The Proximon: Representation, Evaluation, and Applications of Metagenomic Functional Interactions

by

Gregory Detlev Alexander Vey

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Biology

Waterloo, Ontario, Canada, 2017

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner	John Parkinson, PhD Associate Professor - Biochemistry & Molecular and Medical Genetics University of Toronto
Supervisor	Trevor C. Charles, PhD Professor - Department of Biology University of Waterloo
Internal Member	Christine Dupont, PhD Continuing Lecturer & Biology Teaching Fellow - Department of Biology University of Waterloo
Internal-external Member	Brian Ingalls, PhD Associate Professor - Department of Applied Mathematics University of Waterloo
Other Member	Brendan J. McConkey, PhD Associate Professor - Department of Biology University of Waterloo

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

The effective use of metagenomic functional interactions represents a key prospect for a variety of applications in the field of functional metagenomics. By definition, metagenomic operons represent such interactions but many operon predictions protocols rely on information about orthology and/or gene function that is frequently unavailable for metagenomic genes. In this thesis, I introduce the proposition of the proximon as a unit of functional interaction that is intended for use in metagenomic scenarios where supplemental information is sparse. The proximon is defined as a series of co-directional genes where minimal intergenic distance exists between any two consecutive member genes within the same proximon. In particular, the proximon is presented here as a biological abstraction aimed at facilitating bioinformatics and computational goals. In this thesis, proximons are constructed as information theoretic entities and employed in a variety of contexts related to functional metagenomics. I begin by implementing a computational representation for proximon data and demonstrate its utility through the deployment of a public database. Next, I perform a formal validation where proximons are contrasted against known operons by using the *Escherichia coli* K-12 model organism as a gold standard to measure the extent to which proximons emulate actual operons. This is followed by a demonstration of how proximon data can be applied to infer potential functional networks and depict potential functional modules. I conclude by enumerating the limitations of the research performed here and I present objectives and goals for future work.

Acknowledgements

A myriad of people has impacted the shaping and formation of this thesis in numerous ways and to varying extents. First, I thank my supervisor Trevor Charles for encouraging my diverse research interests and allowing me to practice an interdisciplinary approach to science and also for his enduring support throughout this journey. Next, I thank Christine Dupont for support and service on my advisory committee throughout the duration of my doctoral studies, as well as Brendan McConkey and Brian Ingalls for their service in the later phases. I also thank Gabriel Moreno-Hagelsieb for collaboration and guidance during the initial stage of my doctoral studies. I acknowledge financial support from an NSERC Strategic Grant, an OGSST grant, and the University of Waterloo. In addition, I thank the following people for scholarly interactions relating to my research and/or general support with my PhD studies: Paul Thagard, John Thompson, Chrysanne DiMarco, Britt Anderson, Doreen Fraser, Michael Lynch, Andre Masella, Andrew Doxey, Cheryl Duxbury, Bruce Wolff, Rick Henderson, Mirko Vucicevich, Marc del Grande, Jennifer Collins, Linda Zepf, April Wettig, Juan Javier Díaz-Mejía, Jason Greatrex, and my colleagues at Information Systems & Technology at the University of Waterloo. I give posthumous thanks to Allan Speare. I also extend my gratitude to the many editors and anonymous reviewers who have assisted in the development of the peer-reviewed papers that comprise the core of this thesis. Finally, I thank the members of my family for their ongoing support and encouragement in this undertaking.

Dedication

Dedicated to my mother, Elisabeth Vey B.A., B.Ed., M.A.

Table of Contents

Examining Committee Membership ii

Author's Declaration iii

Abstract iv

Acknowledgements v

Dedication vi

Table of Contents vii

List of Figures x

List of Tables xii

Chapter 1 Introduction 1

1.1 The Rise of the Metagenomic Era 2

1.2 Challenges and Prospects in Metagenomics 6

1.3 Challenges in Context-Based Functional Inference 20

1.4 The Proximon Proposition 24

1.5 Goals and Objectives 26

Chapter 2 Representation: The MetaProx Database 29

2.1 The Big Data Challenge in Computational Biology 30

2.1.1 Dissemination and Representation 30

2.1.2 Large-Scale Data Analysis Protocols 33

2.2 Computational Representation of Proximon Data 33

2.3 Implementation 35

2.3.1 Data Model 39

2.3.2 Data Generation 40

2.4 Deployment 47

2.5 Applications 50

2.6 Future Directions 52

Chapter 3 Evaluation: Mapping Proximons to Operons 54

3.1 Introduction 55

3.2 Methods 57
3.2.1 Genes 57
3.2.2 Proximons 58
3.2.3 Operons 58
3.2.4 Metrics 59
3.2.5 Intergenic Distance Profiles 60
3.3 Results 60
3.4 Discussion 66
Chapter 4 Applications: Metagenomic Annotation Networks 70
4.1 Introduction 71
4.2 Materials and Methods 73
4.2.1 Data Preparation 73
4.2.2 Network Prediction 75
4.2.3 Network Translation 77
4.2.4 Annotation Frequency Analysis 78
4.3 Results 79
4.3.1 Target Perspective Networks 79
4.3.2 Source Perspective Networks 85
4.3.3 Comparative Networks 91
4.4 Discussion 93
Chapter 5 Conclusion 98
5.1 The Proximon Proposition 99
5.2 Computational Representation of Biological Data 100
5.3 Evaluation of Proximons 105
5.4 Metagenomic Annotation Networks 107
5.5 Final Remarks 110
Bibliography 112
Appendix A MetaProx UML Architecture 128
Appendix B IMG/M Datasets 132

Appendix C Functional Modules 170

Appendix D Publications and Permissions 197

D.1 Publications Included in this Thesis 197

D.2 Other Doctoral Publications 199

List of Figures

Figure 1-1 The classical genome versus the microbial pan-genome 13

Figure 1-2 Abstract metagenome.. 21

Figure 1-3 Inference and annotation of metagenomic functional interactions 23

Figure 1-4 The proximon proposition.. 25

Figure 2-1 Relational data modeling.. 32

Figure 2-2 Abstract data model.. 39

Figure 2-3 Proximon selection criteria 42

Figure 2-4 IMG/M sample record.. 43

Figure 2-5 Contig hierarchy.. 44

Figure 2-6 Distribution of proximon lengths.. 46

Figure 2-7 Application deployment perspectives 48

Figure 2-8 MetaProx graphical user interface 49

Figure 2-9 Purine degradation network 51

Figure 3-1 Proximon mapping configurations... 61

Figure 3-2 Proportion of mapping configurations 62

Figure 3-3 Distribution of operon hits 63

Figure 3-4 Multi-hit operon 63

Figure 3-5 Distribution of operon coverage... 64

Figure 3-6 Intergenic distance distributions 65

Figure 4-1 Data source diversity.. 74

Figure 4-2 Network construction workflow 76

Figure 4-3 Target stringency versus network coverage... 81

Figure 4-4 Metagenomic cellulase networks 83

Figure 4-5 Annotation hierarchy chart... 84

Figure 4-6 Human gut networks 86

Figure 4-7 Comparative gut networks 92

Figure 5-1 Representation versus inflation.. 101

Figure 5-2 Annotative cohesion of known operons... 109

Figure A-1 MetaProx UML overview 128
Figure A-2 ADT UML diagram.. 129
Figure A-3 GUI UML diagram... 130
Figure A-4 Chipset UML diagram... 131

List of Tables

Table 1-1 Online operon resources 18

Table 2-1 Sparsity of metagenomic functional annotations 38

Table 2-2 MetaProx database composition 45

Table 4-1 Summary of network features 80

Table 4-2 Top ranked functional modules 88

Table B-1 IMG/M dataset usage and descriptions 132

Table C-1 Metagenomic functional modules with abbreviated annotations 170

Table C-2 Top ranked metagenomic functional modules with verbose annotations 185

Chapter 1

Introduction

Metagenomics is the culture-independent genomic analysis of a particular environment or community of microorganisms. A fundamental benefit of this approach is that it offers a means to investigate the genomic properties of the large proportion of bacteria, archaea, and viruses that are not amenable to standard culturing techniques. As a result, metagenomics has the potential to greatly extend our understanding of microbial ecology by revealing new insights with respect to both phylogenetic and functional perspectives. In particular, metagenomic studies have shed light on key issues such as characterizing genetic variability within and between microbial species, as well as enumerating the functional repertoires and ecological roles of both individual species and whole communities. However, the application of a culture-independent paradigm also simultaneously entails a variety of unique challenges and caveats.

1.1 The Rise of the Metagenomic Era

Historically, metagenomics as a field of study is preceded by bacterial genomics which in turn is preceded by microbiology. Unlike the two *omics* fields, microbiology has presided over microbial investigation and study for a considerable time period, beginning with the bacteriologists of the 19th Century (see Winslow, 1950) who followed from microscopists such as Antonie van Leeuwenhoek two centuries earlier (Bulloch, 1938). Late in the 19th Century, pioneering microbiologists such as Robert Koch were motivated to employ a pure-culture protocol in their research in an effort to draw a clear causal connection between bacteria and disease (Mazumdar, 1995). In turn, the precise information gleaned from model organisms in pure culture progressively established the general body of prevailing knowledge in microbiology over the next century (Handelsman, 2004). However, the dominance of the pure-culture paradigm began to undergo challenge as a consequence of numerous unprecedented findings, most notably the "great plate-count anomaly" pointed out by Staley & Konopka (1985) based on findings from several earlier works, as well as the groundbreaking work by Woese & Fox (1977) and Woese (1987) that revolutionized perspectives on prokaryotic taxonomy by applying quantitative molecular analysis, rather than phenotypic characterization. As a result of these highly impactful findings, interest in uncultured microbes began to increase in the mid-1980s, propelled as well by the advent of PCR[1] technology. In particular, microbiologists began to recognize the bias produced by

[1] Polymerase Chain Reaction (PCR): A process for the generation of numerous copies (e.g. thousands or even millions) of a DNA sequence produced from a single or low number of source sequences.

culturing limitations and began to confirm the breadth of the uncultured majority as revealed by evidence from ribosomal RNA studies (Handelsman, 2004).

The closing decades of the last century saw the emergence of genomics, a field that provides a whole-organism perspective of hereditary material in the form of a genomic sequence (Medini et al., 2008). The sequencing of the first microbial genome, *Haemophilus influenzae* (Fleischmann et al., 1995), was followed by steady decreases in the overall cost of sequencing. This resulted in an exponential increase in the number of sequenced genomes (Handelsman, 2004). Although these were primarily microbial genomes, the Human Genome (Venter et al., 2001) was a highly notable inclusion. Genomics has greatly enhanced the field of microbiology, particularly in terms of clarifying the relationship between traditional phenotypic characteristics and their underlying DNA sequences (Achtman & Wagner, 2008; Joyce et al., 2002). Furthermore, genomics has also raised serious questions about the validity of traditional taxonomy and evolution of microbes, especially in light of HGT[2] (Achtman & Wagner, 2008; Joyce et al., 2002). Despite being pragmatic, the notion of discrete species, which implies relatively static and discrete genomes, is difficult to reconcile with dynamic views of microbial genomic composition such as pan-genomes[3] (Medini et al., 2008). These propositions have been so compelling that the study of microbes has been irreversibly propelled into a post-genomic era.

[2] Horizontal Gene Transfer (HGT): in prokaryotes, the transfer of genes by means of bacteriophages or plasmids, rather than through successive duplication involving binary fission

[3] See Section 1.2 for a description of the pan-genome concept

While the contributions of genomics have been remarkable, as a research field it has historically shared one particularly salient artifact with traditional microbiology: most sequences are determined from pure cultures to avoid ambiguities during sequence assembly (Schloss & Handelsman, 2005), although more genomes have begun to be cloned from metagenomic sequences. Therefore, like microbiology, genomics cannot provide a holistic viewpoint necessary to adequately understand diverse microbial communities. This is because the pure-culture paradigm is necessarily limited by how much of microbial life is amenable to culturing. However, it has been estimated that more than 99% of microorganisms are culture-resistant (Ferrer et al., 2005; Tringe & Rubin, 2005; Riesenfeld et al., 2004a). Metagenomics circumvents this limitation by sequencing heterogeneous samples of DNA amplified directly from the environment, and thus containing a variety of genomic sources, rather than a single source organism (Tringe & Rubin, 2005; Handelsman, 2004). The obvious benefit of this method is that it provides access to previously inaccessible organisms (Tringe & Rubin, 2005). For example, symbionts and obligate pathogens cannot survive outside of their hosts and environmental microbes are often unable to grow in pure culture (Tringe & Rubin, 2005). However, DNA can be directly extracted from such organisms while they are in their natural habitats, thereby yielding a heterogeneous mixture of DNA that can be separated into libraries of sequence data (Tringe & Rubin, 2005). Metagenomic libraries provide insight into community dynamics by revealing the complement of genes that occur with respect to a particular environment (Tringe & Rubin, 2005). In turn, such knowledge can drive specific studies, like the search for quorum sensing (QS) cell–cell communication systems beyond those found in cultured microorganisms (Hao et al., 2010). Overall, metagenomics offers a means to exceed the current limitations of genomics by disregarding the pure-culture paradigm, thereby extending the amount of usable sequence data.

Pace (1985) put forth the proposition of using DNA obtained directly from environmental samples and this concept was implemented several years later by Schmidt et al. (1991) by utilizing cloning in a phage vector. This initiative was subsequently followed by more elaborate metagenomic library construction efforts such as Stein et al. (1996). Metagenomics as its own distinct field of research began to take shape at the end of the 20th Century and the term metagenome was first coined by Handelsman et al. (1998) with respect to the concept of meta-analysis being applied to similar but not identical datasets. Interest in metagenomics flourished in the new millennium, sparking several landmark projects. The Sargasso Sea metagenomic survey (Venter et al., 2004) represented an effort to better understand oceanic microbial populations. The 1,214,207 putative protein-encoding genes that were identified constituted an enormous contribution, both in terms of novelty and volume (Venter et al., 2004). This project alone yielded almost as many proteins as existed in the combined curated protein databases (non-redundant SWISSPROT, TREMBL and TREMBLnew) of the same time period (Tress et al., 2006). Previously, the acid mine drainage project (Tyson et al., 2004) assessed the microbial community associated with acid resulting from the oxidation of sulfide minerals produced by mining and provided an example of a low complexity community, as it is dominated by only five microbial species. The soil-resistome project (D’Costa et al., 2006) attempted to identify antibiotic resistance genes by screening DNA fragments for their potential expression of antibiotic resistance. Specific environmental niches, or microbiomes, also began to be compared and contrasted, such as phylogenetic contrasts between the guts of lean versus obese mice (Turnbaugh et al., 2006). Since these early projects, a vast number of metagenomic datasets have been produced that characterize an incredibly rich range of environments.

1.2 Challenges and Prospects in Metagenomics

With an estimated 10^{30} microbial cells on Earth (Turnbaugh & Gordon, 2008), microbes represent the most abundant contributors to life on Earth, both from the perspective of their sheer numbers and with respect to the biological processes that they mediate. The interdependent processes arising from integrated microbial communities drive the biosphere in fundamental ways ranging from providing bioavailability to carrying out biogeochemical processes. Microbes also play a key role for numerous human interests and technologies, such as agricultural enhancement, antibiotic production, food fermentation, and biofuel production (Simon & Daniel, 2011). Therefore, gaining access to the novel metabolic repertoire contained within the uncultured majority represents a paramount objective in modern biology.

Metagenomic research bypasses the limitations of culturing because it is based on the isolation of DNA obtained directly from environmental samples. Metagenomic studies begin with sample collection from a habitat of interest and typically employ filtration by size to reduce contamination by viruses or eukaryotes (Teeling & Glöckner, 2012). The biodiversity of the particular habitat itself can exert a powerful effect on the quality of the final metagenomic data. Both the sheer number of different species in a sample as well as the evenness of their relative proportions will impact the efficacy of sequence assembly such that increasing complexity impedes assembly resolution (Teeling & Glöckner, 2012). Furthermore, genomic coherence can pose an additional challenge when a habitat contains species that exhibit a low level of population clonality as a result of a rich pan-genome (Teeling & Glöckner, 2012).

DNA sequencing technology has progressed across the past half-century and is commonly categorized with respect to three periods known as first generation, second generation, and third generation. First generation sequencing efforts concerned the sequencing of clonal DNA populations and involved a series of technical increments, most notably Sanger's chain-termination technique (Sanger & Nicklen, 1977). Propelled by the emergence of PCR, improvements to Sanger’s method permitted first generation sequencers to produce reads approaching one kilobase in length that could be extended further by computationally overlapping separately sequenced DNA fragments to produce a longer contiguous sequence (Heather & Chain, 2016). Second generation sequencing, often referred to as next-generation sequencing (NGS), is highlighted by the parallelization of reactions in order to produce huge gains in sequencing throughput. Driven largely by the adoption of pryrosequencing[4] methodology (Nyrén, 1987), real-time NGS technology incrementally improved, ultimately yielding substantial decreases in the overall cost of sequencing (Heather & Chain, 2016). As a result, NGS is recognized as a key contributor to the shaping of genomic era (Heather & Chain, 2016). The progression from NGS to third generation

[4] Pyrosequencing is a DNA sequencing protocol that exploits the detection of light emission caused by pyrophosphate release that occurs during iterative nucleotide incorporation.

sequencing is less pronounced than the previous demarcation; however, it is often distinguished by the inclusion of single molecule sequencing (SMS) technology (Braslavsky et al, 2003). SMS operates in a manner similar to Illumina (a dominant NGS platform) but with the notable difference that bridge amplification is not required, thereby removing potential biases and errors (Heather & Chain, 2016). The third generation continues to evolve, including the ongoing improvement of nanopore sequencing (Haque et al., 2013). In addition, current sequencing technology has facilitated single-cell genomics studies by offering improved resolution and accuracy in variant calling, in comparison to microarrays (Macaulay & Voet, 2014). In turn, these studies could reveal a new understanding of complex biological systems with a rich range of application domains (Gawad et al., 2016).

The removal of noise caused by PCR and sequencing errors represents a key quality improvement step in sequence analysis. Noise removal typically involves tracking information on erroneous sequences through the retention of representative reads (Kim et al., 2013). A variety of tools and algorithms exist such as Denoiser (Reeder & Knight, 2010) and PyroNoise (Quince et al., 2011). Chimera detection is another important process aimed at increasing data quality. Chimeras occur when prematurely terminated fragments reanneal to other template DNA during PCR amplification and result in artificial recombinants formed from multiple sources (Bradley & Hillis, 1997). Moreover, increased read length, while desirable, also increases the risk of chimeric assemblies (Teeling & Glöckner, 2012). Several tools are available for chimera detection including UCHIME (Edgar et al., 2011) and DECIPHER (Wright et al., 2012).

The prediction of protein-coding genes is a key objective following DNA sequence analysis. In the case of bacteria, a gene is typically comprised of an uninterrupted span of DNA ranging between a start codon and a stop codon (Koonin & Galperin, 2003). Similarly, with respect to an open reading frame (ORF) which is also a span of DNA between a start and stop codon, a gene is commonly defined as the longest ORF occurring in a given region of DNA (Koonin & Galperin, 2003). While gene length itself is highly variable, ORFs that are less than 100 bases in length are typically ignored as candidates for protein-coding genes. However, the gene length heuristic can fail in uncommon cases where the shorter of two overlapping ORFs represents the real gene (Koonin & Galperin, 2003). Several well-known gene prediction algorithms have been developed, such as GeneMark (Borodovsky & McIninch, 1993) and Gene Locator and Interpolated Markov Modeler (GLIMMER) (Salzberg et al., 1998).

Taxonomic examinations of metagenomic data typically involve any number of sequence-based analyses aimed organizing a given collection of contigs with respect to phylogenetic bins or clusters. Gene-based classification exploits potential similarity between sequences within metagenomic contigs and the sequences of known genes and/or proteins. In particular, sequence alignment algorithms and tools such as Basic Local Alignment Search Tool (BLAST) queries (Altschul et al., 1990) can provide taxonomic indicators, especially when BLAST hits are subjected to further processing by more taxonomically oriented tools such as MEtaGenome ANalyzer (MEGAN) (Huson et al., 2007). However, gene-based analysis does require the existence of at least remotely comparable sequences within the reference database in order to drive taxonomic inference.

An alternative to gene-based analysis involves inferring taxonomic information on the basis of patterns in DNA sequence composition. This approach exploits the detection of recognizable phylogenetic signals determined using normalized frequencies of short DNA oligomers (Abe et al., 2003; Abe et al., 2005). This method offers taxonomic profiling of metagenomes while circumventing the requisites of the previously mentioned homology driven approach. However, the accuracy of binning metagenomic contigs is contingent upon a minimum length of assembly and this can impact the inclusion of data consisting of short fragments such as pyrosequencing reads (McHardy & Rigoutsos, 2007). It should also be pointed out that both of the approaches described here are also susceptible to chimeric contigs where assembly has occurred using reads from different taxonomic origins.

Taxonomic analysis can also be accomplished using conserved marker genes, such as *recA* or 16S rRNA genes (Simon & Daniel, 2011). In particular, rRNA gene-based studies have been widely applied toward inferring diversity and composition of a broad range of microbial communities. Moreover, 16S rRNA genes analyses are supported by large databases of reference sequences, such as Ribosomal Database Project II (RDP II) (Cole et al., 2003). However, fragments carrying rRNA genes are infrequent (less than 0.1% of a typical collection) (McHardy & Rigoutsos, 2007) and depending on the specific approach taken other caveats can arise such as primer bias or differing proportions of rRNA operons depending on taxonomic origin (Teeling & Glöckner, 2012). Nevertheless, the comparative analysis of 16S rRNA gene sequence data has had a profound impact on taxonomic efforts in metagenomic research and this trend is likely to continue as the diversity of phylogenetic markers increases.

Facilitated by the previously described taxonomic protocols, one of the most compelling insights that metagenomics has to offer concerns our understanding of the completeness of microbial biodiversity. Although it is typically not possible to exhaustively determine the complete biodiversity of a microbial community, environmental samples can still provide valuable indications of the number of taxa in a community, as well as their relative abundance (Shaw et al., 2008). Similarly, the aggregation of multiple data sets can be applied to important large-scale analyses such as Hug et al.’s (2016) new view of the tree of life. Improved resolution of microbial biodiversity has important consequences for reducing biases that currently exist in the composition of many databases (Pignatelli et al., 2008). The limited spectrum of culturable microbes combined with applied research interests has yielded a skewed representation of recognized microbial biodiversity (Wu et al., 2009). Metagenomes offer an opportunity to better depict the diversity of genes and proteins, as well as organisms, thereby leading to greater database completion (Pignatelli et al., 2008). Improved database completion would have a subsequent impact on the effectiveness of various pursuits, including the functional assignment of proteins and the taxonomic classification of metagenomic sequences (Pignatelli et al., 2008).

Conventional views on species and genomes, as well as their relationship to one another, are also being impacted by ongoing metagenomic findings. It has been suggested that some of these data have demonstrated a general weakness in our accepted views of simplified linear evolution and the concept of a bifurcating tree of life (Bapteste et al., 2009). This problem is further compounded by recent challenges to the concept of adaptation and its role in evolutionary thought (Depew, 2011), as well as various semantic and philosophical concerns (Krohs, 2012; O'Malley & Soyer, 2012; Callebaut, 2012; Calvert, 2012; Strasser, 2012). Metagenomic datasets have recently demonstrated that members within a given species can exhibit striking genomic plasticity, despite being considered taxonomically equivalent (Mira et al., 2010; Tettelin et al., 2008; Medini et al., 2005). This recurrent finding is believed to be strongly driven by horizontal gene transfer and has given rise to the concept of a *pan-genome* for microbial species, rather than a fixed and singular genomic identity (Mira et al., 2010; Tettelin et al., 2008; Medini et al., 2005). Sequence data from multiple conspecific instances can be used to construct a pan-genome by taking the union of the sets of genes that correspond to each source genome (Mira et al., 2010; Tettelin et al., 2008; Medini et al., 2005). Therefore, any given instance of that particular species will have a genome that contains a subset of the genes found in the total pan-genomic collection (Mira et al., 2010; Tettelin et al., 2008; Medini et al., 2005). Furthermore, by identifying the intersection between conspecific genomes, a mutually occurring set of genes can be identified as the core genome for a given species, while the remaining genes are considered to be auxiliary or strain-specific genes (Mira et al., 2010; Tettelin et al., 2008; Medini et al., 2005) (see Figure 1-1). For example, various strains of *Escherichia coli* are known to exhibit a mutual core genome that does not exceed 40% of their combined set of genes (Mira et al., 2010; Bapteste et al., 2009). This is in stark contrast to eukaryotic scenarios where genomic instances are highly conserved within a given species (Mira et al., 2010; Bapteste et al., 2009) (see Figure 1-1). Therefore, it has been argued that the prokaryotic species definition should differ from that of eukaryotes (Fraser et al., 2009; Achtman & Wagner, 2008).

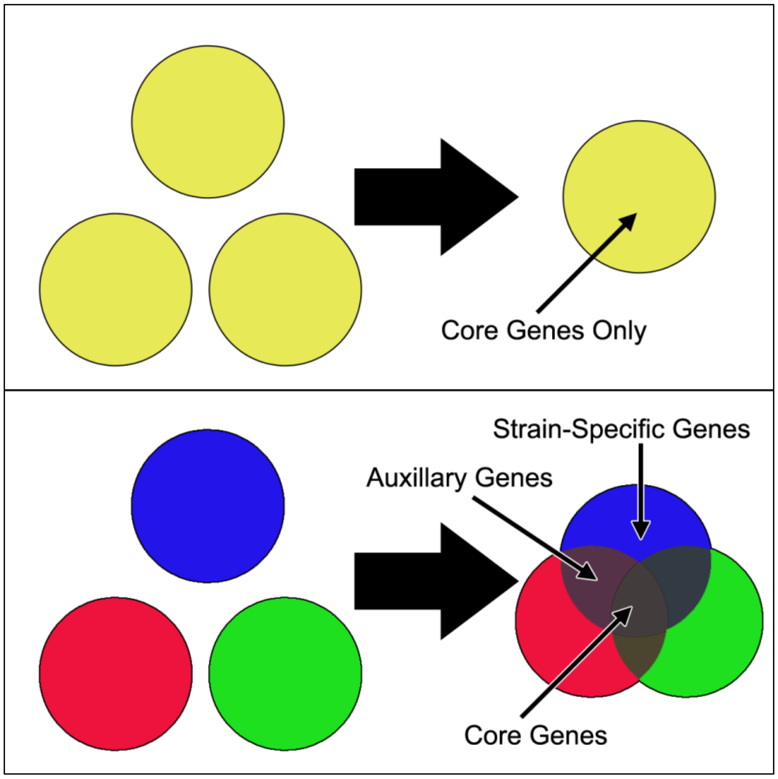


Figure 1-1 The classical genome versus the microbial pan-genome. *The upper panel shows three conspecific and identical genomes. Their resulting set theoretic comparison produces a single set of genes equivalent to any and each of the source genomes. This relationship corresponds to the classical genome; a singular core of fixed genes where species-genome cardinality is one-to-one. The lower panel shows three conspecific but non-identical genomes. Their resulting set theoretic comparison produces a superset of genes greater than any and each of the source genomes. This relationship corresponds to the microbial pan-genome; a core of mutual genes in combination with additional auxiliary and strain-specific genes where species-genome cardinality is one-to-many.*

Microbial communities exhibit complex taxonomic and structural arrangements that are a reflection of their highly organized interspecies interactions (Wilmes et al., 2009; Allen & Banfield, 2005). These dynamics stem from the particular metabolic requirements associated with the effective exploitation of a given niche (Wilmes et al., 2009; Allen & Banfield, 2005). Furthermore, achieving metabolic capacity makes no guarantees about the underlying species composition and individual species members can vary in their functional contributions both between and within communities (Wilmes et al., 2009; Allen & Banfield, 2005). In general, community dynamics complement the issue of genomic plasticity by affirming that microbes also possess a capacity for functional plasticity. The situation becomes further exacerbated by the moonlighting capabilities of certain proteins (Jeffery, 2009; Jeffery, 1999), as well as the multifunctional interactions of some genes (Gillis & Pavlidis, 2011).

Functional metagenomics represent another major aspect of metagenomic analyses where the primary objectives can range from the annotation of specific genes to understanding the overall functional repertoire for a given microbiome. Functional analyses can be accomplished in the absence of sequence information through a variety of screening techniques involving the use of metagenomic library containing clones. Function-based screening employs heterologous complementation of host strains or mutants of host strains that require specific targeted genes for survival given selective conditions, such as genes that confer a specific antibiotic-resistance (Riesenfeld et al., 2004b). Phenotypical detection involves the incorporation of chemical dyes fused with enzyme substrates as a component of the growth medium, thereby revealing metabolic capabilities of individual clones (Ferrer et al., 2009). Other elaborate approaches such as product-induced gene expression (PIGEX) (Uchiyama & Miyazaki, 2010) have also been used to detect gene expression in metagenomic clones.

Functional analyses can also be carried out when sequence information is available by exploiting bioinformatics resources. In particular, a substantial collection of homology-based tools is available that employ either BLAST (including variations like BLASTX, BLASTP, or BLAT) or alternatively a hidden Markov model algorithm such as HMMER (Finn et al., 2011), or similar statistical approach. Examples of homology-based annotation tools include the integrated metagenome data management and comparative analysis system (IMG/M) (Markowitz et al., 2012), the databases of clusters of orthologous groups (COGs) (Tatusov et al., 1997), the COGNIZER framework (Bose et al., 2015), the Pfam protein families database (Punta et al., 2012), the TIGRFAMs database of protein families (Selengut et al., 2007), the KEGG PATHWAY Database (Kanehisa et al., 2012), and many more. Similarly, motif-based annotation databases such as InterPro (Hunter et al., 2012) and PROSITE (Sigrist et al., 2010) search protein sequences for motifs or patterns that correspond to structural and/or functional qualities necessary for a given category of proteins to maintain their properties and/or activities. An alternative to homology-based and motif-based approaches are the context-based methods such as gene fusions (Enright et al., 1999; Marcotte et al., 1999), conservation of adjacency (i.e. gene neighbourhoods) (Dandekar et al., 1998; Overbeek et al., 1999), and phylogenetic profiling (Pellegrini et al., 1999).

The computational prediction of operons is another context-based strategy that is particularly well-suited for use with metagenomic sequence data. In accordance with Jacob &

Monod’s seminal work (1961) an operon is generally regarded as a collection of genes that are mutually regulated and transcribed as a single polycistronic unit. Attempts to explain the existence of operons typically involve considerations of the selective advantages that they might provide and constraints that would favour the physical proximity of their member genes. Operon organization has been traditionally considered as advantageous due to the coordinated expression of genes involved in a common function (Jacob & Monod, 1961). The selfish operon theory (Lawrence & Roth, 1996) asserts that for an operon to confer a function its full complement of genes must be acquired as a unit and the probability of transferring multiple genes increases with gene proximity. In addition, the added constraint of whether or not operons can be overlapping must also be considered. In particular, inclusion or exclusion of this qualifier has important ramifications for operon prediction protocols that rely primarily on intergenic distance and co-direction. Recent research (Conway et al., 2014) has clearly shown that operons can exhibit differential expression where a single operon acts as a complex of transcription units (TUs), due to the presence of internal promoters or terminators. This relationship introduces a degree of ambiguity that can be clarified from a set theoretic perspective where a TU cluster (TUC) is a set of one or more TUs that are connected to one another by way of shared member genes (Mao et al., 2015). In other words, a TUC is a set of contiguous genes and any given member TU is also a set of genes such that TU $\subseteq$ TUC. Moreover, this definition allows for a variety of TU configurations where a TU can span its entire TUC, begin with the leading gene but terminate prior to the final gene, begin after the leading gene but terminate with the final gene, or both begin after the leading gene and terminate prior to the final gene (Mao et al., 2015). In this thesis, I generally use the terms *operon* and *proximon* (see Section 1.4) to refer to an instance of a TU.

Operon prediction in prokaryotes has been undertaken using a variety of perspectives and techniques. Ermolaeva et al. (2001) employed a conservation of adjacency method that requires the identification of operons that are conserved across multiple species and exploits the premise that genes that remain adjacent after long periods of evolution are likely to be in the same operon. Moreno-Hagelsieb & Collado-Vides (2002) utilized intergenic distance to predict operons based on the observation that genes in the same operon tend to be separated by fewer base pairs. Using data from *Escherichia coli* operons, they created a probabilistic distance model that is considered transferable to other species and they validated this premise using *Bacillus subtilis* (Moreno-Hagelsieb & Collado-Vides, 2002). Bockhorst et al. (2003) used Bayesian networks in combination with sequence data and expression data to predict operons using a probabilistic approach. These founding approaches have given rise to the robust collection of currently available online resources dedicated to prokaryotic operons (see Table 1-1).

Harnessing functional metagenomics offers many useful prospects and applications. For example, microbial communities play a key role in many agricultural pursuits, both as beneficial agents and as dangerous contaminants (Kyrpides et al., 2014). Also, microbial ecosystems can be used as predictive models to understand large-scale environmental processes or as indicators of environmental damage, as well as potential facilitators of environmental remediation (Handelsman, 2004). Microbes also have potential biomedical applications including revealing novel treatments for disease based on a better understanding

Table 1-1 Online operon resources. Online resources for prokaryotic operons are summarized including name, URL, authors, and a description of the information that is available.

Resource	Authors	Description
DOOR http://csbl.bmb.uga.edu/DOOR	Mao et al., 2009	A comprehensive operon database covering containing 1,323,902 operons from 2,072 bacterial genomes.
MicrobesOnline http://microbesonline.org	Alm et al., 2005	Provides a variety of bioinformatics tools including operon predictions for every bacterial and archaeal genome.
ODB http://operondb.jp	Okuda et al., 2006	Contains over 400,000 conserved operons from more than 1,000 bacterial genomes, as well as various graphical interfaces for analyses and visualization.
OperonDB http://operondb.cbcb.umd.edu	Pertea et al., 2009	Contains predicted gene pairs for 1,059 bacterial and archaeal genomes.
ProOpDB http://operons.ibt.unam.mx/ OperonPredictor	Taboada et al., 2012	Uses a novel operon identification algorithm and contains operon predictions for more than 1,200 prokaryotic genomes.

of the relationship between health and the human microbiome (Handelsman, 2004). Similarly, biotechnology can benefit from the novel biocatalytic and biosynthetic abilities of microbial communities (Kyrpides et al., 2014). Even the future of energy generation stands to benefit from the viability of microbially generated biofuels (Kyrpides et al., 2014).

The Human Microbiome Project (HMP) launched by the National Institutes of Health (NIH) is a prime example of a large-scale project rooted in applied metagenomics. This venture is composed of numerous core initiatives, each of which includes its own set of research projects. These initiatives span a broad range of concerns, from computational and technical issues to ethical and social considerations. It is estimated that microbes in the human body outnumber their host cell count by a tenfold factor and that these microbes may collectively encode 100 times the number of unique genes contained in the human genome (Qin et al., 2010). Therefore, understanding the contributions of these microbes is essential to realizing the complexity of our nutritional, physiological, and immunological capacities and how these facets arise as consequences of our interaction with our own microbiome (Qin et al., 2010). Moreover, changes in the composition of the human microbiome, particularly the gut microbiome, may serve as indicators of disease or obesity (Qin et al., 2010). Overall, research on the human microbiome is likely to yield a wealth of information that will simultaneously advance applied research interests and general knowledge about microbes.

Given the opportunities made available by functional metagenomics combined with the possibility to reach previously inaccessible organisms, the remainder of this thesis is dedicated to context-based functional analyses. In particular, the present research focuses on the use of computationally-driven prediction strategies to infer functionally linked groups of metagenomic genes that are analogous to operons or TUs. The motivation for this undertaking is to generally facilitate, augment, and extend the current infrastructure and protocols available for exploratory and comparative research involving functional topology and across different levels of scope, ranging from simple functional units to elaborate functional networks.

1.3 Challenges in Context-Based Functional Inference

Metagenomic genes fundamentally differ from genomic genes in that they provide limited contextual information because they are situated within variable length fragments of DNA (see Figure 1-2). This is because metagenomic DNA is obtained from an environmental sample that represents a heterogeneous community, rather than an isolated population, and therefore the derived sequence data is typically limited to being assembled into contigs (contiguous genomic subsections) instead of complete genomes (Kunin et al., 2008). Moreover, the properties of species richness and species abundance interact to produce an effect of overall community complexity that subsequently affects the resolution of the assembly process such that contig length generally decreases as community complexity increases (Kunin et al., 2008). As a result, metagenomic genes provide reduced information about features such as absolute genomic position and conditions like orthology or paralogy (Vey & Moreno-Hagelsieb, 2010).

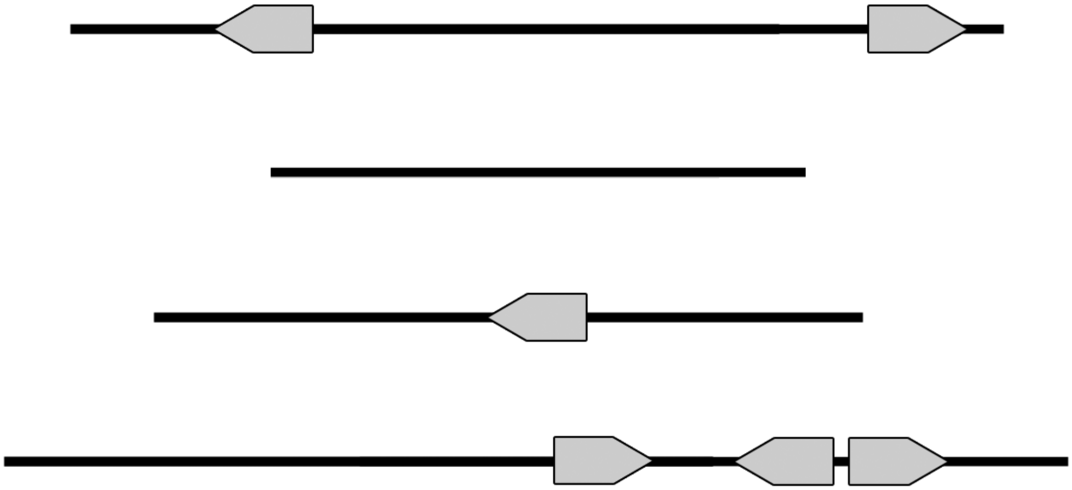


Figure 1-2 Abstract metagenome. *Metagenomic DNA is assembled into variable length contigs that can contain either multiple (two or more) genes in varying configurations with respect to proximity and direction, a single gene in either direction, or no genes at all.*

The processes of metagenomic gene prediction and subsequent annotation also differ from their genomic analogs because of the fragmentary and anonymous nature of metagenomic sequences. Unassembled reads and very short contigs are prone to fragmented gene predictions where one or both ends of a predicted gene exist beyond the read or contig that has spawned the initial prediction (Liu et al., 2013). Even if a contig is sufficiently long so as to contain multiple genes, these genes typically occur in very low numbers in comparison to a genomic scenario, thereby eliminating model training required by supervised prediction methods that have been previously applied to single genomes (Noguchi et al., 2006). These problems are exacerbated by the fact that the taxonomic origins of most metagenomic fragments are unknown and/or completely novel, thus impeding the construction of statistical models intended to exploit aspects of feature detection (Liu et al., 2013). Similarly, both gene prediction and corresponding functional assignments are frequently reliant on homology-based tools like Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1990) or hidden Markov models (HMMs) (Yoon, 2009). However, this type of approach is limited to identifying genes that already have known homologs (Liu et al., 2013). Alternatively, *ab initio* gene identification algorithms (Hyatt et al., 2012; Kelley et al., 2012; Zhu et al., 2010; Rho et al., 2010; Hoff et al., 2008; Noguchi et al., 2008; Noguchi et al., 2006) have also been developed to circumvent the requisite of homology in order to better address the aspect of novelty that is a hallmark of metagenomic data.

A functional interaction can be defined by a mutually cooperative relationship that functionally links two or more genes and necessarily indicates a state of functional association. Such configurations are demonstrated among the member genes of a given type of functional unit, such as the co-transcribed protein coding genes within an operon (Jacob & Monod, 1961; Miller & Reznikoff, 1978). Therefore, metagenomic functional interactions can be used for the inference of unknown functional characteristics in a manner that involves aspects of both homology searching and *ab initio* methods. Specifically, once gene predictions and functional annotations have been assigned as previously described, potential metagenomic functional interactions can be determined using a standard operon detection protocol (Salgado et al., 2000; Moreno-Hagelsieb & Collado-Vides, 2002) that has been previously demonstrated with metagenomic data (Vey & Moreno-Hagelsieb, 2010). Next, the functional annotations of interacting genes can then be used to derive networks that portray functional interdependence and modularity as depicted through various features of network

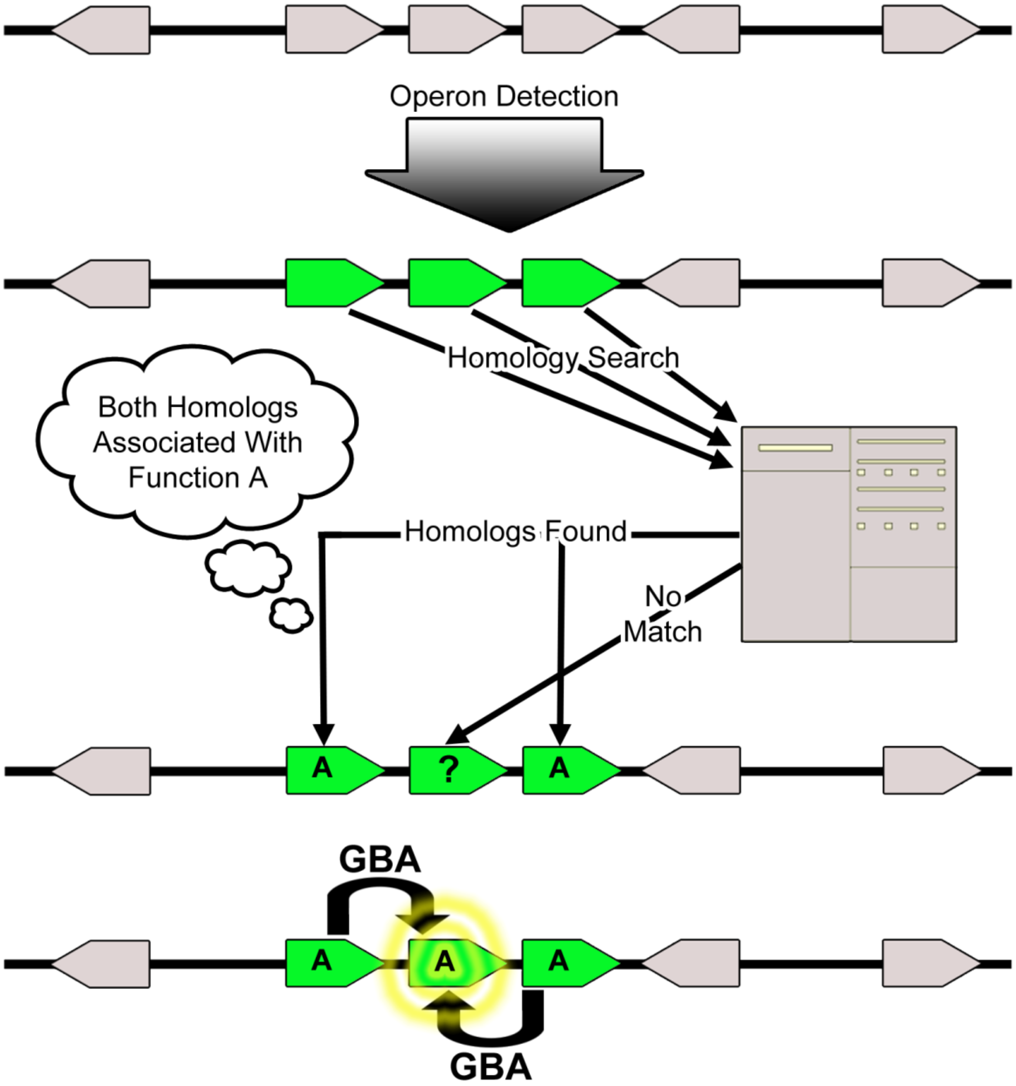


Figure 1-3 Inference and annotation of metagenomic functional interactions. *A metagenomic contig is subjected to an operon detection protocol which can be preceded by or followed by functional annotation using various homology-based tools. Remaining unannotated genes can optionally have putative functions potentially inferred using the guilt by association paradigm.*

connectivity. In addition, existing annotations can be used to infer putative functions for genes that lack an annotation but have functional linkages to other annotated genes by way of the guilt by association paradigm (Aravind, 2000; Oliver, 2000) (see Figure 1-3). Overall, the effective use of metagenomic functional interactions represents a key prospect for a variety of applications in the field of functional metagenomics.

1.4 The Proximon Proposition

Experimental validation performed using *Escherichia coli* (Salgado et al., 2013) and *Bacillus subtilis* (Sierro et al., 2008) has helped to identify key features of operon member genes, particularly co-direction and proximity with respect to intergenic distance. Therefore, by using the coordinates of detected genes, metagenomic functional interactions can be subsequently predicted using an operon detection protocol (Salgado et al., 2000; Moreno-Hagelsieb & Collado-Vides, 2002) that has been previously demonstrated with metagenomic data (Vey, 2013; Vey & Moreno-Hagelsieb, 2012; Vey & Moreno-Hagelsieb, 2010).

However, while metagenomic functional interactions offer utility for various pursuits in functional metagenomics, it is nevertheless inaccurate to qualify sets of co-directional and co-proximal genes as necessarily being operons. Although the same can be said for genomic operon candidates, the protocols used to predict these candidates often augment their selections with additional evidence such as equivalent arrangements of orthologous genes (Moreno-Hagelsieb & Janga, 2008; Janga & Moreno-Hagelsieb, 2004; Moreno-Hagelsieb & Collado-Vides, 2002) or functional relationships between known protein products (Taboada et al., 2010) that are typically not available in metagenomic scenarios. Thus, while the metagenomic functional packets that are identified using solely direction and proximity are not guaranteed to be operons, they are more significant than the general case of directons (series of contiguous co-directional genes) because their member genes exhibit close proximity with respect to adjacent pairwise distances. Therefore, it is proposed that these structures represent their own unique class situated as a subset of the directon class and a superset of the operon class (see Figure 1-4) and the term *proximon* is proposed here to denote a proximally significant directon.

It is important to explicitly clarify that the proximon proposition is intended to be used primarily as a biological abstraction. In other words, a proximon is meant to serve as a conceptual entity aimed at facilitating bioinformatics and/or computational goals. Therefore, in this thesis proximons are constructed as information theoretic entities generated from

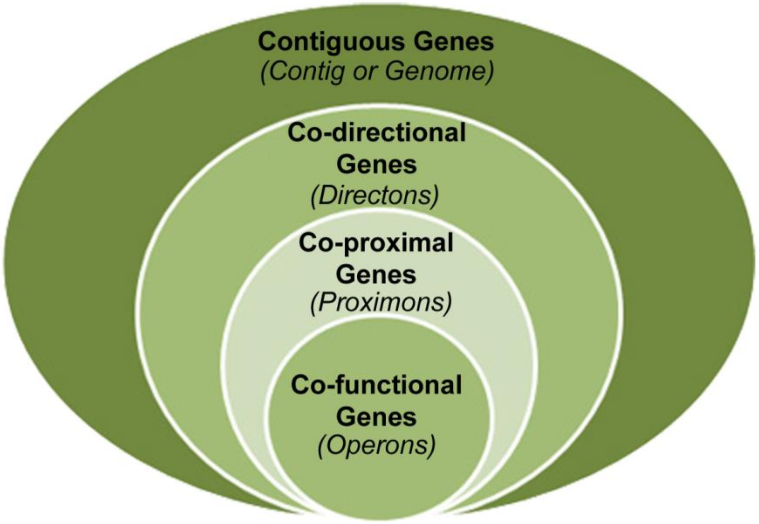


Figure 1-4 The proximon proposition. *The proximon class (co-proximal genes) is shown from a set theoretic perspective as a subset of the directons (co-directional genes) and a superset of the operons (co-functional genes).*

digital data for further usage in downstream computational analyses. While the biological underpinnings of the proximon must be real and ultimately mappable to actual genes, the paradigms, protocols, and validations that are implemented and investigated in this thesis reside within the aforementioned layer of computational abstraction and any direct connection to wetlab applications or validations is beyond the scope of the present work.

1.5 Goals and Objectives

This thesis is aimed at evaluating the proximon proposition by demonstrating and examining its utility with respect to functional metagenomics. Specifically, this investigation uses a manuscript-based approach to present a series of studies focused on the following research areas:

1. Representation: In response to current trends in the effective management of large-scale biological data, alternatives to the relational data model will be investigated. In particular, a robust object-based representation will be devised, as well as a corresponding means to perform queries on data rendered in this form. The proposed data model will then be used to represent a large-scale repository of metagenomic proximons derived from previously identified metagenomic genes. The finalized data will be offered in the form of a publicly available online database and accompanying frontend search tool, thereby also addressing topics immediately adjacent to data representation, such as challenges in dissemination and deployment. It should be noted that the primary considerations of this goal relate to the modelling, persistence,

and distribution of data but not the data prediction protocol itself because this process is an implementation of a previously established method.

2. Evaluation: Although the primary goal of predicting proximons is the identification of potential functional interactions between metagenomic genes, the proximon proposition itself is predicated on the assertion that proximons are abstractions of real operons. Therefore, a formal evaluation will be carried out where proximons are contrasted against known operons. Specifically, using the *Escherichia coli* K-12 as a gold standard predicted proximons will be compared against known operons and the cardinalities and configurations of their respective mappings will be measured. In particular, the metric of operon coverage will be analyzed to determine the extent to which proximons emulate actual operons. The reciprocal perspective will also be considered in order to determine the proportion of operon data that is not captured by proximons.

3. Applications: To demonstrate the utility of metagenomic proximons as collections of functional interactions, a protocol will be devised where proximons can be used as an informative source to infer broader functional modules through network formation on the basis of mutual functional annotations for any given set of metagenomic proximons that represent an environment and/or functional category of interest. These modules will be intended to characterize the functional relationships within data of interest and to facilitate functional comparisons between metagenomic datasets by way of set theoretic contrasts and/or quantitative analysis of various network features. However, it is important to reiterate that such modules ultimately represent a

computational proof of concept and validating the veracity of these predicted modules using corroboration by wetlab experimentation or other similar undertakings is beyond the scope of the work presented in this thesis.

Chapter 2

Representation: The MetaProx Database

† *The following chapter contains previously published material*.[5]

In this chapter I address the challenges related to the computational representation of proximon data, while the following chapters are devoted to their utilization. Given the volume and format of the available metagenomic gene data, addressing the representation, storage, and effective dissemination of proximon data became a necessary pursuit in order to drive the investigations carried out in subsequent chapters. Here, I explore the factors that affect the modeling of biological data, such as genes and their interactions, and propose a novel object-oriented approach to storage, retrieval, and deployment that is inspired by the recent emergence of the big data trend that currently dominates the Life Sciences. In particular, the utility and feasibility of these ideas are demonstrated through the development of a publicly available online database.

[5] Vey G, Charles TC (2014) MetaProx: the database of metagenomic proximons. Database (Oxford) 2014: bau097 (see Appendix D).

2.1 The Big Data Challenge in Computational Biology

The exponential increase in computing capacity[6] that has occurred during recent decades has revolutionized many facets of science. Biology has been particularly impacted by the advent of computationally driven fields such as the *omics* fields discussed here. Driven in conjunction by increments in next generation sequencing technologies, it is now the status quo for computational biologists to handle volumes of data that require interpretation and processing vastly beyond manual human capabilities: Thus, a new era of big data has emerged in biology and many other fields. Similarly, applications, operating systems, and even programming languages have begun to progress in a direction that allows greater usage and accessibility by non-computational users. Given that the applications of metagenomic functional interactions already involve computational protocols, accommodating the current climate of big data represents both a necessity and an opportunity, with respect to how research is implemented and what new discoveries are now possible. As a result, several principal challenges need to be addressed in order to optimize the current prospects for research involving the use of metagenomic functional interactions.

2.1.1 Dissemination and Representation

Trends such as cloud computing and cluster-based computing have shaped recent attitudes concerning the dissemination of biological data (Schadt et al., 2010) and spawned novel perspectives of utility supplied resources such as *Data as a Service* where data are provided on demand to any user under a provider/consumer model where the provider is not concerned

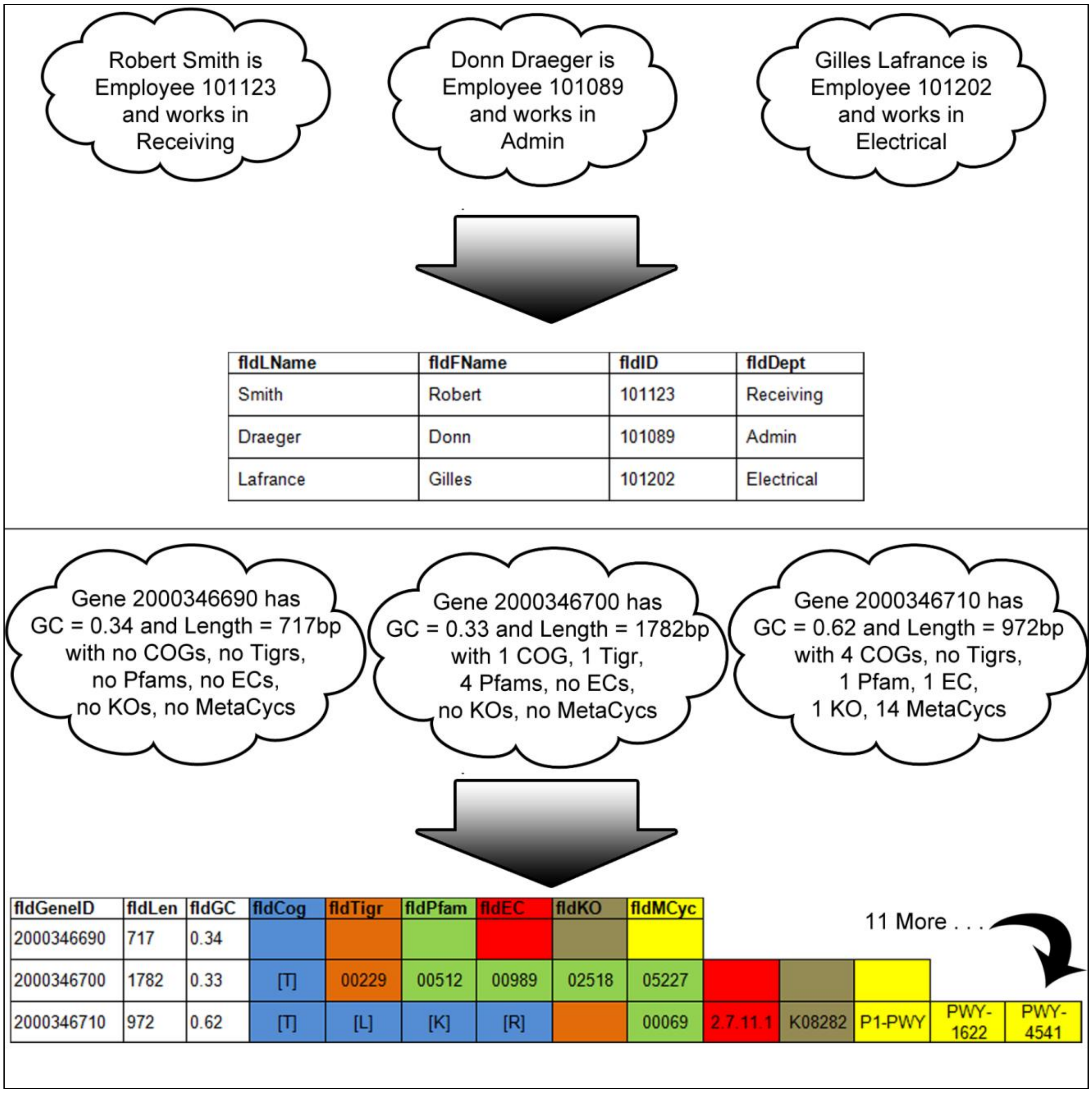
[6] See Moore's Law, Kryder's Law, and Nielsen's Law

with the geographic location or organizational status of the consumer (Dai et al., 2012). Currently, online databases remain an effective and popular means to publish and offer distribution of specialty data (Howe et al., 2008). Given that there are no existing databases that deal with the prediction, characterization, and warehousing of metagenomic functional interactions, the establishment of such a resource represents a keystone venture.

While big data has escalated a hardware arms race featuring petabyte-scale storage capacities, the efficacy of the underlying data representation is often questionable. In fact, this issue has been largely ignored in favour of simply throwing bigger and better hardware at challenges that could be dramatically alleviated by a more thorough understanding of data representation options and consequences. In the case of biological data, the crux of the representation problem rests in the fact that these data typically are not amenable to the tabular representations[7] that are required for a relational data model (O’Driscoll et al., 2013). In particular, relational models that have dominated business domains are very effective for portraying data where each record has regular and recurrent fields. In contrast, biological data can be highly variable both in the number and types of properties that need to be represented (see Figure 2-1). Therefore, investigating the factors governing the effective and economic[8] representation of metagenomic functional interactions is just as important as devising an online resource to store and disseminate them.

[7] A table of records where each record is a row of columns and each column represents a particular field or property

[8] Economic with respect to computational resources


Robert Smith is Employee 101123 and works in Receiving
Donn Draeger is Employee 101089 and works in Admin
Gilles Lafrance is Employee 101202 and works in Electrical

fldLName	fldFName	fldID	fldDept
Smith	Robert	101123	Receiving
Draeger	Donn	101089	Admin
Lafrance	Gilles	101202	Electrical

Gene 2000346690 has GC = 0.34 and Length = 717bp with no COGs, no Tigrs, no Pfams, no ECs, no KOs, no MetaCycs
Gene 2000346700 has GC = 0.33 and Length = 1782bp with 1 COG, 1 Tigr, 4 Pfams, no ECs, no KOs, no MetaCycs
Gene 2000346710 has GC = 0.62 and Length = 972bp with 4 COGs, no Tigrs, 1 Pfam, 1 EC, 1 KO, 14 MetaCycs
11 More . . .

fldGeneID	fldLen	fldGC	fldCog	fldTigr	fldPfam	fldEC	fldKO	fldMCyc					
2000346690	717	0.34											
2000346700	1782	0.33	[T]	00229	00512	00989	02518	05227					
2000346710	972	0.62	[T]	[L]	[K]	[R]		00069	2.7.11.1	K08282	P1-PWY	PWY-1622	PWY-4541

Figure 2-1 Relational data modeling. *The upper panel shows employee data with regular and recurrent properties being modeled into a relational data table with four fields containing atomic data values. The lower panel shows data from metagenomic genes with an attempt to model a corresponding relational data table. While some properties are regular and recurrent (in white), others are irregular and variable (in colour) and prevent the materialization of fixed fields that contain only atomic data values.*

2.1.2 Large-Scale Data Analysis Protocols

Large-scale data require analysis protocols that are capable of iterating over them and condensing knowledge from information. Again, hardware-centric solutions are popularly asserted including clouds, clusters, and GPUs[9]. The alternative to hardware-based strategies is to dedicate research and effort toward algorithmic and implementational advances (Schatz et al., 2010). While this avenue of research has been largely ignored in favour of the aforementioned hardware arms race, there remain serious obstacles to implementing and using analysis protocols that rely on 'big hardware' to handle big data. In the case of parallelization, only certain types of problems can be effectively ported to the GPU environment and such a migration involves the use of specialized programming languages like CUDA[10] that require domain-specific expertise (Schadt et al., 2010). Similarly, cloud computing has been criticized for a variety of concerns ranging from privacy and security to the induction of dependence upon its services (Pearson & Benameur, 2010). Therefore, protocols for the analysis and utilization of metagenomic proximons should be constructed with respect to the previously discussed considerations and should challenge the veracity of the current climate of overbearing hardware requirements, rather than acquiescing to what remains a largely rhetorical stance on computation.

2.2 Computational Representation of Proximon Data

Currently, much interest exists in the field of computational biology regarding the effective storage, dissemination, and harnessing of large datasets. In particular, there is a concern that

[9] Graphics Processing Units: GPUs are intended as a low cost parallel computing alternative to the conventional use of CPUs (Central Processing Units)

[10] Compute Unified Device Architecture: a proprietary programming model for NVIDIA GPUs

the current tools and approaches no longer scale up to the present volume of data, thus resulting in a bottleneck in the synthesis of knowledge from data (Marx, 2013). Metagenomic data are no exception to this trend with open-access reads in the Sequence Read Archive (SRA) (Leinonen et al., 2011) exceeding 100 Terabases by 2011, with metagenomic sequences accounting for 11% of all bases (Kodama et al., 2012). Open-access reads in the SRA as of June 2014 totaled more than 1,200 Terabases. Although the functional annotation and analysis of these data are crucial, the tools currently available to accomplish these tasks have not evolved to match the rate of data generation capabilities (Prakash & Taylor, 2012). Therefore, the development of protocols and tools that can capitalize on the vast availability of metagenomic data represents a major goal for computational biologists.

The prediction of metagenomic operons offers a means to reveal functional interactions in the absence of knowledge about orthologous relationships (Vey & Moreno-Hagelsieb, 2012; Vey & Moreno-Hagelsieb, 2010), thereby potentially driving numerous research interests in functional metagenomics. Therefore, the effective computational representation of metagenomic functional interactions (i.e. proximons) combined with the founding of a publicly available data source would offer a means to facilitate these kinds of research efforts. Although established resources already exist with respect to predicted operons from genomic sources (Pertea et al., 2009; Taboada et al., 2012), I am not aware of any analogous tools that operate at the metagenomic level. Consequently, I have developed MetaProx: the database of metagenomic proximons. MetaProx provides a searchable repository of proximon objects conceived with the goal of accelerating research involving metagenomic functional interactions (see Applications). MetaProx currently includes 4,210,818 proximons consisting of 8,926,993 total member genes (see Data Generation, Section 2.3.2). The following sections describe the implementation, deployment, and applications of the MetaProx database.

2.3 Implementation

Relational databases are based on an underlying relational model (RM) and they traditionally offer numerous strengths such as low data redundancy, data consistency, and physical data independence (Ward & Dafoulas, 2006). In addition, logical database independence and expandability combined with the general ease and robustness of query operations permit relational databases to support a broad range of purposes (i.e. views) and be accessible across a wide range of skillsets (Ward & Dafoulas, 2006). Furthermore, from an implementation perspective, the RM rests on a formal mathematical basis[11] and relational database design is well described through a formal normalization process (Ward & Dafoulas, 2006). However, there are several key facets of data representation and entity-relationship (ER) modeling[12] that are not effectively portrayed by the RM.

Relational databases cannot directly represent many real-world objects, particularly those that are complex and composed of other objects. This stems from the inability of the RM to distinguish between entities versus relationships because relationships identified during ER modelling do not endure using direct representation in the RM. In other words, the RM does not offer a direct means to recover the relationships between entities, such as the *Works In*

[11] Tuple relational calculus is a declarative language designed to provide a formal description of a domain or data model.

[12] Entity–relationship modeling uses entity types to describe objects or things while specifying the relationships that can exist between instances of given entity types, in order to describe a specific domain of knowledge.

relationship between *Employee* and *Department* entities. Consequently, this requires users to possess prior knowledge about such relationships in order to compensate for the resulting semantic overloading where relations from the RM are used to represent both the entities and the relationships from the corresponding ER model (Ward & Dafoulas, 2006). Similarly, the decomposition of entities via standard normalization can lead to excessive fragmentation that manifests as spurious relations that do correspond well to actual real-world entities (Ward & Dafoulas, 2006). This type of fragmentation can also impose numerous join requirements for queries, in order to recover the original information describing a given entity (Ward & Dafoulas, 2006). In addition, standard normalization requirements, particularly First Normal Form, mandate that all attributes in the RM must be atomic. Therefore, it is not possible to directly include a composite attribute in the relational schema, such as *Name*, which might contain constituent member attributes like *First Name* and *Last Name* (Ward & Dafoulas, 2006). Similarly, it is not possible to directly represent list or sets in the RM, even if the members of such structures are in fact atomic in nature. Furthermore, the range of available datatypes is limited and there is no way to create user-defined types intended to meet to specific application needs (Ward & Dafoulas, 2006). In addition to being ineffective at portraying complex and composite objects, the RM cannot depict hierarchical or inheritance associations. For example, there is no way to denote that entities like *Employee* and *Student* both inherit the attributes of a mutual parent entity like *Person* or that the set of all *Employees* is a subset of all *Persons*. Finally, the RM is unable to enforce domain-specific organization constraints, such as setting an upper bound for the number of students that can be enrolled in a course (Ward & Dafoulas, 2006).

In comparison, object-oriented databases provide flexible and direct modelling of real-world entities, which can be composed of simple attributes but also list, sets, or even other objects, while relationships are encapsulated directly within objects via their exposed methods. Similarly, concepts such as hierarchical relationships and inheritance follow naturally from the object-oriented paradigm. In addition, the general ability to accommodate completely novel user-defined types plus domain-specific organization constraints offer robust utility that is not available in the RM. However, object-oriented databases do not necessarily support complex queries to the same extent afforded by the RM and enforcement of reliability paradigms such as ACID properties[13] require additional programmatic implementation by the application.

For the present purposes, there are several key facets of the data and the queries that operate on them that have shaped the implementation of MetaProx. The data are composite and irregular with metagenomic genes exhibiting a high degree of variability in the number and type of annotations that they contain (see Table 2-1). Next, the required data retrieval patterns are known. That is, a generalized and robust query system is not required because any query result will always be a collection of proximons that is retrieved according to functional and/or environmental qualifiers. Finally, MetaProx as an application needs to read data in order to provide results to a user but it never needs to provide write access to the underlying database. Therefore, ACID considerations have no bearing the database requirements. Given the semi-structured and composite features of the data, the fixed data

[13] ACID (Atomicity, Consistency, Isolation, Durability) is a set of database transaction properties aimed at enforcing validity in the event of errors, such as a power failure.

Table 2-1 Sparsity of metagenomic functional annotations. Excerpts from several metagenomic gene annotation records are shown with counts for their respective functional annotations across six different annotation categories. All records were obtained from the Sludge/US Phrap Assembly metagenome, publicly available from the IMG/M: Taxon Object ID 2000000000.

Record	COG Cat.	Pfam	TIGRfam	KEGG Mod.	MetaCyc Path.	EC Num.
2000000060	2	2	-	-	-	-
2000000140	2	1	-	2	17	1
2000000300	4	1	-	-	-	-
2000000320	-	-	-	-	-	-
2000000360	-	-	1	-	-	-
2000001710	1	2	-	-	3	1

retrieval pattern, and the read-only nature of the user application, MetaProx has been implemented as serialized object repository, rather than a relational database.

Beyond structuring considerations, the present approach to data representation has been inspired by specifications like the Common Object Request Broker Architecture (CORBA) (Object Management Group, 2012) and a *Data as Data* policy is advocated here where the same serialized objects persist across all levels, including the database layer, the application layer, and even for the materialization of saved user files. This is in contrast to the lingering perception that biological data should be both transformable and humanly readable, considerations that fuel the persistence of verbose textual and markup-based representations. It is asserted here that data-centric research could be generally accelerated if developers begin to adopt the exchange of serialized objects, rather than maintaining the status quo of inflating computational data into delimited text or markups and then deflating them again during subsequent computation.

2.3.1 Data Model

MetaProx uses a hierarchy of Java classes (see Appendix A for a formal UML[14] depiction) to support the representation and functionality of proximon objects where the manipulation or extraction of data occurs directly by way of a specified application programming interface

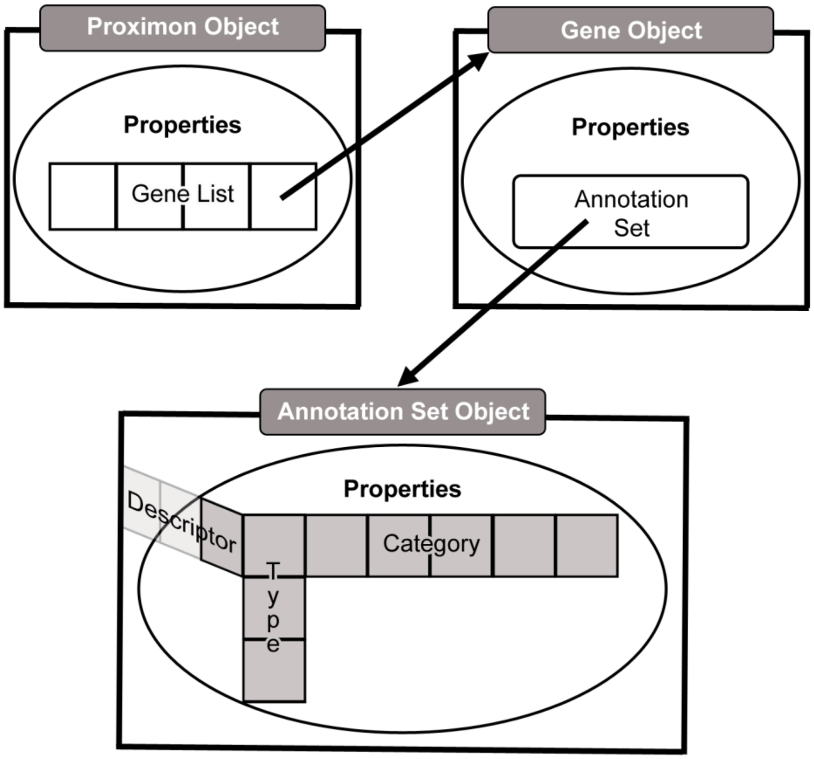


Figure 2-2 Abstract data model. *The top level proximon object is shown with a list of gene objects encapsulated among its properties such that a gene object encapsulates an annotation set object that subsequently encapsulates a three-dimensional list of annotation.*

[14] Unified Modeling Language

(API). Given the aforementioned irregularities of semi-structured data, a proximon object is essentially a multidimensional list where dimensionality is constant but the length and contents of a list at any given dimension are highly variable (see Figure 2-2). At the top level, a proximon object contains a list of gene objects that correspond to its member genes. In turn, each encapsulated gene object contains its own collection of functional annotations in the form of a variable length list of annotation types, each of which contains one or more categorical values and corresponding functional descriptors.

Queries execute by iterating over a subset of proximons where each proximon subsequently iterates over its member genes and in turn each member gene iterates over its particular collection of functional annotations. Specifically, a query object uses the API to perform comparison operations and/or check substring occurrences for each candidate proximon, in a manner similar to the db4o native query (Versant Corporation, 2014). Qualifying proximons are added to a sorted results queue and the queue is returned when all proximons are exhausted.

2.3.2 Data Generation

Proximon predictions were based on the *Escherichia coli* distance model for operon prediction in prokaryotes, as described by Moreno-Hagelsieb & Collado-Vides (2002). In their model, the authors examined pairs of adjacent genes within directons (WD pairs) in order to contrast pairs of genes within operons (WO pairs) against pairs at TU boundaries (TUB pairs)[15]. Specifically, using the intergenic distances of experimentally known WO

[15] A pair of genes consisting of the last gene in a given TU and the first gene in the next TU.

pairs versus known TUB pairs, Moreno-Hagelsieb & Collado-Vides (2002) calculated the log-likelihood of two adjacent genes being in an operon as the logarithm of the fraction of WO pairs divided by the fraction of TUB pairs containing genes separated by a distance within the interval, using an interval size of 10 base pairs. As a result, using only intergenic distances, they were able to discriminate genes within operons from those in different TUs with an accuracy of above 82%. Moreover, Moreno-Hagelsieb & Collado-Vides (2002) showed that the *E. coli* distance model performed equally well when applied to *Bacillus subtilis,* despite the evolutionary distance between these organisms. Further still, using operon data available at the time, the authors were able to demonstrate that the intergenic distance distributions of most genomes exhibit a characteristic peak between −20 and 30 base pairs, due to the presence of operons in the same range (Moreno-Hagelsieb & Collado-Vides, 2002). Thus, they were able to show the general extensibility of the *E. coli* distance model to prokaryotic genomes that do not have their own baseline data for TUB pairs, although exceptions for genomes such as *Halobacterium NRC-1* and *Helicobacter pylori* have been observed in other studies (Price et al., 2005).

The proximon data and corresponding metagenomic genes were derived from metagenomic data obtained from the Integrated Microbial Genomes with Microbiome Samples metagenomics database (IMG/M) (Markowitz et al., 2012) (see Appendix B). Specifically, proximons were generated from available metagenomic gene coordinates using a previously published metagenomic implementation of Moreno-Hagelsieb & Collado-Vides' original intergenic distance model (see Vey, 2013; Vey & Moreno-Hagelsieb, 2012; Vey &

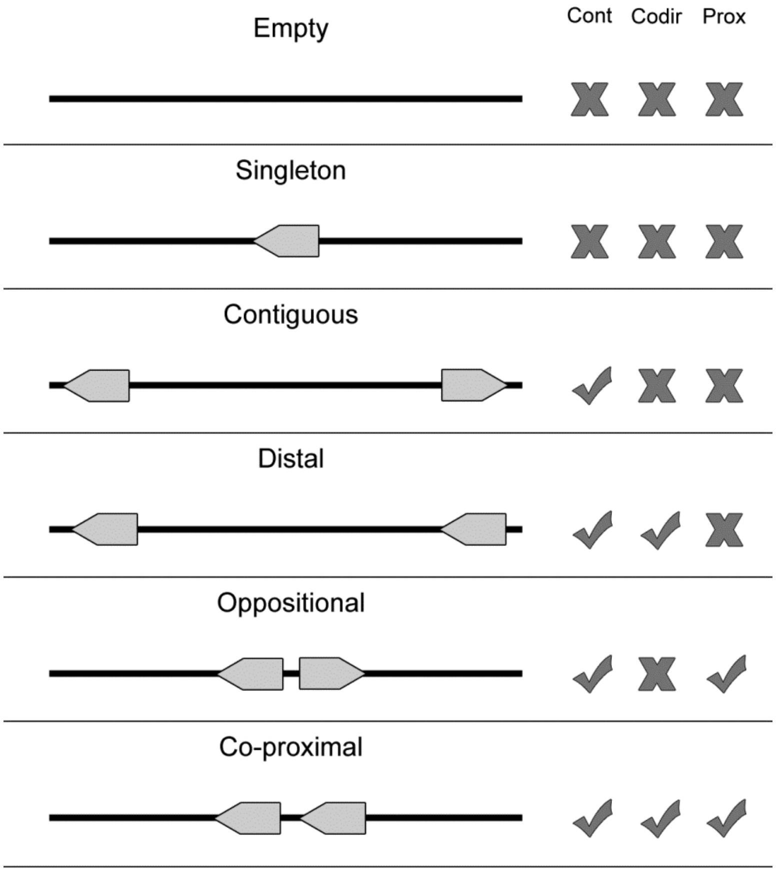
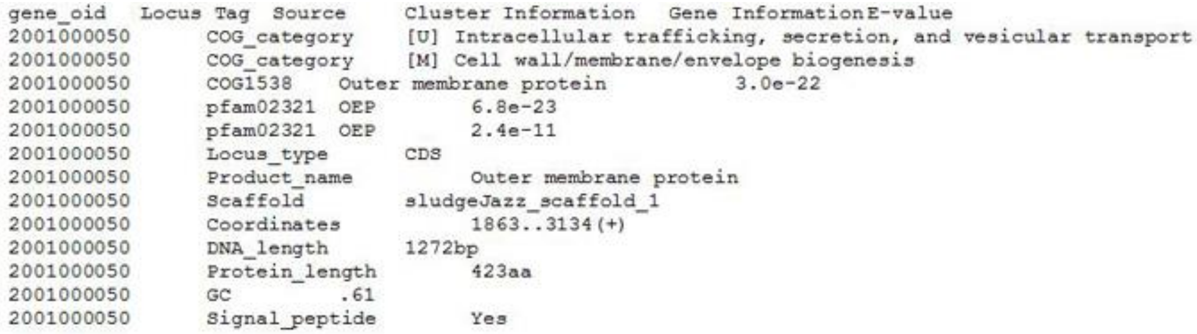


Figure 2-3 Proximon selection criteria. *Various configurations are shown for a metagenomic scaffold that contains either zero (Empty), one (Singleton), or two (all other cases) genes. Each configuration is considered with respect to whether or not it exhibits multiple contiguous genes (Cont), genes that are co-directional (Codir), and genes that are co-proximal (Prox). Only the last configuration meets all of the criteria required by the proximon definition.*

Moreno-Hagelsieb, 2010) intended for identifying metagenomic operon candidates based solely on the intergenic distances

between adjacent co-directional genes (see Figure 2-3). All proximons included in MetaProx were obtained using a minimum threshold of confidence that is equivalent to a positive predictive value of 0.90: In other words, 90% of the proximons are expected to represent true metagenomic operons based on evidence from known operons of *Escherichia coli* K-12 contained in RegulonDB (Gama-Castro et al., 2011). Specifically, this level of confidence corresponds to intergenic distances of co-directional genes falling within the window of -20 to 10 base pairs. However, it is important to point out that the accuracy of any predicted proximon is contingent upon the corresponding accuracy of the coordinates of its member genes and metagenomic gene prediction represents an inherently challenging task. For example, metagenomic gene prediction can be effected by the ability to correctly assemble metagenomic sequence reads into longer contigs and this process can be subsequently impacted by factors such as sequencing coverage and chimerism (Hoff, 2009). With this caveat in mind, specific proximon predictions were generated by parsing metagenomic data files using a computational pipeline, as described below.

For the metagenomes listed in Appendix B, corresponding tab delimited text files were downloaded from the publicly available IMG/M data repository. Each file contained information about protein coding genes occurring within a given metagenome, such as gene coordinates, strand indicator, and functional annotations (see Figure 2-4). For each file, gene data were parsed on the basis of known delimiters and regular expressions to produce a list of



```
gene_oid   Locus Tag  Source     Cluster Information   Gene InformationE-value
2001000050      COG_category    [U] Intracellular trafficking, secretion, and vesicular transport
2001000050      COG_category    [M] Cell wall/membrane/envelope biogenesis
2001000050      COG1538    Outer membrane protein          3.0e-22
2001000050      pfam02321  OEP         6.8e-23
2001000050      pfam02321  OEP         2.4e-11
2001000050      Locus_type      CDS
2001000050      Product_name           Outer membrane protein
2001000050      Scaffold        sludgeJazz_scaffold_1
2001000050      Coordinates            1863..3134(+)
2001000050      DNA_length      1272bp
2001000050      Protein_length         423aa
2001000050      GC         .61
2001000050      Signal_peptide         Yes
```

Figure 2-4 IMG/M sample record. *An excerpt from an IMG/M data file is shown where the contents describe features and annotations for a single gene.*

corresponding gene objects, stored as an in-memory representation for further processing. Next, the list of genes was parsed on the basis of co-direction and intergenic distance, *IGD* = *gene2_start* − (*gene1_end* + 1), with respect to contigs and their member directons (see Figure 2-5). This allowed genes to be combined into composite proximons and each proximon was successively added to a list of proximons in memory, for further processing. After each gene in the file had been processed, each proximon in the finalized list of proximons was serialized to produce a byte encoded representation that was subsequently materialized to external storage in corresponding file that constituted a database file system block. Block size was determined on a sliding scale where a given source file could produce

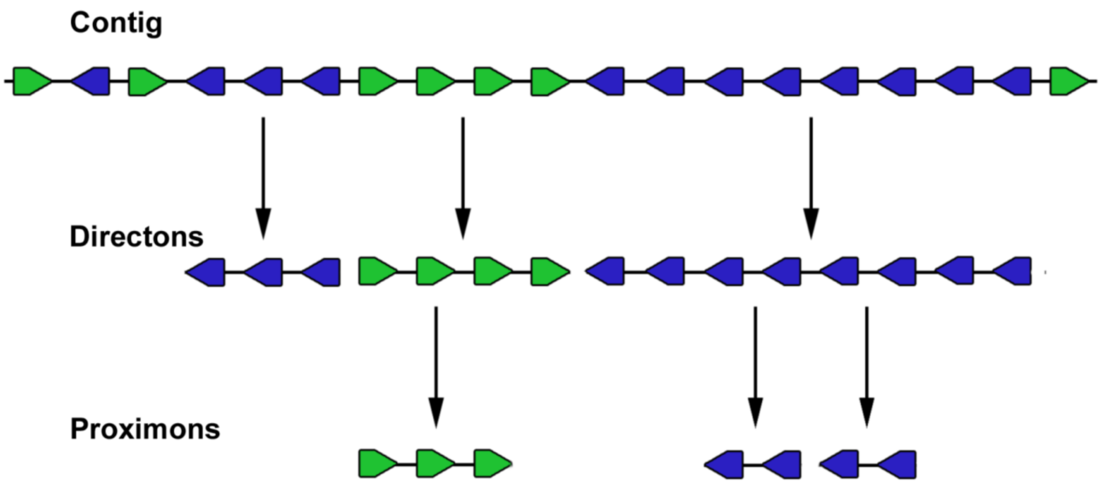


Figure 2-5 Contig hierarchy. *An abstract representation of the YNP19_C2070 contig from the hot spring microbial communities (Yellowstone National Park) is shown with respect to gene order and direction (relative gene length and intergenic distance are not depicted). The non-unary member directons are also shown along with their corresponding member proximons.*

a block size on the interval (0, 6500000] bytes. In general, if a source file required greater than 6500000 bytes of storage it was split into multiple blocks. However, there were 12

Table 2-2 MetaProx database composition. The database composition is shown according to proximon count and proportion (% of total count) versus metagenomic ecosystem and also for the categories within each respective ecosystem.

Ecosystem	Count	%	Category	Count	%
Engineered	246,919	5.9%	Bioremediation	48,111	1.1%
			Biotransformation	94,339	2.2%
			Solid waste	65,156	1.5%
			Wastewater	39,313	0.9%
Environmental	3,188,109	75.7%	Air	6,647	0.2%
			Aquatic	2,258,143	53.6%
			Terrestrial	923,319	21.9%
Host-associated	775,790	18.4%	Arthropoda	395,549	9.4%
			Birds	63,329	1.5%
			Human	5,075	0.1%
			Mammals	150,050	3.6%
			Microbial	5,183	0.1%
			Mollusca	27,761	0.7%
			Plants	128,843	3.1%

source files that were exempt from the block splitting policy in order to avoid very small trailing blocks. Upon completion of the overall batch process, the complete set of source files had been translated into a collection of database blocks that stored serialized proximons and their respective member genes. Finally, an index was generated that served as a mapping between metagenome features, such as ecosystem or category (see Table 2-2), and block identifiers so that the search space could be reduced whenever possible.

MetaProx currently consists of 4,210,818 total proximon objects and all data are categorized according to the taxonomic system used by the IMG/M (see Table 2-2). Proximon lengths ranged from 2 to 25 member genes with no proximons of length 22 or 23.

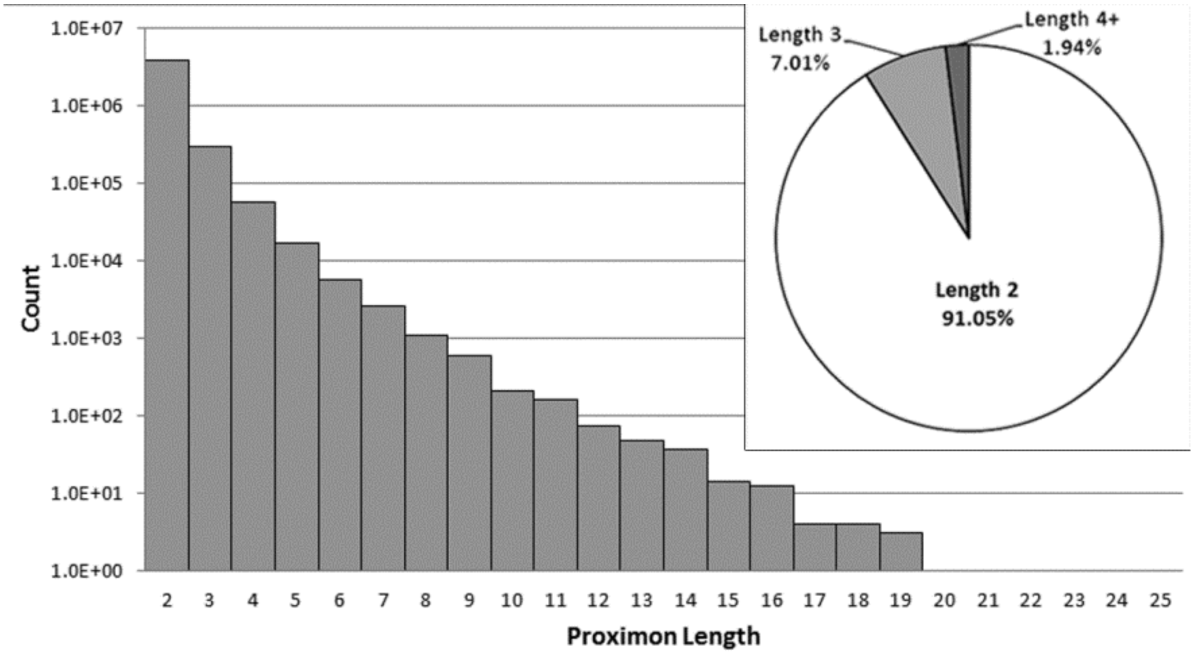


Figure 2-6 Distribution of proximon lengths. *The main panel shows the distribution of proximon lengths with respect to frequency of occurrence using a log (base 10) scale. The inset shows the relative proportion (%) of binary proximons, ternary proximons, and proximons with lengths greater than three member genes.*

Given that the complete set of proximons is composed of 8,926,993 total member genes, the vast majority of proximons are binary proximons (i.e. consist of two member genes) with only 9% of all proximons containing more than two member genes (see Figure 2-6).

2.4 Deployment

MetaProx is deployed using a distributed client-server model. Commonly, client-server interaction involves a client-side web interface that is used to request server-side processing that often involves subsequent retrieval from a backend database (Kurose & Ross, 2005) (see Figure 2-7). MetaProx, however, uses a distribution where the client owns the application (i.e. the search tool) that in turn invokes the server solely for access to the database (see Figure 2-7). Specifically, the MetaProx database responds to client requests by sending indexed blocks of proximon objects, thereby minimizing physical I/O while emulating a logical perspective where all data is readable by any given application instance (Ramakrishnan & Gehrke, 2003). The received blocks are subsequently subjected to additional query criteria that are carried out by the client's unique application instance, running on their own local machine. The benefit of this distributed approach is that clients provide many of their own resources (e.g. memory and CPU) therefore allowing them to take advantage of their own hardware capabilities while simultaneously alleviating the limitations of server-imposed quotas. For example, the maximum number of proximon objects that can be returned by any given search is greatly affected by the amount of memory that the client

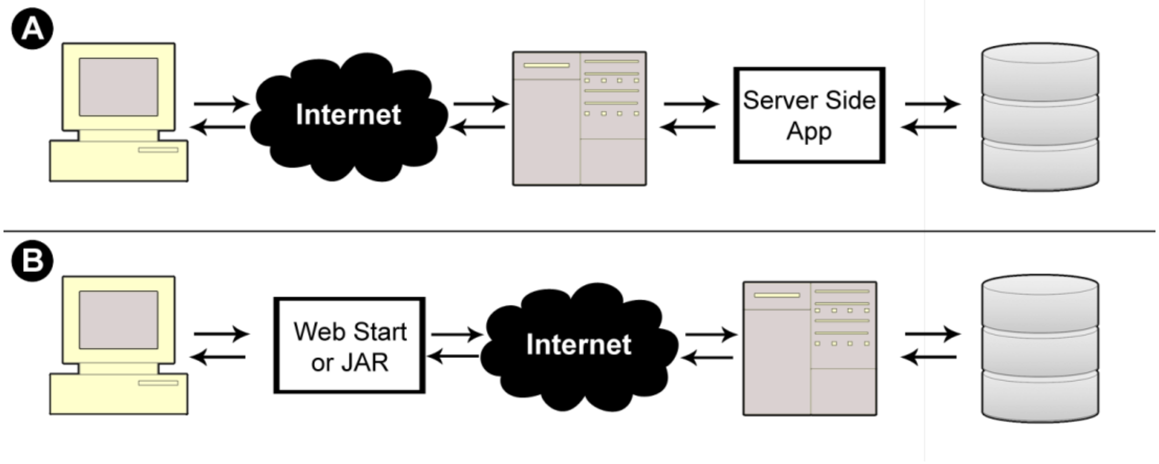
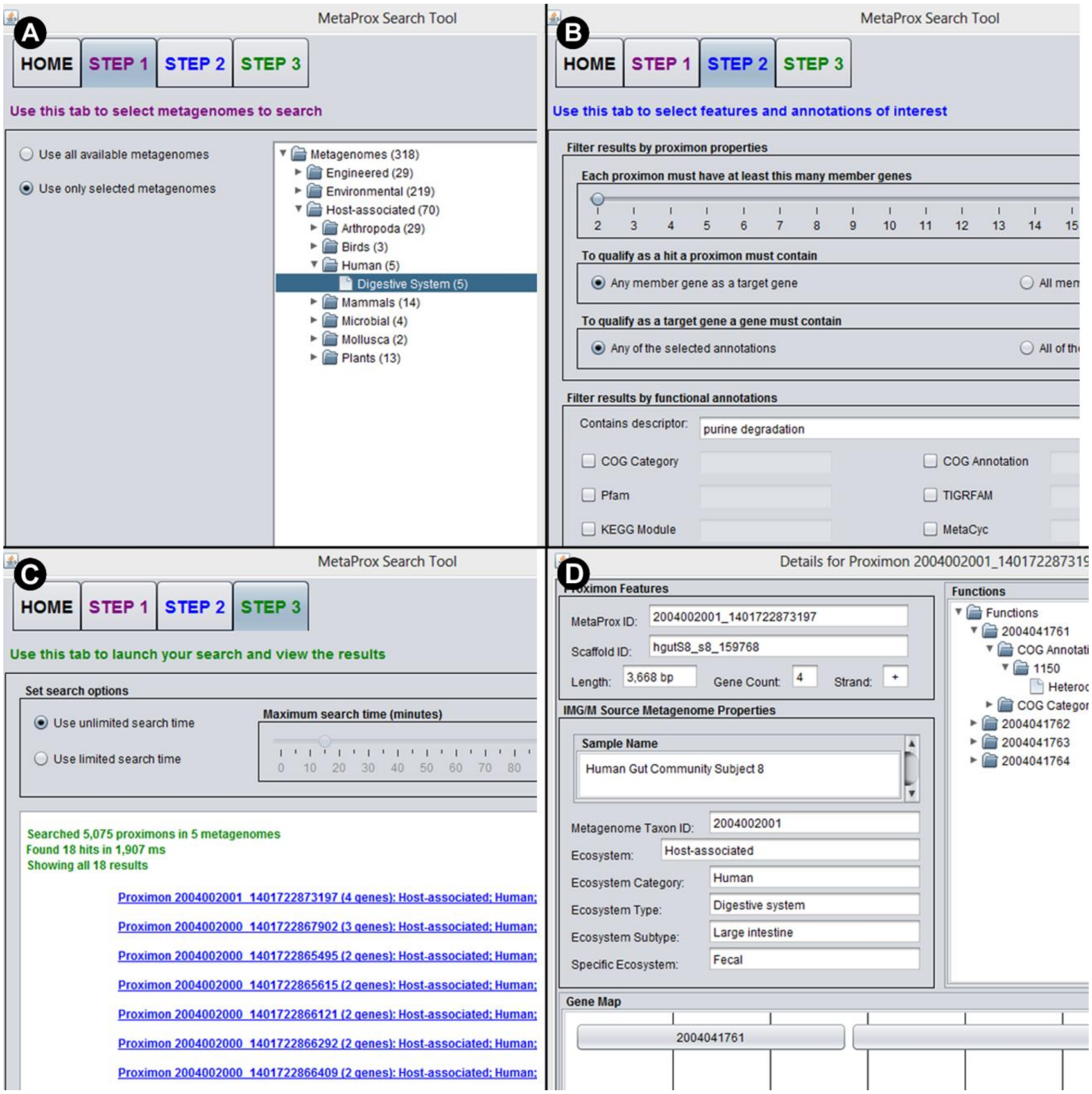


Figure 2-7 Application deployment perspectives. *(A) In a typical deployment scenario a web interface is used to invoke a server-side application that subsequently queries a backend database. (B) In contrast, MetaProx deployment provides a client-side JAR or Java Web Start application that directly interacts with a server-side database.*

has elected to allocate for the Java Virtual Machine. Using modest hardware, the performance of the search tool has been benchmarked and the search rate has been determined to be roughly 2,400 proximon objects per second, although the rate at any given time can be highly variable.

The MetaProx search tool is deployed as a JAR[16] that can be either downloaded from the website or launched directly from the browser using Java Web Start Technology (Oracle Corporation, 2011). Although the JAR is identical for both search modes, using a local downloaded JAR can typically circumvent the permissions and security issues that can arise from Java Web Start launches. In either case, the JAR will run a GUI application on the client machine that provides a simple stepwise search protocol (see Figure 2-8). Search results can be saved using the MetaProx serialized object format or alternatively saved as

[16] Java Archive: A compressed file format that aggregates multiple Java class files, along with associated metadata and resources.



Use this tab to select metagenomes to search

○ Use all available metagenomes

● Use only selected metagenomes

- Metagenomes (318)
  - Engineered (29)
  - Environmental (219)
  - Host-associated (70)
    - Arthropoda (29)
    - Birds (3)
    - Human (5)
      - Digestive System (5)
    - Mammals (14)
    - Microbial (4)
    - Mollusca (2)
    - Plants (13)

Use this tab to select features and annotations of interest

Filter results by proximon properties

Each proximon must have at least this many member genes: 2 3 4 5 6 7 8 9 10 11 12 13 14 15

To qualify as a hit a proximon must contain

● Any member gene as a target gene ○ All mem

To qualify as a target gene a gene must contain

● Any of the selected annotations ○ All of th

Filter results by functional annotations

Contains descriptor: purine degradation

□ COG Category □ COG Annotation

□ Pfam □ TIGRFAM

□ KEGG Module □ MetaCyc

Use this tab to launch your search and view the results

Set search options

● Use unlimited search time

○ Use limited search time

Maximum search time (minutes): 0 10 20 30 40 50 60 70 80

Searched 5,075 proximons in 5 metagenomes
Found 18 hits in 1,907 ms
Showing all 18 results

Proximon 2004002001_1401722873197 (4 genes): Host-associated; Human;
Proximon 2004002000_1401722867902 (3 genes): Host-associated; Human;
Proximon 2004002000_1401722865495 (2 genes): Host-associated; Human;
Proximon 2004002000_1401722865615 (2 genes): Host-associated; Human;
Proximon 2004002000_1401722866121 (2 genes): Host-associated; Human;
Proximon 2004002000_1401722866292 (2 genes): Host-associated; Human;
Proximon 2004002000_1401722866409 (2 genes): Host-associated; Human;

Details for Proximon 2004002001_140172287319

Proximon Features

MetaProx ID: 2004002001_1401722873197

Scaffold ID: hgutS8_s8_159768

Length: 3,668 bp Gene Count: 4 Strand: +

IMG/M Source Metagenome Properties

Sample Name: Human Gut Community Subject 8

Metagenome Taxon ID: 2004002001

Ecosystem: Host-associated

Ecosystem Category: Human

Ecosystem Type: Digestive system

Ecosystem Subtype: Large intestine

Specific Ecosystem: Fecal

Functions

- Functions
  - 2004041761
    - COG Annotati
      - 1150
        - Heteroc
    - COG Categor
  - 2004041762
  - 2004041763
  - 2004041764

Gene Map

2004041761

Figure 2-8 MetaProx graphical user interface. *Portions of the MetaProx graphical user interface are shown including the Source tab (A), the Target tab (B), and the Query tab (C). Clicking on a proximon link in the Query tab will display the corresponding Proximon Details panel (D) and clicking on a gene link in the Proximon Details panel will display the corresponding Gene Details panel (not shown).*

delimited text for further processing with other tools and pipelines. It is also possible to extract various annotation categories to expedite the construction of metagenomic annotation networks (Vey & Moreno-Hagelsieb, 2012) (see Applications, Section 2.5).

2.5 Applications

MetaProx has been designed to facilitate the retrieval of metagenomic functional annotations. For example, a user might want to gain insight about cellulase genes from soil metagenomes. The corresponding MetaProx search would provide proximons that meet these constraints and reveal information about the targeted genes but also about the genes that are potentially interacting with the targets. Furthermore, MetaProx offers features to save retrieved proximon data and also to extract specific functional annotations for easy construction of metagenomic annotation networks using network analysis software such as Cytoscape (Smoot et al., 2011).

Here, a working example is provided using the MetaProx search tool where purine degradation genes are contrasted from a network perspective using human digestive system metagenomes versus soil metagenomes. First, the source metagenomes are selected from the metagenome tree in Step 1: *Host-associated → Human → Digestive System* (see Figure 2-8). Next, the target genes are constrained by entering the keywords “purine degradation” in the descriptor textbox in Step 2 (see Figure 2-8). Executing this search (Step 3) will return 18 qualifying proximons composed of 39 member genes (see Figure 2-8). Using the *Save* command followed by the *Save Annotations Only* option allows functional annotations to be saved according to common annotation categories such as COG (Tatusov et al., 2003), Pfam (Punta et al., 2012), TIGRFAM (Haft et al., 2013), MetaCyc (Caspi et al., 2012), etc. Here

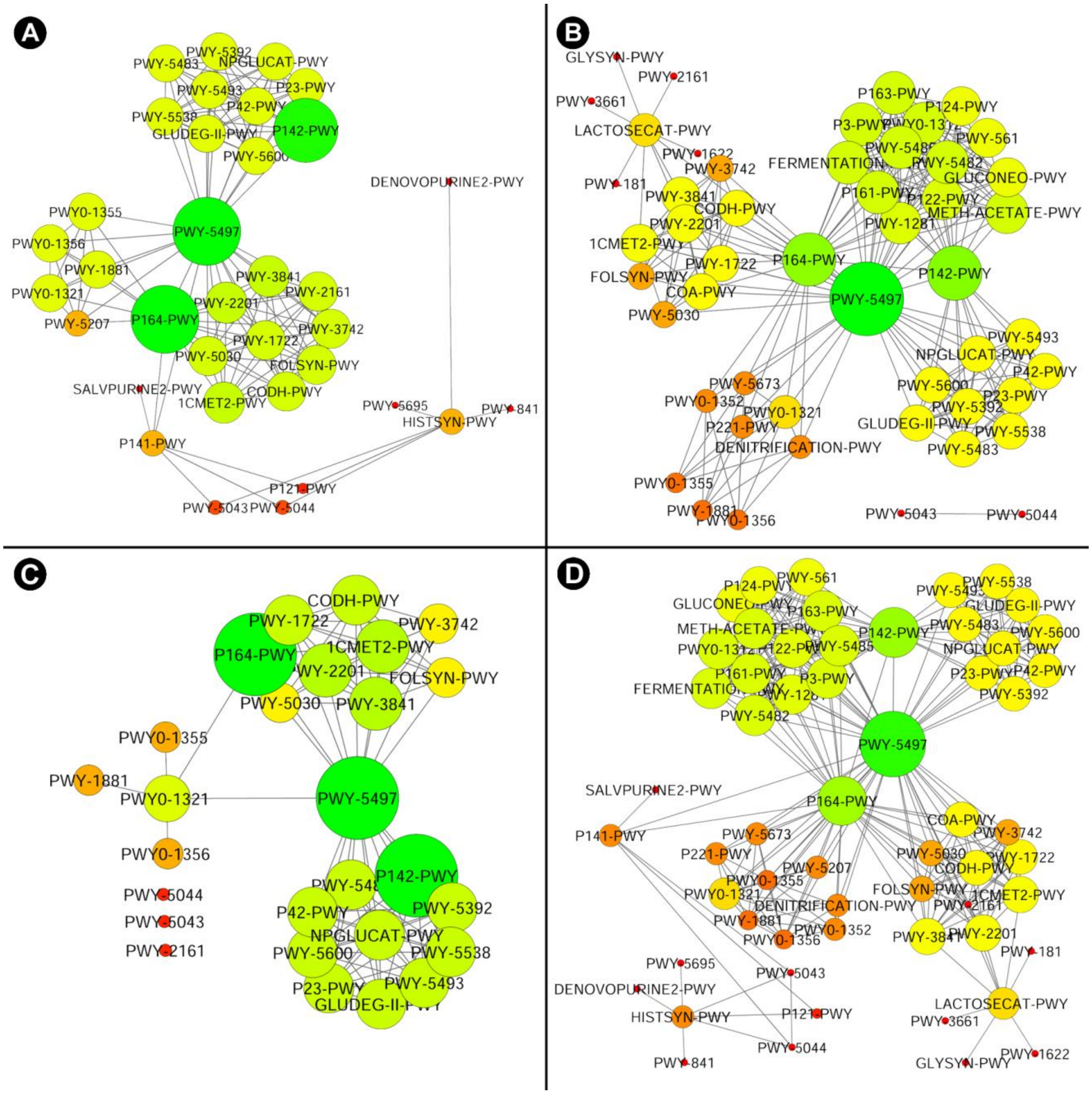


Figure 2-9 Purine degradation network. *Purine degradation networks are shown for MetaCyc pathways from human digestive metagenomes (A), soil metagenomes (B), their inter-section (C), and their union (D) where node diameter and brightness (greenness) increase with increasing edge count.*

the MetaCyc pathways were selected and their annotations were used to construct a metagenomic annotation network using Cytoscape 2.8.2 and the resulting network contains 35 nodes and 142 edges (see Figure 2-9). The previous search is repeated but new source metagenomes are selected from the metagenome tree in Step 1: *Environmental* → *Terrestrial* → *Soil*. The 44 qualifying proximons provide MetaCyc pathways that produce a network with 50 nodes and 254 edges (see Figure 2-9). These networks can be subsequently contrasted and their intersection (27 nodes and 99 edges) and union (58 nodes and 297 edges) are depicted in Figure 2-9. This example demonstrates the ease of producing novel functional interaction networks and it is estimated that a novice user could have accomplished this task in roughly half an hour, while an experienced user could have completed it in just a few minutes. The resulting interaction network can then lead to hypothesis generation and experimental validation.

2.6 Future Directions

Future directions for MetaProx include increasing the number of proximons contained in the database and expanding the functionality of the search and visualization tools according to user feedback. An increase in the number of available search settings is also planned in conjunction with additional result filtering options. Query optimization for serialized objects will also be a key focus of future development with the goal of reducing database search times. Similarly, a database block caching policy will also be considered.

MetaProx will also implement support for the JSON[17] format. This will include the ability to save search results in JSON, rather than custom delimited text, because JSON is highly portable due to the wide availability of parsing tools and utilities. For example, Gson, is an open source library developed by Google to provide conversion between Java objects and JSON. Moreover, a web service might be developed that would allow other applications to poll MetaProx. This would allow the retrieval of data for consumption in other processes where the format of the provided query responses would also use JSON. In general, MetaProx will be aimed at supporting robust data dissemination in the form of Java objects and/or JSON.

[17] JavaScript Object Notation (JSON): A human-readable text file format designed for the transmission of data objects consisting of attribute–value pairs

Chapter 3

Evaluation: Mapping Proximons to Operons

† *The following chapter contains previously published material*.[18]

In the previous chapter, the MetaProx data were derived under the assertion that proximons are useful for inferring functional linkages because their member genes are synonymous to operon member genes, with respect to a given degree of confidence. In this chapter, I corroborate this assertion by performing a formal validation aimed at measuring the extent to which proximons emulate actual operons. This is accomplished by using the *Escherichia coli* K-12 genome to compare proximons and operons within the same genome and observe the configurations and cardinalities among their corresponding mappings. A statistical analysis of operon coverage is also carried out, along with an examination of metagenomic directon pairs. I conclude by examining intergenic distance profiles in order to understand the extensibility of results from the model to general metagenomic data.

[18] Vey G, Charles TC (2016) An analysis of the validity and utility of the proximon proposition. Functional & Integrative Genomics 16(2): 215-220. (see Appendix D).

3.1 Introduction

A functional interaction can be interpreted as a mutually cooperative relationship that functionally links two or more genes and necessarily defines a state of functional association. Such arrangements are exemplified among the member genes of a given type of functional unit, such as the co-transcribed protein coding genes within an operon (Jacob & Monod, 1961; Miller & Reznikoff, 1978). In the case of functional metagenomics, such interactions can be used in a variety of contexts ranging from the inference of broad functional modules to the assignment of a putative function to an individual gene. For example, homology methods such as BLAST (Altschul et al., 1990), as well as ab initio protocols, can be used to identify metagenomic gene occurrences and potentially assign corresponding functional annotations. Using the coordinates of detected genes, metagenomic functional interactions can be subsequently predicted using an operon detection protocol (Salgado et al., 2000; Moreno-Hagelsieb & Collado-Vides, 2002) that has been previously demonstrated with metagenomic data (Vey, 2013; Vey & Moreno-Hagelsieb, 2012; Vey & Moreno-Hagelsieb, 2010). Next, the functional annotations of interacting genes can be used to derive networks that portray functional interdependence and modularity as depicted through various features of network connectivity (Vey & Moreno-Hagelsieb, 2012; Rhee & Mutwil, 2014; De Filippo et al., 2012; Liu & Pop, 2011). In addition, existing annotations can be used to infer putative functions for genes that lack an annotation but have functional linkages to other annotated genes by way of the guilt by association paradigm (Aravind, 2000; Oliver, 2000). Overall, the effective use of metagenomic functional interactions represents a key prospect for a variety of applications in the field of functional metagenomics.

Recently, the concept of the metagenomic proximon was proposed (Vey & Charles, 2014). Whether metagenomic or genomic in origin, a proximon is a series co-directional genes and therefore it is necessarily a type of directon (a contiguous span of co-directional genes). However, a proximon has the added constraint that all of its member genes are also co-proximal where minimal intergenic distance exists between any two consecutive member genes within the same proximon. Thus, for any given metagenome or genome the set of proximons will be a subset of the set of directons. Similarly, there will be a subset of proximons that represent true operons, where the complete set of operons can include additional non-intersecting elements. Thus, proximons serve as strong operon candidates as inferred by evidence from known operons of *Escherichia coli* K-12 contained in RegulonDB (Salgado et al., 2013). Moreover, the proximon represents a key conceptual demarcation that was motivated by previous works (Vey, 2013; Vey & Moreno-Hagelsieb, 2012; Vey & Moreno-Hagelsieb, 2010) involving the detection of metagenomic operon candidates. In particular, the previous metagenomic prediction process has been relegated to the use of co-direction and proximity while various genomic prediction protocols augment their selections with additional evidence such as equivalent arrangements of orthologous genes (Moreno-Hagelsieb & Janga, 2008; Janga & Moreno-Hagelsieb, 2004; Moreno-Hagelsieb & Collado-Vides, 2002) or functional relationships between known protein products (Taboada et al., 2010). Therefore, it is tenuous to imply or infer equivalence between metagenomic versus genomic operon candidates and this is reflected in the set theoretic relationship between proximons versus operons.

Proximons are well suited for use in metagenomic scenarios where supplemental information about orthology and/or gene function is often sparse (Vey & Charles, 2014). However, the extent to which proximons effectively emulate operons is currently unclear. In this chapter, I aim to shed light on the validity and utility of the proximon proposition. Here, operons from the *Escherichia coli* K-12 model organism are used as a gold standard for comparison against proximons predicted from the same genome. In turn, this contrast is used to establish the characteristics of proximons with respect to operon coverage and equivalence. I conclude by examining intergenic distance profiles in order to understand the extensibility of results from the model to general metagenomic data.

3.2 Methods

Protein-coding genes, proximons, and operons were obtained or predicted for the *Escherichia coli* K-12 MG1655 genome and a variety of comparisons were carried out in order to contrast genomic proximons against genomic operons (see Results). All file parsing routines and computational predictions were implemented using Java and run on a Gateway NV59 laptop using an Intel Core i3-330M processor.

3.2.1 Genes

Gene data for the *Escherichia coli* K-12 MG1655 genome were obtained from the National Center for Biotechnology Information (NCBI) FTP directory of bacterial genomes (NCBI, 2014). Specifically, the .ptt file was downloaded from the corresponding directory on July 7th 2014. This file included coordinate information and functional annotations for 4,140 protein-coding genes. The coordinate data were subsequently used to generate proximon predictions (see Proximons) used in this study.

3.2.2 Proximons

The gene data (see Genes) were used to predict genomic proximons using a process identical to the one previously described in Vey & Charles (2014) where co-direction and proximity were used based on the metagenomic operon detection process previously described in (Vey, 2013; Vey & Moreno-Hagelsieb, 2012; Vey & Moreno-Hagelsieb, 2010). Specifically, intergenic distance (IGD) was iteratively measured for consecutive genes in the same strand using the number of base pairs between the end of the current gene and the start of the next gene, as determined using the formula that was previously defined in Section 2.3.2: $IGD = gene2\_start - (gene1\_end + 1)$. A total of 556 proximons were predicted for the *Escherichia coli* K-12 MG1655 genome using a positive predictive value of 0.90 (i.e. 90% of the predictions were expected to represent actual operons from the same genome). Specifically, this level of confidence corresponds to intergenic distances of co-directional genes falling within the window of -20 to 10 base pairs.

3.2.3 Operons

The complete set of operons for the *Escherichia coli* K-12 MG1655 genome was downloaded from RegulonDB (Salgado et al., 2013) (Release 8.6) on July 7th 2014. This file included gene information and evidence rankings for 2,640 operons. However, the gene information included only the identity of the member genes without specific features or functional annotations. Therefore, operon member genes had properties transferred from the gene data (see Genes) if they had a matching identity, otherwise the operon was removed if it contained one or more anonymous member genes, leaving a total of 729 operons with fully recognized genes, where each operon contained at least two member genes. This reduction was necessary in order to evaluate the mapping of proximons to operons on a gene-by-gene basis (see Metrics).

3.2.4 Metrics

In order to measure the extent to which proximons represent operons, operon coverage was used as the primary metric. Operon coverage was defined as the quotient of the number of matching genes between an operon and a proximon divided by the total number of member genes in the operon:

$$c = \frac{|O \cap P|}{|O|}$$

However, in cases where an operon was covered by more than one proximon the definition of operon coverage was adapted to:

$$c = \frac{|O \cap (P_1 \cup P_2 \cup \ldots P_n)|}{|O|}$$

where { P_1, P_2, … P_n } was the set of covering proximons and each proximon was itself a set of genes. Therefore, operon coverage was measured as a real number on the interval [0, 1] where 0 represented no proximon coverage and 1 represented full proximon coverage. Similarly, the number of hits (i.e. covering proximons) required to produce the coverage score was also recorded as a secondary metric. Both coverage and hits were evaluated by iteratively matching each proximon from the complete set of proximons against each given operon.

3.2.5 Intergenic Distance Profiles

Intergenic distances (IGDs) were calculated for consecutive gene pairs within the same directon (WD pairs). For example, a directon with genes {*a*, *b*, *c*} would yield two WD pairs, namely *ab* and *bc*, where each pair would provide a single IGD. Specifically, this value was the number of base pairs (bp) between the end of the leading gene and the start of the trailing gene, as determined using the formula that was previously defined in Section 2.3.2: $IGD = gene2\_start - (gene1\_end + 1)$. For *Escherichia coli* K-12 MG1655, WD pairs were calculated using the same coordinate data that was used for proximon prediction (see Section 3.2.1) and the resulting 2,899 pairs were measured to determine their corresponding IGDs. Outliers were excluded when an IGD was <-400bp or >400bp, leaving a total of 2,733 values (94.3% of the original data). For the metagenomes, WD pairs were derived using the same coordinate data that was used to construct MetaProx (see Section 2.3.2) and the resulting 12,918,643 pairs were measured to determine their corresponding IGDs. Outliers were excluded when an IGD was <-400bp or >400bp, leaving a total of 12,766,020 values (98.8% of the original data). WD pairs were used instead of within operon pairs because MetaProx does not contain operon data and using within proximon pairs would not be informative because the IGDs between proximon member genes are necessarily constrained by the proximon definition itself.

3.3 Results

Proximons were mapped to operons and a variety of configurations were observed (see Figure 3-1). Nearly 40% of proximons were identical matches to exactly one operon where each member gene exhibited a one-to-one mapping between the proximon and its

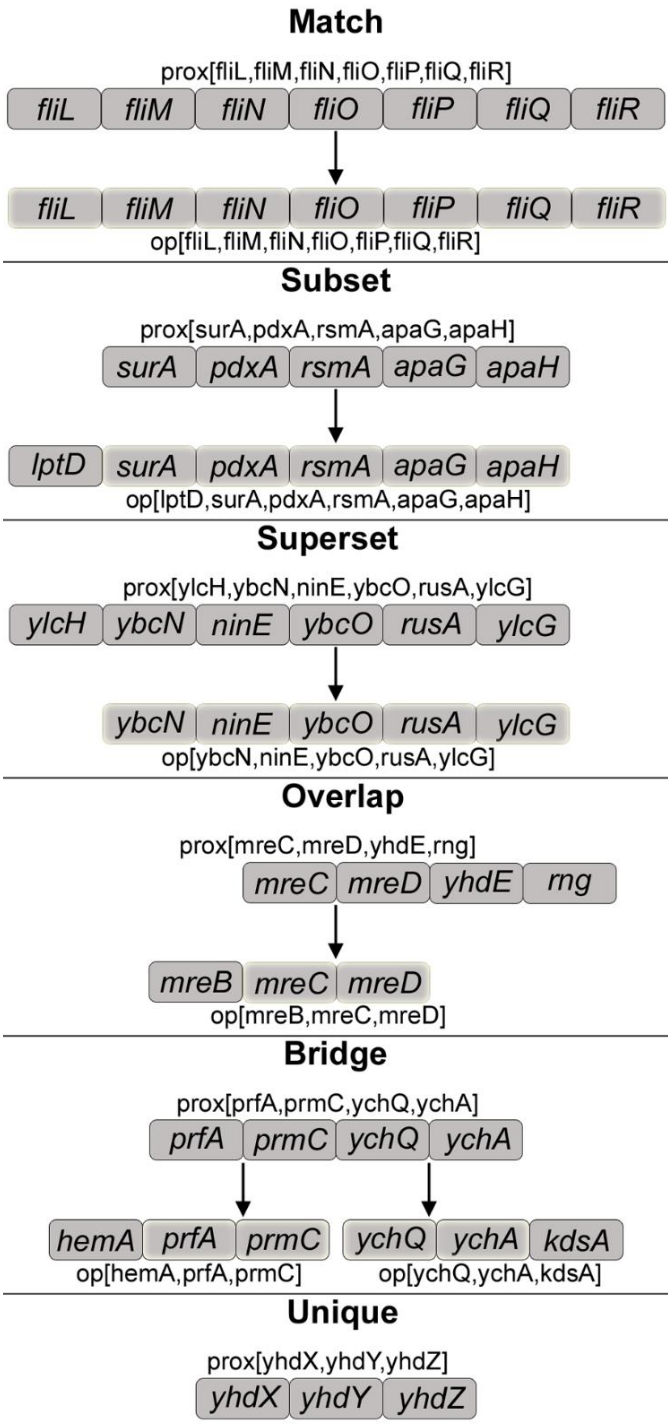


Figure 3-1 Proximon mapping configurations. *Examples of proximons are shown with respect to their corresponding operons where the mappings between respective sets of member genes exhibit various configurations including match, subset, superset, overlap, bridge, and unique (i.e. no mapping).*

corresponding operon. An additional 50% of proximons mapped to exactly one operon in a subset relationship where all of the member genes from the proximon mapped to member genes in the corresponding operon but the operon also contained one or more additional member genes. Approximately 1% of proximons exhibited a superset relationship where all of the genes from exactly one operon mapped to a corresponding proximon but the proximon also contained one or more additional member genes. Nearly 3% of proximons had an overlap with exactly one operon where the proximon and operon had an intersection of member genes but both the proximon and operon contained at least one exclusive member gene. Less than 1% of proximons showed a bridge configuration where the proximon shared an overlap with exactly two operons. The remaining 6% of proximons were composed solely

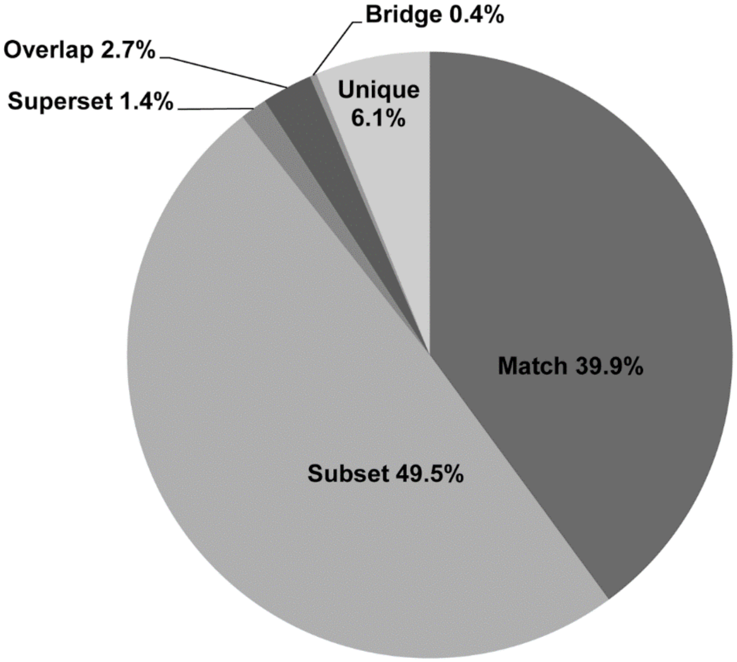


Figure 3-2 Proportion of mapping configurations. *The relative proportions for the observed categories of proximon mapping configurations are shown as percentage values.*

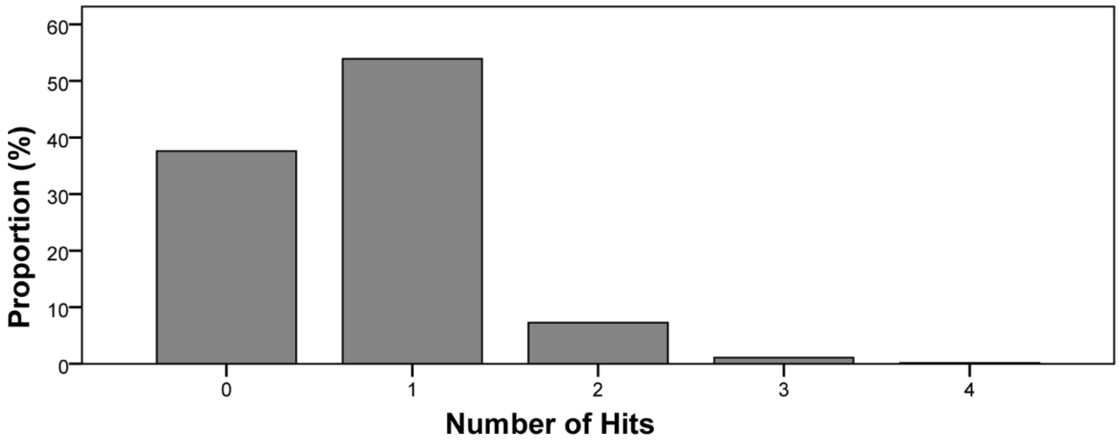


Figure 3-3 Distribution of operon hits. *The distribution of operon hits is shown where the horizontal axis represents the number of hits (i.e. proximons mapping to a single operon) and the vertical axis represents the relative proportion of operons occurring in each hit category as a percentage value.*

of exclusive member genes and had no match to any operons. Figure 3-2 shows the relative proportions for the various observed proximon mapping configurations.

While the vast majority of proximons (94%) mapped discretely to a single operon, in comparison, mapping from the operon perspective was much more variable with only 54% of operons mapping to only one specific proximon (see Figure 3-3). The large proportion of proximons existing in subset relationships with respect to their corresponding operons

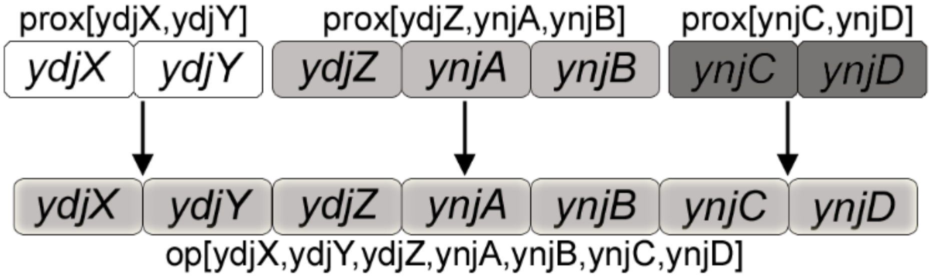


Figure 3-4 Multi-hit operon. *An example of a multi-hit operon is shown that has three hits where each of the corresponding proximons is fully contained within the multi-hit operon and these hits cumulatively provide 100% coverage of the operon.*

permitted mappings where operons were covered by multiple proximons (see Figure 3-4), with nearly 9% of operons exhibiting hits from two or more proximons. Of particular interest was the observation that almost 38% of operons had no hits at all.

For the 455 operons (62% of the total pool) that had one or more hits, the proportion of operon coverage was measured (see Methods). The proportion of coverage exhibited a non-normal distribution ranging from 0.22 to 1.00 with $\mu = 0.84$ and $\sigma = 0.21$. Figure 3-5 shows

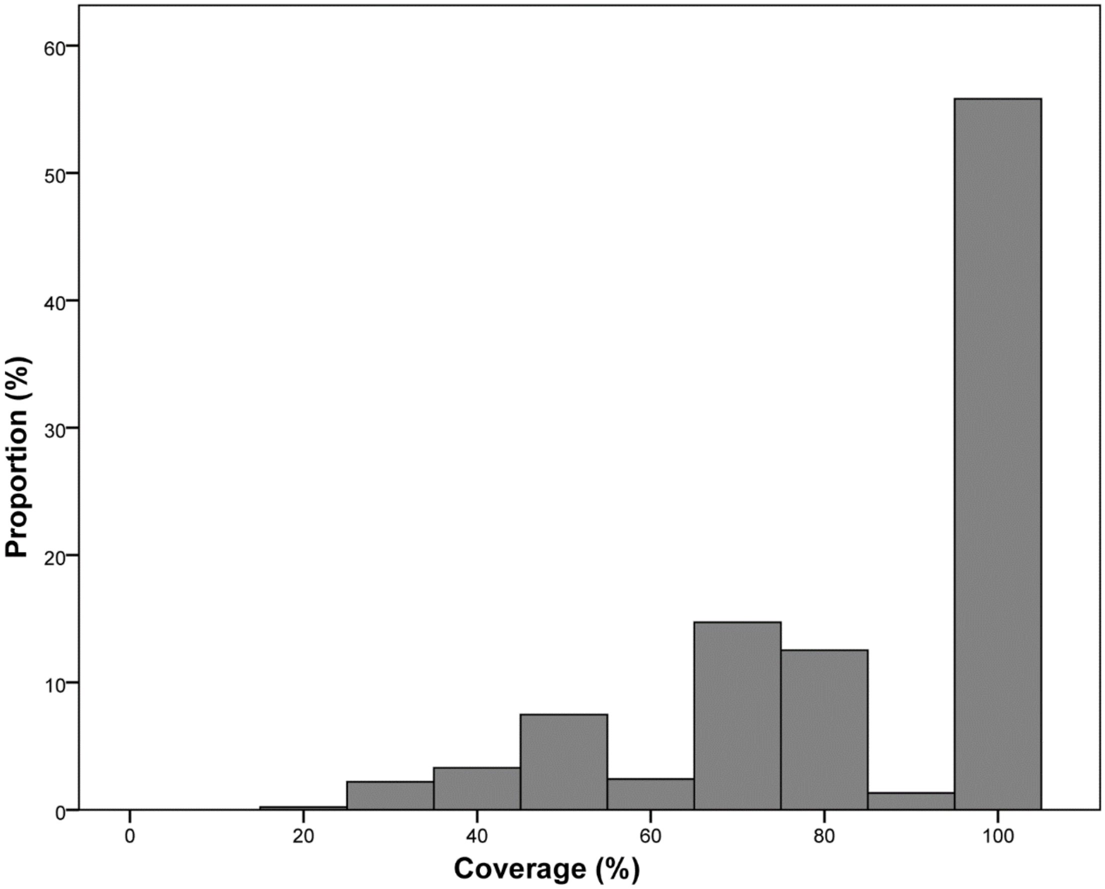


Figure 3-5 Distribution of operon coverage. *The distribution of operon coverage is shown where the horizontal axis represents bins depicting the proportion of operon coverage in 10% intervals and the vertical axis represents the relative proportion of operons in each coverage bin as a percentage value.*

the distribution of coverage converted to percentage and binned at 10% intervals, with 56% of cases falling into the highest bin. The difference between mapped coverage (i.e. coverage produce by mapping proximons to operons) and true coverage (i.e. every operon necessarily has a coverage of 1.00 with respect to itself) was analyzed using the Wilcoxon signed-rank test where each operon had its coverage score paired a constant value of 1.00. The analysis showed that true coverage was significantly higher than mapped coverage, $Z = -12.36$, $p < 0.001$.

In order to calibrate the extensibility of the present work to metagenomic data, intergenic distance (IGD) profiles were examined. Moreno-Hagelsieb & Collado-Vides (2002) had

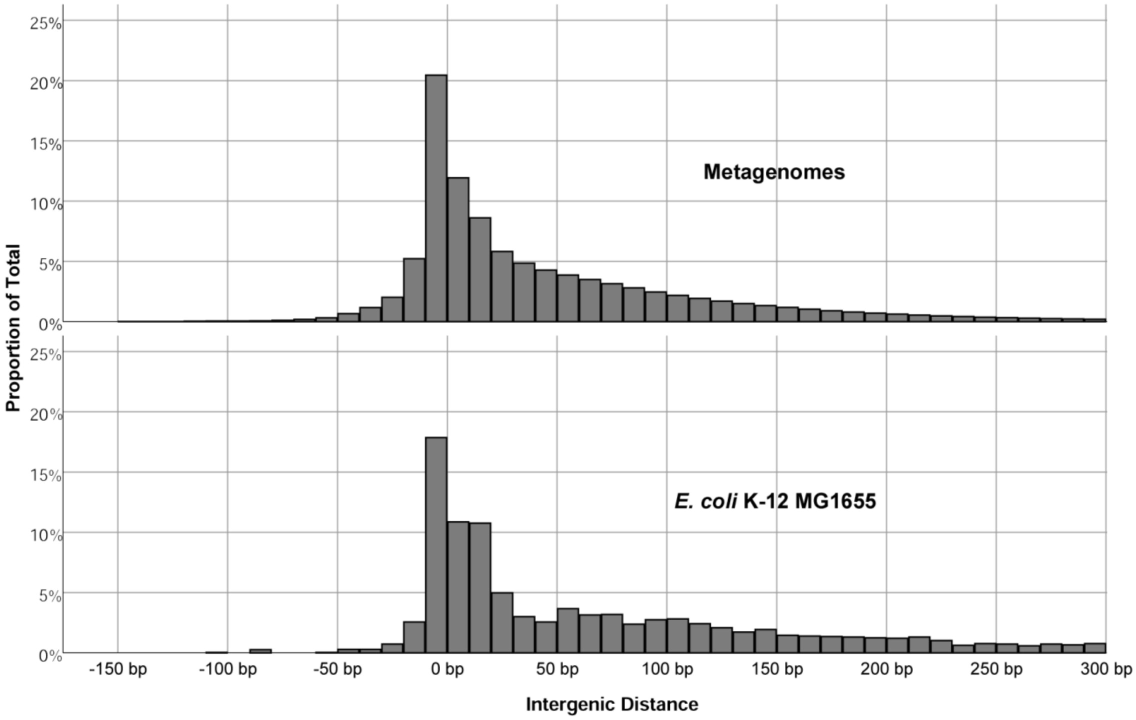


Figure 3-6 Intergenic distance distributions. *The distribution of intergenic distances is shown for within directon gene pairs for E. coli K-12 MG1655 and for the complete set of metagenomes from MetaProx.*

previously demonstrated the applicability of their IGD paradigm to *Bacillus subtilis*, as well as other prokaryotic data available at the time. By comparing IGD profiles, they showed that the distributions for IGD were highly similar between *E. coli* K-12 MG1655 and other prokaryotes (Moreno-Hagelsieb & Collado-Vides, 2002). Here, I extend this same comparison to contrast *E. coli* K-12 MG1655 against the metagenomic data from MetaProx. Figure 3-6 shows the comparison of IGD profiles with the metagenomes following the same trend as the other data sources previously examined by Moreno-Hagelsieb & Collado-Vides (2002) where the distribution closely resembles that of *E. coli* K-12 MG1655.

3.4 Discussion

The obtained results demonstrate that the vast majority of proximons do in fact map to operons and that these mappings include a variety of configurations and cardinalities. Moreover, 90% of all proximons exhibit a one-to-one mapping to a specific operon where the set of proximon member genes is either equivalent to the set of operon member genes or it is a subset. In other words, 90% of proximons are composed entirely of true operonic genes while the remaining 10% contain one or more superfluous genes. This finding demonstrates that proximon member genes offer a strong degree of confidence for inferring functional interactions, thereby confirming the utility of this approach in scenarios where gene position and direction are the predominant data. However, when conversely mapping operons to proximons, the results are far less conclusive with nearly 40% of operons having no corresponding proximons. This raises an important caveat in that while proximon data are both useful and reliable for inferring functional interactions, they capture only a portion of the total collection of functional linkages.

Since the results show that 6% of proximons are entirely composed of member genes that have no intersection with any operonic genes, the set theoretic perspective of operons as a subset of proximons (i.e. not every proximon maps to an operon) is confirmed. However, the results also show that this assertion is a simplification that requires elaboration based on several findings. First of all, given the large number of operons that do not have corresponding proximons, the broader set theoretic perspective shows an intersection where most proximons match some operons and conversely, the symmetric difference is composed of very few proximons but still a notable proportion of operons. Second, this perspective is a categorical perspective where operon-proximon correspondence is viewed as a simple binary state (i.e. match or no-match). However, the present results show that the configurations and proportions of coverage are variable on a member gene basis and corresponding operons and proximons can exhibit their own variety of set theoretic relationships when considered individually. Moreover, qualifying the existence of any intersection between sets of member genes as a match, even abstractly, must be tempered against the highly significant reduction in coverage when using proximons to emulate operons. Thus while it is valid to answer the question "*How do proximons relate to operons?*" with the response "*Proximons are a superset with respect to operons.*" it must be pointed out that this assertion is accurate from the perspective of the set of proximons but not from a broader perspective where both sets are fully considered. Again, it is crucial to reiterate that this is a categorical perspective where operons and proximons are treated as discrete elements rather than sets of member genes because when asking the same question from the perspective of any given mapping between a proximon and its corresponding operon then it is clear that a proximon is actually a subset of that operon, in the vast majority of cases. At this juncture it necessary to remember that the goal of proximon prediction is ultimately the inference of functional interactions by exploiting the features of proximity and co-direction exhibited by many operonic genes, but not necessarily operon prediction itself. Nevertheless, proximons can also be regarded as and utilized as a class of operon candidates.

The current results are based on exclusive comparison using only the *Escherichia coli* K-12 MG1655 genome and the specific scope and limitations of generalizing such an outcome to metagenomic data remain unclear, although the IGD profile results do support and extend the original IGD model put forth by Moreno-Hagelsieb & Collado-Vides (2002). However, this distance model is known to be less effective for certain genomes, such as *Halobacterium NRC-1* or for *Helicobacter pylori* (Price et al., 2005). Also, given the potential for genomic novelty within metagenomic data, there exists the possibility of alternative operon organization, as demonstrated in Kagan et al. (2008). An improved understanding of operonic configurations across a wide range of bacteria will be essential in order to determine how accurately metagenomic proximons represent actual metagenomic operons. Similarly, even based on the present *E. coli* results, the existence of multi-hit mapping configurations such as the one shown in Figure 3-4 suggest that there can be cases where metagenomic proximons can be concatenated to form larger entities. In particular, in a case where two proximons occur consecutively with no other interleaved genes and the proximons are also co-directional then such a case is a candidate for aggregation. However, additional knowledge characterizing the frequency and probability of these occurrences will be necessary in order to derive a confidence for these types of fusions.

Overall, the evidence presented here supports the validity and utility of the proximon for inferring potential functional interactions among member genes. This offers a powerful addendum to functional annotation strategies, particularly for metagenomic scenarios where functional inference by homology methods can be limited. In general, proximons represent reliable but conservative predictions of true operons, where a typical proximon is synonymous to an equivalent or truncated operon. As a result, proximon member genes can be used for the inference of functional interactions that can be subsequently used to drive functional annotation efforts. However, functional predictions derived from proximons represent only a portion of the total available linkages and whenever possible additional supplementary should be used to augment such predictions.

Chapter 4

Applications: Metagenomic Annotation Networks

† *The following chapter contains previously published material*.[19]

In this chapter I demonstrate how proximon data can be used to drive research in functional metagenomics. This is accomplished by aggregating functional interactions between member genes within their respective proximons to produce composite functional interaction networks. Moreover, any given network can be filtered so that its interactions are qualified with respect to a function and/or environment of interest. Networks can be further examined to infer member modules and they can also be compared to one another using a set theoretic perspective. Finally, I show how the annotations within modules can be subjected to various text-based analyses to examine annotative cohesion and semantic models.

[19] Vey G, Moreno-Hagelsieb G (2012) Metagenomic annotation networks: construction and applications. PLoS One 7(8): e41283. doi: 10.1371/journal.pone.0041283 (see Appendix D).

4.1 Introduction

The ubiquity of next-generation sequencing projects has vastly accelerated the accumulation of metagenomic sequence data. Recently, the Sequence Read Archive (Leinonen et al., 2011) exceeded 100 Terabases of open-access reads produced by next-generation sequencing efforts (Kodama et al., 2012). A common goal in attempting to understand the functional capabilities of newly sequenced microbial communities involves the annotation of putative genes through the assignment of biological functions. Such functional annotation relies heavily on homology-based annotation transfer using tools such as BLAST[20], HMMs[21], and motif finding (Wooley et al., 2010). In turn, the success of these approaches is necessarily bounded by the diversity of the reference databases that are used to find candidate annotations. However, it has been estimated that more than 99% of microorganisms are not amenable to common laboratory culturing conditions (Ferrer et al., 2005; Tringe & Rubin, 2005). This limited spectrum of microbial diversity combined with biases in applied research interests has yielded a skewed representation within sequence annotation databases (Pignatelli et al., 2008). Because metagenomes represent an attempt to gain access to the uncultured majority, homology-based annotation methods rooted in limited experimental knowledge about the functional roles of gene products are insufficient to adequately address the influx of unknown genes (Janga et al., 2011).

Given the difficulties in the annotation of individual metagenomic genes, the derivation and comparison of biological interaction networks represents a promising prospect for

[20] Basic Local Alignment Search Tool (BLAST)

[21] Hidden Markov Models (HMMs)

metagenomic data sources. Nevertheless, interaction networks can reveal vital information about functional organization and activity (Sun & Kim, 2011). For example, studies of interaction networks in *Escherichia coli* (Peregrin-Alvarez et al., 2009) and *Saccharomyces cerevisiae* (Hsu et al., 2011) have provided a systems perspective of these genomes by enumerating their respective functional modules. Recently there have been several attempts to capture metagenomic analogs of traditional interaction networks through the prediction of metabolic pathways and functional modules. MetaPath (Liu & Pop, 2011) uses prior knowledge of metabolic pathways in conjunction with metagenomic sequence data to predict the occurrence of metabolic pathways in metagenomic data sources. In contrast, Konietzny *et al.* (2011) used a Bayesian approach to find co-occurrence patterns for functional descriptors contained in microbial genome annotations in order to infer functional modules.

In the present work, proximons are used to derive functional interactions that are translated and categorized according to their associated functional annotations. The result is a collection of discrete networks of weighted annotation linkages that are subsequently examined for the occurrence of annotation modules that portray functional and hierarchical organization, with respect to a function and/or environment of interest. Finally, I show how the annotations within modules can be subjected to various text-based analyses to examine annotative cohesion and semantic models. However, while these analyses can yield insight into functional organization, they are provided as one possible example of numerous applications for network-based analyses of proximon data.

4.2 Materials and Methods

Metagenomic genes were parsed from downloaded raw data and used in a two-phase protocol consisting of network prediction followed by network translation. All operations were computationally implemented in Java and run on a Gateway NV59 laptop using an Intel Core i3-330M processor.[22]

4.2.1 Data Preparation

The raw data consisted of the complete set of public metagenomes available from the Integrated Microbial Genomes with Microbiome samples (IMG/M) metagenomics database (Markowitz et al., 2008) as of late August 2011. This included 224 datasets comprised of 40,189,394 total genes, distributed across 40,325,419 scaffolds (see Appendix B). The simulated datasets (simLC, simMC, simHC) were removed, as well as any datasets that did not contain gene coordinate information (DRU, VLU, Yorkshire Pig Fecal Sample 266, Yorkshire Pig Fecal Sample 267), since these coordinates are required for the network prediction phase (see Network Prediction). The remaining 217 datasets included 39,660,386 total genes, from which 207,097 rRNA genes were excluded, leaving an aggregate working dataset of 39,453,289 protein-coding genes.

[22] The metagenomic functional interactions used in this chapter were derived prior to the public release of the MetaProx database. They were produced in a manner very similar to the data generation protocol previously described for MetaProx but with some differences in source data and stringency for qualifying interactions. Therefore, the complete Materials and Methods section from the original paper is included here to facilitate experimental replication.

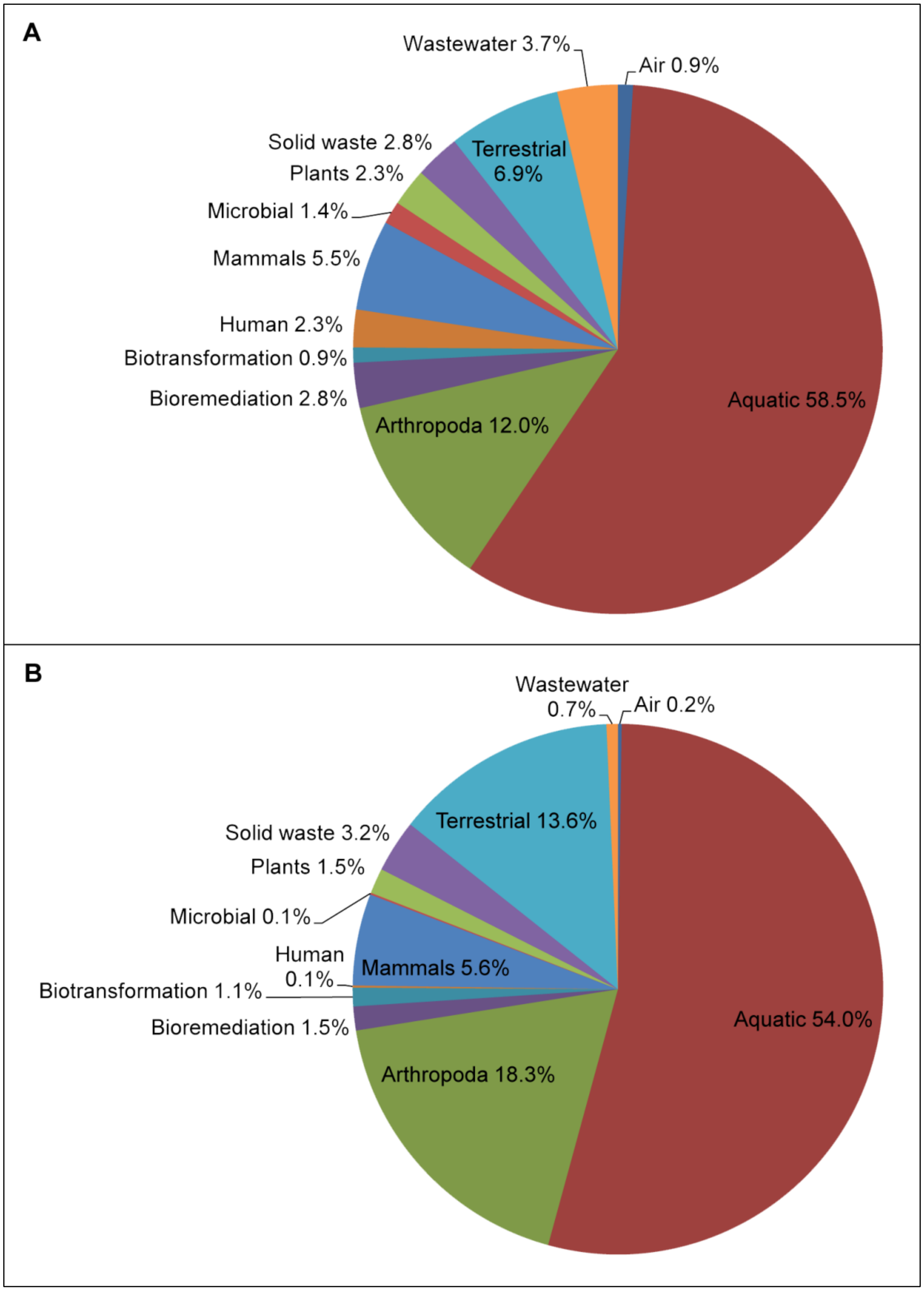


Figure 4-1 Data source diversity. *The relative proportions (%) of various data source types that were used (see Methods) are shown categorized according to IMG/M microbiome taxons at the class level. Panel A shows the proportions (%) with respect to the total number of datasets while Panel B shows the proportions (%) with respect to the total number of genes.*

The IMG/M was selected as the raw data source for three reasons: (i) It offered a very large amount of data from a diverse range of environments (see Figure 4-1). (ii) Virtually all of the annotated genes (> 99.5%) included information about their position and strand within the scaffolds in which they occurred. (iii) There was a high proportion of sufficiently assembled scaffolds such that multiple genes could occur within a single scaffold. This is in stark contrast to repositories that primarily offer data from short reads that frequently lack a single gene, let alone multiple genes. Overall, these factors are indicative of a current dichotomy in sequence databases: submitter-biased, such as MG-RAST[23] (Meyer et al., 2008), which cater to needs of authors that require a public depository of their data; *versus* query-biased, such as the IMG/M, which are focused on offering the expedient retrieval of data.

4.2.2 Network Prediction

Proximons were predicted in scaffolds containing two or more adjacent genes in the same strand (see Figure 4-2, Panel A) using a previously published method based on intergenic distances [$D = gene2\_start - (gene1\_end + 1)$], where the likelihood for two genes to be in the same proximon given the distance between them is assigned based on the ratio of known genes in operons to known genes in different transcription units found at such distance (Salgado et al., 2000; Moreno-Hagelsieb & Collado-Vides, 2002; Vey & Moreno-Hagelsieb, 2010). A minimum threshold of confidence was selected that is equivalent to a positive predictive value of 0.85 (meaning that 85% of the predictions are expected to consist of true

[23] MG-RAST metagenomics analysis server: http://metagenomics.anl.gov

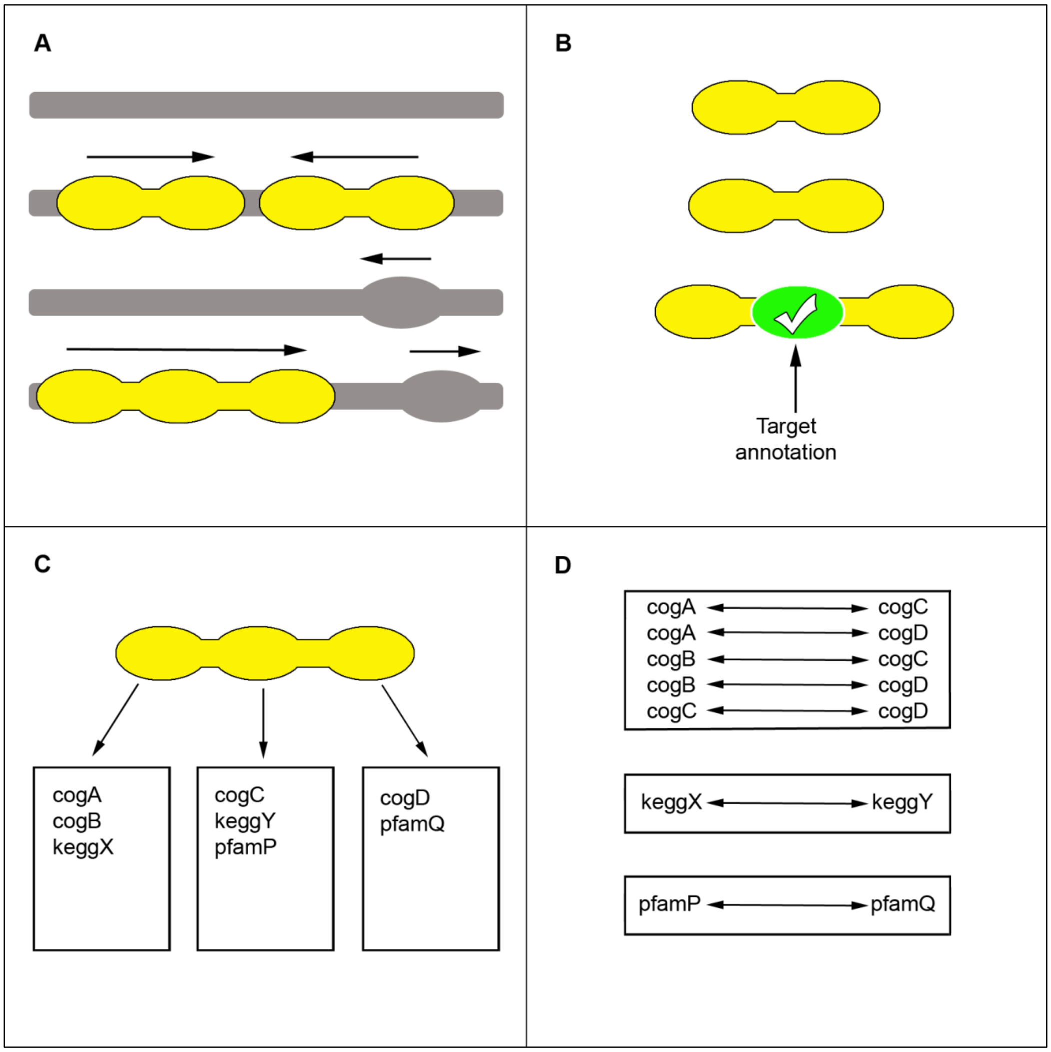


Figure 4-2 Network construction workflow. *Proximon member genes are predicted on the basis of co-direction and intergenic proximity using scaffolds containing more than one gene (Panel A). Proximons and their constituent genes can be filtered according to the presence or absence of a target annotation such that at least one member of an proximon is required to possess a target descriptor (Panel B). Note that the filter step is optional and can applied to obtain target perspective networks while being omitted in the construction of source perspective networks. Each gene in a given proximon is mined for its various types of functional annotations where any particular type has a domain of existing values (Panel C). For each proximon, the obtained functional annotations are used to infer bidirectional functional interactions for annotations having the same type but different values (Panel D). Note that interactions are inferred directly for immediately adjacent gene pairs and also transitively for downstream members within the same proximon.*

positives), as evaluated against known operons of *Escherichia coli* K-12 found in RegulonDB (Gama-Castro et al., 2011). Next, functional interactions were defined in a pairwise manner for all member genes within a given proximon. For example, a proximon with the consecutive gene members *a*, *b*, and *c* would yield predicted functional interactions for the adjacent pairs *ab* and *bc*, plus an additional transitive functional interaction, namely *ac*. In the case of target perspective networks (see Results) proximons were filtered according to the presence or absence of a target annotation by requiring a minimum number of member genes to contain a specific keyword descriptor (see Figure 4-2, Panel B). The effects of target stringency (i.e. the size of the minimum number) were also tested (see Results).

4.2.3 Network Translation

Each gene in a given proximon was mined for the following types of functional annotations: MetaCyc pathways (Caspi et al., 2012), COGs (Tatusov et al., 2003), KEGG pathways (Kanehisa et al., 2012), and TIGRFAMs (Selengut et al., 2007). A gene may have multiple annotation types and also have multiple values for a given type (see Figure 4-2, Panel C). For each proximon, the obtained functional annotations were used to infer functional interactions for annotations having the same type but different values. Translated interactions were inferred directly for immediately adjacent gene pairs but also transitively for downstream members within the same proximon (see Figure 4-2, Panel D). Note that the use of transitive translations is necessarily a reflection of the transitivity implemented in the network prediction phase.

The interactions were sorted by annotation type in order to derive a collection of discrete annotation networks for any given data source where each network had a particular annotative basis, such as MetaCyc or COG. This was possible because the translation of interacting genes into interacting annotations generated a unique set of nodes and edges with respect to each of the annotative bases. Moreover, specific annotation values (e.g. COG1363) were considered to be synonymous with their textual descriptors (e.g. cellulase M and related proteins) thereby providing a means for the conversion of nodes into a more verbose form. It is noted that it would have been possible to use the descriptors that were already available in the source data, rather than using the categorized annotations. The raw descriptors were not used in order to contrast the differences between specific annotative bases and also to avoid inflation caused by the redundant duplication of synonymous descriptors that varied only in terms of minor formatting features (i.e. lexicographical redundancy). Moreover, using specific categorized annotations produced connections between otherwise disjoint subgraphs thereby yielding a more connected network. However, future works may utilize the raw descriptors if the goal is to create a single global network of annotation linkages, regardless of annotation category.

4.2.4 Annotation Frequency Analysis

Functional annotations from modules of interest were subjected to word frequency queries using NVivo 11 for Windows. Each query used the same settings; all words were included, minimum word length was set to set four characters, and words were grouped by stem. The query results were then used to produce corresponding word clouds where frequently occurring words were depicted using increasing font sizes. In this context, a word cloud also represents a semantic model that depicts the diversity and relative dominance of annotations within a given module.

4.3 Results

In order to demonstrate the utility of metagenomic annotation networks, networks employing a variety of perspectives and annotation categories were constructed and compared. The network construction protocol proposed in this work is confined to a process of network prediction followed by network translation (see Materials and Methods). Subsequent analyses of the resulting networks were performed in order to demonstrate potential uses and applications but not as part of the network construction protocol itself. Therefore, examples provided here involve the use of Cytoscape 2.8.1 (Smoot et al., 2011) for network analyses and the MINE plugin (Rhrissorrakrai & Gunsalus, 2011) for the identification of putative annotation modules. However, these tools were selected on the basis of potential familiarity for readers and it is certainly possible to use any other software, plugins, or algorithms that might be required for particular investigations.

4.3.1 Target Perspective Networks

Networks can be constructed from a target perspective by using a keyword or series of keywords joined by logical operators to filter and reduce a set of results based on keyword occurrence, or target hits. The goal is to constrain the resulting functional interactions so that they reflect a target-centric view for a domain of interest, such as interactions relating to cellulases. In the following examples single keywords of general interest, namely “polyketide” and “cellulase”, were used to select specific proximons from the complete set of

Table 4-1 Summary of network features. The general features of each metagenomic functional network are shown including the type of network, the category of annotations used to construct the network, the network perspective, the number of nodes and edges that compose the network, and the number of predicted functional modules contained within the network.

Network Type	**Annotation**	**Perspective**	**Nodes**	**Edges**	**Modules**
Cellulase	MetaCyc	Target	213	779	5
Cellulase	COG	Target	301	763	33
Human Gut	KEGG	Source	153	192	11
Human Gut	TIGRFAM	Source	543	607	57
Gut Intersection	TIGRFAM	Comparative	407	278	19
Gut Difference	TIGRFAM	Comparative	356	329	20

available proximons thereby reducing the overall network into specific target perspective networks. However, it is possible to construct a target perspective network from a smaller and more specific range of datasets, such as using only human gut microbiomes (see Source Perspective Networks).

Prior to evaluating any target perspectives networks, the effects of target stringency were investigated. Specifically, proximons can be qualified as target hits if a fixed number or scalable proportion of their member genes has an annotation that contains the target. To determine the effects of target stringency versus network coverage, four polyketide target perspective networks were constructed and the stringency for qualification was progressively increased. Proximons in the first network were required to have at least one target hit,

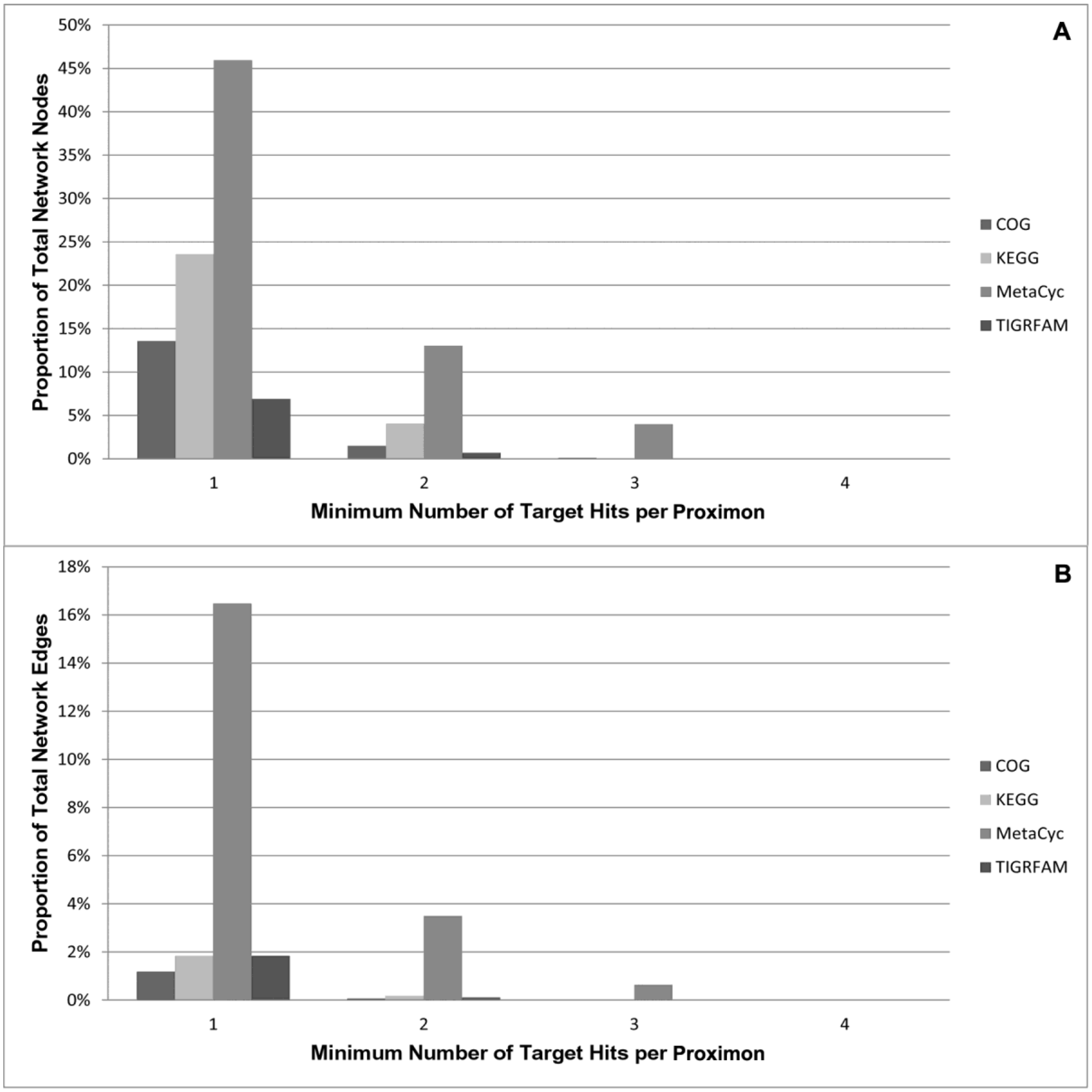


Figure 4-3 Target stringency versus network coverage. *Four polyketide target perspective networks were constructed with progressively increasing target stringency and each network was translated into each of the four annotation categories. The proportion of nodes and edges in each polyketide network was compared to its corresponding overall network. Panel A shows that coverage for nodes decreased for all annotation categories with increasing target stringency and Panel B shows that coverage for edges also decreased for all annotation categories with increasing target stringency.*

proximons in the second network were required to have at least two target hits, and so on, up to and including a stringency of requiring at least four target hits. Furthermore, each of the four networks was translated into each the four different annotation categories (see Materials and Methods) resulting in four sets of four target perspective networks. Figure 4-3 shows the proportion of nodes and edges recovered from the equivalent overall network (i.e. no filtering) with respect to increasing target stringency across each annotation category. The results illustrated that coverage for both nodes and edges decreased for all annotation categories with increasing target stringency. Therefore, the target perspective networks that follow used the least stringent requirement (i.e. at least one target hit) in an attempt to maximize the diversity and number of putative functional interactions available for subsequent analyses.

A MetaCyc cellulase network was constructed that consisted of 213 nodes and 779 edges (see Table 4-1). A highly connected central hub was observed that had the annotation PWY-1001: cellulose biosynthesis (see Figure 4-4, Panel A). Five modules were identified within the network (see Appendix C). The top ranked module (see Figure 4-4, Panel B) contained annotations relating to amino acid degradation and biosynthesis (see Table 4-2). The precise annotation terms were analyzed for more general themes and a highly cohesive module emerged that described aliphatic amino acid metabolism, with particular emphasis on branched-chain amino acids (BCAAs) (see Figure 4-5). BCAA metabolism is consistent with functional expectations for ruminal bacteria such as members of the genus *Peptostreptococcus* (Chen & Russell, 1989). Likewise, data from ruminal environments would be expected to contribute interactions to a metagenomic cellulase network.

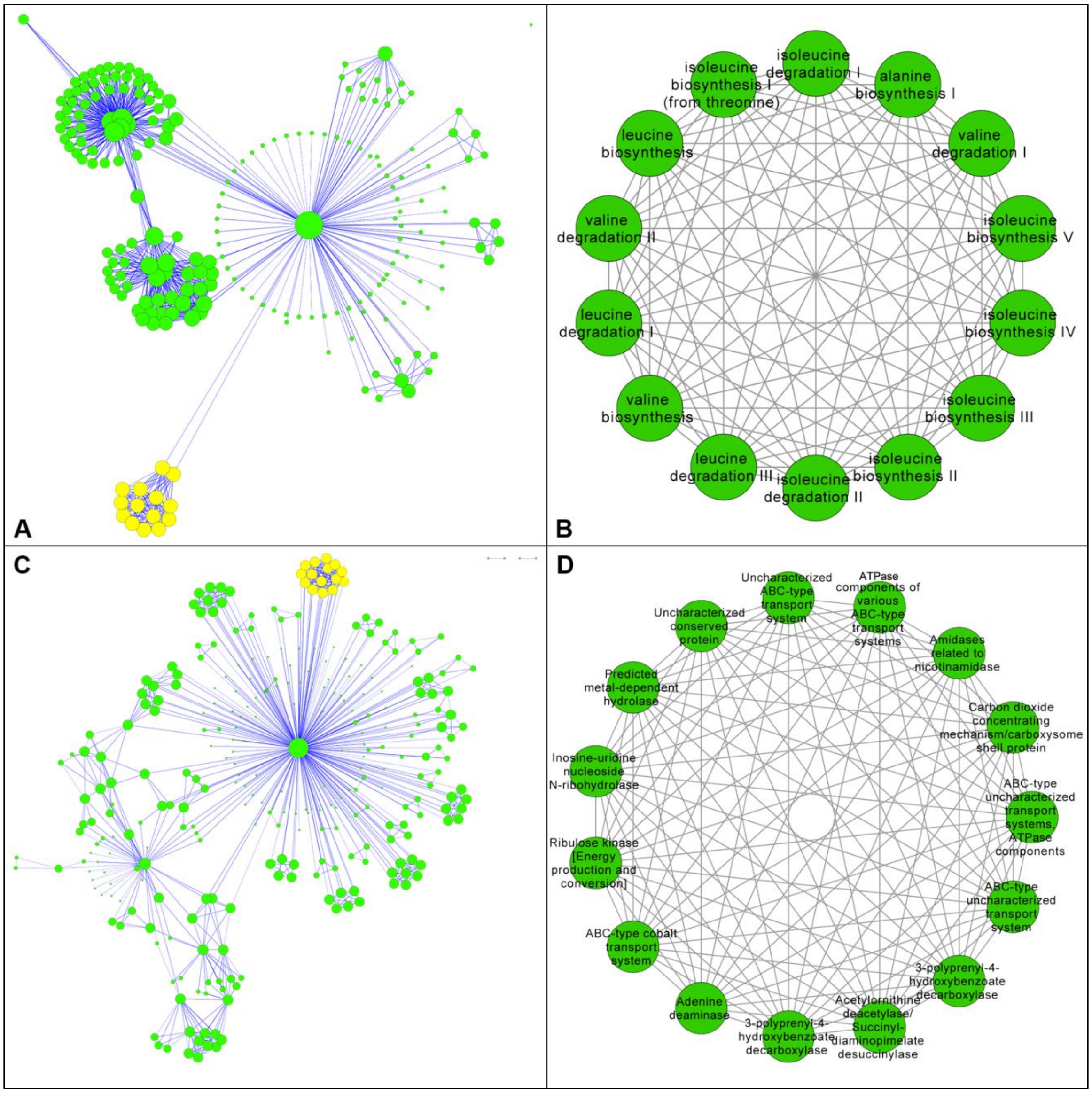

A
isoleucine biosynthesis I (from threonine)
isoleucine degradation I
alanine biosynthesis I
leucine biosynthesis
valine degradation I
valine degradation II
isoleucine biosynthesis V
leucine degradation I
isoleucine biosynthesis IV
valine biosynthesis
isoleucine biosynthesis III
leucine degradation III
isoleucine degradation II
isoleucine biosynthesis II
B
C
D
Uncharacterized ABC-type transport system
ATPase components of various ABC-type transport systems
Uncharacterized conserved protein
Amidases related to nicotinamidase
Predicted metal-dependent hydrolase
Carbon dioxide concentrating mechanism/carboxysome shell protein
Inosine-uridine nucleoside N-ribohydrolase
ABC-type uncharacterized transport systems, ATPase components
Ribulose kinase [Energy production and conversion]
ABC-type uncharacterized transport system
ABC-type cobalt transport system
3-polyprenyl-4-hydroxybenzoate decarboxylase
Adenine deaminase
Acetylornithine deacetylase/ Succinyl-diaminopimelate desuccinylase
3-polyprenyl-4-hydroxybenzoate decarboxylase

Figure 4-4 Metagenomic cellulase networks. *The target perspective networks for cellulase functional interactions are shown where large node diameter represents high node degree within each respective network. Panel A shows a network constructed using MetaCyc annotations with a highly connected central hub having the annotation PWY-1001: cellulose biosynthesis. The highlighted nodes represent the top ranking module which is enlarged in Panel B. Panel C shows a network constructed using COG annotations and features a highly connected central hub with the annotation COG1363: cellulase M and related proteins. The highlighted nodes represent the top ranking module which is enlarged in Panel D.*

A COG cellulase network was constructed that consisted of 301 nodes and 763 edges (see Table 4-1). A highly connected central hub was observed that had the annotation COG1363: cellulase M and related proteins (see Figure 4-4, Panel C). A total of 33 modules were identified within the network (see Appendix C). The top ranked module (see Figure 4-4,

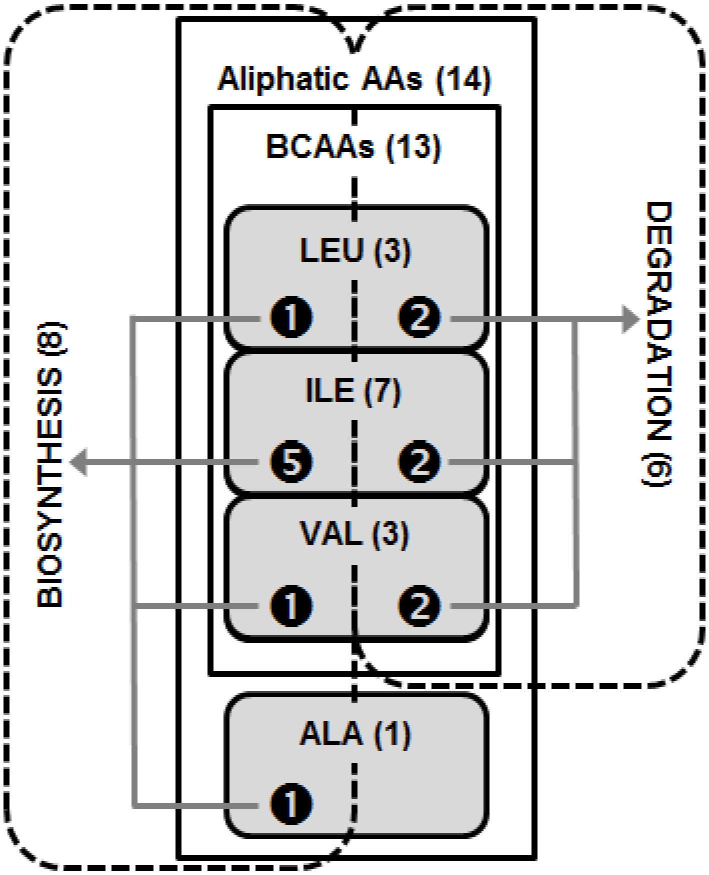


Figure 4-5 Annotation hierarchy chart. *The annotative themes for the top ranked MetaCyc module are depicted where the numeric values indicate the number of annotations belonging to a thematic category. Specifically, amino acid categories are represented vertically and metabolic categories are represented horizontally. Note, the vertical themes are encapsulatory while the horizontal themes are mutually exclusive. A variety of functional perspectives can be simultaneously visualized by way of the interacting and overlapping thematic sets.*

Panel D) contained annotations relating to ABC-type transport, permease, ATPase, as well as various other terms (see Table 4-2). Furthermore, the term “uncharacterized” was observed in conjunction with several instances of the previously listed annotations. The precise annotation terms were analyzed for more general themes resulting in a less cohesive module than the top ranked MetaCyc module. Nevertheless, these annotations are generally consistent with secretion and transfer activities such as multienzyme secretion in the cellulolytic bacterium *Clostridium thermocellum* (Nataf et al., 2009) and glycoside hydrolase secretion in *Thermobifida fusca*, a soil bacterium involved in the degradation of plant cell walls (Lykidis et al., 2007).

4.3.2 Source Perspective Networks

In contrast to the target perspective, a source perspective network involves generating all possible functional interactions but from a particular range or collection of datasets. In this case, the goal is to constrain functional interactions so that they reflect a source-centric view for a domain of interest, such as human gut interactions. Moreover, it is possible to integrate target and source perspectives by constructing a target perspective network from a particular collection of source related datasets. This approach can be used to find functional interactions that are simultaneously target-centric and source-centric, such as cellulase interactions occurring in the human gut. In the present work a human gut microbiome (Gill et al., 2006) was used to produce two source perspective networks.

A KEGG gut network was constructed that consisted of 153 nodes and 192 edges (see Table 4-1). Unlike the target perspective networks, no central hub was observed (see Figure

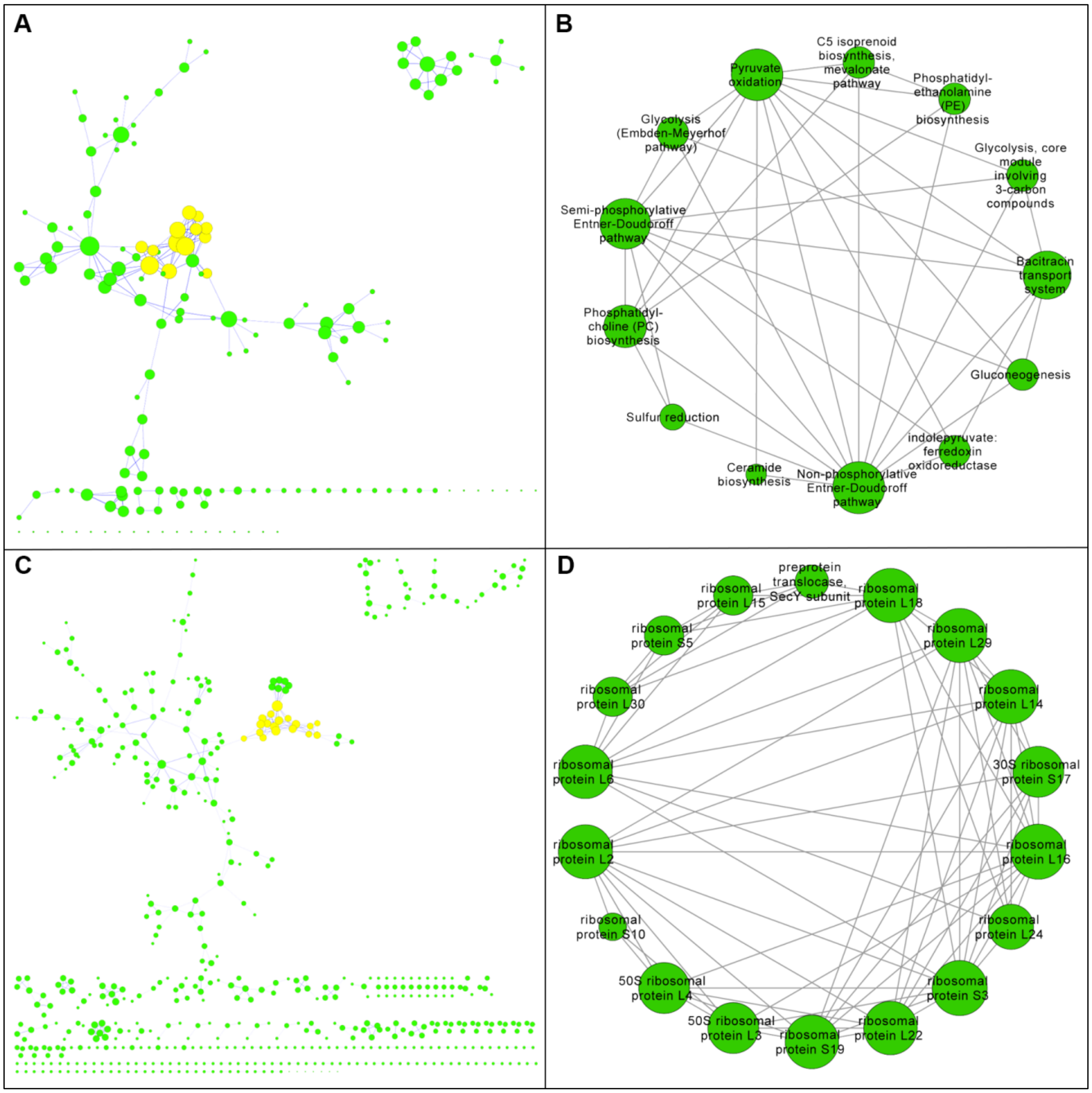
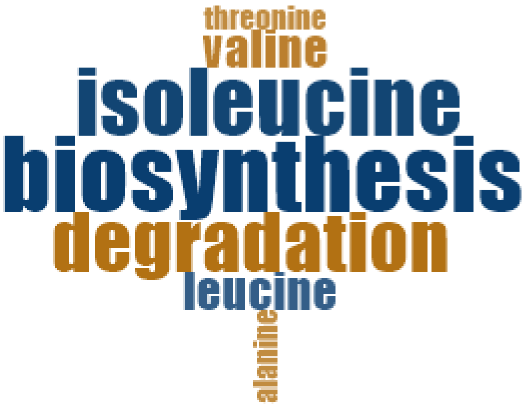
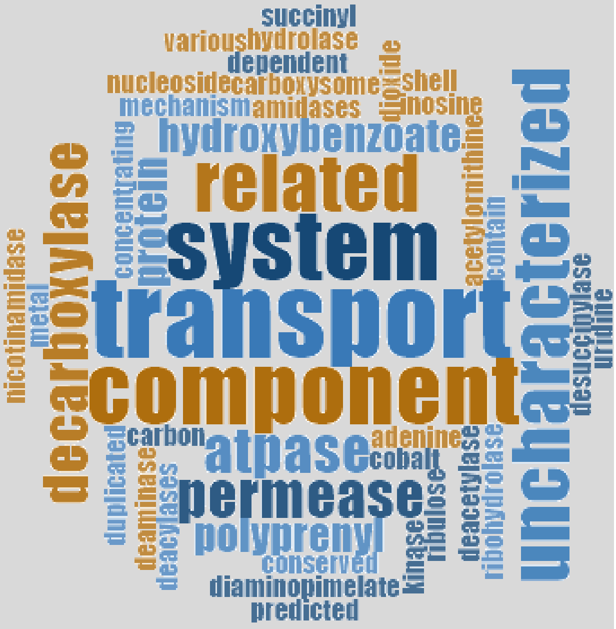

A
B
Pyruvate oxidation
C5 isoprenoid biosynthesis, mevalonate pathway
Phosphatidyl-ethanolamine (PE) biosynthesis
Glycolysis (Embden-Meyerhof pathway)
Glycolysis, core module involving 3-carbon compounds
Semi-phosphorylative Entner-Doudoroff pathway
Bacitracin transport system
Phosphatidyl-choline (PC) biosynthesis
Gluconeogenesis
Sulfur reduction
indolepyruvate: ferredoxin oxidoreductase
Ceramide biosynthesis
Non-phosphorylative Entner-Doudoroff pathway
C
D
preprotein translocase, SecY subunit
ribosomal protein L15
ribosomal protein L18
ribosomal protein S5
ribosomal protein L29
ribosomal protein L30
ribosomal protein L14
ribosomal protein L6
30S ribosomal protein S17
ribosomal protein L2
ribosomal protein L16
ribosomal protein S10
ribosomal protein L24
50S ribosomal protein L4
ribosomal protein S3
50S ribosomal protein L3
ribosomal protein S19
ribosomal protein L22

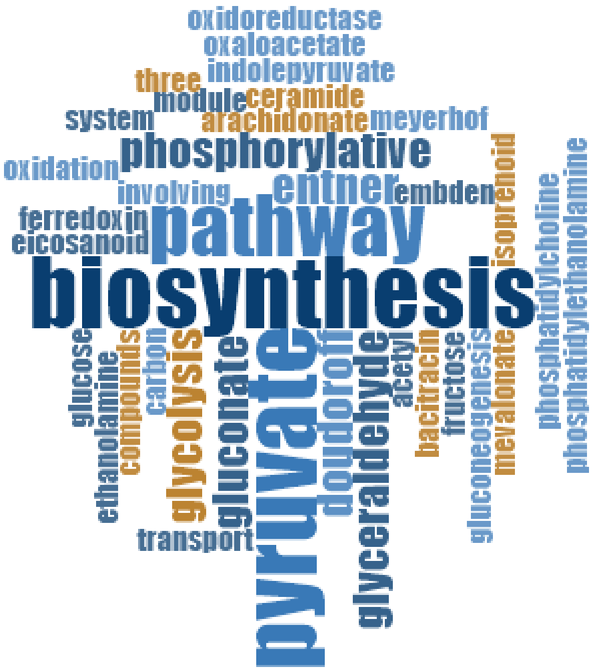
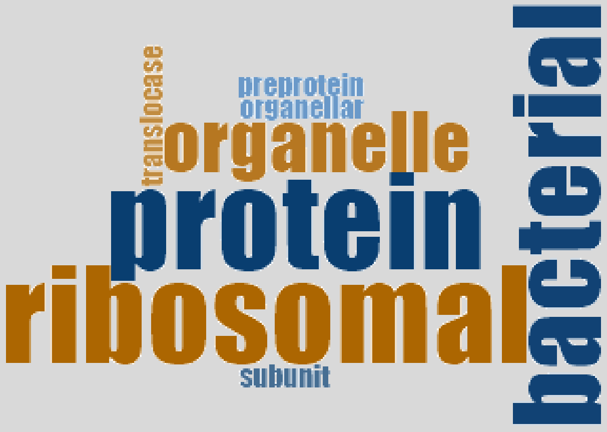
Figure 4-6 Human gut networks. *The source perspective networks for human gut functional interactions are shown where large node diameter represents high node degree within each respective network. Panel A shows a network constructed using KEGG annotations where the highlighted nodes represent the top ranking module which is enlarged in Panel B. Panel C shows a network constructed using TIGRFAM annotations where the highlighted nodes represent the top ranking module which is enlarged in Panel D.*

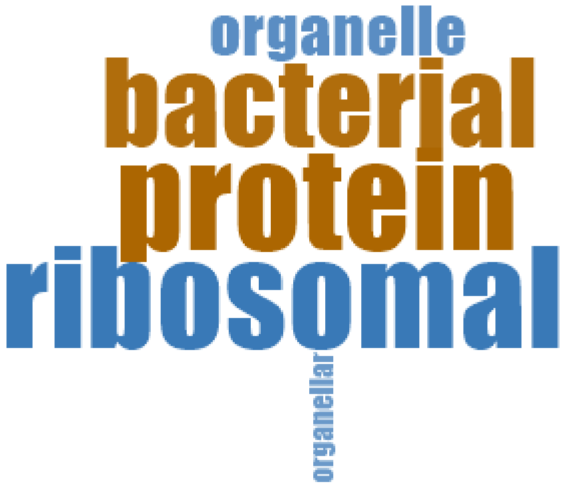
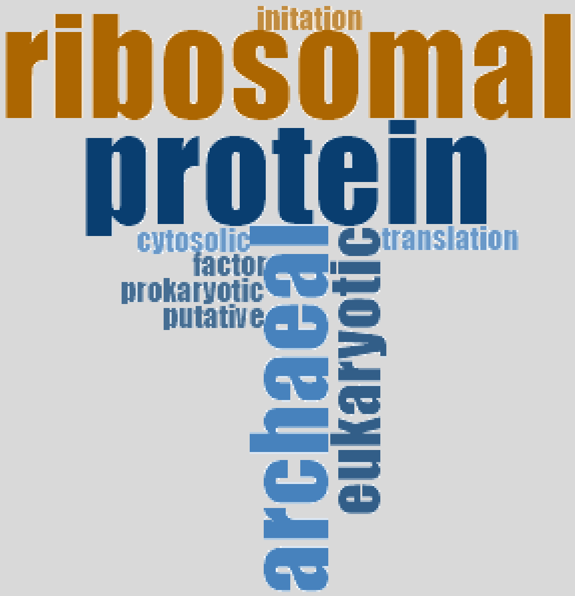
4-6, Panel A). A total of 11 modules were identified within the network (see Appendix C). The top ranked module (see Figure 4-6, Panel B) scored lower than either of the top ranked cellulase modules and contained a diverse range of annotation terms (see Table 4-2). The terms glycolysis and pyruvate occurred frequently in the pathway annotations of this module and are likely indicative of core metabolic activities across the gut community. Additional terms like isoprenoid biosynthesis and mevalonate pathway may be associated with cholesterol and possibly the statin pathway of the host liver. In fact, recent evidence suggests that the enteric microbiome can moderate response to statins (Kaddurah-Daouk et al., 2011). Moreover, the term phosphatidylcholine biosynthesis may offer another potential link to the host liver as phosphatidylcholine from non-microbial sources has been reported to be associated with significant liver protection as part of the silybin-phosphatidylcholine complex (Kidd & Head, 2005).

A TIGRFAM gut network was constructed that consisted of 543 nodes and 607 edges (see Table 4-1). Like the KEGG network, no central hub was observed (see Figure 4-6, Panel C). A total of 57 modules were identified within the network (see Appendix C). The top ranked module (see Figure 4-6, Panel D) scored slightly higher than the top ranked KEGG module but lower than either of the top ranked cellulase modules (see Table 4-2). The annotation term ribosomal protein dominated this module and often occurred in conjunction with the term bacterial/organelle. The result was a highly cohesive module that involved bacterial ribosomal proteins. Like the glycolysis features of the aforementioned KEGG module, this is potentially indicative of core metabolic activities across the gut community.

Table 4-2 Top ranked functional modules. The general features of the highest scoring functional module from each network are shown including the source network, the score assigned by MINE, the number of nodes and edges that compose the module, and the member annotations derived from the nodes. Member annotations are displayed using word clouds where frequently occurring words are depicted using increasing font sizes.

Source	Score	Nodes	Edges	Word Cloud
MetaCyc Cellulase Network	14.0	14	91	 threonine valine isoleucine biosynthesis degradation leucine alanine
COG Cellulase Network	15.0	15	105	 succinyl varioushydrolase dependent nucleosidecarboxysome dioxide shell mechanism amidases inosine hydroxybenzoate related concentrating protein system acetylornithine contain uncharacterized nicotinamidase metal decarboxylase transport desuccinylase uridine component carbon adenine atpase cobalt duplicated deaminase deacylases permease ribulose deacetylase ribohydrolase polyprenyl kinase conserved diaminopimelate predicted

KEGG Human Gut Network	6.2	13	37	
TIGRFAM Human Gut Network	8.2	18	70	

Human Gut Intersection Network	8.2	13	49	 organelle bacterial protein ribosomal organellar
Human Gut Difference Network	7.7	8	27	 initation ribosomal protein cytosolic translation factor prokaryotic putative archaeal eukaryotic

4.3.3 Comparative Networks

Provided that two or more networks share the same perspective and a common annotative basis, it is possible to perform set theoretic operations that result in newly generated comparative networks. In the present work a second source perspective TIGRFAM network was generated using another human gut microbiome from the same study (Gill et al., 2006) that was used to produce the other source perspective networks. The TIGRFAM networks were compared to produce two new networks, an intersection network and a difference network.

A gut intersection network was constructed that consisted of 407 nodes and 278 edges (see Table 4-1). This network contained a much lower ratio of edges to nodes than the non-comparative networks (see Figure 4-7, Panel A). A total of 19 modules were identified within the network (see Supplementary Materials). The top ranked module (see Figure 4-7, Panel B) scored the same as the top ranked TIGRFAM module (i.e. module derived using only one gut microbiome) and was composed of the same annotation terms (see Table 4-2). In fact, the top ranked intersection module was a subset with all 13 nodes occurring in the superset of the 18 nodes that comprised the top ranked TIGRFAM module. Compared to the TIGRFAM module, the result was a reduced but highly cohesive module that similarly involved bacterial ribosomal proteins.

A gut difference network was constructed that consisted of 356 nodes and 329 edges (see Table 4-1). This network contained a higher ratio of edges to nodes than the intersection network but was still slightly lower than the ratio in the TIGRFAM network (see Figure 4-7,

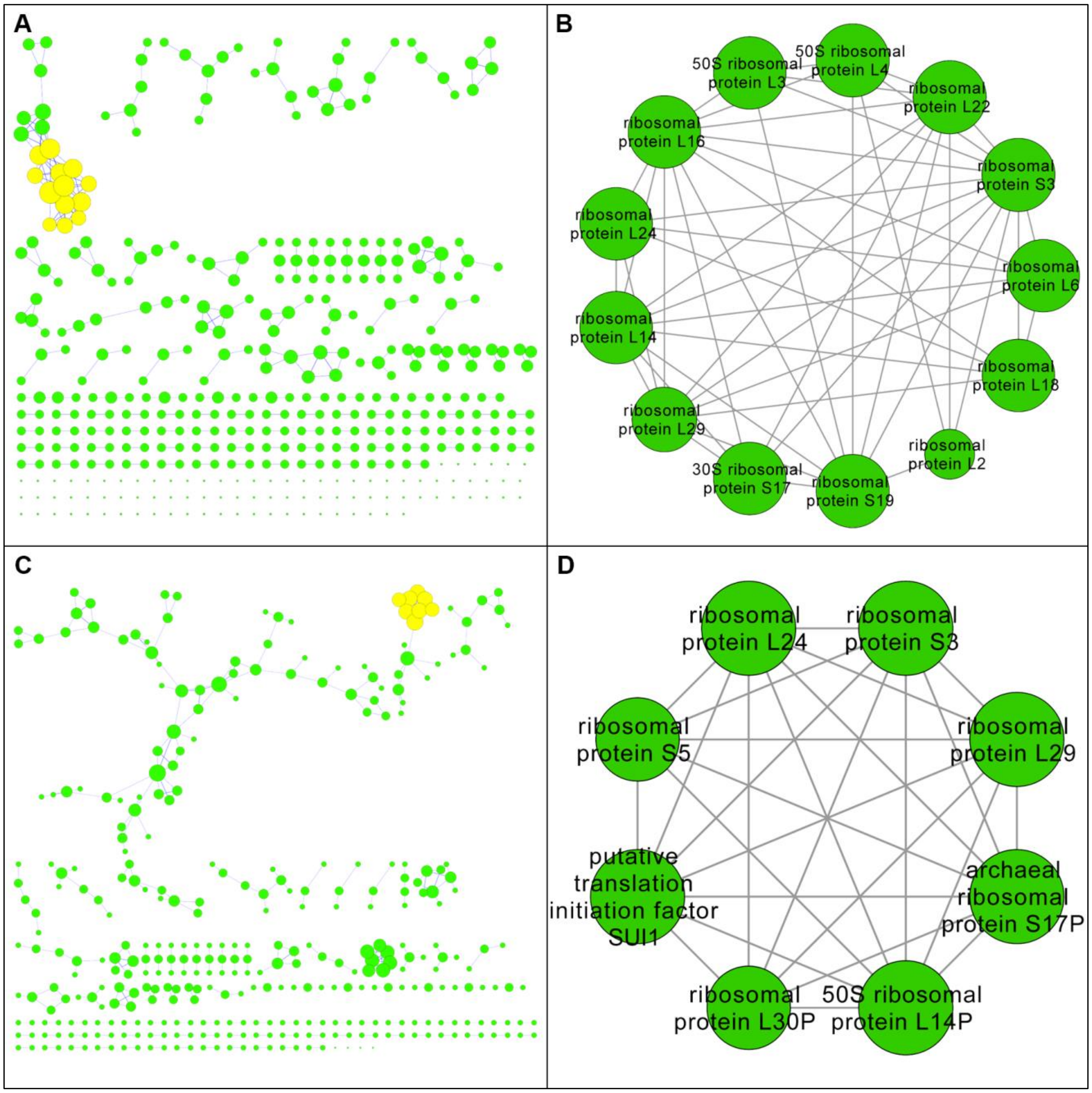


Figure 4-7 Comparative gut networks. *The comparative networks for human gut functional interactions are shown where large node diameter represents high node degree within each respective network. Specifically, two networks were constructed using TIGRFAM annotations and compared for mutual versus exclusive nodes. Panel A shows the intersection of the networks where the highlighted nodes represent the top ranking module which is enlarged in Panel B. Panel C shows the difference of the networks where the highlighted nodes represent the top ranking module which is enlarged in Panel D.*

Panel C). A total of 20 modules were identified within the network (see Supplementary Materials). The top ranked module (see Figure 4-7, Panel D) scored roughly the same as the top ranked TIGRFAM module and was composed of similar annotation terms (see Table 4-2). While this module was also dominated by the theme of ribosomal proteins it was however more diverse and the ribosomal proteins terms frequently occurred in conjunction with the terms eukaryotic and/or archaeal, rather than bacterial.

4.4 Discussion

The modules derived in this work are of particular interest because they represent functional metamodules. This is because it is not possible to resolve whether the activities of a single module are accomplished by a single microbial species or if they represent composite functionality produced by the greater microbial community. Therefore, metamodules provide a systems perspective at the community level. In addition, these modules provide a direct characterization of functional capability and organization as opposed to an inferred characterization on the basis of taxonomic composition. This marks an important departure from previous taxonomy driven approaches because they are susceptible to effects of community functional plasticity (Dinsdale et al., 2008; Manichanh et al., 2010) that can cloud the taxonomy versus function relationship. However, this does not exclude the incorporation of concurrent taxonomic information that could bolster the interpretation of certain datasets. As a result, metamodules can provide crucial functional insight for a variety of applied pursuits.

Modules from target perspective networks have the potential to reveal novel metabolic relationships that can subsequently assist in the hunt for new biocatalyst candidates. This process can be regarded as a metagenomic analog to the guilt by association principle (Aravind, 2000) that has been previously used to infer contextual information at the genomic level. In the case of the presented cellulase networks, modules that contain annotations with keywords like unknown or uncharacterized can be used to highlight genes of particular interest since annotation values (e.g. COG1699) can be easily traced back to their source genes in the raw data. This provides an expedient method to recover a shortlist of promising genes from among a raw dataset that may contain tens of millions of otherwise indistinguishable records. Mining candidate genes that can be subjected to more rigorous analyses can be applied to a broad collection of interests ranging from novel glycoside hydrolase detection for biomass degradation (Li et al., 2011) to prebiotic molecule discovery for human health applications (Candela et al., 2010).

Modules from source perspective networks have the potential to reveal how particular microbial environments orchestrate functional interactions to achieve specific functional capacities and hierarchical organization. Although gene-centric analyses have been previously applied to metagenomic functional evaluation (Tringe et al., 2005), they lack the ability to provide a systems perspective of functional organization. This is because gene content analyses cannot reveal the functional interactions that are essential in understanding how various microbial communities cooperatively achieve their specific functional capabilities. In the case of the presented human gut networks, it becomes possible to speculate not only on how the gut microbiome interacts among its constituents but also on how it exerts a collective effect on host metabolic activities. Currently this is a topic of tremendous interest and many research ventures could be served by analysis and interpretation of metamodules recovered from source perspective networks. For example, metamodules from various human microbiomes could be compared to functional modules from disease related functional linkage networks (Rende et al., 2011) in order to provide complementary analyses.

The motivation for comparative networks follows logically from the utility of source perspective networks since modules from comparative networks can expose commonalities and differences in functional configurations between different data sources. Such comparisons can be used to contrast vastly different microbial environments or to find mutual cores within closely related habitats, such as the human gut of various individuals. In addition, the approach taken in the current work differs from past studies involving comparative metagenomics because it is not affected by the previously discussed limitations of taxonomy based methods and it provides information beyond the previously mentioned gene-centric analyses. In the case of the presented comparative gut networks, it is possible to see that essential core modules could be developed for a variety of human microbiomes by deriving respective intersection networks from sets of multiple participants. The ability to directly contrast and compare metagenomic functional repertoires can offer tremendous utility to existing comparative research areas such as obese versus lean gut microbiomes (Turnbaugh et al., 2009) and control versus autistic gut microbiomes (Finegold et al., 2010).

The implementation presented here was based on several simplifications and assumptions that could be addressed by future works. The use of transitive functional interactions favoured the formation of complete subgraphs (i.e. a component where each node has an edge to every other node). Although this was done to maximize functional information, it could also have contributed to an inflation of network edges that can bias module finding algorithms. Other implementations should consider the prospect of constrained transitivity as a comparison. Similarly, the confidence thresholds for defining proximons should be further tested in a metagenomic context and this could be performed in conjunction with limits for transitivity in order to characterize the interaction of these two essential factors. Further still, the operon reference data obtained from RegulonDB (Gama-Castro et al., 2011) represents knowledge derived from a classic model organism. Given that metagenomes offer access to the uncultured microbial majority, the applicability of such reference data remains to be established, although some evidence of extensibility was provided in the previous chapter. In general, an improved understanding of the properties of metagenomic proximons and/or metagenomic operons (e.g. size, composition, frequency, etc.) would benefit metagenomic annotation networks and related interests.

Metagenomic annotation networks offer a novel taxonomy-free approach for understanding the functional capacity and hierarchical organization of integrated microbial communities. In particular, these networks can be analyzed for functional metamodules that subsequently provide a systems perspective at the microbial community level. Modules from target perspective networks can be used to infer interactions for a given gene or protein of interest. In turn, these interactions can be instrumental in revealing novel metabolic relationships that can subsequently assist in the hunt for new biocatalyst candidates. Modules from source perspective networks reveal how particular microbial environments orchestrate functional interactions to achieve specific functional capacities and hierarchical organization. This offers a mechanism of functional characterization that goes beyond gene-centric analyses. Modules from comparative networks can expose commonalities and differences between functional configurations from different data sources. These comparisons can be used to contrast vastly different microbial environments or to find mutual cores within closely related habitats, such as the human gut of various individuals. Comparing the functional repertoire of human microbiomes will be especially informative for future works of medical interest. In conclusion, the metagenomic annotation networks developed in this chapter demonstrate the application and utilization of metagemomic proximons for the purpose functional investigation. Moreover, numerous other designs and protocols are certainly possible based on proximons as an informative source of potential metagenomic functional interactions.

Chapter 5
Conclusion

The various projects carried out during the course of this thesis have been directed toward three particular areas representing current challenges involved with the use of metagenomic functional interactions: the computational representation of metagenomic proximons (i.e. metagenomic functional interactions) and their corresponding dissemination in the big data era; the evaluation of the relationship between proximons and operons; the utilization of metagenomic proximons for applications in functional metagenomics. In this final chapter, I enumerate potential gains to metagenomic research resulting from these implementations and investigations. I also list the limitations and experimental assumptions that were involved, as well as proposing future directions for research in each of the examined areas.

5.1 The Proximon Proposition

The results of mapping between proximons and operons have shown that while proximons frequently represent actual operons, many operons are not captured as proximons. In turn, this demonstrates the viability of a conceptual demarcation that stems from predictions relying exclusively on co-direction and proximity. In turn, this distinction is important because gene orientation and position are typically the only data ubiquitous to all metagenomic datasets. Therefore, while the use of these properties is inevitable for the prediction of metagenomic operon candidates, by no means does every set of co-directional and proximal genes represent an actual operon, particularly in the case of binary configurations. Moreover, the relationship between a given proximon and its corresponding transcription unit cluster remains unclear. In other words, more work needs to be performed to determine how often proximons represent specific member transcription units of a greater cluster versus the complete cluster itself.

The set theoretic nature of the proximon renders it as both a tangible abstraction and an empirically defined entity. However, while the property of gene direction can be represented as a discrete variable, intergenic distance is represented as a continuous variable and therefore the condition of proximity requires an operational definition. In other words, while co-direction is absolute, proximity can be defined to varying degrees (i.e. on a continuum) and the set theoretic nature of the proximon is purely a product of establishing a threshold for intergenic distance that represents an operational definition for proximity. In this thesis, the threshold for intergenic distance was based on existing knowledge about known operons from *Escherichia coli* K-12 found in RegulonDB (Gama-Castro et al., 2011). Future work should consider the applicability of this model and carry out comparative analyses on the threshold for proximity versus the properties and reliability of the corresponding proximons.

It is important to reiterate that the primary purpose of proximon prediction is to provide a source of metagenomic functional interactions. Therefore, the validity and usefulness of the work carried out in this thesis is not contingent upon the acceptance or adoption of the proximon proposition. Irrespective of nomenclature, the sets of genes identified here still represent strong candidates for mutual functional linkages based on their directional and positional properties. Thus, the contents of MetaProx offer metagenomic functional interactions that can be used to drive a variety of interests and pursuits in functional metagenomics and the investigations performed here offer valuable information on usage and limitations of the guilt by association paradigm with respect to these data.

5.2 Computational Representation of Biological Data

MetaProx provides two primary contributions. First it serves as a publicly available repository of metagenomic functional interactions that can be used to accelerate research in various areas of functional metagenomics. Second, it explores representations for semi-structured biological data that can offer an alternative to the traditional relational database approach. In particular, a serialized object implementation is used that advocates a *Data as Data* policy where the same serialized objects can be used at all levels (database, search tool, saved user file) without conversion or the use of human-readable markups.

The optimal exploitation of data representation and transmission has traditionally eluded scientists in the past, largely due to the absence of necessity. Previously, small-scale ad hoc data formats were sufficient for occasional distribution to a small number of interested individuals. Alternatively, when broader standardization has been implemented, such as FASTA format[24] data, it has required a centralized entity and/or data repository to drive the adoption of a standard that is still an inflexible and inflated representation, albeit uniform. Bioinformaticians commonly spend a significant amount of time materializing binary data into an inflated representation that is transmitted to other bioinformaticians who subsequently parse and deflate this data back into a binary format for their own particular usage (see Figure 5-1). This type of approach to data representation and dissemination perpetuates a cumbersome mindset and in order to accelerate data-centric research the following

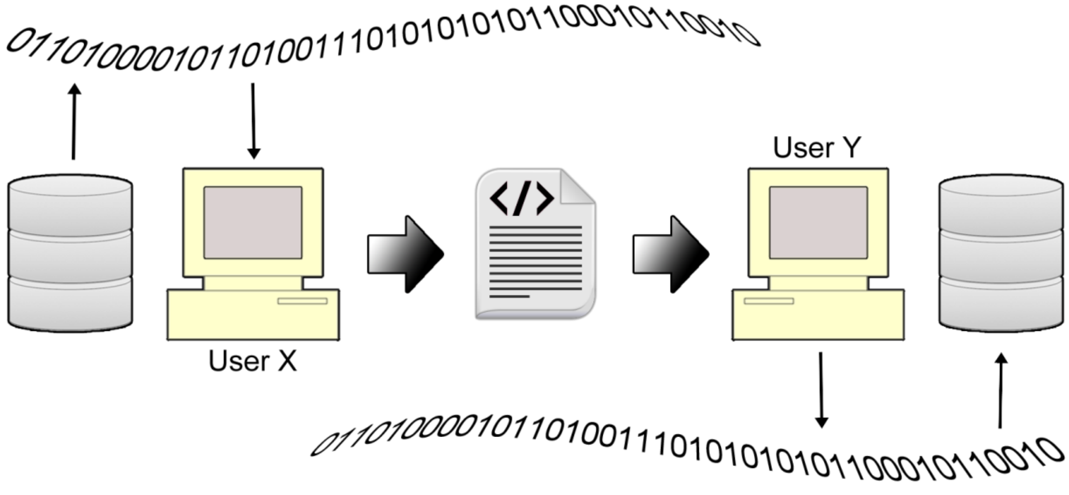


Figure 5-1 Representation versus inflation. *An abstract depiction of data exchange between users where User X has materialized data into an inflated XML representation that User Y must deflate prior to usage.*

[24] A text-based representation commonly used for either nucleotide or peptide sequences

conceptual obstacles must be addressed: readability, representation, and standardization. Although these factors will be discussed in terms of biological *omics* data, their consideration is generally extensible to any data-centric field.

While there have been notable efforts toward binary data representation, such as the BAM file format[25], many data repositories still dispense downloadable data that is human-readable. As a result, verbose textual and markup-based representations of biological data still play a large role in *omics* research. Whether this circumstance stems from methodological legacy or the inability to achieve consensus on a superior format is unclear. However, what is clear is the redundancy in converting compact bytewise machine-ready representations into larger less economical bytewise representations to support the contingency of manual human usage, especially since in most cases the size and number of files precludes this event, at least in terms of any kind of reasonable time frame. Therefore, human readability should be forever deprecated as consideration of file format specification. However, it is important to clarify that these arguments are made with respect to static data transmission, such as providing files for researchers to download and use in their computational pipelines. Using a well-accepted standard like JSON to drive a dynamic web application is a wholly reasonable solution, despite the human-readable nature of the representation.

File formats for *omics* data should strive for compactness and utility in that they represent immediately usable information without the requisite of transformation, such as parsing, that

[25] Binary Alignment/Map (BAM): A binary representation of a corresponding human-readable Sequence Alignment/Map (SAM) file

is synonymous with present representations. A strong candidate for a solution would be to adopt the use of serialized object data. Specifically, programming languages like Java and the .NET languages offer the functionality to materialize objects (i.e. data structures) from memory, thereby providing a compressed storage format that can be read directly back into memory as instantiated objects, without the need for cumbersome file parsing. In addition, objects are query-ready through the methods of their corresponding API[26], thus providing a human handle for rapid manipulation of a computationally optimized representation.

The standardization of file and exchange formats can be regarded as double-edged sword in that mandating a standard format vastly increases its recognition and adoption by users but simultaneously robs them of the flexibility to devise representations that best suit their specific purposes. Again, object serialization can mitigate this conflict by providing low-level standardization for basic constructs (e.g. a gene class) while allowing users to combine these entities in whatever fashion they require. Then a simple wrapper class can be developed to extract the standard serialized objects for use according to their standardized API. For example, a MetaProx query returns a list of proximon objects, which in turn contain gene objects. Ultimately, the most expedient use of such query results would be to download the serialized genes and perform some type of further analysis by invoking the methods of their API. This represents a much more intuitive and transparent process than the opacity of saving as text, then parsing text back into memory, then filtering qualifying cases.

[26] Application Programming Interface (API): a specification for the interaction between components or classes that allows users (i.e. other developers) to make use of a component or class without the need to understand its underlying implementation

The object deployment paradigm used for MetaProx is also amenable to application layer protocols (in contrast to transport layer protocols like TCP or UDP) such as Internet InterORB Protocol[27] which could potentially expedite the transmission of object data. This is important because while MetaProx does not use a standard relational database approach to data representation, the object management system in the current implementation does have a key limitation: Although the distributed deployment strategy allows users to harness their own computational resources, it does so with the requisite of user bandwidth. This is potentially limiting because the efficacy of searching MetaProx is constrained by both network performance (i.e. download speed) and network availability (i.e. user access to unlimited or sufficiently large bandwidth). This limitation could be at least partially mitigated by the use of a hybridized relational-object database where a conventional relational database is used to store proximon data based on the recurrent features (e.g. proximon identifier, metagenomic sample name, etc.) while storing the irregular features (e.g. variable lists of annotation objects) as serialized Java objects that would be housed as BLOBs[28]. This would permit some portion of server-side pre-processing that would lead to a reduction in the amount of data that needs to be sent across the network.

Overall, future work on the computational representation of biological data should address the following key interests. First, data representation needs to be subjected to critical scrutiny where the null hypothesis of human-readable file generation is discarded in favour

[27] Internet InterORB Protocol (IIOP): an abstract protocol that provides a mapping between object-level transactions and the TCP/IP layer

[28] Binary Large OBject (BLOB): a conglomeration of binary data stored as a single entity in a database management system

of a Biological Object Exchange specification that emulates other existing standards like CORBA. Next, as with CORBA, a compliant mapping needs to be specified so that standard biological objects can be rendered into TCP/IP layer transactions. Furthermore, experimentation needs to be carried out on how best to leverage both object and relational facets of data representation and management. Finally, it is important that such investigations carefully mind the significant work that has already been accomplished for document-oriented databases, like MongoDB, as well as other NoSQL implementations that are steadily gaining recognition and adoption.

5.3 Evaluation of Proximons

The results presented in this thesis characterize proximons as being conservative and reliable representations of actual operons. Similarly, those results also clearly demonstrate that a large proportion of operons are not represented by corresponding proximons. While these findings generally support the viability of a distinction between these two classes, the present results were produced within a fixed experimental domain and future work should strive to test the applicability and usefulness of proximons in a broader scope.

The evaluation of proximons in this thesis was confined to a single model organism because it provided a gold standard for mapping to known operons. However, the identification of alternative operon configurations beyond the *Escherichia coli* model could extend the applicability of proximon usage but represents an inherently difficult challenge. Earlier genome-scale studies (Ermolaeva et al., 2001; Chen et al., 2004) used a comparative genomic approach to identify potential operons through the detection of recurrent gene sequences (i.e. sequences of orthologous genes) and a similar method could be applied to metagenomic data in order to find recurrent patterns of interest, albeit from a context of homology rather than orthology. However, unlike the genomic scenario it would be ambiguous whether repeated patterns were bona fide single instances from multiple different genomic sources versus multiple repeated instances from a single genomic source, or some combination of these two extremes (Vey & Moreno-Hagelsieb, 2010). One solution would be to use only instances that contain flanking genes that could be used to disambiguate the issue of genomic cardinality. Either way, a comparative metagenomic study would have the potential to reveal noteworthy repeats in gene cluster configuration that deviate from the standard *Escherichia coli* model by identifying genes exhibiting greater than expected intergenic distances between functionally linked members or perhaps distributed arrangements that include a lagging member gene or even a bipartite cluster.

Future work should evaluate proximon predictions using other operon repositories, such as DOOR: Database for prOkaryotic OpeRons (Mao et al., 2009) and also consider cross-validation of operon predictions by aggregating data from multiple sources. As mentioned in previous sections, these analyses should also incorporate varying thresholds for intergenic distance and could also be contrasted against randomized and/or synthetic datasets to identify potential artifacts of the proximon prediction process itself. Also, *Bacillus subtilis* represents another significant model organism with respect to operon data and future work should investigate the use of resources such as DBTBS: a database of transcriptional regulation in *Bacillus subtilis* (Sierro et al., 2008).

5.4 Metagenomic Annotation Networks

The metagenomic annotation networks produced in this thesis offer useful demonstrations of how proximons can provide metagenomic functional interactions that can be aggregated to produce broader network constructs that can subsequently reveal functional relationships and information. However, the approach taken here could be improved by considering several aspects network generation that are already well explored topics with respect to biological interaction networks. Specifically, future work should contrast network evolution algorithms, the calculation of network edge values, and also consider the annotation schema used to describe the interactions.

Models for network evolution attempt to emulate the features of experimentally derived networks by attaining several key topological properties including scale-free topology, hierarchical modularity, and degree dissortativity (Sun & Kim, 2011; Zhu et al., 2007). This is illustrated by the preferential attachment model and the gene duplication and divergence model which can both produce a scale-free topology where a small number of nodes form hubs that have relatively high connectivity to other nodes in the network (Sun & Kim, 2011; Zhu et al., 2007; Jeong et al., 2000; Chung et al., 2003). In addition, various physical constraint models have shown hierarchical modularity and degree dissortativity, while also producing a scale-free topology (Sun & Kim, 2011). In comparison, the networks derived in this thesis were produced in a non-iterative fashion rather than progressively evolving in accordance with algorithmic constraints. Therefore, it is likely that more sophisticated metagenomic annotation networks could be inferred if some aspects of network evolution algorithms were employed.

Related to network configuration and topology is the assignment of values that describe the degree (i.e. strength of connection) and direction of the respective network edges. In the current work, edge degree was calculated by the cumulative number of observations for a given binary interaction such as $A \leftrightarrow B$. Moreover, in accordance with the guilt by association paradigm all interactions were assumed to represent bidirectional connections. As with network evolution, it is likely that more sophisticated networks will need to utilize more complicated methods, especially with respect to the determination of edge degree. Here, edge values are positive integers reflecting the sum of binary instances but each individual instance is always an all-or-nothing outcome based on exceeding a fixed threshold. Instead, edge values could be more accurately depicted using real numbers, depending on the context of the network. Specifically, the intergenic distance for a given interaction such as $A \leftrightarrow B$ could be used to provide a variable degree of confidence instead of being transformed into a binary value.

Like network edge values, the values of the network vertices (i.e. annotations) are also subject to interpretation. In this case, the vertices are dependent on the specific annotation schema that was used to determine the annotation labels. Future work should strive to understand how the use of varying annotation schemas can cause networks to fluctuate (i.e. exhibit the loss or emergence of vertices) in response to variations in their underlying annotative schemas.

The accuracy and the applicability of the guilt by association paradigm has a direct connection to metagenomic annotation network construction and interpretation. Understanding how gene functions co-occur within operons could be quantitatively inferred using known genomic operons. Figure 5-2 provides examples that illustrate existing extremes in annotative cohesion for operons found in the *Escherichia coli* K-12 MG1655 genome, with the *trp* operon exhibiting nearly perfect annotative cohesion while the *lac* operon shows no annotative cohesion. While an initial examination was undertaken in research not included in this thesis (see Appendix D), more elaborate analyses on this type of information could be used to compile frequencies for known annotative co-occurrences that could be used to adjust and augment the assignment of putative functional annotations to unknown but functionally linked genes, by way of guilt by association. In other words, it would be possible to answers questions about functionally linked genes like "*Given that gene X is a permease, how likely is it that gene Y is a transferase?"* Quantifying these types of probabilities is essential in

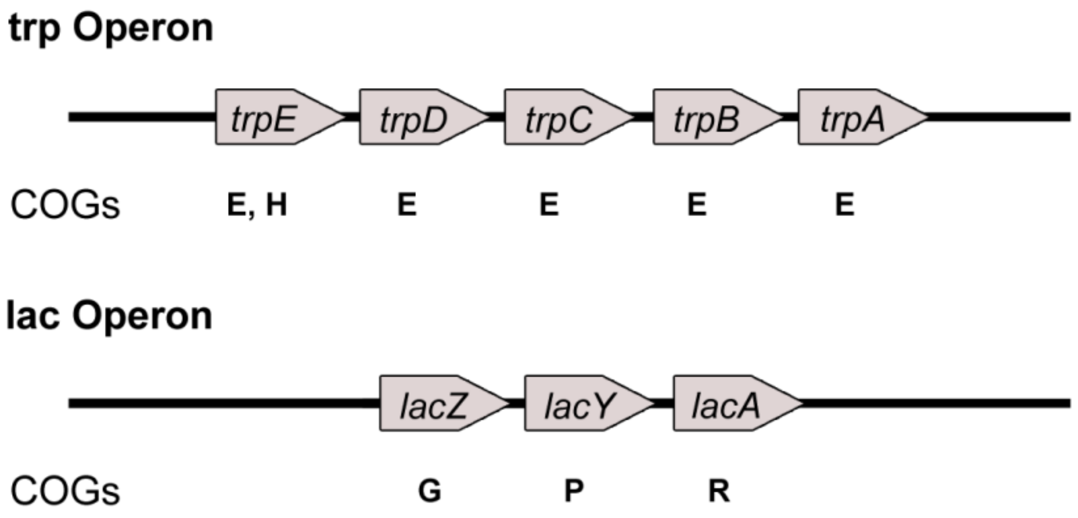


Figure 5-2 Annotative cohesion of known operons. *The trp and lac operons from the Escherichia coli K-12 MG1655 genome are shown with respect to their member genes and the corresponding COG category annotations for each member gene.*

understanding whether annotative cohesion is a property of an underlying deterministic phenomenon or if annotative co-occurrence represents a stochastic process. Moreover, such a determination might have important implications toward our understanding of the organization of functional linkages.

5.5 Final Remarks

Metagenomic research has had a profound impact on our fundamental understanding of microbial ecology and our expectations for microbial genomic plasticity. The far reaching hand of metagenomic inquiry will remain a driving methodology in the science of the 21st Century, being both augmented and shaped by the prevailing focus on big data and cloud-driven resources. Similarly, the continued study of metagenomic functional interactions has the potential to guide the discovery of novel functional relationships and expand our understanding of genomic functional organization beyond the limited scope provided by standard model organisms. Tremendous work remains to be done with respect to identifying and characterizing the currently unknown microbial majority that is responsible for facilitating and mediating many of the fundamental processes of life on our planet. As their complete portrait continues to materialize, we must be vigilant in both our maintenance and perpetuation of accepted paradigms while simultaneously listening for the earnest evidence that signals the need for conceptual reformation. Critical thought and dispassionate objective analysis serve as our best weapons against the inherently human need to describe and solve problems through the application of anthropomorphism and teleology. However, if meticulously devised and rigorously implemented, the combination of computation and metagenomics, along with whatever future protocols that they might spawn, offers a previously unseen opportunity for large-scale data analysis that will subsequently lead to the inference of unprecedented knowledge.

Bibliography

Abe T, Kanaya S, Kinouchi M, Ichiba Y, Kozuki T, Ikemura T (2003) Informatics for unveiling hidden genome signatures. Genome Res 13: 693-702.

Abe T, Sugawara H, Kinouchi M, Kanaya S, Ikemura T (2005) Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples. DNA Res 12: 281-290.

Achtman M, Wagner M (2008) Microbial diversity and the genetic nature of microbial species. Nat Rev Microbiol 6: 431-440.

Allen EE, Banfield JF (2005) Community genomics in microbial ecology and evolution. Nat Rev Microbiol 3: 489-498.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403-410.

Aravind L (2000) Guilt by association: contextual information in genome analysis. Genome Res 10: 1074-1077.

Bapteste E, O'Malley MA, Beiko RG, Ereshefsky M, Gogarten JP, et al. (2009) Prokaryotic evolution and the tree of life are two different things. Biol Direct 4: 34.

Bose T, Haque MM, Reddy C, Mande SS (2015) COGNIZER: A Framework for Functional Annotation of Metagenomic Datasets. PLoS ONE. 10: e0142102.

Bockhorst J, Craven M, Page D, Shavlik J, Glasner J (2003) A Bayesian network approach to operon prediction. Bioinformatics 19: 1227-1235.

Borodovsky M, McIninch J (1993) GeneMark: parallel gene recognition for both DNA strands. Comp Chem 17: 123-133.

Bradley RD, Hillis DM (1997) Recombinant DNA sequences generated by PCR amplification. Mol Biol Evol 14: 592-593.

Braslavsky I, Hebert B, Kartalov E, Quake SR (2003) Sequence information can be obtained from single DNA molecules. Proc Natl Acad Sci U S A 100: 3960-3964.

Bulloch W (1938) The history of bacteriology. Oxford University Press, New York, N.Y.

Callebaut W (2012) Scientific perspectivism: a philosopher of science's response to the challenge of big data biology. Stud Hist Philos Biol Biomed Sci 43: 69-80.

Calvert J (2012) Systems biology, synthetic biology and data-driven research: a commentary on Krohs, Callebaut, and O'Malley and Soyer. Stud Hist Philos Biol Biomed Sci 43: 81-84.

Candela M, Maccaferri S, Turroni S, Carnevali P, Brigidi P (2010) Functional intestinal microbiome, new frontiers in prebiotic design. Int J Food Microbiol 140: 93-101.

Caspi R, Altman T, Dreher K, Fulcher CA, Subhraveti P, et al. (2012) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res 40: D742-D753.

Chen GJ, Russell JB (1989) Sodium-dependent transport of branched-chain amino acids by a monensin-sensitive ruminal peptostreptococcus. Appl Environ Microbiol 55: 2658-2663.

Chen X, Su Z, Dam P, Palenik B, Xu Y, Jiang T (2004) Operon prediction by comparative genomics: an application to the Synechococcus sp. WH8102 genome. Nucleic Acids Res 32:2147-2157.

Chung F, Lu L, Dewey TG, Galas DJ (2003) Duplication models for biological networks. J Comput Biol 10: 677-687.

Cole JR, Chai B, Marsh TL, Farris RJ, Wang Q, et al. (2003) The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. Nucleic Acids Res 31: 442-443.

Conway T, Creecy JP, Maddox SM, Grissom JE, Conkle TL, et al. (2014) Unprecedented high-resolution view of bacterial operon architecture revealed by RNA sequencing. MBio 5: e01442-14.

Dai L, Gao X, Guo Y, Xiao J, Zhang Z (2012) Bioinformatics clouds for big data manipulation. Biology Direct 7: 43.

Dandekar T, Snel B, Huynen M, Bork P (1998) Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci 23: 324-328.

D'Costa VM, McGrann KM, Hughes DW, Wright GD (2006) Sampling the antibiotic resistome. Science 311: 374-377.

De Filippo C, Ramazzotti M, Fontana P, Cavalieri D (2012) Bioinformatic approaches for functional annotation and pathway inference in metagenomics data. Brief Bioinform 13: 696-710.

Depew DJ (2011) Adaptation as process: the future of Darwinism and the legacy of Theodosius Dobzhansky. Stud Hist Philos Biol Biomed Sci 42: 89-98.

Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, et al. (2008) Functional metagenomic profiling of nine biomes. Nature 452: 629-632.

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011) UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 27: 2194-2200.

Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA (1999) Protein interaction maps for complete genomes based on gene fusion events. Nature 402: 86-90.

Ermolaeva MD, White O, Salzberg SL (2001) Prediction of operons in microbial genomes. Nucleic Acids Res 29:1216-1221.

Ferrer M, Beloqui A, Timmis KN, Golyshin PN (2009) Metagenomics for mining new genetic resources of microbial communities. J Mol Microbiol Biotechnol 16: 109-123.

Ferrer M, Martínez-Abarca F, Golyshin PN (2005) Mining genomes and 'metagenomes' for novel catalysts. Curr Opin Biotechnol 16: 588-593.

Finegold SM, Dowd SE, Gontcharova V, Liu C, Henley KE, et al. (2010) Pyrosequencing study of fecal microflora of autistic and control children. Anaerobe 16: 444-453.

Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 39: W29-W37.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, et al. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269: 496-512.

Fraser C, Alm EJ, Polz MF, Spratt BG, Hanage WP (2009) The bacterial species challenge: making sense of genetic and ecological diversity. Science 323: 741-746.

Gama-Castro S, Salgado H, Peralta-Gil M, Santos-Zavaleta A, Muñiz-Rascado L, et al. (2011) RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). Nucleic Acids Res 39: D98-D105.

Gawad C, Koh W, Quake SR (2016) Single-cell genome sequencing: current state of the science. Nat Rev Genet 17: 175-188.

Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, et al. (2006) Metagenomic analysis of the human distal gut microbiome. Science 312: 1355-1359.

Gillis J, Pavlidis P (2011) The impact of multifunctional genes on "guilt by association" analysis. PLoS One 6: e17258.

Haft DH, Selengut JD, Richter RA, Harkins D, Basu MK, et al. (2013) TIGRFAMs and Genome Properties in 2013. Nucleic Acids Res 41: D387-D395.

Handelsman J (2004) Metagenomics: application of genomics to uncultured microorganisms. Microbiol Mol Biol Rev 68: 669-685.

Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem Biol 5: R245-R249.

Hao Y, Winans SC, Glick BR, Charles TC (2010) Identification and characterization of new LuxR/LuxI-type quorum sensing systems from metagenomic libraries. Environ Microbiol 12: 105-117.

Haque F, Li J, Wu HC, Liang XJ, Guo P (2013) Solid-state and biological nanopore for real-time sensing of single chemical and sequencing of DNA. Nano Today 8: 56-74.

Heather JM, Chain B (2016) The sequence of sequencers: The history of sequencing DNA. Genomics 107: 1-8.

Hoff KJ (2009) The effect of sequencing errors on metagenomic gene prediction. BMC Genomics, 10: 520.

Hoff KJ, Tech M, Lingner T, Daniel R, Morgenstern B, et al. (2008) Gene prediction in metagenomic fragments: a large scale machine learning approach. BMC Bioinformatics 9: 217.

Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, et al. (2008) Big data: The future of biocuration. Nature 455: 47-50.

Hsu JT, Peng CH, Hsieh WP, Lan CY, Tang CY (2011) A novel method to identify cooperative functional modules: study of module coordination in the Saccharomyces cerevisiae cell cycle. BMC Bioinformatics 12: 281.

Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, et al. (2016) A new view of the tree of life. Nat Microbiol 1: 16048.

Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, et al. (2012) InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res 40: D306-D312.

Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. Genome Res 17 :377-386.

Hyatt D, Locascio PF, Hauser LJ, Uberbacher EC (2012) Gene and translation initiation site prediction in metagenomic sequences. Bioinformatics 28: 2223-2230.

Jacob F, Monod J (1961) Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol 3: 318-356.

Janga SC, Diaz-Mejia JJ, Moreno-Hagelsieb G (2011) Network-based function prediction and interactomics: the case for metabolic enzymes. Metab Eng 13: 1-10.

Janga SC, Moreno-Hagelsieb G (2004) Conservation of adjacency as evidence of paralogous operons. Nucleic Acids Res 32: 5392-5397.

Jeffery CJ (1999) Moonlighting proteins. Trends Biochem Sci 24: 8-11.

Jeffery CJ (2009) Moonlighting proteins--an update. Mol Biosyst 5: 345-350.

Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási AL (2000) The large-scale organization of metabolic networks. Nature 407: 651-654.

Joyce EA, Chan K, Salama NR, Falkow S (2002) Redefining bacterial populations: a post-genomic reformation. Nat Rev Genet 3: 462-473.

Kaddurah-Daouk R, Baillie RA, Zhu H, Zeng ZB, Wiest MM, et al. (2011) Enteric microbiome metabolites correlate with response to simvastatin treatment. PLoS One 6: e25482.

Kagan J, Sharon I, Beja O, Kuhn JC (2008) The tryptophan pathway genes of the Sargasso Sea metagenome: new operon structures and the prevalence of non-operon organization. Genome Biol 9: R20.

Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M (2012) KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res 40: D109-D114.

Kelley DR, Liu B, Delcher AL, Pop M, Salzberg SL (2012) Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. Nucleic Acids Res 40: e9.

Kidd P, Head K (2005) A review of the bioavailability and clinical efficacy of milk thistle phytosome: a silybin-phosphatidylcholine complex (Siliphos). Altern Med Rev 10: 193-203.

Kim M, Lee K-H, Yoon S-W, Kim B-S, Chun J et al. (2013) Analytical Tools and Databases for Metagenomics in the Next-Generation Sequencing Era. Genomics & Informatics 11: 102-113.

Kodama Y, Shumway M, Leinonen R, International Nucleotide Sequence Database Collaboration (2012) The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Res 40: D54-D56.

Konietzny SG, Dietz L, McHardy AC (2011) Inferring functional modules of protein families with probabilistic topic models. BMC Bioinformatics 12: 141.

Koonin EV, Galperin MY (2003) Sequence - Evolution - Function: Computational Approaches in Comparative Genomics. Boston: Kluwer Academic.

Krohs U (2012) Convenience experimentation. Stud Hist Philos Biol Biomed Sci 43: 52-57.

Kunin V, Copeland A, Lapidus A, Mavromatis K, Hugenholtz P (2008) A bioinformatician's guide to metagenomics. Microbiol Mol Biol Rev 72: 557-578.

Kurose JF, Ross KW (2005) Computer Networking: A Top-Down Approach Featuring the Internet (3rd ed.). Massachusetts: Addison-Wesley.

Kyrpides NC, Hugenholtz P, Eisen JA, Woyke T, Göker M, et al. (2014) Genomic encyclopedia of bacteria and archaea: sequencing a myriad of type strains. PLoS Biol 12: e1001920.

Lawrence JG, Roth JR (1996) Selfish operons: horizontal transfer may drive the evolution of gene clusters. Genetics 143: 1843-1860.

Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database Collaboration (2011) The sequence read archive. Nucleic Acids Res 39: D19-D21.

Li LL, Taghavi S, McCorkle SM, Zhang YB, Blewitt MG, et al. (2011) Bioprospecting metagenomics of decaying wood: mining for new glycoside hydrolases. Biotechnol Biofuels 4: 23.

Liu B, Pop M (2011) MetaPath: identifying differentially abundant metabolic pathways in metagenomic datasets. BMC Proc 5: S9.

Liu Y, Guo J, Hu G, Zhu H (2013) Gene prediction in metagenomic fragments based on the SVM algorithm. BMC Bioinformatics 14: S12.

Lykidis A, Mavromatis K, Ivanova N, Anderson I, Land M, et al. (2007) Genome sequence and analysis of the soil cellulolytic actinomycete Thermobifida fusca YX. J Bacteriol 189: 2477-2486.

Macaulay IC, Voet T (2014) Single Cell Genomics: Advances and Future Perspectives. PLoS Genet 10: e1004126.

Manichanh C, Reeder J, Gibert P, Varela E, Llopis M, et al. (2010) Reshaping the gut microbiome with bacterial transplantation and antibiotic intake. Genome Res 20: 1411-1419.

Mao F, Dam P, Chou J, Olman V, Xu Y (2009) DOOR: a database for prokaryotic operons. Nucleic Acids Res 37: D459-D463.

Mao X, Ma Q, Liu B, Chen X, Zhang H, et al. (2015) Revisiting operons: an analysis of the landscape of transcriptional units in E. coli. BMC Bioinformatics 16: 356.

Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, et al. (1999) Detecting protein function and protein-protein interactions from genome sequences. Science 285: 751-753.

Markowitz VM, Chen IM, Chu K, Szeto E, Palaniappan K, et al. (2012) IMG/M: the integrated metagenome data management and comparative analysis system. Nucleic Acids Res 40: D123-D129.

Markowitz VM, Ivanova NN, Szeto E, Palaniappan K, Chu K, et al. (2008) IMG/M: a data management and analysis system for metagenomes. Nucleic Acids Res 36: D534-538.

Marx V (2013) Biology: the big challenges of big data. Nature 498: 255-260.

Mazumdar PMH (1995) Species and specificity: an interpretation of the history of immunology. Cambridge University Press, New York, N.Y.

McHardy AC, Rigoutsos I (2007) What’s in the mix: phylogenetic classification of metagenome sequence samples. Curr Opin Microbiol 10: 499-503.

Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R (2005) The microbial pan-genome. Curr Opin Genet Dev 15: 589-594.

Medini D, Serruto D, Parkhill J, Relman DA, Donati C, et al. (2008) Microbiology in the post-genomic era. Nat Rev Microbiol 6: 419-430.

Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, et al. (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9: 386.

Miller JH, Reznikoff WS (1978) The Operon. New York: Cold Spring Harbor Laboratory.

Mira A, Martín-Cuadrado AB, D'Auria G, Rodríguez-Valera F (2010) The bacterial pan-genome: a new paradigm in microbiology. Int Microbiol 13: 45-57.

Moreno-Hagelsieb G, Collado-Vides J (2002) A powerful non-homology method for the prediction of operons in prokaryotes. Bioinformatics 18: S329-336.

Moreno-Hagelsieb G, Janga SC (2008) Operons and the effect of genome redundancy in deciphering functional relationships using phylogenetic profiles. Proteins 70: 344-352.

Nataf Y, Yaron S, Stahl F, Lamed R, Bayer EA, et al. (2009) Cellodextrin and laminaribiose ABC transporters in Clostridium thermocellum. J Bacteriol 191: 203-209.

National Center for Biotechnology Information (2014) FTP directory of bacterial genomes. ftp://ftp.ncbi.nih.gov/genomes/Bacteria/

Noguchi H, Park J, Takagi T (2006) MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. Nucleic Acids Res 34: 5623-5630.

Noguchi H, Taniguchi T, Itoh T (2008) MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. DNA Res 15: 387-396.

Nyrén PJ (1987) Enzymatic method for continuous monitoring of DNA polymerase activity. Anal Biochem 238: 235-238.

Object Management Group (2012) CORBA 3.3. http://www.omg.org/spec/CORBA/3.3

O'Driscoll A, Daugelaite J, Sleator RD (2013) 'Big data', Hadoop and cloud computing in genomics. J Biomed Inform 46: 774-781.

Oliver S (2000) Guilt-by-association goes global. Nature 403: 601-603.

O'Malley MA, Soyer OS (2012) The roles of integration in molecular systems biology. Stud Hist Philos Biol Biomed Sci 43: 58-68.

Oracle (2011) Java Web Start Technology. http://docs.oracle.com/javase/6/docs/technotes/guides/javaws/developersguide/overview.html.

Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N (1999) The use of gene clusters to infer functional coupling. Proc Natl Acad Sci U S A. 96: 2896-2901.

Pace NR, Stahl DA, Lane DJ, Olsen GJ (1985) Analyzing natural microbial populations by rRNA sequences. ASM News 51: 4-12.

Pearson S, Benameur A (2010) Privacy, security and trust issues arising from cloud computing. Cloud Computing Technology and Science (CloudCom), 2010 IEEE Second International Conference: 693-702.

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A 96: 4285-4288.

Peregrin-Alvarez JM, Xiong X, Su C, Parkinson J (2009) The Modular Organization of Protein Interactions in Escherichia coli. PLoS Comput Biol 5: e1000523.

Pertea M, Ayanbule K, Smedinghoff M, Salzberg SL (2009) OperonDB: a comprehensive database of predicted operons in microbial genomes. Nucleic Acids Res 37: D479-D482.

Pignatelli M, Aparicio G, Blanquer I, Hernández V, Moya A, et al. (2008) Metagenomics reveals our incomplete knowledge of global diversity. Bioinformatics 24: 2124-2125.

Prakash T, Taylor TD (2012) Functional assignment of metagenomic data: challenges and applications. Brief Bioinform 13: 711-727.

Price MN, Huang KH, Alm EJ, Arkin AP (2005) A novel method for accurate operon predictions in all sequenced prokaryotes. Nucleic Acids Res 33: 880-892.

Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, et al. (2012) The Pfam protein families database. Nucleic Acids Res 40: D290-D301.

Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, et al. (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464: 59-65.

Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ (2011) Removing noise from pyrosequenced amplicons. BMC Bioinformatics 12: 38.

Ramakrishnan R, Gehrke J (2003) Database Management Systems (3rd ed.). Massachusetts: McGraw-Hill.

Reeder J, Knight R. (2010) Rapidly denoising pyrosequencing amplicon reads by exploiting rank-abundance distributions. Nat Methods 7: 668-669.

Rende D, Baysal N, Kirdar B (2011) A novel integrative network approach to understand the interplay between cardiovascular disease and other complex disorders. Mol Biosyst 7: 2205-2219.

Rhee SY, Mutwil M (2014) Towards revealing the functions of all genes in plants. Trends Plant Sci 19: 212-221.

Rho M, Tang H, Ye Y: FragGeneScan (2010) predicting genes in short and error-prone reads. Nucleic Acids Res 38: e191.

Rhrissorrakrai K, Gunsalus KC (2011) MINE: Module Identification in Networks. BMC Bioinformatics 12: 192.

Riesenfeld CS, Schloss PD, Handelsman J (2004a) Metagenomics: Genomic analysis of microbial communities. Annu Rev Genet 38: 525-552.

Riesenfeld CS, Goodman RM, Handelsman, J (2004b) Uncultured soil bacteria are a reservoir of new antibiotic resistance genes. Environ Microbiol 6: 981–989.

Salgado H, Moreno-Hagelsieb G, Smith TF, Collado-Vides J (2000) Operons in Escherichia coli: genomic analyses and predictions. Proc Natl Acad Sci U S A 97: 6652-6657.

Salgado H, Peralta-Gil M, Gama-Castro S, Santos-Zavaleta A, Muñiz-Rascado L, et al. (2013) RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. Nucleic Acids Res 41: D203-D213.

Salzberg SL, Delcher AL, Kasif S, White O (1998) Microbial gene identification using interpolated Markov models. Nucleic Acids Res 26: 544-548.

Sanger SF, Nicklen ARC (1977) DNA sequencing with chain-terminating. Proc Natl Acad Sci 74: 5463–5467.

Schadt EE, Linderman MD, Sorenson J, Lee L, Nolan GP (2010) Computational solutions to large-scale data management and analysis. Nat Rev Genet 11: 647-657.

Schatz MC, Langmead B, Salzberg SL (2010) Cloud computing and the DNA data race. Nat Biotechnol 28: 691-693.

Schloss PD, Handelsman J (2005) Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. Genome Biol 6: 229.

Schmidt TM, DeLong EF, Pace NR (1991) Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. J Bacteriol 173: 4371-4378.

Selengut JD, Haft DH, Davidsen T, Ganapathy A, Gwinn-Giglio M, et al. (2007) TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. Nucleic Acids Res 35: D260-264.

Shaw AK, Halpern AL, Beeson K, Tran B, Venter JC, et al. (2008) It's all relative: ranking the diversity of aquatic bacterial communities. Environ Microbiol 10: 2200-2210.

Sierro N, Makita Y, de Hoon M, Nakai K (2008) DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information. Nucleic Acids Res 36: D93-D96.

Sigrist CJA, Cerutti L, de Castro E, Langendijk-Genevaux PS, Bulliard V, et al. (2010) PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res 38: D161-D166.

Simon C, Daniel R (2011) Metagenomic analyses: past and future trends. Appl Environ Microbiol 77: 1153-1161.

Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T (2011) Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 27: 431-432.

Stein JL, Marsh TL, Wu KY, Shizuya H, DeLong EF (1996) Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment front a planktonic marine archaeon. J Bacteriol 178: 591–599.

Strasser BJ (2012) Data-driven sciences: from wonder cabinets to electronic databases. Stud Hist Philos Biol Biomed Sci 43: 85-87.

Sun MG, Kim PM (2011) Evolution of biological interaction networks: from models to real data. Genome Biol 12: 235.

Taboada B, Ciria R, Martinez-Guerrero CE, Merino E (2012) ProOpDB: Prokaryotic Operon DataBase. Nucleic Acids Res 40: D627-D631.

Taboada B, Verde C, Merino E (2010) High accuracy operon prediction method based on STRING database scores. Nucleic Acids Res 38: e130.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, et al. (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4: 41.

Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 278: 631-637.

Teeling H, Glöckner FO (2012) Current opportunities and challenges in microbial metagenome analysis--a bioinformatic perspective. Brief Bioinform 13: 728-742.

Tettelin H, Riley D, Cattuto C, Medini D (2008) Comparative genomics: the bacterial pan-genome. Curr Opin Microbiol 12: 472-477.

Tress ML, Cozzetto D, Tramontano A, Valencia A (2006) An analysis of the Sargasso Sea resource and the consequences for database composition. BMC Bioinformatics 7: 213.

Tringe SG, Rubin EM (2005) Metagenomics: DNA sequencing of environmental samples. Nat Rev Genet 6: 805-814.

Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, et al. (2005) Comparative metagenomics of microbial communities. Science 308: 554-557.

Turnbaugh PJ, Gordon JI (2008) An invitation to the marriage of metagenomics and metabolomics. Cell 134: 708-713.

Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, et al. (2009) A core gut microbiome in obese and lean twins. Nature 457: 480-484.

Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. Nature 444: 1027-1031.

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, et al. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428: 37-43.

Uchiyama, T., and K. Miyazaki. 10 September 2010. Product-induced gene expression (PIGEX): a product-responsive reporter assay for enzyme screening of metagenomic libraries. Appl Environ Microbiol 76: 7029-7035.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. (2001) The sequence of the human genome. Science 291: 1304-1351.

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304: 66-74.

Versant Corporation (2013) db4o: database for objects. http://www.db4o.com.

Vey G (2013) Metagenomic guilt by association: an operonic perspective. PLoS ONE 8: e71484.

Vey G, Charles TC (2014) MetaProx: the database of metagenomic proximons. Database (Oxford) 2014: bau097.

Vey G, Moreno-Hagelsieb G (2010) Beyond the bounds of orthology: functional inference from metagenomic context. Mol Biosyst 6: 1247-1254.

Vey G, Moreno-Hagelsieb G (2012) Metagenomic annotation networks: construction and applications. PLoS ONE 7: e41283.

Ward P, Dafoulas G (2006) Database Management Systems. London, England: Thompson.

Wilmes P, Simmons SL, Denef VJ, Banfield JF (2009) The dynamic genetic repertoire of microbial communities. FEMS Microbiol Rev 33: 109-132.

Winslow CE (1950) Some leaders and landmarks in the history of microbiology. Bacteriol Rev 14: 99-114.

Woese CR (1987) Bacterial evolution. Microbiol Rev 51: 221-271.

Woese CR, Fox GE (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A 1977 74: 5088-5090.

Wooley JC, Godzik A, Friedberg I (2010) A primer on metagenomics. PLoS Comput Biol 6: e1000667.

Wright ES, Yilmaz LS, Noguera DR (2012) DECIPHER, a search-based approach to chimera identification for 16S rRNA sequences. Appl Environ Microbiol 78: 717-725.

Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, et al. (2009) A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. Nature 462: 1056-1060.

Yoon BJ (2009) Hidden Markov models and their applications in biological sequence analysis. Curr Genomics 10: 402-415.

Zhu W, Lomsadze A, Borodovsky M (2010) Ab initio gene identification in metagenomic sequences. Nucleic Acids Res 38: e132.

Zhu X, Gerstein M, Snyder M (2007) Getting connected: analysis and principles of biological networks. Genes Dev 21: 1010-1024.

Appendix A

MetaProx UML Architecture

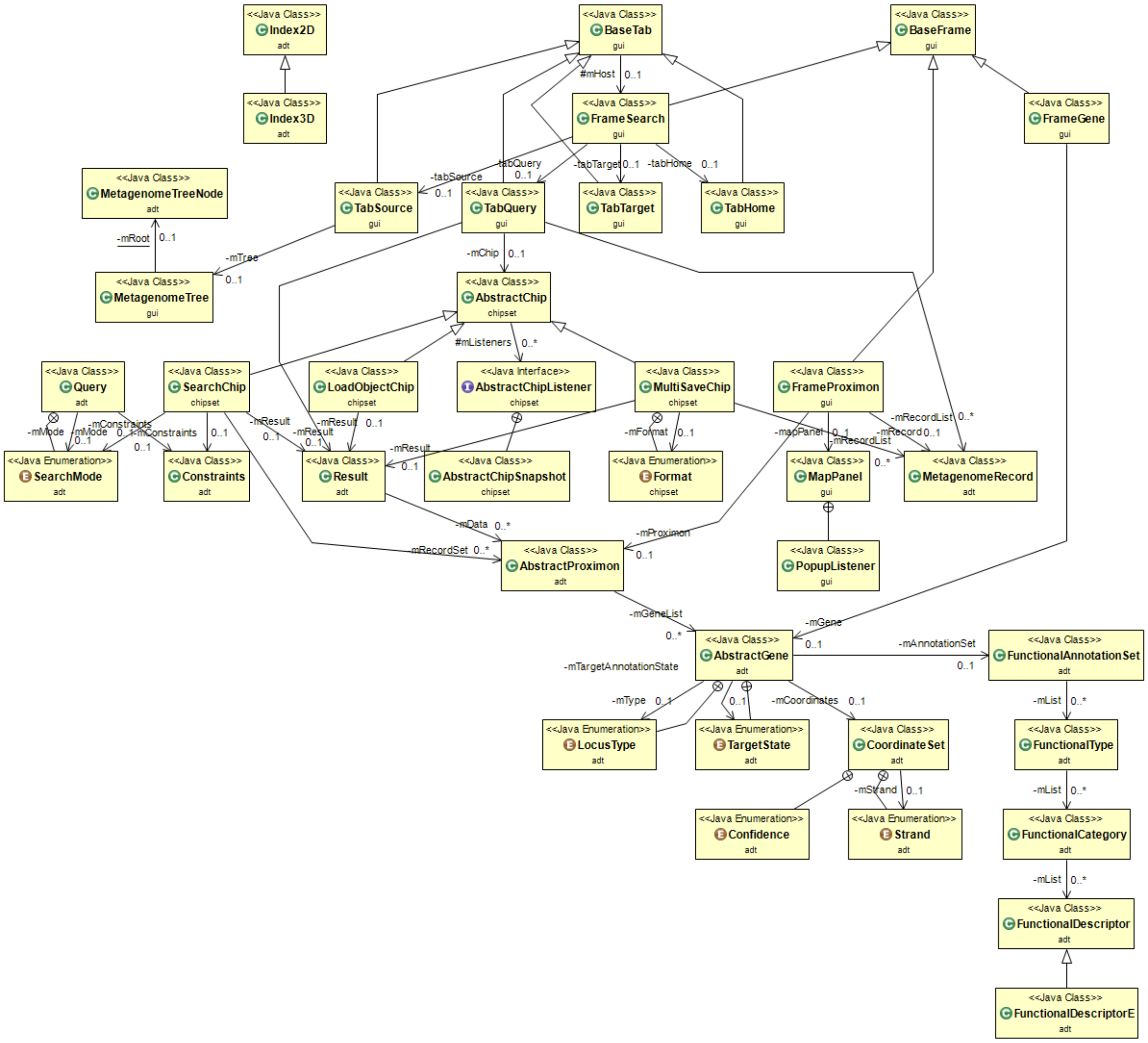


Figure A-1 MetaProx UML overview. *The dependencies and cardinalities of the complete collection of Java classes, interfaces, and enumerations are shown for the MetaProx search tool.*

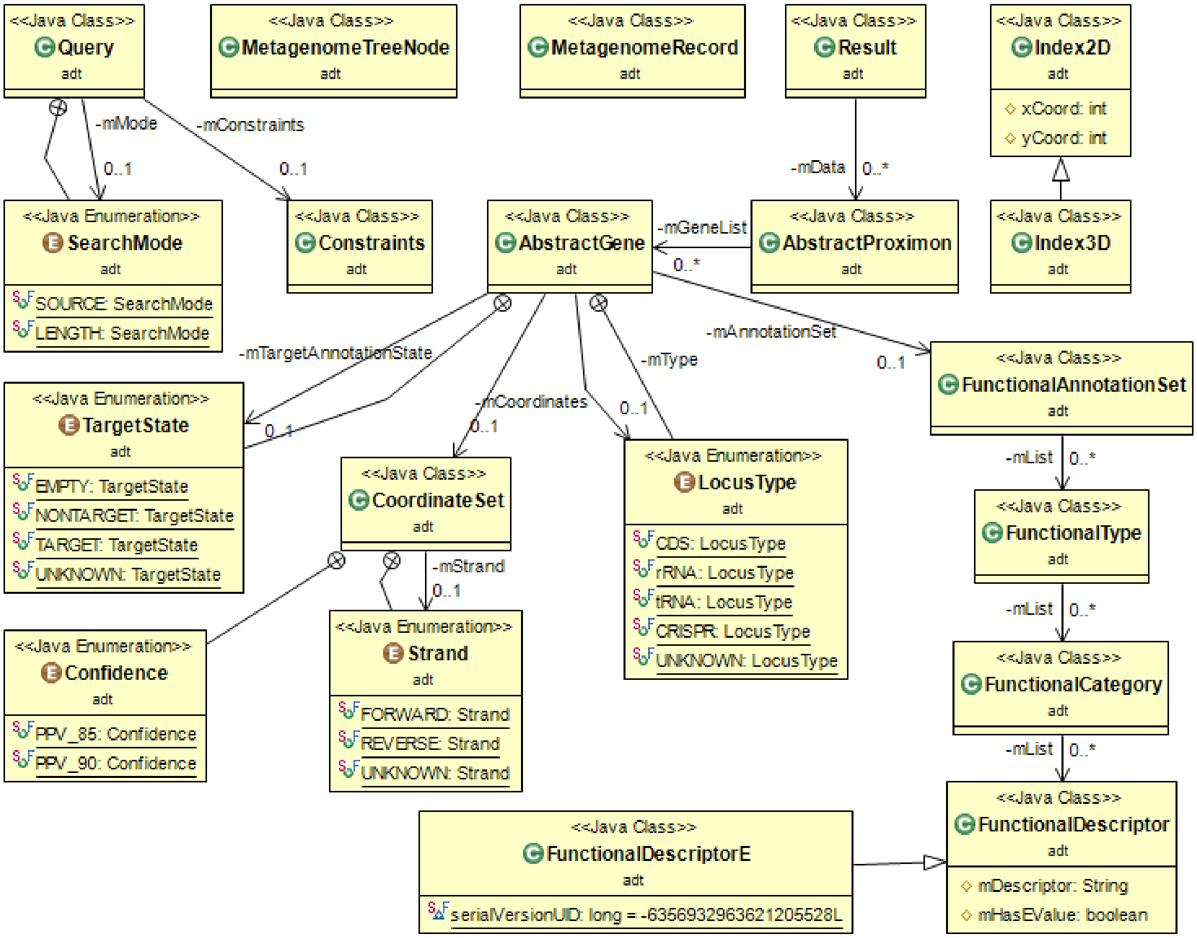


Figure A-2 ADT UML diagram. *The dependencies and cardinalities of the Java classes, interfaces, and enumerations of the Abstract Data Types (ADT) package are shown for the MetaProx search tool.*

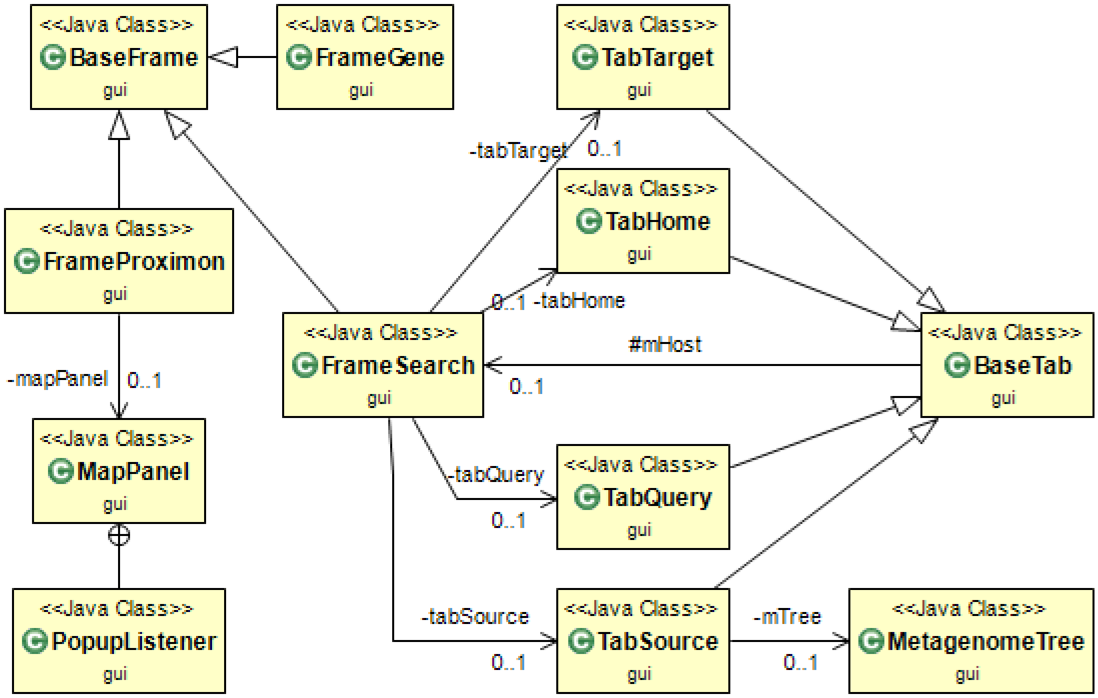


Figure A-3 GUI UML diagram. *The dependencies and cardinalities of the Java classes, interfaces, and enumerations of the Graphical User Interface (GUI) package are shown for the MetaProx search tool.*

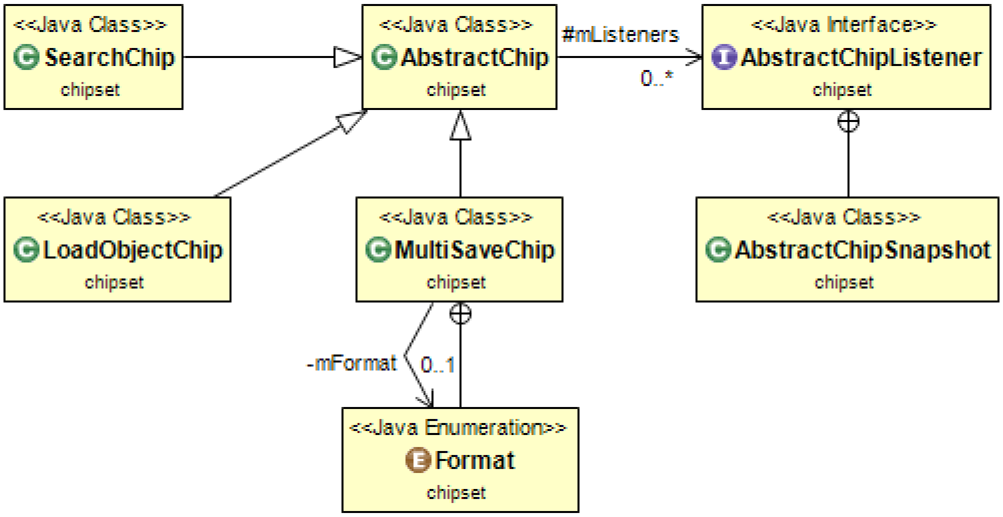


Figure A-4 Chipset UML diagram. *The dependencies and cardinalities of the Java classes, interfaces, and enumerations of the Chipset package are shown for the MetaProx search tool.*

Appendix B
IMG/M Datasets

Table B-1 IMG/M dataset usage and descriptions. Integrated Microbial Genomes with Microbiome samples (IMG/M) metagenomics datasets (Markowitz et al., 2012) are listed here and their specific usage is indicated with respect to Chapter 2 and Chapter 4. A description of each dataset is also included. All data were obtained by publicly available download.

Dataset	Chapter 2	Chapter 4	Description
2000000000	✓	✓	Sludge/US, Phrap Assembly
2000000001	✓	✓	Sludge/Australian, Phrap Assembly
2001000000	✓	✓	Sludge/US, Jazz Assembly
2001200000	✓	✓	Acidic water microbial communities from Richmond acid mine drainage
2001200001	✓	✓	Soil microbial communities from Waseca County, Minnesota Farm
2001200002	✓	✓	Fossil microbial community from Whale Fall, Santa Cruz Basin of the Pacific Ocean
2001200003	✓	✓	Fossil microbial community from Whale Fall, Santa Cruz Basin of the Pacific Ocean

2001200004	✓	✓	Fossil microbial community from Whale Fall, Santa Cruz Basin of the Pacific Ocean
2003000006	✓	✓	Air microbial communities from Singapore
2003000007	✓	✓	Air microbial communities from Singapore
2004000001	✓	✓	Oral TM7 microbial communities of Human
2004002000	✓	✓	Fecal microbiome of Human from distal gut of healthy adults
2004002001	✓	✓	Fecal microbiome of Human from distal gut of healthy adults
2004080001	✓	✓	Gut microbiome of Costa Rica Nasutitermes termites from P3 luminal contents
2004175000	✓	✓	Gut microbiome of Costa Rica Nasutitermes termites from P3 luminal contents
2004175001	✓	✓	Sediment archaeal communities from Eel River Basin

2004178001	✓	✓	Olavius algarvensis microbiome from Mediterranean sea
2004178002	✓	✓	Olavius algarvensis microbiome from Mediterranean sea
2004178003	✓	✓	Olavius algarvensis microbiome from Mediterranean sea
2004178004	✓	✓	Olavius algarvensis microbiome from Mediterranean sea
2004230000	✓	✓	Intestinal microbiome of Mouse lean and obese
2004230001	✓	✓	Intestinal microbiome of Mouse lean and obese
2004230002	✓	✓	Intestinal microbiome of Mouse lean and obese
2004230003	✘	✓	Intestinal microbiome of Mouse lean and obese
2004230004	✓	✓	Intestinal microbiome of Mouse lean and obese
2004247000	✓	✓	Saline water microbial communities from Guerrero Negro hypersaline mats

2004247001	✓	✓	Saline water microbial communities from Guerrero Negro hypersaline mats
2004247002	✓	✓	Saline water microbial communities from Guerrero Negro hypersaline mats
2004247003	✓	✓	Saline water microbial communities from Guerrero Negro hypersaline mats
2004247004	✓	✓	Saline water microbial communities from Guerrero Negro hypersaline mats
2004247005	✓	✓	Saline water microbial communities from Guerrero Negro hypersaline mats
2004247006	✓	✓	Saline water microbial communities from Guerrero Negro hypersaline mats
2004247007	✓	✓	Saline water microbial communities from Guerrero Negro hypersaline mats
2004247008	✓	✓	Saline water microbial communities from Guerrero Negro hypersaline mats
2004247009	✓	✓	Saline water microbial communities from Guerrero Negro hypersaline mats
2004247010	✓	✓	Oral TM7 microbial communities of Human

2005503000	✓	✓	Single-cell genome from subgingival tooth surface TM7b
2005560000	✓	✗	Gut microbiome of Costa Rica Nasutitermes termites from P3 luminal contents
2006207000	✓	✓	Methylotrophic community from Lake Washington sediment Methane enrichment
2006207001	✓	✓	Sediment methylotrophic communities from Lake Washington
2006207002	✓	✓	Sediment methylotrophic communities from Lake Washington
2006207003	✓	✓	Sediment methylotrophic communities from Lake Washington
2006207004	✓	✓	Sediment methylotrophic communities from Lake Washington
2006543005	✓	✓	Sediment methylotrophic communities from Lake Washington
2006543007	✓	✓	Groundwater microbial community from Contaminated well in in Oak Ridge, TN
2007300000	✓	✓	Sludge/US Virion (fgenesb)

2007309000	✓	✓	Bath Hot Springs, filamentous community
2007309001	✓	✓	Bath Hot Springs, planktonic community
2007427000	✓	✓	Groundwater microbial community from Contaminated well in in Oak Ridge, TN
2007915000	✓	✓	Wastewater Terephthalate-degrading communities from Bioreactor
2009439000	✓	✓	Hot spring microbial communities from Yellowstone Bison Hot Spring Pool
2009439003	✓	✓	Hot spring microbial communities from Yellowstone Bison Hot Spring Pool
2010170001	✓	✓	Hot spring microbial communities from Yellowstone Bison Hot Spring Pool
2010170002	✓	✓	Hot spring microbial communities from Yellowstone Bison Hot Spring Pool
2010170003	✓	✓	Hot spring microbial communities from Yellowstone Bison Hot Spring Pool
2010170004	✓	✓	Hot Spring microbial communities from Yellowstone Obsidian Hot Spring
2010388001	✓	✓	Poplar biomass bioreactor microbial communities from Brookhaven National Lab, NY

2010483000	✓	✓	Freshwater microbial communities from Lake Kinneret
2010483001	✓	✓	Freshwater microbial communities from Lake Kinneret
2010483002	✓	✓	Freshwater microbial communities from Lake Kinneret
2010483003	✓	✓	Freshwater microbial communities from Lake Kinneret
2010483004	✓	✓	Freshwater microbial communities from Lake Kinneret
2010483005	✓	✓	Freshwater microbial communities from Lake Kinneret
2010483006	✓	✓	Freshwater microbial communities from Lake Kinneret
2010483007	✓	✓	Freshwater microbial communities from Lake Kinneret
2010549000	✓	✓	Endophytic microbiome from Rice
2012990003	✓	✓	Marine microbial communities from six Antarctic regions
2013338003	✓	✓	Macropus eugenii forestomach microbiome from Canberra, Australia

2013515000	✓	✓	Hot spring microbial communities from Yellowstone National Park, US
2013515001	✓	✓	Hot spring microbial communities from Yellowstone National Park, US
2013515002	✓	✓	Hot spring microbial communities from Yellowstone National Park, US
2013843001	✓	✗	Activated sludge plasmid pools from Switzerland
2013843002	✓	✓	Groundwater dechlorinating community (KB-1) from synthetic mineral medium in Toronto, ON
2013843003	✓	✓	Hot spring microbial communities from Yellowstone National Park, US
2013954000	✓	✓	Hot spring microbial communities from Yellowstone National Park, US
2013954001	✓	✓	Hot spring microbial communities from Yellowstone National Park, US
2014031002	✓	✓	Hot spring microbial communities from Yellowstone National Park, US
2014031003	✓	✓	Hot spring microbial communities from Yellowstone National Park, US

2014031004	✓	✓	Hot spring microbial communities from Yellowstone National Park, US
2014031005	✓	✓	Hot spring microbial communities from Yellowstone National Park, US
2014031006	✓	✓	Hot spring microbial communities from Yellowstone National Park, US
2014031007	✓	✓	Hot spring microbial communities from Yellowstone National Park, US
2014613002	✗	✓	Marine planktonic communities from Hawaii Ocean Times Series Station (HOT/ALOHA)
2014613003	✓	✓	Marine planktonic communities from Hawaii Ocean Times Series Station (HOT/ALOHA)
2014642000	✓	✓	Marine planktonic communities from Hawaii Ocean Times Series Station (HOT/ALOHA)
2014642001	✓	✓	Marine planktonic communities from Hawaii Ocean Times Series Station (HOT/ALOHA)

2014642002	✓	✓	Marine planktonic communities from Hawaii Ocean Times Series Station (HOT/ALOHA)
2014642003	✓	✓	Marine planktonic communities from Hawaii Ocean Times Series Station (HOT/ALOHA)
2014642004	✓	✓	Marine planktonic communities from Hawaii Ocean Times Series Station (HOT/ALOHA)
2014730001	✓	✓	Soil microbial community from bioreactor at Alameda Naval Air Station, CA, contaminated with Chloroethene
2015219000	✓	✓	Hot spring microbial communities from Yellowstone National Park, US
2015219001	✓	✓	Hot spring microbial communities from Yellowstone National Park, US
2015219002	✓	✓	Hot spring microbial communities from Yellowstone National Park, US
2015219006	✓	✓	Archaeal viriome from Yellowstone Hot Springs

2015391000	✓	✓	Archaeal viriome from Yellowstone Hot Springs
2015391001	✓	✓	Hot spring microbial communities from Yellowstone National Park, US
2016842003	✓	✓	Hot spring microbial communities from Yellowstone National Park, US
2016842004	✓	✓	Hot spring microbial communities from Yellowstone National Park, US
2016842005	✓	✓	Hot spring microbial communities from Yellowstone National Park, US
2016842008	✓	✓	Hot spring microbial communities from Yellowstone National Park, US
2017108002	✓	✓	Freshwater propionate Anammox bacterial community from bioreactor in Nijmegen, The Netherlands
2019105001	✓	✓	Fecal microbiome of Canis familiaris
2019105002	✓	✓	Fecal microbiome of Canis familiaris
2020627002	✓	✓	Poplar biomass bioreactor microbial communities from Brookhaven National Lab, NY

2020627003	✓	✓	Benzene-Degrading Methanogenic communities from Bioreactor
2021593001	✓	✓	Macropus eugenii forestomach microbiome from Canberra, Australia
2021593002	✓	✗	Soil microbial pyrene-degrading mixed culture
2021593003	✓	✓	Trichonympha termites gut microbiome from Mt. Pinos, Los Padres National Forest, California
2021593004	✓	✓	Switchgrass rhizosphere bulk soil microbial community from Michigan, US
2022004001	✓	✗	Wastewater treatment Type I Accumulibacter community from EBPR Bioreactor in Madison, WI
2029527000	✓	✓	Green-waste compost microbial community from soild state bioreactor
2029527002	✓	✓	Soil microbial communities from FACE and OTC sites in USA
2029527003	✓	✓	Fungus garden microbial communities from Apterostigma dentigerum

2029527004	✓	✓	Fungus garden microbial communities from Atta cephalotes in Gamboa, Panama
2029527005	✓	✓	Atta columbica fungus garden (ACOFG)
2029527006	✓	✓	Atta columbica fungus garden (Fungus garden bottom)
2029527007	✓	✓	Fungus gallery microbial communities from Dendroctonus ponderosae
2030936000	✓	✗	Amitermes wheeleri gut microbiome from Arizona, USA, collected from P3 segment hindgut in fecal pellets under cow dung
2030936001	✓	✗	Laboratory Nasutitermes corniger gut microbiome from Florida, USA
2030936003	✓	✓	Freshwater propionate Anammox bacterial community from bioreactor in Nijmegen, The Netherlands
2030936005	✓	✓	Fungus garden microbial communities from Cyphomyrmex longiscapus
2030936006	✓	✓	Atta texana internal waste dump (Dump top)
2032320001	✓	✗	PCE-dechlorinating microbial communities from Ithaca, NY

2032320002	✓	✓	Soil microbial communities from FACE and OTC sites in USA
2032320003	✓	✓	Soil microbial communities from FACE and OTC sites in USA
2032320004	✓	✓	Soil microbial communities from FACE and OTC sites in USA
2032320005	✓	✓	Soil microbial communities from FACE and OTC sites in USA
2032320006	✓	✓	Soil microbial communities from FACE and OTC sites in USA
2032320007	✓	✓	Atta texana internal waste dump (Dump bottom)
2032320008	✓	✓	Fungus gallery microbial communities from Dendroctonus ponderosae
2032320009	✓	✓	Mountain Pine Beetle microbial communities from Alberta and British Columbia, Canada
2035918000	✓	✓	Fungus garden microbial communities from Acromyrmex echinatior in Panama
2035918001	✓	✓	Activated sludge plasmid pools from Switzerland

2035918002	✓	✓	Activated sludge plasmid pools from Switzerland
2035918003	✓	✓	Mountain Pine Beetle microbial communities from Alberta and British Columbia, Canada
2035918004	✓	✓	Soil microbial communities from FACE and OTC sites in USA
2035918005	✓	✓	Soil microbial communities from FACE and OTC sites in USA
2035918006	✓	✓	Soil microbial communities from FACE and OTC sites in USA
2038011000	✓	✓	Atta columbica fungus garden and dump (Dump top)
2040502000	✓	✓	Atta columbica fungus garden and dump (Dump bottom)
2040502001	✓	✓	Soil microbial communities from FACE and OTC sites in USA
2040502002	✓	✗	Switchgrass rhizosphere bulk soil microbial community from Michigan, US
2040502004	✓	✓	Marine Bacterioplankton communities from Antarctic

2040502005	✓	✓	Marine Bacterioplankton communities from Antarctic
2043231000	✓	✓	Xyleborus affinis microbiome from Bern, Switzerland
2044078000	✓	✓	Switchgrass, Maize and Miscanthus rhizosphere microbial communities from University of Illinois Energy Farm, Urbana, IL
2044078001	✓	✓	Switchgrass, Maize and Miscanthus rhizosphere microbial communities from University of Illinois Energy Farm, Urbana, IL
2044078002	✓	✓	Switchgrass, Maize and Miscanthus rhizosphere microbial communities from University of Illinois Energy Farm, Urbana, IL
2044078003	✓	✓	Switchgrass, Maize and Miscanthus rhizosphere microbial communities from University of Illinois Energy Farm, Urbana, IL

2044078004	✓	✓	Switchgrass, Maize and Miscanthus rhizosphere microbial communities from University of Illinois Energy Farm, Urbana, IL
2044078005	✓	✓	Switchgrass, Maize and Miscanthus rhizosphere microbial communities from University of Illinois Energy Farm, Urbana, IL
2044078006	✓	✓	Dendroctonus frontalis microbial community from Southwest Mississippi
2044078007	✓	✓	Dendroctonus frontalis Fungal community
2044078011	✓	✓	Xyleborus affinis microbiome from Bern, Switzerland
2046860004	✓	✓	Sediment microbial communities from Lake Washington, Seattle, for Methane and Nitrogen Cycles
2046860005	✓	✓	Sediment microbial communities from Lake Washington, Seattle, for Methane and Nitrogen Cycles

2046860006	✓	✓	Sediment microbial communities from Lake Washington, Seattle, for Methane and Nitrogen Cycles
2046860007	✓	✓	Sediment microbial communities from Lake Washington, Seattle, for Methane and Nitrogen Cycles
2046860008	✓	✓	Sediment microbial communities from Lake Washington, Seattle, for Methane and Nitrogen Cycles
2048955003	✓	✓	Poplar biomass bioreactor microbial communities from Brookhaven National Lab, NY
2049941001	✓	✓	Mixed alcohol bioreactor microbial communities from Texas A&M University
2051774008	✓	✓	Saline water microbial communities from Great Salt Lake, Utah
2053563001	✓	✓	Switchgrass and industrial compost incubating bioreactor microbial community from JBEI, CA, that is aerobic and thermophilic

2053563014	✓	✓	Sediment and Water microbial communities from Great Boiling Spring, Nevada
2058419001	✓	✓	Saline water microbial communities from Great Salt Lake, Utah
2058419002	✗	✓	Saline water microbial communities from Great Salt Lake (South Arm Stromatolite)
2058419003	✓	✓	Saline water microbial communities from Great Salt Lake, Utah
2058419004	✓	✓	Sediment and Water microbial communities from Great Boiling Spring, Nevada
2061766000	✓	✓	Benzene-Degrading Methanogenic communities from Bioreactor
2061766001	✓	✓	Switchgrass and industrial compost incubating bioreactor microbial community from JBEI, CA, that is aerobic and thermophilic
2061766005	✓	✗	Saline water microbial communities from Elkhorn Slough hypersaline mats

2061766006	✓	✗	Saline water microbial communities from Elkhorn Slough hypersaline mats
2061766008	✓	✓	Freshwater microbial communities from Antarctic Deep Lake
2065487013	✓	✓	Fungus-growing Termite worker microbial community from South Africa
2065487014	✓	✓	Fungus garden microbial community from termites in South Africa
2067725009	✓	✗	Permafrost microbial communities from Central Alaska
2070309010	✓	✗	Bankia setacea gill microbiome from Puget Sound, WA
2077657003	✓	✓	Sediment and Water microbial communities from Great Boiling Spring, Nevada
2077657005	✓	✓	Mixed alcohol bioreactor microbial communities from Texas A&M University
2077657006	✓	✓	Freshwater microbial communities from Mississippi River

2077657007	✓	✓	Freshwater microbial communities from Mississippi River
2077657008	✓	✓	Bovine rumen viral communities from University of Illinois Dairy Farm in Urbana, IL
2077657009	✓	✓	Bovine rumen viral communities from University of Illinois Dairy Farm in Urbana, IL
2077657010	✓	✓	Saline water microbial communities from Great Salt Lake, Utah
2077657013	✓	✗	Marine Bacterioplankton communities from Antarctic
2077657014	✓	✓	Marine sediment archaeal communities from Santa Barbara Basin, CA, that are methane-oxidizing
2077657018	✓	✓	Marine sediment archaeal communities from Santa Barbara Basin, CA, that are methane-oxidizing
2077657019	✓	✓	Marine sediment archaeal communities from Santa Barbara Basin, CA, that are methane-oxidizing

2077657020	✓	✗	Marine Bacterioplankton communities from Antarctic
2077657023	✓	✓	Hot spring microbial communities from Yellowstone National Park, US
2077657024	✓	✓	Hot spring microbial communities from Yellowstone National Park, US
2081372006	✓	✓	Soil microbial communities from FACE and OTC sites in USA
2081372007	✓	✓	Freshwater microbial communities from Antarctic Deep Lake
2081372008	✓	✓	Wastewater bioreactor microbial communities from Singapore and Univ of Illinois at Urbana, that are terephthalate-degrading
2084038000	✓	✓	Bovine rumen viral communities from University of Illinois Dairy Farm in Urbana, IL
2084038008	✓	✓	Xyleborus affinis microbiome from Bern, Switzerland

2084038009	✓	✓	Sediment microbial communities from Lake Washington, Seattle, for Methane and Nitrogen Cycles
2084038011	✓	✓	Freshwater microbial communities from Antarctic Deep Lake
2084038012	✓	✓	Marine sediment microbial communities from Kolumbo Volcano mats, Greece
2084038013	✓	✓	Anoplophora glabripennis gut microbiome from Worchester, MA
2084038018	✓	✓	Fungus garden microbial communities from Trachymyrmex in Gamboa, Panama
2084038019	✓	✓	Freshwater microbial communities from Antarctic Deep Lake
2084038020	✓	✓	Sediment and Water microbial communities from Great Boiling Spring, Nevada
2084038021	✓	✓	Marine sediment archaeal communities from Santa Barbara Basin, CA, that are methane-oxidizing

2088090005	✓	✓	Sediment microbial communities from Lake Washington, Seattle, for Methane and Nitrogen Cycles
2088090006	✓	✓	Sediment microbial communities from Lake Washington, Seattle, for Methane and Nitrogen Cycles
2088090007	✓	✓	Sediment microbial communities from Lake Washington, Seattle, for Methane and Nitrogen Cycles
2088090009	✓	✓	Sediment microbial communities from Lake Washington, Seattle, for Methane and Nitrogen Cycles
2088090012	✓	✓	Coastal water and sediment microbial communities from Arctic Ocean, off the coast from Alaska
2088090013	✓	✓	Sediment microbial communities from Lake Washington, Seattle, for Methane and Nitrogen Cycles
2088090016	✓	✗	Hoatzin crop microbial communities from Cojedes, Venezuela

2088090017	✓	✓	Marine microbial communities from Deepwater Horizon Oil Spill
2088090019	✓	✗	PCE-dechlorinating microbial communities from Ithaca, NY
2088090027	✓	✓	Sediment and Water microbial communities from Great Boiling Spring, Nevada
2088090030	✓	✓	Marine sediment microbial communities from Kolumbo Volcano mats, Greece
2088090031	✓	✓	Freshwater microbial communities from Lake Sakinaw,Canada
2088090036	✓	✗	Hoatzin crop microbial communities from Cojedes, Venezuela
2100351001	✓	✓	Coastal water and sediment microbial communities from Arctic Ocean, off the coast from Alaska
2100351002	✓	✗	Hoatzin crop microbial communities from Cojedes, Venezuela
2100351005	✓	✗	Arabidopsis rhizosphere microbial communities from University of North Carolina

2100351006	✓	✓	Coastal water and sediment microbial communities from Arctic Ocean, off the coast from Alaska
2100351007	✓	✓	Sediment microbial communities from Lake Washington, Seattle, for Methane and Nitrogen Cycles
2100351008	✓	✓	Hot spring microbial communities from Yellowstone National Park, US
2100351009	✓	✓	Sediment and Water microbial communities from Great Boiling Spring, Nevada
2100351010	✓	✓	Groundwater dechlorinating microbial community from Kitchener, Ontario, containing dehalobacter
2100351011	✓	✓	Coastal water and sediment microbial communities from Arctic Ocean, off the coast from Alaska
2100351012	✓	✓	Coastal water and sediment microbial communities from Arctic Ocean, off the coast from Alaska

2100351014	✓	✗	Freshwater microbial communities from Antarctic Deep Lake
2100351015	✓	✓	Freshwater microbial communities from Antarctic Deep Lake
2100351016	✓	✓	Sirex noctilio microbiome from Pennsylvania
2119805007	✓	✗	Hot spring microbial communities from Yellowstone National Park, US
2119805009	✓	✗	Soil microbial communities from FACE and OTC sites in USA
2119805010	✓	✗	Soil microbial communities from FACE and OTC sites in USA
2119805011	✓	✗	Soil microbial communities from FACE and OTC sites in USA
2119805012	✓	✗	Soil microbial communities from FACE and OTC sites in USA
2124908000	✓	✓	Sediment microbial communities from Lake Washington, Seattle, for Methane and Nitrogen Cycles
2124908001	✓	✗	Soil microbial communities from FACE and OTC sites in USA

2124908006	✓	✗	Soil microbial communities from FACE and OTC sites in USA
2124908007	✓	✗	Soil microbial communities from FACE and OTC sites in USA
2124908008	✓	✗	Soil microbial communities from FACE and OTC sites in USA
2124908009	✓	✗	Soil microbial communities from FACE and OTC sites in USA
2124908018	✓	✗	Switchgrass rhizosphere microbial community from Michigan, US
2124908019	✓	✗	Switchgrass rhizosphere microbial community from Michigan, US
2124908021	✓	✗	Switchgrass rhizosphere microbial community from Michigan, US
2124908023	✓	✗	Switchgrass rhizosphere bulk soil microbial community from Michigan, US
2124908025	✓	✗	Miscanthus rhizosphere microbial communities from Kellogg Biological Station, MSU

2124908027	✓	✗	Miscanthus rhizosphere microbial communities from Kellogg Biological Station, MSU
2124908038	✓	✗	Soil microbial communities from permafrost in Bonanza Creek, Alaska
2124908040	✓	✗	Soil microbial communities from permafrost in Bonanza Creek, Alaska
2124908041	✓	✗	Soil microbial communities from permafrost in Bonanza Creek, Alaska
2124908043	✓	✗	Soil microbial communities from permafrost in Bonanza Creek, Alaska
2124908044	✓	✗	Soil microbial communities from permafrost in Bonanza Creek, Alaska
2140918001	✓	✗	Hot spring microbial communities from Yellowstone National Park, US
2140918003	✓	✗	Marine sediment archaeal communities from Santa Barbara Basin, CA, that are methane-oxidizing
2140918004	✓	✗	Marine sediment archaeal communities from Santa Barbara Basin, CA, that are methane-oxidizing

2140918005	✓	✗	Coastal water and sediment microbial communities from Arctic Ocean, off the coast from Alaska
2140918006	✓	✗	Soil microbial communities from permafrost in Bonanza Creek, Alaska
2140918012	✓	✗	Sediment microbial communities from Lake Washington, Seattle, for Methane and Nitrogen Cycles
2140918017	✓	✗	Freshwater microbial communities from Antarctic Deep Lake
2140918027	✓	✗	Freshwater microbial communities from Antarctic Deep Lake
2149837004	✓	✗	Sediment and Water microbial communities from Great Boiling Spring, Nevada
2149837005	✓	✗	Sediment and Water microbial communities from Great Boiling Spring, Nevada
2149837010	✓	✓	Fresh water microbial communities from LaBonte Lake, Laramie, Wyoming

2149837011	✓	✓	Fresh water microbial communities from LaBonte Lake, Laramie, Wyoming
2149837029	✓	✗	Sediment microbial communities from Lake Washington, Seattle, for Methane and Nitrogen Cycles
2149837030	✓	✗	Sediment microbial communities from Lake Washington, Seattle, for Methane and Nitrogen Cycles
2156126002	✓	✗	Biofuel metagenome
2156126005	✓	✗	Marine Trichodesmium cyanobacterial communities from the Bermuda Atlantic Time-Series
2156126009	✓	✗	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones
2156126010	✓	✗	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones
2156126011	✓	✗	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones

2156126012	✓	✗	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones
2156126013	✓	✗	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones
2162886003	✓	✗	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones
2162886004	✓	✗	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones
2162886005	✓	✗	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones
2162886006	✓	✗	Switchgrass rhizosphere microbial community from Michigan, US
2162886007	✓	✗	Switchgrass rhizosphere microbial community from Michigan, US

2162886011	✓	✘	Miscanthus rhizosphere microbial communities from Kellogg Biological Station, MSU
2162886012	✓	✘	Miscanthus rhizosphere microbial communities from Kellogg Biological Station, MSU
2162886013	✓	✘	Switchgrass rhizosphere microbial community from Michigan, US
2166559021	✓	✓	Fresh water microbial communities from LaBonte Lake, Laramie, Wyoming
2166559022	✓	✓	Fresh water microbial communities from LaBonte Lake, Laramie, Wyoming
2166559023	✓	✓	Fresh water microbial communities from LaBonte Lake, Laramie, Wyoming
2166559024	✓	✘	Biofuel metagenome
2166559025	✓	✘	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones
2189573006	✓	✘	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones

2189573007	✓	✘	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones
2189573008	✓	✘	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones
2189573009	✓	✘	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones
2189573010	✓	✘	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones
2189573011	✓	✘	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones
2189573012	✓	✘	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones
2189573013	✓	✘	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones

2189573014	✓	✘	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones
2189573015	✓	✘	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones
2189573016	✓	✘	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones
2189573017	✓	✘	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones
2189573018	✓	✘	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones
2189573019	✓	✘	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones
2189573022	✓	✘	Soil microbial communities from Puerto Rico rain forest, that decompose switchgrass

2189573023	✓	✓	Fresh water microbial communities from LaBonte Lake, Laramie, Wyoming
2189573029	✓	✗	Bankia setacea gill microbiome from Puget Sound, WA
2199034001	✓	✗	Freshwater microbial communities from Trout Bog Lake, WI and Lake Mendota, WI
2199034002	✓	✗	Soil microbial community from bioreactor at Alameda Naval Air Station, CA, contaminated with Chloroethene
2199352000	✓	✗	Freshwater microbial communities from Trout Bog Lake, WI and Lake Mendota, WI
2199352001	✓	✗	Freshwater microbial communities from Trout Bog Lake, WI and Lake Mendota, WI
2199352002	✓	✗	Freshwater microbial communities from Trout Bog Lake, WI and Lake Mendota, WI

2199352003	✓	✗	Freshwater microbial communities from Trout Bog Lake, WI and Lake Mendota, WI
2199352004	✓	✗	Freshwater microbial communities from Trout Bog Lake, WI and Lake Mendota, WI
2199352005	✓	✗	Freshwater microbial communities from Trout Bog Lake, WI and Lake Mendota, WI
2199352006	✓	✗	Soil microbial communities from four geographically distinct crusts in the Colorado Plateau and Sonoran desert
2199352009	✓	✗	Marine subseafloor sediment microbial communities from Peru Margin, Ocean Drilling Program Site 1229
2199352035	✓	✗	Decomposing wood compost microbial communities from rain forest habitat in Puerto Rico, that are thermophilic
2209111000	✓	✗	Soil microbial communities from four geographically distinct crusts in the Colorado Plateau and Sonoran desert

2209111006	✓	✗	Arabidopsis rhizosphere microbial communities from University of North Carolina
2222084007	✓	✗	Freshwater microbial communities from Lake Vostok at Ice accretion
2222084012	✓	✗	Wild Panda gut microbiome from Shaanxi, China
2222084013	✓	✗	Wild Panda gut microbiome from Shaanxi, China
2222084014	✓	✗	Wild Panda gut microbiome from Shaanxi, China
2225789020	✓	✗	Freshwater propionate Anammox bacterial community from bioreactor in Nijmegen, The Netherlands

Appendix C
Functional Modules

Table C-1 Metagenomic functional modules with abbreviated annotations. The general features for the complete collection of inferred functional modules from each network in Chapter 4 are shown including the source network, the score assigned by MINE, the number of nodes and edges that compose the module, and the member annotations derived from the nodes. Member annotations remain in their original abbreviated form.

Type	Module	Score	Nodes	Edges	Members
Cellulase MetaCyc	1	14.00	14	91	VALSYN-PWY, VALDEG-PWY, PWY-5108, PWY-5104, PWY-5103, PWY-5101, PWY-5078, PWY-5076, PWY-5057, LEUSYN-PWY, LEU-DEG2-PWY, ILEUSYN-PWY, ALANINE-VALINESYN-PWY, ILEUDEG-PWY
	2	13.11	28	177	YEAST-4AMINOBUTMETAB-PWY, TOLSULFDEG-PWY, PWY0-1221, PWY-6473, PWY-5537, PWY-5482, PWY-5305, PWY-5195, PWY-4321, PWY-321, PWY-282, PWY-1121,

				P105-PWY, GLYCOLYSIS-TCA-GLYOX-BYPASS, 4TOLCARBDEG-PWY, 4AMINOBUTMETAB-PWY, 3-HYDROXYPHENYLACETATE-DEGRADATION-PWY, ANARESP1-PWY, PWY-5154, GLUTORN-PWY, PWY-0, DAPLYSINESYN-PWY, PWY-1822, LYSDEGII-PWY, ECASYN-PWY, PWY-5784, ARGSYNBSUB-PWY, PWYQT-4475
3	6.00	9	24	PWY-882, PWY-5659, PWY-3881, PWY-3861, MANNCAT-PWY, ARO-PWY, PWY-6164, PWY-2681, PWY-5381
4	5.00	5	10	PWY-2781, PWY-6471, PWY-6470, PWY-5265, PWY-6385
5	3.71	8	13	PWY-1001, ASPARTATESYN-PWY, CYSTEINE-DEG-PWY,

					GLUTDEG-PWY, MALATE-ASPARTATE-SHUTTLE-PWY, PWY-5913, PWY-6318, RUMP-PWY
Cellulase COG	1	15.00	15	105	cog3845, cog4603, cog1079, cog1335, cog1123, cog0619, cog0163, cog0043, cog4577, cog1001, cog1878, cog3665, cog1957, cog1069, cog0624
	2	10.60	11	53	cog0836, cog3594, cog1595, cog2148, cog1596, cog3206, cog0728, cog3664, cog1215, cog0438, cog1216
	3	9.00	9	36	cog0735, cog1059, cog1376, cog3185, cog1014, cog1013, cog0674, cog0541, cog1146
	4	9.00	9	36	cog1363, cog1923, cog0324, cog0323, cog0123, cog0249, cog1691, cog0621, cog0768

5	7.00	7	21	cog1725, cog1131, cog0301, cog1058, cog3393, cog0771, cog4100
6	6.67	10	30	cog1206, cog0122, cog1636, cog1482, cog0850, cog0337, cog1555, cog1611, cog1137, cog2825
7	6.00	6	15	cog3480, cog0669, cog0742, cog4471, cog0588, cog0772
8	6.00	6	15	cog0601, cog1173, cog1192, cog4608, cog3405, cog0444
9	5.00	5	10	cog0622, cog0596, cog1024, cog0813, cog1250
10	5.00	5	10	cog1216, cog0438, cog1134, cog1538, cog1091
11	5.00	11	25	cog4124, cog2211, cog1482, cog3458, cog2942, cog3459, cog4206, cog3345, cog2152, cog3934, cog0747

12	4.33	7	13	cog3414, cog1762, cog1299, cog0235, cog1349, cog3775, cog0036
13	4.00	4	6	cog4186, cog4848, cog0073, cog1986
14	4.00	4	6	cog0644, cog0642, cog2755, cog4771
15	4.00	4	6	cog3622, cog1477, cog3590, cog0673
16	4.00	4	6	cog1445, cog1299, cog0006, cog1080
17	4.00	4	6	cog0542, cog3669, cog4206, cog3250
18	4.00	4	6	cog4986, cog1116, cog0105, cog4754
19	4.00	4	6	cog2211, cog1874, cog3507, cog2730
20	3.67	7	11	cog1192, cog5010, cog3405, cog3063, cog0340, cog0812, cog1696
21	3.00	3	3	cog2894, cog0719, cog0432

	22	3.00	3	3	cog0548, cog4992, cog0002
	23	3.00	3	3	cog2159, cog0318, cog1167
	24	3.00	3	3	cog2893, cog3715, cog3716
	25	3.00	3	3	cog0553, cog0769, cog1285
	26	3.00	3	3	cog2877, cog1063, cog1630
	27	3.00	3	3	cog2442, cog3635, cog2402
	28	3.00	3	3	cog0572, cog0745, cog5002
	29	3.00	3	3	cog1175, cog1653, cog0395
	30	3.00	3	3	cog0180, cog1181, cog2919
	31	3.00	3	3	cog0489, cog0436, cog0794
	32	3.00	3	3	cog2058, cog0081, cog0244
	33	3.00	3	3	cog0364, cog1629, cog3940
Gut KEGG	1	6.17	13	37	keggM00095, keggM00003, keggM00002, keggM00001, keggM00092, keggM00091, keggM00313, keggM00314, keggM00308, keggM00716, keggM00307, keggM00309, keggM00094
	2	5.33	7	16	keggM00030, keggM00029, keggM00027, keggM00249,

				keggM00095, keggM00037, keggM00028
3	5.00	5	10	keggM00103, keggM00051, keggM00025, keggM00023, keggM00024
4	4.50	9	18	keggM00683, keggM00011, keggM00012, keggM00010, keggM00009, keggM00682, keggM00679, keggM00302, keggM00684
5	4.00	4	6	keggM00296, keggM00294, keggM00004, keggM00007
6	4.00	6	10	keggM00370, keggM00366, keggM00376, keggM00375, keggM00388, keggM00250
7	3.33	4	5	keggM00031, keggM00270, keggM00286, keggM00293
8	3.00	3	3	keggM00037, keggM00210, keggM00118
9	3.00	3	3	keggM00275, keggM00159, keggM00160

	10	3.00	3	3	keggM00035, keggM00036, keggM00033
	11	3.00	3	3	keggM00649, keggM00245, keggM00648
Gut TIGRFAM	1	8.24	18	70	tigr03635, tigr01050, tigr01044, tigr01171, tigr01009, tigr01164, tigr01067, tigr00012, tigr03953, tigr01049, tigr03625, tigr03654, tigr01079, tigr00060, tigr00967, tigr01021, tigr01071, tigr01308
	2	8.00	8	28	tigr01309, tigr01020, tigr01080, tigr03673, tigr03630, tigr01158, tigr00012, tigr01008
	3	7.00	7	21	tigr01148, tigr01112, tigr03256, tigr03259, tigr03264, tigr03257, tigr03260
	4	5.20	6	13	tigr00978, tigr00036, tigr00657, tigr00674, tigr00656, tigr00683
	5	5.00	5	10	tigr00736, tigr01819, tigr01916, tigr00904, tigr01922

6	4.50	5	9	tigr02469, tigr01444, tigr00312, tigr01465, tigr02467
7	4.00	4	6	tigr03677, tigr03680, tigr00231, tigr00491
8	4.00	4	6	tigr01038, tigr03636, tigr03626, tigr03672
9	4.00	4	6	tigr01260, tigr01091, tigr01145, tigr01144
10	4.00	4	6	tigr00123, tigr01165, tigr01506, tigr02454
11	4.00	4	6	tigr01017, tigr03632, tigr00059, tigr02027
12	3.60	6	9	tigr00081, tigr01162, tigr01134, tigr00639, tigr00877, tigr00878
13	3.50	5	7	tigr01216, tigr01145, tigr01039, tigr00962, tigr01146
14	3.50	5	7	tigr00158, tigr00165, tigr00621, tigr02937, tigr00166
15	3.33	4	5	tigr00888, tigr00566, tigr00564, tigr01245

16	3.33	4	5	tigr03992, tigr01208, tigr01141, tigr02495
17	3.33	4	5	tigr01947, tigr01944, tigr01943, tigr01948
18	3.33	4	5	tigr01128, tigr03725, tigr00150, tigr01575
19	3.33	4	5	tigr01071, tigr00008, tigr00500, tigr01351
20	3.33	4	5	tigr00038, tigr00494, tigr00478, tigr01951
21	3.33	4	5	tigr00922, tigr00186, tigr00964, tigr02937
22	3.00	3	3	tigr00560, tigr03455, tigr01163
23	3.00	5	6	tigr00174, tigr00585, tigr01070, tigr01574, tigr03156
24	3.00	3	3	tigr01496, tigr00277, tigr00063
25	3.00	3	3	tigr00092, tigr00453, tigr03064
26	3.00	3	3	tigr00670, tigr00857, tigr00240
27	3.00	3	3	tigr00281, tigr00275, tigr00093
28	3.00	3	3	tigr00157, tigr01163, tigr01378
29	3.00	3	3	tigr01463, tigr00234, tigr00311

30	3.00	3	3	tigr00005, tigr02227, tigr00648
31	3.00	3	3	tigr00088, tigr02273, tigr02227
32	3.00	3	3	tigr00097, tigr00408, tigr03552
33	3.00	3	3	tigr00255, tigr00690, tigr03263
34	3.00	3	3	tigr02127, tigr00336, tigr01037
35	3.00	3	3	tigr01032, tigr00001, tigr00168
36	3.00	3	3	tigr00128, tigr03151, tigr00517
37	3.00	3	3	tigr02209, tigr02892, tigr02893
38	3.00	3	3	tigr02124, tigr01287, tigr03959
39	3.00	3	3	tigr01035, tigr01470, tigr03277
40	3.00	3	3	tigr03740, tigr03732, tigr03733
41	3.00	3	3	tigr00252, tigr00368, tigr00732
42	3.00	3	3	tigr00253, tigr03595, tigr00488
43	3.00	3	3	tigr00665, tigr01203, tigr02432
44	3.00	3	3	tigr01114, tigr01149, tigr02507
45	3.00	3	3	tigr02135, tigr00974, tigr00972
46	3.00	3	3	tigr00184, tigr00877, tigr00762
47	3.00	3	3	tigr01088, tigr00033, tigr01357
48	3.00	3	3	tigr01560, tigr01563, tigr01554
49	3.00	3	3	tigr01979, tigr00420, tigr01994
50	3.00	3	3	tigr00246, tigr00180, tigr00637

	51	3.00	3	3	tigr01855, tigr00007, tigr00735
	52	3.00	3	3	tigr01085, tigr03188, tigr00735
	53	3.00	3	3	tigr00017, tigr00216, tigr00530
	54	3.00	3	3	tigr02065, tigr00291, tigr03633
	55	3.00	3	3	tigr00069, tigr00070, tigr01141
	56	3.00	3	3	tigr01968, tigr01215, tigr02210
	57	3.00	3	3	tigr00133, tigr00132, tigr00135
Gut Intersection	1	8.17	13	49	tigr03635, tigr01050, tigr01044, tigr03654, tigr00060, tigr01079, tigr03953, tigr01171, tigr03625, tigr01009, tigr01164, tigr00012, tigr01067
	2	4.00	4	6	tigr00967, tigr01021, tigr01308, tigr01071
	3	4.00	4	6	tigr01017, tigr03632, tigr02027, tigr00059
	4	3.50	5	7	tigr02467, tigr02469, tigr01444, tigr00312, tigr01465
	5	3.50	5	7	tigr01216, tigr01039, tigr00962, tigr01145, tigr01146

	6	3.33	4	5	tigr00683, tigr00674, tigr00978, tigr00036
	7	3.33	4	5	tigr03992, tigr01208, tigr02495, tigr01141
	8	3.00	3	3	tigr00157, tigr01378, tigr01163
	9	3.00	3	3	tigr01245, tigr00566, tigr00564
	10	3.00	3	3	tigr01260, tigr01144, tigr01145
	11	3.00	3	3	tigr03740, tigr03733, tigr03732
	12	3.00	3	3	tigr00665, tigr02432, tigr01203
	13	3.00	3	3	tigr02135, tigr00972, tigr00974
	14	3.00	3	3	tigr00735, tigr01855, tigr00007
	15	3.00	3	3	tigr00736, tigr01819, tigr01916
	16	3.00	3	3	tigr00017, tigr00530, tigr00216
	17	3.00	3	3	tigr00008, tigr01351, tigr00500
	18	3.00	3	3	tigr00133, tigr00135, tigr00132
	19	3.00	3	3	tigr01070, tigr00585, tigr00174
Gut Difference	1	7.71	8	27	tigr01309, tigr01020, tigr00012, tigr01008, tigr01158, tigr03673, tigr01080, tigr03630

2	6.67	7	20	tigr01148, tigr01112, tigr03259, tigr03257, tigr03256, tigr03264, tigr03260
3	4.00	4	6	tigr01038, tigr03636, tigr03672, tigr03626
4	3.67	7	11	tigr00231, tigr01574, tigr00089, tigr00174, tigr03677, tigr03680, tigr00491
5	3.50	5	7	tigr00736, tigr01819, tigr01916, tigr01922, tigr00904
6	3.50	5	7	tigr00158, tigr02937, tigr00165, tigr00166, tigr00621
7	3.33	4	5	tigr00123, tigr01165, tigr02454, tigr01506
8	3.00	3	3	tigr00291, tigr02065, tigr03633
9	3.00	3	3	tigr00275, tigr00281, tigr00093
10	3.00	3	3	tigr02210, tigr01968, tigr01215
11	3.00	3	3	tigr00420, tigr01979, tigr01994
12	3.00	3	3	tigr01171, tigr03625, tigr01049
13	3.00	3	3	tigr00964, tigr00186, tigr00922
14	3.00	3	3	tigr00092, tigr03064, tigr00453

15	3.00	3	3	tigr00670, tigr00240, tigr00857
16	3.00	3	3	tigr00150, tigr01575, tigr01128
17	3.00	3	3	tigr00200, tigr01125, tigr00560
18	3.00	3	3	tigr00877, tigr00184, tigr00762
19	3.00	3	3	tigr03455, tigr01163, tigr00560
20	3.00	3	3	tigr00180, tigr00246, tigr00637

Table C-2 Top ranked metagenomic functional modules with verbose annotations. The general features of the top-three highest scoring functional module from each network in Chapter 4 are shown including the source network, the score assigned by MINE, the number of nodes and edges that compose the module, and the member annotations derived from the nodes. Member annotations have been translated into their corresponding verbose form and are sorted in ascending lexicographical order and delimited using the semicolon symbol. For all top ranked modules each verbose annotation is unique.

Type	Module	Score	Nodes	Edges	Members
Cellulase MetaCyc	1	14.00	14	91	Alanine biosynthesis I; Isoleucine biosynthesis I (from threonine); Isoleucine biosynthesis II; Isoleucine biosynthesis III; Isoleucine biosynthesis IV; Isoleucine biosynthesis V; Isoleucine degradation I; Isoleucine degradation II; Leucine biosynthesis; Leucine degradation I; Leucine degradation III; Valine biosynthesis; Valine degradation I; Valine degradation II
	2	13.11	28	177	4-aminobutyrate degradation IV; 4-hydroxyphenylacetate degradation; 4-toluenecarboxylate degradation; 4-toluenesulfonate degradation I;

	Arginine biosynthesis II (acetyl cycle); Arginine biosynthesis III; Artemisinin biosynthesis; Bixin biosynthesis; Cuticular wax biosynthesis; Cutin biosynthesis; Enterobacterial common antigen biosynthesis; Glucosinolate biosynthesis from hexahomomethionine; Glutamate degradation IV; IAA biosynthesis II; IAA conjugate biosynthesis II; Lysine biosynthesis I; Lysine degradation III; Ornithine biosynthesis; Pathway: TCA cycle variation I; Putrescine degradation II; Putrescine degradation III; Pyruvate fermentation to acetate II; Pyruvate fermentation to acetate V; Respiration (anaerobic); Suberin biosynthesis; Superpathway of 4-aminobutyrate degradation; Superpathway of glycolysis, pyruvate

					dehydrogenase, TCA, and glyoxylate bypass
	3	6.00	9	24	3-dehydroquinate biosynthesis I; Ascorbate biosynthesis I (L-galactose pathway); Chorismate biosynthesis I; D-mannose degradation; GDP-mannose biosynthesis; Mannitol biosynthesis; Mannitol degradation II; Pyridine nucleotide cycling (plants); Trans-zeatin biosynthesis
Cellulase COG	1	15.00	15	105	3-polyprenyl-4-hydroxybenzoate decarboxylase; 3-polyprenyl-4-hydroxybenzoate decarboxylase and related decarboxylases; ABC-type cobalt transport system, permease component CbiQ and related transporters; ABC-type uncharacterized transport system, permease component; ABC-type uncharacterized transport systems, ATPase components; ATPase components of various ABC-type

				transport systems, contain duplicated ATPase; Acetylornithine deacetylase/Succinyl-diaminopimelate desuccinylase and related deacylases; Adenine deaminase; Amidases related to nicotinamidase; Carbon dioxide concentrating mechanism/carboxysome shell protein; Inosine-uridine nucleoside N-ribohydrolase; Predicted metal-dependent hydrolase; Ribulose kinase; Uncharacterized ABC-type transport system, permease component; Uncharacterized conserved protein
2	10.60	11	53	Beta-xylosidase; DNA-directed RNA polymerase specialized sigma subunit, sigma24 homolog; Fucose 4-O-acetylase and related acetyltransferases; Glycosyltransferase; Glycosyltransferases, probably involved in cell wall biogenesis; Mannose-1-phosphate

				guanylyltransferase; Periplasmic protein involved in polysaccharide export; Predicted glycosyltransferases; Sugar transferases involved in lipopolysaccharide synthesis; Uncharacterized membrane protein, putative virulence factor; Uncharacterized protein involved in exopolysaccharide biosynthesis
3	9.00	9	36	4-hydroxyphenylpyruvate dioxygenase and related hemolysins; Fe2+/Zn2+ uptake regulation proteins; Ferredoxin; Pyruvate:ferredoxin oxidoreductase and related 2-oxoacid:ferredoxin oxidoreductases, alpha subunit; Pyruvate:ferredoxin oxidoreductase and related 2-oxoacid:ferredoxin oxidoreductases, beta subunit; Pyruvate:ferredoxin oxidoreductase and related 2-oxoacid:ferredoxin oxidoreductases, gamma subunit;

					Signal recognition particle GTPase; Thermostable 8-oxoguanine DNA glycosylase; Uncharacterized protein conserved in bacteria
Gut KEGG	1	6.17	13	37	Bacitracin transport system; C5 isoprenoid biosynthesis, mevalonate pathway; Ceramide biosynthesis; Eicosanoid biosynthesis, arachidonate => 8(S)-HETE; Gluconeogenesis, oxaloacetate => fructose-6P; Glycolysis (Embden-Meyerhof pathway), glucose => pyruvate; Glycolysis, core module involving three-carbon compounds; Indolepyruvate:ferredoxin oxidoreductase; Non-phosphorylative Entner-Doudoroff pathway, gluconate => glyceraldehyde + pyruvate; Phosphatidylcholine (PC) biosynthesis, PE => PC; Phosphatidylethanolamine (PE) biosynthesis, ethanolamine => PE;

				Pyruvate oxidation, pyruvate => acetyl-CoA; Semi-phosphorylative Entner-Doudoroff pathway, gluconate => glyceraldehyde-3P + pyruvate
2	5.33	7	16	C5 isoprenoid biosynthesis, mevalonate pathway; Capsular polysaccharide transport system; GABA (gamma-Aminobutyrate) shunt; Lysine biosynthesis, 2-oxoglutarate => 2-aminoadipate => lysine; Melatonin biosynthesis, tryptophan => serotonin => melatonin; Ornithine biosynthesis, glutamate => ornithine; Urea cycle
3	5.00	5	10	Cholecalciferol biosynthesis; Phenylalanine biosynthesis, chorismate => phenylalanine; Tryptophan biosynthesis, chorismate => tryptophan; Tyrosine biosynthesis, chorismate => tyrosine; Uridine monophosphate biosynthesis, glutamine (+ PRPP) => UMP

Gut TIGRFAM	1	8.24	18	70	30S ribosomal protein S17; 50S ribosomal protein L3, bacterial; 50S ribosomal protein L4, bacterial/organelle; Preprotein translocase, SecY subunit; Ribosomal protein L14, bacterial/organelle; Ribosomal protein L15, bacterial/organelle; Ribosomal protein L16, bacterial/organelle; Ribosomal protein L18, bacterial type; Ribosomal protein L2, bacterial/organellar; Ribosomal protein L22, bacterial type; Ribosomal protein L24, bacterial/organelle; Ribosomal protein L29; Ribosomal protein L30, bacterial/organelle; Ribosomal protein L6, bacterial type; Ribosomal protein S10, bacterial/organelle; Ribosomal protein S19, bacterial/organelle; Ribosomal protein S3, bacterial type;

					Ribosomal protein S5, bacterial/organelle type
	2	8.00	8	28	50S ribosomal protein L14P; 50S ribosomal protein L30P, archaeal; Archaeal ribosomal protein S17P; Ribosomal protein L24p/L26e, archaeal/eukaryotic; Ribosomal protein L29; Ribosomal protein S3, eukaryotic/archaeal type; Ribosomal protein S5(archaeal type)/S2(eukaryote cytosolic type); Translation initation factor SUI1, putative, prokaryotic
	3	7.00	7	21	Methyl-coenzyme M reductase I operon protein C; Methyl-coenzyme M reductase operon protein D; Methyl-coenzyme M reductase, alpha subunit; Methyl-coenzyme M reductase, beta subunit; Methyl-coenzyme M reductase, gamma subunit; Tetrahydromethanopterin S-methyltransferase, subunit C;

					Tetrahydromethanopterin S-methyltransferase, subunit D
Gut Intersection	1	8.17	13	49	30S ribosomal protein S17; 50S ribosomal protein L3, bacterial; 50S ribosomal protein L4, bacterial/organelle; Ribosomal protein L14, bacterial/organelle; Ribosomal protein L16, bacterial/organelle; Ribosomal protein L18, bacterial type; Ribosomal protein L2, bacterial/organellar; Ribosomal protein L22, bacterial type; Ribosomal protein L24, bacterial/organelle; Ribosomal protein L29; Ribosomal protein L6, bacterial type; Ribosomal protein S19, bacterial/organelle; Ribosomal protein S3, bacterial type
	2	4.00	4	6	Preprotein translocase, SecY subunit; Ribosomal protein L15, bacterial/organelle; Ribosomal protein

					L30, bacterial/organelle; Ribosomal protein S5, bacterial/organelle type
	3	4.00	4	6	30S ribosomal protein S11; DNA-directed RNA polymerase, alpha subunit, bacterial and chloroplast-type; Ribosomal protein L17; Ribosomal protein S4, bacterial/organelle type
Gut Difference	1	7.71	8	27	50S ribosomal protein L14P; 50S ribosomal protein L30P, archaeal; Archaeal ribosomal protein S17P; Ribosomal protein L24p/L26e, archaeal/eukaryotic; Ribosomal protein L29; Ribosomal protein S3, eukaryotic/archaeal type; Ribosomal protein S5(archaeal type)/S2(eukaryote cytosolic type); Translation initation factor SUI1, putative, prokaryotic
	2	6.67	7	20	Methyl-coenzyme M reductase I operon protein C; Methyl-coenzyme M reductase operon protein D; Methyl-coenzyme M reductase, alpha subunit;

					Methyl-coenzyme M reductase, beta subunit; Methyl-coenzyme M reductase, gamma subunit; Tetrahydromethanopterin S-methyltransferase, subunit C; Tetrahydromethanopterin S-methyltransferase, subunit D
	3	4.00	4	6	50S ribosomal protein L4P; Archaeal ribosomal protein L23; Archaeal ribosomal protein L3; Ribosomal protein L22(archaeal)/L17(eukaryotic/archaeal)

Appendix D
Publications and Permissions

D.1 Publications Included in this Thesis

Chapter 2

Title: MetaProx: the database of metagenomic proximons

Authors: Gregory Vey, University of Waterloo; Trevor Charles, University of Waterloo

Journal: Database: The Journal of Biological Databases and Curation

Volume: 2014

DOI: 10.1093/database/bau097

Date of Submission: December 17th, 2013

Date of Acceptance: September 2nd, 2014

Date of Publication: October 6th, 2014

Permissions: Reproduced here under the Oxford Open model and the Creative Commons Attribution Non-Commercial License: *This license permits users to use, reproduce, disseminate or display the article provided that the author is attributed as the original creator and that the reuse is restricted to non-commercial purposes i.e. research or educational use.*

Chapter 3

Title: An analysis of the validity and utility of the proximon proposition

Author: Gregory Vey, University of Waterloo; Trevor Charles, University of Waterloo

Journal: Functional & Integrative Genomics

Volume: 16(2): 215-220

DOI: 10.1007/s10142-016-0478-z

Date of Submission: August 6th, 2015

Date of Acceptance: January 19th, 2016

Date of Publication: February 2nd, 2016 (online); March 1st, 2016 (print)

Permissions: Reproduced here in accordance with the Springer Copyright Transfer Agreement which states: "*The author retains the right to use his/her article for his/her further scientific career by including the final published journal article in other publications such as dissertations and postdoctoral qualifications provided acknowledgement is given to the original source of publication.*" The original publication is available through Springer International Publishing at: http://link.springer.com/article/10.1007/s10142-016-0478-z

Chapter 4

Title: Metagenomic annotation networks: construction and applications

Authors: Gregory Vey, University of Waterloo; Gabriel Moreno-Hagelsieb, Wilfrid Laurier University

Journal: PLoS ONE

Volume: 7(8): e41283

DOI: 10.1371/journal.pone.0041283

Date of Submission: March 1st, 2012

Date of Acceptance: June 19th, 2012

Date of Publication: August 7th, 2012

Permissions: Reproduced here under the PLoS ONE Open Access model and the Creative Commons Attribution License: *This license permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.*

D.2 Other Doctoral Publications

Title: Metagenomic guilt by association: an operonic perspective

Author: Gregory Vey, University of Waterloo

Journal: PLoS ONE

Volume: 8(8): e71484

DOI: 10.1371/journal.pone.0071484

Date of Submission: February 27th, 2013

Date of Acceptance: June 28th, 2013

Date of Publication: August 6th, 2013

Permissions: No portion of this publication is included or reproduced in this thesis. The original publication is available at www.plosone.org at http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0071484

Title: Gene coexpression as Hebbian learning in prokaryotic genomes

Author: Gregory Vey, University of Waterloo

Journal: Bulletin of Mathematical Biology

Volume: 75(12): 2431-2449

DOI: 10.1007/s11538-013-9900-z

Date of Submission: February 7th, 2013

Date of Acceptance: August 27th, 2013

Date of Publication: September 27th, 2013 (online); December 1st, 2013 (print)

Permissions: No portion of this publication is included or reproduced in this thesis. The original publication is available at www.springerlink.com at http://link.springer.com/article/10.1007/s11538-013-9900-z#page-1.