevey *et al.* 2004). From a study on YESS group phylogeny in 2009 (Haggerty *et al.* 2009), we concluded that "Assessing deep-level phylogenetic relationships is fraught with difficulties related to HGT and erosion of phylogenetic signal; however, assessing shallow relationships is no less difficult." Three groups were consistently recovered: the *Yersinia* group, the *Salmonella* group and the *Escherichia/Shigella* group. Beyond that there was very little agreement between trees built from different methods (Figures A4-A7, Appendix).

1.3: A Species Concept for Bacteria

1.3.1: Tree-thinking

Until the mid-eighteenth century the order of nature had been depicted as a chain of being. It was Bonnet (1764) and Pallas (1766) that first asked if invoking branches might better describe life on earth. The idea that the relationships among organisms resemble a tree-like structure was beginning to take hold. It was not until 1809, however when the first evolutionary tree was conceptualized by Jean-Baptiste Lamark (Lamarck 1809). Lamark's Figures were the first to depict the origins of various animals as a branching structure allowing for the theory of ongoing spontaneous creation of primitive forms (Figure 1.1 and 1.2). The tree metaphor was made famous in 1859 when Darwin suggested that the natural system is necessarily genealogical and represented this on an abstract tree diagram (Darwin 1859). Ernst Haeckel was responsible for promoting Darwin's ideas; he agreed that all organisms branched from one or a few original ancestors (Haeckel 1866). He used morphology to reconstruct the evolutionary history of life, and in doing so invented the terms phylogeny and phylogenetics. Haeckel's work sparked a determination to reconstruct a "universal tree of life" and the search for this unique tree continues to this day.

Carl Woese and colleagues (Woese *et al.* 1990) revolutionized the tree metaphor by using a universally distributed marker. They used indirect methods of oligonucleotide cataloguing from small-subunit rRNA (ssrRNA) to build a "Tree of Life" (Figure 1.3) and chose the root according to Gogarten (1989). The tree was extremely well received as they found a split within the prokaryotes separating

Archaebacteria from Eubacteria. Studies of prokaryotic evolution really took off following this news.

Although it has been prevalent in evolutionary studies for quite some time, the monistic Tree of Life depends on evolution following a tree-like structure across all forms of life. In reality genes are inherited vertically but can also be acquired through horizontal transfer (HGT), which does not adhere to the conventional understanding of speciation events and so is problematic when building phylogenies.

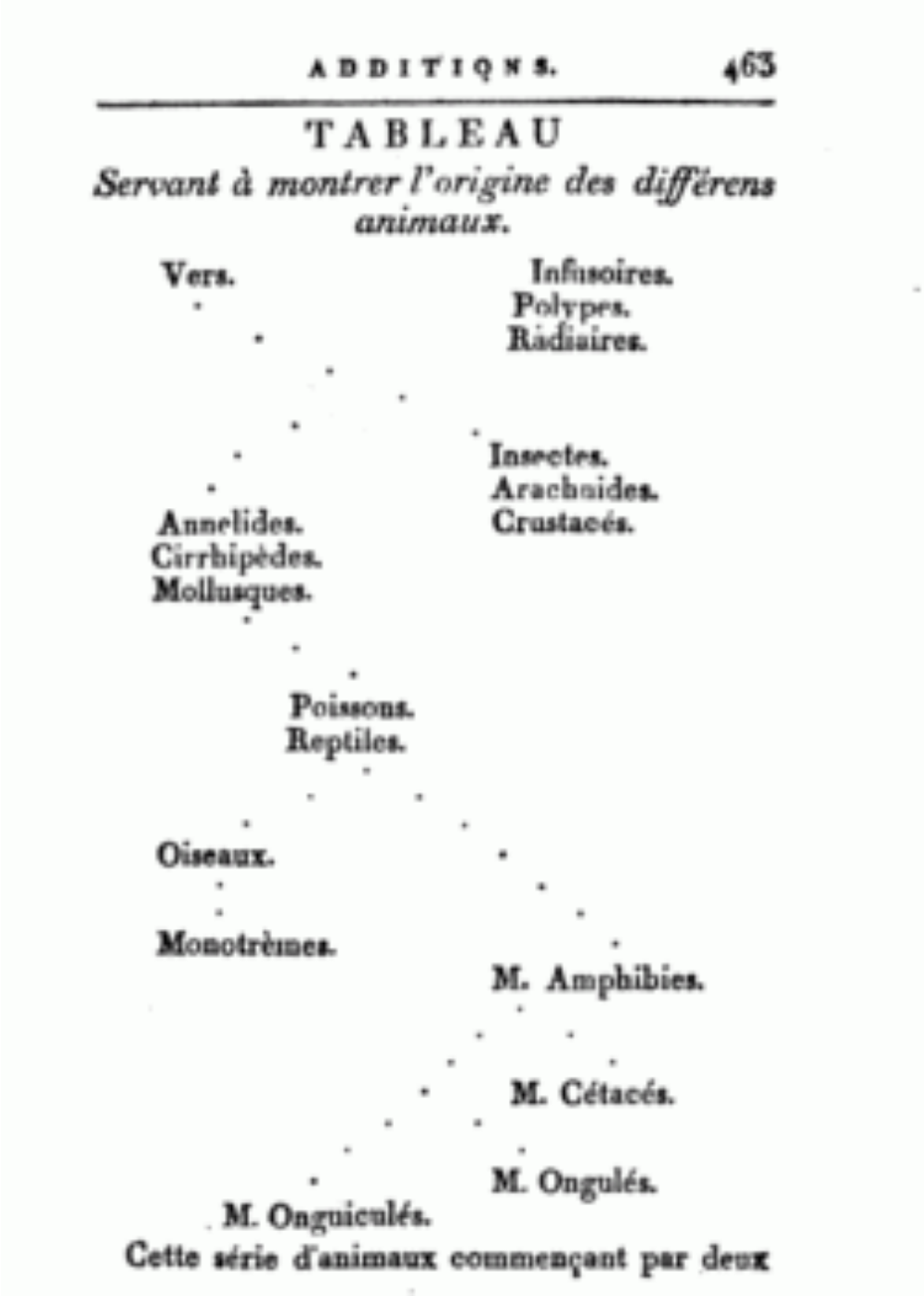


Figure 1.1: Lamarck's Tree depicting the origin of animals, from the *Philosophie zoologique* (Lamarck 1809).

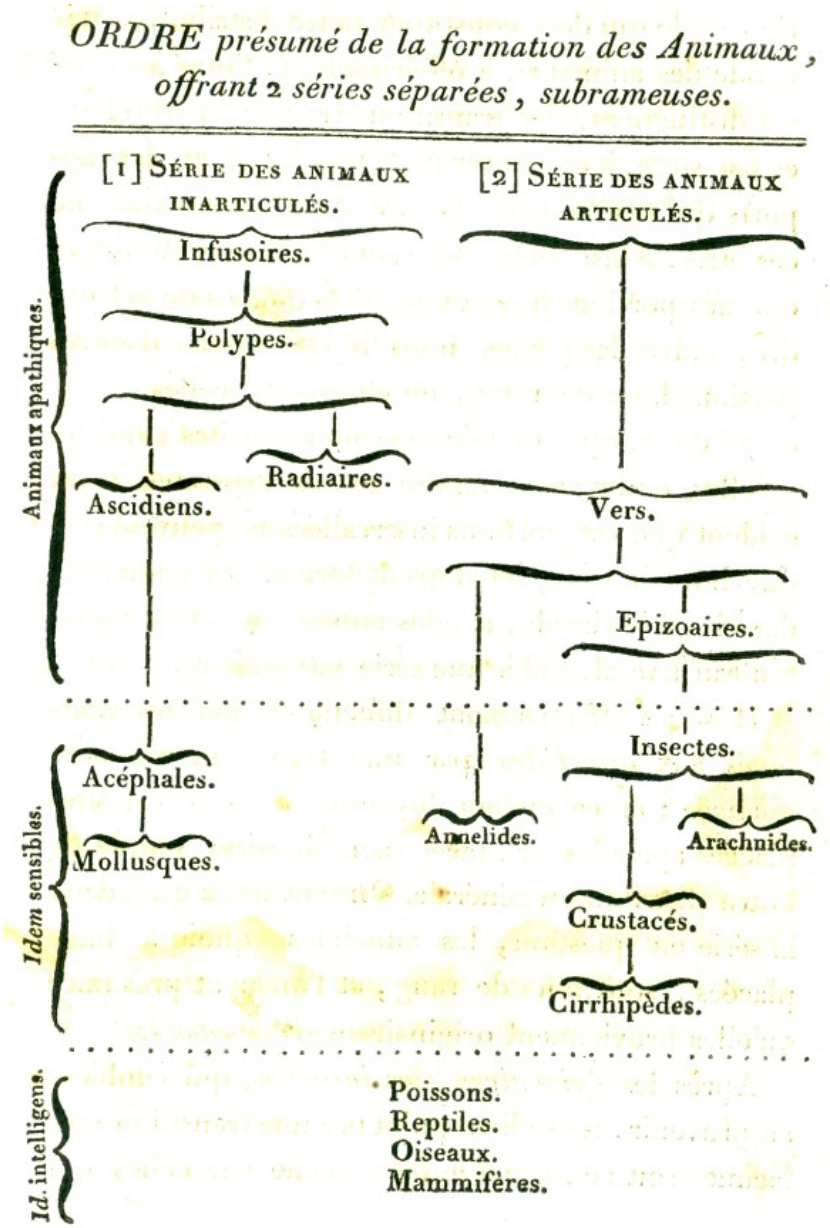


Figure 1.2: Lamarck's tree depicting two branching series of animal origins, from the *Histoire naturelle des animaux sans vertèbre* (Lamarck 1815-1822).

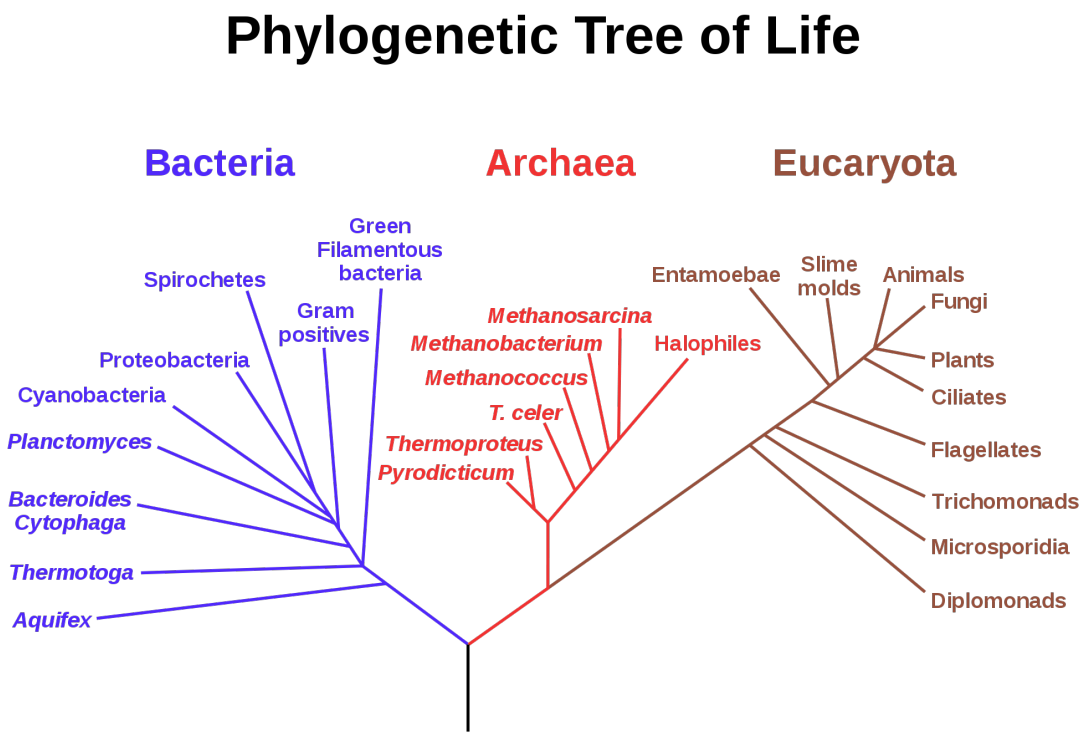


Figure 1.3: Phylogenetic tree based on small-subunit ribosomal RNA sequences showing three domains of life. Figure from Wikimedia Commons after Carl Woese and colleagues **(1990)**.

1.3.2: Horizontal Gene Transfer

For a long time it was widely accepted that bacterial serological types were fixed und unchangeable within a generation. In 1928, however a physician named Frederick Griffith made a revolutionary discovery (Griffith 1928). Griffith found that when he injected mice with a mixture of non-virulent live bacteria and virulent dead bacteria, it was in fact fatal. He used two strains of pneumococcus bacteria, the type III-S strain evades the immune system with a protective polysaccharide capsule whereas the type II-R strain has no capsule and is defeated by the immune system. When Griffith infected mice with heat killed III-S they survived, but when he added II-R the dose was deadly. Griffith concluded that there was a "transforming principle" whereby II-R was "transformed" to III-S, rendering it virulent.

It was not until 1944 that this "transforming principle" was discovered to be DNA (Avery *et al.* 1944). Avery, MacLeod and McCarty lysed the heat killed S-cells and used the lysate for transformation assays. The components of the lysate were the capsule coating, protein, RNA and DNA. Each was removed sequentially from the lysate and the resulting solution was tested for transformation capabilities. Finally they discovered that, in the absence of DNA, transformation was not possible.

This discovery gave scientists a new understanding of inheritance at the molecular level. But, more surprisingly, it provided the first suggestion that DNA is exchangeable and can alter bacterial cells, even at maturity.

Horizontal or lateral gene transfer (HGT) is any process by which an organism transfers genetic material to another cell that is not its offspring. This can occur through a number of mechanisms:

- Conjugation: is the transfer of genetic material by direct cell-to-cell contact or by a bridge-like connection. The donor cell extends a tubular structure, called a pilus, which attaches to the recipient cell. A conjugative or mobolizable genetic element allows the transfer of a single strand of DNA to the recipient cell. Finally, both cells synthesize a complementary strand to produce double stranded DNA.
- Transduction: is achieved via bacteriophages (viruses that infect bacteria). A bacteriophage infects a bacterial cell in order to use its replicational, transcriptional and translation machinery to produce virions and viral particles. Often while the bacteriophage is using the cell's machinery bacterial DNA can be inserted into the viral capsid and when the bacteriophage removes itself from the chromosome it can bring bacterial DNA along with it. The next bacterium to be infected by the bacteriophage will often receive the non-viral DNA.
- Transformation: is the stable uptake, integration and functional expression of extracellular DNA. This is the only mechanism of HGT that is independent of mobile genetic elements and bacteriophages. A cell that has acquired time-limited competence in response to environmental conditions can receive intact DNA that has been released from decomposing cells, disrupted cells or through excretion by living cells. The extracellular DNA binds, non-covalently, to the cell surface and is translocated across the membrane.

We have known about conjugation, transduction and transformation for decades now but there other methods of gene transfer continually being discovered. Genes can be transferred by temporary fusion followed by chromosomal recombination and plasmid exchange. It was shown by Dubey *et al.* (Dubey and Ben-Yehuda 2011) that bacterial

communication could be mediated by nanotubes that bridge neighbouring cells. These nanotubes were found between *Bacillus subtilis* and *Staphlococcus aureus* as well as between *B. subtilis* and the evolutionary distant *E. coli*. The release of membrane vesicles containing DNA from phages, plasmids or chromosomes can merge with nearby cells, thus allowing integration of the DNA into the new host. A Gene Transfer Agent (GTA) is a phage-like element that contains random pieces of the host genome from which it came. Genes encoding the phage-like structure are contained in the cell that produces the GTA. DNA that is successfully transferred by one of the gene transfer mechanisms is integrated into a recipient genome by recombination, the mechanism whereby segments of DNA are exchanged between two sequences (Lawrence and Retchless 2009).

1.3.3: Recombination

Recombination is the exchange of genetic material between multiple chromosomes or different regions of the same chromosome. In diploid eukaryotes, recombination between newly duplicated chromosomes during meiosis is essential for maximizing genetic diversity. It is also important in DNA repair and in DNA replication where it assists in filling gaps and preventing stalling of the replication fork.

Recombination in prokaryotes is essential in the incorporation of acquired DNA into a recipient genome and in DNA repair and replication. After DNA has been injected into the cytoplasm of the recipient cell it is subjected to processes that either allow the DNA to be integrated into the cell or remove it from the cell altogether.

Firstly restriction endonucleases cleave and exonucleases degrade the double stranded DNA (dsDNA) to reduce its size. Following this, if the donor DNA has sequence similarity with the recipient then homologous recombination can occur and the DNA is integrated into the chromosome, replacing the resident allele at its cognate position. However, if there is no sequence similarity, illegitimate recombination occurs and the DNA is placed anywhere in the chromosome or site-specific recombinases catalyse recombination into specific locations.

Homologous recombination occurs where the donor DNA is highly similar to the recipient DNA. In prokaryotes homologous recombination has been best studied in *E. coli* and shown that it can occur by two different pathways: one for the repair of double stranded DNA (dsDNA) and another for single stranded DNA (ssDNA). dsDNA breaks are repaired by he RecBCD pathway (Michel *et al.* 2007). RecBCD is a three-subunit enzyme complex, it initiates recombination by binding to the broken dsDNA. It then begins to unzip the DNA duplex through helicase activity until it reaches a complex called the Chi site (χ-site). The χ-site is a short stretch of DNA found in the genome of a bacterium and is unique to each group of closely related organisms, e.g. enteric bacteria share the sequence 5'-GCTGGTGG-3' (Dillingham and Kowalczykowski 2008). At the χ-site the unzipping is halted and restarted at a slower rate. RecA is then loaded onto ssDNA cut from the duplex. The RecA coated nucleoprotein filament searches for an area of homology elsewhere in the genome and moves into the recipient duplex by strand invasion. During strand invasion the recipient DNA is cut and the invading strand inserted to create a Holliday junction. The RuvAB complex arranges itself around the junction. Strands from both duplexes are unwound on the surface of the RuvAB complex as they are guided from one

duplex to the other. The Holliday junction is resolved to form two recombinant DNA molecules with reciprocal genetic types (Kowalczykowski *et al.* 1994).

For the repair of ssDNA the RecF pathway is utilized. RecQ unwinds the DNA and RecJ nucleases degrade the 5' strand. RecA binds to the 3' strand and the nucleoprotein filament searches for a homologous sequence. The Holliday junction is formed as in the RecBCD pathway and a new duplex is formed where the donor DNA has replaced the broken strand (Morimatsu and Kowalczykowski 2003).

Non-homologous recombination occurs by ligation of break ends without the need for a template. Short sequence repeats (SSRs) act as a guide for repair of damaged DNA. SSRs are often present as single-stranded overhangs on the end of the double-strand breaks (Weller *et al.* 2002). Bacteria often lack the proteins required for non-homologous recombination but they have been discovered in *B. subtilis* and *Mycobacteria* (Moeller *et al.* 2007). Bacterial non-homologous recombination is mostly utilized by bacteria that spend a significant portion of their life cycle in stationary haploid phase, where no there is no template available.

1.3.4: HGT and Bacterial Phylogeny

Although it has been seen among Protists, most eukaryotes, particularly animals and fungi, are largely unaffected by HGT (Andersson 2005). In 2001 the scientific world was stirred by reports of an unusual number of genes in the human genome that appeared to have been acquired through horizontal transfer (Lander *et al.* 2001; Salzberg *et al.* 2001). These studies were quickly proved to be flawed and so the focus centered on HGT in prokaryotes (Stanhope *et al.* 2001).

In bacteria barriers to HGT are very low (McInerney *et al.* 2008), and so genes are exchanged frequently (Gogarten *et al.* 2002; McDaniel *et al.* 2010; Popa *et al.* 2011). The process does not require the donor and recipient to be of the same species, and since it can be achieved via an intermediate such as a bacteriophage, it allows the exchange of genes between strains far outside closed gene pools (Ereshefsky 2010). It has been estimated that 81 ± 15% of all genes in prokaryotic genomes have undergone HGT at some point in their evolutionary history (Dagan *et al.* 2008). Other estimates for the proportion of protein families affected by HGT range between 60 (Kunin *et al.* 2005; Dagan and Martin 2007) and 90% (Mirkin *et al.* 2003). There are also studies which find these frequencies to be much lower (Ge *et al.* 2005). It was shown however, that of 246,045 genes that were transferred into *E. coli* via a plasmid, 99.4% were successfully integrated into the recipient cell (Sorek *et al.* 2007).

HGT confuses and confounds prokaryotic relationships by implying different, incongruent relationships within a set of taxa. Some authors question the meaning of trees as a representation of the evolutionary history of species affected by HGT (Gogarten *et al.* 2002; Doolittle and Bapteste 2007). Many have accepted that HGT was a ‘‘rampant’’ phenomenon concluding that ‘‘the history of life cannot properly be represented as a tree’’ (Doolittle 1999).

1.4: Alternatives to the Tree Of Life Hypothesis

There is much need for a theory or system that recognizes more than one ultimate principle when describing prokaryotic evolution. By employing the attitude that conflict between two different models does not necessarily invalidate one we can describe every aspect of organization within the prokaryotes.

1.4.1: The "Forest Of Life"

Puigbo et al. (2009) constructed 6,901 maximum likelihood trees from prokaryotic genes in an attempt to find a central trend that could be considered an approximation of the tree of life (Figure 1.4). An inconsistency score, measuring how representative a given tree is of the whole forest, allowed them to see a trend without a given species tree. They named the universal or close to universal trees NUTs (nearly universal trees). The 102 NUTs agreed quite well; at a 50% similarity cutoff, they found that almost all trees agreed with almost all others. They also found that, in many cases, the topology of the NUTs was similar to others in the forest. Within the forest however, there seemed to be a lot of inconsistency, where the shallow relationships appeared to remain constant, at a deeper level they were no more similar than a random dataset. They conclude a weak vertical trend displayed by the NUTs. It has been shown before that a vertical trend may only exist at the tips of prokaryotic trees and that there is very little confidence in deeper relationships (Creevey *et al.* 2004). The fact that many of the gene trees display different topologies has made it clear that independent processes have impacted the evolutionary history of genes and genomes.

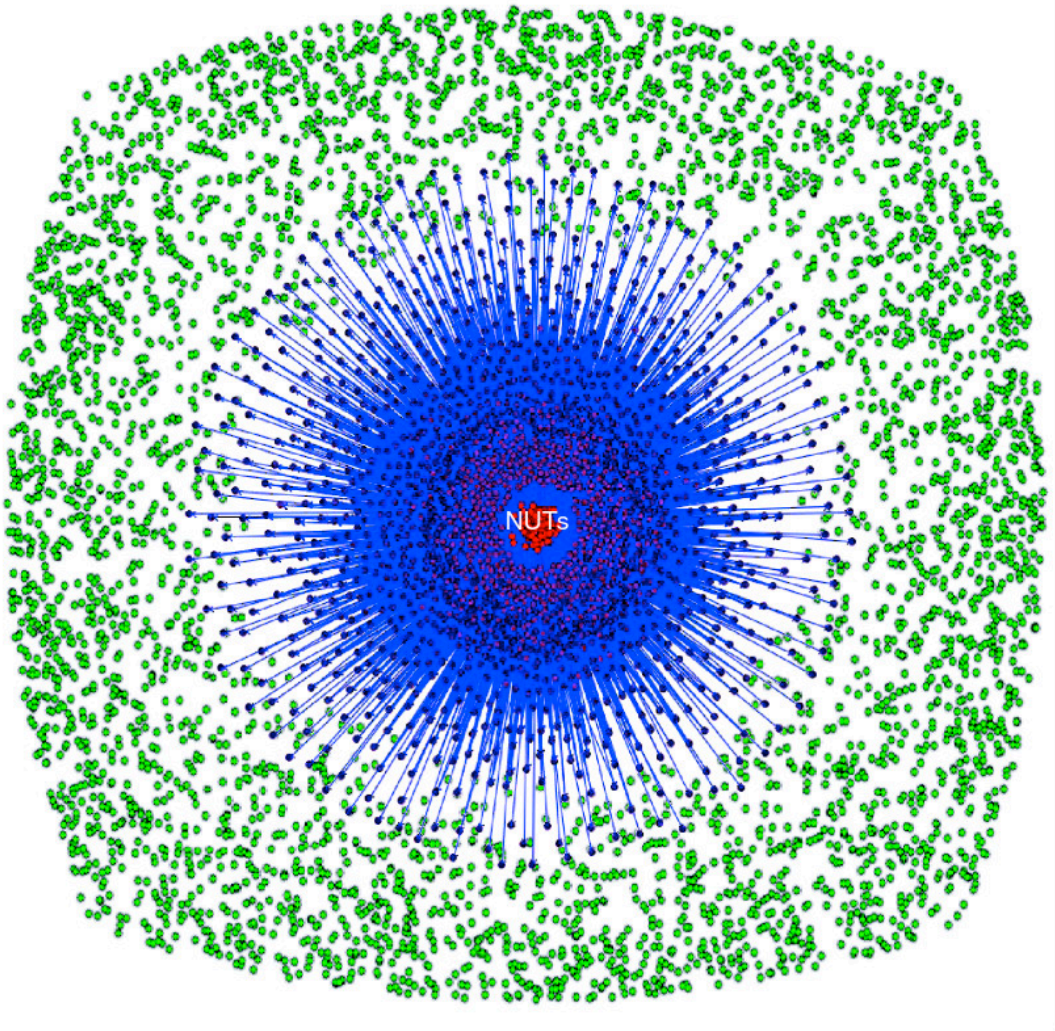


Figure 1.4: Network representation of the 6,901 trees of the forest of life. Each node represents one gene tree and the edges connect trees with at least 50% similarity between their topologies. The 102 NUTs are shown as red nodes located towards the centre of the network, these are quite densely connected; there is a high level of similarity between the NUTs. Purple nodes are non-NUTs that are connected to at least one NUT, i.e. have at least 50% similarity to one or more of the NUTs, the connections to these are somewhat sparser. The rest of the trees are shown as green nodes and are less than 50% similar to the NUTs in topology, there has been sufficient HGT to cause a large amount of inconsistency between gene tree topologies. Taken from Puigbo et al. (Puigbo and Koonin 2009).

1.4.2: The "Net Of Life"

Hilario and Gogarten (1993) were the first to propose a "net of life". They used three types of the ATPases to root the "universal tree of life": vacuolar, archaebacterial and eubacterial. An archeabacterial type ATPase was found in the eubacteria, and vice versa, suggesting that both types of ATPase may have already been present in the last common ancestor. They suggest that horizontal gene transfer can explain the data.

The rooted net of life genome phylogeny (Williams *et al.* 2011) accommodates for numerous examples of reticulated histories. An initial scaffold of predominately vertical descent is inferred from a supermatrix of combined ribosomal genes. Unrooted phylogenies of gene families are then superimposed over the scaffold.

1.4.3: The "Rhizome Of Life"

Merhej at al. (2011) propose a new representation for the evolutionary history of *Rickettsia felis* in the form of a rhizome. In a comparison to ten other *Rickettsia* genomes they found *R. felis* to be a collection of genes potentially having different evolutionary histories. Although the majority of genes agreed with the phylogeny based on *Rickettsia* core gene concatenation, the data showed that 12% of the *R. felis* genome comes from non-vertical inheritance. Multiple origins of the *R. felis* gene repertoire make it impossible to represent the evolutionary history of this genome as a tree. Merhej and colleagues use a rhizome of life to show multiple roots and intertwining origins of currently living species.

Many now agree that HGT is such a powerful force that the evolutionary history of the prokaryotes would be better represented using a network in which edges represent HGTs.

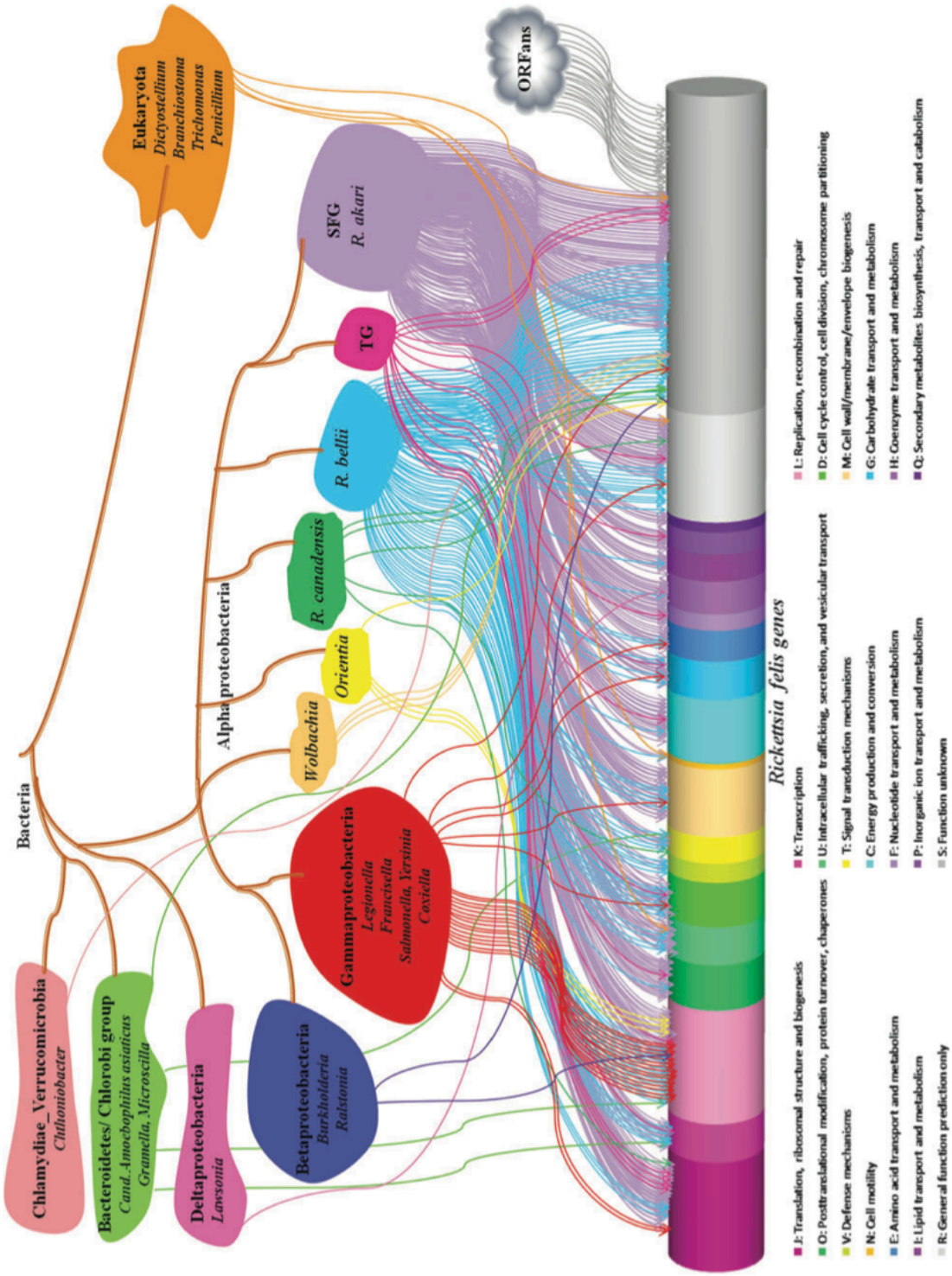


Figure 1.5: The different origins of *R. felis* genes. The genome of *R. felis* is represented at the bottom of the Figure, the genes classified into functional categories of COGs. Arrows link *R. felis* genes to its closest related species on the organismal phylogeny (top of the Figure) based on the corresponding gene phylogeny. Taken from Merhej *et al.* (2011).

1.5: Networks

In this thesis I refer to the structure and analysis of a number of networks. Therefore it is important to establish the definition and mathematics of networks. In theory any economic agent has the potential to interact with any other, directly or indirectly (Newman 2010). In reality these agents will have preferences and biases depending on need and social position. The patterns of interactions between agents form a network. A network is a set of vertices or nodes connected via edges. It can be directed or undirected, weighted or non-weighted, cyclic or acyclic etc. (Newman 2010).

An empirical study is required to reveal the relationships between agents and from this, the network structure is created. A network has properties: size, overall connectivity, mean distance from any agent to any other agent, etc. The agents also have properties pertaining to the number of relationships they have with other agents. These properties reveal how important an agent is in the network and which communities it belongs to. With the structure in place, mathematical analyses will answer questions about the agents and their relationships.

1.5.1: Network Centralities

Hereafter I will refer to agents as nodes and the relationships between them as edges on the network. Measures of centrality of a node indicate its importance in the network, in other words, who in the network is most central and therefore important? (Newman 2004) (Figure 1.6). Each node in a network has a number of interactions, known in total as the degree of the node. The degree of a node, therefore, is the

number of edges directly connected (Borgatti 2005). This centrality measure is representative of a node's direct relationships, i.e. its neighbours that are no more than one edge away. The location of a node within the network is also important. In a network a path is a sequence of nodes bridged by edges. Therefore even though a node may not have a direct connection with another node it may be reached "through the grapevine". A geodesic path then, is the shortest path between two nodes in terms of the number of intervening nodes traversed (Borgatti 2005). A measure called "closeness centrality", like degree centrality, helps to reveal the most central nodes in the network but unlike the degree it includes information from all relationships to a node, direct and indirect. The closeness centrality is the mean geodesic distance from one node to all other reachable nodes. Also reliant on paths in networks is the measure of "betweenness centrality". Betweenness is a kind of measure of flow in the network, it reveals how often a node lies on the shortest path between two random nodes in the network (Borgatti 2005). To discover how "between" a node is on a network, one must find all geodesic paths in the network and calculate the number on which the falls.

The nature of geodesic paths gave rise to the small-world network concept (Boccaletti *et al.* 2006). In general the mean geodesic path is small compared to the size of the network, if you think of network size in terms of the number of nodes. Stanley Milgram famously discovered that if he asked a random person to get a message to a specified target, the message would pass through an average of 6 people (Milgram 1967). Thus coining the phrase "six degrees of separation". Evidence suggests that in most real-world networks nodes tend to create tightly knit groups characterized by a relatively high density of connections, this likelihood tends to be greater than the

average probability of a connection randomly established between two nodes (Jin *et al.* 2001).

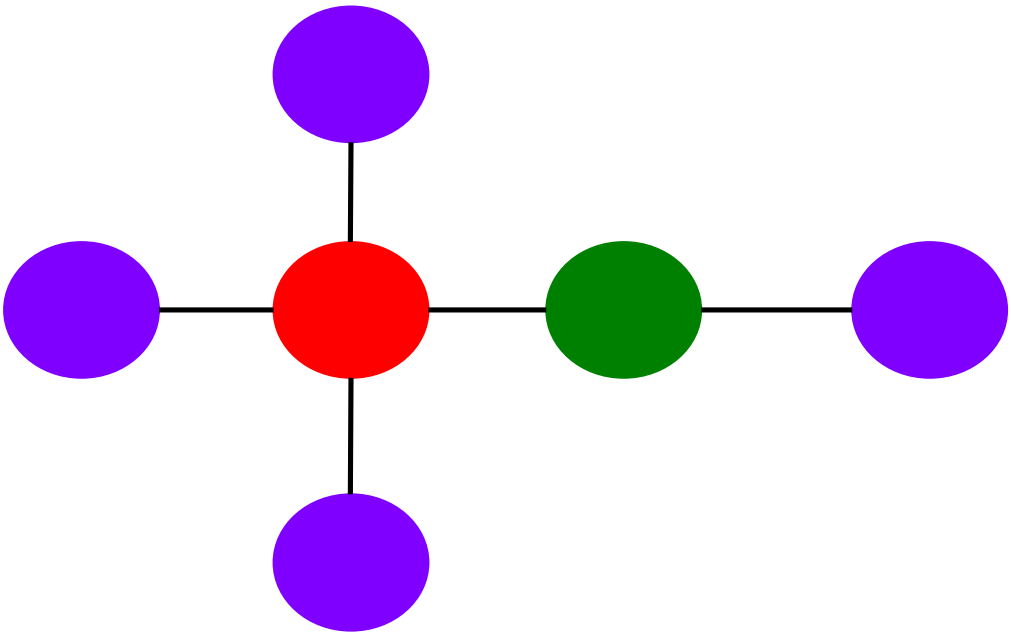


Figure 1.6: Network centrality measures: The red node has a high degree- it has 4 direct neighbours compared to the green node that only has 2 direct neighbours. The red node also has a higher closeness than the green. The red node can reach all other nodes bar one by passing through just one edge. The green node however has the highest betweenness centrality as nodes must pass through it to reach the purple node on the far right.

1.5.2: Communities in Networks

Communities, clusters or modules are found where there are more dense connections or a higher concentration of edges within groups than between groups (Porter *et al.* 2009). Nodes within a community will share some kind of common trait.

The clustering coefficient of a node characterizes the density of connections surrounding the node (Soffer and Vazquez 2005). The clustering coefficient is calculated as the ratio between the total number of edges connecting a node's nearest neighbours and the total number of all possible edges between those neighbours. This is a measure of the mutual acquaintance of a node's "friends", in other words it asks: of a node's "friends" how many are also "friends" with one another? The clustering coefficient can also be thought as the cliqueishness of a node. A clique is a maximally connected subgraph, i.e. all nodes are connected to all other nodes. If a node has a high clustering coefficient it is likely to be contained in a clique.

A network may consist of many disjoint parts, these are known as connected components (Hopcroft and Tarjan 1973). On a connected component all nodes are mutually reachable one way or another and the size of the component is simply the number of nodes it contains. The giant connected component is the one with the most nodes.

1.5.3: Networks in Biology

The use of networks in biology is more prevalent than is commonly thought and dates back to the 18th century. Comte de Buffon made use of networks to describe animal breeding and the diversity of forms it could produce (Loveland 2004). His network of genealogical relationships among dog breeds is his most well known (Figure 1.7). Later in the 18th and early in the 19th century networks were used to describe the detail of affinities among organisms. They included affinities among plants (Ruling 1774) (Figure 1.8), animals (Hermann 1783) (Figure 1.9) and vegetables (Batsch 1791) (Figure 1.10). In an attempt to resolve the problems in bringing algae to a phylogenetic system, Georg Klebs (Klebs 1892) made the network of relationships among groups of algae and protozoa (Figure 1.11).

There are a number of areas in biology that rely on networks to answer questions. In order to understand cells and diseases at a system-wide level many are turning to the study of protein interactions (Pellegrini *et al.* 2004). The full view of interacting proteins is best displayed on a network (Zhang 2009). On a protein interaction network nodes are proteins and the edges indicate a physical interaction between the two it connects (Figure 1.12).

The metabolism of an organism encompasses the basic chemical system that generates essential components such as amino acids, sugars and lipids, and the energy required to synthesize them and to use them in creating proteins and cellular structures (Reddy 2007). A metabolic network represents all chemical reactions and physical process performed by a cell (Figure 1.13). Nodes on a metabolic network are the enzymes and metabolites involved in various processes; the edges indicate the

relationship between them. Sub-networks are used to describe subsystems of metabolism and pathways of enzymatic activity (Jeong *et al.* 2000).

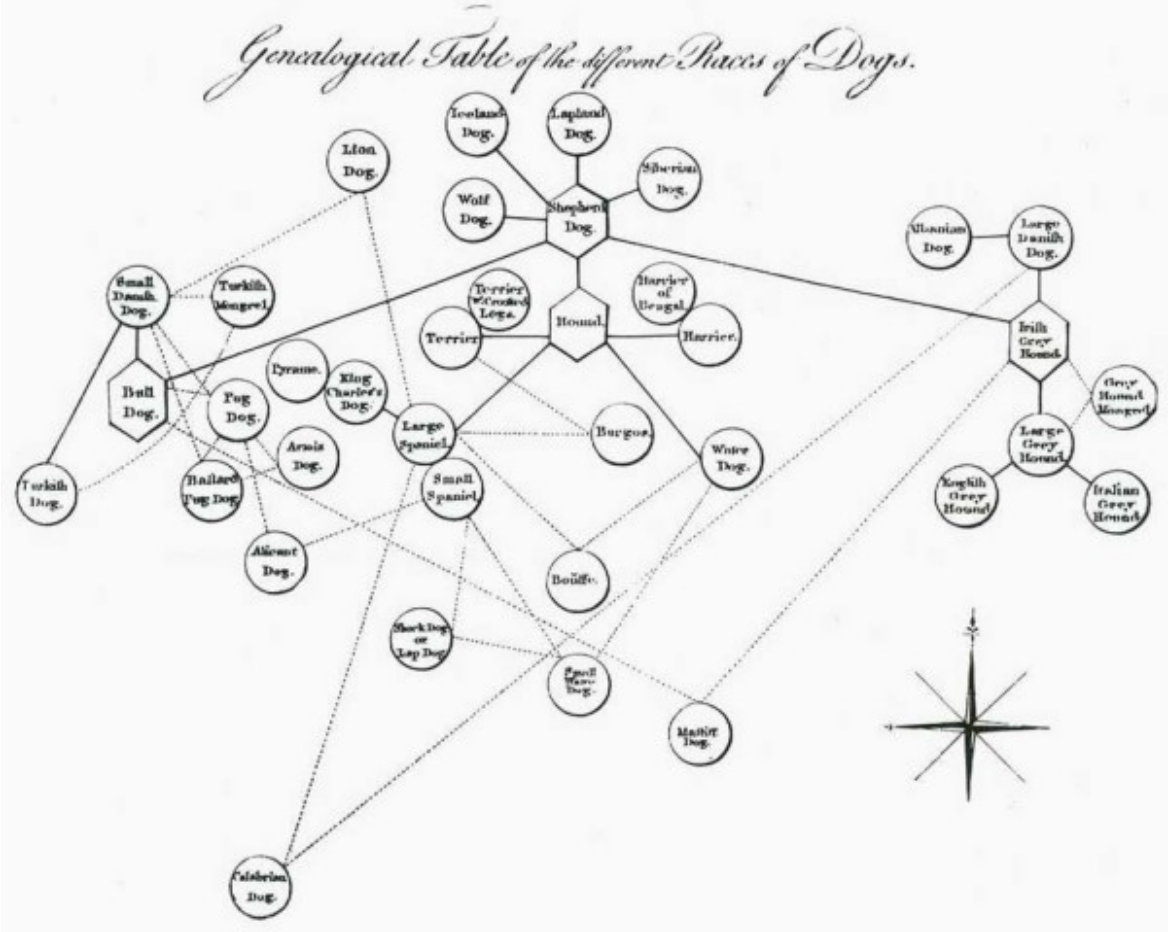


Figure 1.7: Network of genealogical relationships among breeds of dogs, from *Histoire Naturelle* of Georges-Louis Leclerc, (Loveland 2004)

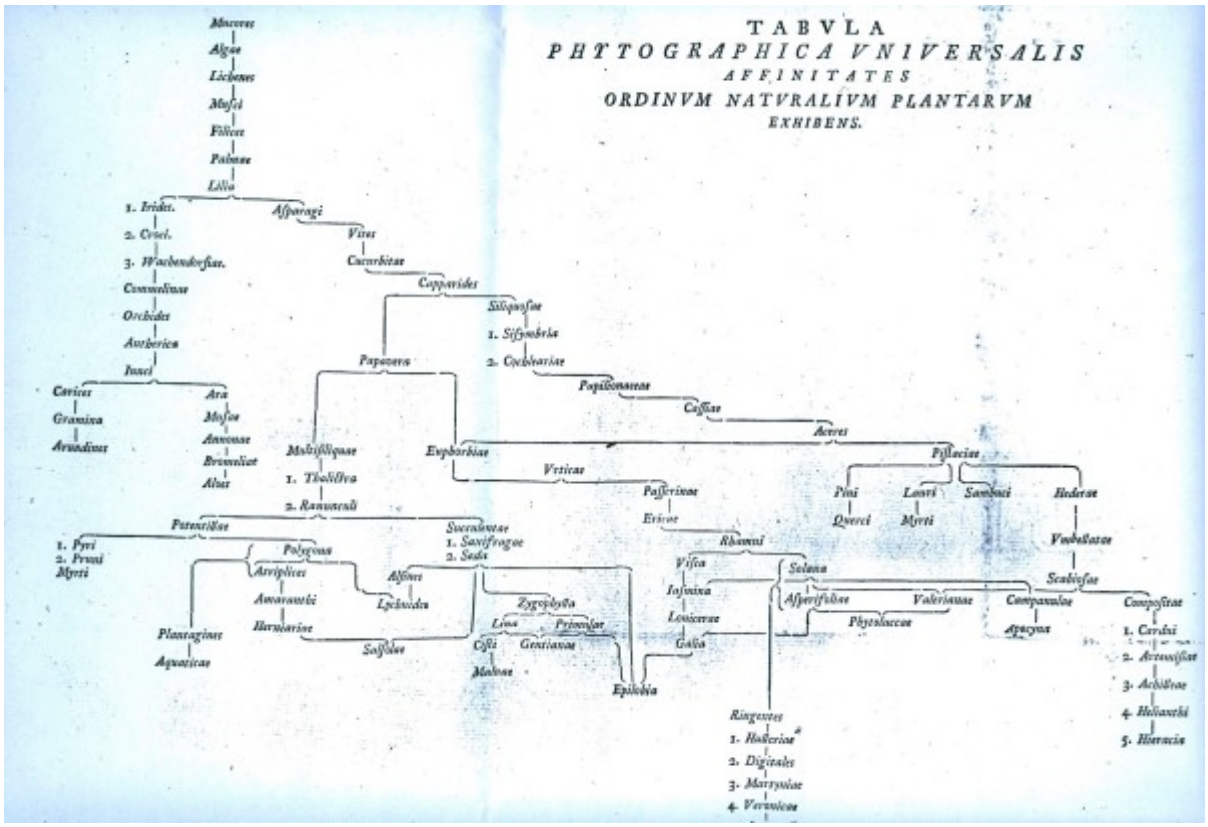


Figure 1.8: Network of affinities among the natural orders of plants, from the *Ordines naturales plantarum commentatio botanica* (Ruling 1774).

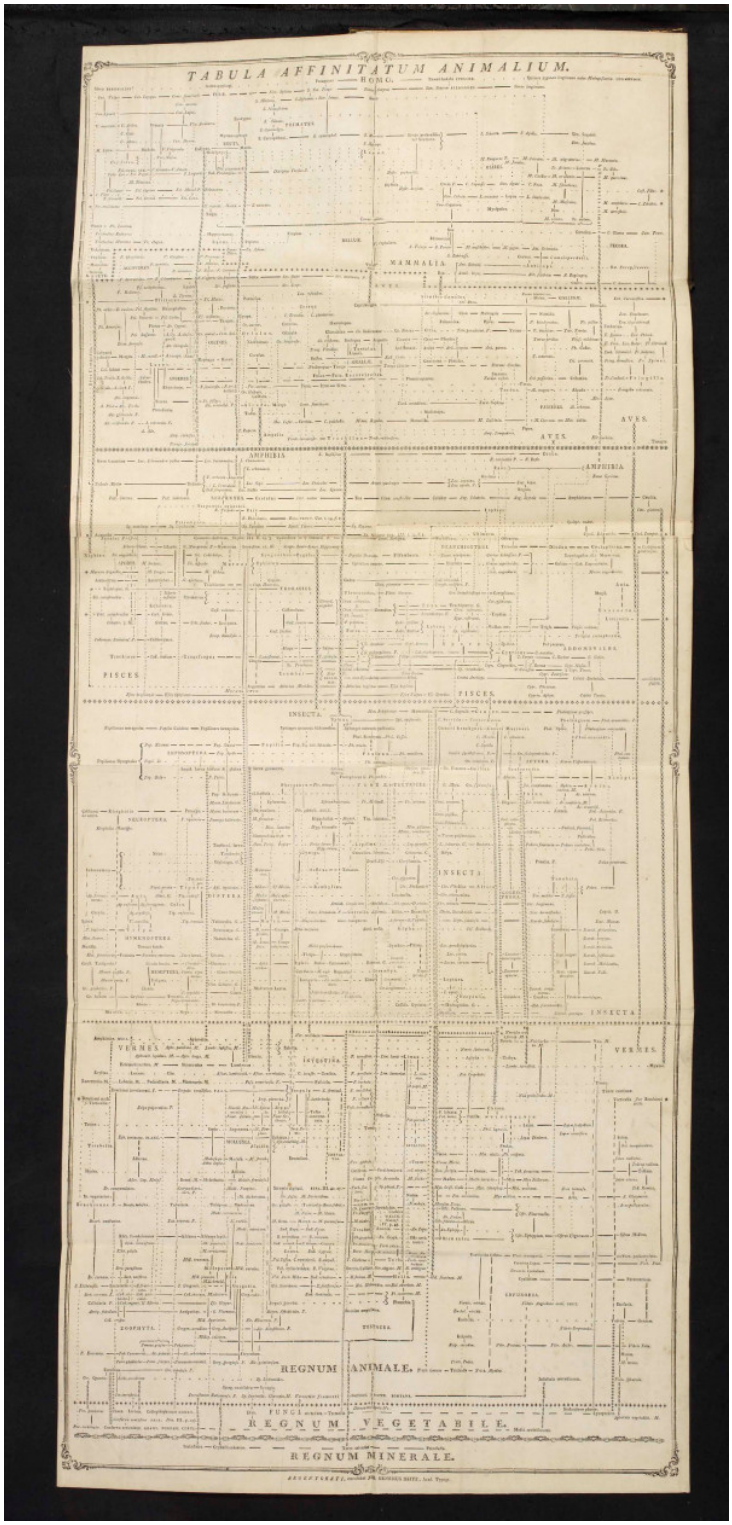


Figure 1.9: Network of affinities among animals, from the *Tabula affinitatum animalium* (Hermann 1783).

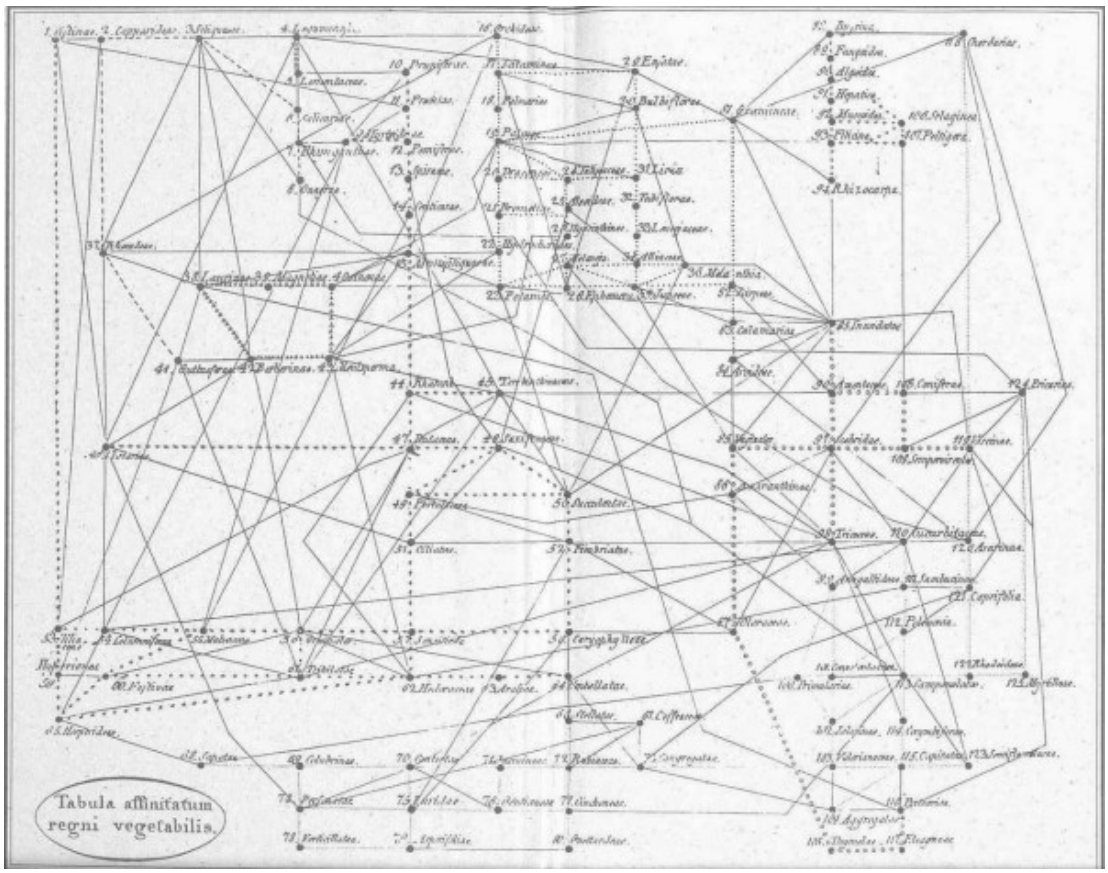


Figure 1.10: Network of affinities within the vegetable kingdom, from the *Tabula affinitatum regni vegetabilis* (Batsch 1791).

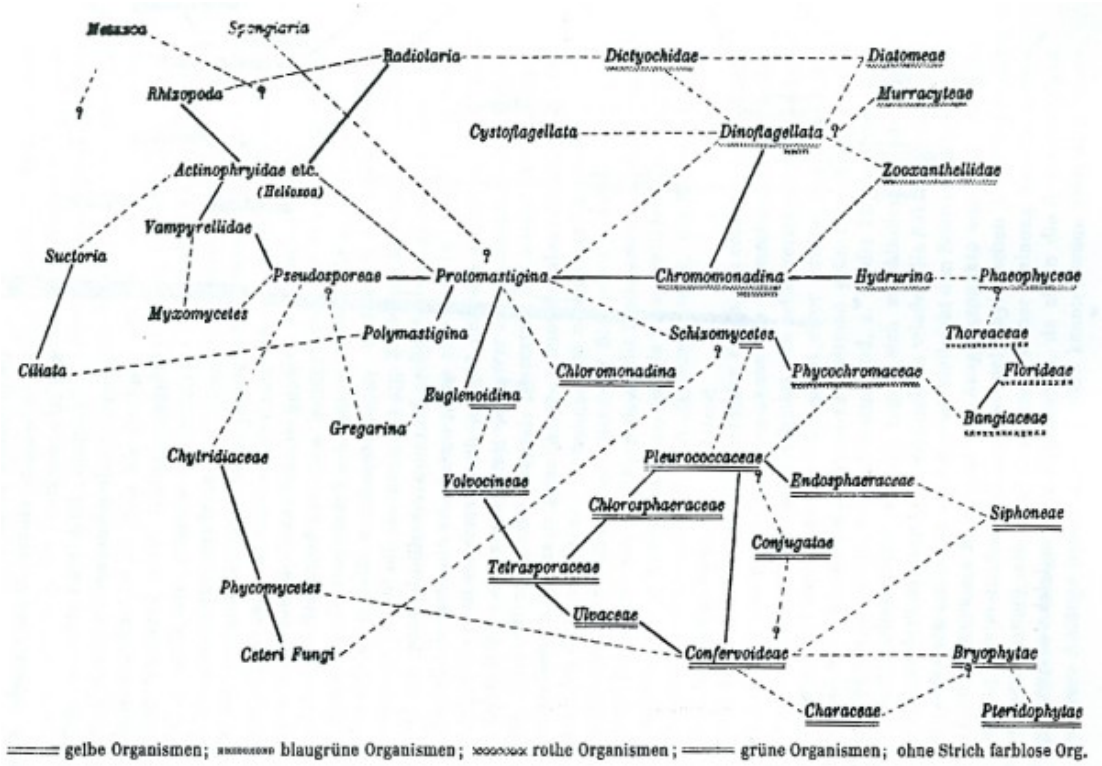


Figure 1.11: Network of lines of relationships among groups of algae and protozoa, by Georg Klebs (1892).

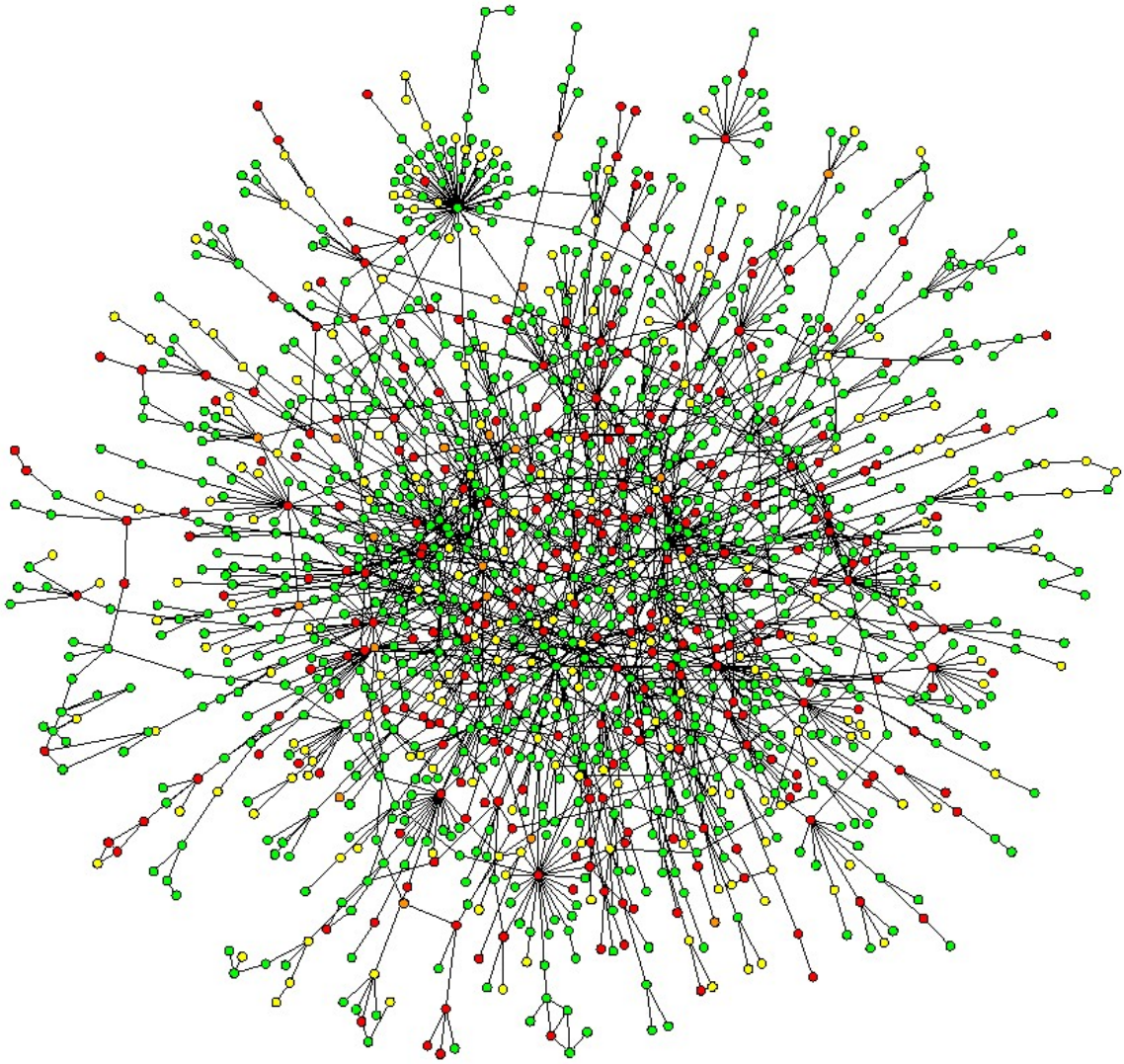


Figure 1.12: Yeast protein-protein interaction network. The nodes represent 1,870 proteins and the edges indicate the 2,240 direct physical interactions between the they connects. The colours signify the phenotypic effect of removing the corresponding protein (red, lethal; green, non-lethal; orange, slow growth; yellow, unknown) (Jeong *et al.* 2001).

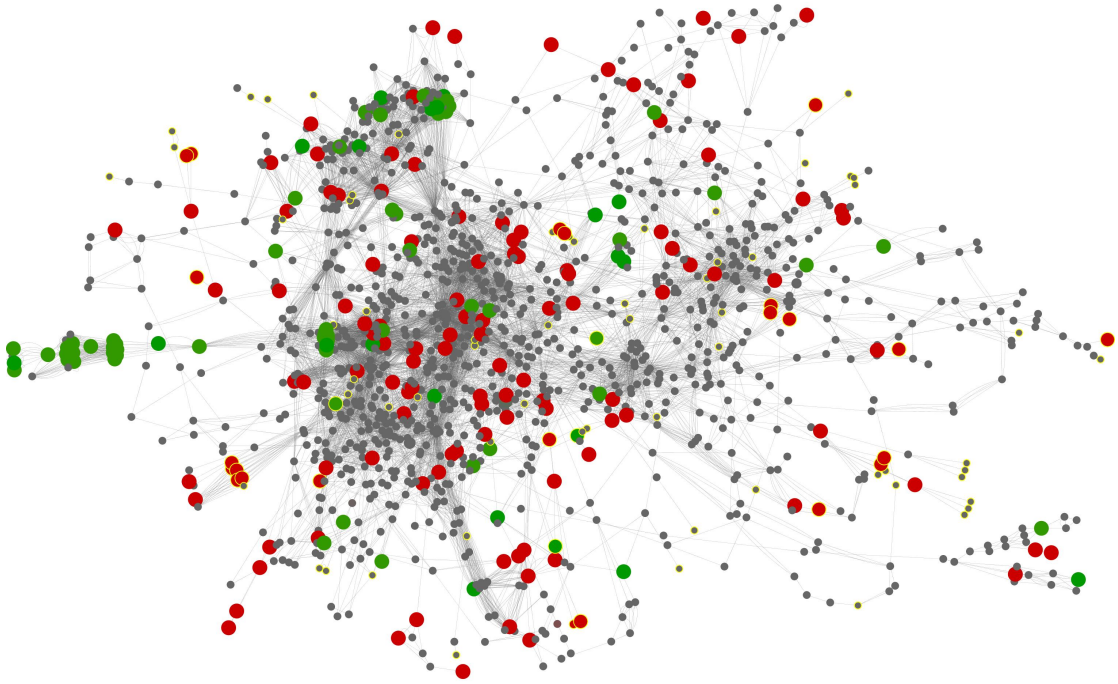


Figure 1.13: A community-level metabolic network of the gut microbiome. Nodes represent enzymes and edges connect enzymes that catalyze successive metabolic steps. Enzymes associated with obesity appear as larger colored nodes (red=enriched, green=depleted). Taken from (Greenblum *et al.* 2012).

1.5.4: Phylogenetic Networks

A phylogenetic network is any graph used to visualize evolutionary relationships between nucleotide sequences, genes, chromosomes, genomes or species (Huson and Bryant 2006; Huson and Scornavacca 2011). The advantage of using phylogenetic networks over phylogenetic trees is the ability to include hybrid nodes i.e. nodes with two parents. Phylogenetic networks are currently used when describing the outcome evolutionary processes that are non-tree like in nature e.g. recombination, genome fusion and HGT (Huson and Bryant 2006; Huson and Scornavacca 2011). These networks can also be used for tree-like analyses where the vertical signals from the data are conflicting with one another (Huson and Bryant 2006).

Splits networks are reconstructed from bipartitions in of taxa implied by the given data (Bryant and Moulton 2004; Huson and Scornavacca 2011). Incompatible splits are those splits that do not agree with the phylogenetic tree of the data and the compatible splits are those that do agree with the tree. All splits are represented on the network, rendering it more informative than a tree of strictly vertical signal (Bryant and Moulton 2004; Huson and Scornavacca 2011).

A splits network was used in a study of *Euglena gracilis* (Ahmadinejad *et al.* 2007) (Figure 1.14). This unicellular flagellate protist has a chimeric genome with some genes inherited from its heterotrophic host and some from a photoautotrophic endosymbiont during secondary endosymbiosis. Ahmadinejad et al. sequenced 2,770 ESTs from the *E. gracilis* genome and found 841 to have eukaryotic homologs, 117 of which are specific to the photoautotrophic eukaryotes. A tree was inadequate to describe their findings so Ahmadinejad and colleagues used a network to show the common origin of *E. gracilis* from kinetoplastid and photoautotrophic ancestors.

The use of networks is becoming more popular in phylogenomics- the study of phylogenetic relationships at the whole genome level (Kunin *et al.* 2005; Dagan and Martin 2007; Lima-Mendez *et al.* 2008; Kloesges *et al.* 2011). Phylogenomic networks are reconstructed from presence or absence patterns of genes. On these phylogenomic or gene-sharing networks the nodes represent genomes. An edge is represents the presence of at least one gene found in common between the two genomes it connects. The edges are weighted based on the number of shared genes or the number of orthologous gene families that are present in both genomes. A phylogenomic network can be reconstructed from complete genomes (Kunin *et al.* 2005; Dagan and Martin 2007; Kloesges *et al.* 2011), plasmids (Fondi *et al.* 2010; Halary *et al.* 2010), phages (Lima-Mendez *et al.* 2008; Halary *et al.* 2010) or even metagenomes (Halary *et al.* 2010).

Halary et al. (2010) reconstructed a phylogenomic network from 111 eukaryotic and prokaryotic genomes along with thousands of phage and plasmid sequences (Figure 1.15). They found that different protein families have their distribution limited to certain vehicles i.e. chromosome, phage or plasmid.

HGT networks are used to study the horizontal component of microbial evolution (Beiko *et al.* 2005; Dagan and Martin 2007; Dagan *et al.* 2008; Kloesges *et al.* 2011; Popa *et al.* 2011). These networks are reconstructed from HGT events inferred from genomic data. On HGT networks the nodes are external and internal nodes of a reference species tree and edges are HGT events between the nodes they connect. Dagan et al. (2008) reconstructed a HGT network from 181 fully sequenced microbial genomes (Figure 1.16) and discovered that, on average, 81±15% of the proteins in each genome are affected by HGT at some point in their evolutionary history.

Beiko at al. (2005) summarised all HGT events inferred from 22,432 phylogenies of orthologous protein families in 144 prokayotic genomes. On their network the nodes represented 21 higher taxonomic groups of microbes and the edges were HGT events. They described 1,398 HGT events on their network and found that 56% of the transfer events were between the Alpha-, Beta- and Gamma-Proteobacteria.

As can be seen, networks are versatile frameworks that can be used for a wide variety of evolutionary questions. They provide perspectives that are often different to those perspectives provided by phylogenetic trees and can often account for additional evolutionary events that are unseen by phylogenetic trees. As a consequence, it was decided to explore networks more thoroughly in an effort to expand their usefulness in evolutionary biology.

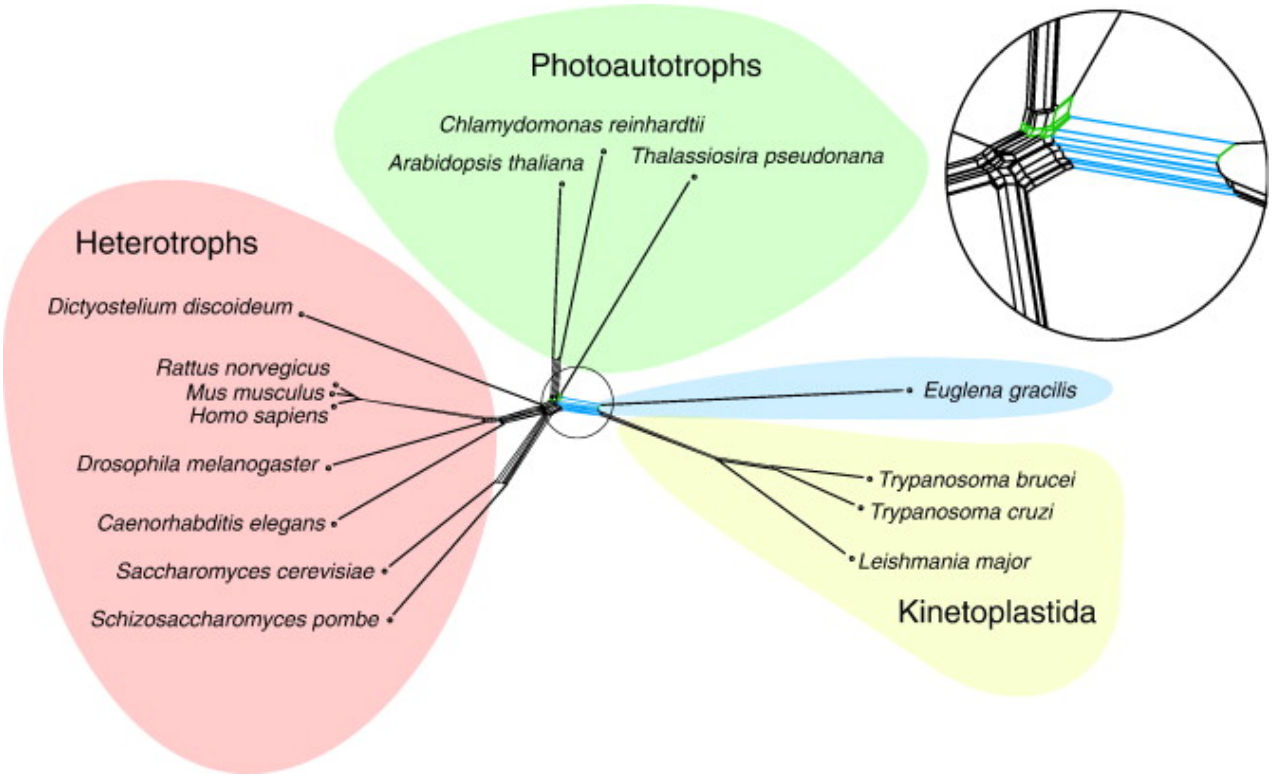


Figure 1.14: A phylogenetic network reconstructed for the concatenated alignments of 259 globally distributed genes. Splits represent disagreement between gene tree topologies. An accumulation of splits indicates that the tips were derived from multiple ancestors. Taken from (Ahmadinejad *et al.* 2007).

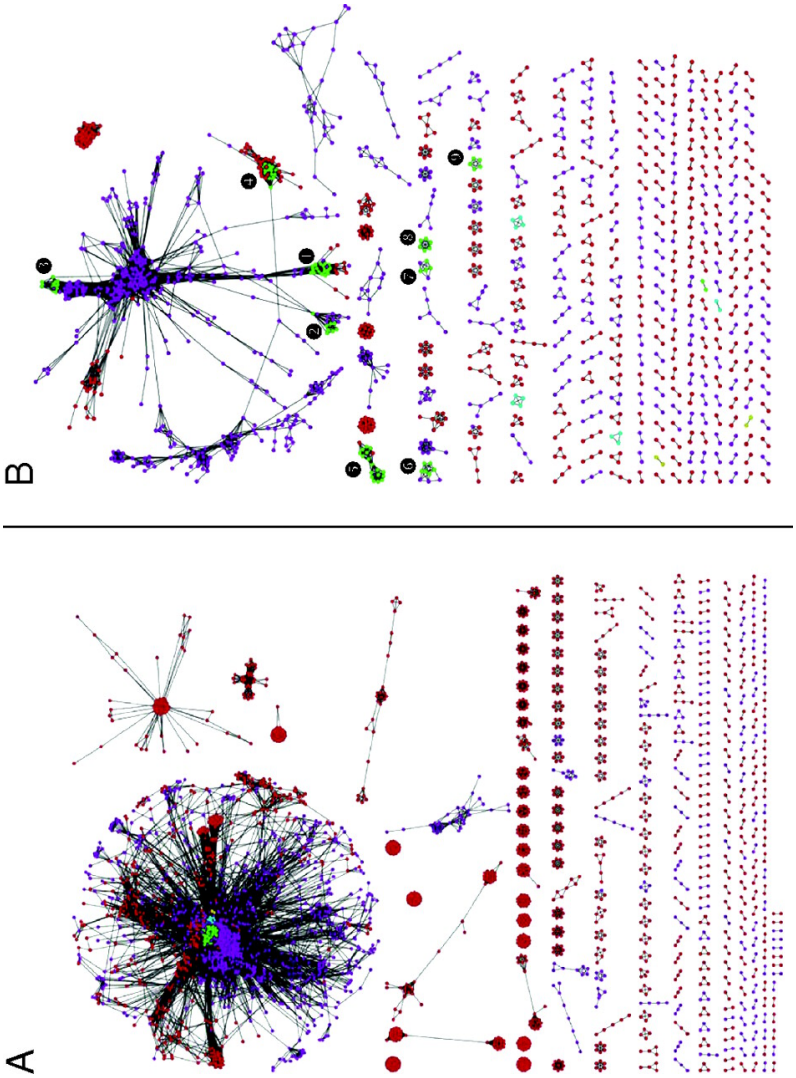


Figure 1.15: Network of shared DNA families among cellular, plasmid, and phage genomes. (*A*) Global network in which each node represents a genome. Two nodes are connected by an edge if they share homologous DNA. Genomes sharing larger proportions of their DNA are closer together and the density of the giant connected component indicates a high level of sharing between most a large number of the genomes in the datset. (*B*) Global network displaying connections between genomes for a minimum of 95% sequence identity. Using only genes at this level of sequence identity roughly filters for recent sharing events. Clusters of nodes of the same colour are indicative of certain protein families having preference for a particular DNA vehicle.

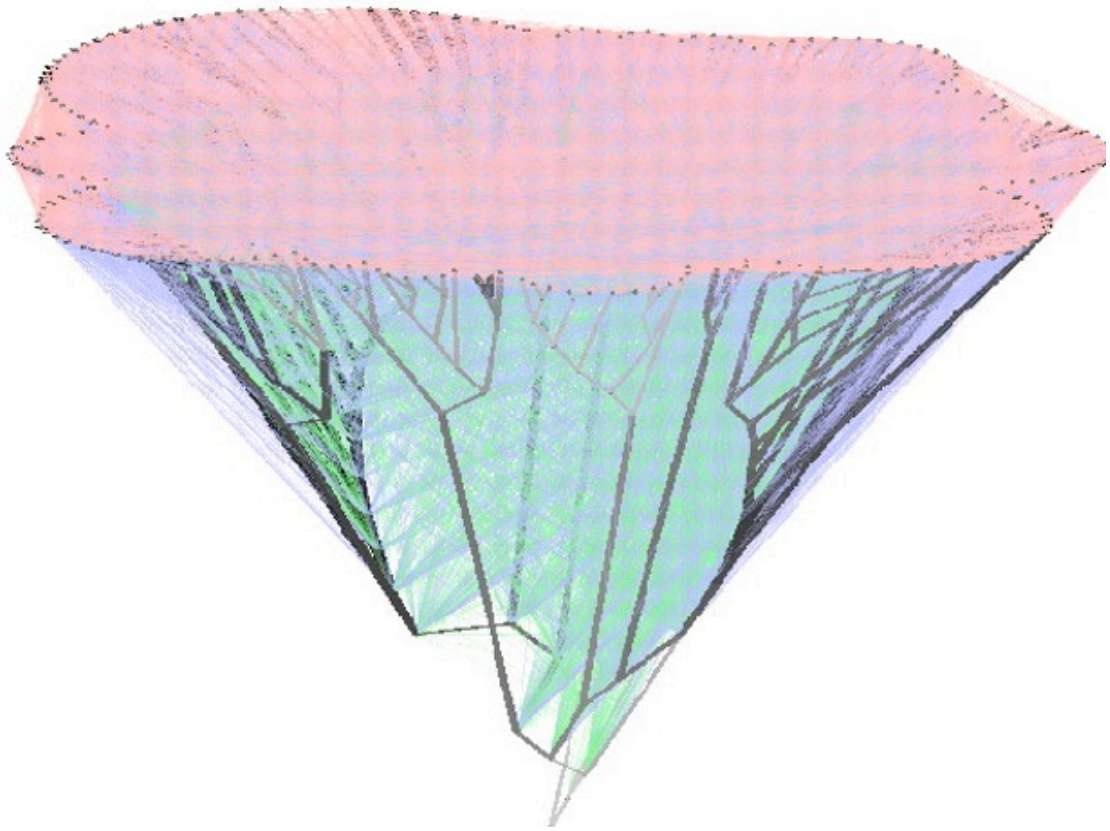


Figure 1.16: A three-dimensional projection of a HGT network. The grey tree represents the vertical component of evolution. The nodes on the network are the external and internal nodes of the tree. A blue, green or pink edge on the network indicates a HGT event between the two nodes it connects. Taken from Dagan *et al*. (2008).

1.6: Aims of this Thesis

In this thesis I wish to discuss the evolutionary processes and entities that are often overlooked in studies of bacterial evolution.

In chapter 2 I discuss the evolutionary phenomenon of gene fusion and present a new method for the detection of fusions of unrelated genes using network structure analysis. I report the capabilities of this method based on tests using simulated and biological data.

Chapter 3 sees the utilization of this method on a number of datasets. The functionality and limitations of the method are put to the test. The success of the method in finding fusions of unrelated genes allows us to quantify fusions in a given genome and discuss the functional categories to which the fusion genes belong.

In chapter 4 I attempt to gain an understanding of the evolutionary history of a group of closely related bacteria while incorporating all aspects of this history. There is particular emphasis in this chapter on the use of networks in creating an all-encompassing view of evolution. I discuss the phylogenetic and non-phylogentic signal elucidated by network visualization of homologous relationships between whole genomes and between individual genes. I also discuss the impact of genetic entities that are acquired through horizontal gene transfer on bacterial evolution as seen in the networks of homologous relationships.

Chapter2 - FUSION: A Network-based Approach to Finding Fusion Genes

2.1: Introduction

The central tree metaphor has been challenged over the last couple of decades with the observation of incongruent trees derived largely from protein-coding genes in prokaryotic genomes (Brochier *et al.* 2002; Zhaxybayeva *et al.* 2006; Galtier and Daubin 2008; Retchless and Lawrence 2010) . In addition to gene tree disagreement, many genes have been found to have sparse and inconsistent patterns of being in different genomes (Nakamura *et al.* 2004; Beiko *et al.* 2005; Dagan and Martin 2007). The Tree of Life model has generally been supported by methods that carry out an *a priori* selection for treelike data. This has involved selecting genes that seem recalcitrant to horizontal transfer as well as restricting the analysis to the comparison of genes that appear to be homologous along the vast majority of their length (Brochier *et al.* 2002; Daubin *et al.* 2003). What this has effectively meant is that novel genetic entities, such as gene fusions are omitted from the analysis.

When making alignments with a view to building trees, the tendency has been to focus on full genes. Consequently, relationships are inferred when the majority of one sequence is homologous to the majority of another. In order to explain gene relationships, often a sequence identity percentage cutoff is applied to the data or anything that is not optimally aligned is "trimmed" from the alignment (DeSantis Jr *et*

al. 2006; Dunn *et al.* 2008). Ultimately, when constructing phylogenetic trees we tend to lose a whole wealth of information.

A fusion gene is the result of an event whereby two previously separate genes are joined, to encode a single, usually multifunctional, protein (Enright *et al.* 1999; Suhre and Claverie 2004; Pasek *et al.* 2006). Natural fusion proteins are a result of complex mutations such as tandem duplications (Jones *et al.* 2008), retrotranspositions (Ruan *et al.* 2007) and chromosomal translocation (Rowley and Beck 1973). The fusion of genes or intra-gene recombination is in fact having a major impact on prokaryotic evolution and it has been seen not only in metabolic enzymes (Tsoka and Ouzounis 2000), but also in housekeeping genes (Suhre and Claverie 2004) and genes that were thought to be resistant to recombination such as rRNA genes (Wang and Zhang 2000; Inagaki *et al.* 2006; Chan *et al.* 2009).

In terms of phylogeny, the presence or absence of fusion genes can provide a distinction between organisms. In animals and fungi, the two genes for dihydrofolate reductase and thymidylate synthase are translated separately, they are also translated separately in eubacteria, although often they are found in one operon (Philippe 2000; Stechmann and Cavalier-Smith 2002). In plants, aveolates and euglenozoa, however, the genes have fused together, resulting in a bi-functional gene with both enzyme activities manifesting in one protein. This multifunctional hybrid is found exclusively in the bikonts and as a result, has been used in studies attempting to root the eukaryote tree of life (Stechmann and Cavalier-Smith 2002; Stechmann and Cavalier-Smith 2003). Therefore, gene fusions are informative and useful markers and their identification provides interesting insights into evolutionary biology.

Gene fusion is an important event in cancer cell biology and detection of these events is important for diagnosis and treatment (Maher *et al.* 2009). Discovered by

Janet Rowley in 1972 (Rowley and Beck 1973) the Philadelphia chromosome is a shortened chromosome 22 as a result of reciprocal exchange of DNA between the long arms of chromosomes 9 and 22. The exchange results in the 3' end of the Abl gene being moved from chromosome 9 to 22 where it is juxtaposed to a segment of the disrupted Bcr gene. The result is a chimeric Bcr-Abl gene, a tyrosine kinase, which, due to loss of the N-terminal is stuck in the "on" position causing unregulated cell growth. In their review of the literature Kurzrock et al. found that more than 90% of patients with chronic myelogenous leukemia tested positive for the chimeric Bcr-Abl gene making it an important diagnostic tool (Kurzrock *et al.* 2003).

Multifunctional genes with novel properties are being continuously discovered in all kinds of areas of biology. For instance, it has also been shown that 1,680 fusion and fission events can be seen across a dataset of 12 fungal genomes, with fusions mostly involving genes of similar function (Durrens *et al.* 2008). Novel gene fusions involving aminoglycoside resistance genes have been discovered, including the bifunctional aminoglycoside 3" adenyltransferase aminoglycoside 6'-N-acetyltransferase on a plasmid in a multiresistant *Serratia marcesens* strain (Centron and Roy 2002) and the bifunctional 6'- aminoglycoside acetyltransferase 2" aminoglycoside phosphotransferase on the *Streptococcus faecalis* plasmid (Ferretti *et al.* 1986). In metagenomic studies, bifunctional multidrug resistance genes have been identified in soil from an orchard (Donato *et al.* 2010). In a review of bifunctional antibiotic resistance genes and mechanisms of generating bifunctional genes, Zhang and co-workers called bifunctional antibiotic resistance elements "harbingers of clinically significant resistance mechanisms of the future" (Zhang *et al.* 2009).

The news is not all bad, however. In some instances, artificial generation of multifunctional proteins through the expression of recombinant DNA

has had beneficial therapeutic effects. For instance, a preclinical study of a fusion of a cancer cell homing protein and a PKCepsilon inhibitory peptide has shown promise in a mouse model for the treatment of head and neck squamous cell carcinoma (Bao *et al.* 2009). Etanercept is a chimeric protein drug for the treatment of autoimmune diseases including rheumatoid and psoriatic arthritis. A TNFα blocker, it is created though combining tumor necrosis factor receptor (TNFR) with Immunoglobulin G1 Fc segment (Rapaka *et al.* 2007). The study of gene sharing is now at the forefront of our understanding of biology and can make a significant impact on the theoretical underpinning in the emerging field of synthetic biology – the engineering of genetic components by designing elements with novel combinations of genes (Khalil and Collins 2010).

In a practical sense, knowledge of gene fusions can be particularly interesting for understanding genome evolution and organismal adaptation. Proteins can form functional connections in metabolic pathways, complexes and regulatory networks (Szklarczyk *et al.* 2011). For a long time, these interactions were only detected experimentally (Phizicky and Fields 1995) but the explosion of genome sequence availability has increased the amount of information for function prediction (Bork *et al.* 1998). Sequence comparison software programs including BLAST (Altschul *et al.* 1990) and methods that detect subtle sequence conservation like HMMer (Durbin *et al.* 1998) were used to define proteins. From these analyses, 70-90% of functions for encoded proteins could be predicted via annotation transfer from well-characterised homologs (Galperin and Koonin 2000). Bioinformatics and comparative genomics also allowed the proposal of various methods to predict functional interactions based on the genomic context of their genes. This meant finding a protein's interacting partners based on the position homologs in one or more genomes (Huynen *et al.*

2000). There are 12 recognized computational methods for predicting a protein's interaction partners (Shoemaker and Panchenko 2007). While some predict a physical interaction, others, including the Rosetta Stone method (Marcotte *et al.* 1999), predicts a functional association.

A Rosetta Stone protein is one that can provide us with information about other proteins (Veitia 2002). The Rosetta Stone method is based on the knowledge that often two interacting proteins in one genome will have homology with a single fused protein chain elsewhere in the genome, or in another genome. In other words, in those instances where two separate genes A and B match one fused gene in a single open reading frame (ORF). In order for a fused gene to become fixed in the genome the two component genes need to be able to function in the same compartment of the cell, at the same developmental stage and in response to the same stimuli (Patthy 2008). It has been shown that more often than not, the two component genes show functional similarity (Sali 1999; Yanai *et al.* 2001; Yanai *et al.* 2002) and are likely to be linked in an attempt to reduce the regulational load in the cell for a particular process (Enright and Ouzounis 2001). Thus, knowing the function of the fused gene provides information about the function of the two component genes. Rosetta Stone proteins or fusion genes have proved to be key in finding potential protein-protein interactions and metabolic or regulatory networks (Sali 1999; Galperin and Koonin 2000).

Marcotte *et al* (1999) and Enright *et al* (1999) were among the first to determine a functional relationship between two proteins in one genome by finding their fused homolog in another genome, thus introducing the "Rosetta Stone" or "fusion" method.

Marcotte *et al* (1999) set out with the goal of determining whether protein function and protein-protein interactions could be indentified computationally from a genome sequence. They asked whether you could predict an physical interaction between two separate proteins in one genome that have sequence similarity to different segments of the same protein in another genome. They searched for this pattern in 4,290 protein sequences from the *E. coli K-12* genome. Initially they found 6,809 pairs (some genes appeared in more than one pair) of non-homologous sequences, both members of the pair having significant similarity to a single protein in another genome. It was predicted that there must be some fusions that went undetected. They defined false negatives as those results that were missing due to a lack of homology with a fusion gene and this failure to identify homologs was either because of divergence or loss of the fusion gene during the course of evolution. They also found that they could not detect fusions of homologous proteins (homologous recombination), therefore could not predict an interaction between proteins that were paralogous. Other false positives included pairs of genes that were homologous to same fusion gene but did not in fact interact. For *E. coli* they predicted that 82% of their results would be false positive.

In order to substantiate their results Marcotte and colleagues implemented three tests of confirmation. The first was to compare the separate component proteins in SWISS-PROT (Bairoch *et al.* 2004) for similar functions. Secondly they compared their findings to the literature and databases. Finally they used phylogenetic profiles to detect functional interactions by analyzing correlated evolution of proteins. At this point they had a robust predictor of functionally linked proteins, but only a subset of the results were physically linked. Following the tests of confirmation, they were left with 749 of the original 6,809 links. Detection and removal of promiscuous domains

i.e. domains that pair with large numbers of other domain types (Basu *et al.* 2008), via the ProDom database (Corpet *et al.* 2000) meant that the rate of false positives in their results dropped to 65%.

Enright *et al* (Enright *et al.* 1999) also looked for functional and physical interactions between proteins that show homology to the same fused protein. They used a similar computational, sequence comparison method to Marcotte and his colleagues. From 7,768 protein sequences they found 215 proteins in *E. coli*, *Haemophilus influenzae* and *Methanococcus jannaschii* that were involved in 85 suspected fusion events of which 64 were confirmed to be unique fusions while 21 were false positives. Both Marcotte and Enright's groups concluded that with more computational power they could use the method described to predict interacting partners in all complete genomes.

Though not on a particularly large scale, the fusion method has been used since as a predictor of functional linkages. Prolinks (Bowers *et al.* 2004) is a database of protein functional linkages. The creators use four algorithms including the fusion method used by Marcotte and Enright to infer protein function and linkages.

The fusion method has also been applied to quantifying fusions. FusionDB (Suhre and Claverie 2004) provides a strict definition of a fusion protein in an attempt to provide a database for in-depth analysis of prokaryotic gene fusion events. A fusion event is accepted and placed in the database if it meets the criteria of the fusion method. Suhre and Claverie (2004) use the same computational methods as Marcotte and Enright as well as introducing the use of COG annotated genomes to find fusion events. These "COG fusions" provide information about the types of genes that are likely to fuse.

When used to quantify fusions, the fusion method is cumbersome, difficult to implement on a large-scale and rife with false positive results (Enright *et al.* 1999; Marcotte *et al.* 1999; Snel *et al.* 2000). However, the data pertaining to gene relationships can be efficiently represented and explored using network-based models of homology. Network structures (Newman 2004) provide a way of examining a gene in terms of its relationship to all other genes. In particular networks can elucidate the hybrid evolutionary signals in genomes that are often overlooked by tree methods.

In this chapter I present a new method of finding fusions using a network-based algorithm. The algorithm consists of three parts:

- Reconstruction of a homology network from input data
- A search through the network to find nodes with a defined set of characteristics and
- A user-friendly report of all fusion nodes detected on the network

For the first part an all-vs-all similarity search is translated into a network structure. The nodes on this network represent the genes of interest and each edge indicates a homology relationship between the two nodes or genes it connects. By searching for specific structures in the network, we can identify which nodes represent fusion genes.

A fusion gene, by definition, is made up of two individual component genes**.** Each of these component genes may have a set of homologs. By default, any homologs of the two component genes should also be homologs of the fusion gene. A group of homologs on a graph will be connected to one another and so will form a maximal clique; a subgraph on which all nodes are connected to all other nodes and which cannot be contained in another clique. The fusion gene node therefore should be connected to all cliques formed by the component genes and their homologs

(Figure 2.1). By this logic a potential fusion gene will always be found in all maximal cliques that contain its components. We refer to the cliques as “component cliques” and a node found in more than one component clique is a potential “fusion node”.

It is important to note however, that our method enriches a dataset in fusions of non-homologous genes. In other words we search for nodes that potentially represent fusions in which the component genes are not homologs of one another.

The reason our search terms are so specific is that fusions of related genes are extremely difficult to identify on a network. This means that there are examples of natural fusion events that will go undetected by this method. False negatives will include fusions as a result of tandem duplications or fusions of homologs of any kind. On a network, component genes that are homologous to one another will be connected to one another. So the nodes representing the fusion and both sets of component genes will all be connected to form one maximally connected clique (Figure 2.2). The component genes will be indistinguishable as different gene families and therefore as separate cliques.

Nodes on networks that seem to satisfy the condition of being in two cliques can be in this position on the network for two reasons: (i) they are fusions, (ii) they represent divergence/loss of different parts of an ancestral gene. We have a lot of difficulty in distinguishing these two kinds of gene. However, when we see nodes of this kind, we can be alerted to the fact that they are interesting.

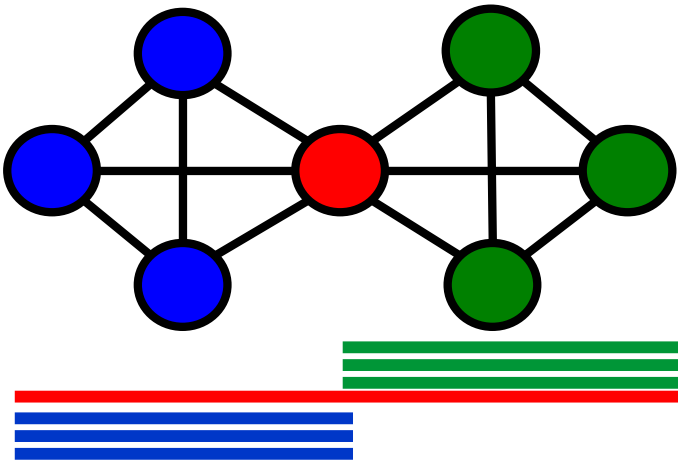


Figure 2.1: Network representation of a fusion event: the fusion gene (in red) is part of two different cliques, the blue clique is one gene family and the green is another. Below the network is a crude representation of how these proteins would align.

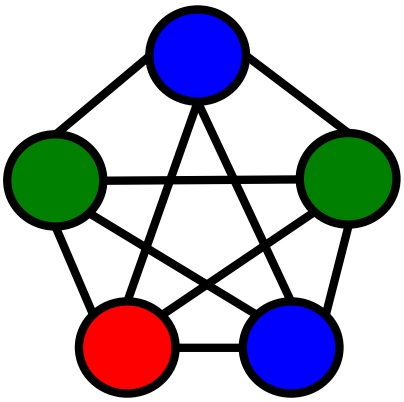


Figure 2.2: Network representation of the fusion of homologous component genes. On this network the green nodes represent genes that arose by tandem duplication of the blue node genes. The red node is a fusion gene made up of a blue node fused to its duplicate green node. Because the duplicate gene is homologous to the original gene they are connected on the network and so form one maximally connected subgraph. In this case, although the red node is a true fusion, it is not identified as a fusion gene by our method.

In some cases we will find that, even though the component genes do not appear to be related, their entire length may overlap substantially while only small regions align with the potential fusion gene. An explanation for this may be that the component genes may have once shown homology along their entire length. They have now diverged beyond the point where they are similar enough to one another to be recognized as homologs in a similarity search, but remain linked by the common gene that has retained the ancestral information. These genes, although false positives in terms of being fusion genes, can provide us with information about the two component genes and should not be ignored.

In order to understand the conditions under which the algorithm will work well and when it is likely to fail we set up a rigorous testing regime. The regime was threefold; firstly testing the effectiveness of the algorithm in building a network from text input and subsequently finding potential fusion nodes on that network. This was achieved using simulated network structures that are representative of how we would expect a fusion event to look. The second test involved using simulated sequence data to test how well the algorithm deals with a BLAST output and whether it can produce a diagrammatic representation of the alignment between fusions and their component genes. To demonstrate the utility of the algorithm, we tested it on all genes from the genome of *E. coli* K-12 MG1655.

2.2: Method

The following algorithm was implemented in the Python programming language to allow the use of Python modules for network manipulation and analysis.

2.2.1: The Algorithm

2.2.1.1: Building the Network Structure

To construct a network we perform an all-versus-all similarity search of a collection of genes using the BLAST algorithm (Altschul *et al.* 1997) with "–m 8" flag. The M8 output consists of 13 tab-delimited columns of information pertaining to the relationship between the query and subject genes. This information includes Query id, Subject id, percentage identity, alignment length, mismatches, gap openings, query start and end, subject start and end, e-value and bit score.

To build the network we require the information from the first two columns of the BLAST output: the query and subject ids. These two columns are used to create a graph structure where each gene in our sample is represented as a node and each edge on the graph is a statement of homology between the two nodes that it connects. The network is constructed from the similarity search output using the python package NetworkX (Hagberg *et al.* 2008). Edges are added sequentially from the similarity search results using the "add_edge" command from the NetworkX package. The query gene becomes one node, the homologous gene is another node and an edge is

drawn between the two, reflecting their homology. “Self hits” i.e. where the same gene is both the query and hit in the BLAST are excluded from the network to reduce its size in memory and also to reduce the time it takes to search through the network.

2.2.1.2: Finding Cliques on the Network

The completed network is traversed to find all possible cliques. In other words the network is searched from one node to next to find any set of nodes that is maximally connected. We use algorithm 457 (Bron and Kerbosch 1973) otherwise known as the Bron-Kerbosch algorithm after its creators, as adapted by Tomita et al. (Tomita *et al.* 2006) with worst-case time complexity of $O(3^{n/3})$. An alternative strategy can be used, based on matrix multiplication, to list all cliques in polynomial time per generated clique (Tsukiyama *et al.* 1977). In other words an output sensitive algorithm which runs in O(mn) time, where m is the number of edges on the network and n is the number of nodes. This algorithm can list all cliques in polynomial time for graphs in which the number of cliques is polynomially bound. Algorithm 457 guarantees worst-case time complexity of $O(3^{n/3})$ and has been shown to be faster than its competitors (Cazals and Karande 2008; Eppstein and Strash 2011).

Algorithm 457 is implemented in the NetworkX python package (Hagberg *et al.* 2008) and is implemented using “find_cliques(*G*)”, where G is the network you wish to search. At its most basic, the Bron-Kerbosch algorithm is a recursive backtracking algorithm whereby, given three sets *compsub*, *candidates* and *not*, it finds the maximal cliques that include all of the vertices from *compsub*, some from *candidates* and none from *not*. By assessing one node at a time the node can either be

added to the clique or to the set of nodes that are excluded from the clique. Those excluded from the clique must have at least one non-neighbour in the final clique.

The clique-finding algorithm returns a list of all possible cliques in the network. For every possible pair of cliques the nodes are divided into three subsets: those only found in clique 1, those only found in clique 2 and those common to both cliques. Any node that is common to two or more cliques is considered a potential fusion gene.

2.2.1.3: Testing for Overlap

I have implemented a test for overlap of non-homologous genes on putative fusion genes. The aim is to see whether the putative fusion gene is a fusion of unrelated genes or a gene that, on the network has the same properties as a fusion gene but may be the result of a number of evolutionary processes.

To reiterate, the putative fusion is found in more than one clique and each different clique represents a different component gene family. If the putative fusion gene is a result of an event whereby two unrelated genes were fused then the different component genes should align to different regions of the fusion gene.

When testing for overlap the putative fusion gene acts as the query and any gene suspected to be a component of this fusion is treated a subject in the BLAST output. For each component gene the query start and end positions are parsed from the BLAST output. In other words we find the start and end positions of the area of homology between the fusion and component.

Figure 2.3 summarizes the alternative alignments that can result from an event that is or resembles fusion event on a network. The red line represents the fusion gene,

the blue and green lines are the areas of overlap between component genes and the fusion: blue is one gene family and green is a different gene family. The black dashed lines represent the area of a component gene that is not homologous to the fusion gene.

Figure 2.3A shows a fusion of unrelated genes. There is no overlap between the blue and green areas nor is there overlap between black dashed lines and any of the component genes. The simplest explanation of this pattern is that the gene represented by the red line is simply a fusion of the blue and the green genes. Therefore, we refer to this as a "true fusion".

Two different situations are observed where the component genes might overlap. The first is outlined in Figure 2.3B. In this case, a small section of a fusion gene shows homology with two component genes that do not seem homologous to one another. The region of overlap is small. In this situation, it is not possible to explain the result without further analysis. We identify these kinds of situation and put the genes aside for further analysis. The second situation is outlined in Figure 2.3C. In this situation there is quite a substantial overlap between the blue and green areas and the full length of the blue genes overlaps with the green genes. This final example could be indicative of homology between the component genes, lost through divergence. For the following study we disregard these results.

Our algorithm for finding fusion genes is a highly conservative method and has been designed to find fusions of unrelated genes only. We have not made any attempt to find other kinds of fusion genes and such an analysis was outside the scope

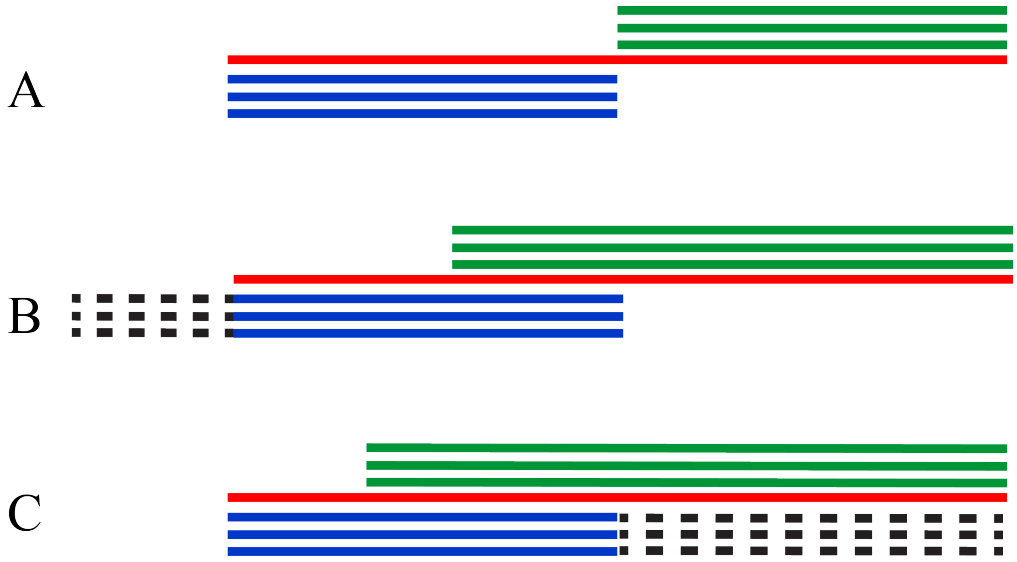


Figure 2.3: Examples of how a fusion event might look as an alignment. The red lines represent fusion genes and the blue and green lines represent the areas of alignment from the component gene families to the fusion gene. The black dashed line then shows the full length of the component genes, in some cases the area of alignment does not span the entire component gene.

of this study. In our study a fusion gene is one in which there is no sign of homology between the component genes.

2.2.2: Output

From here on we deal only with results that revealed fusions of unrelated genes or "true fusions". For each of the results, the algorithm produces a postscript file containing a crude drawing of the alignment between the fusion gene and its component genes. Again, using the information from the BLAST output we provide coordinates for each of the genes involved in the fusion event. The fusion gene is represented in red and the length of the line is proportional to the length of the gene. Each component gene is added to the diagram. The part that it homologous to the putative fusion gene is coloured either in solid red or solid greed. Parts of any component gene that does not appear to be homologous to the fusion gene are represented as a dashed line. Therefore, many component genes are found in the diagram with part of the gene represented as a solid colour and part as a dashed line.

The algorithm produces a corresponding information file for each postscript diagram. The information file contains the names, functions (if available) and gene lengths of the fusion gene and its homologs, i.e. more fusions of the same kind. It also contains information pertaining to the genes in each clique – the component genes. For each component gene the start and end positions of the area of alignment are provided as well as the start and end positions of the full length of the gene. These numbers are relative to the fusion gene.

Filename: 1.ps

Articulation point: Gene_1162 (891)
Function: unavailable

More Fusions:
Gene_1162, function: unavailable

Clique1:
Gene_2322 3e-37 aln(465 858) full(430 825)
Function: unavailable

Gene_3386 1e-40 aln(466 846) full(448 831)
Function: unavailable

Gene_2643 1e-44 aln(456 857) full(448 831)
Function: unavailable

Clique2:
Gene_2324 6e-40 aln(52 373) full(-25 442)
Function: unavailable

Figure 2.4: An example of an alignment diagram and its corresponding information file. The "Articulation point" in the information file refers to the fusion gene chosen to create the diagram. The red line in the diagram represents this fusion gene. Clique 1 from the information file is drawn in blue on the diagram and Clique 2 in green.

2.2.3: Producing Test Data

Testing the constituent parts of the algorithm separately allows us to find out where potential pitfalls lie and how much time and computational power are required to execute each part.

2.2.3.1. Simulated Network Data

The first simulated dataset was created to test the accuracy of our algorithm in (i) constructing a network from a text file, (ii) searching the network and finding all maximal cliques and (iii) reporting instances of overlapping cliques. For this test we did not use biological data, but simulated data that has properties found in the real data.

The first step was to simulate a network with structures that represented how we would expect a fusion event to look. The input data used to create the simulated network consisted of a text file that resembled a simplified BLAST output. In other words each line of the text file denotes a homologous relationship between two genes. In this case the file contains only the query and subject IDs, given that this is the minimum information needed to construct the network.

The next step was to execute the part of the algorithm that deals in constructing and searching the network. To keep it simple we did not include the test for overlap or production of the output diagrams. The output of this test is simply a list of nodes that are found to exist in more than one clique, i.e. potential fusion nodes.

2.2.3.1. Simulated Biological Data

The second test dataset was created to test the successfulness of our algorithm in producing output diagrams that are a true representation of the fusion events. Again we used simulated data so that we were aware of how many and what types of fusion events had occurred.

Firstly, we simulated component gene data. In other words we simulated multiple sequence alignments for eight different gene families (named gene family 1 to 8) were generated using Seq-Gen (Rambaut and Grass 1997), a program that simulates the evolution of nucleotide or amino acid sequences along a phylogeny, using common models of the substitution process.

The user must provide Seq-Gen with an input tree topology and a number of parameters. We used a simple bifurcating tree topology for nine taxa (named taxa 1 to 9) (Figure A1, Appendix), so each gene family contained nine homologous genes. The component genes were simulated using the HKY model with TS/TV rate of 2.5 and base frequencies set to equal.

The user may also specify a shape for the gamma rate heterogeneity called alpha (-a). The default is no site-specific rate heterogeneity. A low value for this parameter (<0.1) simulates a large degree of site-specific rate heterogeneity and as this value increases the simulated data becomes more rate-homogenous. Consistent with results from a previous study on rate heterogeneity in bacteria (Worobey 2001) we alpha set to <1.0.

The user may also specify a random seed number (-z). Using the same seed number with the same input topology will result in identical datasets. If unspecified, Seq-Gen generates a seed number based on the number of milliseconds passed since

January 1st 1970 (UNIX time). To create eight different gene families it was necessary to ensure a different seed for each iteration through random number generation.

Seq-Gen also requires that the user input the desired length of the alignment (-l) and the output type (-o) e.g. nexus file format. The sequences in this dataset are 1,000 base pairs (bp) long and the output type is nexus. The final command line input is:

./seq-gen –mHKY –t2 –fe –a0.09 –l1000 –*z(random)* –on <*(tree)* >*(ouput)*

In order to simulate a number of fusion events, seven genes from seven different gene families were randomly concatenated. This provided us with three simulated fusion genes. Two fusion genes consisted of two component genes from two different simulated gene families while a third fusion gene consisted of three component genes from three different families. To provide a comparison, we did not include genes from the eighth gene family in any of the simulated fusion events.

We then used Seq-Gen to simulate divergence after a fusion event. The user can specify a sequence as the ancestral sequence at the root (otherwise a random sequence is used) (-k). We simulated fusion gene families in the same way we simulated the eight component gene families, but this time the ancestral sequence was specified as one of the three concatenated genes that represent fusion genes and the tree contained just 6 taxa (Figure A2, Appendix):

./seq-gen –mHKY –t2 –fe –a0.09 –k1 –l1000 –*z(random)* –on <*(tree)* >*(ouput)*

2.2.3.1. Real Biological Data

After the chief elements had been scrutinized, the final test was to understand the functionality of our algorithm. In other words we wanted to see how the algorithm would fair when real genes that had been involved in real evolutionary were provided as input. Using real data would also allow us to test the effectiveness of the test for overlap. We expect to see the effects of different evolutionary processes and so we may find putative fusions with overlapping component genes.

We chose to quantify fusions within the genome for *E. coli* K-12 MG1655. The input for this analysis was 4,145 full genes from the *E. coli* genome. The network would be created from the output of an all-vs-all similarity search of the set of genes from within the *E. coli* genome. From this test we would find out whether the algorithm was successful in reporting fusions of unrelated genes within a single genome.

2.3: Results

2.3.1: Simulated Network Data

Our first analysis of the accuracy of our algorithm focused on constructing a network from a text file. From the input text file the algorithm was able to construct the network shown in Figure 2.5. The network consists of 41 nodes and 94 edges and is divided into five connected components. Each of the connected components is representative of fusion event.

The algorithm was also successful in searching the network and finding all maximal cliques. It was reported that there were thirteen cliques in total (Figure 2.5). Finally, all instances of overlapping cliques were reported in the form of a list. The list contained all nodes found in more than one clique, i.e. all potential fusions. The clique-finding algorithm works as expected, at least on small, simple datasets and the logic behind detecting fusion node is reliable.

2.3.2: Simulated Biological Data

To see how well the algorithm can obtain the information needed to produce postscript diagrams of the alignment area between fusions and their component genes, we used simulated data pertaining to fusion events. The simulated data consists of 72 gene sequences, from eight different gene families and each sequence is 1,000bps in length. In addition to this there are 18 fusion gene sequences relating to three different

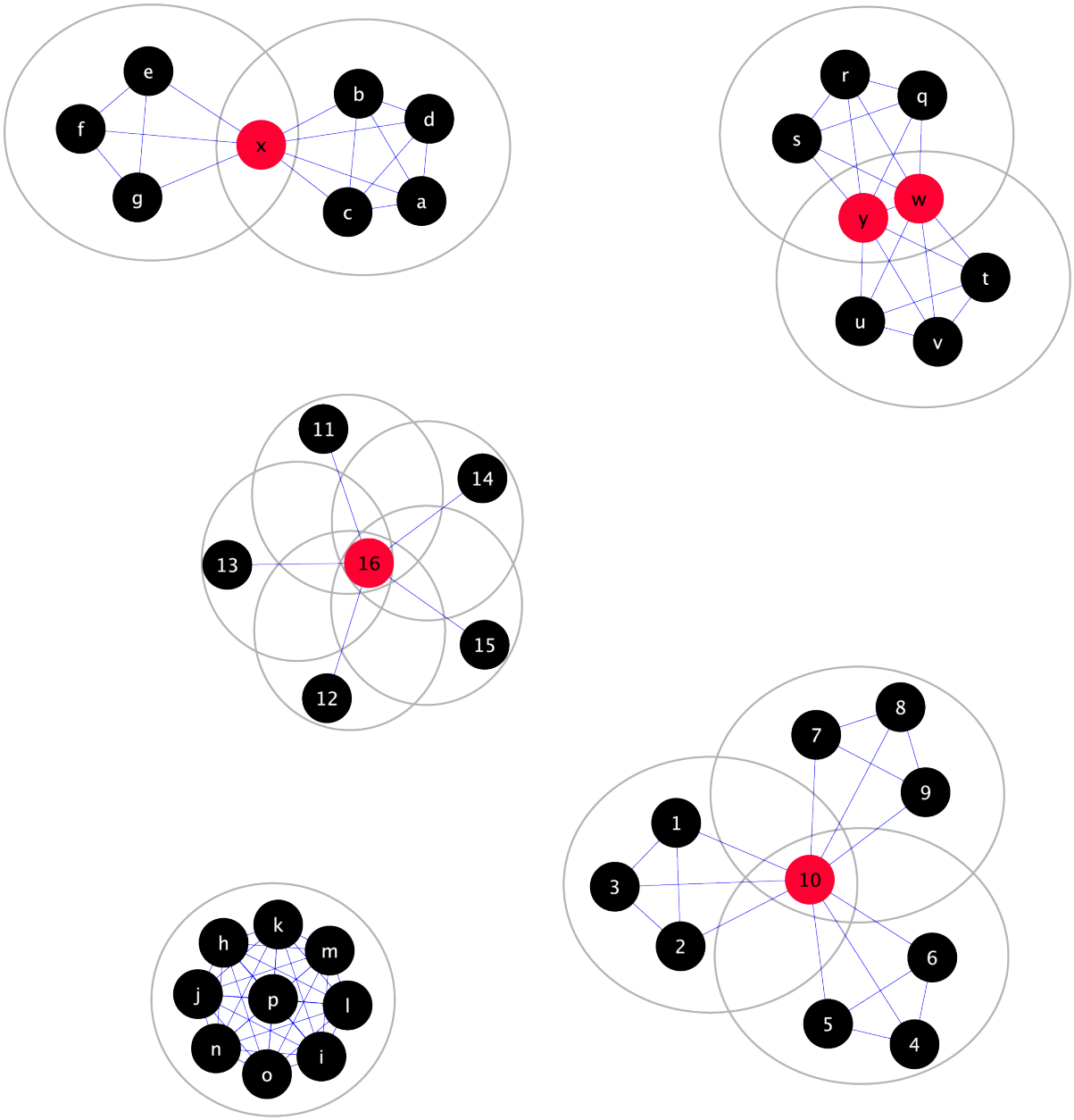


Figure 2.5: Simulated network. Each of the five components represents a different fusion event. Nodes in red represent putative fusion genes and nodes in black represent component genes or non-fusion genes. One of the components does not contain a red node. This component represents a fusion event whereby the two component genes are homologous to one another, e.g. a fusion as a result of duplication. Using network structure analyses there is no way to detect which node is the fusion in this case. The grey circles each encompass one maximal clique. The nodes that fall into the overlapping area of two or more circles are fusion nodes

fusion events. Two of the fusion genes are 2,000bps in length while the third is 3,000bps long. That is a total 90 genes.

The all-vs-all BLAST search returned 1,512 homologous relationships. The network produced (Figure 2.6) from this BLAST output consists of 90 nodes and 1,512 edges. This network has four components, three of which are representative of a fusion event. The fourth component represents the eighth gene family, of which no members are involved in any of the simulated fusion events.

We know from the simulated network data that the algorithm successfully detects fusion nodes. Our second analysis of the accuracy of our algorithm is focused on the ability of the algorithm to use information from the BLAST output to produce a diagrammatic representation of the alignment between the fusion genes and its components. Figures 2.7-2.9 shows the three diagrams produced by the algorithm. Each diagram has a corresponding information file containing the names, functions (if available) and lengths of the fusion and component genes. It also contains the start and end positions of the alignment areas. It can be seen from the information files that each diagram represents a different simulated fusion event. Figure 2.7 shows the diagram and information file for the fusion gene called F3_Taxon5, Figure 2.8 is of fusion F1_Taxon5 and 2.9 is of fusion F2_Taxon5. Thus, for every fusion event that we simulated, the algorithm was successful in extracting the necessary alignment information from the BLAST output and subsequently producing a diagram of the alignment.

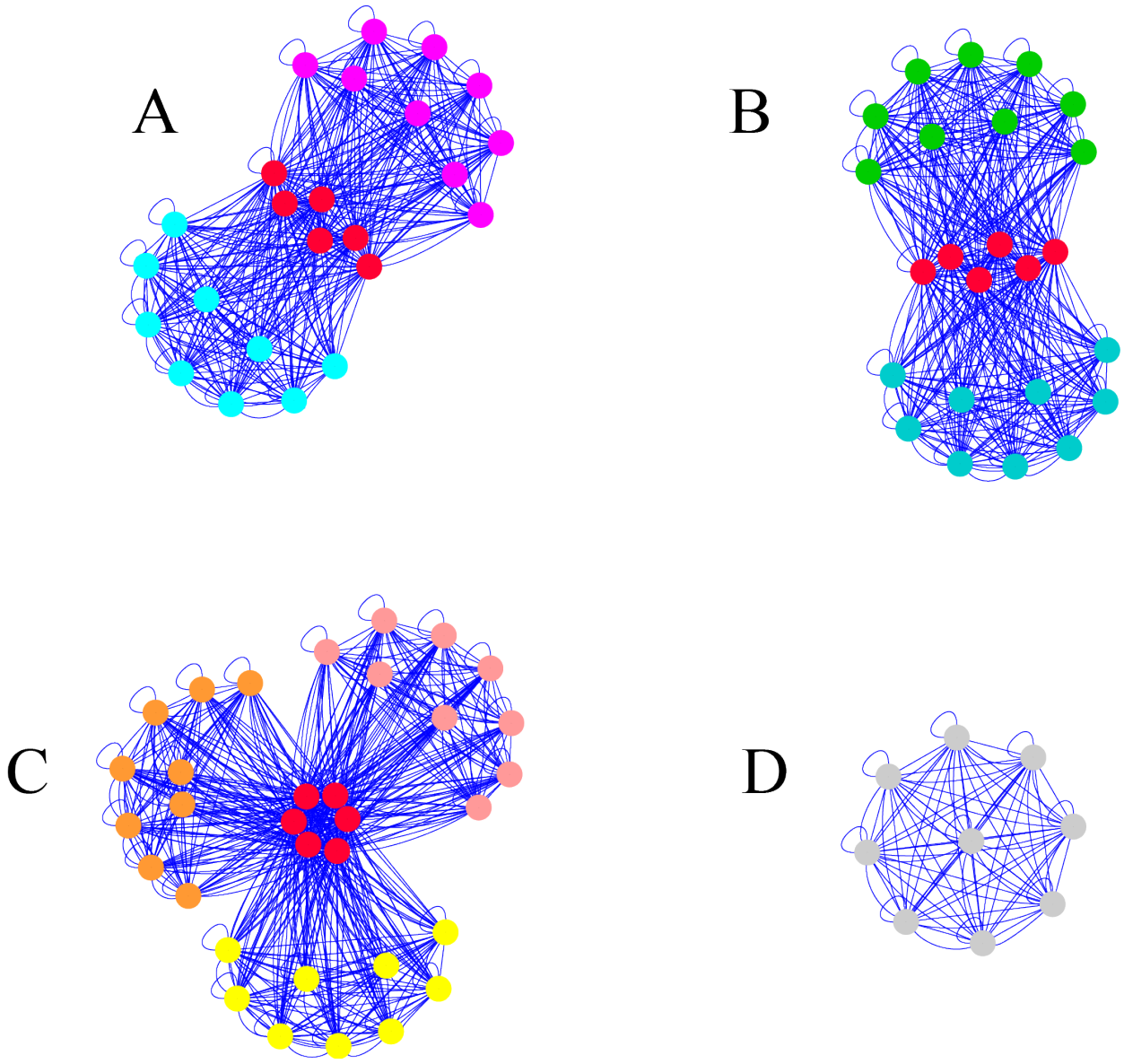


Figure 2.6: Network constructed from simulated sequences data. The red nodes represent the fusion genes. Each of the other colours corresponds to a different gene family. The fusions in A and B are comprised of two component genes while in C the fusion is comprised of three component genes. In D there is no fusion to report.

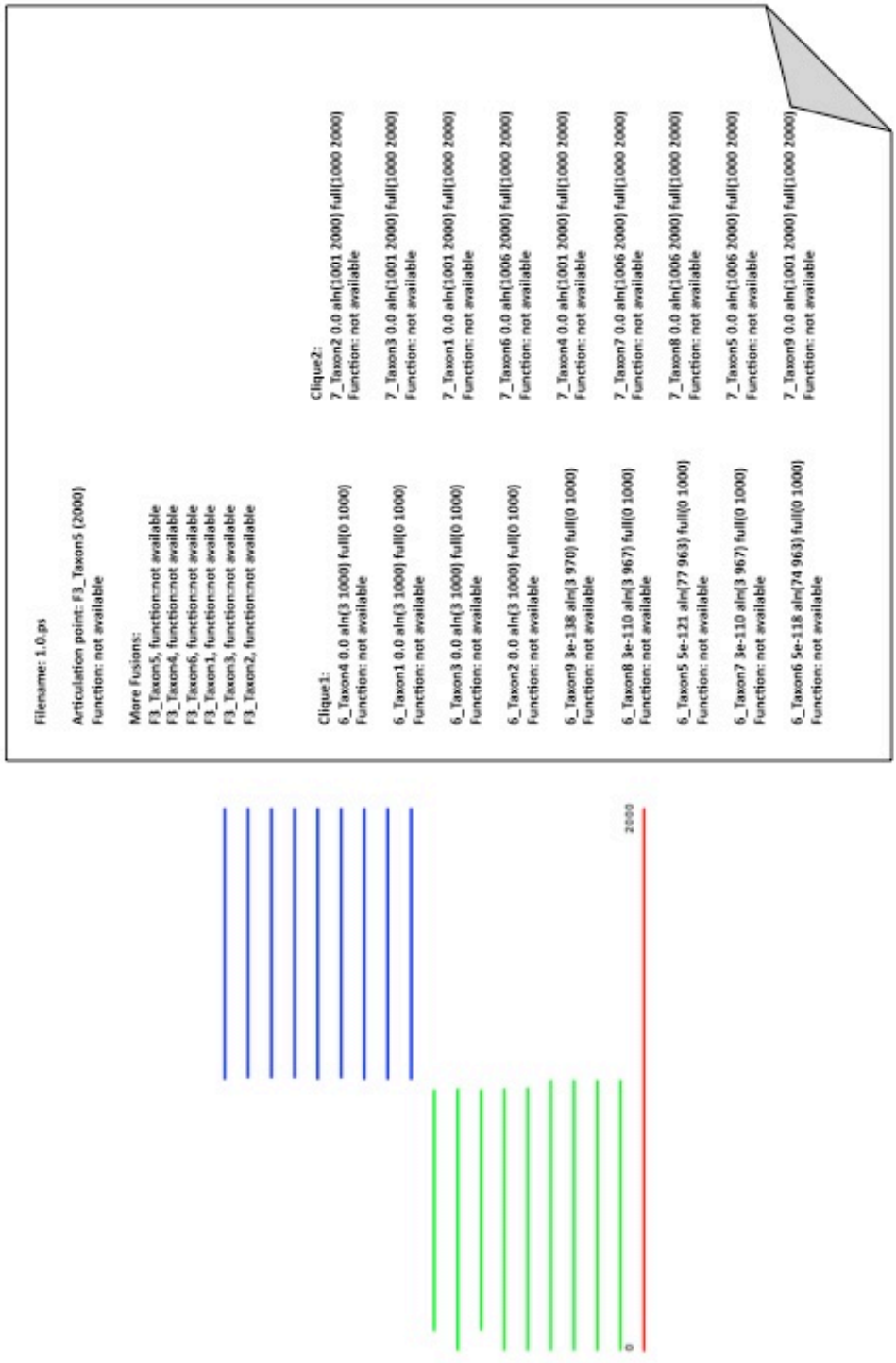


Figure 2.7: Diagram and information file for fusion F3_Taxon5.

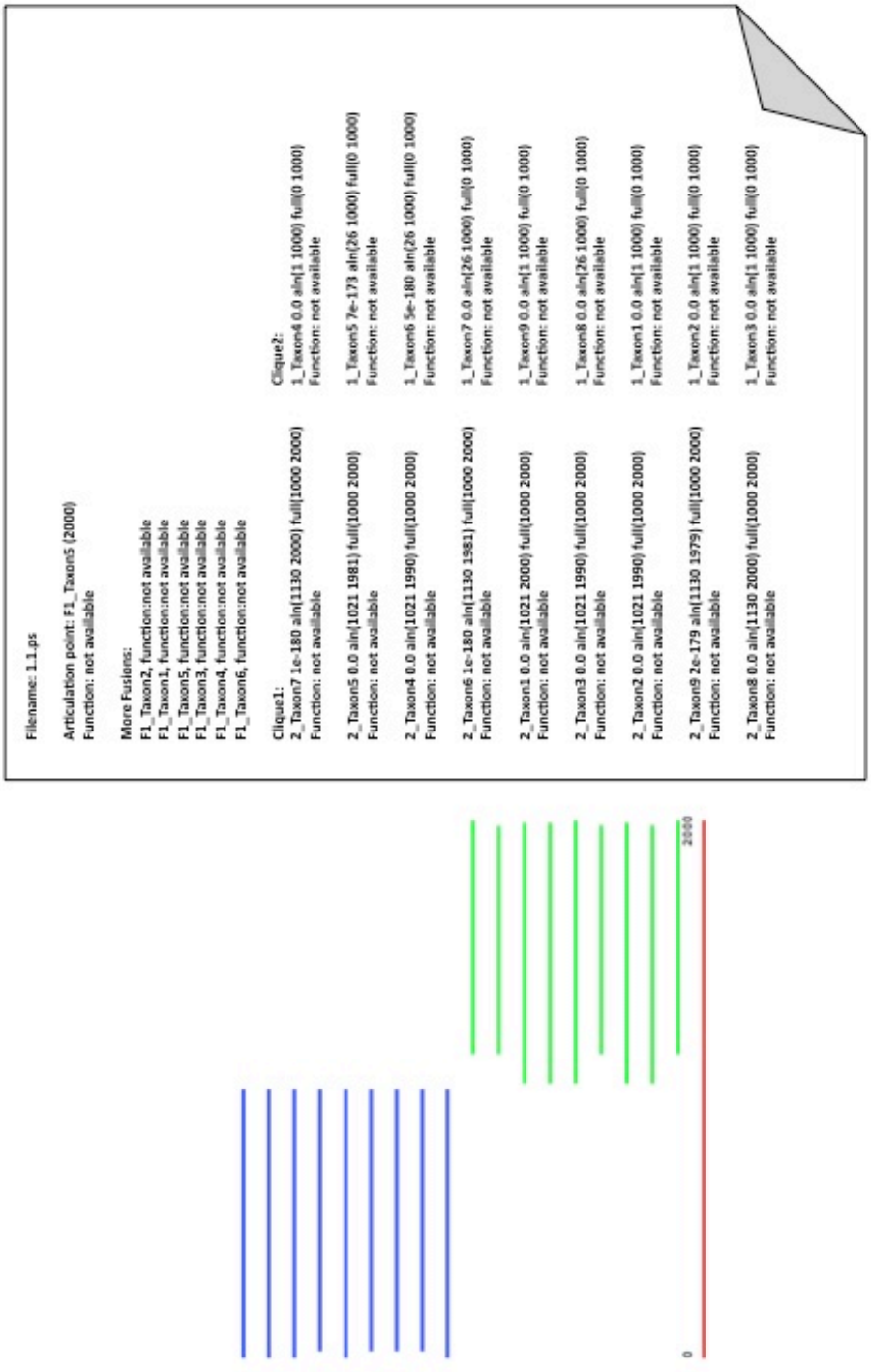


Figure 2.8: Diagram and information file for fusion F1_Taxon5.

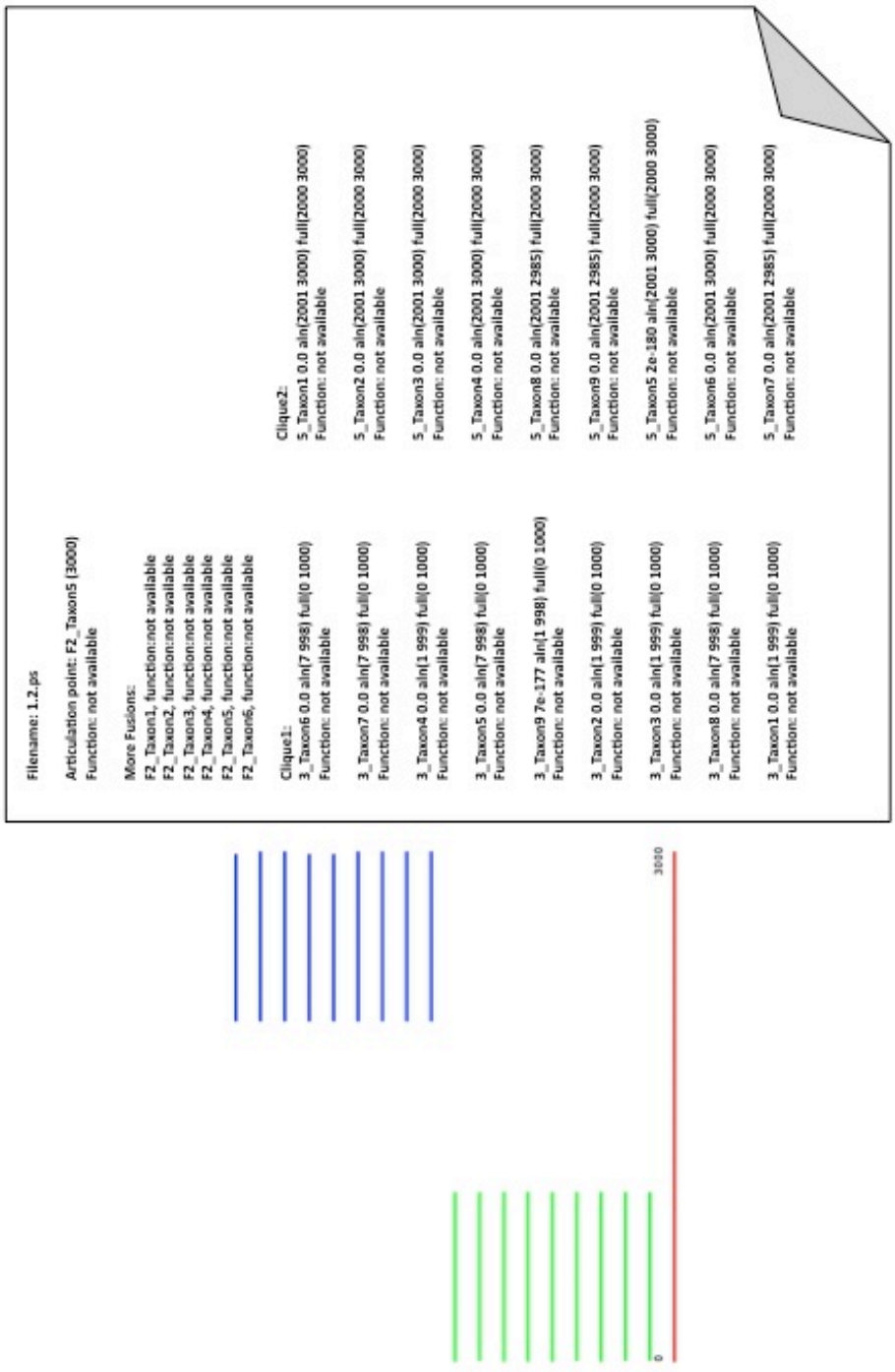


Figure 2.9: Diagram and information file for fusion F2_Taxon5

2.3.3: Real Biological Data

Our final analysis of the accuracy of our algorithm includes testing the execution of the overlap test as well understanding the functionality of our algorithm. For this we use all 4,145 genes from the genome for *E. coli K-12 MG1655.*

The 17,181,025 (4,145 X 4,145) BLAST searches returned 11,152 homologous pairs. Removal of "self hits" reduced this number of homologous pairs to 6,805. The reduced BLAST output produced a network of 1,565 nodes and 6,805 edges (Figure 2.10). This network consists of 458 connected components, the largest of which contains 76 nodes and 2,485 edges. It took only seconds for the algorithm to search the network, report all cliques and perform pairwise comparisons of said cliques.

For this dataset the algorithm returned 10 potential fusion genes. Of the 10 potential fusions, two were "true fusions", i.e. there was no overlap between the different component genes (Figure 2.11). For the other eight results we cannot be sure of their evolutionary history. We can however, be sure that our algorithm is successful in detecting overlap between the component genes (Figure 2.12).

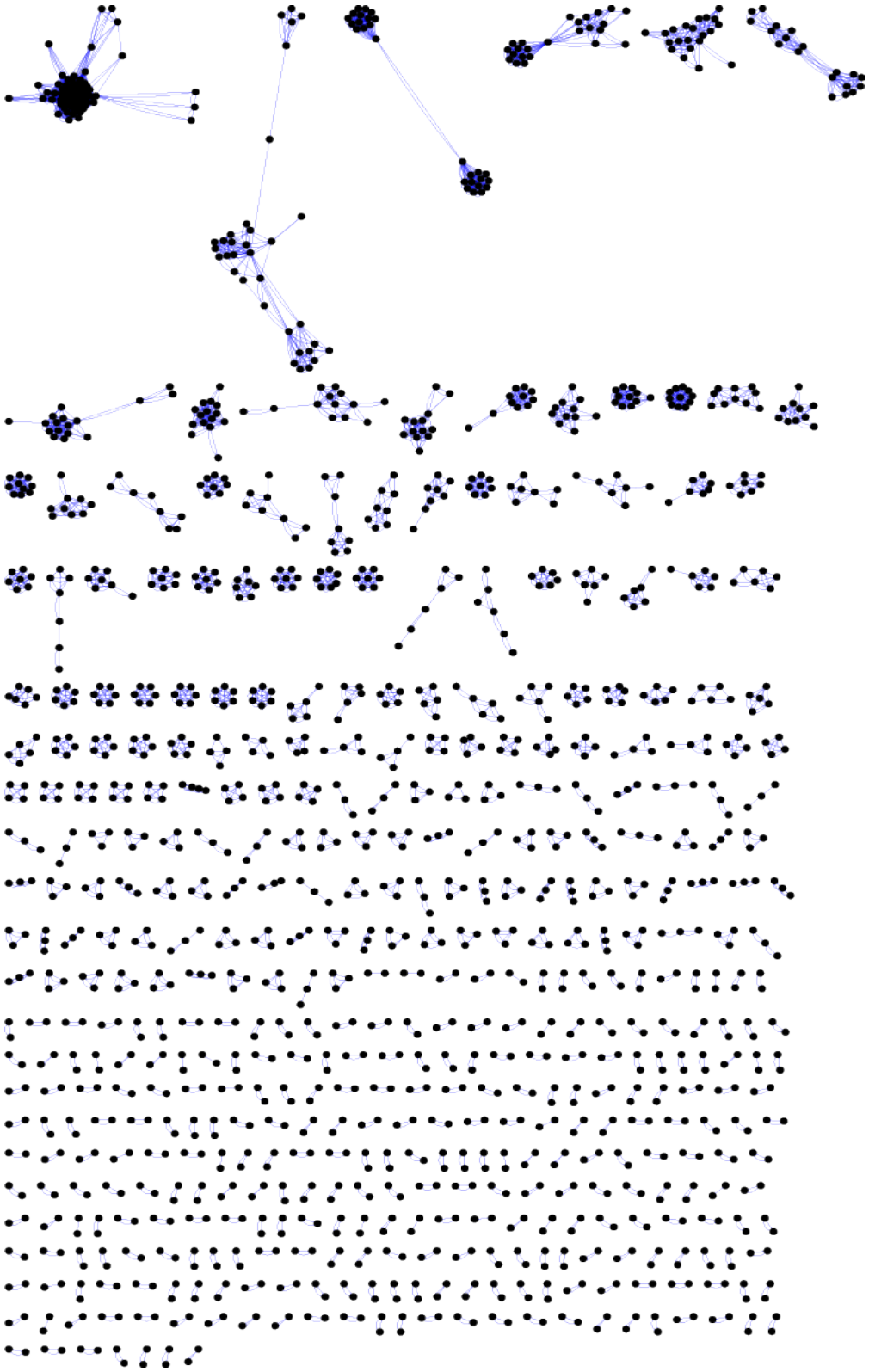


Figure 2.10: Network representation of the all-vs-all BLAST of all 4,145 genes from the genome of *E. coli K-12 MG1655.*

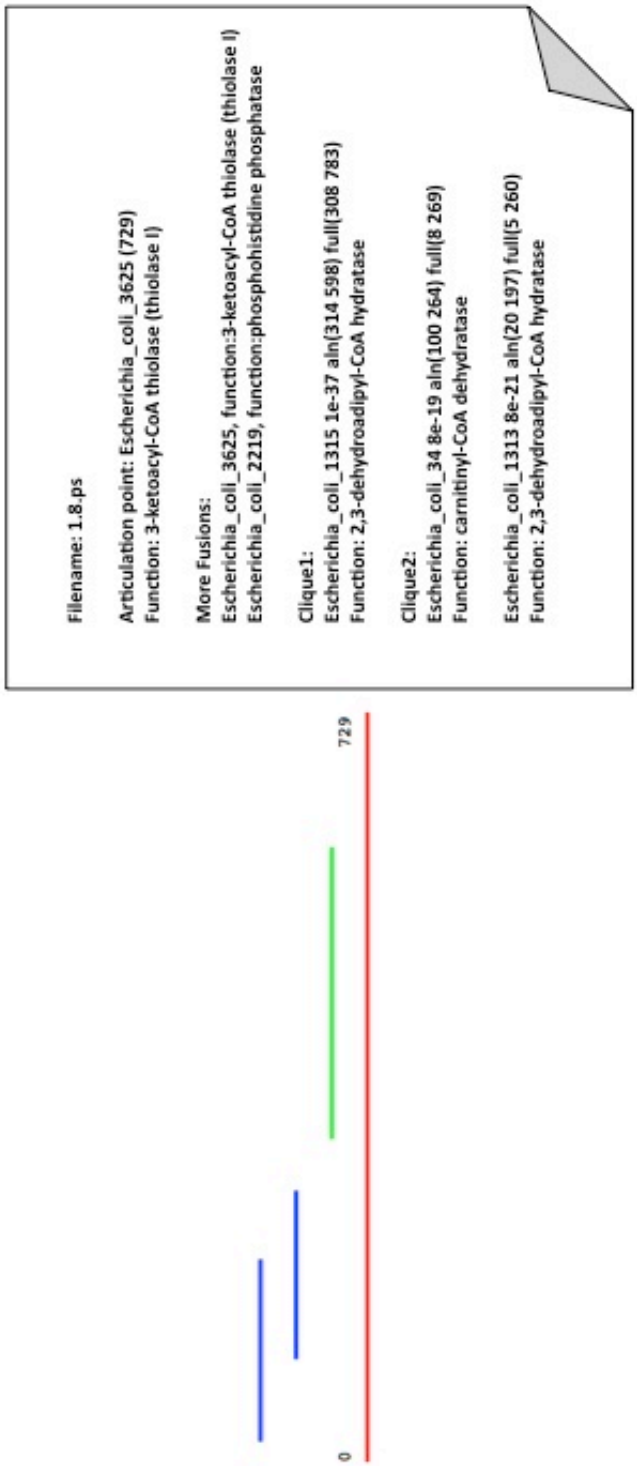


Figure 2.11: Diagram and information file for a fusion of non-overlapping genes from *E. coli*.

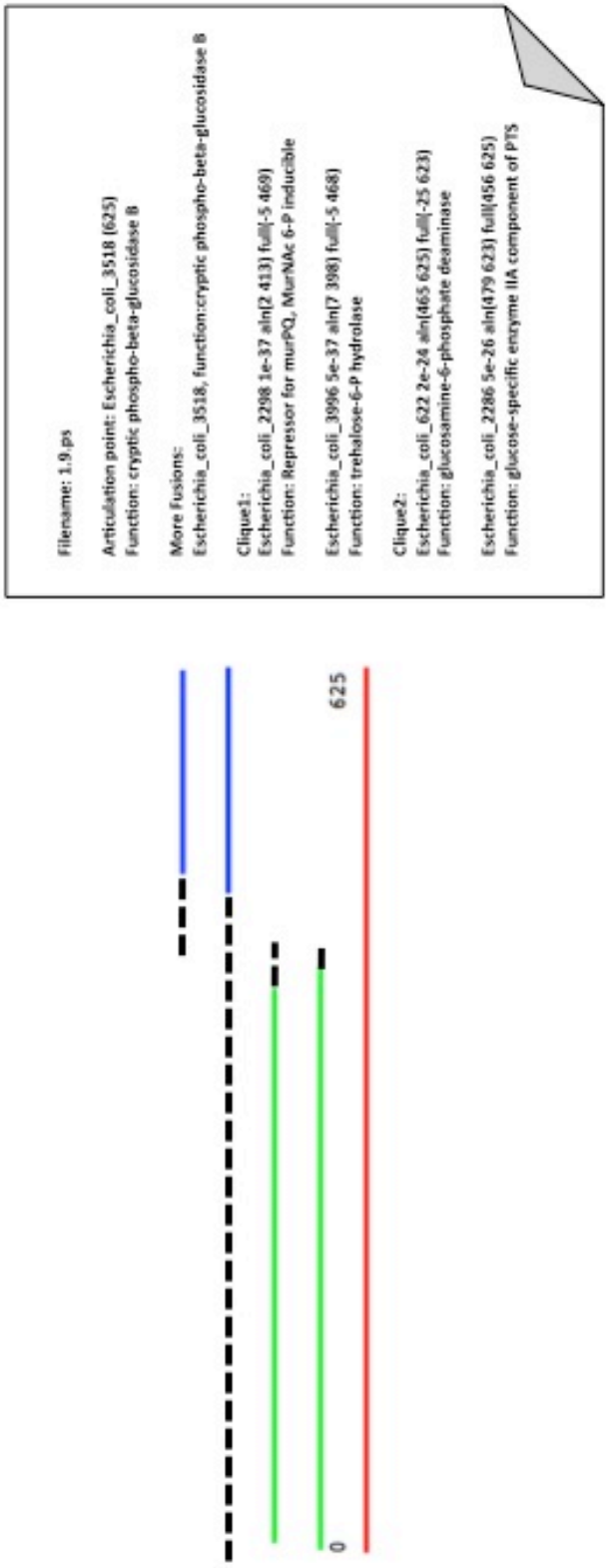


Figure 2.12: Diagram and information file for a result with overlapping genes from *E. coli*.

2.4: Discussion

In this chapter, I report the development of an algorithm to detect fusions of unrelated genes using network structure analysis. The attributes of fusion genes render them inappropriate for constructing branching structures of evolution. By their very nature, fusion genes originate from multiple sources, trying to represent their evolution in a strictly vertical way would result in ambiguities. By constructing networks of evolutionary patterns we create an all-encompassing view of the relationships between genes from which we can detect fusion events. We expect that there will be instances of fusion genes that are undetectable through network structure analysis. Fusions of homologous genes evade our method of detection.

Notwithstanding these elusive fusion genes, we have observed some desirable properties of this approach that make it a useful in quantifying fusions within a specified dataset. Previous fusion detection algorithms rely on non-overlapping, side-by-side BLAST matches of two genes from a reference genome to a single open reading frame (ORF) in a target genome. As with our method, fusions of homologous genes go undetected. However, perhaps the most striking shortcoming of fusion detection algorithms is that the results are highly limited by the input data. A reference genome is always chosen and compared to a target genome. This means that component genes are strictly limited to the reference genome in question and fusion genes to the target genome. Fusions of genes in the target genome that are composed of genes from outside the source genome are entirely overlooked. By representing all gene relationships on a network we remove the restriction of component genes originating in one specified genome. The all-versus-all BLAST from which the network is formulated ensures that all genes are described by their relationships to all

other genes. In theory our method has no limitations in the detection of fusions of unrelated genes, given any number of input genomes we should be able to describe all gene relationships and thus find all potential fusion genes. Unfortunately, computational power is highly limited. The clique-finding element of the algorithm alone has worst-case time complexity of $O(3^{n/3})$, as the network size increases it becomes impossible to traverse its entirety. In this chapter we did not reach the limits of computational power, this is discussed further in chapter 3.

Ultimately we have presented a method with the potential, if there were enough computational power, to report all instances of fusions of unrelated genes. We have proven the accuracy of our method through a number of tests. We are confident that we can take input data, create a true representation of that data on a network and as a matter of course, find structures pertaining to fusion events within the network. This confidence persists from using simulated data invented from our understanding of fusion events all the way to using real biological data that has undergone real evolutionary events. We have formulated a way to represent the data in a user-friendly format in the hope that this algorithm can be developed into a practical software program.

There is much room for improvement on our method. In light of the problems in searching very large networks we began to evaluate ways of reducing the size of the network in memory. By splitting the network into smaller parts the algorithm could be executed in parallel on all parts. However, in preliminary tests of this we found that the giant connected component of the network grows exceedingly large as more data is added. Splitting a connected component into smaller parts runs the risk of loosing information concerned with important relationships. Nevertheless, we are

contemplating alternative ways of searching for fusion genes on a network structure, which I shall discuss in more detain in Chapter 5.

To conclude, in this chapter I present a method of detecting fusions of unrelated genes using the network structure analysis. So far I have proven that this method is accurate in reporting fusion from simulated data and biological data. The functionality and limitations of this method are discussed further in the following chapter.

Chapter 3 - Using FUSION: Homology Network Properties Reveal Fusions in *S. enterica*

In chapter two, I presented the algorithmic basis for a new means of discovering fusions of unrelated genes. It was shown that this produced favourable results when used with simulated data. In this chapter I will present results obtained from using the network based approach to find fusion genes in bacterial genomes.

3.1: Introduction

Fusion studies have often recognised and described individual cases of fusion genes (Ferretti *et al.* 1986; Zakharova *et al.* 1999; Tenorio *et al.* 2001). These studies provide little information in terms of how often this phenomenon is occurring. With the accumulation of completely sequenced genomes it has become possible to study evolution on a genome-wide scale. However, while many have moved on to quantifying fusions in completely sequenced genomes (Snel *et al.* 2000; Kummerfeld and Teichmann 2005) a clear and detailed picture of gene fusion in genomes is still lacking. Although gene fusion estimates have been made across a limited number of genomes, problems such as false positive results (Enright *et al.* 1999; Marcotte *et al.* 1999; Snel *et al.* 2000) and sampling bias (Enright *et al.* 1999; Marcotte *et al.* 1999; Suhre and Claverie 2004).

Snel and colleagues (Snel *et al.* 2000) obtained estimates of gene fusion and gene fission of orthologous genes. They found that, in general, there were more fusions than fissions per genome. They predicted that for an *E. coli* strain containing 4,290 genes, there were 33 fusion genes, which is less than 0.8% of the genome. Somewhat in agreement with Snel et al., was a study by Kummerfeld and colleagues (Kummerfeld and Teichmann 2005). In 131 genomes from all three domains of life, they identified 2,869 multi-domain proteins as a single protein in certain organisms and as two or more smaller proteins in other organisms. This suggests an average of 21.9 fusions per genome. They also concluded that the dominant process for evolution of composite proteins and their components is fusion, occurring four times more often than fission.

Pasek et al. (2006) used 'DomainTeam', dedicated to the search for microsyntenies of domains to find 'reshaped proteins' – proteins encoded by genes derived from a common ancestor that have evolved by modular assembly of domains. From 28 bacterial genomes they found 141 sets of homologous proteins that contained at least one 'reshaped protein'. They quantified the relative contribution to these events and found that, in 38 cases, gene fusion was the driver of evolution of the multi-domain protein. In conclusion they estimated that the contribution of fusion events to the evolution of multi-domain proteins is at least 27%.

It is evident that fusion events are occurring in bacterial genomes, and in many cases it is frequent (Enright, Iliopoulos et al. 1999; Suhre and Claverie 2004; Pasek, Risler et al. 2006; Fani, Brilli et al. 2007). The bigger question concerning fusion events is related to the types of gene that can or cannot, or are more or less inclined to fuse.

Tsoka and Ouzounis (Tsoka and Ouzounis 2000) had noted that most pairs of fusion component proteins of known function were involved in metabolic pathways (Enright *et al.* 1999). They retrieved 636 protein monomers or protein subunits from multimeric enzymes involved in small-molecule metabolism from the EcoCyte database (Karp *et al.* 1999). When they compared these proteins to 22 genomes they found 106 components that were involved in 96 fusion events. It would appear that 1 in 6 metabolic enzymes in *E. coli* are involved in a fusion event. Finally, when compared to control sets, they showed that metabolic enzymes from *E. coli* exhibit a threefold preference in fusion events.

Yanai et al. (2001) looked for fusion links, i.e. pairs of genes that have an alignment of at least 80 residues with the same protein but have a max overlap of 20 residues with one another. They found in the parasitic bacterium *Mycoplasma genitalium,* a genome of only 468 genes that there was 20 fusion links. From the genes involved in the 20 links they found 5 performed adjacent steps in metabolic pathways and 3 genes encoded sequential steps in glycolysis. Another 3 genes encoded RNA helicases.

The FusionDB (Suhre and Claverie 2004) website provides a matrix that displays the number of functional COG pairs that are found to have fused in a given set of genomes (web link 7). To create the matrix, Suhre and Claverie checked all annotated genes from 51 bacterial and archaean genomes against 89 fully sequenced bacterial and archaeal genomes. Two genes from a given reference genome were considered to be involved in a putative fusion event if they both matched the same open reading frame (ORF) in the target genome as their highest scoring BLAST hit. Any putative fusion events for which there was COG category information available

were added to the matrix. The matrix represents 12,724 distinct pairs of Clusters of Orthologous Groups of proteins (COGs).

Although there are fusions from all 20 categories in the matrix, those involved in metabolism appear to be overrepresented in comparison. It can also be seen that certain types of genes seem less likely to fuse together. For instance, from their results, Suhre and Claverie find that no instances where genes involved in energy production and conversion have fused with genes involved in cell motility. In fact there are very low numbers for fusions of genes involved in cell motility with any other type of gene, while genes involved in energy production and conversion appear to fuse much more readily.

The impact of known gene fusions has been particularly dominant in the emergence of novel antibiotic resistance genes. Since their discovery in 1910, the medical significance of antibiotics has been marred by the emergence of resistant microbes (Donadio *et al.* 2010). Antibiotic-resistance is a natural and ancient phenomenon. Genes for resistance to beta-lactam, tetracycline and glycopeptides were discovered in DNA from 30,000 year-old permafrost sediments (D'Costa *et al.* 2011). However, the evidence is overwhelming that evolutionary pressure from overuse of antibiotics has played a major role in the development of multidrug resistance (Livermore 2005). By the 1990s reports were issued warning people against overuse, misuse and use in animal feeds as growth promoters (Soulsby 2005). Despite knowing how it was caused, reducing the use of antibiotics did not reduce the resistance potential of bacteria (Wise 2004).

Bacteria outnumber humans by a factor of 10^{22} and they can go through as many as 500,000 generations to every 1 of ours (Schaechter *et al.* 2004). Today we find ourselves in the mist of multi-drug resistant “superbugs” that show no sign of an

evolutionary plateau (Spellberg *et al.* 2008). One strain in particular has epitomized the struggle against antibiotic resistance and found itself very much in the public eye. *Staphylococcus aureus* occurs as commensal on humans under ordinary conditions, however it does have the ability to cause infection and acquisition of resistance genes has rendered this organism a lethal pathogen (Foster 2004). The introduction of Penicillin in 1940 drastically reduced the number of *S. aureus* infections but only two years later resistant strains were discovered (Rammelkamp and Maxon 1942). Methicillin was introduced to treat *S. aureus* in 1959 and by the 1960s several isolates had acquired resistance (Barber 1961; Rolinson 1961). The mecA gene responsible for methicillin resistance, is part of a mobile genetic element found in all methicillin resistant *S. aureus* (MRSA) strains (Lowy 2003). This gene is not native to *S. aureus* but was acquired from an extraspecies source (Beck *et al.* 1986).

The human gut is a reservoir for antibiotic resistance genes (Cheng *et al.* 2012). Many of the bacteria living within the human microbiome were once considered commensal and relatively harmless but have now have emerged as multidrug resistant, disease-causing organisms (Sommer *et al.* 2009). Bacteria living the same environment can share genes easily and it has been shown that the intestinal microbiota plays a role in the development and transmission of antibiotic resistance via HGT (Donskey 2004). Despite their low inherent pathogenicity, the gram-negative opportunist strains are problematic in intensive care (Livermore 2009).

Bacteria are more commonly making use of bi-functional enzymes: single polypeptides with multiple catalytic activities by separate active sites (Ferretti *et al.* 1986; Deka *et al.* 2002; Allen *et al.* 2008; Zhang *et al.* 2009) Beta-Lactamases catalyze the hydrolysis of beta-lactam antibiotics rendering the antibiotics deactivated. More recently, as a result of evolution in response to antibiotic chemotherapy, beta-

lacatamases have been found to be constituents of fusion proteins with multiple types of resistance. The first bi-functional beta-lactamase was discovered in 2002 (Deka *et al.* 2002). Tp47, an enzyme possessing two active sites, came from *T. palladium*, the causative bacterium of syphilis. One of the enzymes active sites is a beta-lactamase while the other binds to penicillin. Despite being the most abundant membrane protein in Treponema palladium, its precise function is unknown; its activity may be insufficient in antibiotic resistance. However, there is no doubt surrounding the resistance potential of the bi-functional beta-lacatmase blaLRA-13 (Allen *et al.* 2008). This enzyme was derived from *E. coli* in soil and confers significant levels of beta-lactam resistance. The N and C terminus of the gene encoding this enzyme both show homology to different classes of beta-lactamase. While the N-terminus domain is involved in amoxicillin, ampicillin and carbenicillin resistance, the C-terminus domain encodes resistance to cephalexin. Together the domains confer a resistance to piperacillin.

Aminoglycosides are a useful antibiotic. They bind to the 30S ribosomal subunit and disrupt bacterial translation (Mingeot-Leclercq *et al.* 1999). Aminoglycoside modifying enzymes are usually monofunctional but a number of bi-functional ones have arisen through fusion events (Donato, 2010, Dubois, 2002, Ferretti, 1986, Kim, 2006, Mendes, 2004). They often have two different aminoglycoside-modifying activities as separate domains of the same gene (Zhang *et al.* 2009). The first to be discovered was the aacA-aphD gene that encodes the bi-functional enzyme 6'-aminoglycoside acetyltransferase (AAC-6') -2" - aminoglycoside phosphotranferase (APH-2") (Ferretti *et al.* 1986). This gene, being the resistance determinant of the *S. aureus* transposon Tn4001, which specifies resistance to gentamicin, tobramycin and kanamycin has been cloned and shown to

express these resistances in *E. coli*. It has also been found to mediate gentamicin resistance in *Enterococcus faecalis*. When the AAC(6')-APH(2") enzyme from Streptococcus plasmid was cloned in *E. coli* it was found that the N-terminal was homologous to chloramphenicol acetyltransferase of *B. plumilus*. The C-termial was homologous to aminoglycoside phosphotransferase of *Streptomyces fradiae*.

From 2002 to 2010 four more bi-functional proteins were found to be involved in aminoglycoside resistance (Donato, 2010, Dubois, 2002, Kim, 2006, Mendes, 2004). The novel proteins resulted from the fusion of a highly specified domain to a domain with a much broader substrate acceptance. The fused protein shows a broader aminoglycoside-modifying enzyme activity compared with either protein alone. One of these fusion proteins was found in the soil of an apple orchard (Donato *et al.* 2010). Of 13 antibiotic resistance genes that were found, 2 encoded bi-functional proteins. While one was involved in aminoglycoside resistance the other conferred resistance to ceftazidine. The ceftadizine resistant bi-functional protein contains a natural fusion between a predicted transcriptional regulator and a beta-lacatmase, the first discovered of its kind.

The overall aim of this chapter is to discover how many and what kinds of fusions can be found in bacterial genomes. First of all, we wished to understand the extent to which gene fusion is a feature of one particular bacterial genome – that of *S. enterica subsp. enterica serovar Paratyphi A*. This particular strain of *Salmonella* is the second most prevalent cause of typhoid, responsible for one third of cases or more in southern and eastern Asia (McClelland *et al.* 2004). *Salmonella* belong to the YESS group, a group of medically and scientifically important bacteria including the genera *Yersinia*, *Escherichia*, *Shigella* and *Salmonella*. This group contains many pathogens and has had an interesting evolutionary history. It has been difficult to

describe the evolutionary relationships between these species. Depending on the method of analysis and the depth of the evolutionary history under consideration, several different kinds of conflicting results can be recovered for this group (Haggerty *et al.* 2009).

As discussed in chapter 2, we make use of algorithm 457 (Bron and Kerbosch 1973) otherwise known as the Bron-Kerbosch algorithm in order to find fusion genes. The worst-case time complexity of the algorithm is $O(3^{n/3})$, which means that the larger the dataset, the larger the network will be and it will take longer to search the network for cliques. To overcome the methodological hurdle of how to do this analysis we restrict the size of the dataset. The optimal sized dataset consists of approximately 20,000 genes, the equivalent of four average sized bacterial genomes. Our initial dataset was sampled from the YESS group and contained the genomes for *E. coli* str. K-12 substr. MG1655*, Shigella sonnei* Ss046*, Yersinia enterocolitica* subsp. enterocolitica 8081 and *Salmonella enterica* subsp. enterica serovar Paratyphi A.

In addition to the initial dataset, sampled from the YESS group, we constructed a further four datasets consisting of quartets of different genomes. The genomes were selected from the major divisions of bacteria as described by Ciccarelli et al (2006) (Figure A3, Appendix). They ranged from shallow relationships between four different gamma-proteobacteria, to much deeper relationships between species from four different major divisions (Table 3.1). The genome for *Salmonella enterica subsp. enterica serovar Paratyphi A* was included in all five datasets so that later we could predict the total number of fusions in a genome.

We also aim to discover whether there is a bias in the kinds of gene fusions we detect. To date there have been many studies to estimate the frequency of gene fusion

events and to find out what types of genes are likely to fuse. However, a clear picture of gene fusion in bacterial genomes is still lacking, despite the availability of enormous numbers of genome sequences. In this chapter I attempt to gain further understanding of gene fusion in bacteria using an approach that, has not been used before.

The objective of this chapter was to focus on one genome – a strain of *Salmonella* – and using a variety of triplets of additional genomes, estimate how many true fusion genes are present in this target genome. Given the computational limitation of approximately four prokaryotic genomes, or 20,000 genes, it was not possible to perform an analysis with, say, several hundred genomes. Therefore, by using an approach that analysed the *Salmonella* genome from the perspective of a variety of other genomes, we were able to monitor the rate at which we discovered new gene fusions and we could quantify the overlap in fusion discovery in the *Salmonella* genome when we used different combinations of other genomes.

Table 3.1: List of genomes, their accession number in GenBank and the dataset they were used in.

Species	**Accession Number**	**Used in datasets…**
Salmonella enterica subsp. enterica serovar Paratyphi	NC_000913	1, 2, 3, 4, 5
Escherichia coli str. K-12 substr. MG1655	NC_011147	1
Shigella sonnei Ss046	NC_008800	1
Yersinia enterocolitica subsp. enterocolitica 8081	NC_007384	1
Petrotoga_mobilis_SJ95	NC_010003	2
Streptomyces_coelicolor_A3(2)	NC_003888	2
Leptospira_borgpetersenii serovar Hardjo-bovis L550	NC_008508	2
Pseudomonas_aeruginosa_PAO1_chromosome	NC_002516	3
Shewanella_oneidensis_MR-1	NC_004347	3
Actinobacillus_pleuropneumoniae_L20	NC_009053	3
Streptobacillus moniliformis	NC_013515	4
Meiothermus ruber	NC_013946	4
Pirellula staleyi	NC_013720	4
Synechocystis sp. PCC 6803	NC_000911	5
Alkaliphilus oremlandii	NC_009922	5
Prosthecochloris vibrioformis	NC_009337	5

3.2: Materials and Methods

3.2.1: Runtime analysis

In chapter 2 I described our analyses of the accuracy of our algorithm. These analyses focused on the ability of the algorithm in constructing a network from a text file, in obtaining the information needed to produce postscript diagrams of the alignment area between fusions and their component genes and finally in the execution of the overlap test. In this chapter we have the opportunity to further our understanding of the functionality of our algorithm and its ability to accurately find fusions in much larger dataset than those presented in chapter 2. In addition we obtain data insights relating to the effect of input data size and complexity and the time it takes to execute all parts of the algorithm on the datasets.

3.2.2: Quantifying fusions in a dataset

Five datasets, each containing all annotated genes from the genome for *S. enterica subsp. enterica serovar Paratyphi A* and three additional bacterial genomes were constructed for this study. For our initial analysis we simply execute the algorithm described in chapter 2 on each of the five datasets in order to obtain an estimate of the frequency of fusion events within a subset of bacteria. By using datasets with different degrees of relatedness we hope to see whether fusion events occur more readily

within a group of genomes that are more closely related than within a group that is quite divergent.

3.2.3: Overlap between datasets

When we obtain a list of fusions found in each of the dataset we expect that there will be a certain amount of overlap. The same fusion genes could be found in more than one analysis for a number of reasons. Firstly we used the same *Salmonella* genome in all five of our datasets. We may find that there is a fusion gene that is found in the *Salmonella* genome and both of its component genes are also found in the *Salmonella* genome. No matter what additional genomes we include in our analysis this particular fusion will be present and therefore detectable. Similarly if a fusion gene in the *Salmonella* genome is made up of component genes that are universally distributed throughout the bacteria, then the fusion and its components will always be contained in the data. There is also the possibility of finding orthologous fusion genes. A fusion gene found in a genome in dataset 1 could be homologous to a fusion gene found in a different genome in dataset 2. This may be the result of a HGT event or the fusion gene may have arisen from multiple independent origins.

Within each dataset we are also likely to find paralogous fusion genes (in-paralogs), e.g. as a result of a duplication event within a genome. It is also very likely that a fusion gene in one genome in a dataset will have homologous fusion genes in other genomes in the same dataset. In order to quantify the number of unique fusion genes in a given dataset we count each unique fusion. In other words, within each dataset just one fusion gene will account for itself and all its fusion gene homologs. Finally, to quantify the number of unique fusions across all five datasets we compare all of the

unique fusion genes from each dataset in an all-vs-all similarity search. When two genes from different datasets are hit in the BLAST search we consider them the same fusion, and therefore contained in the overlap of the two datasets.

3.2.4: Estimating the number of fusions in *S. enterica subsp. enterica serovar Paratyphi A*

We want to use four different analyses of *Salmonella* gene fusions. Each analysis should provide a particular perspective on what genes have been involved in a fusion event. Detecting these fusions will naturally require that the two donor genes are present in the dataset as well as at least one instance of the fusion gene. We wanted to compare the different analyses in terms of the overlap in identified fusion genes. If each analysis found completely different fusion genes, then this would tell us one thing and if each analysis found the exact same fusion genes, it would tell us something different. Our goal was to find and record gene fusions, the rate at which we find them and the biochemical and physiological functions of their protein products.

The numbers of *Salmonella* fusions found in all datasets are then used to make a prediction of the overall number of fusion gene in the *Salmonella* genome. We use a "mark and recapture" population size estimation method called the Schnabel estimator to predict the number of fusions overall in the *Salmonella* genome. The Schnabel estimator is defined as:

$$N = \frac{\sum_{t=1}^{s} \left(C_t M_t \right)}{\left(\sum_{t=1}^{s} R_t \right) + 1} \qquad 1.$$

Where C_t is the total number of individuals caught in sample t or in our case the total number of *Salmonella* fusions found in the current dataset. R_t is the number of individuals already marked when caught in sample t or for our purposes, the number of *Salmonella* fusions found in the current dataset that were already found in the previous datasets. M_t is the number marked individuals in the population just before the current sample was taken or the number of *Salmonella* fusions found in all datasets so far. S is the number of the sample in the series, i.e. the number of the dataset.

Finally we established an estimate of the number of unique fusions that could be discovered as more datasets are analysed. We begin with the results from dataset 1 and as we include the results from further datasets we count the number of fusions that we had not already see in previous datasets. We fit the increasing numbers of new fusions found as each dataset was added to a logarithmic curve. At the plateau of the curve we find an estimate of the number of unique fusions within the *Salmonella* genome.

3.2.5: COG category enrichment of fusion genes

As well as estimating the frequency of fusion genes in a genome, we hoped to discover which kinds of genes come about as a result of fusion events. In other words for the fusion genes that we found in our five datasets we want to find out which functional categories they are involved in.

In order to find the functional category to which each fusion gene belongs, we performed a similarity search of fusion genes against a database of COG-categorised genes. For each fusion gene we found the functional category to which it was most likely to belong. Our final analysis involved discovering whether there are specific functional categories to which most fusions belong. In other words of the 22 COG categories, we want find out which contain more fusions than expected and which contain fewer than we would expect.

In order to evaluate whether some categories of genes are over represented or underrepresented in the collection of fusion genes, we carried out a chi-squared test. The "expect" value for this test was obtained by first of all categorizing all genes in all datasets according to their function. We calculated the percentage of genes from each dataset that falls into each category. We also obtained the functions of the fusion genes and expressed these as a percentage of overall functions. We could then compare these values using a chi-squared test to see if there was a significant difference between the two values.

3.3: Results

3.3.1: Network of genes in each dataset

As discussed in chapter 2, the first step of the FUSION algorithm is the construction of a network from an all-versus-all BLAST search of all annotated genes from the specified genomes of the dataset. Different datasets yield networks of different size and shape. In dataset 1, for instance there are 16,415 genes, which means there are 16,415 nodes on the network. The network constructed from gene relationships in dataset 2 has 16,443 nodes, 28 more than dataset 1. However, the number of edges on the network of dataset 1 exceeds the number of edges on the network of dataset 2 by 93,050. Despite there being fewer genes in dataset 1, the network is far more highly connected than that of dataset 2. This is explained by the fact that genomes used in dataset 1 are more closely related than the genomes used in dataset 2. All of the genomes in dataset 1 come from the YESS group, which contains 4 enteric *Gamma-proteobacterial* species. Whereas the genomes in dataset 2 have been sampled from 4 different phyla of bacteria- one from the *Thermatoga*, one from the *Actinobacteria*, one from the *Spirochaetes* and the *Salmonella* genome from the *Proteobacteria*. The sizes of each network in terms of nodes and edges are displayed in table 3.2.

Table 3.2: Sizes of each network, constructed from an all-versus-all BLAST search of all annotated genes in each dataset. The size of each network is reported as number of nodes and as number of edges.

Dataset	Taxonomy	No. Nodes	No. Edges
1	*Yersinia, Shigella, Escherichia coli, Salmonella*	16,415	171,315
2	*Leptospira, Streptomyces, Petrotoga, Salmonella*	16,443	78,263
3	*Salmonella, Pseudomona, Shewanella, Actinobacillus*	15,975	89,567
4	*Salmonella, Streptobacillus, Meiothermus, Pirellula*	13,239	54,383
5	*Salmonella, Synechocystis, Alkaliphilus, Prosthecochloris*	11,842	47,459

3.3.2: Runtime Analysis

In chapter 2 we used relatively small datasets, the largest yielded a network of 4,145 nodes and 11,152 edges. When we executed the algorithm on such a small dataset it completed in just seconds. Our intention for the analyses in this chapter was to use as many genomes as possible. Unfortunately we found that when we attempted to execute the algorithm on a preliminary dataset containing large numbers of genomes it was unable to find results in real time. This is most a likely a result of the clique-finding step of the algorithm that has a worst-case time complexity of $O(3^{n/3})$. This means that the larger the dataset, the larger the network will be and it will take longer to search the network for cliques. Through trial and error we found that the algorithm will complete in real time on a dataset of four genomes or approximately 20,000 genes. The size of each dataset in memory and the runtime for the algorithm on each dataset are displayed in table 3.3.

3.3.3: Fusion Genes

Overall the analyses of 16 different genomes yielded 157 fusion genes that were unique within their dataset. In other words each fusion event is only counted once, even if there are multiple instances of the fusion gene in the dataset. The number of unique fusions per dataset is displayed in table 3.4.

Table 3.3: Size in memory and runtime for each of the five datasets.

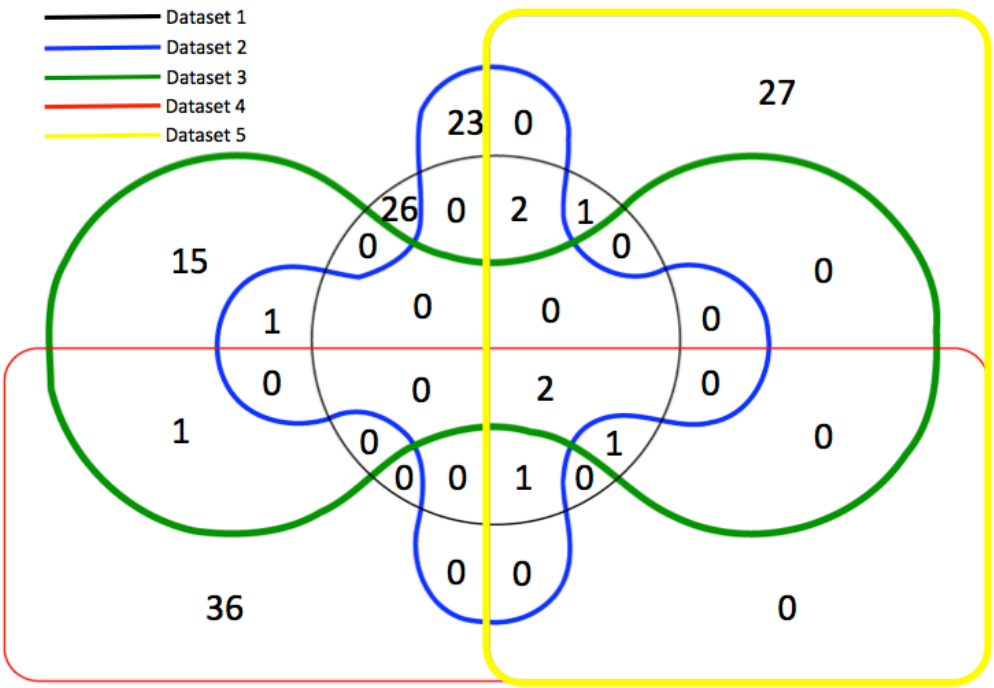
Dataset	**Taxonomy**	**Size in memory (megabytes)**	**Runtime (hours)**
1	*Yersinia, Shigella, Escherichia coli, Salmonella*	11.2	13.5
2	*Leptospira*, *Streptomyces*, *Petrotoga*, *Salmonella*	7.2	13
3	*Salmonella*, *Pseudomona*, *Shewanella*, *Actinobacillus*	9.1	19
4	*Salmonella*, *Streptobacillus*, *Meiothermus*, *Pirellula*	3.7	19
5	*Salmonella*, *Synechocystis*, *Alkaliphilus*, *Prosthecochloris*	3.4	10

Table 3.4: Number of fusion genes found in each of the five datasets. Fusion genes reported here are fusions of unrelated genes and unique within their dataset.

Dataset	1	2	3	4	5
No. Fusions	33	29	20	41	34

3.3.4: Overlap Between Datasets

When the datasets were tested for overlapping fusions it was shown that only two were common to every dataset. Just nine of the 157 fusions were found in more than one dataset. The overlap between all five datasets is represented in the five-way Venn diagram (Figure 3.1).



n = 136

Figure 3.1: Five-way Venn diagram representing the overlap of fusion genes between the five datasets.

3.3.5: Quantifying Fusions in a Genome

Overall, we found 186 fusion genes in the *Salmonella* genome, distributed among the five datasets as shown in table 3.4. However, there is a substantial amount of overlap between the datasets (Figure 3.2), in other words the same *Salmonella* fusion gene is often detected in more than one dataset. In fact 16 of the *Salmonella* fusions are found in all five datasets and only 24 *Salmonella* fusions are unique to one dataset.

In order to gain an estimate of the number of fusion genes in the *Salmonella* genome we used the numbers of fusions found in each of the five datasets (Table 3.6) in a mark and recapture population size estimation method called the Schnabel estimator:

$$N = \frac{\sum_{t=1}^{s}\left(C_t M_t\right)}{\left(\sum_{t=1}^{s} R_t\right)+1}$$

$$N = \frac{(0\times 46)(46\times 30)(76\times 35)(111\times 39)(140\times 36)}{(0+23+28+35+32)+1} = 112.7$$

From the Schnabel estimator we predict that the *Salmonella* genome contains 112.7 fusion genes. There are 4074 genes in this particular strain so this equates to 2.8% of the genome.

To add support to our estimation of the number of fusions in *Salmonella*, we quantified the number of new unique fusions found after each dataset is analysed. The best-fitting curve for the data has the equation **y = 13.733ln(x) + 45.05** (Figure 3.3). The data plateaus at approximately 118 fusion genes, in close agreement with our estimation from the capture-recapture method.

Table 3.5: Number of *Salmonella* fusion genes found in each of the five datasets. Fusion genes reported here are fusions of unrelated genes found specifically in the *Salmonella* genome.

Dataset	1	2	3	4	5
No. Fusions	46	30	35	39	36

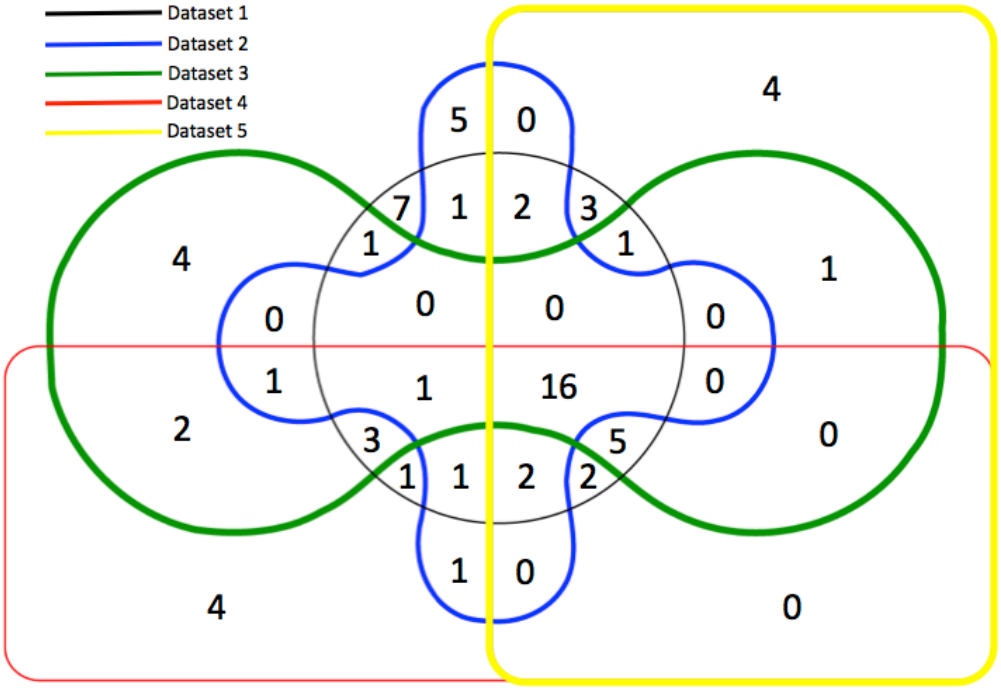


Figure 3.2: Five-way Venn diagram representing the overlap of *Salmonella* fusion genes between the five datasets.

Table 3.6: Numbers used in the Schnabel estimator to predict the number of fusion in the *Salmonella* genome. Where ***t*** is the dataset number, C_t is the total number of individuals caught in sample ***t*** or in our case the total number of *Salmonella* fusions found in the current dataset. R_t is the number of individuals already marked when caught in sample ***t*** or for our purposes, the number of *Salmonella* fusions found in the current dataset that were already found in the previous datasets. M_t is the number marked individuals in the population just before the current sample was taken or the number of *Salmonella* fusions found in all datasets so far.

	Dataset 1	**Dataset 2**	**Dataset 3**	**Dataset 4**	**Dataset 5**
C_t	46	30	35	39	36
R_t	0	23	28	35	32
M_t	0	46	76	111	140

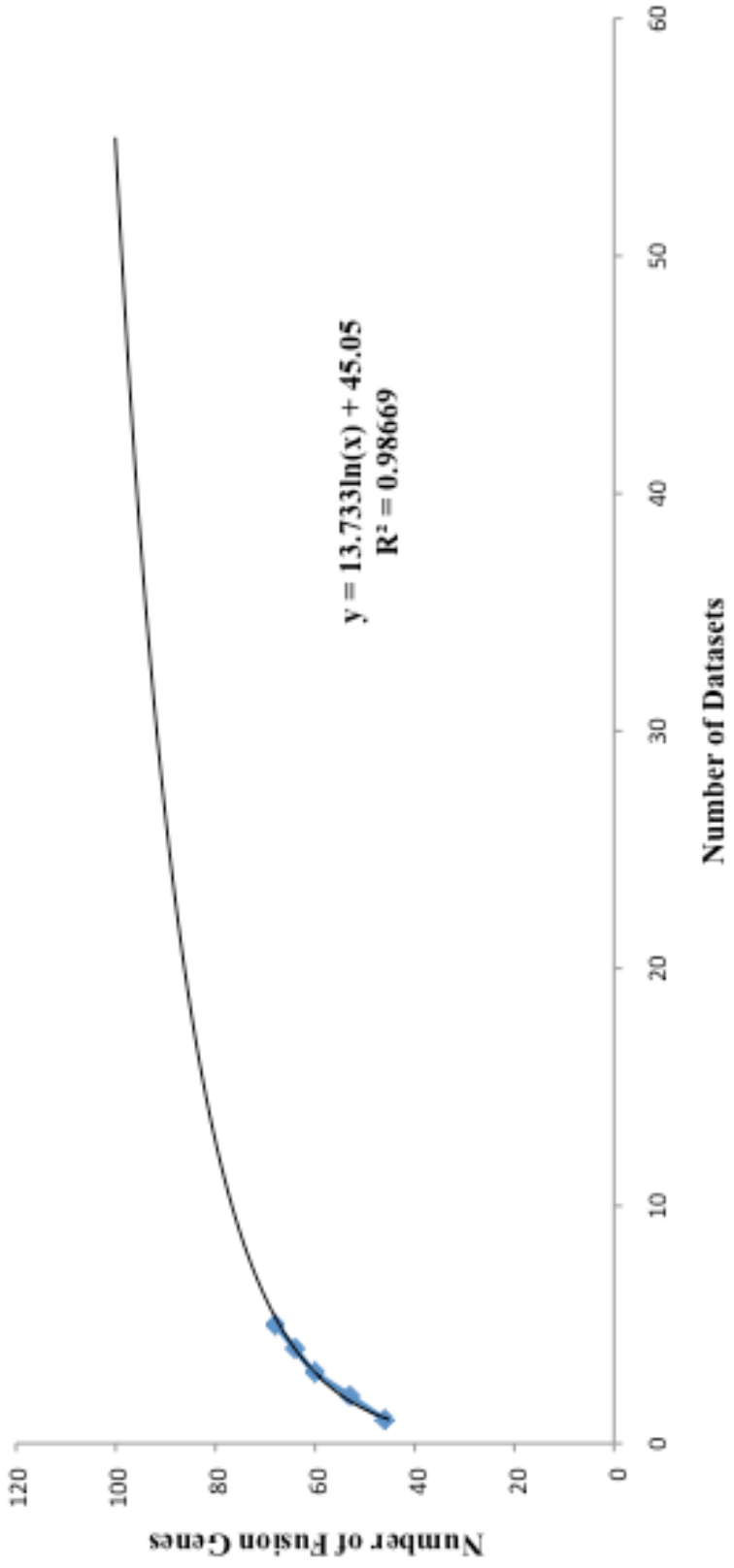


Figure 3.3: Increasing numbers of new fusions as more datasets are analysed.

3.3.6: COG function enrichment

The results of the COG enrichment analysis are displayed in table 3.7 and 3.8 on Figure 3.4. We found that for all 5 datasets there were significantly more than expected fusions (highlighted in black in tables 3.7 and 3.8 and marked with stars on Figure 3.4) involved in defense mechanisms, this would include mechanisms of antibiotic resistance. For 3 of the 5 datasets, Carbohydrate transport and metabolism was significantly enriched.

There are a few categories for which there are very few or no fusions, these include RNA processing and modification (significantly less than expected results are highlighted in orange in tables 3.7 and 3.8).

Table 3.7: Number of fusions in each COG category for datasets 1 to 3. The expected value is calculated as the expected number of fusions in a given category based on the percentage of the all the genes in the dataset that fall into that category. Black boxes indicate that there are more fusions than expected for that category and orange boxes indicate that there are less than expected fusions in that category.

	Dataset 1			dataset 2			dataset 3		
COG Categories	OBS	EXP	P-value	OBS	EXP	P-value	OBS	EXP	P-value
INFORMATION STORAGE AND PROCESSING									
Translation, ribosomal structure and biogenesis	0	1.44	0.23	1	1.09	0.93	0	0.89	0.34
RNA processing and modification	0	0.01	0.91	0	0.004	0.95	0	0.006	0.94
Transcription	0	1.92	0.17	0	1.39	0.24	0	1.18	0.28
Replication, recombination and repair	4	2.4	0.3	0	0.98	0.32	2	0.78	0.17
CELLULAR PROCESSES AND SIGNALING									
Cell cycle control, cell division, chromosome partitioning	0	0.29	0.59	0	0.16	0.69	0	0.15	0.7
Defense mechanisms	5	0.37	1.84E-14	4	0.41	1.82E-08	4	0.25	7.89E-14
Signal transduction mechanisms	0	1.07	0.3	1	0.7	0.72	2	0.82	0.2
Cell wall/membrane/envelope biogenesis	0	1.59	0.2	1	1	1	0	0.89	0.34
Cell motility	0	0.18	0.67	1	0.12	0.01	0	0.09	0.76
Intracellular trafficking, secretion, and vesicular transport	0	0.94	0.33	1	0.33	0.25	0	0.52	0.47
Posttranslational modification, protein turnover, chaperones	0	1.04	0.31	0	0.6	0.44	0	0.67	0.41
METABOLISM									
Energy production and conversion	0	1.98	0.16	3	1.29	0.13	2	1.14	0.42
Carbohydrate transport and metabolism	12	2.19	3.34E-11	5	1.59	0.007	0	0.76	0.38
Amino acid transport and metabolism	0	2.26	0.13	0	1.49	0.22	0	1.32	0.25
Nucleotide transport and metabolism	0	0.6	0.44	1	0.42	0.38	0	0.35	0.56
Coenzyme transport and metabolism	1	1.09	0.93	2	0.68	0.11	1	0.65	0.67
Lipid transport and metabolism	1	0.52	0.5	0	0.5	0.48	0	0.43	0.51
Inorganic ion transport and metabolism	0	1.59	0.2	1	0.89	0.91	0	0.91	0.34
Secondary metabolites biosynthesis, transport and catabolism	1	0.39	0.33	1	0.38	0.32	1	0.3	0.19
POORLY CHARACTERIZED									
General function prediction only	5	3.75	0.52	4	2.65	0.4	3	2.21	0.6
Function unknown	1	2.54	0.33	0	1.26	0.26	2	1.51	0.69
No hit in COG database	3	4.84	0.4	3	11.07	0.02	3	4.14	0.57
Total fusions in dataset			33			29			20

Table 3.8: Number of fusions in each COG category for datasets 4 and 5. The expected value is calculated as the expected number of fusions in a given category based on the percentage of the all the genes in the dataset that fall into that category. Black boxes indicate that there are more fusions than expected for that category and orange boxes indicate that there are less than expected fusions in that category.

	Dataset 4			dataset 5		
COG Categories	OBS	EXP	P-value	OBS	EXP	P-value
INFORMATION STORAGE AND PROCESSING						
Translation, ribosomal structure and biogenesis	0	1.75	0.19	0	1.71	0.2
RNA processing and modification	0	0.006	0.94	0	0.002	0.96
Transcription	0	1.38	0.24	0	1.41	0.24
Replication, recombination and repair	0	1.3	0.25	2	1.53	0.7
CELLULAR PROCESSES AND SIGNALING						
Cell cycle control, cell division, chromosome partitioning	0	0.26	0.61	0	0.3	0.58
Defense mechanisms	2	0.53	**0.04**	3	0.49	**0.0003**
Signal transduction mechanisms	1	1.06	0.96	2	1.32	0.56
Cell wall/membrane/envelope biogenesis	1	1.36	0.76	2	1.58	0.74
Cell motility	0	0.11	0.75	0	0.11	0.74
Intracellular trafficking, secretion, and vesicular transport	0	0.71	0.4	0	0.66	0.42
Posttranslational modification, protein turnover, chaperones	0	1.01	0.32	1	0.99	0.99
METABOLISM						
Energy production and conversion	2	1.68	0.43	2	1.93	0.96
Carbohydrate transport and metabolism	1	1.98	0.49	4	1.39	**0.03**
Amino acid transport and metabolism	2	2.24	0.87	1	1.9	0.51
Nucleotide transport and metabolism	1	0.73	0.75	0	0.64	0.42
Coenzyme transport and metabolism	2	1.03	0.34	3	1.25	0.12
Lipid transport and metabolism	0	0.61	0.43	2	0.53	**0.04**
Inorganic ion transport and metabolism	0	1.45	0.23	1	1.5	0.68
Secondary metabolites biosynthesis, transport and catabolism	1	0.37	0.29	2	0.29	**0.002**
POORLY CHARACTERIZED						
General function prediction only	9	3.64	**0.005**	3	3.51	0.78
Function unknown	1	1.97	0.49	0	2.25	0.13
No hit in COG database	18	15.8	0.58	6	8.69	0.36
Total fusions in dataset			41			34

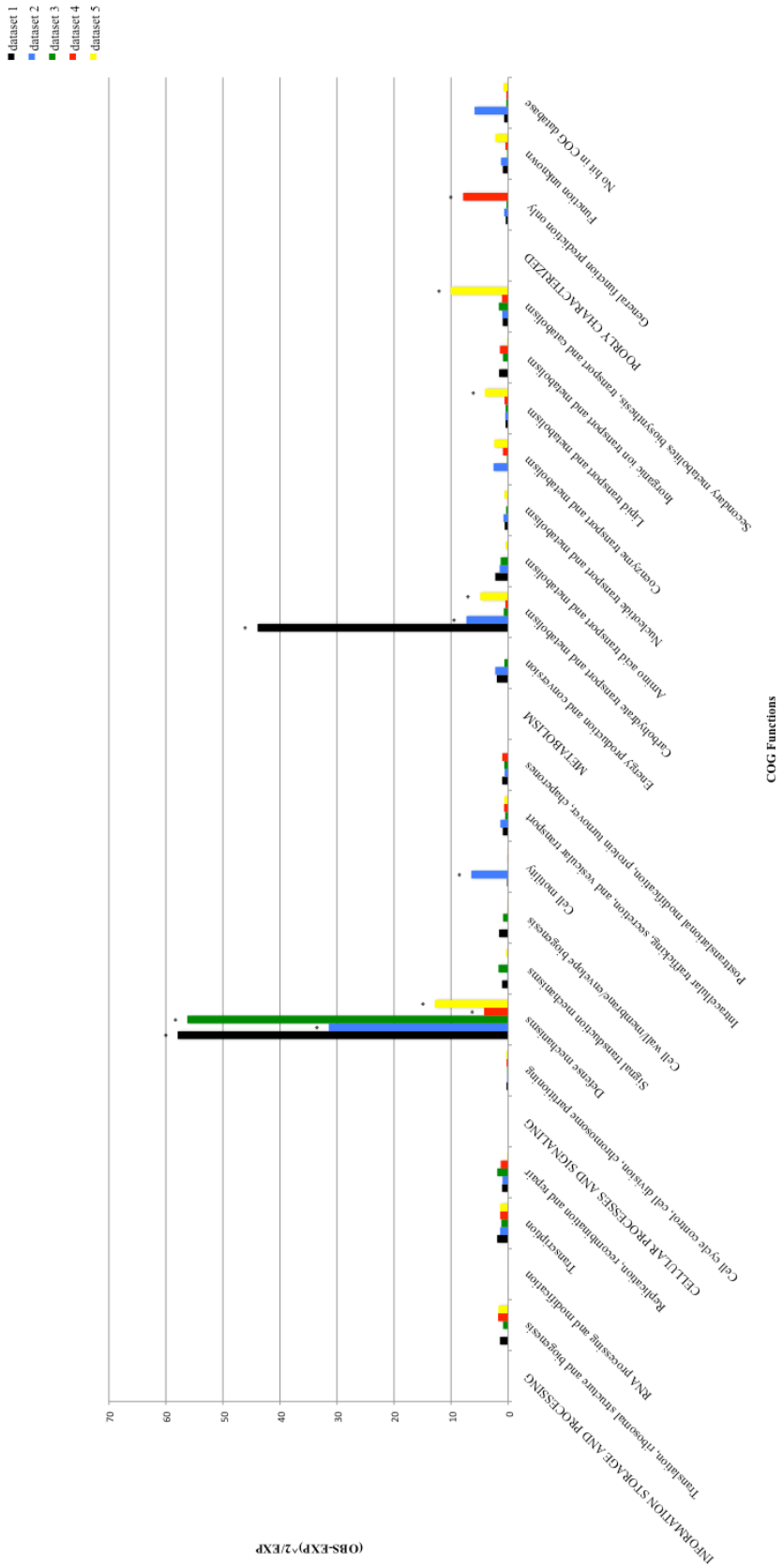


Figure 3.4: COG categories for all fusion genes. Stars denote categories where we observe significantly more fusions than expected.

3.4: Discussion

In this chapter I explore the use of a new method of detecting fusions using similarity searches and network mathematics. In chapter 2 I raised the subject of limitations in the method in terms of computational power. The worst time complexity of the clique-finding algorithm is $O(3^{n/3})$, where n is the number of nodes. As the amount of input data increases so does the size of the network increases and in turn the time it takes to search the network for all cliques. Through trial and error we found that the optimal dataset size on which the algorithm will run is approximately 20,000 genes or the equivalent of four average-sized bacterial genomes.

Using manageable sized datasets meant that our results would be limited by the input data. For four genomes any fusions that were detected would be exclusively found in those four genomes as would their component genes. In an attempt to overcome this we created an overlap of one genome in every dataset. The hope was that as we assessed more data, the picture of fusion events in this one genome would become clearer. In the future the hope is that we can describe all genomes in terms of all over genomes by means of overlapping datasets or otherwise.

In terms of quantifying fusions we found that there are between 20 and 41 fusions per 4 genomes. This averages at 7.85 fusion genes per genome, which at face value seems to be fewer fusions than were found in previous studies but with the potential to increase with the addition of more data. However, the use of the overlapping *Salmonella* genome analyses provided us with a means of assessing the extent of fusion within a genome in relation to 5 triplets of different genomes. This approach, while not being an exhaustive analysis of all genomes simultaneously

increased the scope of our analysis and provided us with excellent insight into the amount of fusion occurring in *S. enterica subsp. enterica serovar Paratyphi A*.

Despite the knowledge that we were only gaining an estimate of fusion occurrence based on a small portion of the data, we still report a considerable result. Use of the Schnabel population estimator provides us with the prediction that this *Salmonella* genome contains around 113 fusions of unrelated genes. This implies that approximately 2.8% of the *Salmonella* genome is made up of fusions of specifically unrelated genes. This last point is quite important – we are only analyzing the fusion of non-homologous genes and it is almost a certainty that this genome is replete with other fusions of homologous genes. Our estimate of 2.8% is supported by the best-fit curve to the data. We can conclude with confidence that fusion events are contributing a notable amount to bacterial genome evolution.

On the question of what types of genes appear as fusions in bacteria: we can somewhat agree with previous studies in saying that fusions are involved in metabolic functions. It is also undeniable that fusion genes are very often involved in defense. We consistently find that fusions of unrelated genes fall into the COG category of defense mechanisms.

Chapter 4 - Phylogenetic and Non-phylogenetic Signals During the Separation of the Enteric Bacterial Species Clouds

4.1: Introduction

It has become increasingly clear that the Tree of Life (ToL) hypothesis has limitations in its ability to describe the evolution of all evolving entities on the planet. The discoveries of conjugation (Lederberg and Tatum 1946), transduction (Zinder and Lederberg 1952), transformation (Griffith 1928), plasmids (Hayes 1953), bacteriophage (Duckworth 1976), Gene-Transfer Agents (GTAs) (Lang and Beatty 2000) and nanotubes (Dubey and Ben-Yehuda 2011) have caused great problems when constructing branching diagrams of life. These processes and associated mobile genetic elements (MGEs), that facilitate horizontal gene transfer (HGT), have the potential to disrupt the vertical inheritance pattern that is expected from the ToL hypothesis.

The most notable shortcoming of using tree structures to describe life is that it does not deal with all the evolving entities on the planet. MGEs have normally been excluded from discussions of the grand schemes of evolution of life. This might be permissible if MGEs played a very small role in the evolutionary history of life on the planet, but the data suggests otherwise.

The success of a gene can be measured by its ability to persist in nature and to be spread throughout genomes or biomes (Orgel and Crick 1980). Genes need to be

flexible enough to adapt to new environments while still maintaining enough sequence conservation to keep their protein structure (Drummond and Wilke 2008; Koonin 2009). Recently a study was conducted whereby the prevalence of all biological functions was estimated (Aziz *et al.* 2010). Aziz et al calculated the abundance and ubiquity of all functions encoded in genomes and ecosystems with the assumption that these values are correlated with gene fitness. Alone, abundance of a gene is simply an indicator of the genes ability to express adaptive, organism-specific functionality. Ubiquity is an indicator of essentiality; usually genes with essential functions are carried in every genome. Abundance and ubiquity together provide a measurement of gene success.

What Aziz et al found was somewhat surprising. They did not find that highly expressed ribosomal proteins were most pervasive, despite the fact that these genes are universally distributed throughout cellular life forms. Even though the enzyme Ribulose-1,5-bisphosphate carboxylase oxygenase (RuBisCo), critical in fixation of carbon dioxide via the Calvin cycle, has been found to be the most successful, most abundant enzyme on the planet (Dhingra *et al.* 2004), they did not find that it came top of their calculations. Nor did they find that DNA polymerase genes, essential for DNA-based life, were the most prevalent. Instead Aziz and colleagues demonstrated that transposases are the most abundant genes in both completely sequenced genomes and environmental metagenomes, and are also the most ubiquitous in metagenomes.

A transposable element or a transposon is a defined segment of DNA that has the ability to move, or copy itself, into a second location without a requirement for DNA homology (Curcio and Derbyshire 2003). A transposase then, is the enzyme that is responsible for the catalysis of transposition. This is but one example of the MGEs that are deliberately omitted from many phylogenetic studies.

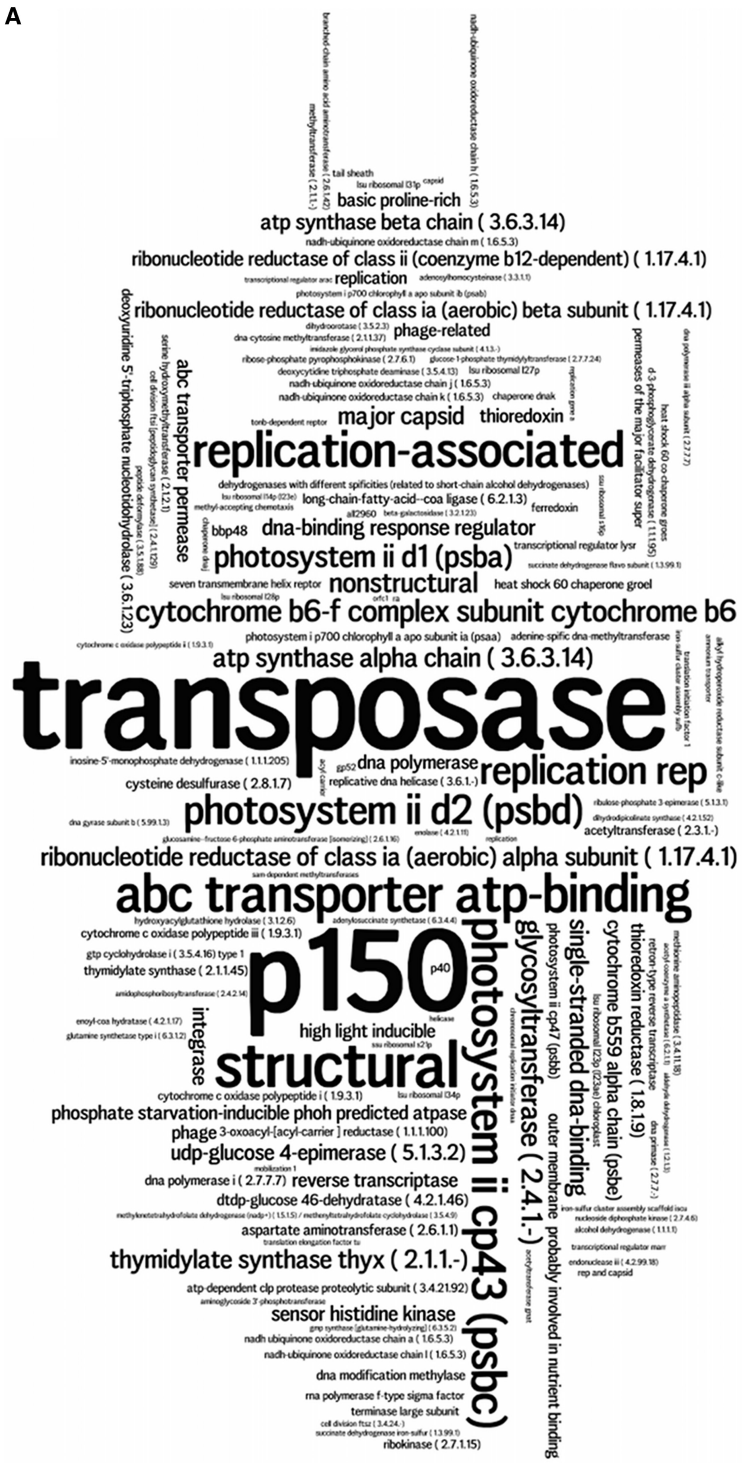


Figure 4.1: Word cloud representing the 100 most abundant functional roles taken from Aziz et al. 2010.

Phages are the most abundant life form on the planet, it is estimated that there are 10^{30} tailed phage particles (Frost *et al.* 2005). Their genetic diversity is enormous (Canchaya *et al.* 2004; Nelson 2004) and at 10^{25} infections per second, they are the most rapidly replicating entity on earth (Frost *et al.* 2005). Their mosaic structure results from their ability to recombine with other prophages and MGEs that reside in the same bacterial host (Hendrix 2003).

There is recognition now that, when the goal is to describe life in its most fundamental way, all evolving entities should be included (McInerney *et al.* 2011; McInerney *et al.* 2011). Also, since it has been observed that genes can be acquired by both vertical and horizontal transmission, a description of life should encompass both the vertical and horizontal components of evolution (Boto, 2010, Dagan, 2009, McInerney, 2011). The public goods hypothesis, proposed by McInerney et al. (2011) describes entities such as nucleotides, genes, operons or even genomes as genetic goods in the same sense as goods are viewed in the discipline of economics (Samuelson 1954). In economics goods can be shared, exchanged modified, etc. A good can also be excludable or rivalrous. One can limit the use of an excludable good by others and the use of a rivalrous good prevents its use by others altogether.

If we apply the "public goods" way of thinking to DNA we can say, for the most part, that DNA is not excludable- the same nucleotides are used by all entities, recombination is global and the machinery of DNA replication and translation is pretty universal. So if DNA is a "good", it can be shared or exchanged between different genetic entities, for example genes or genomes. The public goods hypothesis does not require ad hoc amendment or qualifications and so incorporates all of the data (McInerney *et al.* 2011).

Networks are capable of displaying both the vertical and horizontal components of evolutionary histories (Ragan 2009). On these DNA-sharing networks nodes represent the various kinds of evolving entities and edges represent identifiably homologous regions that are shared. The structural properties of these DNA-sharing networks open up new insights into genome evolution (Dagan 2011). Halary at al. (2010) constructed a gene-sharing network on which nodes represented genomes sampled from 111 prokaryotes and Eukaryotes along with thousands of phage and plasmid protein sequences. These genomes and protein sequences are joined by an edge on the network if they share at least one homologous DNA fragment. On one hand they found small genetic worlds. Most protein families preferred a specific type of DNA carrier and thus a boundary was created between the different types. On the other hand, their network contained a large highly interconnected component containing genes from chromosomes, phages and plasmids. The high frequency of links between chromosomes and plasmids in the large connected component suggests a prevalence of conjugal HGT in nature.

Node connectivity quantifies how many direct neighbours a node on a network has and node centrality quantifies how often a node falls on a shortest path between two nodes (Newman 2010). These measures in a gene-sharing network show how preferential attachment between nodes can result from the evolutionary history of the network. As expected, a large genome will be more highly connected than a smaller genome because it has more genes to share (Dagan and Martin 2007). Plasmids tend to have higher centrality than phages (Halary *et al.* 2010), indicating that HGT is more frequently mediated by conjugation than transduction. Lack of connections between phages and plasmids, suggested that the two vehicles rarely carry the same genes (Halary *et al.* 2010).

Community structures on gene-sharing networks correspond to taxonomic classification of the connected species (Dagan 2011). Genes tend to be shared more within species than between (Popa *et al.* 2011). So communities on a network will often correspond to clades on a tree. However, communities of distant relatives indicate high frequency of HGT between species living in the same environment or affected by the same phage. Gene-sharing networks then, can reveal relationships that remain unreported by branching structures. In a study of gene-sharing networks, Kloseges at al. (2011) found a split between the Alpha-, Delta- and Epsilon-Proteobacteria and the Beta- and Gamma- Proteobacteria, that is yet undetected by phylogenetic methods.

The goal of this analysis is solely to look at the way in which a group of organisms that would once have been in a single species cloud have diverged from one another. We wished to explore how smoothly this process occurs and the kinds of genes that are last to diverge. By using gene-sharing networks we can gain an understanding of the evolutionary history of a group of bacterial species at a level that includes all aspects of this history.

We start with a gene-sharing network on which each node is representative of a genome and the edges are a statement of homology between any pair of genomes that are connected. If two genomes have at least one homologous gene in common then they will be connected by an edge. In this case we think of the genomes as genetic entities and the goods they are sharing are genes. The genomes in this analysis are sampled from the YESS group, a group of medically and scientifically important bacteria including *Yersinia*, *Escherichia*, *Shigella* and *Salmonella*. Also included in the dataset are genomes from the *Pectobacteria*, primarily a plant pathogen and opportunistic human pathogen. However, it has been shown that plant-associated

enterobacteria share a high proportion of their genome with human pathogenic strains and have even been know to jump across kingdoms (Holden *et al.* 2009). This group of closely related genomes often shares the same environment of the lower intestinal tract of mammals. It has been shown that there is a high rate horizontal transfer of genes between them.

Homologous recombination in bacteria is analogous to sex in eukaryotes. It generates genetic diversity and improves the response of bacterial populations to natural selection (Vos 2009). As a prerequisite to homologous recombination the donor DNA must have regions of high similarity to the recipient genome. Calculations show that genomes will share adequate regions of high similarity for up to 30% divergence (Townsend *et al.* 2003). It is accepted that when DNA similarity levels between two strains are greater than 70% they can be assigned to the same species (Achtman, 2008, Cho, 2001, Konstantinidis, 2006, Stackebrandt, 1994, Staley, 2006). It has also been shown that there is a sharp decline in recombination frequencies as sequence divergence increases (Majewski and Cohan 1999; Majewski *et al.* 2000). As a consequence genes can speciate at different times, some early whereas others, that are still recombining, diverge much later. By assessing levels of sequence similarity we can show which genes have the least divergence i.e. those that are still recombining. High levels of sequence similarity between two genomes that might otherwise show moderate levels of divergence might be explained by HGT (Kloesges *et al.* 2011). Another explanation is strong conservation of ancestral sequences. However, given what we know about inter-species gene transfer, plasmids (Hayes 1953), phages (Duckworth 1976), Gene Transfer Agents (GTAs) (Lang and Beatty 2000), etc., HGT is a strong candidate for these regions.

To discover which genomes are still capable of recombination with one another we remove relationships between genes that have less than 70% sequence similarity from our initial gene-sharing network. At the point where only very similar sequences are included in the network, is there an uneven distribution of edges? In other words are some genomes sharing genes up to a much higher similarity threshold than others? We further raise the threshold in order to understand at what level of similarity the genomes in this group stop sharing genes. At different levels of similarity we would expect to see different pairs genomes loosing the edge between them. Therefore highly similar genomes will remain highly connected. We find highly connected modules on our networks and determine whether these correspond to species classification.

As a genome looses more connections its centrality in the network will be reduced. We can therefore use measures of centrality to reveal how divergent these species are. In this analysis we rely on measures of degree centrality, defined as the fraction of nodes in the network that are incident upon a node, closeness, calculated as 1 over the sum of its distances to all other nodes and betweenness, equal to the number of shortest paths from all vertices to all others that pass through that node. When the network is maximally connected, i.e. when all genomes are sharing at least one gene with all other genomes, all nodes have equal centrality. As the number of edges on a network decreases some nodes become less connected and the measure of their degree centrality and closeness centralities will be smaller. In some cases as the number of connections to a node drop off, the measure of betweenness will also decrease. However there are instances where a node of low degree centrality and closeness can be found on a path that connects two disjoint parts of the network and therefore have a higher betweenness than perhaps expected. A genome that has low

degree centrality and closeness centralities by comparison to the rest of the network has lost more edges than other genomes in the network. We can assume that this genome has diverged more than the others.

Finally we assess the kinds of genes that are last to diverge by creating a network of homologous gene relationships. This time the nodes or genetic entities are genes and the good they share is a segment of DNA, if two genes are sharing they are connected by an edge. By determining which genes remain at each level of similarity we can discover what is last to diverge in our dataset.

4.2: Materials and Methods

4.2.1: Data

A dataset of 66 completed γ-proteobacteria genomes was retrieved from NCBI (web link 8). This included 29 *Escherichia*, 16 *Salmonella*, 7 *Shigella*, 2 *Pectobacterium* and 12 *Yersinia* (Table A2, Appendix). A database containing all 320,395 DNA gene sequences from this dataset was created.

The database was split into 100 input files each containing roughly the same number of sequences using fastaspltn (web link 9). Homologs were identified using an all vs. all BLASTN of the 320,395 nucleotide sequences with an e-value threshold of e^{-6}.

For this study we examined the relationships between whole genes. The BLAST output was split into two; the genes that have similarity across at least 80% of the query sequence and those that had similarity across less than 80% of the query sequence, i.e. genes with partial sequence homology. Of the 16,468,419 pairs of homologs, ~40% represented partial homology.

The percentage similarity for each pair of homologs was found using the Smith-Waterman algorithm implemented in SSEARCH (part of the FASTA package, web link 10). Smith-Waterman is a dynamic programming algorithm; it is guaranteed to find the optimal local alignment with respect to the scoring system provided. Although slower than BLAST; it takes O(mn) time to align two sequences, m and n, it is more rigorous. BLAST reduces the time required by identifying conserved regions using rapid lookup strategies, at the cost of exactness.

4.2.2: Network of Genomes

Cytoscape (Shannon *et al.* 2003) was used to visualize the data, producing a network representation of the 66 genomes (nodes) and their homology relationships with one another (edges). Each pair of homologs was scored based on their similarity, so the strength of a connection between two genomes is the sum of the similarity scores for all the genes they have in common. The strength score was normalized across the data by dividing by all strength scores by the highest strength score. Every pair of genomes now has a score between 0 and 1 (0 implies no genes in common between the two genomes: these will not be connected by an edge on the network, 1 is the score between the two genomes with the most genes in common). On the network the edges are coloured according to the strength of connection. Genomes sharing the most genes i.e. with a score of 1, will be connected by a blue edge, the edges will progressively get darker as the strength of connection decreases. Genomes connected by darker edges have weaker connections.

4.2.3: Filtered Networks

In order to examine the genes still capable of recombination a similarity threshold was set. In other words only pairs of genes with 70% or more sequence similarity were included in the data and visualized as a network. To elucidate which genes were diverging at slower rates the threshold was raised and the network was analysed at various levels of sequence identity. Networks were visualised for genes with 90% or more similarity, 95% or more and finally for genes that are 100%

identical across at least 80% of the query gene (from the "removing partial homologs" step).

At each level of similarity we quantify the number of edges lost in removing gene relationships below the given threshold. On the initial, maximally connected network the number of edges projecting from each genome (herein referred to as 'outgoing edges') is 66. Therefore for a given genus, the maximum number of outgoing edges is the number of genomes in that genus multiplied by 66. We find the percentage of the maximum overall number of outgoing edges that can be accounted for by each genus. In other words, for 66 genomes there can be a maximum possible 4,356 outgoing edges (66 x 66). Of these 4,356 edges what percentage are accounted for by *Escherichia* genomes, what percentage are accounted for by *Pectobacteria* genomes, etc. As we increase the percentage sequence similarity we re-examine the percentage of outgoing edges from each genus. At each level of similarity we find the total number of outgoing edges remaining. The expected number of outgoing edges for a given genus at a given threshold should make up the same percentage of the overall number of outward edges as it did for the initial network. If a particular genus of genomes has fewer outgoing edges in the network then it has lost more relationships than other genomes in the network, i.e. it is more divergent than others in the network or it has lost more genes than expected.

4.2.4: Network Community Detection

To find communities or clusters in the networks of genomes we use NeMo (Rivera *et al.* 2010), a cytoscape plugin for unweighted network clustering (available

for download at web link 11). The method is based on a score that estimates the likelihood that a pair of nodes has more common neighbours than expected by chance (Rivera *et al.* 2010).

4.2.5: Measures of Centrality

Measures of centrality were computed using the NetworkX python package (Hagberg *et al.* 2008). The degree centrality of a node *v* is calculated as the fraction of nodes to which it is connected. Degree centrality values are normalized by dividing the number nodes to which a given node is connected by the maximum possible degree for that network (n-1 where n is the number of nodes in the network G). For graphs with self-loops, like our network of genomes, the maximum degree might be higher than n-1 and values of degree centrality greater than 1 are possible.

The closeness centrality of a node is calculated as 1 divided by the average distance to all other nodes. Closeness centrality is normalized to the maximum possible degree divided by the size of the network (n-1 /s-1 where n is the number of nodes in the connected part of graph containing the node and s is the size of the graph in number of nodes). If the graph is not completely connected, i.e. there are disjoint parts of the network, this algorithm computes the closeness centrality for each connected part separately.

Betweenness centrality of a node is calculated as the sum of the fraction of all-pairs shortest paths that pass through said node. The NetworkX package makes use of the algorithm by Ulrik Brandes, to compute the betweenness values of nodes on a

graph (Brandes 2001). If we wish to pass through the network from one given node to another we are likely to pass through a node of high betweenness. The node with high betweenness may have many or few connections and is likely to lie between two parts of the network. These two parts may be highly connected within but they have few edges between them so the node with high betweenness acts as a bridge between the parts.

4.2.6: Network of Genes

The network of genes was constructed from the same data as the original network of genomes. Instead of summing the pairs of homologs between each pair of genomes, this time we treated every pair of homologs as a separate relationship. The network of genes, therefore, consists of 320,395 nodes, each representing a gene and edges that indicate a homologous relationship between to two genes it connects.

A network of 320,395 genes is too large to visualize in Cytoscape. The network of genes was analysed through stats and parts of the network containing unexpected information were visualized in Cytoscape.

4.2.7: Kinds of Genes that are Last to Diverge

To assess the kinds of genes that are still homologous at each level of divergence we first considered the COG categories of all genes remaining at each similarity threshold previously specified for the genome networks. A database of COG categorized genes was obtained from ftp://ftp.ncbi.nih.gov/pub/COG/COG/.

Every gene in the dataset was searched using BLAST against the COG database in order to obtain the gene's best suited COG category. There were 22 categories overall and a 23rd to account for any genes that did not have a homolog in the COG database.

Initially we determined the number of genes found in each COG category for the entire database and reported this number as a percentage of the overall number of genes in the dataset. For each similarity threshold we established how many genes from each COG category remained, this is our observed number of genes. The expected number genes in each COG category, at each level of similarity, was the number of genes remaining at that level that would be expected to fall into this category based on the percentage of genes in this category for the overall dataset. We test for significantly higher or lower numbers of genes than expected, using the Chi-squared test.

We also assessed the GenBank functions at each level of similarity. For each gene in the dataset its function was parsed from the full GenBank file. There are a vast variety of functional annotations in GenBank, in this analysis we focus on those that are dominant, i.e. functions that appear more than any other. We assess the top 25 occurring GenBank functions at each similarity threshold.

4.2.8: Percentage Similarity of Homologous Genes

It is convenient to group pairs of homologous genes with similar percentage sequence similarity together. In this study we divide all pairs of homologous genes into a specified number of partitions, or bins. Pairs are placed into bins depending on their percentage sequence similarity. The lowest sequence similarity is 55% and the

highest is 100%, bins are constructed by splitting the pairs into equal partitions. We use 20 bins, each with a range of 2.25% sequence similarity.

When the homologous pairs are divided into bins we can we can determine the average rate of divergence between homologs in this dataset. Over time we would expect to see homologs diverge to an extent that distinguishes them as belonging to different species. If we see a swell in the number of homologs at a given bin or range of bins we can assume that this is the expected amount of sequence similarity for homologs in this dataset. Homologs that are less similar than this have diverged rapidly and homologs that are more similar are strongly conserved or recently obtained through HGT. A gene obtained recently through HGT has not had sufficient time to diverge from its homolog.

To further gain insight into gene sharing, we examine the rate of divergence within each genus and between each pair of genera. A large number of homologous pairs at the higher levels of percentage similarity, i.e. close to 100%, for a two different genera should indicate a substantial amount of recent gene sharing.

4.3 Results

4.3.1 Network of Genomes

The network of 66 genomes, connected by genes with 70% similarity or more has 4,356 edges, is maximally connected (66 x 66). In other words every genome in the network has at least one gene that is at least 70% similar to at least one homologous gene in every other genome. At the 70% similarity threshold, all genomes have the maximum number of outgoing edges.

The strongest edges (darkest edges on Figure 4.2) are found between genomes from the same genus. In particular *Shigella* and *Yersinia* appear to have strong within-genus connections. Modules detected by NEMO on the 70% similarity network are indeterminate in terms of taxonomic classification. Each module contains genomes from all genera; there is no phylogenetic distinction between modules at this level.

At this level of similarity, the measures of centrality are non-informative. If a network is maximally connected then all nodes are connected to all other nodes, therefore they all have the same measure of degree centrality. Similarly, the measures for closeness and betweenness will be the same for every node.

When there are no genes being shared between two genomes above the given threshold an edge does not exist between the two. We find that up to 90% sequence identity the network is maximally connected. This means that every genome has at least one homologous gene in every other genome that is at least 90% similar.

Between 90 and 95%, however the number of edges starts to decrease. We find that, to an extent, the edges are being lost in correlation with the relationships

found in phylogenetic studies. According to ribosomal RNA gene tree phylogenetics, the *Pectobacteria* is the most distant from the group followed by the *Yersinia*. We see the edges connecting the *Pectobacteria* and *Yersinia* to the rest of the group disappear first as the similarity threshold is raised (Figures 4.4, 4.9, 4.14 and 4.19).

At higher percentages of sequence similarity we begin to see a correlation between modules found by the NeMo software and the classification of species through phylogenetic studies. The *Pectobacteria*, being the most dissimilar to the group can quickly be seen to from a separate community from the rest of the network. Following this begin to see that there is a clear boundary between the *Yersinia* and the rest of the network. However, all the way to the 100% similarity threshold, the software cannot distinguish between the three genera *Escherichia*, *Salmonella* and *Shigella*.

As we increase the similarity threshold to 95% and beyond we see patterns in the centrality values for genomes on the network. The *Escherichia*, *Salmonella* and *Shigella* remain highly connected to one another, displaying high values of degree centrality and closeness. The *Yersinia*, being more divergent, and thus having fewer outgoing edges, has much lower values for degree centrality and closeness. The Pectobacteria genomes, surprisingly, begin at opposite ends of the degree centrality and closeness value distribution. They do however become more consistent with one another as the threshold is raised. The betweenness values are the most interesting as they reveal the nodes at each level of similarity that are still connected to parts of the network that are otherwise separating from one another. The genomes with highest betweenness values vary as the similarity threshold increases.

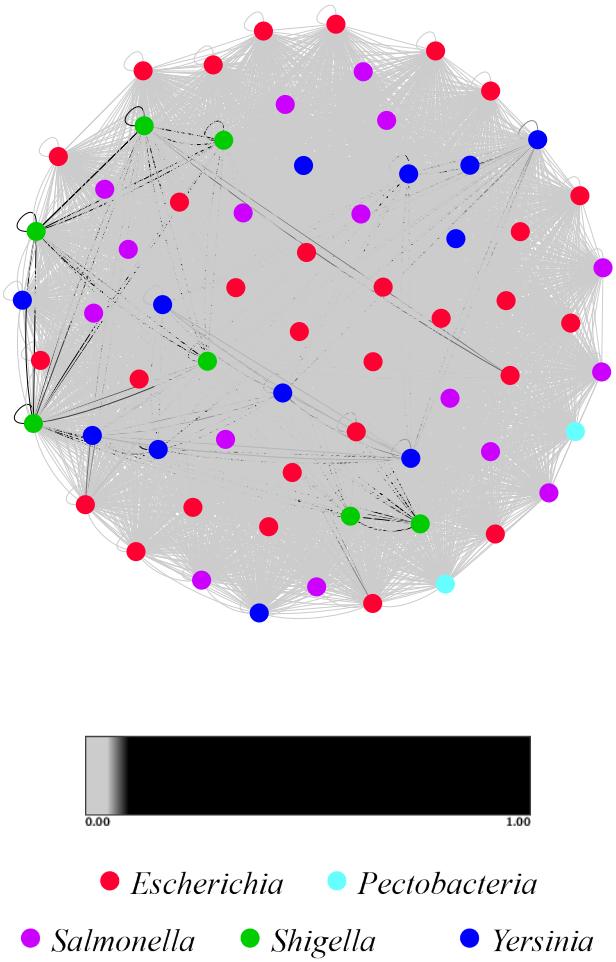


Figure 4.2: Networks of genomes at 90% similarity threshold. Lighter edges are weaker by comparison (very close to 0 on the scale bar). The darker edges are the strongest in the network (closer to 1 on the scale bar).

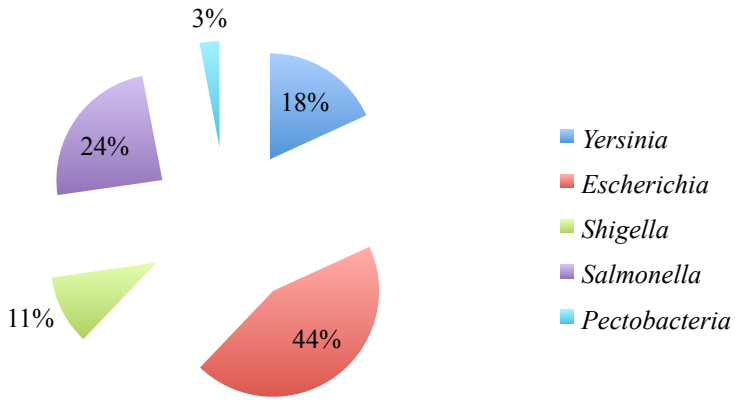


Figure 4.3: Pie chart of the percentages of overall outgoing edges represented by each genus for the network built from genes that have at least 90% sequence similarity.

Table 4.1: The number of outgoing edges and corresponding percentages for each genus in the network built from genes that have at least 90% sequence similarity.

Genus	**No. Outgoing Connections**	**Percentage of All Connections**
Yersinia	792	18.18%
Escherichia	1914	43.93%
Shigella	462	10.60%
Salmonella	1056	24.24%
Pectobacteria	132	3.0%
	4356	100

4.3.1.1: Network of genomes with 95% Similarity Threshold

At the 95% similarity threshold the network begins to disconnect (Figure 4.2). The network is no longer maximally connected, for 66 nodes there are now 3,681 edges (Table 4.2). Different genera begin to form distinct groups as the between-genus connections begin to disappear. The *Pectobacteria* have suffered a noteworthy loss of between-genus edges, i.e. there is a drop in the number of edges between the *Pectobacteria* genomes and the rest of the network. The *Pectobacteria* genomes are no longer sharing genes with many of the *Escherichia*, *Shigella* and *Salmonella* genomes. However, there has been no loss of edges within the *Pectobacteria*, nor has there been a loss of edges between the *Pectobacteria* and the *Yersinia*.

The *Yersinia* genus is starting to separate from the rest of the network, there are far fewer edges between *Yersinia* and the rest of the genomes and those remaining edges have low weights. The edges within the *Yersinia* module, however, have much higher weights by comparison, i.e. there are far more genes being shared within the *Yersinia* than between *Yersinia* and other genera.

The *Escherichia, Shigella* and *Salmonella* have remained maximally connected up to this level of similarity. Every *Escherichia, Shigella* and *Salmonella* genome has at least one gene with 90% sequence similarity to at least one gene in every other *Escherichia, Shigella* and *Salmonella* genome. These results are in line with phylogenetic studies in that the *Pectobacteria* and the *Yersinia* are separating from the group first, suggesting that they are the most divergent. However it is important to note the massive amount of sharing that is still occurring at this high level of similarity. It took raising the threshold to 95% for us to begin to see this phylogenetic signal.

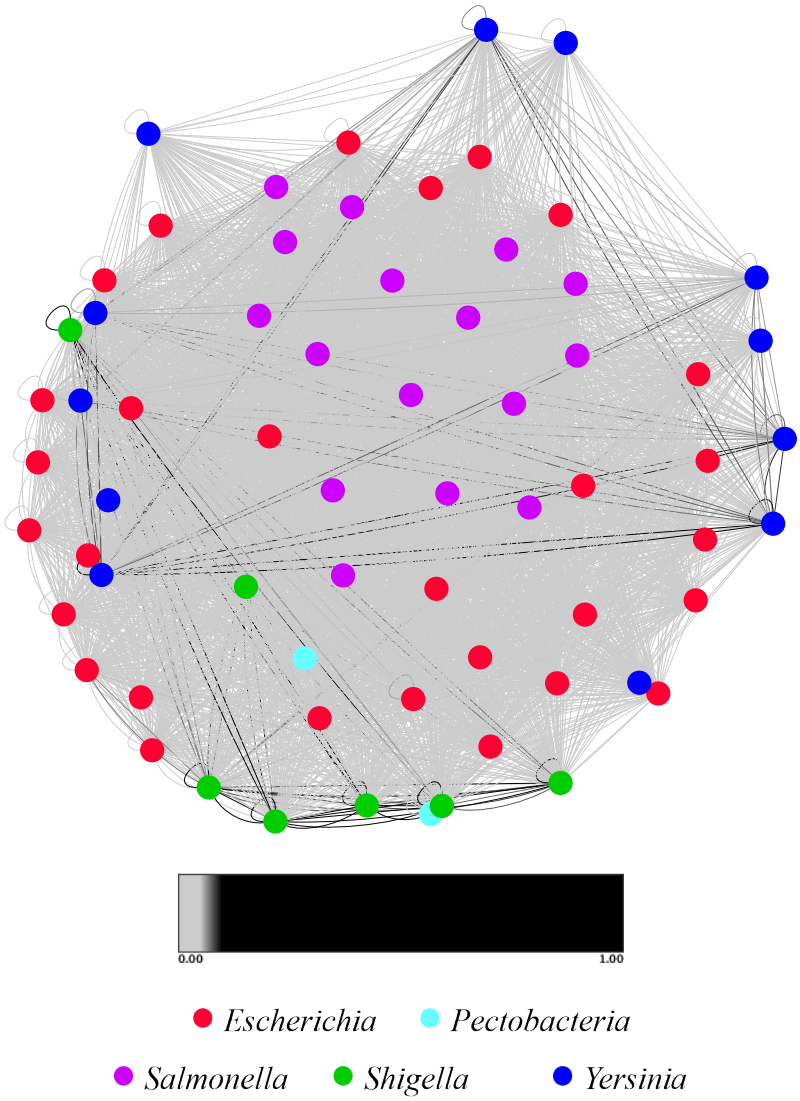


Figure 4.4: Network of genomes with 95% similarity threshold. Lighter edges are weaker by comparison (very close to 0 on the scale bar). The darker edges are the strongest in the network (closer to 1 on the scale bar).

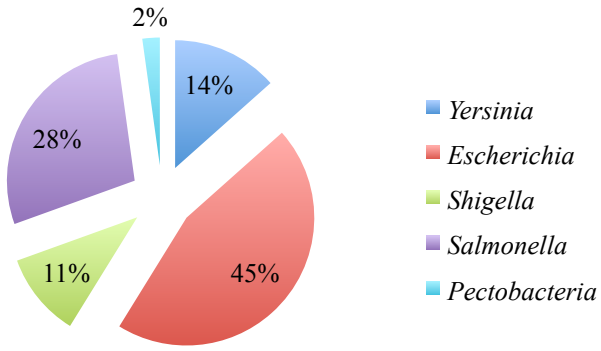


Figure 4.5: Pie chart of the percentages of overall outgoing edges represented by each genus for the network built from genes that have at least 95% sequence similarity.

Table 4.2: The number of outgoing edges and corresponding percentages for each genus in the network built from genes that have at least 95% sequence similarity. The number of outgoing connections is calculated from the initial percentages indicated by the maximally connected network. The Chi-squared test provides significance scores. For genera that have less outgoing edges than expected the P value is highlighted in orange for those with more than expected outgoing edges the P value is highlighted in black.

Genus	**No. Outgoing Connections**	**Percentage of All Connections**	**Expected No. Outgoing Connections**	**P value**
Yersinia	494	13.42026623	669.2727273	1.2436E-11
Escherichia	1671	45.39527302	1617.409091	0.182681937
Shigella	393	10.67644662	390.4090909	0.895674822
Salmonella	1042	28.30752513	892.3636364	5.46638E-07
Pectobacteria	81	2.200488998	111.5454545	0.003826165
	3681	100	3681	

Table 4.3: Number of outgoing edges from the source genera labeled on the left to the target genera labeled on top for the network built from genes that have at least 95% sequence similarity. Cells highlighted in yellow represent the maximum number of outgoing edges a given genus can have to towards the target genus, i.e. the genomes in this genera are maximally connected.

	Yersinia	***Escherichia***	***Shigella***	***Salmonella***	***Pectobacteria***	**Total**
Yersinia	144	132	2	192	24	494
Escherichia	130	841	203	464	33	1671
Shigella	22	203	49	112	7	393
Salmonella	192	464	112	256	18	1042
Pectobacteria	24	29	7	17	4	81
						3681

4.3.1.2: Modules on the Network of genomes with 95% Similarity Threshold

The modules found by NeMo at 95% similarity reveal some phylogenetic signal. There is a hint of genomes from the same species grouping together. Module 1, the highest scoring module (from Table 4.4), contains all but one of the *Yersinia* genomes and one of the *Pectobacteria*. This module is almost representative of the *Yersinia* genus. Every other module contains all 12 of the *Yersinia* genomes; clearly these genomes form a very tight-knit community on the network.

When including only pairs of genes with 95% or more similarity it could be assumed that the highly connected modules would contain highly similar genomes and thus be indicative of a species-level set of relationships. An interesting result from the NeMo analysis is that *Pectobacterium carotovorum subsp. Carotovorum* appears in none of the modules, yet *Pectobacterium carotovorum subsp. Atroseptica* appears in every module. This is because at the 95% similarity threshold, *Pectobacterium carotovorum subsp. Atroseptica* is still sharing genes with all *Yersinia* genomes on the network and so is part of the highly connected subgroup containing all the *Yersinia*. This is the kind of signal that could lead to the assumption that certain strains of bacteria could have been misclassified. Had we not known its previous classification, from the network it might be reasonable to assume that *Pectobacterium carotovorum subsp. Atroseptica* is a *Yersinia* genome.

Modules 2-8 contain at least one genome from all genera. It is hard to distinguish one species from another when so much gene sharing is occurring and at such high levels of sequence similarity.

Table 4.4: Modules according to NeMo for the network of genomes at 95% similarity threshold.

Cluster	Score (Density*#Nodes)	Nodes	Edges	Node IDs
1	-111.938	12	144	YERSPBM, YERSPPF, YERSPCO92, YERSPANT, YERSTBPB1, YERSEN8081, YERSPNEPAL, YERSTBIP31, YERSTBYP3, PECTCARATR, YERSPKIM, YERSPANG
2	-169.805	23	269	ECOREL606, SHIGDYS, ECOTW4359, YERSTB32, YERSPBM, YERSPPF, YERSPCO92, YERSPANT, YERSTBPB1, YERSEN8081, YERSPNEPAL, YERSTBIP31, YERSTBYP3, PECTCARATR, YERSPKIM, YERSPANG, ECOCATCC, ECONEW, ECOKMG1655, SHIGBOYDCDC, ECOEDL933, ECOSAKAI, ECOEC4115
3	-169.805	18	194	SHIGDYS, ECOTW4359, YERSTB32, YERSPBM, YERSPPF, YERSPCO92, YERSPANT, YERSTBPB1, YERSEN8081, YERSPNEPAL, YERSTBIP31, YERSTBYP3, PECTCARATR, YERSPKIM, YERSPANG, ECOCATCC, ECONEW, ECOKMG1655
4	-169.805	26	338	ECOREL606, SHIGDYS, ECOTW4359, YERSTB32, YERSPBM, YERSPPF, YERSPCO92, YERSPANT, YERSTBPB1, YERSEN8081, YERSPNEPAL, YERSTBIP31, YERSTBYP3, PECTCARATR, YERSPKIM, YERSPANG, ECOCATCC, ECONEW, ECOKMG1655, SHIGBOYDCDC, ECOEDL933, ECOSAKAI, ECOEC4115, ECOSE11, ECOKDH10B, ECOHS
5	-169.805	20	218	SHIGDYS, ECOTW4359, YERSTB32, YERSPBM, YERSPPF, YERSPCO92, YERSPANT, YERSTBPB1, YERSEN8081, YERSPNEPAL, YERSTBIP31, YERSTBYP3, PECTCARATR, YERSPKIM, YERSPANG, ECOCATCC, ECONEW, ECOKMG1655, SHIGBOYDCDC, ECOEDL933
6	-243.101	42	1153	ECOUMN026, ECOS88, ECOE24377A, ECOH6E2348, ECOLF82, ECOAPEC01, SHIGF301, SHIGFLEX5,

				SHIGF245, ECOSMS35, ECOBL21DE3, SHIGSON, ECOIAI1, ECOBW2952, ECOKW3110, SHIGBOY227, ECOREL606, SHIGDYS, ECOTW4359, YERSTB32, YERSPBM, YERSPPF, YERSPCO92, YERSPANT, YERSTBPB1, YERSEN8081, YERSPNEPAL, YERSTBIP31, YERSTBYP3, PECTCARATR, YERSPKIM, YERSPANG, ECOCATCC, ECONEW, ECOKMG1655, SHIGBOYDCDC, ECOEDL933, ECOSAKAI, ECOEC4115, ECOSE11, ECOKDH10B, ECOHS
7	-263.58	56	2476	SALENSCHW, SALTYPHI, SALENSTY, ECO55989, ECOED1A, ECOIAI39, ECO536, ECOUTI89, ECOCFT073, ECOUMN026, ECOS88, ECOE24377A, ECOH6E2348, ECOLF82, ECOAPEC01, SHIGF301, SHIGFLEX5, SHIGF245, ECOSMS35, ECOBL21DE3, SHIGSON, ECOIAI1, ECOBW2952, ECOKW3110, SHIGBOY227, ECOREL606, SHIGDYS, ECOTW4359, YERSTB32, YERSPBM, YERSPPF, YERSPCO92, YERSPANT, YERSTBPB1, YERSEN8081, YERSPNEPAL, YERSTBIP31, YERSTBYP3, PECTCARATR, YERSPKIM, YERSPANG, ECOCATCC, ECONEW, ECOKMG1655, SHIGBOYDCDC, ECOEDL933, ECOSAKAI, ECOEC4115, ECOSE11, ECOKDH10B, ECOHS, SALENCHOL, SALTYLT2, SALENGAL, SALENPARAA, SALENENTER
8	-263.58	54	2260	SALENSCHW, SALTYPHI, SALENSTY, ECO55989, ECOED1A, ECOIAI39, ECO536, ECOUTI89, ECOCFT073, ECOUMN026, ECOS88, ECOE24377A, ECOH6E2348, ECOLF82, ECOAPEC01, SHIGF301, SHIGFLEX5, SHIGF245, ECOSMS35, ECOBL21DE3, SHIGSON, ECOIAI1, ECOBW2952, ECOKW3110, SHIGBOY227, ECOREL606, SHIGDYS, ECOTW4359, YERSTB32, YERSPBM, YERSPPF, YERSPCO92, YERSPANT, YERSTBPB1, YERSEN8081, YERSPNEPAL, YERSTBIP31, YERSTBYP3, PECTCARATR, YERSPKIM, YERSPANG, ECOCATCC, ECONEW, ECOKMG1655, SHIGBOYDCDC, ECOEDL933, ECOSAKAI, ECOEC4115, ECOSE11, ECOKDH10B, ECOHS, SALENCHOL, SALTYLT2, SALENGAL

4.3.1.3: Centrality Measures of Nodes on the Network of genomes with 95% Similarity Threshold

At the 95% similarity threshold the highest degree centrality value is 1.03 for the genome of *S. enterica subsp. Arizonae*. This genome has more outgoing edges than any other on this network but the largest genome is *S. enterica subsp. enterica serovar Paratyphi B*. With 5,592 genes, this genome is surprisingly not sharing genes with as many different genomes as some others are at this level of similarity.

The values for all *Escherichia, Salmonella* and *Shigella* range down to the lowest value of 0.83. The *Escherichia, Salmonella* and *Shigella* remain completely connected to one another. For these three genera there is little if any relationship between the length of the genome and its measure of degree centrality (Figure 4.6).
It can be seen in the distribution of degree centrality values for the *Yersinia* genomes, have comparably lower values. All values for the *Yersinia* genomes fall beneath the lowest value for the *Escherichia*, *Salmonella* and *Shigella*. This means that the *Yersinia* genomes are far less connected in the network and thus sharing with far fewer genomes than any of the *Escherichia*, *Salmonella* or *Shigella*. Interesting to note is the fact that all the *Yersinia* genomes fall to the lower end of the scale in terms of genome size ranging from 3,832 genes to 4,192 genes. It is possible to interpret this result in one of two ways. Either the genes in *Yersinia* are more rapidly evolving on average, the *Yersinia* have been losing genes and therefore, by chance, they have fewer genes that might help in keeping the cluster together or *Yersinia* might be an older group and consequently.

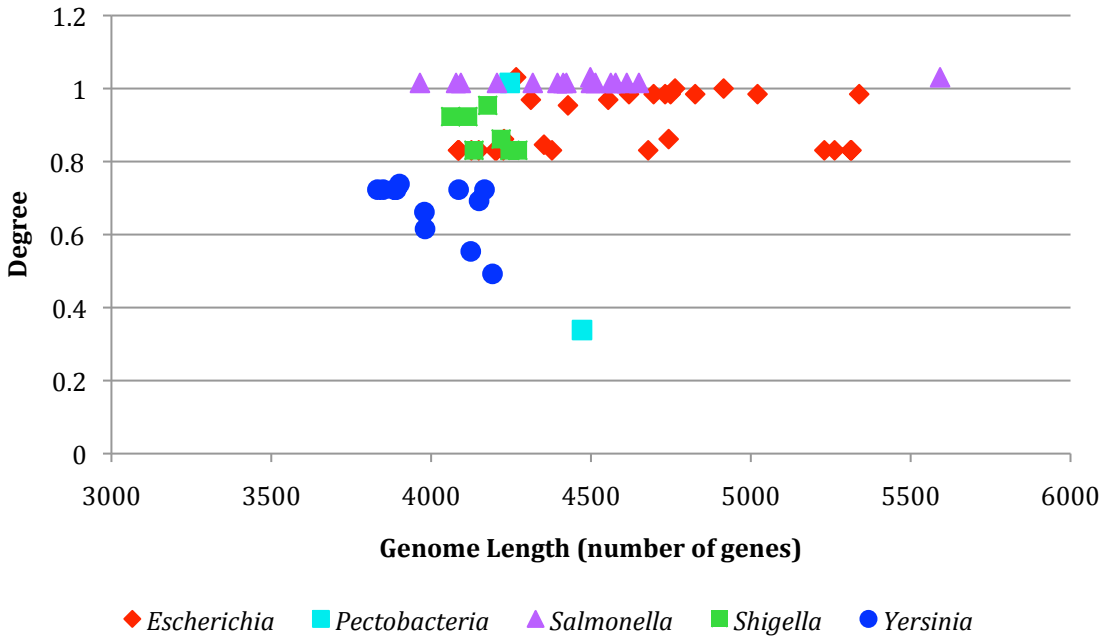


Figure 4.6: Degree centrality against genome length for the network of genomes with 95% similarity threshold.

The two *Pectobacteria* genomes tell two very different stories. While the degree centrality value for the genome for *Pectobacterium carotovorum subsp. carotovorum* falls amongst the highest, the value for *Pectobacterium carotovorum subsp. atroseptica* is by far the lowest at 0.34. Perhaps most surprising is that *Pectobacterium carotovorum subsp. atroseptica,* is in fact the larger of the two genomes, with close to 200 more genes than *P. carotovorum subsp. carotovorum.*

The values for closeness centrality at the 95% similarity threshold tell a very similar story to that told by the degree centrality values. Closeness values for this network range between 0.59 and 1. Again, the genome with the highest closeness value is *S. enterica subsp. Arizonae,* despite not being the largest genome. The closeness values for the *Escherichia, Salmonella* and *Shigella* fall not far below the maximum and there is little correlation between these values and the size of the genomes.

The *Yersinia* closeness values range between 0.65 and 0.76, placing them below the rest of the group, perhaps accounted for by their smaller genome sizes.

Finally, as before, the genome for *Pectobacteria carotovorum subsp. carotovorum* has a high closeness value while the other *Pectobacteria* genome, *P. carotovorum subsp. atroseptica* has the lowest closeness value for this network.

The distribution of values for betweenness centrality for genomes on this network has similarities and differences to the distribution of values for degree centrality and closeness. For the third time, the genome with the highest betweenness value (a value of 0.006) is *S. enterica subsp. Arizonae.* Not only is this genome highly connected in the network but it also falls on most paths between two

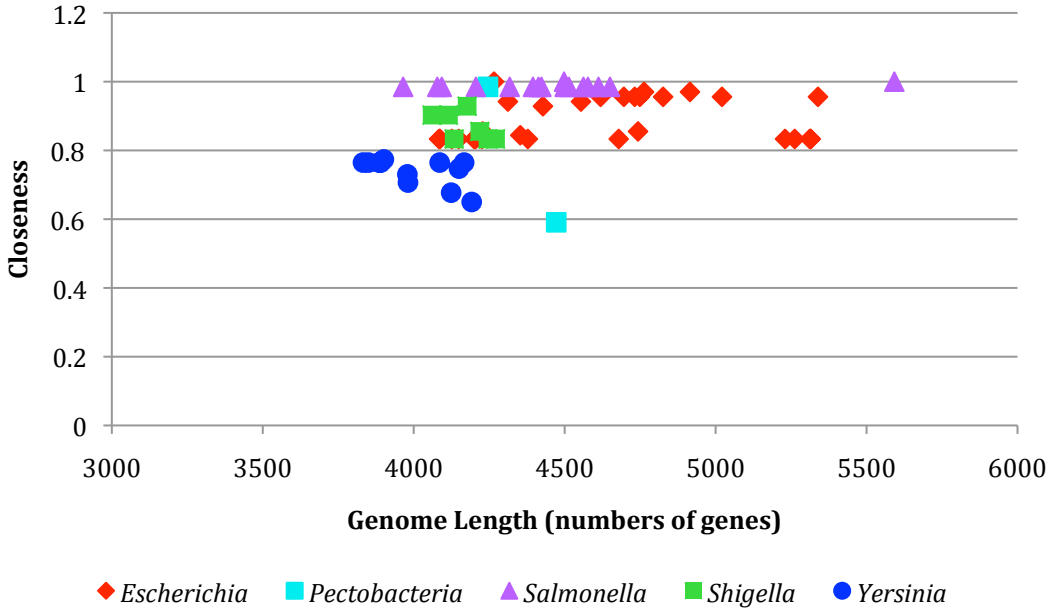


Figure 4.7: Closeness centrality against genome length for the network of genomes with 95% similarity threshold.

given nodes. At this point the network is beginning to separate, there is less gene sharing between certain genomes. Genomes with high betweenness are still sharing genes with genomes from all parts of the network, so *S. enterica subsp. Arizonae* is still sharing with the *Yersinia* genomes even as they loose connections with the rest of the network. In other words a path between a *Yersinia* genome and a genome from the *Escherichia*, *Salmonella* or *Shigella* is likely to contain the node for *S. enterica subsp. Arizonae.* Many of the *Escherichia*, *Salmonella* and *Shigella* have betweenness values not far below the highest value. However, in contrast to the results for degree centrality and closeness we also find a number of *Escherichia* and *Shigella* genomes have very low betweenness values compared to the rest of the genomes in the network. The genomes with low betweenness are likely to have lost connections with the *Yersinia* that are moving away from the rest of the network.

As was seen for degree centrality and closeness, the *Yersinia* genomes all have relatively low values for betweenness centrality.

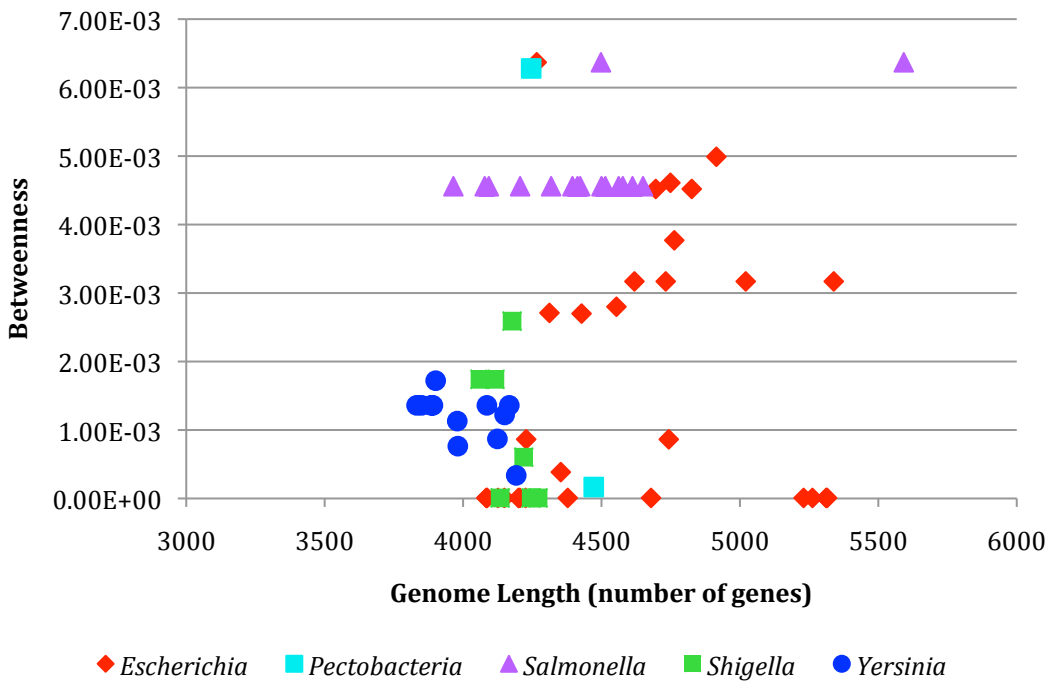


Figure 4.8: Betweenness centrality against genome length for the network of genomes with 95% similarity threshold.

4.3.1.4: Network of genomes with 98% Similarity Threshold

At the 98% similarity threshold the number of outgoing edges has fallen by another 529. Again the *Yersinia* and *Pectobacteria* appear to be underrepresented, an indicator of their divergence from the rest of the group. Between 95 and 98% similarity the number of outgoing edges for these two genera have fallen substantially compared with the numbers for *Escherichia*, *Shigella* and *Salmonella*. At this point all of the latter three genera have significantly more outgoing edges than expected. At 98% similarity the *Pectobacteria* no longer have any outgoing edges to genomes from other species. However there is a connection from the *Escherichia* genome *E. coli E24377* to the *Pectobacteria* genome *P. carotovorum subsp. atroseptica* . This is indicative of a gene (or genes) in the *Escherichia* genome that has an area covering 80% of its length that is homologous with at least 98% sequence identity to at least one gene in the *Pectobacteria* genome.

The *Yersinia* genomes have very few edges left that connect them to *Salmonella* or *Shigella* genomes. They remain connected to the rest of the network mostly through relationships with *Escherichia* genomes. We expect the *Yersinia* to be most divergent in the YESS group and so least connected to the rest of the group on the network. However, given the results from phylogenetic studies (Haggerty *et al.* 2009) we would also expect the *Yersinia* to be sharing equally with the *Escherichia*, *Salmonella* and *Shigella*, given that on a tree structure *Yersinia* is equidistant from all three genera. This result could be explained by the disparate taxon sampling; there are more *Escherichia* genomes than either *Salmonella* or *Shigella* genomes. On the other hand this could indicate a bias in gene sharing between *Yersinia* and *Escherichia*.

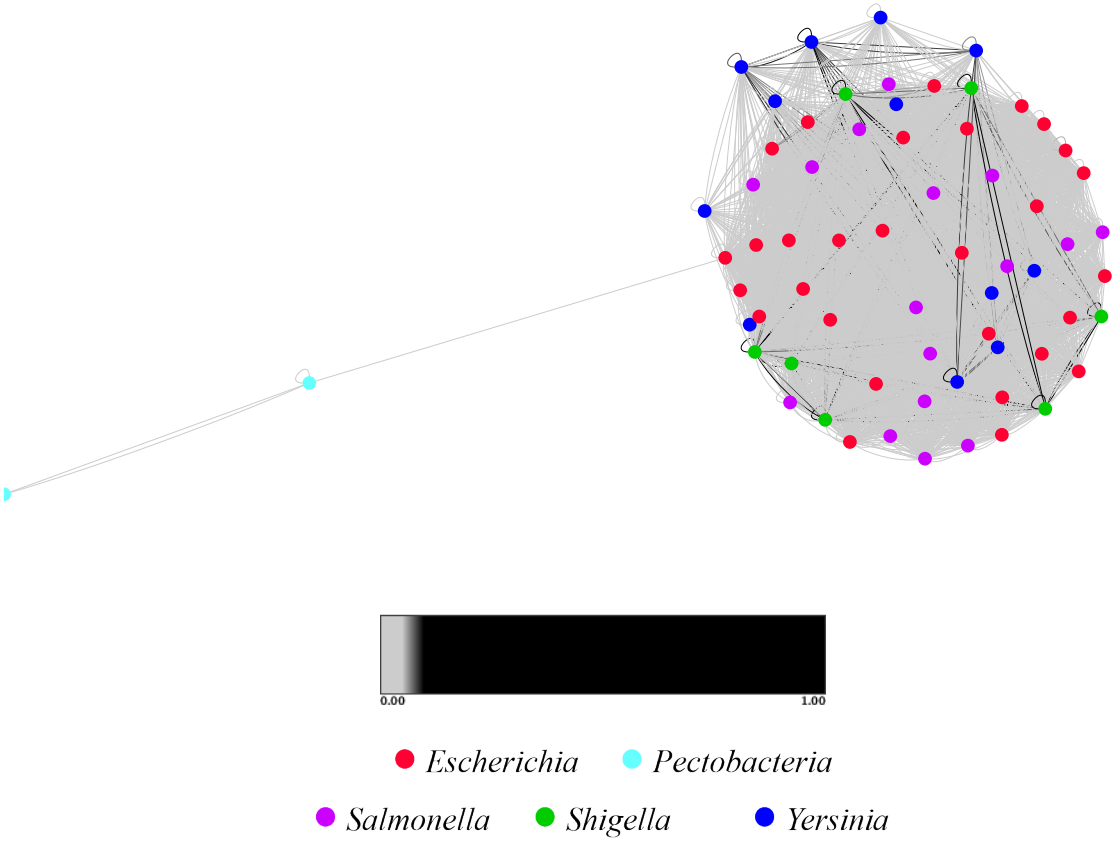


Figure 4.9: Networks of genomes at 98% similarity threshold. Lighter edges are weaker by comparison (very close to 0 on the scale bar). The darker edges are the strongest in the network (closer to 1 on the scale bar).

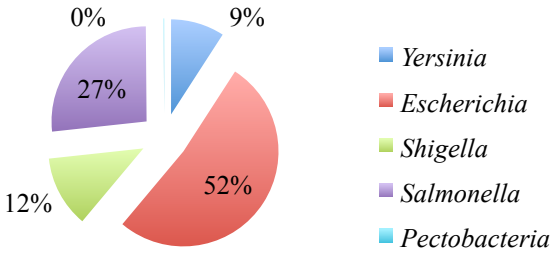


Figure 4.10: Pie chart of the percentages of overall outgoing edges represented by each genus for the network built from genes that have at least 98% sequence similarity.

Table 4.5: The number of outgoing edges and corresponding percentages for each genus in the network built from genes that have at least 98% sequence similarity. The number of outgoing connections is calculated from the initial percentages indicated by the maximally connected network. The Chi-squared test provides significance scores. For genera that have fewer outgoing edges than expected the P value is highlighted in orange for those with more than expected outgoing edges the P value is highlighted in black.

Genus	**No. Outgoing Connections**	**Percentage of All Connections**	**Expected No. Outgoing Connections**	**P value**
Yersinia	288	9.13705584	573.0909091	1.06373E-32
Escherichia	1638	51.96700508	1384.969697	1.05264E-11
Shigella	384	12.18274112	334.3030303	0.006566505
Salmonella	838	26.58629442	764.1212121	0.007525949
Pectobacteria	4	0.12690355	95.51515152	7.6845E-21
	3152	100	3152	

Table 4.6: Number of outgoing edges from the source genera labeled on the left to the target genera labeled on top for the network built from genes that have at least 98% sequence similarity. Cells highlighted in yellow represent the maximum number of outgoing edges a given genus can have to towards the target genus, i.e. the genomes in this genera are maximally connected.

	Yersinia	***Escherichia***	***Shigella***	***Salmonella***	**Pectobacteria**	**Total**
Yersinia	144	131	2	11	0	288
Escherichia	129	841	203	464	1	1638
Shigella	20	203	49	112	0	384
Salmonella	6	464	112	256	0	854
Pectobacteria	0	0	0	0	4	4
						3152

4.3.1.5: Modules on the Network of genomes with 98% Similarity Threshold

Modules 1 and 2, found by NeMo, at the 98% similarity threshold are made up exclusively of *Yersinia* genomes. The number of persisting relationships within the *Yersiniae* at this level of similarity is far more than the number of relationships between the *Yersinia* and others in the network. There is a clear correspondence, in this case, between taxonomic classification and community structure.

Subsequent modules are less conclusive. Modules 3 and 5 contain a small sample of genomes from *Escherichia*, *Salmonella* and *Shigella*. It is not surprising that these three genera would form modules, as at the 98% similarity threshold, there is still maximal connectivity within the group. There has been much debate in the past about the distinction between *Escherichia* and *Shigella* species (Escobar-Paramo, 2003, Pupo, 2000, Yang, 2007) but far less concern for the *Escherichia-Salmonella* divide. At such a high level of similarity there is still a substantial amount of sharing going on between all three of these species, casting doubt on the boundaries that have been previously established between them (Haggerty *et al.* 2009).

Finally modules 4 and 6 contain genomes from all but the *Pectobacteria*. This is indicative of the massive amount of sharing that is still occurring across the entire network, exclusive of the *Pectobacteria*, which is expected to be the most divergent genera of the group.

Table 4.7: Modules according to NeMo for the network of genomes at 98% similarity threshold.

Cluster	Score (Density*#Nodes)	Nodes	Edges	Node IDs
1	-27.998	9	81	YERSTBPB1, YERSEN8081, YERSTBYP3, YERSPNEPAL, YERSPANG, YERSPBM, YERSPPF, YERSPCO92, YERSPANT
2	-39.639	12	144	YERSTBIP31, YERSTB32, YERSPKIM, YERSTBPB1, YERSEN8081, YERSTBYP3, YERSPNEPAL, YERSPANG, YERSPBM, YERSPPF, YERSPCO92, YERSPANT
3	-173.68	10	100	ECOREL606, SHIGDYS, SALENGAL, ECOKMG1655, ECOEDL933, SALENENTER, ECONEW, ECOCATCC, ECOSAKAI, ECOEC4115
4	-173.68	39	873	ECOBW2952, ECOKW3110, SALENATCC, ECOSE11, ECOKDH10B, ECOHS, SHIGBOYDCDC, SHIGSON, SALENSTY, SALENARIZ, YERSTBIP31, YERSTB32, YERSPKIM, YERSTBPB1, YERSEN8081, YERSTBYP3, YERSPNEPAL, YERSPANG, YERSPBM, YERSPPF, YERSPCO92, YERSPANT, ECOIAI1, SALENSCHW, ECOTW4359, SALENPARAC, SHIGBOY227, SALENNEW, SALENDUB, ECOREL606, SHIGDYS, SALENGAL, ECOKMG1655, ECOEDL933, SALENENTER, ECONEW, ECOCATCC, ECOSAKAI, ECOEC4115
5	-173.68	7	49	SHIGDYS, SALENGAL, ECOKMG1655, ECOEDL933, SALENENTER, ECONEW, ECOCATCC
6	-173.68	18	180	SHIGSON, SALENSTY, SALENARIZ, YERSTBIP31, YERSTB32, YERSPKIM, YERSTBPB1, YERSEN8081, YERSTBYP3, YERSPNEPAL, YERSPANG, YERSPBM, YERSPPF, YERSPCO92, YERSPANT, ECOIAI1, SALENSCHW, ECOTW4359

4.3.1.6: Centrality Measures of Nodes on the Network of genomes with 98% Similarity Threshold

At the 98% similarity threshold the highest degree centrality value is 1, the genome with this score, and therefore the most central in the network is *E. fergusonii ATCC*. Despite not having the largest genome, with 4,266 genes, the *E. fergusonii* genome is still sharing with all but the *Pectobacteria.* There are 28 *E. coli* strains in the dataset and only one *E. fergusonii* strain, yet this one strain has more connections in the network than any of the *E. coli* strains. This result is discussed further when we evaluate the relationships between genes remaining at this similarity threshold (Section 4.3.2.3).

For the rest of the *Escherichia* genomes, along with *Salmonella* and *Shigella*, the distribution of degree centrality values is similar to the distribution at the 95% similarity threshold. They range from slightly below the highest value down to 0.82 and there is no pattern associated between the length of the genomes and their degree centrality value. The degree centrality values for *Yersinia*, again, are much smaller than those for the *Escherichia*, *Salmonella* and *Shigella,* ranging between 0.22 and 0.52. At 98% similarity, the two Pectobacteria genomes have the two lowest values for degree centrality. This is obvious when looking at the network (Figure 4.9), *P. carotovorum subsp. carotovorum* has just one outgoing edge connected to the other Pectobacteria genome, *P. carotovorum subsp. atroseptica* and there is only one connection from *P. carotovorum subsp. atroseptica* to the rest of the network, via *E. coli E24377*. It is however surprising that the genome of *P. carotovorum subsp. carotovorum* which previously had one of the highest degree centrality values, and thus was connected to a high proportion of the

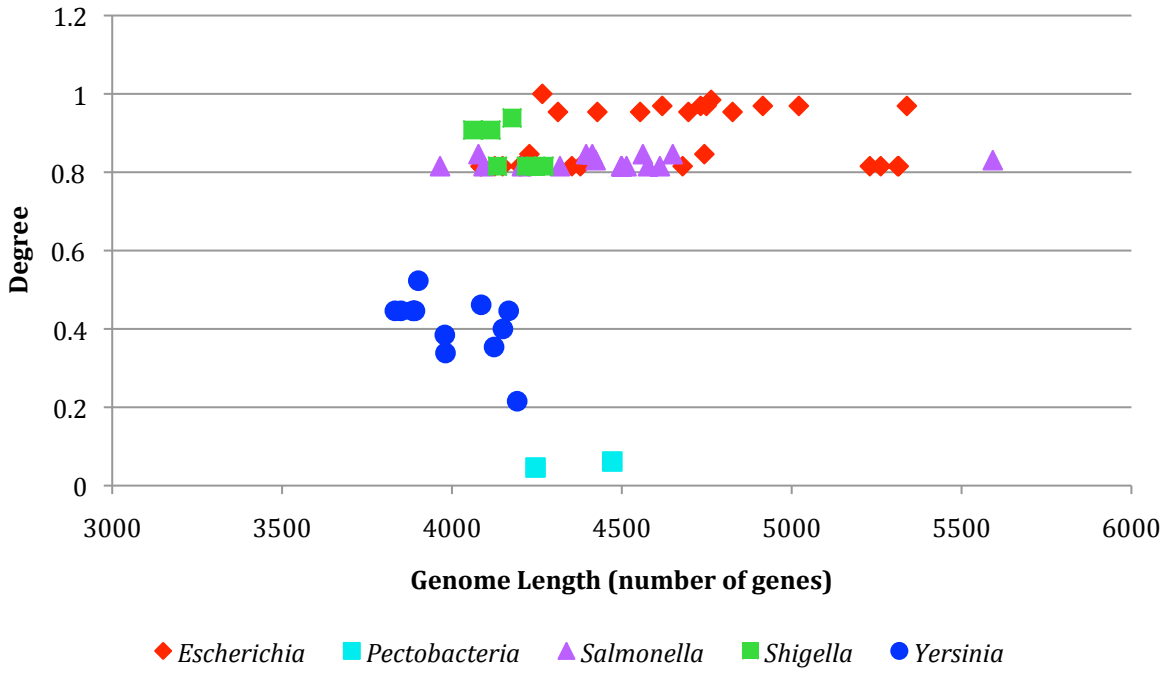


Figure 4.11: Degree centrality against genome length for the network of genomes with 98% similarity threshold.

genomes in the network, now has the lowest degree centrality value of 0.05. When we include only pairs of homologs with 98% similarity or more we find that *P. carotovorum subsp. carotovorum* is exclusively sharing with the other *Pectobacteria* genome, *P. carotovorum subsp. atroseptica* .

As was the case for the network of genomes at the 95% similarity threshold, the values for closeness centrality at 98% similarity are very similar to the values for degree centrality at 98% similarity.

The genome with the highest closeness value is *E. fergusonii ATCC.* The values for the rest of the *Escherichia*, *Salmonella* and *Shigella* genomes fall just below the highest value. Below these come the values for *Yersinia* genome and finally the two *Pectobacteria* have the lowest values for closeness centrality.

The betweenness centrality values do not correspond to the degree centrality and closeness values. The highest betweenness value, at 0.07, belongs to the genome for *E. coli E24377.* This genome is the only bridge between the bulk of the network, containing all the genomes from the YESS group, and the part of the network formed by the Pectobacteria. If we wish to trace a path from any genome from *Escherichia, Salmonella*, *Shigella* or *Yersinia* to either of the genomes from *Pectobacteria* then we must pass through the node for *E. coli E24377.* Following this, *Pectobacterium carotovorum subsp. atroseptica* has the second highest value for betweenness centrality, at 0.03. The node for this genome creates a bridge between *E. coli E24377* and the almost disconnected *Pectobacterium carotovorum subsp. carotovorum.* All paths from the bulk of the network to *Pectobacterium carotovorum subsp. carotovorum* pass through *Pectobacterium carotovorum subsp. atroseptica* .

Betweenness values for some of the *Escherichia* genomes are close to that of *Pectobacterium carotovorum subsp. atroseptica* but most genomes in the network

have a very low value for betweenness centrality on the network of genomes at the 98% similarity threshold.

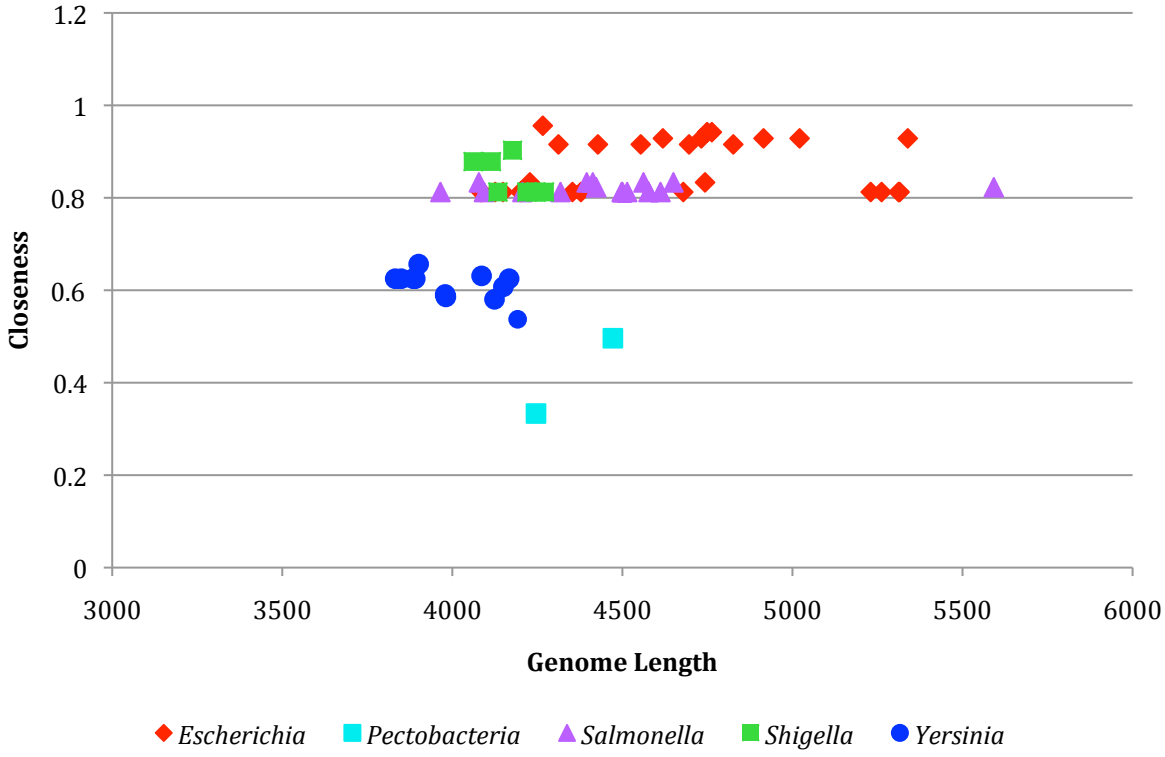


Figure 4.12: Closeness centrality against genome length for the network of genomes with 98% similarity threshold.

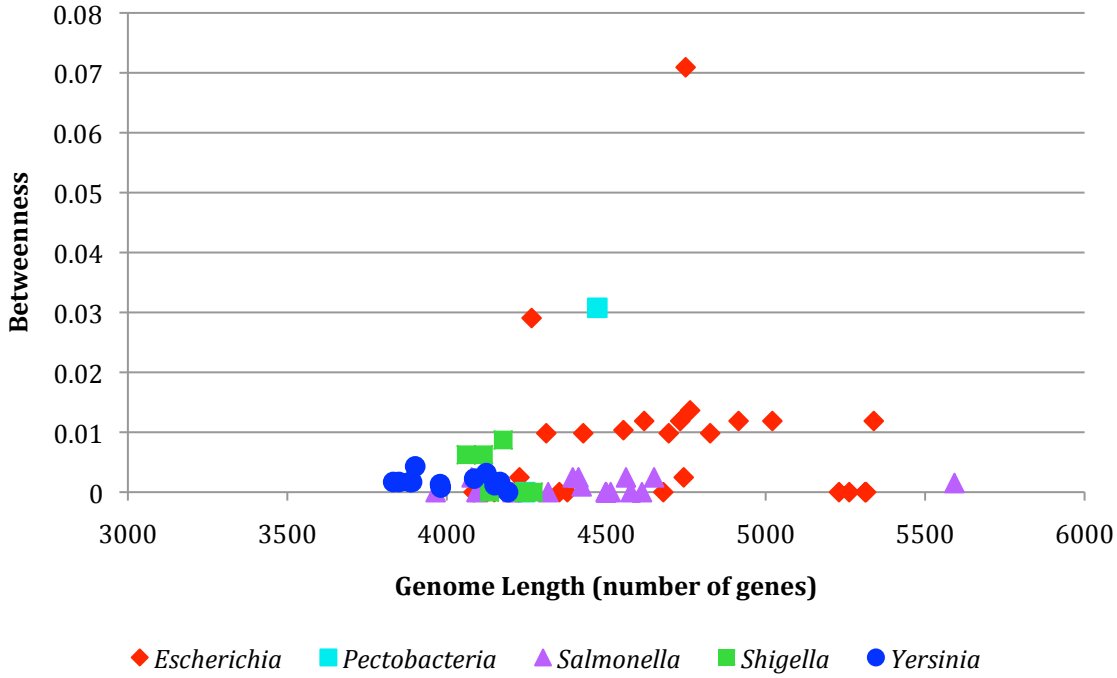


Figure 4.13: Betweenness centrality against genome length for the network of genomes with 98% similarity threshold.

4.3.1.7: Network of genomes with 99% Similarity Threshold

From 98 to 99% similarity there is a loss of a further 41 outgoing edges. The story stays much the same in that the *Yersinia* and *Pectobacteria* have fewer outgoing edges than expected while *Escherichia*, *Shigella* and *Salmonella* show the opposite trend. The genomes within each genus remain maximally connected to one another. The *Escherichia*, *Shigella* and *Salmonella* also remain maximally connected to one another.

The strongest edges (Figure 4.4) appear within the *Yersinia* demonstrating strong within-genus relationships. However, the lack of connections between the *Yersinia* and the rest of the network suggest that this genera has diverged further from the rest if the group.

The *Pectobacteria* has completely diverged from the rest of the group at this level of similarity. Neither of these two genomes have any genes in common with any genomes from the *Escherichia*, *Salmonella*, *Shigella* or *Salmonella*, that are 99% or more similar.

It is important to note at this stage the fact that even if we only include pairs that are almost identical across the extent of their sequences we still retain a highly connected network. There is no way, from just looking at the network, that we can distinguish species barriers between the *Escherichia*, *Salmonella* and *Shigella*. Every genome in these three genera has at least one gene with at least one homolog in every other genome within the three genera, that is 99% identical.

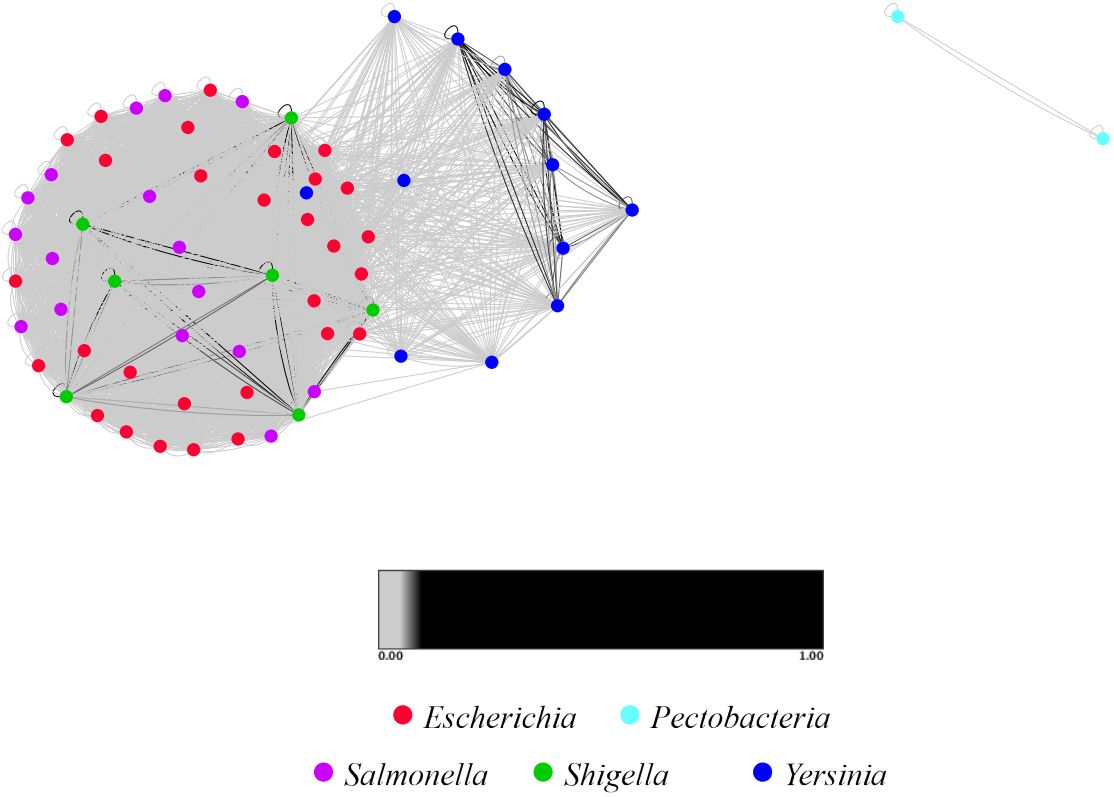


Figure 4.14: Networks of genomes at 99% similarity threshold. Lighter edges are weaker by comparison (very close to 0 on the scale bar). The darker edges are the strongest in the network (closer to 1 on the scale bar).

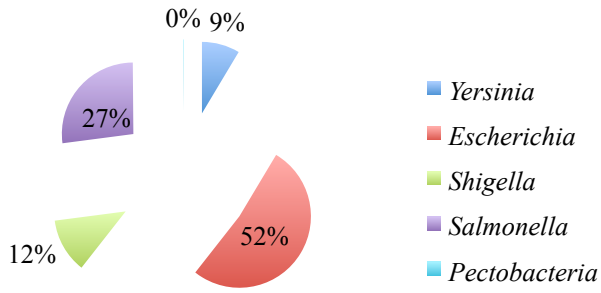


Figure 4.15: Pie chart of the percentages of overall outgoing edges represented by each genus for the network built from genes that have at least 99% sequence similarity.

Table 4.8: The number of outgoing edges and corresponding percentages for each genus in the network built from genes that have at least 99% sequence similarity. The number of outgoing connections is calculated from the initial percentages indicated by the maximally connected network. The Chi-squared test provides significance scores. For genera that have less outgoing edges than expected the P value is highlighted in orange for those with more than expected outgoing edges the P value is highlighted in black.

Genus	**No. Outgoing Connections**	**Percentage of All Connections**	**Expected No. Outgoing Connections**	**P value**
Yersinia	267	8.582449373	565.6363636	3.6545E-36
Escherichia	1618	52.00900032	1366.954545	1.12066E-11
Shigella	384	12.34329797	329.9545455	0.002926933
Salmonella	838	26.93667631	754.1818182	0.002272387
Pectobacteria	4	0.128576021	94.27272727	1.43864E-20
	3111	100	3111	

Table 4.9: Number of outgoing edges from the source genera labeled on the left to the target genera labeled on top for the network built from genes that have at least 99% sequence similarity. Cells highlighted in yellow represent the maximum number of outgoing edges a given genus can have to towards the target genus, i.e. the genomes in this genera are maximally connected.

	Yersinia	***Escherichia***	***Shigella***	***Salmonella***	***Pectobacteria***	**Total**
Yersinia	144	111	2	10	0	267
Escherichia	110	841	203	464	0	1618
Shigella	20	203	49	112	0	384
Salmonella	6	464	112	256	0	838
Pectobacteria	0	0	0	0	4	4
						3111

4.3.1.8: *Modules on the Network of genomes with 98% Similarity Threshold*

At this almost extreme similarity threshold we would expect that communities of tightly connected nodes would be synonymous with species boundaries. It is obvious at this stage that the *Pectobacteria* have formed a species module, this much is evident from looking that network. NeMo does not detect the *Pectobacteria* module because two genomes connected exclusively to one another are not considered to be "densely connected".

The modules that are found by NeMo include three that correspond to the *Yersinia* species grouping (modules 1-3 on table 4.9) and three that advocate the grouping of *Escherichia*, *Salmonella* and *Shigella*.

Table 4.10: Modules according to NeMo for the network of genomes at 99% similarity threshold.

Cluster	Score (Density*#Nodes)	Nodes	Edges	Node IDs
1	-23.997	6	36	YERSPBM, YERSPPF, YERSPCO92, YERSPANT, YERSPKIM, YERSPANG
2	-24.757	10	100	YERSTBPB1, YERSPNEPAL, YERSTBYP3, YERSEN8081, YERSPBM, YERSPPF, YERSPCO92, YERSPANT, YERSPKIM, YERSPANG
3	-34.713	12	144	YERSTBIP31, YERSTB32, YERSTBPB1, YERSPNEPAL, YERSTBYP3, YERSEN8081, YERSPBM, YERSPPF, YERSPCO92, YERSPANT, YERSPKIM, YERSPANG
4	-174.757	12	144	SHIGSON, SALENSTY, SALENARIZ, ECOIAI1, SALENSCHW, ECOTW4359, ECOEDL933, SALTYLT2, SALENGAL, ECOKMG1655, ECONEW, ECOCATCC
5	-174.757	5	25	SALTYLT2, SALENGAL, ECOKMG1655, ECONEW, ECOCATCC
6	-174.757	6	36	EFERATCC, ECOREL606, SHIGDYS, SALENENTER, ECOSAKAI, ECOEC4115
7	-174.757	29	841	ECOBW2952, ECOKW3110, SALENATCC, ECOSE11, ECOKDH10B, ECOHS, SHIGBOYDCDC, SHIGSON, SALENSTY, SALENARIZ, ECOIAI1, SALENSCHW, ECOTW4359, ECOEDL933, SALTYLT2, SALENGAL, ECOKMG1655, ECONEW, ECOCATCC, SALENPARAC, SHIGBOY227, SALENNEW, SALENDUB, EFERATCC, ECOREL606, SHIGDYS, SALENENTER, ECOSAKAI, ECOEC4115
8	-174.757	6	36	SHIGSON, SALENSTY, SALENARIZ, ECOIAI1, SALENSCHW, ECOTW4359
9	-246.168	25	463	ECOLF82, ECOS88, SHIGF301, SHIGFLEX5, ECOSMS35, SALTYPHI, SALENCHOL, SALENPARAB, YERSTBIP31, YERSTB32, YERSTBPB1, YERSPNEPAL, YERSTBYP3, YERSEN8081, YERSPBM, YERSPPF, YERSPCO92, YERSPANT, YERSPKIM, YERSPANG, ECOIAI39, ECOH6E2348, ECOUTI89, ECOE24377A, ECO536

4.3.1.9: Centrality Measures of Nodes on the Network of genomes with 99% Similarity Threshold

At the 99% similarity threshold the genome with the highest degree centrality and closeness values is *E. coli 55989*. This genome is one of the larger, with 4,763 genes. The values just below the highest belong to the genomes from *Escherichia*, *Salmonella* and *Shigella*. The *Yersinia* genomes have values that are lower still and the *Pectobacteria* again, have the lowest values for degree centrality and closeness centrality.

In terms of betweenness, *E. coli CFT073* is the most central at the 99% similarity threshold. This is closely followed by many of the other genomes. At this high level of similarity the *Escherichia* nodes act as bridges between many of the genomes that are no longer sharing genes with one another. As can be seen in table 4.9, *Yersinia* no longer has a large number of connections with either the *Shigella* or *Salmonella* genomes. The *Escherichia*, however, remain relatively well connected to the *Yersinia*, so that many paths from a *Salmonella* or *Shigella* genome to a *Yersinia* genome will pass through an *Escherichia* genome.

The genome for *S. flexneri 2a str. 301* also has a comparatively high betweenness value. This *Shigella* genome is still sharing with 8 of the 12 *Yersinia* genomes and so appears on many of the paths between *Yersinia* and genomes from the *Escherichia*, *Salmonella* and *Shigella*. There are two more *Yersinia* and two more *Shigella* genomes that show a mid-range value for betweenness that are also positioned on the network at a bridging point between the two distinct parts. All other genomes remain at low values for betweenness centrality.

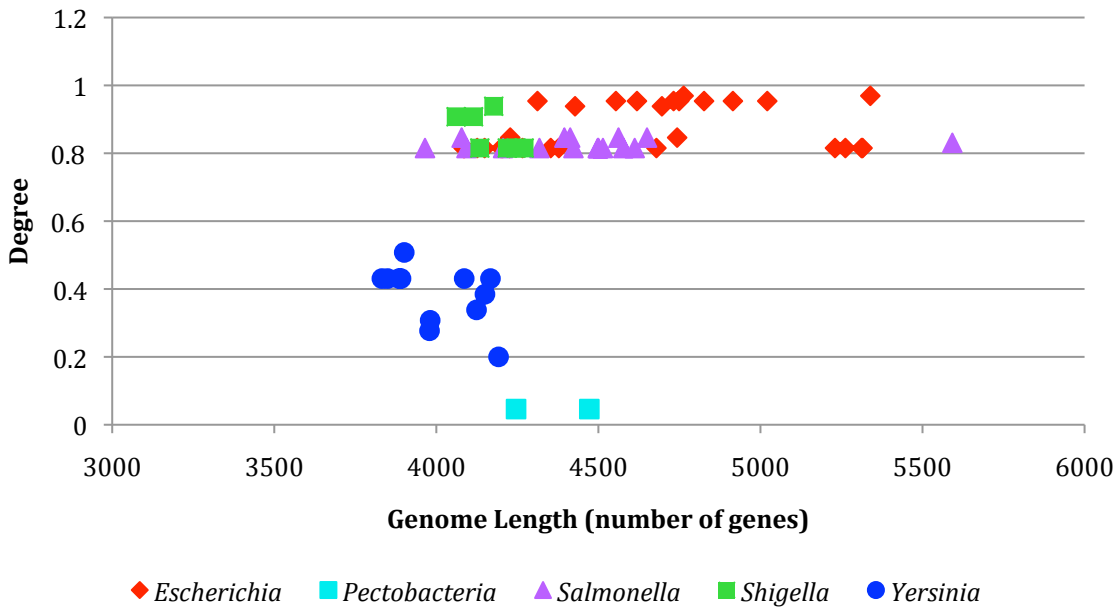


Figure 4.16: Degree centrality against genome length for the network of genomes with 99% similarity threshold.

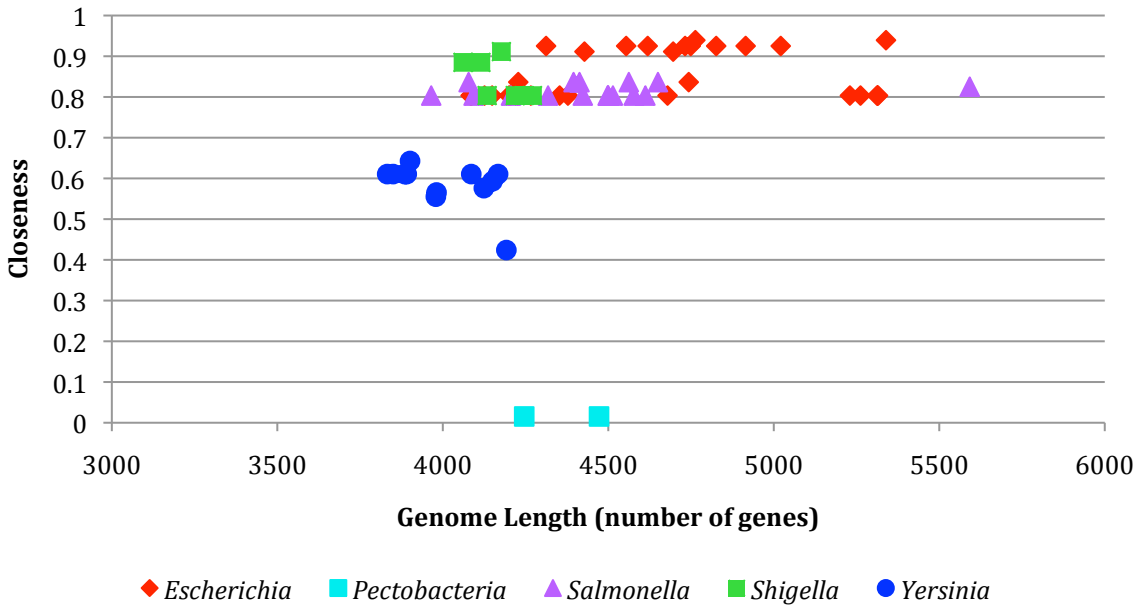


Figure 4.17: Closeness centrality against genome length for the network of genomes with 99% similarity threshold.

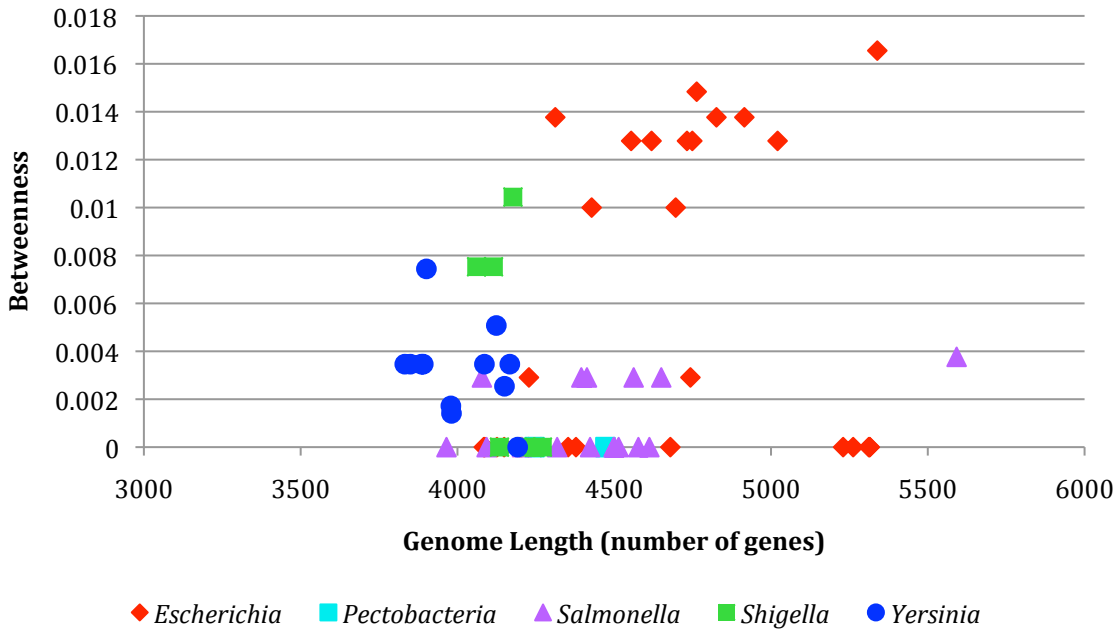


Figure 4.18: Betweenness centrality against genome length for the network of genomes with 99% similarity threshold.

4.3.1.10: Network of genomes with 100% Similarity Threshold

At the extreme threshold of 100% sequence identity, although many connections are lost, there is still a significant number of edges. In other words, there are many genomes, within and between-genus, with homologs that are identical across at least 80% of the gene. On this network there looks to be some clear species boundaries. *Escherichia, Salmonella* and *Shigella*, form a distinctive "ball" on the network. The *Pectobacteria,* of course, remain completely separated from the rest of the group and are only sharing genes with one another. The *Yersinia* have separated further from the group i.e. they have even weaker links with the genomes outside of the *Yersinia* genus. Even within the *Yersinia* community it appears as though many of the connections are not so strong any more (indicated by light coloured edges on Figure 4.19).

Interesting to note at 100% is the absence of an edge between any of the *Salmonella* and any of the *Yersinia*. There are no genes between *Yersinia* and *Salmonella* that are 100% similar. In fact *Yersinia* is sharing exclusively with *Escherichia* at this threshold. This would suggest that *Yersinia* is more closely related to *Escherichia* than to *Shigella* and to *Salmonella*. On the contrary, previous phylogenetic studies have suggested that *Yersinia* is equally closely related to both genera (Haggerty *et al.* 2009).

Finally, at the 100% similarity threshold, where all genes still included in the network are identical across 80% of their length there are still 2,366 outgoing edges. By comparison with the original percentages, *Yersinia* and *Pectobacteria* are less connected than expected and *Escherichia*, *Shigella* and *Salmonella* are more connected than expected. Yet at the threshold where only identical pairs of genes are

holding the network together we still find that *Escherichia* and *Shigella* are maximally connected to one another. *Salmonella* still has a large number of connections with the *Escherichia* and *Salmonella* but it has lost all edges to *Yersinia*. In fact *Yersinia* is now exclusively sharing genes within genus or with *Escherichia* genomes.

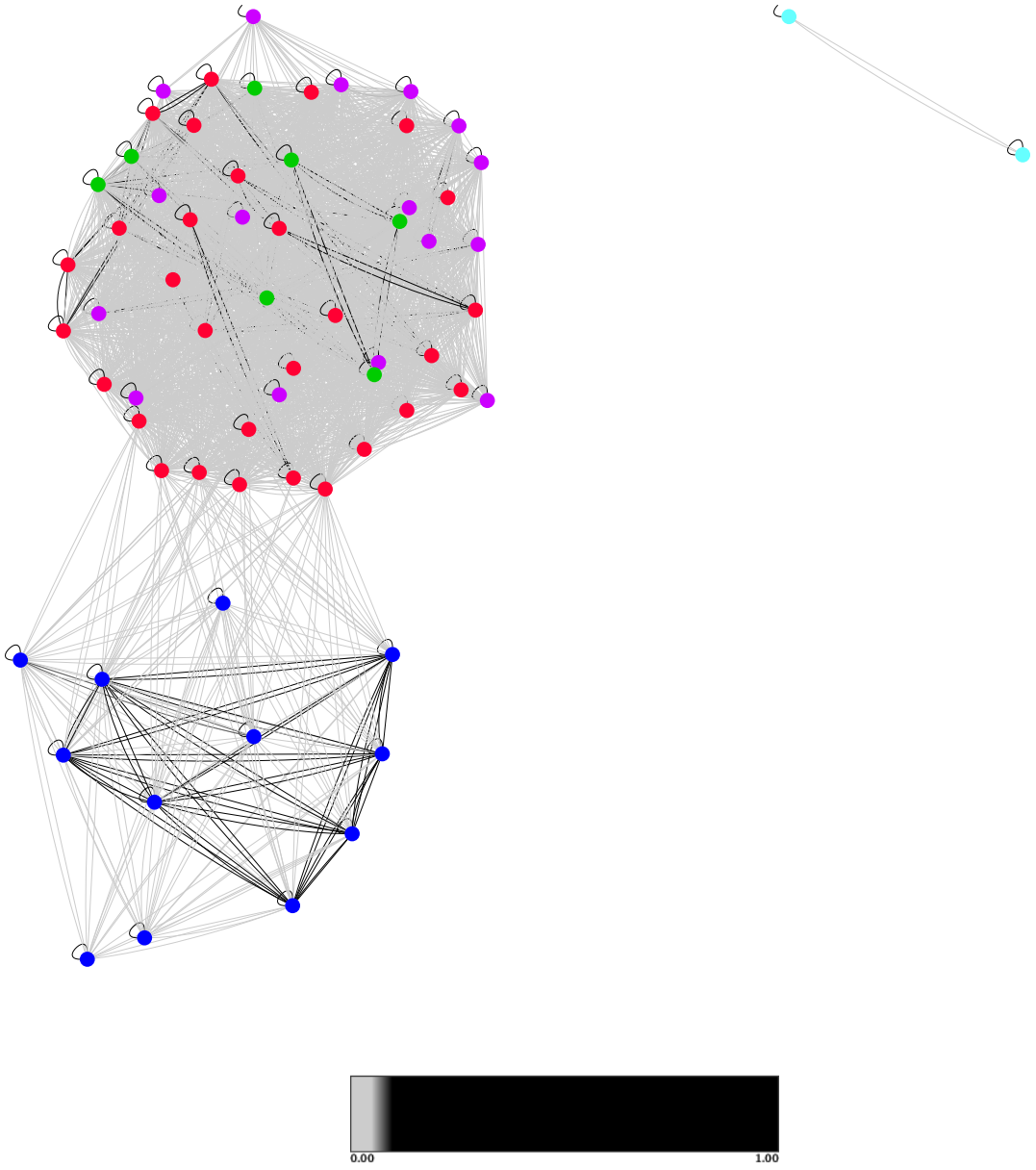


Figure 4.19: Networks of genomes at 100% similarity threshold. Lighter edges are weaker by comparison (very close to 0). The darker edges are the strongest in the network (closer to 1).

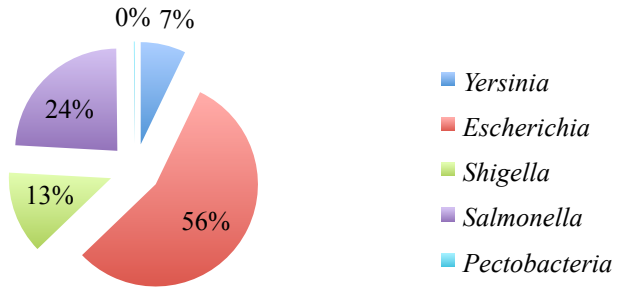


Figure 4.20: Pie chart of the percentages of overall outgoing edges represented by each genus for the network built from genes that have at least 100% sequence similarity.

Table 4.11: The number of outgoing edges and corresponding percentages for each genus in the network built from genes that have 100% sequence similarity. The number of outgoing connections is calculated from the initial percentages indicated by the maximally connected network. The Chi-squared test provides significance scores. For genera that have less outgoing edges than expected the P value is highlighted in orange for those with more than expected outgoing edges the P value is highlighted in black.

Genus	**No. Outgoing Connections**	**Percentage of All Connections**	**Expected No. Outgoing Connections**	**P value**
Yersinia	169	7.142857143	430.1818182	2.31778E-36
Escherichia	1317	55.6635672	1039.606061	7.74896E-18
Shigella	309	13.06001691	250.9393939	0.00024715
Salmonella	567	23.96449704	573.5757576	0.783647921
Pectobacteria	4	0.169061708	71.6969697	1.29567E-15
	2366	100	2366	

Table 4.12: Number of outgoing edges from the source genera labeled on the left to the target genera labeled on top for the network built from genes that have 100% sequence similarity. Cells highlighted in yellow represent the maximum number of outgoing edges a given genus can have to towards the target genus, i.e. the genomes in this genera are maximally connected.

	Yersinia	***Escherichia***	***Shigella***	***Salmonella***	***Pectobacteria***	**Total**
Yersinia	137	32	0	0	0	169
Escherichia	40	841	203	233	0	1317
Shigella	0	203	49	57	0	309
Salmonella	0	251	60	256	0	567
Pectobacteria	0	0	0	0	4	4
						2366

4.3.1.11: Modules on the Network of genomes with 100% Similarity Threshold

Detection of modules at the 100% similarity threshold reveals two distinct groupings of genomes. The first is the *Yersinia* group, corresponding to the *Yersinia* clade on phylogenetic trees (Haggerty et al., 2009). The second is the *Escherichia*, *Salmonella* and *Shigella* group. On a phylogenetic tree this group is represented as two or three separate clades depending on the separation of the *Escherichia* and *Shigella*.

Table 4.13: Modules according to NeMo for the network of genomes at 99% similarity threshold.

Cluster	Score (Density*#Nodes)	Nodes	Edges	Node IDs
1	-1.464	7	46	YERSPNEPAL, YERSTBIP31, YERSTBYP3, YERSEN8081, YERSPANT, YERSPKIM, YERSPCO92
2	-2.744	5	25	YERSPANG, YERSTBPB1, YERSTB32, YERSPPF, YERSPBM
3	-3.438	12	137	YERSPNEPAL, YERSTBIP31, YERSTBYP3, YERSEN8081, YERSPANT, YERSPKIM, YERSPCO92, YERSPANG, YERSTBPB1, YERSTB32, YERSPPF, YERSPBM
4	-126.859	15	123	ECOSMS35, ECOEC4115, ECOS88, EFERATCC, ECOIAI1, ECO55989, SHIGBOYDCDC, ECOAPEC01, SALENHEID, SALENPARAB, SALENDUB, SALENSCHW, SALENAGONA, SALENNEW, SALENARIZ
5	-132.419	20	229	SHIGSON, SHIGF245, ECOTW4359, ECOSAKAI, ECOEDL933, ECOSMS35, ECOEC4115, ECOS88, EFERATCC, ECOIAI1, ECO55989, SHIGBOYDCDC, ECOAPEC01, SALENHEID, SALENPARAB, SALENDUB, SALENSCHW, SALENAGONA, SALENNEW, SALENARIZ
6	-137.624	31	659	SALENPARAC, SALENCHOL, SALENENTER, SALENPARAA, SALENGAL, SHIGBOY227, SHIGDYS, ECOUTI89, ECOHS, ECOE24377A, SHIGFLEX5, SHIGSON, SHIGF245, ECOTW4359, ECOSAKAI, ECOEDL933, ECOSMS35, ECOEC4115, ECOS88, EFERATCC, ECOIAI1, ECO55989, SHIGBOYDCDC, ECOAPEC01, SALENHEID, SALENPARAB, SALENDUB, SALENSCHW, SALENAGONA, SALENNEW, SALENARIZ
7	-139.413	33	763	SALENPARAC, SALENCHOL, SALENENTER, SALENPARAA, SALENGAL, SHIGBOY227, SHIGDYS, ECOUTI89, ECOHS, ECOE24377A, SHIGFLEX5, SHIGSON, SHIGF245, ECOTW4359, ECOSAKAI,

		ECOEDL933, ECOSMS35, ECOEC4115, ECOS88, EFERATCC, ECOIAI1, ECO55989, SHIGBOYDCDC, ECOAPEC01, SALENHEID, SALENPARAB, SALENDUB, SALENSCHW, SALENAGONA, SALENNEW, SALENARIZ, SHIGF301, ECOSE11

4.3.1.12: Centrality Measures of Nodes on the Network of genomes with 100% Similarity Threshold

On the network of genomes made from pairs of identical genes only, the genome for *E. coli 536* has the highest degree centrality at 0.83. This genome, all of the other *Escherichia* and *Shigella* and many of the *Salmonella* genomes have remained highly connected in the network. These genomes also have a corresponding high closeness centrality. However, 5 of the 16 *Salmonella* genomes have now lost contact with so much of the network that their degree centrality and closeness values have fallen quite drastically. At this level of similarity these five *Salmonella* genomes have similar degree centrality and closeness values to the *Yersinia*.

For the first time with our data, the genome with the highest degree centrality and closeness values is also the most central in terms of betweenness. *E. coli 536* has retained many of its outgoing edges to genomes within the *Escherichia* as well as those in other genera. At the same time many of the between genus edges have been lost at this level of similarity. Since *E. coli 536* is connected to parts of the network that have lost connections with one another, i.e. many of the between-genera connections are no longer present, it acts as the best bridging node at this point in the analysis. Six other *Escherichia* genomes have a similar role in the network. By retaining relationships with genomes from all genera they connect otherwise separate components.

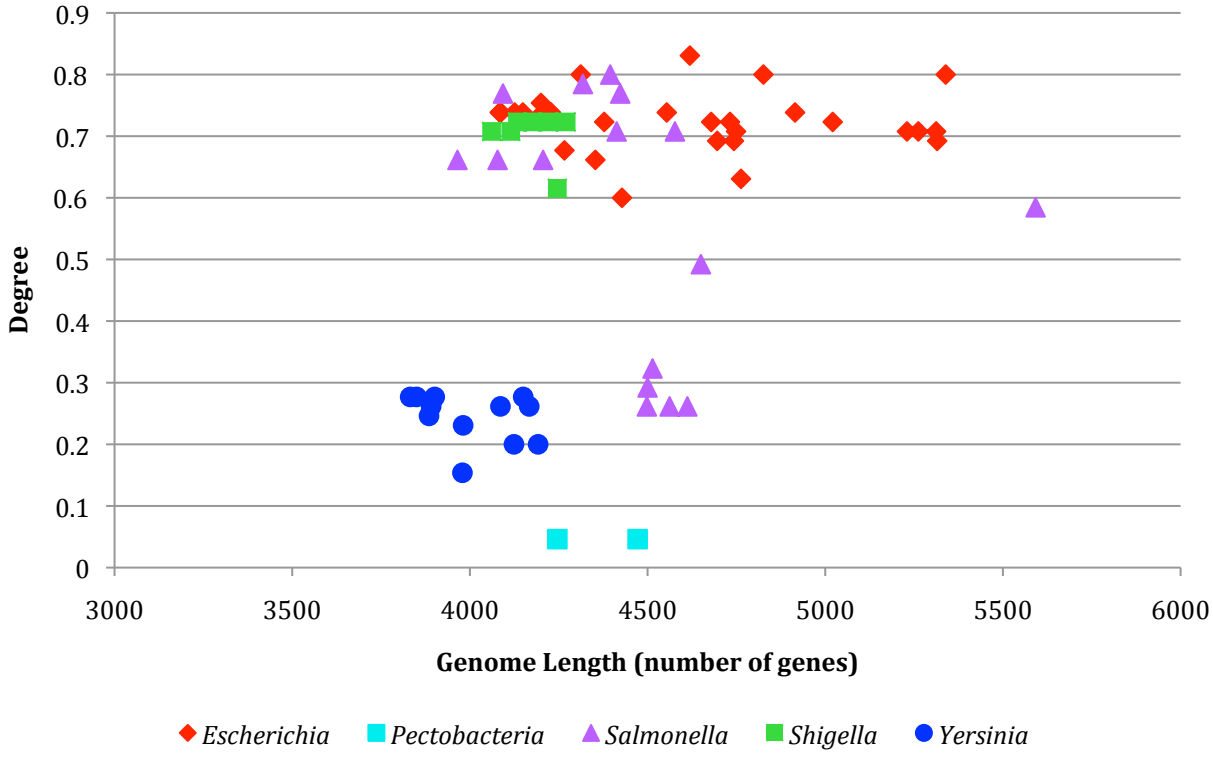


Figure 4.21: Degree centrality against genome length for the network of genomes with 100% similarity threshold.

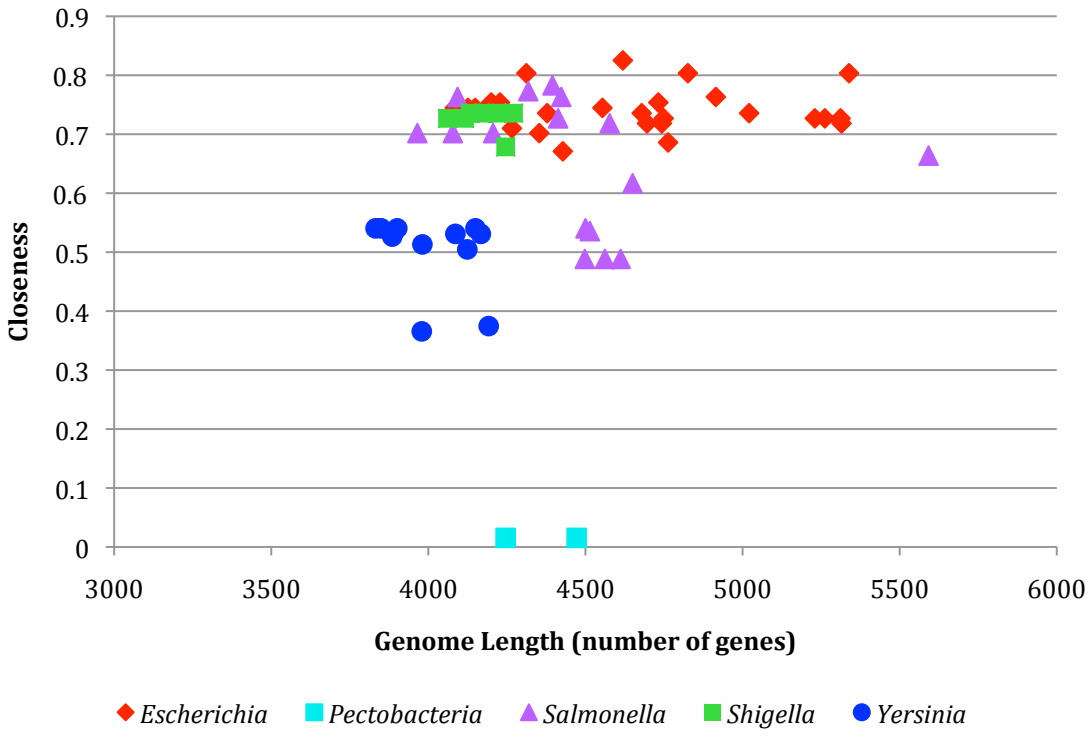


Figure 4.22: Closeness centrality against genome length for the network of genomes with 100% similarity threshold.

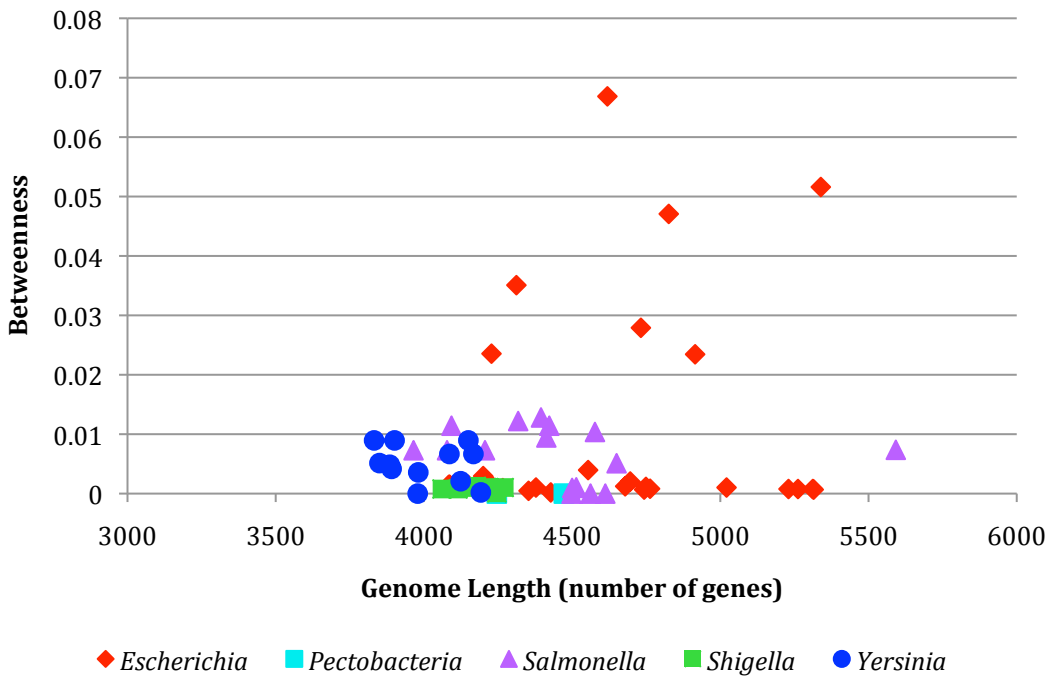


Figure 4.23: Betweenness centrality against genome length for the network of genomes with 100% similarity threshold.

4.3.2: Network of Genes

4.3.2.1: COG Category Analysis

When we refer to the results from the COG category analysis we find that a considerable portion of the genes have no hit in the COG database. In other words many of the genes from our dataset are not similar enough to any the genes in the COG-categorised database to merit a reliable prediction of their COG category. Of the 291,172 genes in our dataset, 54,840 have no homolog in the COG database. A further 22,520 are predicted to have an unknown function and 32,225 have a general function prediction only. The equivalent of one third of the genes in our database could not be assigned a COG category.

There are 22 COG categories in total including unknown function and general function prediction. We find higher numbers of genes in categories associated with metabolism, transcription and replication, recombination and repair (Figure 4.24). As we increase the similarity threshold to 95%, i.e. the point at which the network of genomes is no longer maximally connected, we find very little change. There is a loss of 2,195 genes, but the losses are distributed fairly evenly across the categories, no category has lost significantly more or less genes than we would have expected. This trend follows on as we raise the threshold further. Between the 99 and 100% thresholds we see a slightly significant loss in the number of genes involved in extracellular structures (P-value = 0.05), in cell cycle control (P-value = 0.003) and in intracellular trafficking (P-value = 0.007).

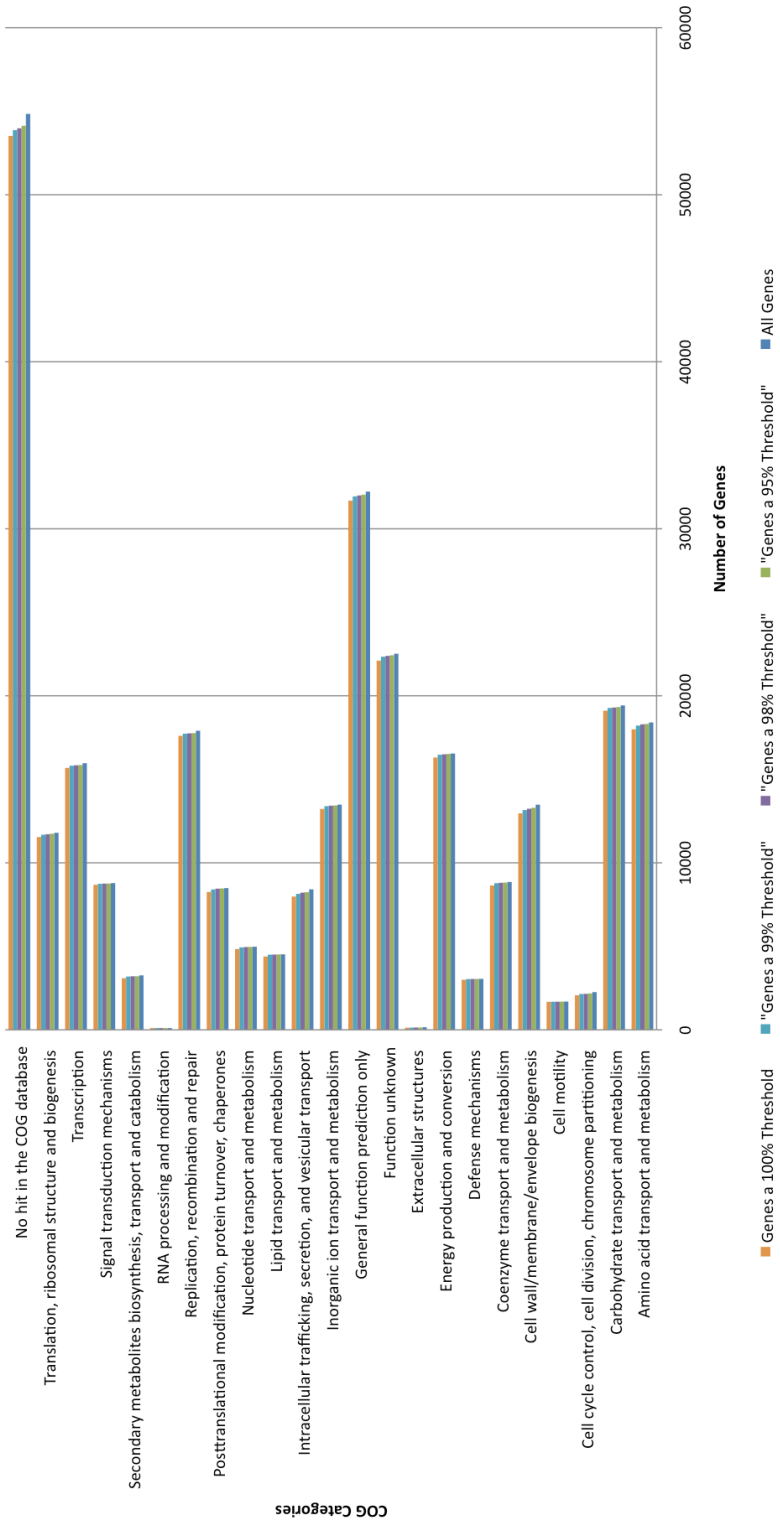


Figure 4.24: Distribution of COG functions for genes at each level of sequence similarity.

4.3.2.2: GenBank Functions

For all 291,172 genes in our dataset there are 20,196 different GenBank functional annotations. The genes are given quite specific functions rather than broad categories. This means that there is redundancy across many of the annotations, often the same or very similar functions can be named several different ways. One example is for proteins involved in transportation, these can be named 'transport protein', 'transporter protein', 'putative transporter', 'predicted transporter', etc. or even something more specific such as 'ABC transporter ATP-binding protein'. We have to bear this in mind when quantifying the different gene functions. However, when we look to the top 25 occurring GenBank functional annotations we see a bias towards certain types of functions.

From the list of GenBank functions for all genes in our dataset we find that the most abundant is the 'hypothetical protein'. There are 163,448 of these non-informative, hypothetical proteins accounting for more than half of the data. For the genes that have a more informative functional annotation we find that the most abundant are the 'putative inner membrane proteins'. An inner membrane is found in all gram-negative bacteria. Interesting to note is the fact that all known conjugative systems make use of an inner membrane protein known as a coupling protein. The coupling protein has a cytoplasmic domain that links secretion systems to relaxosome-bound DNA during transfer (Frost, 2005). In other words this type of protein is essential in conjugative HGT.

The second most abundant functional annotation is 'putative lipoprotein'. Lipoproteins emulsify lipids allowing them to move through the water inside and

outside of the cell. Bacterial lipoproteins are membrane-anchored and typically account for approximately 2% of the bacterial proteome. They have a range of functions including promoting antibiotic resistance, cell signaling and substrate binding in ABC transport systems, and bacterial conjugation.

Within the top 25 occurring functional annotations for all genes in our dataset we find a number of instances of transporter proteins namely 'putative transport' proteins, 'putative transporters', 'predicted transporters', and 'ABC transporter related' proteins. Transporter proteins are involved in moving substances within an organism. In particular the ABC transporter proteins are often involved in iron uptake systems that are important in virulence and often transferred horizontally between enteric bacteria, I discuss this further in the following section (4.3.2.3).

Also within the top 25 occurring functional annotations are transposases and insertion sequences (IS). An IS is a short DNA sequence that acts as a transposable element, unlike most transposable elements they do not carry accessory genes. Instead an IS will only code for a transposase and a regulatory protein which either stimulates or inhibits the transposition activity. Transposases and insertions sequences make up 3,938 or 14.78% of the 26,652 genes in the list of the top 25 occurring GenBank functions.

There are large numbers of genes involved in mechanisms of HGT in the highest occurring functional annotations, but we also see a substantial number of genes involved transcriptional regulation. These informational genes make up a total 5,556 of all the genes in the dataset.

As we raise the similarity threshold and only include pairs of genes with 90% similarity across 80% of their length, we are left with 289,418 genes to assess. Still we find the same categories of functions at the top of the list of GenBank functions.

Further more, when we raise the similarity threshold to 95, 98, 99 and 100% we see no change in the functional annotations that occur most.

At every level of similarity, the genes in our dataset that are most dominant in terms of quantity are mobile genes, those involved in the mobilization of other genes and informational genes. The fact that these types of genes have remained abundant to the highest level of similarity, suggests that they are highly similar and among the last to diverge.

4.3.2.3: Levels of Similarity between Homologs

When we evaluate the number of genes that fall into each of the bins we find a pattern in the distribution that corresponds to the pattern of divergence between two homologous genes. In Figure 4.25 we see a hump in the data that corresponds to an area of increased quantity of homologous pairs. This hump falls between 75.25% and 88.75% sequence similarity and accounts for approximately 25% of the homologous relationships in the dataset. In other words a 25% of homologs have diverged 11.2 to 24.75%. We expect that the minimum level of divergence between two genes from different genomes in our dataset would fall within the range of 11.2 to 24.75%. However, on Figure 4.25 we also see a reversal of the trend in decreasing numbers of genes connecting a pair of genomes resulting in a kick up at the end that corresponds to a large number of homologous genes that have between 93.25 and 100% sequence similarity. This kick up accounts for 72% of the homologous relationships in the dataset, in fact 56% of all the homologous relationships have more than 97.75% sequence similarity.

The extremely high percentage similarity for such a large number of homologous pairs cannot be explained by self-hits, i.e. every gene has 100% sequence similarity with itself. These relationships were removed before the data was assessed. However, the kick up could possibly be an artifact of within-genus relationships. In other words we would expect two genes from genomes in the same genus to be highly similar. To test for this explanation we quantify the number of homologous pairs that fall into each bin for within- and between-genus relationships.

For the within-genus relationships (Figure 4.26) we find that, for the most part, there is small a hump within the area of 10% divergence. Within the *Yersinia* there are very few relationships that have diverged by more than 5%. Approximately 41% of the within-*Pectobacteria* relationships fall within the range of 84.25 – 95.5% sequence similarity. Finally, as expected, for every within-genus relationship there is a substantial increase in the number of homologous relationships at and beyond 95.5% sequence similarity. The percentage of relationships that fall into the top bin, i.e. have sequence similarity of 97.75% or more, ranges from approximately 56% for the *Pectobacteria* to just below 93% for the within-*Yersinia* relationships. It is obvious from these results that, within each genus, the genomes are very closely related.

We find for the between-genus relationships (Figure 4.27), that for every pair of genera bar *Escherichia* and *Shigella*, there is an increase in the number homologous pairs in the range of 73 to 88.75%. From the homologous pairs between *Escherichia* and *Shigella*, only 2.7% fall into this range. On the contrary, pairs of homologs within this range make up between 87 and 95% of the overall number of homologous pairs between any other given pair of genera. The majority of homologous pairs between *Escherichia* and *Shigella* have more than 95.5% sequence similarity. Overall, the

Escherichia-Shigella relationship mimics the within-genus relationships. In fact, with approximately 67% of their homologs falling into the uppermost bin, *Escherichia* and *Shigella* appear to be more similar than the *Pectobacteria* are within-genus.

For seven of the ten between-genus relationships the number of homologous pairs begins to drop at 93.25% sequence similarity. For the *Escherichia-Salmonella* and *Escherichia-Yersinia* relationships however, there is a slight kick up from 95.5 to 100% sequence similarity. For the *Escherichia* and *Salmonella*, this kick up accounts for less than 5% of the homologous pairs, that means that 66,743 of *Escherichia-Shigella* homologs are more than 95.5% similar. For the *Escherichia* and *Yersinia* approximately 10% or 22,296 of the homologous pairs share more than 95.5% sequence similarity. Despite the substantial amount of gene sharing at very high levels of similarity between *Escherichia* and *Salmonella* and between *Escherichia* and *Yersinia*, there is almost no sharing between *Salmonella* and *Yersinia*. Only 55 pairs of homologs fall into the top bin for the *Salmonella-Yersinia* relationship. That means that less than 0.05% of the homologous pairs between *Salmonella* and *Yersinia* have 97.5% or more sequence similarity.

Previous phylogenetic studies have suggested that *Yersinia* is equally closely related to *Escherichia* and *Salmonella (Haggerty et al. 2009)*. However, our work suggests that *Yersinia* genomes share many more highly similar homologs with genomes from the *Escherichia* than with those from the *Salmonella*. This suggests that *Yersinia* is in fact more closely related to *Escherichia* than to *Salmonella*. Also, because we see a notable amount of gene sharing between *Escherichia* and *Salmonella* and between *Escherichia* and *Yersinia* but not between *Yersinia* and *Salmonella* we can assume that each pair of genera are sharing different types of genes. If the *Escherichia*

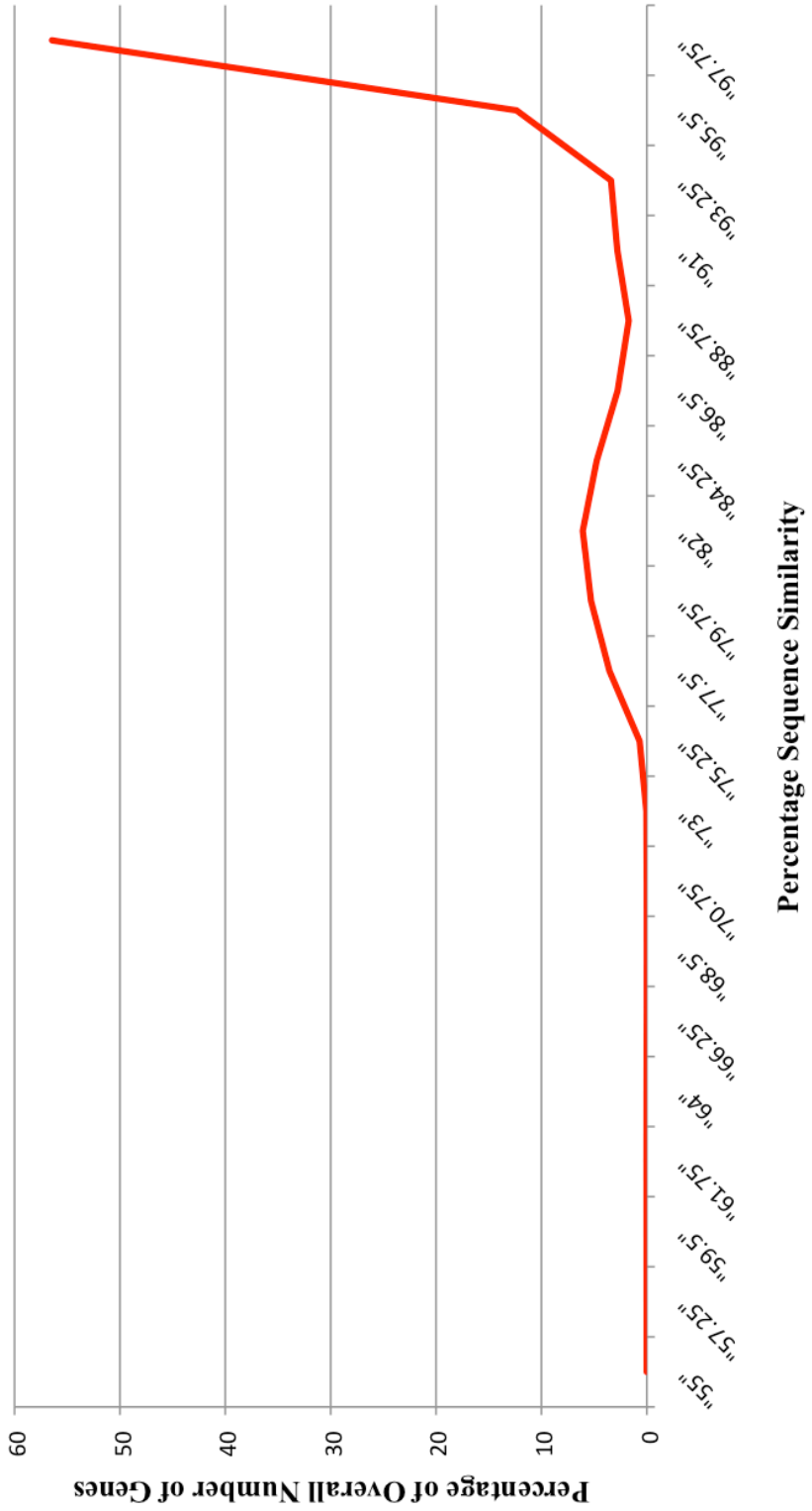


Figure 4.25: Percentage of homologous relationships at each level of sequence similarity.

genomes were sharing the same genes with *Salmonella* and with *Yersinia*, then surely these genes would be homologs between *Salmonella* and *Yersinia*.

When we look to the homologous pairs between *Escherichia* and *Salmonella* that fall into the top bin we find two distinct types of genes; informational genes (those involved in transcription, translation, and related processes) and genes that are likely to have been acquired recently through horizontal transfer.

Figure 4.28 shows the network of homologous genes between *Escherichia* and *Salmonella* that have 97.5% or more sequences similarity. Each node is a gene; red nodes come from *Escherichia* genomes and purple nodes come from *Salmonella*. There is an edge between any pair of genes that share regions of homology for more than 97.5% of the residues over at least 80% of their length. Only relationships between the two different genera and not within are represented on this graph, in other words there are no edges between two genes from the *Escherichia* or between two genes from the *Salmonella*.

Groups of homologs form clusters or connected components. On the network different clusters are enclosed in different coloured circles. These circles indicate different types of genes. Inside the pink circle are all the clusters containing informational genes. Informational genes are involved in important processes such as transcription and translation so they are needed in all genomes and are likely to be highly conserved. We can see that, on the network, the clusters in the pink circle, for the most part, contain lots of genes from both genera i.e. they are generally universally distributed.

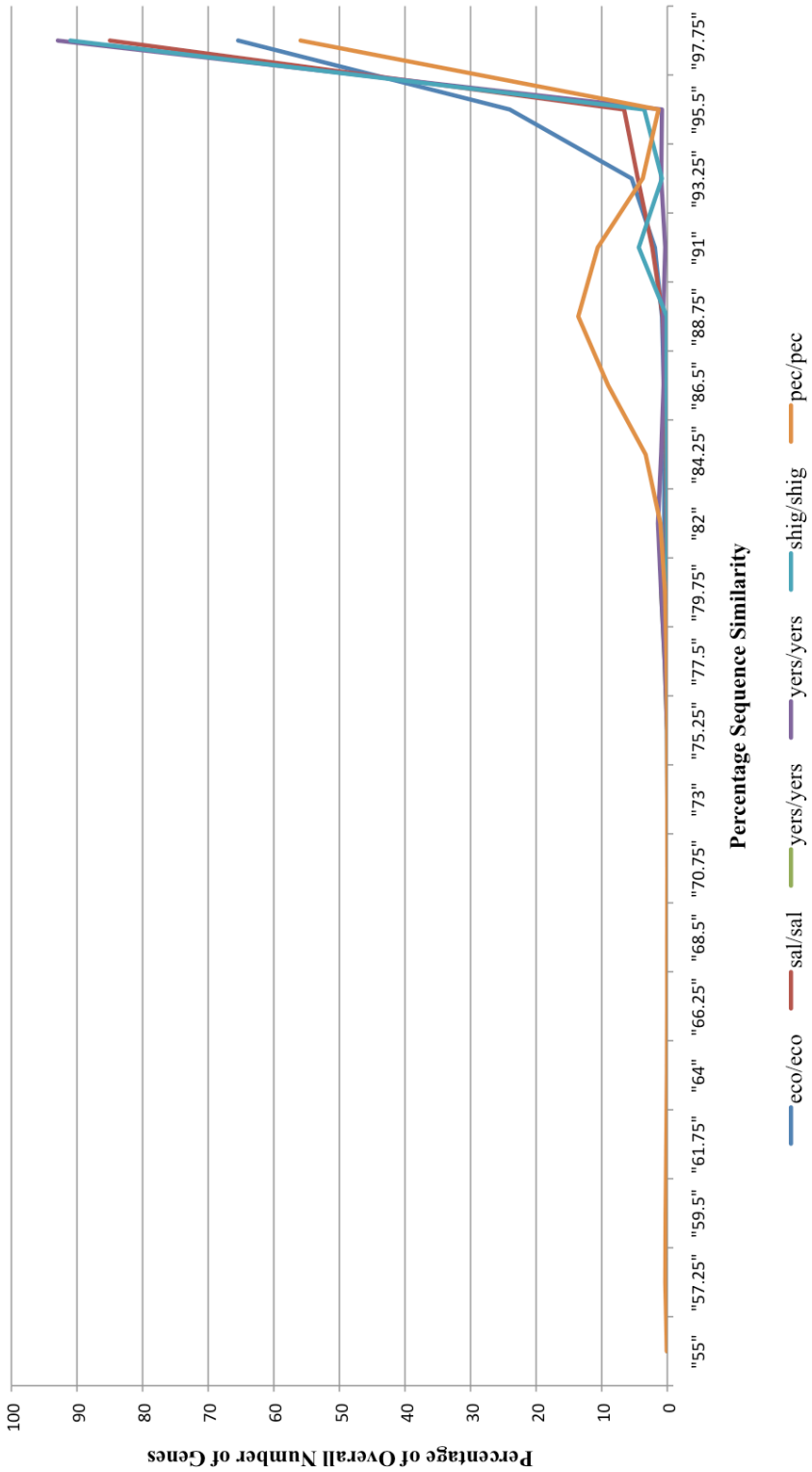


Figure 4.26: Percentage of homologous relationships at each level of sequence similarity for all within-genus relationships.

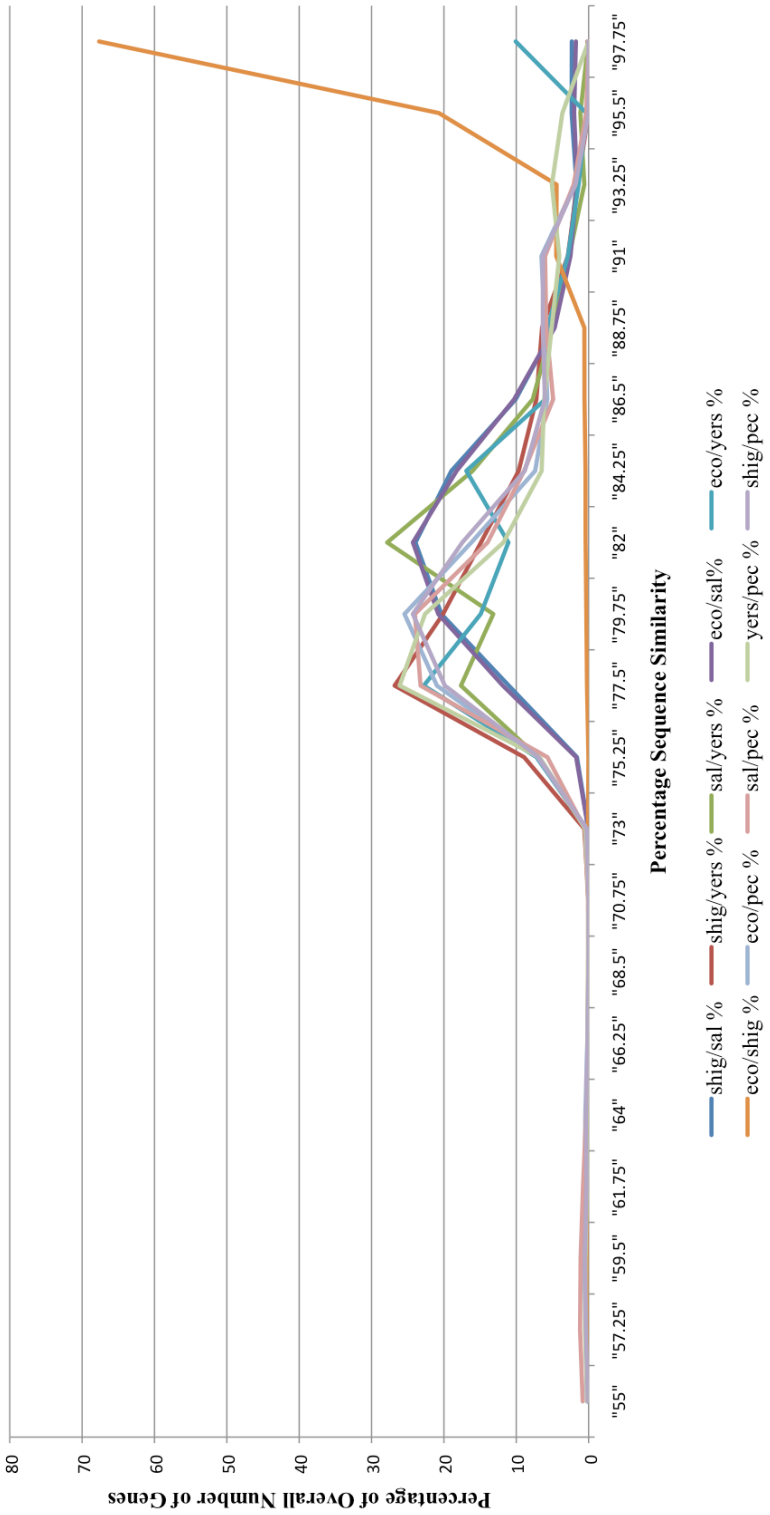


Figure 4.27: Percentage of homologous relationships at each level of sequence similarity for all between-genus relationships.

The large blue circle contains all phage related genes. It can be see in Figure 4.28 that these mobile genetic elements are far more sparsely distributed than the informational genes. Most of the clusters of phage related genes contain genes from only a few different genomes and have strange patterns of relationships, e.g. in some cases two genes from *Salmonella* are homologous to the same gene from *Escherichia*, but just one of the *Salmonella* genes is also homologous to a different *Escherichia* gene.

The patterns of relationships between transposases (within the green circle) from the two genera are somewhere in between the patterns of relationships for the informational gene and for the phage related genes. There is no doubt that there is a large number of transposases and that the level of similarity between those from different genomes is very high, but some appear to be ubiquitous while others are more sparsely distributed. Some of the clusters of transposases contain representatives of many genomes from both genera whereas others contain many from one genus and just a few or one from the other genus. This is indicative of a relationship in which the transposase arose in the genus for which there is many representatives and was horizontally transferred to a select few genomes from the other genus.

The genes inside the small orange circle are acquired antibiotic resistance genes. Genes for Beta-lactamase and ethidium bromide resistance as well as the multidrug efflux are shared between the *Escherichia* and *Salmonella*. The yellow circle contains genes for which the high level of similarity is most likely explained by a recent transfer event. They include operon leader peptides, inner membrane proteins, transporter proteins and genes involved in the hok/sok system of a plasmid.

Figure 4.29 shows the network of homologous genes between *Escherichia* and *Yersinia* that have 97.5% or more sequences similarity. Each node is a gene; red

nodes come from *Escherichia* genomes and blue nodes come from *Yersinia*. There is an edge between any pair of genes that share regions of homology for more than 97.5% of the residues over at least 80% of their length and the network is only representative of between genus relationships.

Again the informational genes are contained within the pink circle; this time there are far fewer informational genes to speak of. Just one type of small ribosomal subunit protein remains between the *Escherichia* and *Yersinia*. This gene in all *Yersinia* genomes has more than 97.5% similarity with the same gene in the genome for *E. fergusonii ATCC.* This may explain why, in the genome networks,

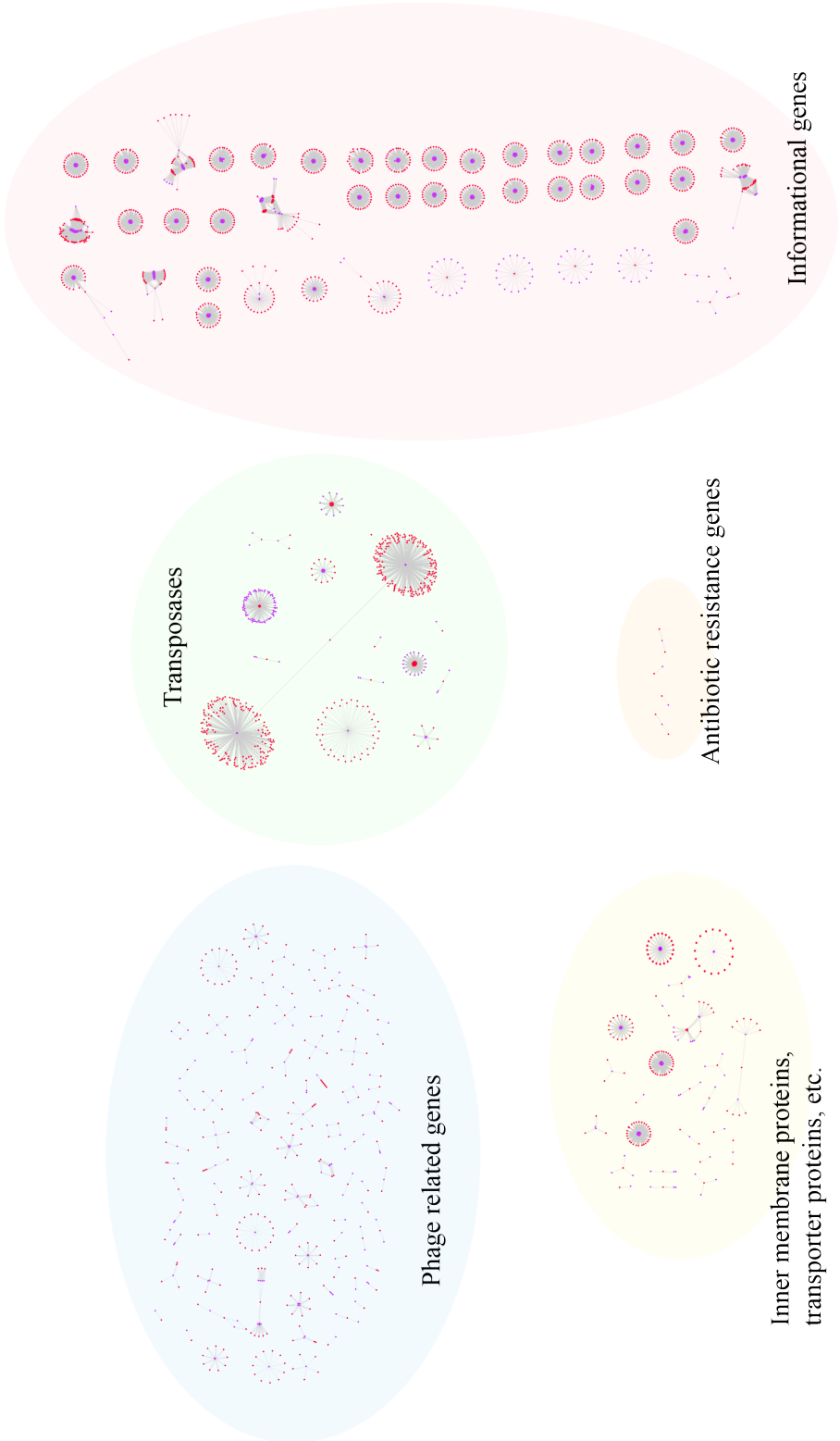


Figure 4.28: Network of homologous genes between *Escherichia* and *Salmonella* that have 97.5% or more sequences similarity. Red nodes are *Escherichia* genes and purple nodes are *Salmonella* genes.

E. fergusonii ATCC retains more connections than the other *Escherichia* genomes with the rest of the network (section 4.3.1.6). The *E. fergusonii* genome contains just one gene that is highly similar to a gene in all 12 *Yersinia* genomes. That means that the node representing *E. fergusonii* has 12 connections in the genome network as a result of just one gene.

There are many transposases shared between the *Escherichia* and *Yersinia* at the highest levels of similarity (green circle). Their distribution is sometimes sparse and sometimes universal, depending on the transposase in question. The high level of similarity between transposases and between genes contained in the yellow circle is most likely explained by a recent transfer event.

The purple circle on Figure 4.29 is contains only genes from the *yersinia*bactin biosynthetic gene cluster including siderophore and receptor proteins. These genes were not seen on the network of homologous genes between *Escherichia* and *Salmonella* that have 97.5% or more sequences similarity. The *yersinia*bactin genes, therefore, are uniquely shared between the *Escherichia* and *Yersinia*. *Yersinia*bactin siderophores are among the strongest iron-binding agents known. When bacteria and fungi are starved of iron they are known to secrete the *yersinia*bactin siderophore to scavenge for ferric ions. The siderophore and receptor genes that we find on the network of homologous genes between *Escherichia* and *Yersinia* that have 97.5% or more sequences similarity, have only been found in highly pathogenic *Yersinia* strains located on a high pathogenicity island (HPI). There have been two groups of HPI distinguished based on DNA comparison, the *Y. pestis* group and the *Y. pseudotuberculosis* group and it is thought that the *Y. pestis* group have been spread throughout the enterics. However we find that the *yersinia*bactin genes have the

highest level of similarity between genomes from the *Escherichia* and both *Y. Pestis* and *Y. pseudotuberculosis* genomes.

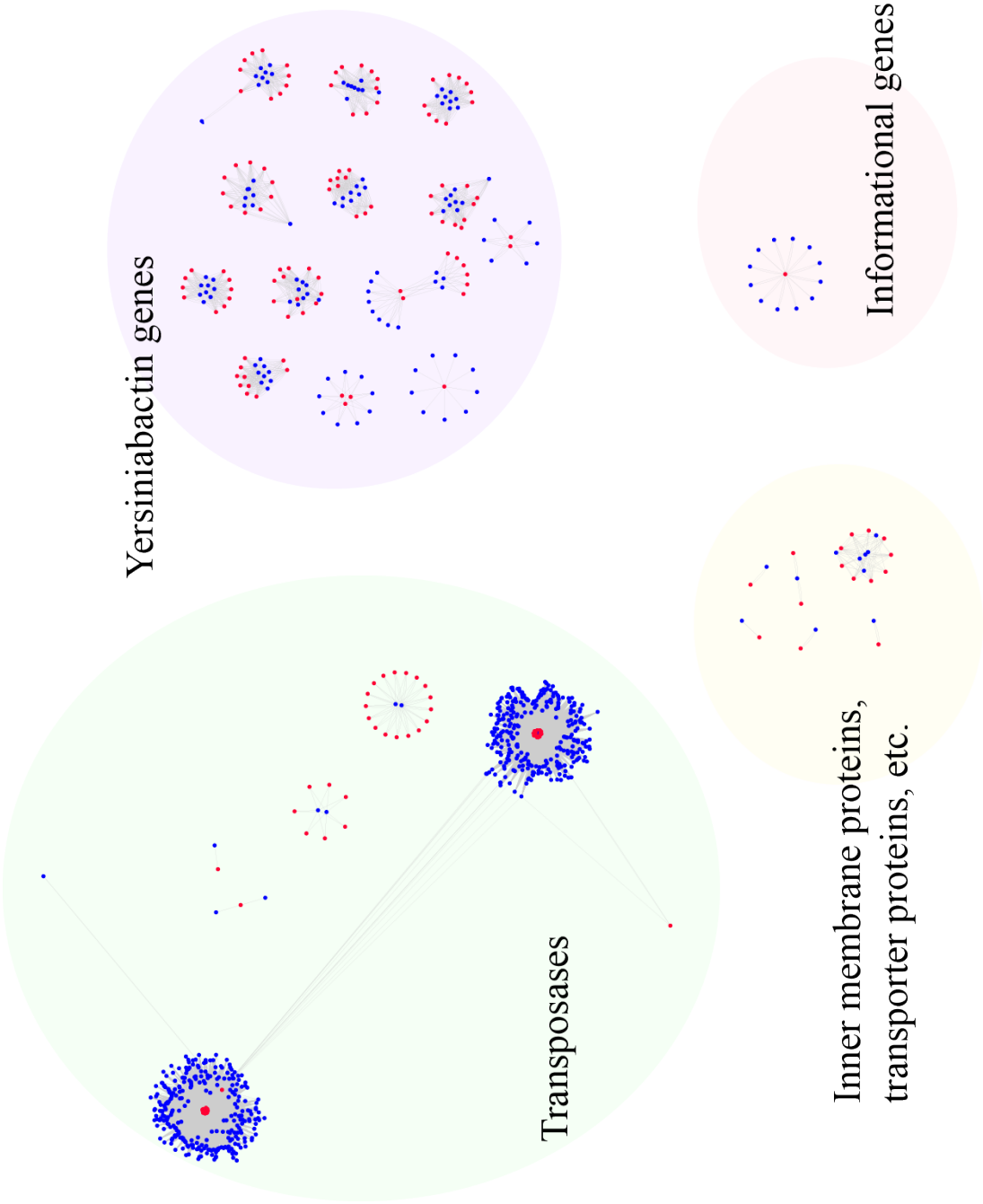


Figure 4.29: Network of homologous genes between *Escherichia* and *Yersinia* that have 97.5% or more sequences similarity. Red nodes are *Escherichia* genes and blue nodes are *Yersinia* genes.

4.4: Discussion

What seems indisputable is that we can identify organisms that have synapomorphies, both genetic and phenotypic. However, even though we recognize groupings, we do not have a bacterial species concept and we do not understand how these groupings (species, subspecies, even genera) form. The kinds of analyses that have shown that there is some structure among currently defined species have the limitation of only examining the evolutionary history of a set of core genes. Not only does this limit the amount of information used in the analysis, core genes are not representative of the rest of the genes in a genome in terms of factors such as functional category and rate mutation. For a modern system of classification to work, it must use complete genomes and be able to accommodate HGT. Staley (2006) suggested that we might consider a species to be an 'irreducible cluster' of organisms.

Assessing deep- and shallow-level phylogenetic relationships within the YESS group has been proven to be fraught with difficulties related to HGT and erosion of phylogenetic signal. The only consistent outcome from phylogenetic studies of the YESS group is the recovery of three groups: the *Yersinia* group, the *Salmonella* group and the *Escherichia/Shigella* group.

In this chapter I report observations on the way in which a group of closely related bacterial genomes have diverged from one another using networks of gene sharing. These networks of gene sharing provide a way to describe a genome in relation to other genomes. Both the vertical and horizontal components of evolution are represented on the gene-sharing network and thus provide an all-encompassing view of evolution.

When we observe the network of all gene sharing for the YESS group of bacteria we find it somewhat uninformative. Every genome has a homologous relationship to every other genome and there is a lack of phylogenetic signal of any kind. As we raise the similarity threshold up to and above 95%, the network begins to elucidate evolutionary signals. In some cases results from network analyses of homologous relationships adhere to the traditional way of thinking, i.e. the ribosomal phylogeny of bacteria. In other cases we find that the networks reveal unexpected insights into the relationships within the group.

More often than not we find that there is substantially more sharing within a genus than between genera. On the networks of genomes, at every level of similarity, the darkest edges, i.e. those between genomes with the most genes in common, are found between genomes from the same genus. When we look to the networks of genes, we find that, within genus, the majority of gene homology relationships fall into the highest bin for sequence similarity. We see that a number of the relationships between genes from different genera also have high similarity but the trend is much stronger within each genus.

The ribosomal phylogeny indicates that the *Yersinia* is the deepest clade and is equidistant from the *Salmonella* and the *Escherichia/Shigella* clades. The *Pectobacteria* is placed outside of this. The networks of genomes, in some way adhere to this signal. The genomes of the *Yersinia* and *Pectobacteria* are the first to move away from the network, i.e. at higher levels of sequence similarity they have fewer genes in common with the rest of the genomes in the network. Also in concordance with the ribosomal phylogeny are some of the modules found in networks of higher similarity thresholds. The *Yersinia* consistently form a tight-knit cluster on the network and the *Pectobacteria* remain connected to one another even

when they have lost all connection with the rest of the network. Clearly there is some evidence of species modules forming.

For the most part genomes from the *Escherichia*, *Salmonella* and *Shigella* are considered most central in the network, based on the three measures on centrality. The *Yersinia* and *Pectobacteria*, considered to have homology with the rest of the genomes, are found towards the peripheral of the network. A genomes position in the network, relates to its relatedness to the rest of the network. Again, this is apparent of the vertical signal within the group.

On the contrary, it is indisputable that forces other than the vertical inheritance of genes are influencing the evolution of this group. At every level of similarity there remains a huge number of connections between genomes. In fact up to the 90% similarity threshold the network is maximally connected. If it can be accepted that when DNA similarity levels between two strains are greater than 70% they can be assigned to the same species (Achtman, 2008, Cho, 2001, Konstantinidis, 2006, Stackebrandt, 1994, Staley, 2006) then there should be problems sub-categorizing the YESS group. Within the group of genomes from the *Escherichia*, *Salmonella* and *Shigella* in particular, there is an excessive amount of sharing. At the point where we consider only those pairs of homologs that are 100% identical across at least 80% of their length, i.e. the network of 100% similarity, the genomes from these three genera remain maximally connected. The 52 genomes from the *Escherichia*, *Salmonella* and *Shigella* comprise a group that, when judged by certain genes, will appear to be part of one species.

We saw from the networks of genes that homologs from within a genus tend to display the highest levels of similarity, in concordance with the ribosomal phylogeny. The distribution of percentage similarity between homologs from different

genera however, opposes the vertical signal. For specific cases there are an uncommon number of homologs from different genera that are similar across the majority of their length. The unusually high numbers of genes with unusually high levels of similarity are found between *Escherichia* and *Salmonella* and between *Escherichia* and *Yersinia* but not between *Salmonella* and *Yersinia*. These results confound expectations that the *Yersinia* would share equally with the *Escherichia* and *Salmonella*.

The genes that are found in the highest bins for between genera relationships are those that cause genomes to appear more closely related than previously reported by the ribosomal phylogeny. These genes are most likely to belong to functional categories involved in informational processes mechanisms of HGT. In fact throughout all of the analyses, at every level of similarity, the genes in our dataset that are most dominant in terms of quantity are mobile genes, those involved in the mobilization of other genes and informational genes. If the genes involved in the mobile portion of the genome are just as influential as the genes involved in the highly conserved proportion, then it would appear that we cannot justifiably describe the evolution of prokaryotes, exclusively, by a set of core informational genes.

Chapter 5: Concluding Remarks and Future Work

Many agree that the Tree of Life (ToL) has become redundant in describing the evolutionary history of prokaryotes. Processes or entities that do not fit the strictly vertical inheritance pattern are often omitted from studies altogether. In fact incongruence between prokaryotic gene trees is so rampant that some believe it is not even possible to create a 'tree of one percent' of the data (Puigbo and Koonin 2009). In this thesis I attempted to gain further understanding of such processes and entities that appear to confuse and confound the ToL hypothesis. Furthermore I explored the alternative of networks to describe the relationships between bacterial genomes and genes.

Fusion genes and their components do not align with one another in the traditionally optimal way. Relationships between genes that are not homologous along their entire length are usually trimmed or removed from the data altogether in order to cater for the branching pattern of the ToL. In chapter 2 I presented the premise of a method for detecting fusions of unrelated genes using network structure analysis. Although a number of fusion detection algorithms precede this (Enright, 1999, Marcotte, 1999, Suhre, 2004), they tend to rely on non-overlapping side-by-side BLAST hits from a source genome to a target genome. Not only are these algorithms limited by the input data and tend to be difficult to replicate on a large scale but they also provide results rife with false positives (Snel et al., 2000). Our method employs an all-versus-all approach that can include all data from multiple sources. Use of networks allows us to describe the relationships between all genes in the entire dataset

of interest. The method is dependant on fusion events forming specific structures on the network and so we recognize false negatives, concordant with those of previous algorithms, in the form of fusions of related genes e.g. as a result of tandem duplication.

I reported successful tests of the accuracy of our method using simulated and small datasets. The test provided confidence in the method's ability to accurately represent input data on a network structure and subsequently search this network and successfully retrieve and report potential fusion genes. Chapter 2 provides an account of further tests of the functionality and limits of this fusion detection method.

It was discovered that the limited availability of computational power would stunt the potential of a network-based algorithm. To overcome the methodological hurdles we restricted the size of our input datasets and presented a formula for predicting the gene fusion content of a particular genome. Despite the limited amount of input data and a strict definition of fusion genes that are reported, we estimated that almost 3% of the *S. enterica subsp. enterica serovar Paratyphi A* genome was made up of fusions of unrelated genes. It seems that the phenomenon of fusion has a more dominant role in bacterial evolution than was previously thought.

Bi-functional proteins have been a vital contributor to acquired antibiotic resistance in bacteria. This is endorsed by the results in chapter 3. We consistently find that fusions of unrelated gene tend to be involved in defense mechanisms.

Whole genome sequencing is a growing research area and produces huge amounts of data daily. Because of this, datasets are getting bigger, both in taxonomic sampling and the number of genes used. Therefore, in the current scientific environment, software implementations of methods are essential. This thesis follows the developmental process of a new method, from the conception of an idea, through

the implementation of that idea and its applications to both simulated and real world datasets. The algorithm described in chapter 2 has the potential to be presented as a user-friendly software program, albeit with some improvements. The most notable impediment of this algorithm is the current limitation on input dataset size. Splitting the network into smaller more easily traversable parts, is a precarious notion. The only way to reduce a network into smaller constituents without cutting away edges is to divide it into its connected components. Since connected components are disjoint from one another separating them does not run the risk of loosing valuable information pertaining to relationships between genes. However, the giant connected component on a network, i.e. the one with the most nodes, grows larger as more data is added. Networks of homology relationships between prokaryote genes have proven to be highly connected and so the giant connected component quickly becomes too large to search in real time.

Chapter 3 sees the parallelization of the algorithm on a number of datasets on a relatively small scale. Five datasets, with an overlap of one genome, were analysed side-by-side in order to obtain a broader view of fusion in bacteria. This has the potential to work on a much larger scale. In a preliminary study I have created a version of the algorithm whereby a genome of interest is specified and others are chosen at random to create a dataset of optimal size. The genome of interest is kept constant in order to obtain a picture of fusion for that genome, as was described for the *Salmonella enterica subsp. enterica serovar Paratyphi A* genome in chapter 3. So far this approach has been tested for the genome of *Aspergillus fumigatus* against 100 other fungal genomes. This preliminary study yielded 238 fusions of unrelated genes in this one genome. These results are yet to be verified and checked for duplicates but

the initial number is surprisingly high and may lead to revelations in relation to fungal evolution.

From the results reported in this thesis I think it is fair to say that the knowledge pertaining fusion genes, their occurrence and their importance, is only the tip of the iceberg. I think that this is an area that is relatively understudied and yet appears to be playing a significant role in the evolution of, at least, bacteria. The final goal of the work described in chapters 2 and 3 is to create a web-based interface for the retrieval of fusion genes in a genome of interest.

Chapter 4 is, in essence, a report of the similarities and differences encountered when comparing tree and network structures in describing the evolutionary history of a group of closely related bacteria. In many ways the information gathered from networks illustrating gene and genome relationships displays a vertical trend. However, these gene and genome networks also support the fact that HGT plays an equally important role in the evolution of bacteria. It becomes more and more apparent as the chapter unfolds that the horizontal component if bacterial evolution cannot be ignored if we truly wish to understand the dynamics of groups of bacteria. The evidence in chapter 4 emphasises that there are phylogenetic and non-phylogenetic signals within the bacteria, and that networks are fully competent in illuminating both.

Furthermore in chapter 4 we reveal support for previous studies that suggest that core informational genes are not necessarily the most abundant (Aziz *et al.* 2010). Alongside the highly conserved, particularly abundant ribosomal proteins and other such informational genes are the equally abundant mobile genetic elements and entities involved in mechanisms of HGT. In the past it has been accepted that the ribosomal phylogeny is appropriate for describing life because the gene is so

successful. In truth many of the mobile genetic elements within a bacterial genome are just as successful as the ribosomal genes in terms of abundance and even ubiquity.

The shift from using tree-like branching structures to describe life, to a more encompassing edifice has well and truly taken hold. Scientists with two very opposing views exist. There are those that endeavor to preserve the ToL and its authority in assigning species and those that feel there is no longer a place in evolutionary biology for such monistic thinking. I think that, either way it is fair to say that the genetic entities and processes that have so often been ignored due to their "inconvenient" existence, are more important than we know. If we are ever to fully understand the evolutionary history of prokaryotes we must find a way to embrace all aspects of this history, no matter how they disagree with our previous assessments of the data.

Chapter 6 – Bibliography

Ahmadinejad, N., T. Dagan, et al. (2007). "Genome history in the symbiotic hybrid *Euglena gracilis*." Gene 402(1): 35-39.

Allen, H. K., L. A. Moe, et al. (2008). "Functional metagenomics reveals diverse Œ≤-lactamases in a remote Alaskan soil." The ISME journal 3(2): 243-251.

Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." Journal of molecular biology 215(3): 403-410.

Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic acids research 25(17): 3389-3402.

Andersson, J. (2005). "Lateral gene transfer in eukaryotes." Cellular and Molecular Life Sciences 62(11): 1182-1197.

Avery, O. T., C. M. MacLeod, et al. (1944). "Studies on the chemical nature of the substance inducing transformation of pneumococcal types." The Journal of experimental medicine 79(2): 137.

Aziz, R. K., M. Breitbart, et al. (2010). "Transposases are the most abundant, most ubiquitous genes in nature." Nucleic acids research 38(13): 4207-4217.

Bairoch, A., B. Boeckmann, et al. (2004). "Swiss-Prot: juggling between evolution and stability." Briefings in Bioinformatics 5(1): 39-55.

Bao, L., M. A. Gorin, et al. (2009). "Preclinical development of a bifunctional cancer cell homing, PKCŒµ inhibitory peptide for the treatment of head and neck cancer." Cancer research 69(14): 5829.

Barber, M. (1961). "Methicillin-resistant staphylococci." Journal of clinical pathology 14(4): 385.

Basu, M. K., L. Carmel, et al. (2008). "Evolution of protein domain promiscuity in eukaryotes." Genome research 18(3): 449-461.

Batsch, A. J. G. K. (1791). Tabula affinitatum regni vegetabilis: quam delineavit, et nunc ulterius adumbratam tradit, Landes-Industrie Comptoir.

Beck, W. D., B. Berger-Bachi, et al. (1986). "Additional DNA in methicillin-resistant Staphylococcus aureus and molecular cloning of mec-specific DNA." Journal of bacteriology 165(2): 373-378.

Beiko, R. G., T. J. Harlow, et al. (2005). "Highways of gene sharing in prokaryotes." Proceedings of the National Academy of Sciences of the United States of America 102(40): 14332.

Bentley, D. R. (2006). "Whole-genome re-sequencing." Current opinion in genetics & development 16(6): 545-552.

Bergey, D. H., F. C. Harrison, et al. (1923). "Bergey's Manual of Determinative Bacteriology." 1st Ed.

Binnewies, T. T., Y. Motro, et al. (2006). "Ten years of bacterial genome sequencing: comparative-genomics-based discoveries." Functional & integrative genomics 6(3): 165-185.

Bland, D. M., N. A. Eisele, et al. (2011). "Novel Genetic Tools for Diaminopimelic Acid Selection in Virulence Studies of Yersinia pestis." PloS one 6(3): e17352.

Boccaletti, S., V. Latora, et al. (2006). "Complex networks: Structure and dynamics." Physics reports 424(4): 175-308.

Bonnet, C. (1764). Contemplation de la nature, MM Rey.

Borgatti, S. P. (2005). "Centrality and network flow." Social networks 27(1): 55-71.

Bork, P., T. Dandekar, et al. (1998). "Predicting function: from genes to genomes and back." Journal of molecular biology 283(4): 707-725.

Bowers, P. M., M. Pellegrini, et al. (2004). "Prolinks: a database of protein functional linkages derived from coevolution." Genome Biology 5(5): R35.

Brandes, U. (2001). "A faster algorithm for betweenness centrality*." Journal of Mathematical Sociology 25(2): 163-177.

Brochier, C., E. Bapteste, et al. (2002). "Eubacterial phylogeny based on translational apparatus proteins." TRENDS in Genetics 18(1): 1-5.

Bron, C. and J. Kerbosch (1973). "Algorithm 457: finding all cliques of an undirected graph." Communications of the ACM 16(9): 575-577.

Bruijn, F. J. d. (2011). Handbook of molecular microbial ecology I : metagenomics and complementary approaches. Hoboken, N.J., Wiley-Blackwell.

Bryant, D. and V. Moulton (2004). "Neighbor-net: an agglomerative method for the construction of phylogenetic networks." Molecular Biology and Evolution 21(2): 255-265.

Canchaya, C., G. Fournous, et al. (2004). "The impact of prophages on bacterial chromosomes." Molecular microbiology 53(1): 9-18.

Cazals, F. and C. Karande (2008). "A note on the problem of reporting maximal cliques." Theoretical Computer Science 407(1): 564-568.

Centron, D. and P. H. Roy (2002). "Presence of a group II intron in a multiresistant Serratia marcescens strain that harbors three integrons and a novel gene fusion." Antimicrobial agents and chemotherapy 46(5): 1402-1409.

Chan, C. X., R. G. Beiko, et al. (2009). "Lateral transfer of genes and gene fragments in prokaryotes." Genome Biology and Evolution 1: 429.

Cheng, G., Y. Hu, et al. (2012). "Functional screening of antibiotic resistance genes from human gut microbiota reveals a novel gene fusion." FEMS Microbiology Letters.

Ciccarelli, F. D., T. Doerks, et al. (2006). "Toward automatic reconstruction of a highly resolved tree of life." Science 311(5765): 1283-1287.

Coburn, B., G. A. Grassl, et al. (2006). "Salmonella, the host and disease: a brief review." Immunology and cell biology 85(2): 112-118.

Corpet, F., F. Servant, et al. (2000). "ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons." Nucleic acids research 28(1): 267-269.

Creevey, C. J., D. A. Fitzpatrick, et al. (2004). "Does a tree-like phylogeny only exist at the tips in the prokaryotes?" Proceedings of the Royal Society of London. Series B: Biological Sciences 271(1557): 2551-2558.

Curcio, M. J. and K. M. Derbyshire (2003). "The outs and ins of transposition: from mu to kangaroo." Nature Reviews Molecular Cell Biology 4(11): 865-877.

D'Costa, V. M., C. E. King, et al. (2011). "Antibiotic resistance is ancient." Nature 477(7365): 457-461.

Dagan, T. (2011). "Phylogenomic networks." Trends in microbiology.

Dagan, T., Y. Artzy-Randrup, et al. (2008). "Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution." Proceedings of the National Academy of Sciences 105(29): 10039.

Dagan, T. and W. Martin (2007). "Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution." Proceedings of the National Academy of Sciences 104(3): 870-875.

Darwin, C. (1859). "On the origin of species by means of natural selection, or The Preservation of Favoured Races in the Struggle for Life." London/Die Entstehung der Arten durch naturliche Zuchtwahl, Leipzig oJ.

Daubin, V., N. A. Moran, et al. (2003). "Phylogenetics and the cohesion of bacterial genomes." Science 301(5634): 829-832.

Deka, R. K., M. Machius, et al. (2002). "Crystal structure of the 47-kDa lipoprotein of Treponema pallidum reveals a novel penicillin-binding protein." Journal of Biological Chemistry 277(44): 41857.

DeSantis Jr, T., P. Hugenholtz, et al. (2006). "NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes." Nucleic acids research 34(suppl 2): W394-W399.

Dhingra, A., A. R. Portis Jr, et al. (2004). "Enhanced translation of a chloroplast-expressed RbcS gene restores small subunit levels and photosynthesis in nuclear RbcS antisense plants." Proceedings of the National Academy of Sciences of the United States of America 101(16): 6315-6320.

Dillingham, M. S. and S. C. Kowalczykowski (2008). "RecBCD enzyme and the repair of double-stranded DNA breaks." Microbiology and molecular biology reviews 72(4): 642-671.

Donadio, S., S. Maffioli, et al. (2010). "Antibiotic discovery in the twenty-first century: current trends and future perspectives." The Journal of antibiotics 63(8): 423-430.

Donato, J. J., L. A. Moe, et al. (2010). "Metagenomic analysis of apple orchard soil reveals antibiotic resistance genes encoding predicted bifunctional proteins." Applied and environmental microbiology 76(13): 4396-4401.

Donskey, C. J. (2004). "The role of the intestinal tract as a reservoir and source for transmission of nosocomial pathogens." Clinical infectious diseases 39(2): 219.

Doolittle, W. F. (1999). "Phylogenetic classification and the universal tree." Science 284(5423): 2124-2128.

Doolittle, W. F. and E. Bapteste (2007). "Pattern pluralism and the Tree of Life hypothesis." Proceedings of the National Academy of Sciences 104(7): 2043.

Drummond, D. A. and C. O. Wilke (2008). "Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution." Cell 134(2): 341-352.

Dubey, G. P. and S. Ben-Yehuda (2011). "Intercellular nanotubes mediate bacterial communication." Cell 144(4): 590-600.

Duckworth, D. H. (1976). "" Who discovered bacteriophage?"." Bacteriological reviews 40(4): 793.

Dunn, C. W., A. Hejnol, et al. (2008). "Broad phylogenomic sampling improves resolution of the animal tree of life." Nature 452(7188): 745-749.

Durbin, R., S. R. Eddy, et al. (1998). Biological sequence analysis: probabilistic models of proteins and nucleic acids, Cambridge university press.

Durrens, P., M. Nikolski, et al. (2008). "Fusion and fission of genes define a metric between fungal genomes." PLoS computational biology 4(10): e1000200.

Ehrenberg, C. G. (1828). "Symbolae physicae-Animalia evertebrata." Phytozoa. Berlin (not seen, quoted from Sperber, 1948).[Links].

Enright, A. J., I. Iliopoulos, et al. (1999). "Protein interaction maps for complete genomes based on gene fusion events." Nature 402(6757): 86-90.

Enright, A. J. and C. A. Ouzounis (2001). "Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions." Genome Biol 2(9): 1ñ0034.0037.

Eppstein, D. and D. Strash (2011). "Listing all maximal cliques in large sparse real-world graphs." Experimental Algorithms: 364-375.

Ereshefsky, M. (2010). "Microbiology and the species problem." Biology and Philosophy 25(4): 553-568.

Escobar-Paramo, P., C. Giudicelli, et al. (2003). "The evolutionary history of Shigella and enteroinvasive Escherichia coli revised." Journal of molecular evolution 57(2): 140-148.

Ferretti, J. J., K. Gilmore, et al. (1986). "Nucleotide sequence analysis of the gene specifying the bifunctional 6'-aminoglycoside acetyltransferase 2"-aminoglycoside phosphotransferase enzyme in Streptococcus faecalis and identification and cloning of gene regions specifying the two activities." Journal of bacteriology 167(2): 631-638.

Fleischmann, R. D., M. D. Adams, et al. (1995). "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd." Science 269(5223): 496-512.

Fondi, M., G. Bacci, et al. (2010). "Exploring the evolutionary dynamics of plasmids: the Acinetobacter pan-plasmidome." BMC Evolutionary Biology 10(1): 59.

Foster, T. J. (2004). "The Staphylococcus aureus "superbug"." J Clin Invest 114(12): 1693-1696.

Fraser, C. M., J. D. Gocayne, et al. (1995). "The minimal gene complement of Mycoplasma genitalium." Science 270(5235): 397-404.

Frost, L. S., R. Leplae, et al. (2005). "Mobile genetic elements: the agents of open source evolution." Nature reviews microbiology 3(9): 722-732.

Galperin, M. Y. and E. V. Koonin (2000). "Who's your neighbor? New computational approaches for functional genomics." Nature biotechnology 18(6): 609-613.

Galtier, N. and V. Daubin (2008). "Dealing with incongruence in phylogenomic analyses." Philosophical Transactions of the Royal Society B: Biological Sciences 363(1512): 4023-4029.

Ge, F., L. S. Wang, et al. (2005). "The cobweb of life revealed by genome-scale estimates of horizontal gene transfer." PLoS biology 3(10): e316.

Gest, H. (2004). "The discovery of microorganisms by Robert Hooke and Antoni Van Leeuwenhoek, fellows of the Royal Society." Notes and Records of the Royal Society of London 58(2): 187-201.

Glenn, T. C. (2011). "Field guide to next‚Äêgeneration DNA sequencers." Molecular Ecology Resources.

Gogarten, J. P., W. F. Doolittle, et al. (2002). "Prokaryotic evolution in light of gene transfer." Molecular Biology and Evolution 19(12): 2226-2238.

Gogarten, J. P., H. Kibak, et al. (1989). "Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes." Proceedings of the National Academy of Sciences 86(17): 6661-6665.

Greenblum, S., P. J. Turnbaugh, et al. (2012). "Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. ." Proceedings of the National Academy of Sciences 109.2: 594-599.

Griffith, F. (1928). "The significance of pneumococcal types." Journal of Hygiene 27(02): 113-159.

Haeckel, E. H. P. A. (1866). Generelle Morphologie der Organismen: allgemeine Grundzuge der organischen Formen-Wissenschaft, mechanisch begrundet durch die von Charles Darwin reformirte Descendenz-Theorie, G. Reimer.

Hagberg, A., P. Swart, et al. (2008). Exploring network structure, dynamics, and function using NetworkX, Los Alamos National Laboratory (LANL).

Haggerty, L. S., F. J. Martin, et al. (2009). "Gene and genome trees conflict at many levels." Philosophical Transactions of the Royal Society B: Biological Sciences 364(1527): 2209-2219.

Halary, S., J. W. Leigh, et al. (2010). "Network analyses structure genetic diversity in independent genetic worlds." Proceedings of the National Academy of Sciences 107(1): 127-132.

Hayes, W. (1953). "Observations on a transmissible agent determining sexual differentiation in Bacterium coli." Journal of general microbiology 8(1): 72-88.

Hendrix, R. W. (2003). "Bacteriophage genomics." Current opinion in microbiology 6(5): 506-511.

Hermann, J. (1783). Tabula affinitatum animalium.

Hilario, E. and J. P. Gogarten (1993). "Horizontal transfer of ATPase genes--the tree of life becomes a net of life." Biosystems 31(2-3): 111-119.

Holden, N., L. Pritchard, et al. (2009). "Colonization outwith the colon: plants as an alternative environmental reservoir for human pathogenic enterobacteria." FEMS microbiology reviews 33(4): 689-703.

Hopcroft, J. and R. Tarjan (1973). "Algorithm 447: efficient algorithms for graph manipulation." Communications of the ACM 16(6): 372-378.

Huson, D. H. and D. Bryant (2006). "Application of phylogenetic networks in evolutionary studies." Molecular Biology and Evolution 23(2): 254-267.

Huson, D. H. and C. Scornavacca (2011). "A survey of combinatorial methods for phylogenetic networks." Genome Biology and Evolution 3: 23.

Huynen, M., B. Snel, et al. (2000). "Predicting protein function by genomic context: quantitative evaluation and qualitative inferences." Genome research 10(8): 1204-1210.

Inagaki, Y., E. Susko, et al. (2006). "Recombination between elongation factor 1α genes from distantly related archaeal lineages." Proceedings of the National Academy of Sciences of the United States of America 103(12): 4528-4533.

Jeong, H., S. P. Mason, et al. (2001). "Lethality and centrality in protein networks." Nature 411(6833): 41-42.

Jeong, H., B. Tombor, et al. (2000). "The large-scale organization of metabolic networks." Nature 407(6804): 651-654.

Jin, E. M., M. Girvan, et al. (2001). "Structure of growing social networks." Physical review E 64(4): 046132.

Jones, D. T. W., S. Kocialkowski, et al. (2008). "Tandem duplication producing a novel oncogenic BRAF fusion gene defines the majority of pilocytic astrocytomas." Cancer research 68(21): 8673-8677.

Karp, P. D., M. Riley, et al. (1999). "Eco Cyc: encyclopedia of Escherichia coli genes and metabolism." Nucleic acids research 27(1): 55-58.

Kedes, L. (2011). "The new date, new format, new goals and new sponsor of the Archon Genomics X PRIZE Competition." Nature genetics 43(11): 1055-1058.

Khalil, A. S. and J. J. Collins (2010). "Synthetic biology: applications come of age." Nature Reviews Genetics 11(5): 367-379.

Klebs, G. (1892). Flagellatenstudien 1. U. 2. Theil, Engelmann.

Kloesges, T., O. Popa, et al. (2011). "Networks of gene sharing among 329 proteobacterial genomes reveal differences in lateral gene transfer frequency at different phylogenetic depths." Molecular Biology and Evolution 28(2): 1057-1074.

Koonin, E. V. (2009). "Darwinian evolution in the light of genomics." Nucleic acids research 37(4): 1011-1034.

Koonin, E. V., A. R. Mushegian, et al. (1996). "Sequencing and analysis of bacterial genomes." Current Biology 6(4): 404-416.

Kotloff, K., J. Winickoff, et al. (1999). "Global burden of Shigella infections: implications for vaccine development and implementation of control strategies." Bulletin of the World Health Organization 77(8): 651-666.

Kowalczykowski, S. C., D. A. Dixon, et al. (1994). "Biochemistry of homologous recombination in Escherichia coli." Microbiological reviews 58(3): 401.

Kummerfeld, S. K. and S. A. Teichmann (2005). "Relative rates of gene fusion and fission in multi-domain proteins." TRENDS in Genetics 21(1): 25-30.

Kunin, V., L. Goldovsky, et al. (2005). "The net of life: reconstructing the microbial phylogenetic network." Genome research 15(7): 954-959.

Kurzrock, R., H. M. Kantarjian, et al. (2003). "Philadelphia chromosome-positive leukemias: from basic mechanisms to molecular therapeutics." Annals of internal medicine 138(10): 819.

Lamarck, J. (1809). Philosophie zoologique, ou Exposition des considerations relatives l'histoire naturelle des animaux Paris: Dentu.

Lamarck, J. B. (1815-1822). "Histoire naturelle des animaux sans vertebres. 7 vols." Verdieere, Paris.

Lander, E. S., L. M. Linton, et al. (2001). "Initial sequencing and analysis of the human genome." Nature 409(6822): 860-921.

Lang, A. S. and J. Beatty (2000). "Genetic analysis of a bacterial genetic exchange element: the gene transfer agent of Rhodobacter capsulatus." Proceedings of the National Academy of Sciences 97(2): 859-864.

Lawrence, J. G. and A. C. Retchless (2009). "The interplay of homologous recombination and horizontal gene transfer in bacterial speciation." Methods Mol Biol 532: 29-53.

Lederberg, J. (1998). "Emerging infections: an evolutionary perspective." Emerging infectious diseases 4(3): 366.

Lederberg, J. and E. L. Tatum (1946). "Gene recombination in Escherichia coli." Nature 158: 558.

Lima-Mendez, G., J. Van Helden, et al. (2008). "Reticulate representation of evolutionary and functional relationships between phage genomes." Molecular Biology and Evolution 25(4): 762-777.

Linnaeus, C. (1774). "Systema Naturae (Laurentius Salvius, Stockholm)." 10th Ed.

Livermore, D. M. (2005). "Minimising antibiotic resistance." The Lancet infectious diseases 5(7): 450-459.

Livermore, D. M. (2009). "Has the era of untreatable infections arrived?" Journal of Antimicrobial Chemotherapy 64(suppl 1): i29-i36.

Loveland, J. (2004). "Georges-Louis Leclerc de Buffon's Histoire naturelle in English, 1775-1815." Archives of natural history 31(2): 214-235.

Lowy, F. D. (2003). "Antimicrobial resistance: the example of Staphylococcus aureus." Journal of Clinical Investigation 111(9): 1265-1274.

Madigan, M. T. and T. D. Brock (2011). Brock biology of microorganisms. San Francisco, Calif., [etc.], Pearson.

Maher, C. A., C. Kumar-Sinha, et al. (2009). "Transcriptome sequencing to detect gene fusions in cancer." Nature 458(7234): 97-101.

Majewski, J. and F. M. Cohan (1999). "DNA sequence similarity requirements for interspecific recombination in Bacillus." Genetics 153(4): 1525-1533.

Majewski, J., P. Zawadzki, et al. (2000). "Barriers to genetic exchange between bacterial species: Streptococcus pneumoniae transformation." Journal of bacteriology 182(4): 1016-1023.

Marcotte, E. M., M. Pellegrini, et al. (1999). "Detecting protein function and protein-protein interactions from genome sequences." Science 285(5428): 751.

Mardis, E. R. (2008). "The impact of next-generation sequencing technology on genetics." TRENDS in Genetics 24(3): 133-141.

Margulies, M., M. Egholm, et al. (2005). "Genome sequencing in microfabricated high-density picolitre reactors." Nature 437(7057): 376-380.

Mazumdar, P. M. H. (2002). Species and specificity : an interpretation of the history of immunology. Cambridge, Cambridge University Press.

McClelland, M., K. E. Sanderson, et al. (2004). "Comparison of genome degradation in Paratyphi A and Typhi, human-restricted serovars of Salmonella enterica that cause typhoid." Nature genetics 36(12): 1268-1274.

McDaniel, L. D., E. Young, et al. (2010). "High frequency of horizontal gene transfer in the oceans." Science 330(6000): 50-50.

McInerney, J., C. Cummins, et al. (2011). "Goods-thinking vs. tree-thinking: Finding a place for mobile genetic elements." Mobile Genetic Elements 1(4): 304-343.

McInerney, J. O., J. A. Cotton, et al. (2008). "The prokaryotic tree of life: past, present, and future?" Trends in ecology & evolution 23(5): 276-281.

McInerney, J. O., D. Pisani, et al. (2011). "The public goods hypothesis for the evolution of life on Earth." Biol Direct 6: 41.

Merhej, V., C. Notredame, et al. (2011). "The rhizome of life: the sympatric Rickettsia felis paradigm demonstrates the random transfer of DNA sequences." Molecular Biology and Evolution 28(11): 3213-3223.

Michel, B., H. Boubakri, et al. (2007). "Recombination proteins and rescue of arrested replication forks." DNA repair 6(7): 967-980.

Milgram, S. (1967). "The small world problem." Psychology today 2(1): 60-67.

Mingeot-Leclercq, M. P., Y. Glupczynski, et al. (1999). "Aminoglycosides: activity and resistance." Antimicrobial agents and chemotherapy 43(4): 727-737.

Mirkin, B. G., T. I. Fenner, et al. (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 Evolutionary Biology 3(1): 2.

Moeller, R., E. Stackebrandt, et al. (2007). "Role of DNA repair by nonhomologous-end joining in Bacillus subtilis spore resistance to extreme dryness, mono-and polychromatic UV, and ionizing radiation." Journal of bacteriology 189(8): 3306-3311.

Morimatsu, K. and S. C. Kowalczykowski (2003). "RecFOR proteins load RecA protein onto gapped DNA to accelerate DNA strand exchange: a universal step of recombinational repair." Molecular cell 11(5): 1337-1347.

Mukhopadhyay, R. (2009). "DNA sequencers: The next generation." Analytical Chemistry 81(5): 1736-1740.

Nakamura, Y., T. Itoh, et al. (2004). "Biased biological functions of horizontally transferred genes in prokaryotic genomes." Nature genetics 36(7): 760-766.

Nelson, D. (2004). "Phage taxonomy: we agree to disagree." Journal of bacteriology 186(21): 7029-7031.

Newman, M. (2010). Networks an introduction. Oxford, Oxford Univ. Press.

Newman, M. E. J. (2004). "Detecting community structure in networks." The European Physical Journal B-Condensed Matter and Complex Systems 38(2): 321-330.

Orgel, L. E. and F. H. Crick (1980). "Selfish DNA: the ultimate parasite." Nature 284(5757): 604.

Pallas, P. S. (1766). Elenchus zoophytorum, van Cleef.

Pasek, S., J. L. Risler, et al. (2006). "Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins." Bioinformatics 22(12): 1418-1423.

Patthy, L. (2008). Protein evolution, Wiley-Blackwell.

Pellegrini, M., D. Haynor, et al. (2004). "Protein interaction networks." Expert review of proteomics 1(2): 239-249.

Philippe, H. (2000). "Early-branching or fast-evolving eukaryotes? An answer based on slowly evolving positions." Proceedings of the Royal Society of London. Series B: Biological Sciences 267(1449): 1213-1221.

Phizicky, E. M. and S. Fields (1995). "Protein-protein interactions: methods for detection and analysis." Microbiological reviews 59(1): 94-123.

Popa, O., E. Hazkani-Covo, et al. (2011). "Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes." Genome research 21(4): 599-609.

Porter, M. A., J. P. Onnela, et al. (2009). "Communities in networks." Notices of the AMS 56(9): 1082-1097.

Puigbo, P. and E. V. Koonin (2009). "Search for a Tree of Life in the thicket of the phylogenetic forest " Journal of biology 8: 59.

Pupo, G. M., R. Lan, et al. (2000). "Multiple independent origins of Shigella clones of Escherichia coli and convergent evolution of many of their characteristics." Proceedings of the National Academy of Sciences 97(19): 10567.

Ragan, M. (2009). "Trees and networks before and after Darwin." Biology Direct 4(1): 43.

Rambaut, A. and N. C. Grass (1997). "Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees." Computer applications in the biosciences: CABIOS 13(3): 235-238.

Rammelkamp, C. H. and T. Maxon (1942). Resistance of Staphylococcus aureus to the Action of Penicillin, Royal Society of Medicine.

Rapaka, R. R., E. S. Goetzman, et al. (2007). "Enhanced defense against Pneumocystis carinii mediated by a novel dectin-1 receptor Fc fusion protein." The Journal of Immunology 178(6): 3702.

Reddy, T. P. J. (2007). Metabolic Network Analysis, Daya Books.

Retchless, A. C. and J. G. Lawrence (2010). "Phylogenetic incongruence arising from fragmented speciation in enteric bacteria." Proceedings of the National Academy of Sciences 107(25): 11453.

Rivera, C. G., R. Vakil, et al. (2010). "NeMo: network module identification in Cytoscape." BMC bioinformatics 11(Suppl 1): S61.

Rolinson, G. (1961). ""Celbenin"-resistant staphylococci." British Medical Journal 1(5219): 125-126.

Rosen, D. A., T. M. Hooton, et al. (2007). "Detection of intracellular bacterial communities in human urinary tract infection." PLoS medicine 4(12): e329.

Rowley, J. D. and W. S. Beck (1973). "Chromosomal patterns in myelocytic leukemia." New England Journal of Medicine 289(4): 220-221.

Ruan, Y., H. S. Ooi, et al. (2007). "Fusion transcripts and transcribed retrotransposed loci discovered through comprehensive transcriptome analysis using Paired-End diTags (PETs)." Genome research 17(6): 828-838.

Ruling, J. P. (1774). Ordines Naturales Plantarum: Commentatio Botanica, Vandenhoeck.

Sali, A. (1999). "Functional links between proteins." Nature 402(23): 25ñ26.

Salzberg, S. L., O. White, et al. (2001). "Microbial genes in the human genome: lateral transfer or gene loss?" Science 292(5523): 1903-1906.

Samuelson, P. A. (1954). "The pure theory of public expenditure." The review of economics and statistics 36(4): 387-389.

Sanger, F., G. Air, et al. (1977). "Nucleotide sequence of bacteriophage φ X174 DNA."

Sanger, F. and A. R. Coulson (1975). "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase." Journal of molecular biology 94(3): 441-448.

Sanger, F., S. Nicklen, et al. (1977). "DNA sequencing with chain-terminating inhibitors." Proceedings of the National Academy of Sciences 74(12): 5463-5467.

Sansonetti, P. J., D. J. Kopecko, et al. (1981). "Shigella sonnei plasmids: evidence that a large plasmid is necessary for virulence." Infection and Immunity 34(1): 75-83.

Sapp, J. (2005). "The prokaryote-eukaryote dichotomy: meanings and mythology." Microbiology and molecular biology reviews 69(2): 292-305.

Sapp, J. (2009). The new foundations of evolution : on the tree of life. New York, Oxford University Press.

Schaechter, M., R. Kolter, et al. (2004). Microbiology in the 21st Century: Where are We and where are We Going?, American Academy of Microbiology.

Scott, S. and C. J. Duncan (2001). Biology of plagues : evidence from historical populations. Cambridge [u.a.], Cambridge Univ. Press.

Shannon, P., A. Markiel, et al. (2003). "Cytoscape: a software environment for integrated models of biomolecular interaction networks." Genome research 13(11): 2498-2504.

Shoemaker, B. A. and A. R. Panchenko (2007). "Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners." PLoS computational biology 3(4): e43.

Smith, N. R. and R. Gordon (1957). " Bergey's Manual of Determinative Bacteriology." 7th ed.: (1957): 1613-1634.

Snel, B., P. Bork, et al. (2000). "Genome evolution-gene fusion versus gene fission." TRENDS in Genetics 16: 9-11.

Soffer, S. N. and A. Vazquez (2005). "Network clustering coefficient without degree-correlation biases." Physical review E 71(5): 057101.

Sommer, M. O. A., G. Dantas, et al. (2009). "Functional characterization of the antibiotic resistance reservoir in the human microflora." Science 325(5944): 1128-1131.

Sorek, R., Y. Zhu, et al. (2007). "Genome-wide experimental determination of barriers to horizontal gene transfer." Science 318(5855): 1449-1452.

Soulsby, E. J. (2005). "Resistance to antimicrobials in humans and animals." Bmj 331(7527): 1219-1220.

Spellberg, B., R. Guidos, et al. (2008). "The epidemic of antibiotic-resistant infections: a call to action for the medical community from the Infectious Diseases Society of America." Clinical infectious diseases 46(2): 155.

Stackebrandt, E. and B. Goebel (1994). "Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology." International Journal of Systematic Bacteriology 44(4): 846-849.

Stackebrandt, E., R. Murray, et al. (1988). "Proteobacteria classis nov., a name for the phylogenetic taxon that includes the "purple bacteria and their relatives"." International Journal of Systematic Bacteriology 38(3): 321-325.

Stanhope, M. J., A. Lupas, et al. (2001). "Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates." Nature 411(6840): 940-944.

Stanier, R. Y. and C. B. Niel (1962). "The concept of a bacterium." Archives of Microbiology 42(1): 17-35.

Stechmann, A. and T. Cavalier-Smith (2002). "Rooting the eukaryote tree by using a derived gene fusion." Science 297(5578): 89.

Stechmann, A. and T. Cavalier-Smith (2003). "The root of the eukaryote tree pinpointed." Current Biology 13(17): 665-666.

Suhre, K. and J. M. Claverie (2004). "FusionDB: a database for in-depth analysis of prokaryotic gene fusion events." Nucleic acids research 32(suppl 1): D273-D276.

Szklarczyk, D., A. Franceschini, et al. (2011). "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored." Nucleic acids research 39(suppl 1): D561-D568.

Tenorio, C., M. Zarazaga, et al. (2001). "Bifunctional enzyme 6'-N-aminoglycoside acetyltransferase-2"-O-aminoglycoside phosphotransferase in Lactobacillus and Pediococcus isolates of animal origin." Journal of clinical microbiology 39(2): 824-825.

Thompson, C., F. Thompson, et al. (2004). "Use of recA as an alternative phylogenetic marker in the family Vibrionaceae." International journal of systematic and evolutionary microbiology 54(3): 919-924.

Tomita, E., A. Tanaka, et al. (2006). "The worst-case time complexity for generating all maximal cliques and computational experiments." Theoretical Computer Science 363(1): 28-42.

Townsend, J. P., K. M. Nielsen, et al. (2003). "Horizontal acquisition of divergent chromosomal DNA in bacteria: effects of mutator phenotypes." Genetics 164(1): 13-21.

Tsoka, S. and C. A. Ouzounis (2000). "Prediction of protein interactions: metabolic enzymes are frequently involved in gene fusion." Nature genetics 26(2): 141.

Tsukiyama, S., M. Ide, et al. (1977). "A new algorithm for generating all the maximal independent sets." SIAM Journal on Computing 6: 505.

Veitia, R. A. (2002). "Rosetta Stone proteins:" chance and necessity"?" Genome Biol 3(2).

Vos, M. (2009). "Why do bacteria engage in homologous recombination?" Trends in microbiology 17(6): 226-232.

Wang, Y. and Z. Zhang (2000). "Comparative sequence analyses reveal frequent occurrence of short segments containing an abnormally high number of non-random base variations in bacterial rRNA genes." Microbiology 146(11): 2845.

Weller, G. R., B. Kysela, et al. (2002). "Identification of a DNA nonhomologous end-joining complex in bacteria." Science 297(5587): 1686-1689.

Williams, D., G. P. Fournier, et al. (2011). "A rooted net of life." Biology Direct 6(1): 45.

Williams, K. P., J. J. Gillespie, et al. (2010). "Phylogeny of gammaproteobacteria." Journal of bacteriology 192(9): 2305-2314.

Winogradsky, S. (1952). On the classification of bacteria. Annales de l'Institut Pasteur.

Wise, R. (2004). "The relentless rise of resistance?" Journal of Antimicrobial Chemotherapy 54(2): 306-310.

Woese, C. R. (1987). "Bacterial evolution." Microbiological reviews 51(2): 221.

Woese, C. R., O. Kandler, et al. (1990). "Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya." Proceedings of the National Academy of Sciences 87(12): 4576.

Worboys, M. (2000). Spreading germs : disease theories and medical practice in Britain, 1865-1900. Cambridge [u.a.], Cambridge Univ. Press.

Worobey, M. (2001). "A novel approach to detecting and measuring recombination: new insights into evolution in viruses, bacteria, and mitochondria." Molecular Biology and Evolution 18(8): 1425-1434.

Yanai, I., A. Derti, et al. (2001). "Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes." Proceedings of the National Academy of Sciences 98(14): 7940.

Yanai, I., Y. I. Wolf, et al. (2002). "Evolution of gene fusions: horizontal transfer versus independent events." Genome Biol 3(5): 1ñ0024.0013.

Yang, J., H. Nie, et al. (2007). "Revisiting the molecular evolutionary history of Shigella spp." Journal of molecular evolution 64(1): 71-79.

Zakharova, N., B. J. Paster, et al. (1999). "Fused and Overlapping rpoB andrpoC Genes in Helicobacters, Campylobacters, and Related Bacteria." Journal of bacteriology 181(12): 3857-3859.

Zhang, B. (2009). "Protein interaction network."

Zhang, W., J. F. Fisher, et al. (2009). "The bifunctional enzymes of antibiotic resistance." Current opinion in microbiology 12(5): 505-511.

Zhaxybayeva, O., J. P. Gogarten, et al. (2006). "Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events." Genome research 16(9): 1099-1108.

Zinder, N. D. and J. Lederberg (1952). "Genetic exchange in Salmonella." Journal of bacteriology 64(5): 679.

Zuckerkandl, E. and L. Pauling (1965). "Molecules as documents of evolutionary history." Journal of theoretical biology 8(2): 357-366.

List of Web Links

1. http://www.polonator.org/vision.aspx
2. http://www.nanoporetech.com/news/press-releases/view/39
3. http://www.lifetechnologies.com/global/en/home/about-us/news-gallery/press-releases/2012/life-techologies-itroduces-the-bechtop-io-proto.html.html
4. http://www.genomesonline.org/cgi-bin/GOLD/index.cgi?page_requested=Statistics
5. http://ecoliwiki.net/colipedia/index.php/Sequenced_E._coli_Genomes
6. http://www.genome.wisc.edu/
7. http://www.igs.cnrs-mrs.fr/FusionDB/fmatrix.html#C-C
8. http://www.ncbi.nlm.nih.gov/
9. ftp://saf.bio.caltech.edu/pub/software/molbio/fastasplitn.c
10. http://fasta.bioch.virginia.edu/fasta_www2/fasta_list2.shtml
11. http://apps.cytoscape.org/apps/nemo

Appendix

Table A1: Genome and accession number in GenBank (Chapter 2).

Species	Accession Number
Escherichia coli K-12 MG1655	NC_000913

Table A2: List of genomes, their accession number in GenBank and their short names (Chapter 4).

Species	Accession Number	Short Name
Escherichia coli 536	NC_008253	ECO536
Escherichia coli 55989	NC_011748	ECO55989
Escherichia coli APEC O1	NC_008563	ECOAPEC01
Escherichia coli BL21	NC_012892	ECONEW
Escherichia coli BL21(DE3)	NC_012947	ECOBL21DE3
Escherichia coli BW2952	NC_012759	ECOBW2952
Escherichia coli ATCC 8739	NC_010468	ECOCATCC
Escherichia coli CFT073	NC_004431	ECOCFT073
Escherichia coli E24377A	NC_009801	ECOE24377A
Escherichia coli O157:H7 str. EC4115	NC_011353	ECOEC4115
Escherichia coli ED1a	NC_011745	ECOED1A
Escherichia coli O157:H7 EDL933	NC_002655	ECOEDL933
Escherichia coli O127:H6 str. E2348/69	NC_011601	ECOH6E2348
Escherichia coli HS	NC_009800	ECOHS
Escherichia coli IAI1	NC_011741	ECOIAI1

Escherichia coli IAI39	NC_011750	ECOIAI39
Escherichia coli str. K-12 substr. DH10B	NC_010473	ECOKDH10B
Escherichia coli str. K-12 substr. MG1655	NC_000913	ECOKMG1655
Escherichia coli str. K-12 substr. W3110	AC_000091	ECOKW3110
Escherichia coli LF82	NC_011993	ECOLF82
Escherichia coli B str. REL606	NC_012967	ECOREL606
Escherichia coli S88	NC_011742	ECOS88
Escherichia coli O157:H7 str. Sakai	NC_002695	ECOSAKAI
Escherichia coli SE11	NC_011415	ECOSE11
Escherichia coli SMS-3-5	NC_010498	ECOSMS35
Escherichia coli O157:H7 str. TW14359	NC_013008	ECOTW4359
Escherichia coli UMN026	NC_011751	ECOUMN026
Escherichia coli UTI89	NC_007946	ECOUTI89
Escherichia fergusonii ATCC 35469	NC_011740	EFERATCC
Salmonella enterica subsp. enterica serovar Agona str. SL483,	NC_011149	SALENAGONA
Salmonella enterica subsp. arizonae serovar 62:z4,z23:--, complete	NC_010067	SALENARIZ
Salmonella enterica subsp. enterica serovar Paratyphi A str. ATCC	NC_006511	SALENATCC
Salmonella enterica subsp. enterica serovar Choleraesuis str.	NC_006905	SALENCHOL
Salmonella enterica subsp. enterica serovar Dublin str.	NC_011205	SALENDUB
Salmonella enterica subsp. enterica serovar Enteritidis str.	NC_011294	SALENENTER
Salmonella enterica subsp. enterica serovar Gallinarum str. 287/91,	NC_011274	SALENGAL
Salmonella enterica subsp. enterica serovar Heidelberg str. SL476,	NC_011083	SALENHEID
Salmonella enterica subsp. enterica serovar Newport str. SL254,	NC_011080	SALENNEW

Salmonella enterica subsp. enterica serovar Paratyphi A str.	NC_011147	SALENPARAA
Salmonella enterica subsp. enterica serovar Paratyphi B str. SPB7,	NC_010102	SALENPARAB
Salmonella enterica subsp. enterica serovar Paratyphi C strain	NC_012125	SALENPARAC
Salmonella enterica subsp. enterica serovar Schwarzengrund str.	NC_011094	SALENSCHW
Salmonella enterica subsp. enterica serovar Typhi Ty2, complete	NC_004631	SALENSTY
Salmonella typhimurium LT2	NC_003197	SALTYLT2
Salmonella enterica subsp. enterica serovar Typhi str. CT18,	NC_003198	SALTYPHI
Shigella sonnei Ss046	NC_007384	SHIGSON
Shigella flexneri 5 str. 8401	NC_008258	SHIGFLEX5
Shigella flexneri 2a str. 301	NC_004337	SHIGF301
Shigella flexneri 2a str. 2457T	NC_004741	SHIGF245
Shigella dysenteriae Sd197	NC_007606	SHIGDYS
Shigella boydii CDC 3083-94	NC_010658	SHIGBOYDCDC
Shigella boydii Sb227	NC_007613	SHIGBOY227
Erwinia carotovora subsp. atroseptica SCRI1043	NC_004547	PECTCARATR
Pectobacterium carotovorum subsp. carotovorum PC1	NC_012917	PECTCARPC1
Yersinia enterocolitica subsp. enterocolitica 8081, complete	NC_008800	YERSEN8081
Yersinia pestis Angola	NC_010159	YERSPANG
Yersinia pestis Antiqua	NC_008150	YERSPANT
Yersinia pestis biovar Microtus str. 91001	NC_005810	YERSPBM
Yersinia pestis CO92	NC_003143	YERSPCO92
Yersinia pestis KIM	NC_004088	YERSPKIM
Yersinia pestis Nepal516	NC_008149	YERSPNEPAL

Yersinia pestis Pestoides F	NC_009381	YERSPPF
Yersinia pseudotuberculosis IP 32953	NC_006155	YERSTB32
Yersinia pseudotuberculosis IP 31758	NC_009708	YERSTBIP31
Yersinia pseudotuberculosis PB1/+	NC_010634	YERSTBPB1
Yersinia pseudotuberculosis YPIII	NC_010465	YERSTBYP3

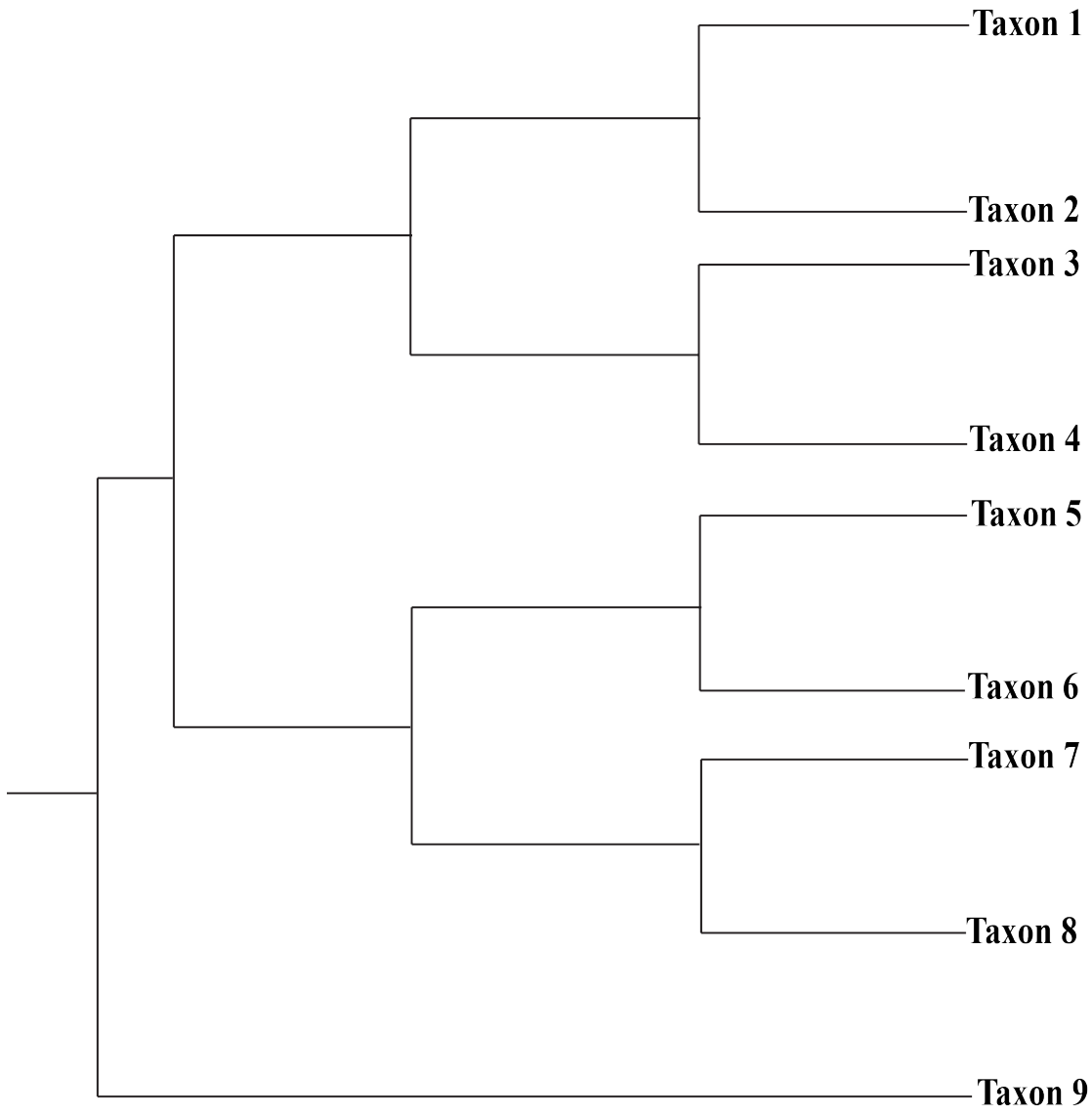


Figure A1: Tree used to simulate nine sequences for eight gene families using SeqGen (Section 2.2.3.1)

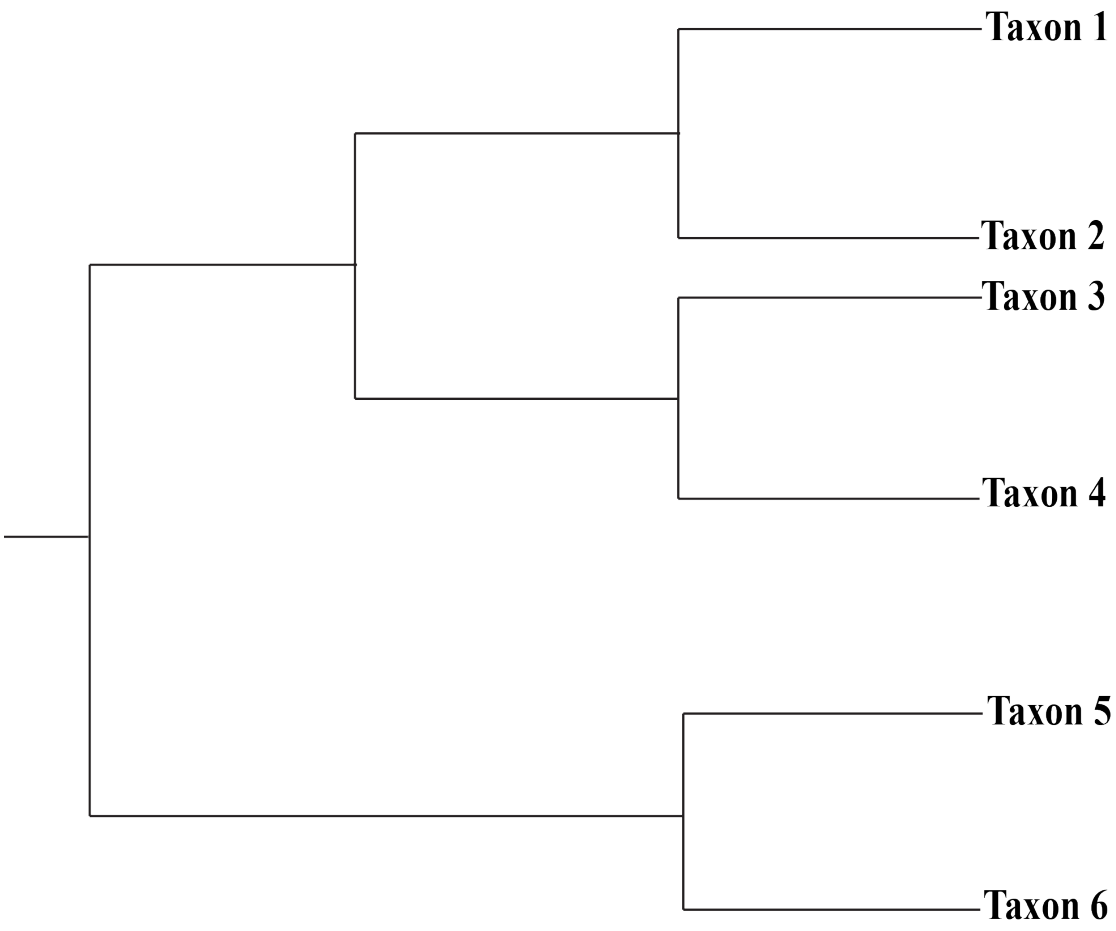


Figure A2: Tree used to simulate six sequences for three fusion genes (Section 2.2.3.1)

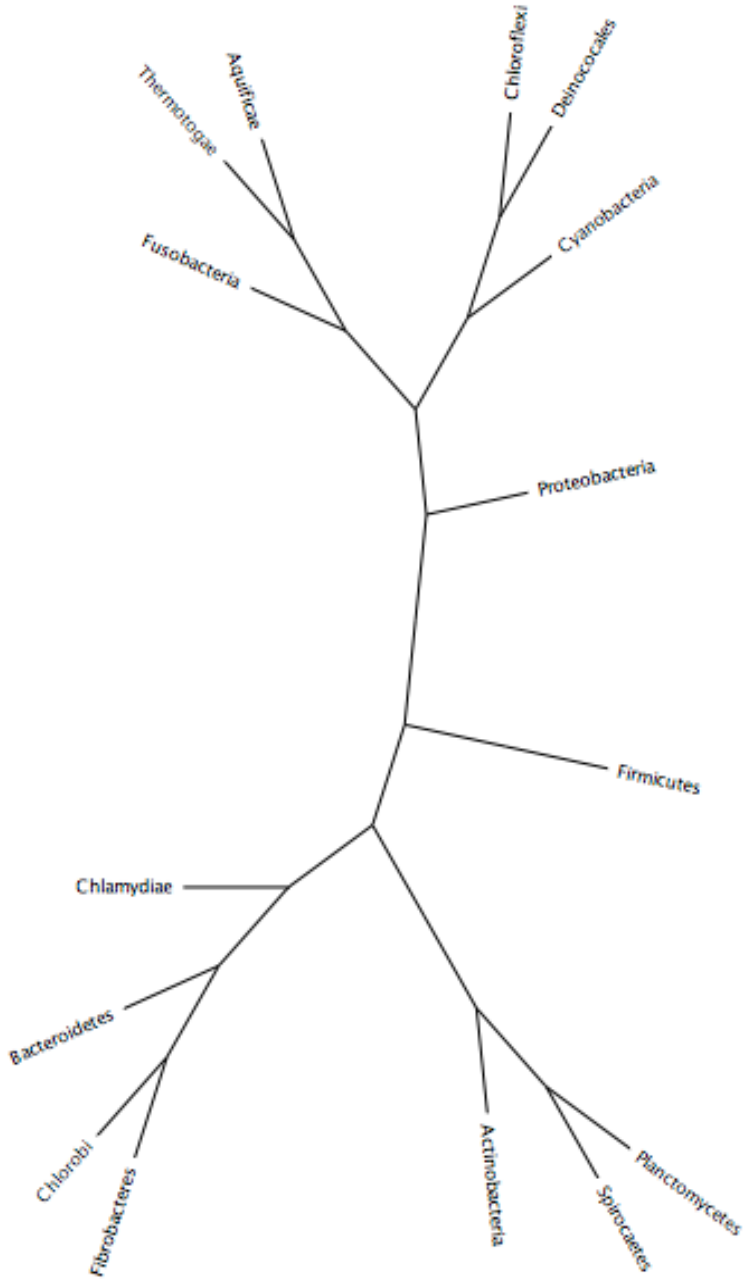


Figure A3: Tree used to guide choice of species in datasets 2, 4 and 5 (see section 3.1) adapted from Ciccarelli *et al.* (2006).

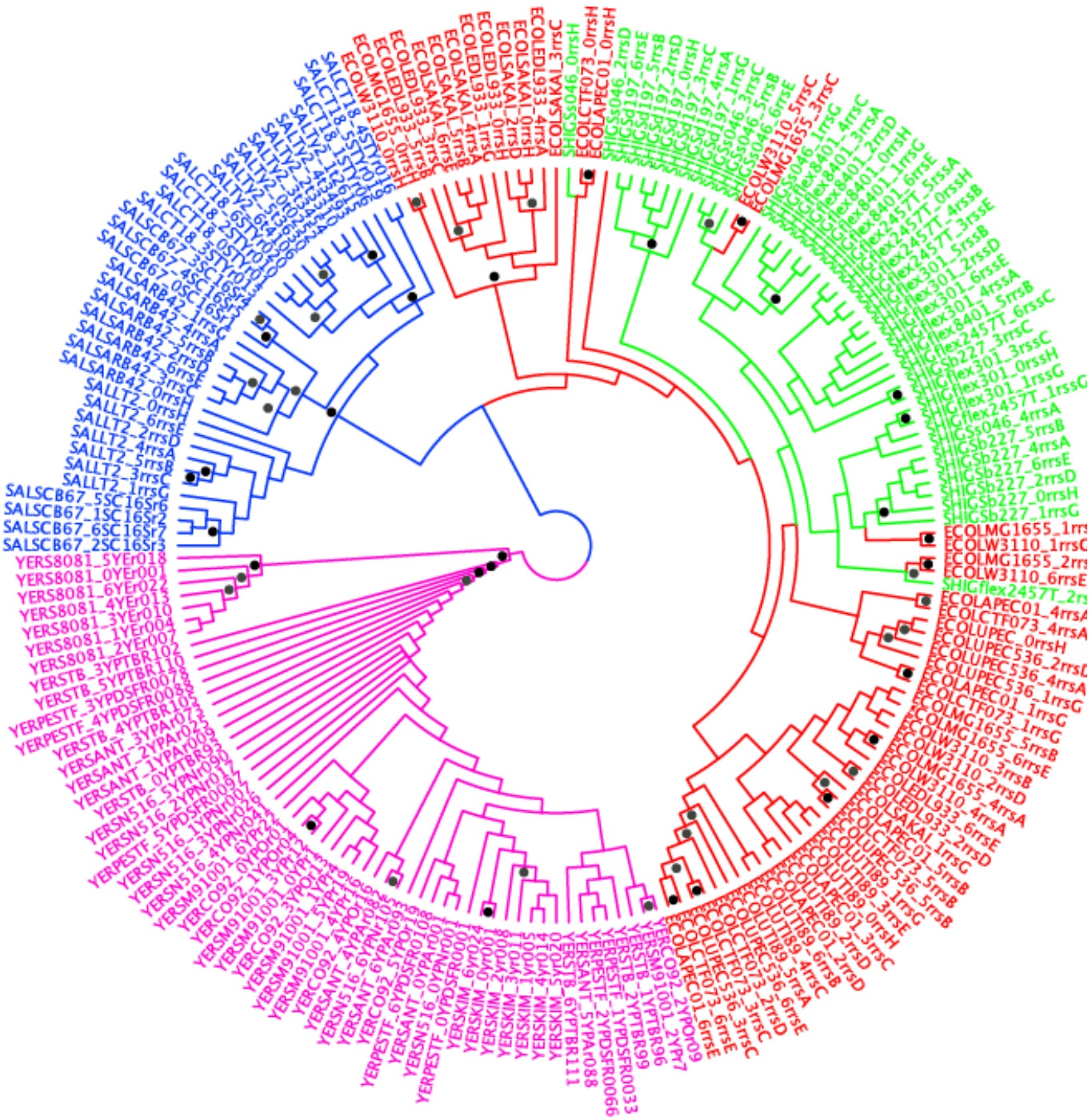


Figure A4: Phylogenetic tree of the YESS group constructed from 187 16S rRNA sequences. Grey nodes denote more than 50 per cent bootstrap support, and black nodes denote more than 70 per cent bootstrap support. Taken from Haggerty *et al.* (2009).

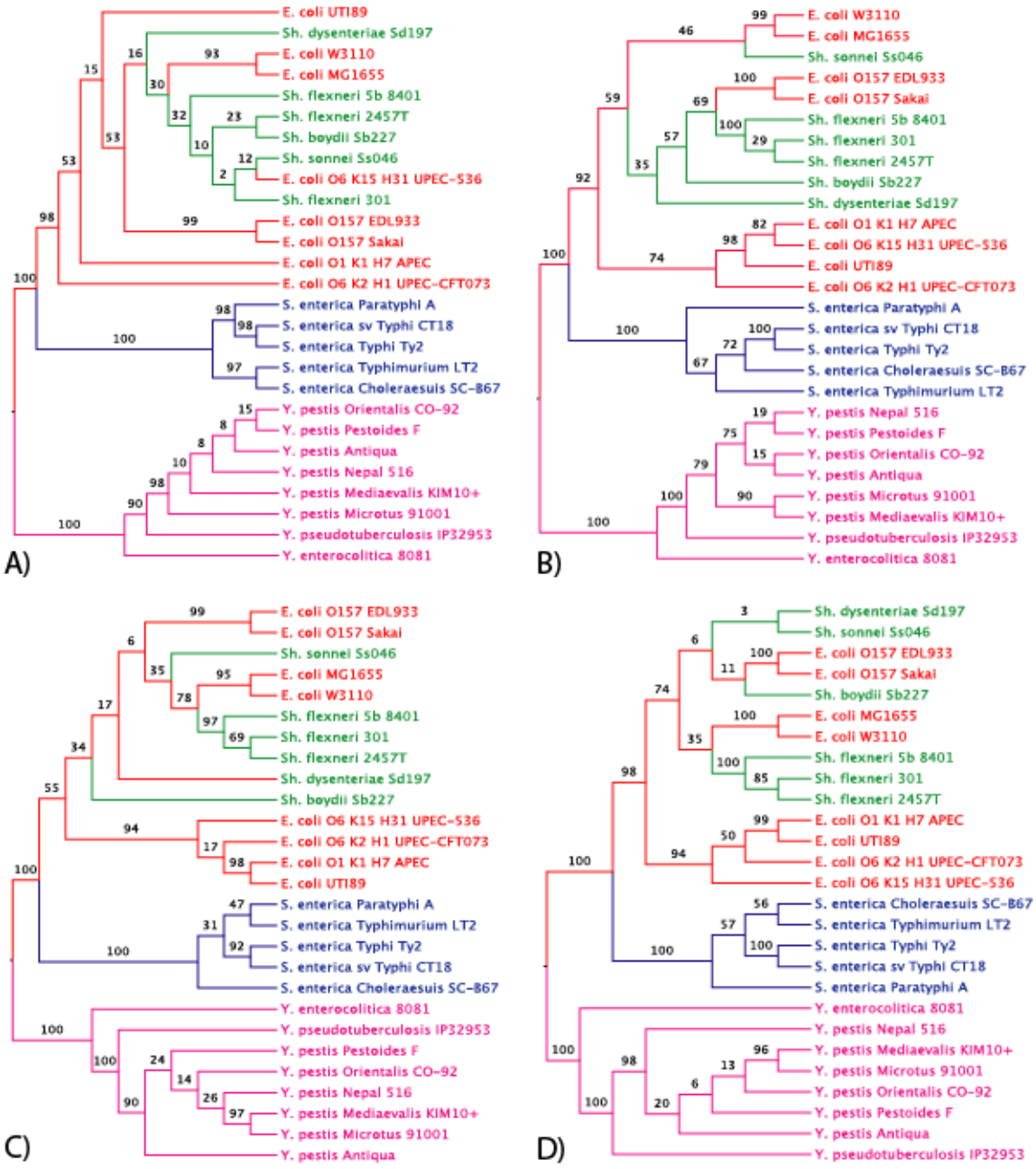


Figure A5: Phylogenetic gene trees for the YESS group. Phylogenetic trees for (A) atpD (B) gyrB, (C) trpB. (D) Phylogenetic tree based on concatenated gene sequences for atpD, gyrB and trpB. Taken from Haggerty *et al*. (2009).

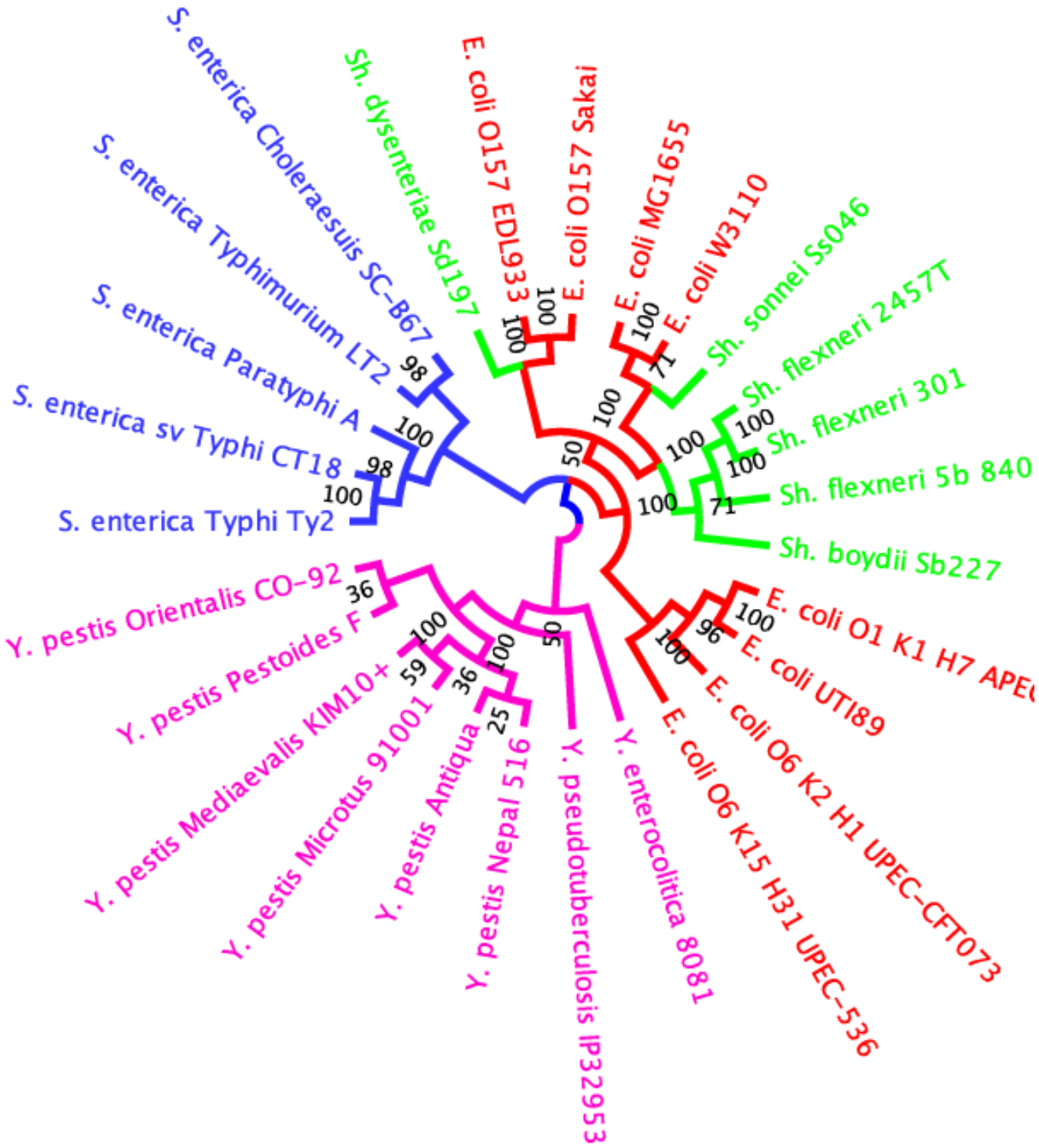


Figure A6: Supertree for the YESS group, constructed from 1408 single-gene families using nucleotide data. Taken from Haggerty *et al*. (2009).

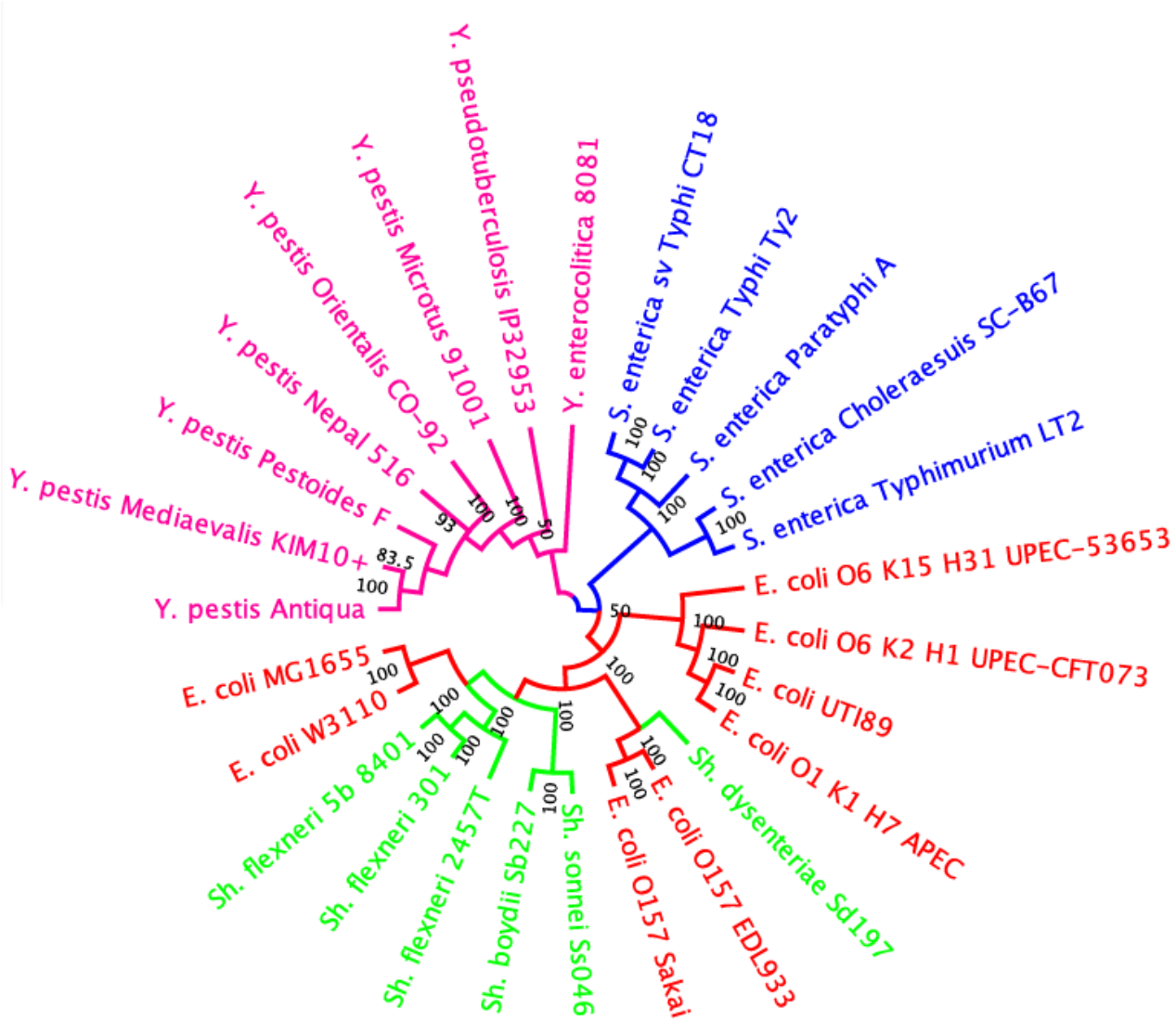


Figure A7: Minimum-evolution tree for the YESS group built from an alignment of 1408 single-gene families. Taken from Haggerty *et al*. (Haggerty *et al.* 2009)