EVALUATING A MICROBIAL COMMUNITY THROUGH FEATURE MATCHING AND GRAPH TOPOLOGY

by

Michael Porter

Submitted in partial fulfilment of the requirements
for the degree of Master of Computer Science

at

Dalhousie University
Halifax, Nova Scotia
August 2012

DALHOUSIE UNIVERSITY

Faculty of Computer Science

The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled “EVALUATING A MICROBIAL COMMUNITY THROUGH FEATURE MATCHING AND GRAPH TOPOLOGY” by Michael Porter in partial fulfilment of the requirements for the degree of Master of Computer Science.

Dated: 16 August 2012

Supervisor: _____

Readers: _____

_____

DALHOUSIE UNIVERSITY

DATE: 16 August 2012

AUTHOR: Michael Porter

TITLE: EVALUATING A MICROBIAL COMMUNITY THROUGH FEATURE MATCHING AND GRAPH TOPOLOGY

DEPARTMENT OR SCHOOL: Faculty of Computer Science

DEGREE: MCSc CONVOCATION: May YEAR: 2013

_____

Signature of Author

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Metagenomics, the sequencing of DNA from environmental samples, has enabled the study of cohabiting microorganisms with a single sequencing experiment. This requires algorithms and techniques specific to metagenomics: Since the environmental sample is not separated by organism before being sequenced, taxonomic classification is required to reveal the taxonomic composition of the sample. The metabolic function of the sample can be determined through functional annotation. Both of these analyses can be done through comparisons to a reference database of sequences with assigned taxonomy and function. Here new techniques for metagenomic analysis are developed. The KB-1 metagenome, representing a microbial community capable of converting toxic chlorinated ethenes into non-toxic ethene, is used as an example for these techniques to determine which KB-1 organism is capable of dechlorination and what metabolic support this organism gets from other community members to sustain its growth.

A new rank-flexible taxonomic classification algorithm called SPANNER (Similarity Profile ANNotatER) is described. Traditional taxonomic classifiers are based on the similarity of a query sequence to sequences in a reference database. SPANNER uses all reference similarities as a feature vector of taxonomic affinities and classifies a query sequence based on affinity similarity. This approach is shown to be less sensitive to events such as lateral gene transfer which can confuse traditional classifiers. Classification using SPANNER is performed on the KB-1 metagenome. SPANNER offers greater control of the trade-off between precision and accuracy compared to other taxonomic classifiers; an appropriate level of precision can therefore be chosen based on the availability of closely related reference genomes. SPANNER classified many taxa at or within one or two ranks of the best possible rank.

Cohabiting microorganisms may interact metabolically via “hand-off points,” the sharing of processed chemicals between organisms. Hand-off points could give an organism access to an otherwise inaccessible biochemical pathway or could split pathways between organisms. This can lead to a community forming where some community members depend on others to provide key metabolites that are essential for survival. A metabolic network representing KB-1 metabolism is reconstructed using newly proposed methods. The topology of this network is analyzed for metabolic interaction and dependencies between microbial organisms. The reconstruction of community metabolism suggests metabolic regions that are complementary or redundant between community members, and hand-off point identification suggests possible dependencies between organisms. This network has topological differences from metabolic networks of single organisms.

Multiple events to the same genome that would normally confuse taxonomic classification, such as lateral gene transfer, create similar patterns of taxonomic affinity across that genome. SPANNER detects these patterns to avoid incorrect assignments from these events, for accurate KB-1 classification. The KB-1 metabolic network has high connectivity between metabolites caused by the complementarily of the metabolism of each community member. This network also identified several putative hand-off points between KB-1 community members, with accurate hand-off point detection being highly sensitive to missing or incorrect functional annotations.

LIST OF ABBREVIATIONS USED

16S	16 Svedberg ribosomal RNA subunit
AIC	Akaike-information criterion
BLAST	Basic Local Alignment Search Tool
DCE	dichloroethene
DNA	Deoxyribonucleic acid
GSC	Giant strong component
LGT	Lateral gene transfer
LCA	Lowest common ancestor
PCE	Perchlorinated ethene
PMK	Pyramid Match Kernel
SPANNER	Similarity Profile ANNotatER
TCE	trichloroethene
VC	vinyl chloride

GLOSSARY

Entries are underlined at their first occurrence in the main text. Cross references in the glossary are underlined.

annotate	The process of assigning to a protein sequence the chemical reaction it performs.
contig	Several short reads of sequenced DNA joined into a longer contig ("contiguous sequence") by a DNA assembly program.
clustering coefficient	The clustering coefficient *C* of a node is the proportion of connections between the node's neighbours.
degree	The degree *k* of a node is the number of edges connecting to it. The in-degree is the number of edges connecting into the node, the out-degree is the number of edges connecting away from the node. The degree equals the sum of the in- and out-degrees.
diameter	The diameter of a network is the longest of all shortest paths.
DNA	Deoxyribonucleic acid, a molecule found in all organisms. The DNA molecule is a polymer, made up of a linear sequence of nucleotide monomers. All genes in an organism are made of DNA.
gene	A subsequence of DNA in a genome. All proteins are encoded by genes.
genome	The entire set of DNA for an organism.
hand-off point	A metabolite (chemical) that is produced by one organism and then excreted into the environment. This metabolite is then acquired by a recipient organism and used as a nutrient. The recipient must not be able to make the metabolite itself.
lineage	The defined taxonomy for an organism, as a vector of labels at each descending rank.
metabolic network	A network representation of an organism's metabolism: all metabolites (chemicals) it can use to live and all reactions it can perform to process those metabolites.
metabolite	A chemical compound used by an organism for growth or sustainment. Proteins convert metabolites (called substrates) into other metabolites (called products).

metagenome	The combined genomes from an environmental sample. For example, all genomes from all organisms in a sample of ocean water form a metagenome.
neighbourhood	The neighbourhood of a node are the nodes connected to it by only one edge.
nucleotide	A molecule of adenine, guanine, cytosine, or thymine used as monomers in DNA.
path	A path from nodes *a* to *b* is series of connected edges and nodes starting at *a* and ending at *b*. The shortest path between two nodes is the path with the fewest edges.
pathway	A series of reactions related by a common biological function.
product	The metabolite (chemical) after a protein modifies it.
protein	A molecule produced by an organism to perform a specific chemical reaction.
taxonomic rank	A level in taxonomy (a hierarchical tree). Ranks at the top of the hierarchy differentiate organisms based on fundamental differences, the bottom ranks differentiate organisms based on lesser differences. The ranks in this thesis from top to bottom are: domain, phylum, class, order, family, genus, species.
rank-flexible	Any taxonomic classification algorithm that chooses the most appropriate rank to assign a sequence.
rank-specific	Any taxonomic classification algorithm that assigns all sequences at a predefined rank.
read	A string of {A, T, C, G} representing a DNA subsequence as produced by a DNA sequencer.
seed set	The metabolites in a network that cannot be produced and enable the production of all other metabolites.
sink	A node with an out-degree of zero.
source	A node with an in-degree of zero.
substrate	A metabolite (chemical) that a protein modifies.
taxon (pl. taxa)	A label at a specific taxonomic rank. This represents a known lineage from domain down to the taxon's rank and unknown taxonomy below that rank. For example, the taxon "Mammalia"

	is at the rank of class. Its lineage is "Eukarytoa, Chordata, Mammalia." A taxon distinguishes the set of organisms beneath it, for example "Mammalia" distinguishes all mammals. In taxonomic classification the assigned taxon is the maximum level of detail for a sequence. For example a sequence classified as "Mammalia" is identified as a mammal but which mammal is unknown.
taxonomic novelty	The rank at which an organism has no siblings. For example, *Homo sapiens* (humans) are novel at the rank of genus, being the only living species of their genus *Homo*.
taxonomy	The differentiation of organisms using different labels at different taxonomic ranks. These ranks and labels form a tree. Taxonomy relates these organisms by their placement on this tree, through their shared ranks. For example, taxonomy defines humans and crows as sharing the ranks domain and phylum (organisms with cellular compartmentalization, tails, and spinal cords) but they do not share the next rank of class. The class differentiates humans (mammals) from birds (feathered creatures who lay amniotic eggs).

ACKNOWLEDGEMENTS

I'd like to express many thanks to Rob Beiko. Several years ago you decided my Python skills were decent and offered me a short term GenGIS job; that decision has defined my career and given a relevance and meaning to computer science that I previously lacked. Thank you for the opportunity to do a master's degree, for your helpful insight and comments while writing this thesis, and for your patience while I try one idea after another. I'd also like to thank Christian Blouin and everyone in the Beiko and Blouin labs, for helping me understand the complexities of bioinformatics and biology. Denis Wong, for his guidance on various bioinformatics databases, and Donovan Parks, for his help with statistics and comparative analysis, made the past two years far more productive. I also want to thank my family who provided endless support and patience and my friends, namely Tara, Kathryn, and Lindy, who made sure I occasionally stopped working and had some fun.

CHAPTER 1 INTRODUCTION

Bioremediation is the process of cleaning polluted sites with microorganisms that are capable of converting pollutants into less harmful chemicals. *In situ* bioremediation involves adding the microorganisms directly to the polluted site where they metabolize the pollutant as part of their normal biological growth. An example toxin that has been cleaned by bioremediation is perchlorinated ethene (PCE), which is a common soil contaminant and is (along with its degraded form trichloroethene or TCE) carcinogenic and toxic to the liver and kidney [1]. KB-1 is a community of microorganisms capable of converting PCE into ethene through a series of biochemical reactions (Figure 1.1). KB-1 is sold commercially by SiREM Labs and has been used successfully for bioremediation of sites contaminated with PCE [2]. This bioremediation involves drilling holes into the soil upstream from the contamination and injecting the KB-1 microbial community along with fermentable organic compounds such as acids or alcohols that stimulate the community's growth; the downstream flow of the groundwater will carry KB-1 and the organic compounds into the contaminated soil. The alternative to *in-situ* bioremediation is excavation of the contaminated soil where it can be sent to a treatment plant; this process is often more expensive and time-consuming [3].

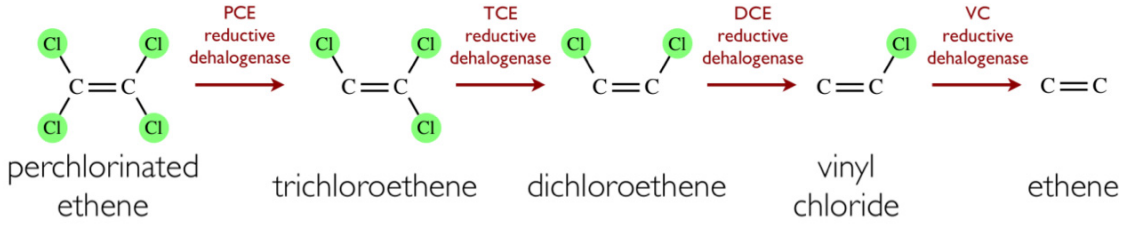


Figure 1.1 PCE degradation as performed by the KB-1 microbial community. Microorganisms in the community convert perchlorinated ethene into trichloroethene, which is then converted into dichloroethene, then into vinyl chloride, and finally vinyl chloride is converted into ethene. The chemical reactions that perform each conversion are named in red.

For bioremediation applications a common interest in the use of microbial communities is identifying which microbial species are responsible for the removal of the pollutant. It is possible that more than one microbe can metabolize the pollutant, or several microbes could be required to work together. Algorithms and analysis pipelines are actively being

developed to help investigate microbial communities: understanding the community composition (which types of microbes are in the community and in what abundance), the biochemical abilities of each community member, and the ecological interactions between them are three questions addressed in this thesis through the development of new algorithms and *in silico* analysis techniques. Answering these questions is necessary to understand KB-1 and its bioremediation potential. For example, KB-1 analysis has shown that the primary community members responsible for the degradation of PCE exhibit poor growth when isolated in a lab culture [4], this suggests those members rely on other community members to sustain it. If these requirements were better understood the KB-1 community might be made more efficient or cost effective for bioremediation.

1.1 Microbial Community Analysis

Microbial communities can be studied through DNA analysis to reveal a community's biochemical and ecological information. DNA, or deoxyribonucleic acid, is a molecule used to store information about the development and function of an organism. DNA is a polymer, a large molecule composed of many repetitions of smaller molecules called monomers. In DNA monomers are repeated in a linear sequence on two anti-parallel strands; these strands are twisted into a double helix (Figure 1.2). DNA monomers can be one of four molecules called nucleotides: adenine, guanine, cytosine, or thymine. The entire set of all DNA molecules within a cell is called a genome. The genome follows a hierarchical organization: a genome may consist of several DNA molecules called chromosomes or plasmids, within these separate polymer molecules are subsections of monomers (sequences of nucleotides) called genes. Figure 1.2 shows the four genes from Figure 1.1 on a hypothetical plasmid; genes can be located on either DNA strand. Genes can be thought of as the *functional encoding* of life: they are translated into proteins, which themselves are the *functional units* of life: the chemical reactions that drive cell metabolism are performed by proteins. For example, in the conversion of PCE into ethene shown in Figure 1.1, each arrow represents a chemical reaction performed by a different protein.

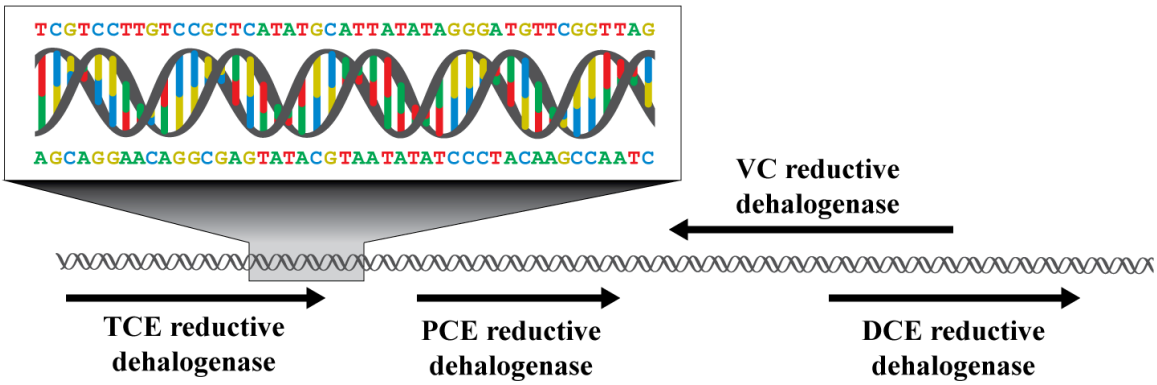


Figure 1.2 Gene sequences on a hypothetical genome. The small grey double helix spanning the width of the figure shows a section of the genome, this section contains the four genes that encode proteins that degrade PCE into ethene (Figure 1.1). These genes could be located at any point on the genome and on either strand, and until identified these gene sequences have an unknown length. The expanded box shows the nucleotides for each helix strand: adenine or "A" in green, guanine or "G" in yellow, cytosine or "C" in blue, and thymine or "T" in red.

DNA sequencing allows the "reading" of these sequences of nucleotides into a computer. Since DNA is composed of any of four nucleotides a DNA sequence can be represented as a string of letters from the alphabet {A G C T} (for adenine, guanine, cytosine, and thymine, respectively) as Figure 1.2 illustrates. This lends genomics, the study of genomes, particularly well to computational analysis. As sequencing technologies advance, the volume of DNA that can be sequenced at a given cost has increased exponentially [5]; this has motivated the development of new algorithms and techniques to study the ever-increasing amount sequenced DNA.

DNA sequencing creates a computer file containing short substrings of the genome called reads which are between (approximately) 40 and 1000 nucleotides in length. The short length of reads poses an analytical problem: Identifying what gene a sequence belongs to becomes harder as the length of the sequence decreases, since the odds of a sequence appearing in multiple places in a genome increases as the length of the sequence decreases. A single read might span only a small percentage of a gene's entire sequence making it difficult to identify what gene each read represents. Assembly is therefore done to increase the length of each read. DNA assembly is an algorithmic technique to join reads together, either using common overlapping regions found on two reads or using

longer reference DNA sequences from previous studies as an assembly template [6]. Assembly joins reads into contiguous sequences (contigs) whose length can be anywhere from hundreds to millions of nucleotides. The extent to which the entire genome is recovered by DNA sequencing depends on the coverage (how much of the DNA was sequenced) and depth (how many times any region of DNA was redundantly sequenced); insufficient coverage or depth means regions of the genome would have gone unsequenced or the assembly algorithm cannot join all reads, and the contigs will represent a subset of the genome. Recovering only a portion of a genome is common in sequencing, this will not hinder further analysis if the sequenced regions of the genome are the regions of analytical interest. Statistical techniques have been developed to predict how much DNA went unsequenced and how it will affect analysis [7]. Once a genome has been sequenced and assembled, the substrings that represent genes can be predicted. Not all of the DNA on a genome is in genes, and only a subset of the genes will encode proteins; protein-coding genes can range from 50 to thousands of nucleotides in length. Protein prediction typically uses models trained by machine learning that predict the location of protein-coding gene sequences on contigs and has an accuracy greater than 90% [8].

1.2 METAGENOMICS

Metagenomics is the study of a collection of genomes (called the metagenome) from an environmental sample [9]. An environmental sample (soil, ocean water, etc) would normally contain a range of different of organisms, in metagenomics these organisms are sequenced together and then analyzed. Metagenomic DNA sequencing differs from single-genome sequencing. A single genome is sequenced by extracting DNA from an organism and sequencing as described above, whereas in metagenomics a filter is used to isolate the organisms of interest (e.g. viruses or microbes) from the environmental sample, and DNA is extracted and sequenced without prior knowledge of the organisms present in the sample. The sequenced DNA is still assembled as described above, however since at no point is the DNA in the environmental sample separated by organism the generated reads will represent a collection of unknown organisms.

The KB-1 community was sampled from Kitchener, Ontario, Canada and sequenced in 2008 by the US Department of Energy Joint Genome Institute (Sample 10166: http://genomeportal.jgi-psf.org/aqukb/aqukb.download.html) and consists of 28506830 nucleotides. JGI assembled these reads using the software programs Lucy [10] and Paracel Genome Assembler into 24990 scaffolds (a scaffold is a series of contigs that are known to be in the correct order, albeit with gaps between them).

1.3 Homology

Once protein-coding gene sequences have been predicted from the sequenced metagenome DNA, comparative analysis can measure the similarity between them; sequence comparison forms the basis of the analysis in the next two sections.
When organisms procreate random mutations can occur in the DNA sequence of their genes, causing an evolutionary divergence of the child gene sequence compared to the parental sequence. This mutated sequence will share a common ancestor with its sibling genes, which may be exact copies of the parental sequence or be versions of the gene sequence with unique mutations. All versions of a gene that share a common ancestor are said to be homologous. As mutations accumulate over time the sequence of each homologous gene will diverge; homologous sequences that are more similar are assumed to have diverged more recently and are therefore more closely related. Measuring sequence similarity as a proxy for divergence requires both sequences to be homologous.

1.4 Functional Assignments

Since proteins are the functional units of the cell, performing chemical reactions that drive cell metabolism and growth, they are often an interest in a metagenomic study. Cells that have unique metabolic properties (such as the degradation of toxic chemicals like PCE into non-toxic ones) derive these properties from the proteins they encode. The chemicals a protein reacts with are called metabolites; the metabolite a protein modifies is called a substrate, and after modification it is called a product. Proteins typically perform only one reaction that converts one or more substrates into one or more products. The reactions performed by proteins can be chained together into a series of reactions,

each one using the previous reaction's product as a substrate for further conversion. This series (from a common source substrate through several reactions into a useful end product) is called a pathway. The four sequential reactions shown in Figure 1.1 represent the PCE degradation pathway, although some reactions in the pathway were omitted from the figure. The first reaction has PCE as a substrate and TCE as a product, the second reaction has TCE as a substrate so these reactions can be connected in series. Multiple pathways intersect to make a metabolic network. The metabolic network is the entire chemical capacity of a cell: all reactions it can perform on all substrates to produce all products. Figure 1.1 represents a small subset of an organism's complete metabolic network of hundreds to thousands of metabolites and reactions [11].

Studying the reactions and pathways in a sequenced metagenome starts by annotating protein sequences with chemical reaction information: assigning a function (the chemical reaction it performs) to all metagenomic protein sequences whose function is unknown. A protein can be annotated with a function in two ways: manual curation involves experimentally verifying the protein's function and is considered more reliable but unable to keep pace with the rate of data acquisition [12]; automated curation compares *in silico* a query sequence of unknown function to a reference database of sequences with assigned function. This comparison is based on sequence similarity or inferred using machine learning and attempts to assess the homology of the sequences. The query and reference sequences are assumed to be homologous if they are highly similar, in which case the query is assigned the function of the closest matching reference sequence. Several reference databases exist that annotate sequences with functional information:

- The Kyoto Encyclopaedia of Genes and Genomes (KEGG) is a database of protein reactions, metabolites, and organism-specific pathways [13]. The reactions in KEGG are either manually curated or automatically inferred. KEGG also maps the relationships between diseases, drugs, organisms, and these metabolic reactions through a hierarchical ontology called BRITE [14]. KEGG also has visualizations for pathways and metabolites.

- SEED is a manually curated database of protein functions with tools for the automated construction of metabolic network models from sequenced genomes [15].
- BiGG is a manually curated database of protein function [16]. BiGG also provides visualizations for organism-specific metabolic network models.
- UniProtKB encompasses two databases: TrEMBL, which is annotated using automatic inference, and Swiss-Prot, which is manually curated [17].
- Genbank NR is a large sequence database whose sequences are annotated using automated methods [18].

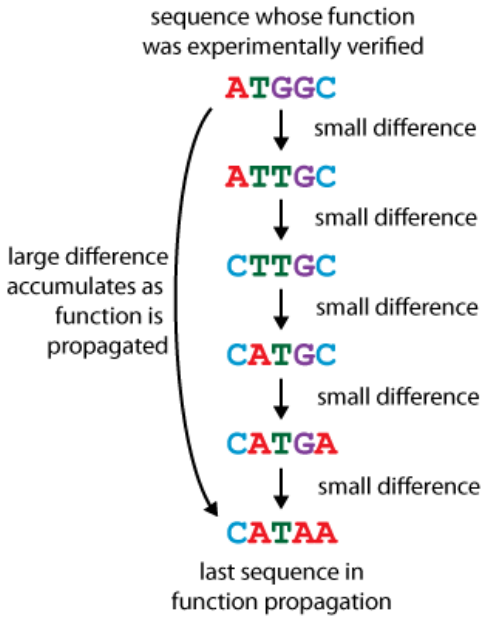


Figure 1.3 Functional propagation is the annotation of function by sequence similarity, across a series of sequences. Small differences between highly similar sequences can accumulate into a large difference between originally annotated sequence and the last sequence in the series.

Many of these databases contain cross-references to other databases, connecting information between them for the metabolites, reactions, etc. Inferring function *in silico* suffers from several limitations: The first is measuring similarity between two sequences. If two sequences are identical it stands to reason they have the same function, but there is no rule for how similar they must be before function can be inferred. Sequence similarity is often measured by aligning the sequences using the Smith-Waterman dynamic

programming algorithm [19]. Smith-Waterman compares two sequences monomer by monomer to calculate similarity, taking into account possible mutations (deviations in either sequence) that have accumulated over generations of evolution. Programs like BLAST [20] use heuristics to approximate Smith-Waterman to compare a single query sequence against a database of many reference sequences. Each comparison of query to reference results in a similarity score (called a "bitscore") and an expectation value or e-value. The e-value is a statistical property describing the number of matches with a bitscore as good or better that can be expected by chance given the size of the database used. (e.g. Doubling the size of the database will double the number of matches expected by chance.) Any measured similarity depends on the query sequence and reference database used and significant similarities could be artefacts of a poor reference database. The second issue is that the correlation between a protein sequence and a protein's function is limited. Highly similar sequences can be assumed homologous and therefore share a common ancestor. These similar sequences have not substantially diverged from this ancestor and both likely retain the ancestor's function. However a single nucleotide mutation can change the function [21], making a high bitscore and significant e-value misleading. This is in spite of the first problem: even with a high quality reference database a significant similarity does not indicate shared function. The third issue is function propagation (Figure 1.3), where the function of a newly sequenced protein in the database was inferred from another protein whose function was inferred from another and so on, until the start of the inference at the original experimentally verified protein. While the difference between any two proteins in function propagation may be small, the difference between the first experimentally verified protein and the protein last inferred in the propagation is large (as the differences accumulate along the chain of propagation). Schnoes *et al.* [12] showed that error rates ranged from 5%, 8%, 3%, 4% to 62%, 65%, 66%, 16% for Genbank NR, TrEMBL, KEGG, and Swiss-Prot respectively across different families of reactions, although they acknowledge they chose protein families that are especially difficult to annotate and that a trade off exists between highly confident annotations for few sequences versus less accurate annotations for many sequences. Schnoes *et al.* [12] suggest researchers choose the acceptable level of "annotation confidence" for their study and use the appropriate reference database. Apart

from annotation errors a subset of the predicted proteins will not be annotated, either because there is no reference sequence similar enough or because the machine learning models returned no result.

1.5 Taxonomic Composition

1.5.1 Motivation

Once functional assignments have been made to the sequenced metagenome proteins, the chemical capabilities of the microbes the metagenome represents can be studied and proteins of interest identified. Excluding proteins whose function could not be annotated, these reactions will represent the entire known chemical capacity of the metagenome (and therefore the microbes it represents). Since the metagenome represents a set of organisms from an environmental sample the annotated chemical functions will represent all organisms in that sample. To understand the role each organism plays the sequenced DNA contigs must be differentiable by organism: contigs need to be labeled with the organism the contig sequence originally came from. For example, functional assignment as outlined above can identify the PCE degradation pathway in KB-1 (Figure 1.1), but not the organism(s) capable of performing the pathway. Assigning a read or contig to an organism is called taxonomic assignment; performing assignment for the entire set of sequences can also be used to estimate the composition (who is present) and abundance (in what relative amounts) of a metagenome.

1.5.2 Taxonomy

Taxonomy is the categorization of life along a hierarchical tree, grouping organisms by physiological or genetic details. The taxonomy of an organism is its membership at each level, or rank, of the tree. Higher ranks separate organisms based on fundamental differences, lower ranks are based on lesser differences. This allows organisms to be described by a series of labels (one for each rank) identifying the categories of life it belongs to, from general to specific. As an example the taxonomy for humans, chimpanzees, common crows, jumping spiders, and the bacterium *Escherichia coli* are

shown in Table 1.1. There are several versions of taxonomy [22], each with different ranks and labels in each rank, so the taxonomic ranks in Table 1.1 are only the main ranks of one of the versions. Humans and chimpanzees share ranks from domain to family, but are assigned to different genera and species. Their lowest rank in common, the family Hominidae, separates the "great apes" (humans, chimps, gorillas, and orangutans) from the "lesser apes" (gibbons), which are largely similar animals but are smaller and do not build nests. All apes are members of the class Mammalia (mammals), sister to the class Aves (birds). At higher ranks, the difference between organisms increases: the lowest common rank between apes and crows is the phylum Chordata, which groups organisms with a tail extending behind the anus and a dorsal neural tube (which develops into a spinal cord). A sister phylum is Arthropoda, a group of organisms with an exoskeleton and segmented body parts, such as the jumping spider *P. audax*. At the highest rank, domain, the difference between organisms are fundamental and at the cellular level: *E. coli* shares no ranks with the others, belonging to the domain Bacteria (single-celled organisms with no internal cell compartmentalization) compared to the domain Eukaryota (organisms with cell compartmentalization). The taxonomy of an organism across all ranks forms a vector of labels, this is called a taxonomic lineage (e.g. the lineage for humans is Eukaryota, Chordata, Mammalia, Primates, Hominidae, *Homo*, *Homo sapiens*).

Taxonomic novelty is defined as the highest rank at which a genome has no known or characterized siblings. If a genome is novel at the rank of class, for example, then there are no other members of that genome's class, order, family, etc. although there could be other members of higher ranks (phylum or domain). In the lineages in Table 1.1 *E. coli* is novel at domain and Human is novel at genus. Novelty is always in relation to a set of known taxa: a genome could be novel at some rank with respect to all known organisms, or it could be novel in relation to a specific sequence database. Members of the same genus are termed "congeners" and "conspecific" organisms are members of the same species.

Table 1.1 Taxonomic lineages for human, chimpanzee, common crow, jumping spider, and *E. coli*. Only the common taxonomic ranks of Species, Genus, Family, Order, Class, Phylum, and Domain are shown.

		Human	**Chimpanzee**	**Crow**	**Spider**	***E. coli***
Taxonomic rank	**Domain**	Eukaryota				Bacteria
	Phylum	Chordata			Arthropoda	Proteobacteria
	Class	Mammalia		Aves	Arachnida	Gammaproteobacteria
	Order	Primates		Passeriformes	Araneae	Enterobacteriales
	Family	Hominidae		Corvidae	Salticidae	Enterobacteriaceae
	Genus	*Homo*	*Pan*	*Corvus*	*Phidippus*	*Escherichia*
	Species	*H. sapiens*	*P. troglodytes*	*C. brachyrhynchos*	*P. audax*	*E. coli*

1.5.3 Defining and Determining Microbial Taxonomy

The taxonomy of microorganisms is unique, and deserves special attention since the analysis in this thesis is applied to microbes. The "biological species concept" defines a species as individuals in populations that can potentially interbreed [23]. While interbreeding is helpful to distinguish animals it cannot delineate microbial species (exceptions also exist for animals). Microbes reproduce both asexually (one parent reproducing an exact duplicate offspring) and via genetic recombination (incorporating the DNA from another organism into its own genome), at various ratios of the two for different microbes [24]. Morphological characteristics, while useful for animal taxonomy (e.g. separating winged creatures from hoofed creatures), are difficult to determine for microorganisms given their small size. While microbial cell shape (rod-shaped, circular, etc) is visible under a microscope, this cannot differentiate all microbial species and other characteristics are too difficult or expensive to detect. DNA sequencing allows a genetic basis for microbial taxonomy; for a time organisms with genome DNA-DNA hybridization greater than 70% were considered the same species [24]. Hybridization is a molecular technique that measures sequence relatedness by its chemical binding affinity; more related sequences have stronger bonds. A taxonomy that relies on the entire genome is also problematic due to Lateral Gene Transfer (LGT), the sharing of DNA between microbes outside of sexual or asexual reproduction. In LGT sections of DNA are released by one organism and incorporated into another organism's genome. This can occur between distantly related organisms making their dissimilar genomes appear more similar. The wide-spread occurrence of LGT makes it a substantial barrier to taxonomic classification by sequence similarity [25]. Some have proposed that microbial species therefore have fuzzy boundaries between overlapping clusters of similar genomes [26; 24].

A newer technique uses a single gene or set of homologous genes present in all members of a group (in this case microbes) as the standard for measuring sequence similarity to define taxonomy. The most common microbial gene used, 16S ("16S ribosomal RNA subunit"), is used in the creation of proteins making it vital to life. 16S has both slow-

evolving and fast-evolving regions, fast-evolving regions accumulate changes to the nucleotide sequence during reproduction fast enough to differentiate species, but slow enough that within any one species there is little genetic variation. Slower-evolving 16S regions can be used to differentiate higher taxonomic ranks. The 16S gene can be extracted and sequenced to determine ancestral relatedness and taxonomy; this differs from metagenomics where all DNA is sequenced regardless of what gene it encodes. LGT of 16S between distant organisms would make them appear as related species and distort the resulting taxonomy, however 16S is less likely to be transferred by LGT than other types of genes [27]. Since 16S is commonly used to identify microbial species, several 16S sequences databases exist and tools are available specifically for analyzing 16S to determine taxonomy [28; 29]. Currently 16S similarity < 97% defines a new species [30].

Table 1.2 The 13 taxa in KB-1 identified using 16S sequencing, their relative abundance, and the taxonomic rank each taxon was identified at.

Taxon in KB-1	Abundance	Rank
Dehalococcoides	56.60%	Genus
Geobacter	7.55%	Genus
Methanomethylovorans	1.33%	Genus
Methanomicrobiales	5.29%	Order
Methanosarcina	1.00%	Genus
Methanosaeta	1.00%	Genus
Sporomusa	4.33%	Genus
Acetobacterium	11.31%	Genus
Spirochaeta SA-8	1.99%	Genus
Spirochaeta SA-8 2	1.00%	Genus
Syntrophus	1.00%	Genus
Chlorobi SJA-28	4.74%	Phylum
OP5	1.00%	Phylum *(novel)*

To determine taxonomic composition 16S was sequenced from KB-1 [31], this 16S profiling suggests there are 13 members in the microbial community. These are referred to as the "13 expected KB-1 taxa," which are shown in Table 1.2, along with their predicted abundance in the metagenome and the taxonomic rank of the 16S assignment. Although 16S can differentiate species, this is limited by the database of reference sequences. Sequences that are similar but not identical to a species' 16S reference sequence indicate the query sequence is from a higher rank such as genus or family. Hence the expected KB-1 taxa are ranks higher than species. OP5 is a novel phylum: no other genome from the same phylum is known, meaning reference genomes can only be related to OP5 at the rank of domain [32].

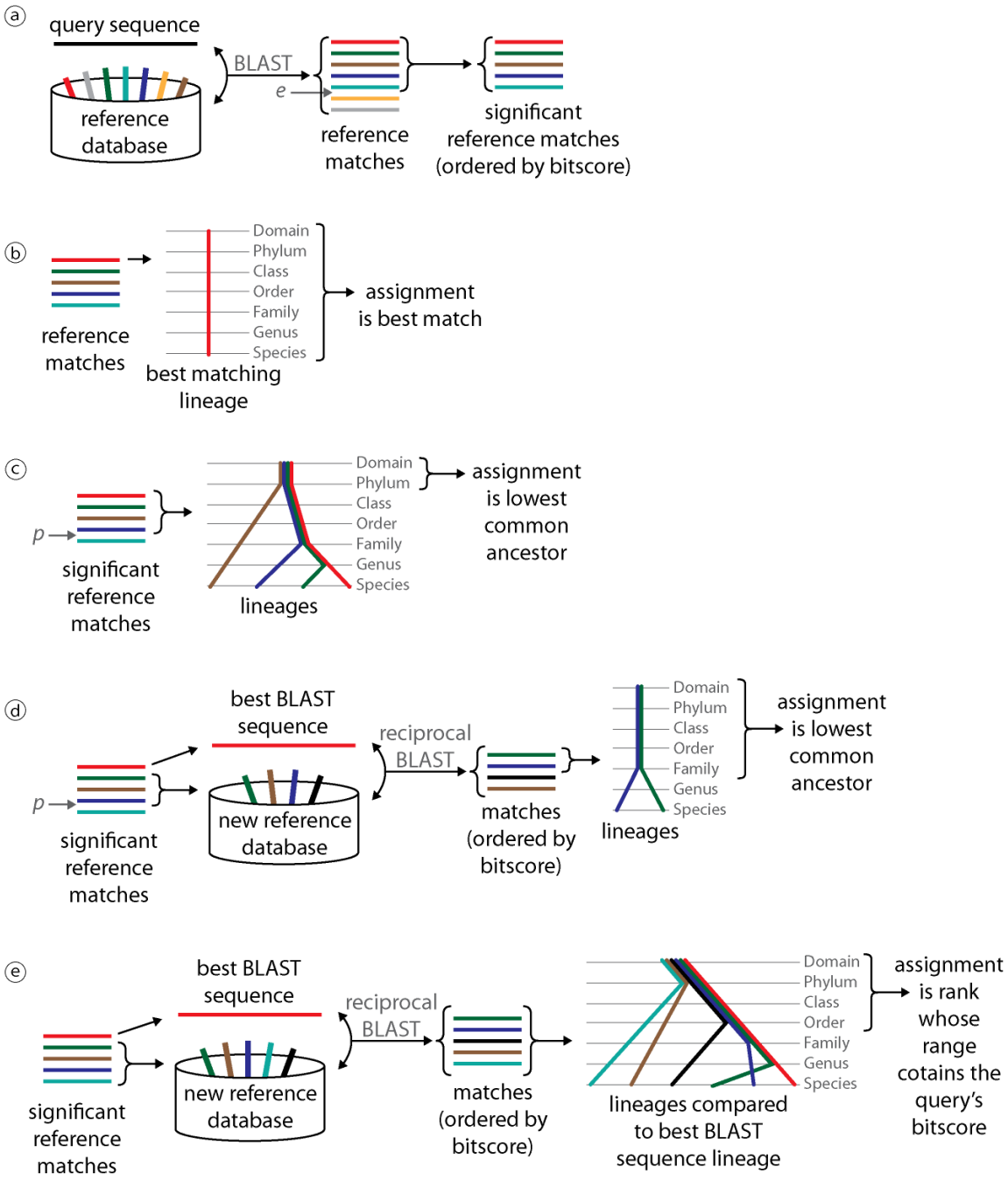
1.5.4 Assigning Taxonomic Information to Metagenome Sequences

In metagenomic studies the DNA is not separated by organism so the taxonomic composition of the metagenome is unknown. While 16S genes can most likely be found in the assembled contigs it is not guaranteed that every organism's 16S gene was sequenced. Furthermore the 16S gene will only be on some contigs; the remaining contigs cannot rely on 16S to determine its taxonomy. For this reason taxonomic assignments are made to all sequences (contigs and unassembled reads) in a metagenome without using 16S, although the 16S profiling in Table 1.2 provides an *a priori* understanding of what taxa the sequences should be assigned to and in what proportions. Classification of sequences without using 16S is a focus of research since many metagenomic studies attempt to describe the composition and taxonomic assignments of all reads and contigs are needed to study how the species might interact with each other in the community; this includes modeling the metabolic networks of microbial communities. This makes accurate taxonomic assignments of reads and contigs critically important to community-wide metabolic analysis [9].

Taxonomic classification exists in two categories, based on the type of learning or assignment procedure: unsupervised and supervised. Unsupervised classification is the clustering of sequences related by some measure into groups, without making

assignments to microbial taxonomy. These groups may represent more than one species or only part of a species depending on the clusters formed. Supervised classification is the comparison of unknown query sequences to a reference database of sequences with assigned taxonomy, which is the focus of Chapter 2. Two forms of supervised classification exist: rank-specific and rank-flexible [33]. In rank-specific classification the taxonomy of the closest matching sequence in the reference database becomes the taxonomy of the query sequence, from domain down to the rank being classified (i.e. all classifications will be at the same rank, e.g. species). In rank-flexible classification the reference sequence comparisons determine the best taxonomic rank to classify the query (i.e. deciding there is insufficient knowledge to classify a sequence at species but enough knowledge to classify at a higher rank such as family and leaving the taxonomy "unknown" at ranks below family). A sequence is classified as "unknown" at all ranks if it cannot even be classified at the rank of domain. Many classifiers add a rank above domain with the label "cellular organisms" to represent these unknown sequences.

Figure 1.4 Reference-based taxonomic classification. **A**: Reference-based taxonomic classification of a query sequence (black) starts with a comparison against a reference database whose sequences have an assigned taxonomy (red, grey, green, cyan, blue, yellow, and brown represent different taxonomic lineages). An example comparison algorithm is BLAST; poor BLAST matches are filtered at an e-value threshold. **B**: Best BLAST classification uses the best matching lineage from the BLAST comparison to assign the query sequence. **C**: LCA takes the lowest common ancestor of the top BLAST matches (filtered by *p*). **D**: SOrt-ITEMS filters matches at *p*, then performs another BLAST comparison of the best match (red) against the remaining matches (green, brown, blue) and the original query (black). The LCA of all matches better than the query is used to assign the query sequence. **E**: CARMA3 performs a reciprocal blast similar to SOrt-ITEMS, but does not filter at *p*. A bitscore range is defined for each rank: the lowest common rank for each matching reference lineage (green, blue, brown, cyan) to the reciprocal query (red) has a range from the lowest bitscore at that rank to the highest. For example, the LCA of both cyan and brown to red is the rank Phylum, so Phylum has a bitscore range from min(bitscore for cyan, bitscore for brown) to max (bitscore for cyan, bitscore for brown). Another example, the rank Genus has a bitscore range equal to the bitscore for green. The query sequence (black) is assigned to whichever rank has a range that encompasses the query's bitscore.



At the heart of all taxonomic classification is the method of comparison between any two sequences. Two widely used classes of comparison techniques are alignment and composition. As two new species arise and diverge from their evolutionary ancestor, mutations to their homologous DNA sequences will accumulate. Highly similar sequences are assumed to be homologous and greater similarity implies less time since the sequences diverged (and therefore greater taxonomic relatedness between the

sequences). Alignment comparisons measure the dissimilarity (i.e. evolutionary distance) between two sequences while taking into account sequence mutations, using an algorithm such as Smith-Waterman. Compositional comparison uses the frequency distribution of substrings (called k-mers) in a sequence, which can provide a characteristic "signature" of a particular genome or taxonomic group [34]. Compositional classifiers have had more difficulty with shorter sequences, although recent classifiers can accurately classify sequences as short as 25 nucleotides [35]. Alignment classifiers are more accurate than composition classifiers, and hybrid classifiers have been built that use both alignment and composition together to make assignments that are more accurate than either type of classifier used in isolation [36].

To avoid the over-specific problem of rank-specific classifiers, rank-flexible classifiers estimate the level of novelty of a query sequence. The Lowest Common Ancestor algorithm (LCA, Figure 1.4c) is an extension to best BLAST [37]. Instead of taking the top match, the lowest common ancestor (the lowest or most precise shared taxonomic rank) of the BLAST matches is used as the assignment. For example if two BLAST matches share taxonomic ranks from domain to family and then diverge, the query sequence will only be assigned the taxonomy common to all matches (in this case from domain to family) and will be unspecified at lower ranks. The parameter p ($0 \leq p \leq 1$) defines how conservative this approach is: only BLAST matches whose bitscore is greater than the best match's bitscore $\times$ p are used when taking the lowest common ancestor. This way p prevents the weakest homology matches from influencing the final taxonomic assignment and at very high values of p LCA approaches best BLAST (by removing all matches but the best). This rank-flexible approach avoids the over-specific problem of best BLAST but suffers from an under-specific problem: events like LGT between two distantly related organisms (brown and the red/green/blue lineage in Figure 1.4c) will appear as strong homology matches, pushing the LCA assignment to the least common ancestor of the distantly related organisms, which would be at higher ranks such as domain or "cellular organisms."

LCA has been used to classify many metagenomes since its development in 2007, however since it suffers from the under-specific problem described above, several extensions to LCA have been proposed. Monzoorul et al. [38] developed SOrt-ITEMS (Figure 1.4d), which like LCA uses *p* to trim low-scoring BLAST matches. Instead of taking the lowest common ancestor of the remaining matches, a reciprocal BLAST comparison is done using the best BLAST match as the query against a new reference database of the original query sequence and all matches greater than *p*. Using the coloured example in Figure 1.4d, if a query sequence Q is compared to a reference database of sequences $\{S_{red}, S_{grey}, S_{green}, S_{cyan}, S_{blue}, S_{yellow}, S_{brown}\}$ and matches to S_{red}, S_{green}, S_{brown}, S_{blue} are not filtered by the e-value threshold and are above the *p* threshold (in descending order: S_{red} being the most similar and S_{blue} being the least), the reciprocal BLAST would compare S_{red} to a reference database of $\{S_{green}, S_{brown}, S_{blue}, Q\}$. The LCA of all reciprocal matches whose bitscore is better than the match to the original query is used in the assignment: in the example the reciprocal matches are S_{green}, S_{blue}, Q, S_{brown} (in descending order) and the lowest common ancestor of S_{green} and S_{blue} will be assigned to Q, since they have more in common with the Q/S_{red} match than with S_{brown}. SOrt-ITEMS includes additional details not covered here, such as what to do if there is no reciprocal match with a bitscore greater than that of the query. SOrt-ITEMS is an extension to LCA and still uses the lowest common ancestor for the assignment, but attempts to reduce the number of lineages that define the lowest common ancestor by filtering some lineages via a reciprocal BLAST. This reduction means SOrt-ITEMS is less likely to include distantly related taxa when taking the lowest common ancestor and the assigned rank should be more precise than that of LCA. SOrt-ITEMS shows an increase in specificity and reduction in false positives compared to LCA.

CARMA3 [39] (Figure 1.4e) is similar to SOrt-ITEMS in that it uses a reciprocal BLAST of the best BLAST match against a new database which consists of the original query sequence and the matched reference sequences, except in CARMA3 all reference matches (except the best) are added to the new database (they are not trimmed to *p*). All sequences in the reciprocal database are then mapped to the taxonomy of the reciprocal query sequence, as shown on the right of Figure 1.4e. For each match in the reciprocal

BLAST except the original query sequence, the rank of the LCA against the reciprocal query is determined. At each rank with one or more LCAs, the minimum and maximum bitscores of the matches defines the bitscore range for that rank. Ranges for ranks with no assigned LCA are determined using linear interpolation from ranges with ranks. Each bitscore range then defines the minimum and maximum bitscore for assignment to that rank: if an unknown sequence matches the reciprocal query with a certain bitscore, that sequence is assigned the reciprocal query's taxonomy from domain down to the rank whose range includes that bitscore (in Figure 1.4e, the query sequence matches the reciprocal query S_{red} with a bitscore that falls in the range for the rank order, this range was defined by linear interpolation between the S_{brown}/S_{cyan} match's bitscores at phylum and the S_{blue} match's bitscore at family). CARMA3 includes additional details, such as what to do if linear interpolation is not possible or if ranges overlap, which are not covered here. CARMA3 outperforms SOrt-ITEMS and LCA, having fewer false positives and increased specificity (queries are classified at lower taxonomic ranks). SOrt-ITEMS filters out reciprocal matches worse than the query, whereas CARMA3 uses information from all reciprocal matches to make an assignment. CARMA3 also does not rely on the lowest common ancestor (where distantly related sequences will decrease specificity by pushing the classification up to higher ranks), instead increasing precision by using bitscore ranges to find an appropriate rank to classify to.

These rank-flexible algorithms for taxonomic classification attempt to assign the lowest rank possible without exceeding the estimated taxonomic novelty of the query sequence. Estimating the novelty for a query sequence is essential to avoid the over-specific assignment problem of rank-specific classification using best BLAST.

1.6 Reconstructing Metabolic Networks

With functional annotation and taxonomic assignments the PCE degradation pathway in Figure 1.1 can be identified in KB-1 and the organism(s) capable of performing the pathway known, but understanding its place in KB-1 metabolism requires a broader view of the metabolic networks of the KB-1 taxa. *In silico* reconstruction of a metabolic network attempts to model the chemical capacity of a cell: all reactions that can be

produced from its genome and all metabolites a cell can use during its life. Four of these reactions and five of these metabolites could be those shown in Figure 1.1; the reconstructed metabolic network should show what other reactions and metabolites they connect to revealing how the PCE degradation pathway fits into the organism's metabolism. Since the metabolic network is reconstructed as a graph of chemical reactions and metabolites, the mathematics of graph theory can be applied to the network [40] enabling computational analysis and modeling of biochemical function and dependencies among pathways.

1.6.1 Network Representation

The metabolic networks developed here are *metabolite-centric* [41], with metabolites represented by nodes and reactions by edges, as seen in Figure 1.1. The network is reconstructed as described in section 1.3 by connecting the products of reactions as the substrates of others, forming a graph. The edges in a network can be directed or undirected. Since protein reactions have a direction (from substrates to products), directed edges are commonly used. Many protein reactions are bidirectional and can also use the products as substrates, performing the reaction in reverse. Bidirectional reactions can be represented as two anti-parallel directed edges.

Reactions often convert more than one substrate into more than one product. Commonly a separate directed edge is used for each substrate-product pair in the reaction, over-representing one edge as $s \times p$ edges, where s is the number of substrates and p is the number of products. The additional substrates and products in a reaction are often "currency metabolites," these are cofactors used in many reactions (such as water or H^+) [42], so that their inclusion in a metabolic network can skew analysis [43]. Currency metabolites under-represent the number of steps needed to convert substrates into products. For example, if ten reactions convert a substrate into a product but the first reaction also produces a currency metabolite that is consumed in the tenth, then the network will connect the substrate and product with only two edges (substrate →

currency metabolite → product) instead of ten. Removing currency metabolites helps correct the over-representation of edges.

Hypergraphs have been used [44] for a more accurate representation of a network: multiple nodes can be represented by a single hypergraph edge, for example a single edge could simultaneously connect two substrates with three products, which would be six separate edges otherwise. The adaptation of network analysis algorithms for use with hypergraphs is ongoing [42]. If a hypergraph is not used the metabolic network is also naturally a multigraph, a graph which allows multiple parallel edges ("multi-edges") between nodes. Multigraphs can be directed or undirected. Multiple proteins performing the same reaction or will create multi-edges between a pair of nodes. Multi-edges are created under even less strict criteria: since an edge is included in the network for every $s \times p$ substrate-product pair in a reaction, proteins performing different reactions will create multi-edges if a subset of the substrates and a subset of the products are shared between the reactions. The network analysis described in the next section applies to directed multigraphs.

1.6.2 Analyzing Networks

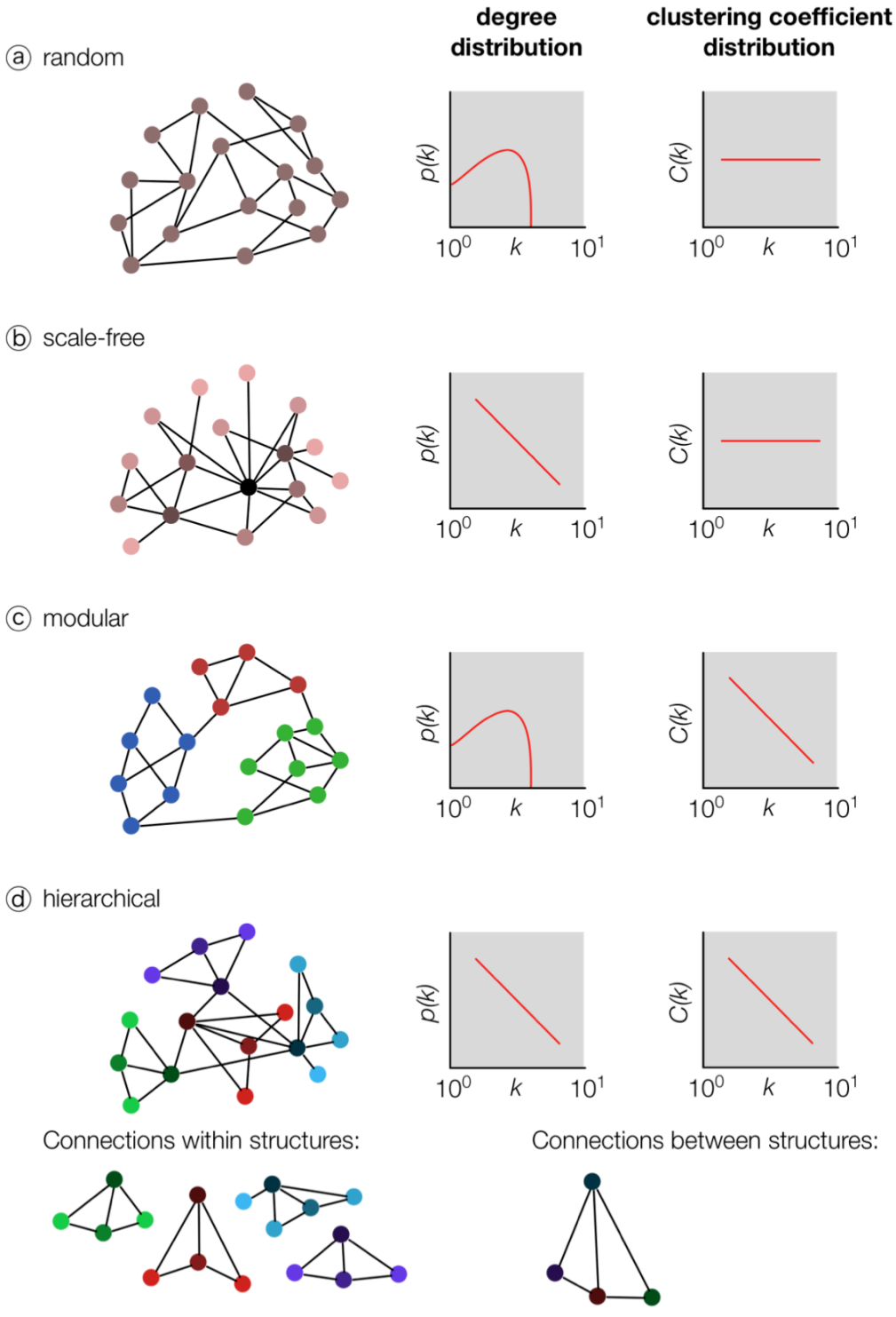
A network $G=(V,E)$ contains a set of nodes (V) and edges (E). The notation e_{abn} indicates the n^{th} edge in a directed multi-edge from node v_a to v_b. The directed multi-edge e_{ab} is the set of all edges it contains. A path between nodes v_a and v_b is a series of edges and nodes starting with v_a and ending at v_b; the shortest path between these nodes is the path with the fewest edges. The diameter of a network is the longest of all its shortest paths and represents the maximal distance to convert one metabolite into another. The degree k_i of a node v_i is the number of edges connecting to it and is the sum of its in-degree (the number of edges connecting into it) and its out-degree (the number of edges connecting away from it). The degree distribution $p(k)$ of a network is the frequency of nodes with degree k=0,1,2,…. A node v_i has a neighbourhood N_i which is the set of nodes connected to v_i by one edge. The clustering coefficient of a node v_i is defined by Equation 1, where

e_{jkn} are all edges connecting nodes in the neighbourhood N_i and $|N_i|$ is the size of the neighbourhood of v_i.

$$C_i = \frac{\left|\left\{e_{jkn} \in E : v_j \in N_i, v_k \in N_i\right\}\right|}{|N_i|\left(|N_i| - 1\right)} \tag{1}$$

The clustering coefficient measures the ratio between the number of edges connecting nodes in the neighbourhood of v_i to each other (numerator) and the maximum possible number of edges the nodes in the neighbourhood could have (denominator). This measures how connected the neighbourhood is, from 0 (no neighbours are connected) to 1 (all neighbours are maximally connected). The clustering coefficient for the entire network is the average of the clustering coefficient for all nodes. The clustering coefficient distribution $C(k)$ is the frequency of nodes with degree k=0,1,2,…, expressed as the average clustering coefficient for each k. The clustering coefficient of nodes with less than two neighbours is zero.

Figure 1.5 Four network topologies. The number of nodes and edges is the same is each network. For visual clarity, the position of each node is the same as well. The expected degree and clustering coefficient distributions are shown on log-log plots. **A:** A network with randomly placed edges. **B:** A scale-free network. Darker nodes have a higher degree (hubs), lighter nodes have a lesser degree. Nodes are more likely to be attached to nodes of a similar degree (meaning hubs tend to attach to other hubs). **C:** A modular network. Three modules shown in blue, red, and green have a lesser number of connections between them than within them. **D:** A hierarchical network. Four structures are shown in green, purple, red, and cyan; darker nodes have a higher degree. The same connections (with minor variations) are found within each structure. Those connections are repeated between structures to form the next level of the hierarchy. This figure is adapted from [40] and [45].


degree distribution
clustering coefficient distribution
ⓐ random
ⓑ scale-free
ⓒ modular
ⓓ hierarchical
p(k)
C(k)
10^0
k
10^1
Connections within structures:
Connections between structures:

Metabolic networks have been suggested to be scale-free (Figure 1.5b) [43], a network topology that is characterized by short average path lengths and a high clustering coefficient; scale-free topologies are also seen in social networks [46] and the World Wide Web [47]. Scale-free networks can be identified by their degree distribution following a power law with a decreasing slope. A distribution fits a power law if it appears as a straight line on a log-log plot. Scale-free networks have a small number of hub nodes, which are node with a statistically significantly high degree. In scale-free networks these hubs tend to connect to other hubs. Scale-free networks are said to be robust to random perturbations since removing a node from the network at random is more likely to remove a node with a smaller degree. If a hub is removed, the lesser-degree hubs surrounding it will remain connected. Jeong *et al.* [48] found that as many as 60 randomly chosen substrates could be removed (representing mutations to the proteins that use them) from the metabolic networks of 43 different organisms with little effect on network diameter.

A network topology can be described by comparing it to a random network (Figure 1.5b). A network having the same number of nodes and edges but with edges placed at random will have a degree distribution that follows the Poisson distribution, indicating nodes tend to have the similar numbers of links [40]. The clustering coefficient for each node in a random network is independent of its degree so its distribution will remain flat across all degrees, a property shared with scale-free networks. In contrast to this, previous studies of metabolic networks have shown that the clustering coefficient decreases as the degree increases, a characteristic of a modular topology [45]. A network with a modular topology is composed of modules that can be definitively and discretely partitioned from the network (Figure 1.5c) using an algorithm that identifies clusters of high connectivity. In metabolic networks these modules would perform separate biological functions. Discrete modules cannot exist if the high-degree hubs in scale-free networks connect uniformly to metabolites from every module; this would cause all modules to become interconnected. To reconcile this Ravasz *et al.* [45] proposed a hierarchical topology (Figure 1.5d) that does not consist of discrete modules but instead contains several small structures with hub nodes. The connections within each structure are similar, and the next

level of the hierarchy is formed by a repetition of these connections between structures. Hierarchical topologies can be identified by both a degree distribution and clustering coefficient distribution following a power law with a decreasing slope; these networks are also scale-free.

Evolutionary models for metabolic networks that generate the observed topology have been proposed. Scale-free networks (whose degree distribution follows a power law) can be generated by "preferential attachment" process in which new nodes are more likely to be connected to nodes with a higher degree [49]. The Big Bang model of network evolution [50] assumes that proteins evolve from one or a few categories of function. In this model all pairs of edges are assigned distance values indicating their sequence similarity. At each time step, the model duplicates an edge and chooses a random number from the interval (0,1) to mutate it. If the mutation is less than a threshold w the new edge retains its functional category and is connected to one of the metabolites of the old edge, otherwise it is assumed to belong to a new functional category and is placed randomly in the network. At the end of each time step sequence mutation is represented by increasing all distances between edges. Any distance now exceeding w implies the proteins have diverged into different categories and the edge is relocated to a random place in the network. Przytycka and Yu [51] expanded this model to also include sequence mutation that can potentially decrease the distance between edges; this model uses preferential attachment which causes a power law degree distribution. While recent studies continue to show metabolic networks are scale-free and hierarchical, several argue the degree distribution does not follow a power law and is best fit by other models. Przytycka and Yu [51] showed that the degree distribution of the scale-free Big Bang model of [50] better fits a Yule distribution (a linear preferential attachment distribution). Stumpf *et al.* [52] tested the degree distributions of metabolic networks from KEGG (the same data as [43]) using log-likelihoods and a goodness of fit test against the power law, log-normal, stretched exponential, and gamma distributions; none were found to match the power law as [43] claimed.

The hierarchical structure of networks has also been a focus of research. Graph clustering is the detection of modules of high connectivity ("communities") within the network. Newman and Girvan [53] proposed an algorithm to find community structures in networks using shortest path edge betweenness, where at each iteration the shortest path between all pairs of nodes is calculated and the edge involved in the greatest number of shortest paths is removed. This algorithm has been implemented in a software package for analyzing metabolic networks [54]. Holme *et al.* [55] applied the edge betweenness algorithm to 43 metabolic networks and found a hierarchy of community structures which themselves are composed of smaller community structures. These structures were related to biological function. Ma *et al.* [56] also found substructures within a metabolic network of *E. coli* related to common biological function.

In graph theory, a connected component is a set of nodes all of which are connected to each other by a path. Networks can have multiple connected components, each disconnected from the rest. A strongly connected component is a set of nodes where a path exists from each node to every other node in the set [43]. Any metabolite can be converted into any other metabolite within the strongly connected component. Substrates a network uses that it cannot produce itself are termed sources, these metabolites must be obtained from the environment (although it is worth noting that metabolites a cell can produce internally (non-sources) might also be available in the environment) (Figure 1.6). Products a cell produces that are not used as substrates for other reactions in the cell are termed sinks. Since sinks are not used for further reactions they are often the constituent chemicals that make up the cell itself ("biomass"), used to expand the cell during growth or reproduction. Sinks can also be waste chemicals a cell discards into the environment.

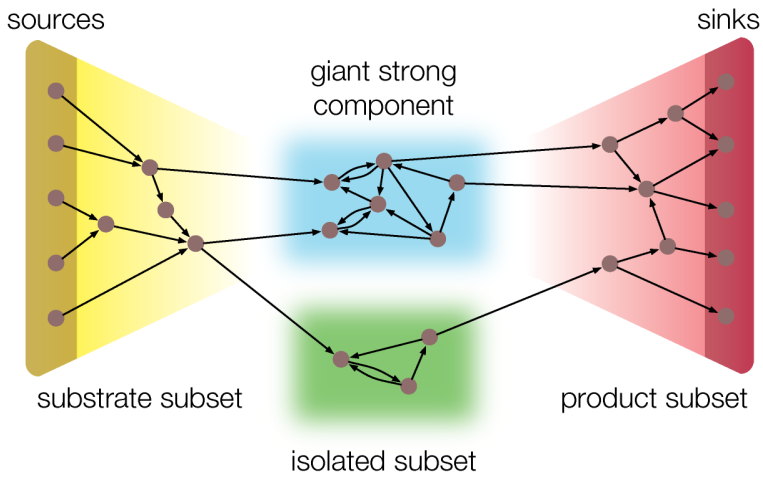


Figure 1.6 The "bow tie" topology structure. The substrate subset consists of sources and carrier metabolites, and supplies the giant strong component and isolated subset with substrates. The product subset consists of carrier metabolites and sinks. This figure is adapted from [56].

Metabolic networks have been described as a having a "bow tie" structure to their topology [43] (Figure 1.6), which is composed of four sets: the "giant strong component" (GSC), a strongly connected component consisting of the core metabolism needed for life, a "substrate subset" of metabolites supplying the network, a "product subset" of metabolites exiting the network, and an "isolated subset" of nodes that do not interact with the GSC. The bow tie analogy is derived from the large number of sources that are available from the environment and are converted by carrier metabolites into a smaller set of products passed to the GSC for use. The GSC itself represents around 30% of the network and performs metabolic functions vital for growth and sustainment [43]. A likewise smaller set of products in the GSC are then converted into a large number of sinks for the products subset, these are biomass metabolites or expelled as waste.

1.7 Identifying Microbial Community Interactions

Microbial communities are defined as "multi-species assemblages [of microbes], in which organisms live together in a contiguous environment and interact with each other" [57]. The organisms contained in an environmental sample and represented by a sequenced metagenome do not necessarily interact. Some organisms could be transient in the environment and not part of a stable community. An environment could also contain

multiple communities, with interactions within but not between them. Microbial community interactions can be metabolic: some organisms in microbial communities cannot survive on their own and rely on other community members to sustain it, perhaps through the removal of an environmental toxin that would otherwise negatively affect their metabolism [58] or by receiving metabolites needed for survival from other members [59].

A "hand-off point" is defined as the exchange of a produced metabolite from one taxon to another that could not otherwise produce it. A hand-off point could create variants of pathways: routing some of the pathway's reactions through the metabolic network of another organism creates new and possibly more efficient versions of the pathway. Hand-off points could also make available new metabolites for other organisms, passed off from organisms that don't need the metabolite, don't need as much of the metabolite as it produces, or who make the metabolite exclusively for handing off to support the growth of another organism. Communities exist where some microorganisms support the life of others who would otherwise die. One example is the glassy-winged sharpshooter (*Homalodisca coagulata*), an insect which hosts two bacterial species: *Baumannia* and *Sulcia* [59]. The sharpshooter provides nutrients to the bacteria who in turn synthesize vitamins and amino acids and provide them to the sharpshooter host; all three organisms depend on the others for survival. The set of potential hand-off metabolites for an organism has been described by [60]. This set is termed a "seed set" and is the minimum set of compounds that (1) an organism cannot produce itself and (2) enables the production of all other compounds in the network. The seed set defines the metabolites an organism needs from the environment, regardless of whether they are being handed-off by another organism. Borenstein and Feldman [61] showed that the seed set of a parasitic organism is more likely to be found in the metabolic network of its host (compared to finding the seed set of a non-parasitic organism). Hand-off points are therefore the subset of seed set metabolites that are produced by another organism, the remaining seed metabolites occurring naturally in the environment.

KB-1 community interactions have been previously described [62]. *Dehaloccoides* is capable of carrying out the PCE degradation pathway shown in Figure 1.1 and *Geobacter* can perform of the first two reactions. *Dehaloccoides* is anaerobic; although *Dehaloccoides* encodes two oxygen scavenging proteins and can survive small amounts of oxygen, significant exposure slows its growth and can kill pure *Dehaloccoides* cultures [63]. *Dehaloccoides* is less sensitive to oxygen exposure when part of the KB-1 community, where other anaerobic members perform additional oxygen scavenging. Apart from increased oxygen sensitivity *Dehaloccoides* shows poor growth in isolation and improved growth when part of the KB-1 community. The observed reliance on other community members suggests two hand-off points: *Dehaloccoides* requires a cobalamin (vitamin B12) cofactor for growth yet encodes an incomplete pathway to produce it [62,64]. *Dehaloccoides* is capable of only the final part of the pathway and must acquire intermediate metabolites from other community members; proteins to acquire these intermediates from the environment have been located in the *Dehaloccoides* genome. In the second hand-off point *Dehaloccoides* requires methionine and is capable of acquiring it from the environment, but a pathway to produce it has not been identified [62].

1.8 Objectives of this Work

Methods to identify and analyze the composition and pathways of a metagenome form the focus of this thesis. Using KB-1 as an example these methods are implemented and refined. In Chapter 2 a new alignment-based rank-flexible taxonomic classifier is developed that is tolerant to events that confuse classification and is used to classify KB-1. In Chapter 3 functional annotations are made to the proteins predicted in KB-1 and an *in-silico* model of the metagenome's metabolic network is built from these annotations, this model is then labeled with the KB-1 taxonomic assignments. This allows a description of the KB-1 community metabolism. Possible metabolic interactions and dependencies between the community members in the form of hand-off points are also identified.

CHAPTER 2 TAXONOMIC CLASSIFICATION BY COMPARING AFFINITIES IN TWO DIMENSIONS

To understand which organisms in KB-1 contribute to the PCE degradation pathway the KB-1 protein sequences need to be predicted and assigned a taxonomic classification. Homology-based taxonomic classifiers start with a set of significant homology matches to a database of reference sequences (Figure 1.4a), where each match is composed of (1) the reference lineage from domain to species and (2) a numerical score, which is both an e-value denoting the statistical confidence of the match and a bitscore denoting the number of nucleotides in the query sequence that match nucleotides in the reference sequence. The bitscore does not describe the likelihood of the match in regards to the entire reference database as the e-value does. Classifiers use these two dimensions in different ways to assign taxonomy, sometimes also estimating the query sequence's level of novelty to make a rank-flexible assignment. LCA uses the e-value and bitscore to filter low quality matches and combines the matched lineages into a lowest common ancestor assignment. Only CARMA3 uses both dimensions concurrently to make an assignment by mapping bitscores from a reciprocal BLAST onto the best matching lineage.

A new algorithm called SPANNER (Similarity Profile ANNotatER) is described to better integrate this two-dimensional set for greater assignment accuracy. SPANNER uses the set of all BLAST matches below an e-value threshold from an unknown protein sequence to a reference database of sequences whose taxonomy is known. This set of BLAST matches is termed an LCA Profile. Each match in the LCA Profile consists of the reference protein's taxonomy and the e-value. A query LCA Profile is built for a query protein by comparing it to the reference database using BLAST (Figure 2.1a), likewise reference LCA Profiles are created by comparing every protein in the reference database to every other reference protein (Figure 2.1b). The bitscore range for inclusion of BLAST matches in an LCA Profile is based on setting a proportion p; only matches with bitscore $\geq p \times$ the highest bitscore in the LCA Profile are included in the LCA Profile (Figure 2.1c, LCA Profiles are also seen in LCA and SOrt-ITEMS after matches are trimmed to p, see the start of Figure 1.4c and Figure 1.4d).

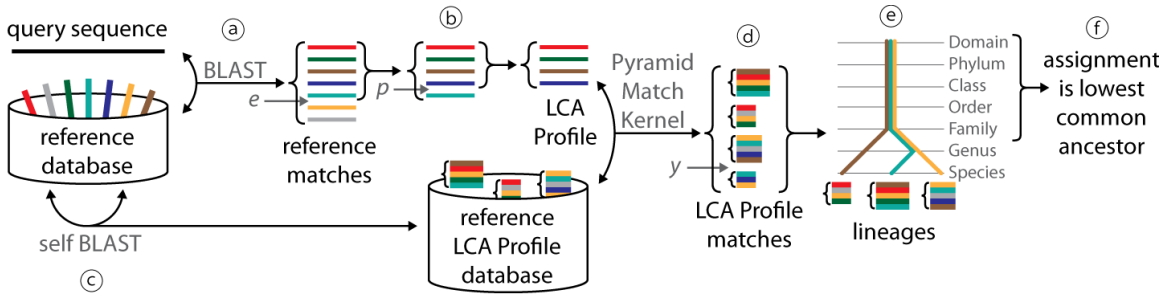


Figure 2.1 The SPANNER classification algorithm. **A**: Like other supervised classifications algorithms (see Figure 1.4a) SPANNER starts with a comparison of a query sequence (black) to a reference database of sequences with assigned taxonomy. Insignificant matches are filtered at an e-value threshold. **B**: Poor matches are again filtered at a bitscore threshold *p*, the remaining set of matches is termed a "LCA Profile." **C**: Every reference sequence is also compared to the reference database using BLAST, these are also filtered at an e-value threshold and at *p* to create reference LCA Profiles. **D**: The Pyramid Match Kernel compares the query LCA Profile to the reference LCA Profiles creating a set of matches, poor matches are filtered at *y*. **E**: The lineages of the remaining matched reference LCA Profiles are compared and **F**: the query is assigned to their lowest common ancestor.

Pairwise scoring of profile similarity is non-trivial, because profiles will have degrees of similarity in terms of both taxonomy and match quality. For example, two profiles may both contain matches to members of the same species, genus, or family, with ranges of e-value matches that are proportionately similar. An appropriate scoring scheme would assign maximum scores to pairs of profiles that are identical in both their taxonomic composition and the relative similarity of the different taxonomic hits. Weaker matches should be recognized but assigned a lower score. The Pyramid Match Kernel (PMK) [65] was adapted to calculate distances. To apply the PMK, match information is embedded in a two-dimensional grid. Genomes matched by the two profiles are placed along one axis (the *taxonomy axis*) in a manner that groups organisms by genus, family, and every additional rank up to domain, and all e-values, normalized to accommodate different rates of substitution in different genes, along the other (the *e-value axis*). The e-values in each profile are normalized separately before being placed on the axis. The key property of the PMK is a hierarchical subdivision of the grid. The taxonomy axis is naturally divided into an eight-level hierarchy (h=8), representing the taxonomic ranks from species to "cellular organisms," while the continuous e-value axis is divided into 2^h sections. Figure 2.2a

shows the initial configuration for the PMK, using a four-level hierarchy for simplicity. The algorithm runs in h iterations. At each iteration the number of multiset intersections between the two LCA Profiles in each section of the grid is counted and multiplied by the weight of that iteration; weights start at 1.0 for iteration one and are halved for each successive iteration (so a weight of $1/2^{i-1}$ for iteration i). The second iteration considers taxonomic matches at the rank of genus and subdivides the e-value axis by 2^{h-1} sections (so the e-value axis sections double in size); this is shown in Figure 2.2b. Again the intersections are counted and multiplied by a weight of ½. The iterations continue (Figure 2.2c), increasing the size of the sections on both axes, counting the intersections and multiplying it by the weight, until at iteration h there is only one section spanning all of both axes (Figure 2.2d). The sum of all the weighted intersection counts is the similarity between the two LCA Profiles.

Each query LCA Profile is compared against a set of reference LCA Profiles whose taxonomy is known, generating a list of PMK matches (Figure 2.1d). To preserve the rank-flexible nature of LCA, the lowest common ancestor of the top LCA Profile matches is used as the final assignment. The set of best-matching profiles is generated in a manner similar to the generation of the original LCA Profiles: the best-matching reference profile is identified, and all other profiles are included whose PMK score is greater than $y \times$ the score of the best-matching profile (Figure 2.1e-f). Taxonomic assignments are therefore based on the similarity between the homology matching pattern of a metagenomic read, and similarity patterns of proteins from the reference database of microbial genomes: if many proteins from a particular genome have unusual patterns of taxonomic similarity due to LGT or other evolutionary or statistical phenomena, then metagenomic reads with similar affinity patterns will be assigned in a manner that is not overly conservative.

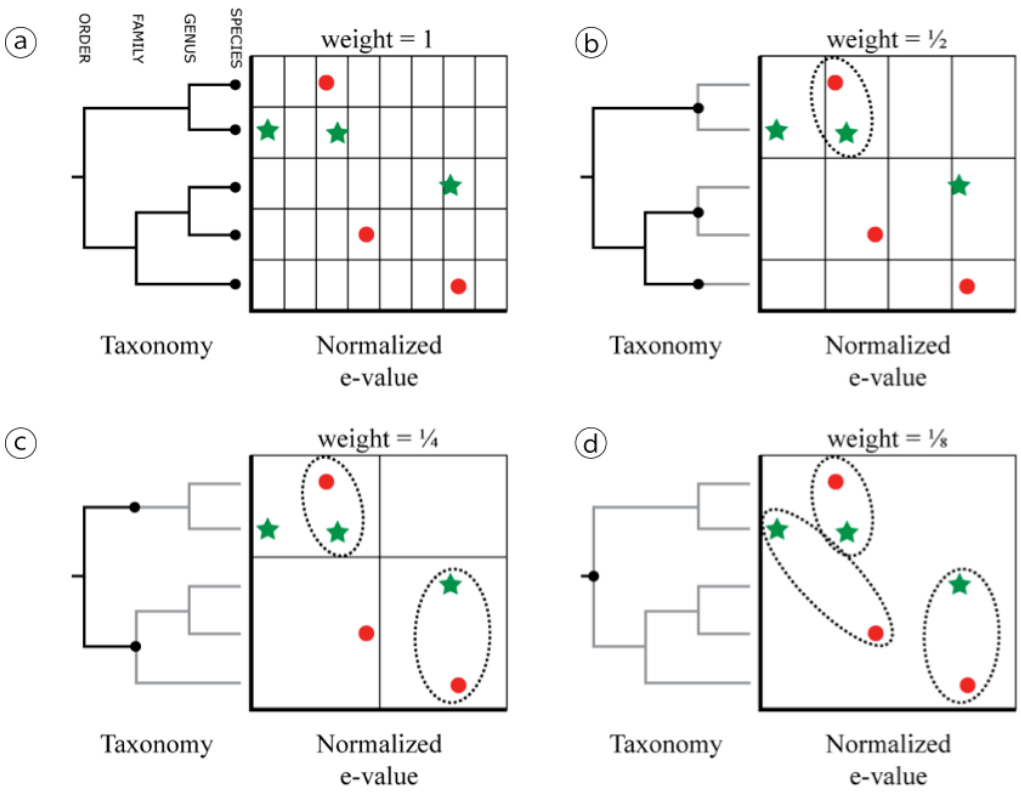


Figure 2.2 Calculating LCA Profile similarity using the Pyramid Match Kernel. In this simplified example, taxonomy consists of only h=4 ranks instead of the usual 8. The green-star profile is being compared against the red-circle profile, each profile consisting of three homology matches at a given species and e-value. **A**: The initial configuration (iteration i=1) of the Pyramid Match Kernel: each species in both profiles is arranged along the taxonomy axis, the hierarchy axis is divided evenly to represent these species. The e-value axis is divided into 2^h sections. The weight of an intersection at this level of granularity is 1; no intersections exist. **B**: Iteration 2 has the e-value axis sections double in size, the taxonomy axis is divided by genus. The weight of the one intersection found is ½. The overall similarity is now ½. **C**: Iteration 3 has the e-value axis sections again double in size, and the taxonomy axis is divided by the next higher rank of family. Two intersections are found at a weight of ¼, making the overall similarity (½) + (¼ + ¼) = 1. **D**: By iteration h the e-value axis sections have doubled until only one section consists of the entire axis, likewise the taxonomy axis has reached the root of the tree (the rank of order) so that axis consists of only one section. Three intersections are counted at a weight of ⅛, making the overall similarity (½) + (¼ + ¼) + (⅛ + ⅛ + ⅛) = 1.375.

2.1 VALIDATION OF SPANNER ALGORITHM

In addition to classifying the KB-1 metagenome, two other datasets were used to assess the performance of SPANNER. All analyses were based on a reference database of 1210 bacterial and archaeal genomes (the two microbial domains found in KB-1) acquired from Genbank in June 2010. SPANNER was first applied to a pseudometagenome (a simulated metagenome) with properties similar to those of KB-1. A small subset of the reference genomes comprised the pseudometagenome; these were chosen to mimic the taxonomic novelty of KB-1, where some community members have close relatives in the reference database of sequenced genomes, while others are members of novel classes or phyla. Simulated sampling of contigs from genomes included in the pseudometagenome was also performed to mimic the distribution of contig lengths and taxonomic abundance found in the KB-1 metagenome. SPANNER was also applied to the real KB-1 metagenome in which the taxonomic affiliations of individual contigs is not known.

For the third dataset, a subset of the reference database was used to generate simulated reads in a taxonomic "leave-one-out" framework similar to that of [66] and [36]. This framework classifies all protein sequences from a set of genomes as queries, using sequences from the remaining gnomes as a reference database. The leave-one-out framework is performed at a specified taxonomic novelty and classified only to ranks above this novelty. The novelty is enforced on query sequences by excluding from the reference database any genome related to the query at that rank or lower. For example, to perform leave-one-out classification at a novelty of family all proteins from each genome are classified in turn, against a reference database comprising all remaining genomes *except* those genomes that share the query's family. This ensures that there is no reference genome related to the query at a rank lower than order, and assignments are made to the rank of order since correct assignments below this rank are not possible. While the "leave-one-out" trials do not generate entire simulated communities, they are useful to assess the performance of SPANNER at different levels of taxonomic novelty.

Comparing the performance of rank-flexible classifiers is challenging, because each prediction has a taxonomic *precision* (the rank at which the classification is made) and

accuracy (the most-precise rank in the prediction that is correct). For example, a given sequence may be classified to the rank of genus, but accurately only to the rank of class, in which case the classification is partially correct and partially incorrect. The strategy used for comparing predictions is illustrated in Figure 2.4. Taxonomic precision, shown on the x-axis, is the number of taxonomic ranks assigned by SPANNER or another algorithm, whether correct or incorrect. The y-axis shows the number of assigned ranks that are incorrect. By treating each rank as a quantitative value from 0 ("cellular organisms") to 7 (species), averages can be computed over all predictions made on a given data set. For example, an average of 3.5 ranks of taxonomic precision means that the central tendency of assignments for all proteins fell between the ranks of class and order. Since accuracy is expressed as the number of ranks that are correct, two classifications are referred to as having equivalent accuracy if they have the same number of ranks correct, whether the remaining ranks are unspecified or incorrect: in Figure 2.4, diagonal lines show equivalent accuracy across ranges of precision and incorrect ranks. However, by mapping accuracy in two dimensions, this evaluation scheme can nonetheless distinguish predictions that are precise but somewhat incorrect, versus less-precise predictions that are completely correct.

2.2 METHODS

2.2.1 KB-1 Analysis

Protein-coding sequences were predicted using MetaGeneMark [8] and compared using BLASTP (a variant of BLAST) with an e-value threshold of 10^{-3} against all 1210 reference genomes to create a set of KB-1 LCA Profiles. Reference proteins were compared against one another to generate reference LCA Profiles. Only the highest (by bitscore) match to a given taxon was kept in any LCA Profile, with lesser matches to the same taxon ignored. The KB-1 LCA Profiles were compared to the reference LCA Profiles (p=0.85 and y=0.95) to obtain rank-flexible assignments. KB-1 was also classified using best BLAST and LCA for comparison.

2.2.2 Pseduometagenome Analysis

A pseudometagenome is an artificial metagenome created specifically to model a real metagenome. This was used for validation purposes: results from analysis on the KB-1 metagenome cannot be validated *in silico* so to validate KB-1 analysis a pseudometagenome was created from 13 microbial genomes to model the KB-1 metagenome as closely as possible (Table 2.1). For each organism found in KB-1 a proxy was chosen from a similar taxonomic group with the same degree of taxonomic novelty as the true KB-1 member. For example, a proxy for a KB-1 methanogen novel at the rank of family would be another methanogen also novel at the family level. The complete genome for all 13 proxies used was retrieved from Genbank. Like KB-1, the pseudometagenome included a genome novel at the phylum level (*Opitutus*), a mix of archaeal and bacterial genomes, four methanogens (all from the same order and three from the same class), and two members from a genus in the Spirochaetaceae family (*Treponema denticola* and *Treponema pallidum*).

Table 2.1 List of the expected KB-1 taxa and the corresponding proxy taxon in the KB-1 pseudometagenome.

Taxon in KB-1	Proxy in pseudometagenome
Dehalococcoides	*Dehalococcoides CBDB1*
Geobacter	*Geobacter lovleyi*
Methanomethylovorans	*Methanohalobium evestigatum*
Methanomicrobiales	*Methanoregula boonei*
Methanosarcina	*Methanosarcina barkeri*
Methanosaeta	*Methanosaeta thermophile*
Sporomusa	*Veillonella parvula DSM 2008*
Acetobacterium	*Moorella thermoacetica ATCC 39073*
Spirochaeta SA-8	*Treponema denticola ATCC 35405*
Spirochaeta SA-8 2	*Treponema pallidum subsp. pallidum SS14*
Syntrophus	*Syntrophus aciditrophicus*
Chlorobi SJA-28	*Chlorobaculum parvum NCIB 8327*
OP5	*Opitutus terrae PB90-1*

A subset of the 1210 completed microbial genomes was used as the reference database for classification purposes. The thirteen genomes used to build the pseudometagenome were excluded from the reference set, as were genomes from the phylum *Verrucomicrobia*, which was necessary to make *Opitutus* unique at the phylum level when comparing it to the reference database. BLASTP version 2.2.18 was used to perform all-versus-all comparisons among reference proteins, with a maximum e-value threshold of 10^{-3} for inference of putative homologs. LCA Profiles for the reference genomes were created by using the p=0.85 bitscore threshold used by [37]. Genomic fragments from each pseudometagenome taxon were sampled in proportion to the estimated abundance of that taxon in the KB-1 culture (see Table 1.2) as well as the average contig length. Proteins and protein fragments were predicted on these sampled fragments using MetaGeneMark version 2.7d and compared to the reference protein set to generate query LCA Profiles. Only the highest (by bitscore) match to a genome was kept in an LCA Profile, all lesser matches to the same genome were discarded. Query and reference LCA Profiles were compared using the PMK, with y, the parameter controlling the number of reference profiles included when choosing a taxonomic rank and label, set to 0.5, 0.6, 0.7, 0.8, and 0.9 in separate trials.

2.2.3 Leave-One-Out Analysis

334 microbial genomes were selected from the larger set of 1210 reference genomes (selecting at least three representatives per genus). To create a query dataset, 1000 fragments of lengths 200 and 1000 bp were sampled from random locations in each of the 334 microbial genomes. Proteins were predicted from these fragments and compared against one another in the same manner as described above to create query LCA Profiles, with secondary matches to particular taxonomic groups ignored. The reference and query LCA Profiles were compared using a range of parameter settings (p=0.65, 0.75, 0.85, 0.95 and y=0.65, 0.75, 0.85, 0.95) to generate taxonomic assignments at three levels of taxonomic novelty: species, genus, and class. The accuracy of SPANNER was assessed by using a "leave-one-out" strategy to classify predicted proteins at different levels of taxonomic novelty. To use LCA Profiles at a species level of novelty, all BLAST

matches to the same species were removed in both query and reference LCA Profiles. Assignments were then made at the genus level, since it is impossible to assign to the correct species. For a genus level of novelty, all BLAST matches to the same genus were removed and assignments were made at the family level. For a class level of novelty, all BLAST matches to the same class were removed and assignments were made at the phylum level.

2.3 RESULTS

2.3.1 KB-1

LCA assigned proteins to a rank between class and order on average, while SPANNER assigned proteins to a rank between family and genus. Best BLAST predictions, being rank-specific, were interpreted at the level of genus since it is likely that no conspecific genomes were present in the reference database. The taxonomic assignments of LCA, SPANNER, and best BLAST are summarized in Figure 2.3a, b, and c respectively, which highlights assignments to the 13 expected KB-1 taxa at all ranks. LCA assigned 10% of the proteins to “cellular organisms”, compared to 0.4% using SPANNER. LCA also assigned more proteins to the rank of domain than SPANNER (28% vs 5% for Bacteria, assignments to Archaea was less than 1% for both). SPANNER assigned more proteins than LCA at the rank of class and below. Although SPANNER had greatly increased taxonomic precision relative to LCA, SPANNER had a higher proportion of assignments to taxonomic groups that are not expected to be present in the KB-1 metagenome (labeled as "other" in Figure 2.3a, b, and c). The increased taxonomic precision comes at the cost of an increased number of incorrectly assigned ranks. For example, at the genus level, approximately 37% of LCA assignments were to genera not known to be present in KB-1, while the corresponding number was approximately 54% for SPANNER and 63% for BLAST.

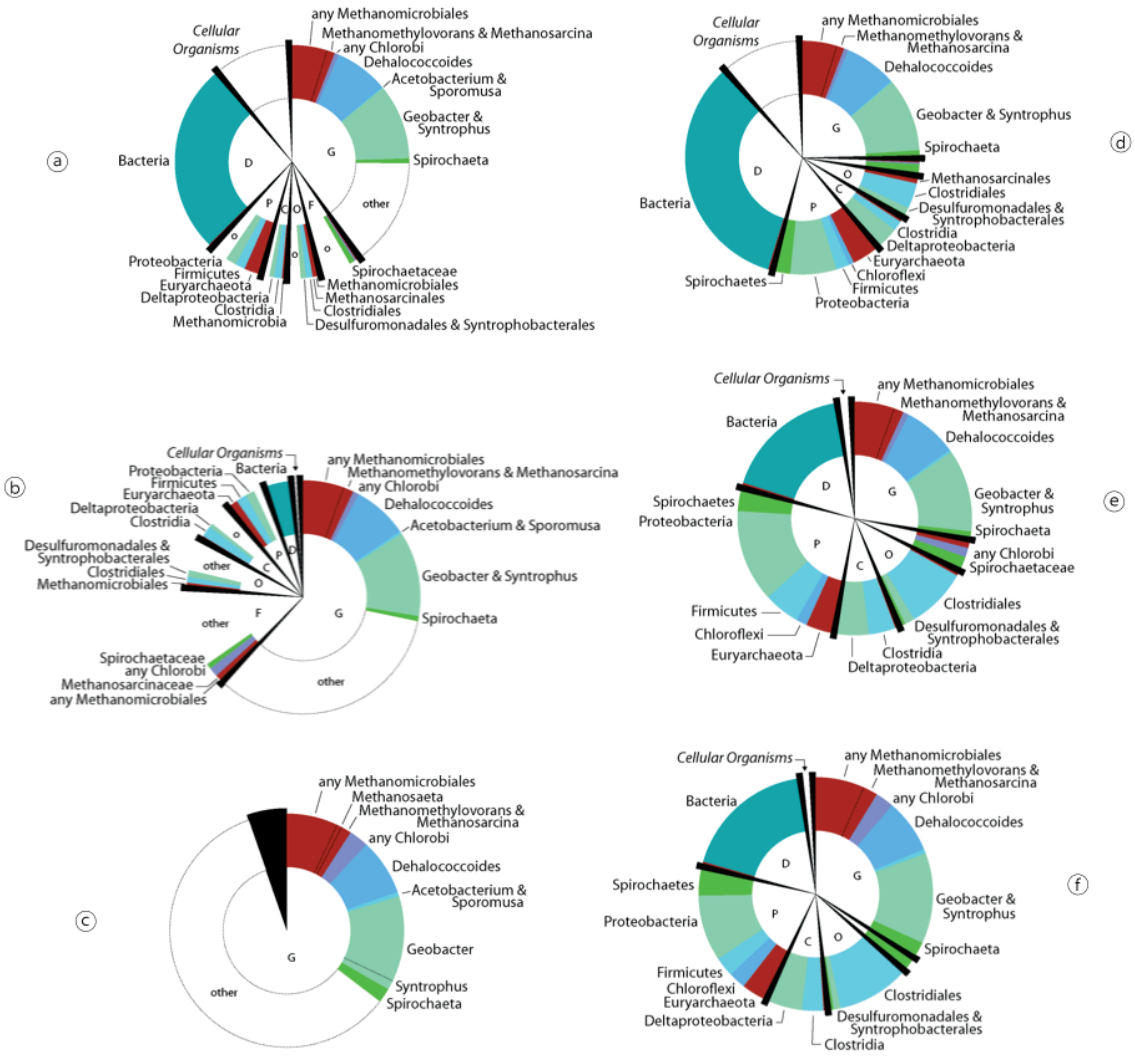


Figure 2.3 Taxonomic predictions of the KB-1 metagenome by SPANNER, LCA, and best BLAST. The predicted taxonomic rank from genus (G) to domain (D) is shown, with all best BLAST assignments made at the rank of genus. The lineages of the 12 expected taxa in KB-1 are shown at the rank predicted; all other predictions are labelled 'other.' Left-hand panels show the distribution of assigned labels and ranks for LCA (**A**), SPANNER (**B**) and best BLAST (**C**). Right-hand panels show the lowest correctly assigned rank for LCA (**D**), SPANNER (**E**) and best BLAST (**F**).

These predictions were also evaluated by considering the lowest *correct* rank for each prediction made by the three algorithms (Figure 2.3d, e, and f for LCA, SPANNER, and best BLAST, respectively). In this case, the assignments based on best BLAST matching resemble those of the rank-flexible classifiers because the lowest correct rank of any BLAST assignment can be anywhere between genus and "cellular organisms". LCA

overall had more assignments that were correct only at the level of domain or "cellular organisms" (i.e., no classification was made, or even the predicted domain was incorrect). SPANNER and BLAST yielded similar distributions of predictions, although SPANNER had more assignments that mapped to "cellular organisms" and fewer correct assignments at the rank of genus.

To assess the accuracy on the metagenome, a “gold standard” of assignments was produced for validation. A reduced reference database of only the closest proxies to the 13 expected KB-1 taxa was created. For example, at least one strain of *Dehalococcoides* is present in KB-1, so all five *Dehalococcoides* genomes from the original reference database were included in the reduced reference database (since the exact *Dehalococcoides* species is unknown all species were included). This reduced reference database consisted of 43 closest proxies to KB-1. Best BLAST results of KB-1 contigs against the reduced database created the “gold standard” of results: best BLAST could only match contigs in the metagenome to one of the 13 expected taxa. Note that this differs from the normal use of best BLAST to classify a metagenome, since the reference database has been limited only to what taxa are *a priori* assumed to be present. Since the reference database has been reduced, a higher confidence can be taken in the best BLAST gold standard for these contigs if they are also long. SPANNER assignments of proteins on contigs 50,000 bp or longer matched the gold standard for 98% of the total *Dehalococcoides* ranks, and 93% of non-*Dehalococcoides* ranks.

2.3.2 Pseudometagenome

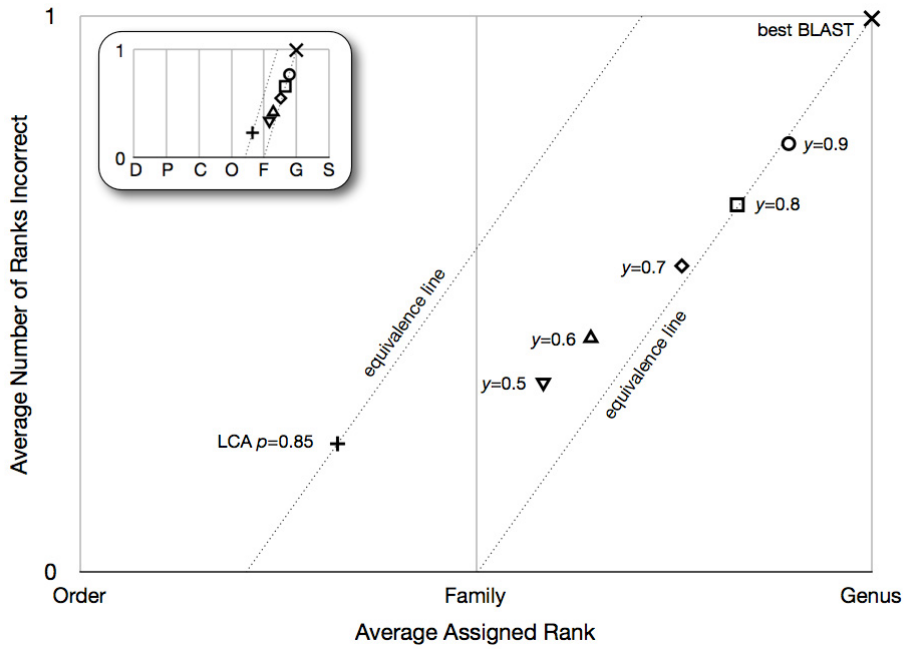


Figure 2.4 SPANNER classification of the KB-1 pseudometagenome to the genus level. Diagonal lines extending from the LCA and best BLAST points connect all other points considered equivalent in terms of overall accuracy: anything on these lines introduces as many assigned ranks as incorrectly assigned ranks, so any point along these lines has the same number of correctly assigned ranks. The inset shows the data relative to all taxonomic ranks, from domain (D) to species (S).

Since none of the species used to build the pseudometagenome had conspecific organisms in the reference database, correct assignments could only be made at the genus level or higher. Although BLAST matches query sequences with targets from specific strains, the taxonomic precision of BLAST matches was limited to the genus level to account for this level of taxonomic novelty; otherwise all BLAST predictions would have been guaranteed at least one incorrect rank (species). Since the use of best matches yields a rank-specific classifier, all predictions were made at the genus level, even for organisms that were novel at much higher taxonomic ranks. Best BLAST matches were on average 0.995 ranks too precise (on average predictions were accurate to just below the level of family), as seen in Figure 2.4. LCA with p=0.85 assigned proteins 1.35 ranks above genus (i.e. the "average" prediction rank was between order and family), with essentially

no incorrect ranks (not over-specific) since the lowest common ancestor almost always encompassed the source of the protein being assigned. LCA avoided the over-specific problem of BLASTP by assigning to higher ranks, decreasing the number of incorrect ranks at a cost of taxonomic precision. SPANNER results for a range of parameter settings (p=0.85 and y=0.5, 0.6, 0.7, 0.8, 0.9) are shown in Figure 2.4, which on average assigned reads between 0.21 (y=0.9) and 0.83 (y=0.5) ranks above genus, and between 0.77 (y=0.9) and 0.34 (y=0.5) ranks incorrect. Unlike best BLAST, rank-flexible classifiers LCA and SPANNER are not guaranteed to have incorrectly assigned ranks for taxa novel at ranks higher than genus since assignments can be made at higher ranks. SPANNER was both less over-specific than best BLASTP and more precise than LCA, although only for a stringent setting of y=0.9 were the SPANNER predictions better overall than those of best BLAST.

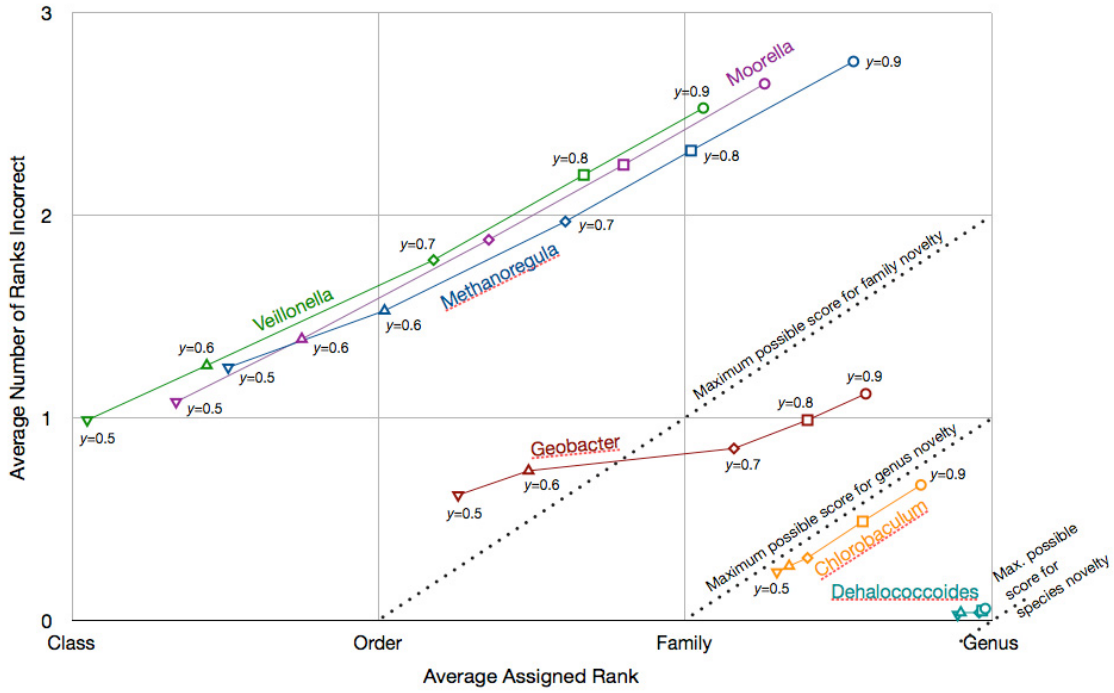


Figure 2.5 SPANNER assignments of the six most abundant taxa in the KB-1 pseudometagenome (p=0.85). The dotted lines represent the maximum possible score for taxa at different levels of novelty. The chosen strains of *Dehalococcoides*, *Chlorobaculum*, and *Geobacter* are novel at the species level, *Moorella* and *Veillonella* are novel at the genus level, and *Methanoregula* is novel at the family level.

The six most abundant taxa (all taxa above 4% in Table 1.2) covered a wide range of taxonomic novelty with respect to the reference database, and accuracy of SPANNER predictions (Figure 2.5). Taxa that were only novel at low ranks (e.g. species-level novelty; having members of the same genus in the reference database) had higher accuracy than taxa novel at higher ranks. Increasing y improved accuracy for some taxa (e.g. *Geobacter* $y \leq 0.6$ versus $y \geq 0.7$) but not others (e.g. *Moorella* and *Veillonella* showed similar performance across all values of y). Figure 2.6 shows the precision of the assignment of each taxon, with pseudometagenome constituents sorted by taxonomic novelty. Taxa novel at ranks above species are more difficult to classify, with the exception of *Opitutus* (novel at the rank of phylum), which was always classified to the correct domain. Although *Geobacter* had an average of 1.1 ranks incorrectly assigned at y=0.9 (Figure 2.5), Figure 2.6 shows most of those incorrect ranks belonged to only 1.8% of *Geobacter* proteins, which suggests a small number of misclassifications of *Geobacter* sequences that are incorrect at high ranks such as phylum, with many of the remainder being correctly assigned to the rank of genus or family. This could be because *Geobacter*, while having congeners, has few or no taxonomic siblings at higher ranks of genus or family, forcing incorrect *Geobacter* predictions to be incorrect up to much higher ranks.

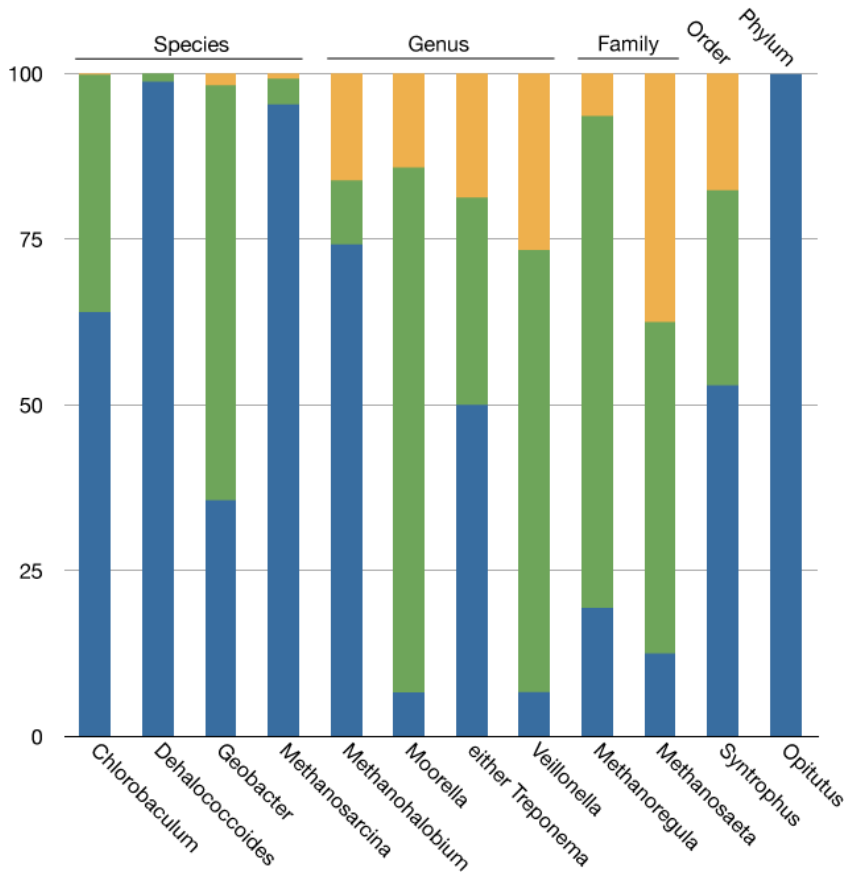
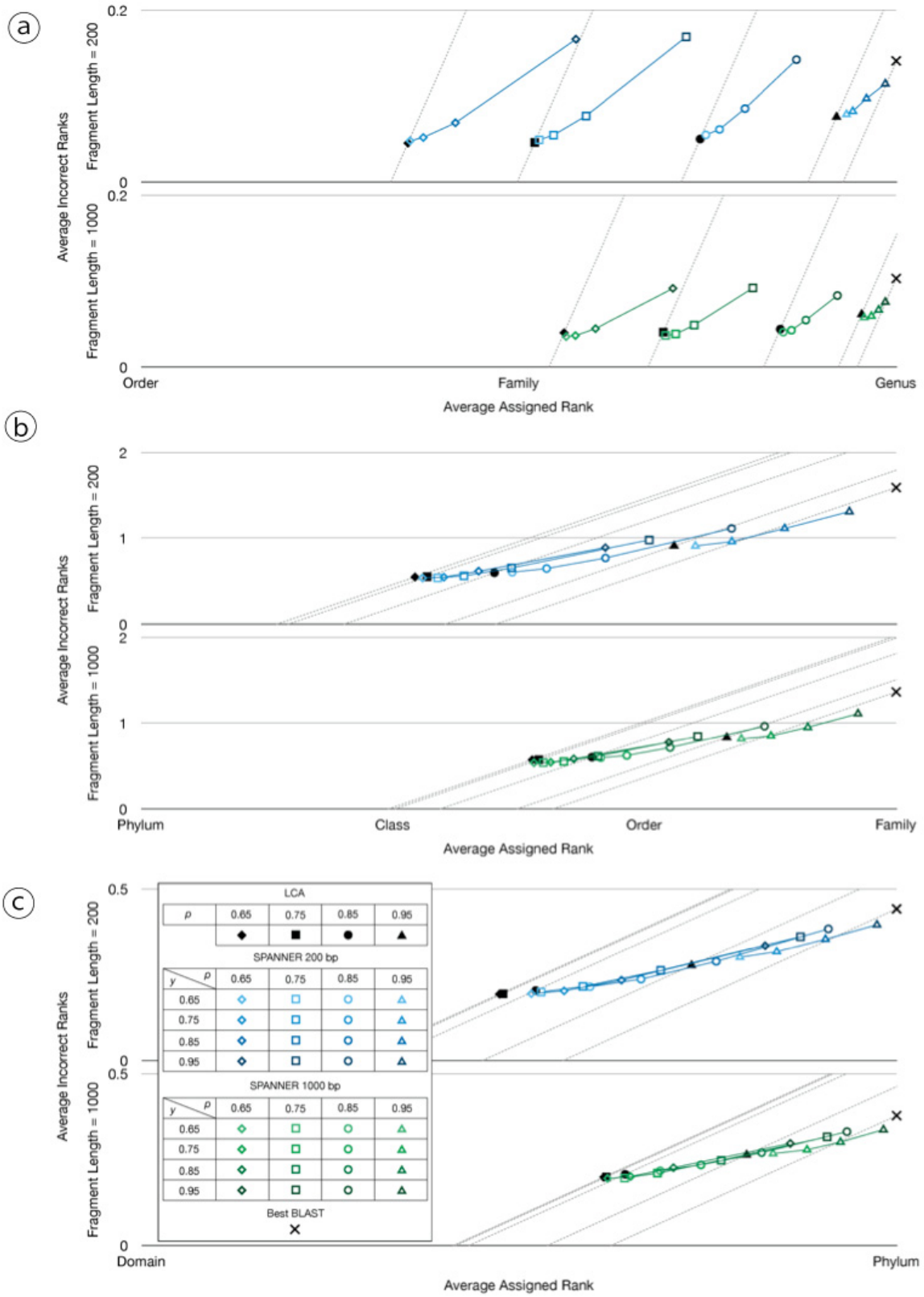


Figure 2.6 SPANNER assignments for each taxon in the KB-1 pseudometagenome (p=0.85, y=0.9). The taxonomic novelty is for each taxon is shown at the top. All genes for each taxon are shown as the percentage of genes correctly assigned (blue), the percentage of genes correctly assigned to a higher rank (green), and the percentage of genes incorrectly assigned (orange). For taxa novel at the species level, for example, assignments correct at the rank of genus are blue; any assignment higher than genus but still within the correct lineage would be green. For taxa novel to the order level, correct assignments to the rank of class are blue.

Figure 2.7 Average taxonomic rank assigned for leave-one-out dataset at three different levels of taxonomic novelty. Panels show ranks assigned for leave-one-out fragments novel at the rank of species (**A**), genus (**B**) and class (**C**). As p increases LCA assignments approach best BLAST, likewise as p and y increases SPANNER assignments approach best BLAST. SPANNER outperformed LCA at all y values and outperformed best BLAST at high values of p and y on genus and class levels of novelty.


a
b
c
Average Incorrect Ranks
Fragment Length = 200
Fragment Length = 1000
Average Assigned Rank
Order
Family
Genus
Phylum
Class
Order
Family
Domain
Phylum
LCA
SPANNER 200 bp
SPANNER 1000 bp
Best BLAST
p
y
0.65
0.75
0.85
0.95

2.3.3 Leave-One-Out Analysis

When results were averaged over all training genomes, best BLAST outperformed LCA at all levels of taxonomic novelty, having more incorrectly assigned ranks but not enough to offset the increased taxonomic precision. SPANNER had higher accuracy than LCA at all levels of novelty over all combinations of parameters *p* and *y*, and outperformed best BLAST at high values of *p* and *y*. SPANNER was always more precise than LCA and less than best BLAST, since all best BLAST matches were assigned at the lowest taxonomic rank possible. The relationship seen between LCA, SPANNER and best BLAST was consistent for fragments of length 200 nt and 1000 nt, overall accuracies were better (1.03, 1.09, and 1.04 times better for species, genus, and class levels of novelty, respectively) on 1000 bp fragments.

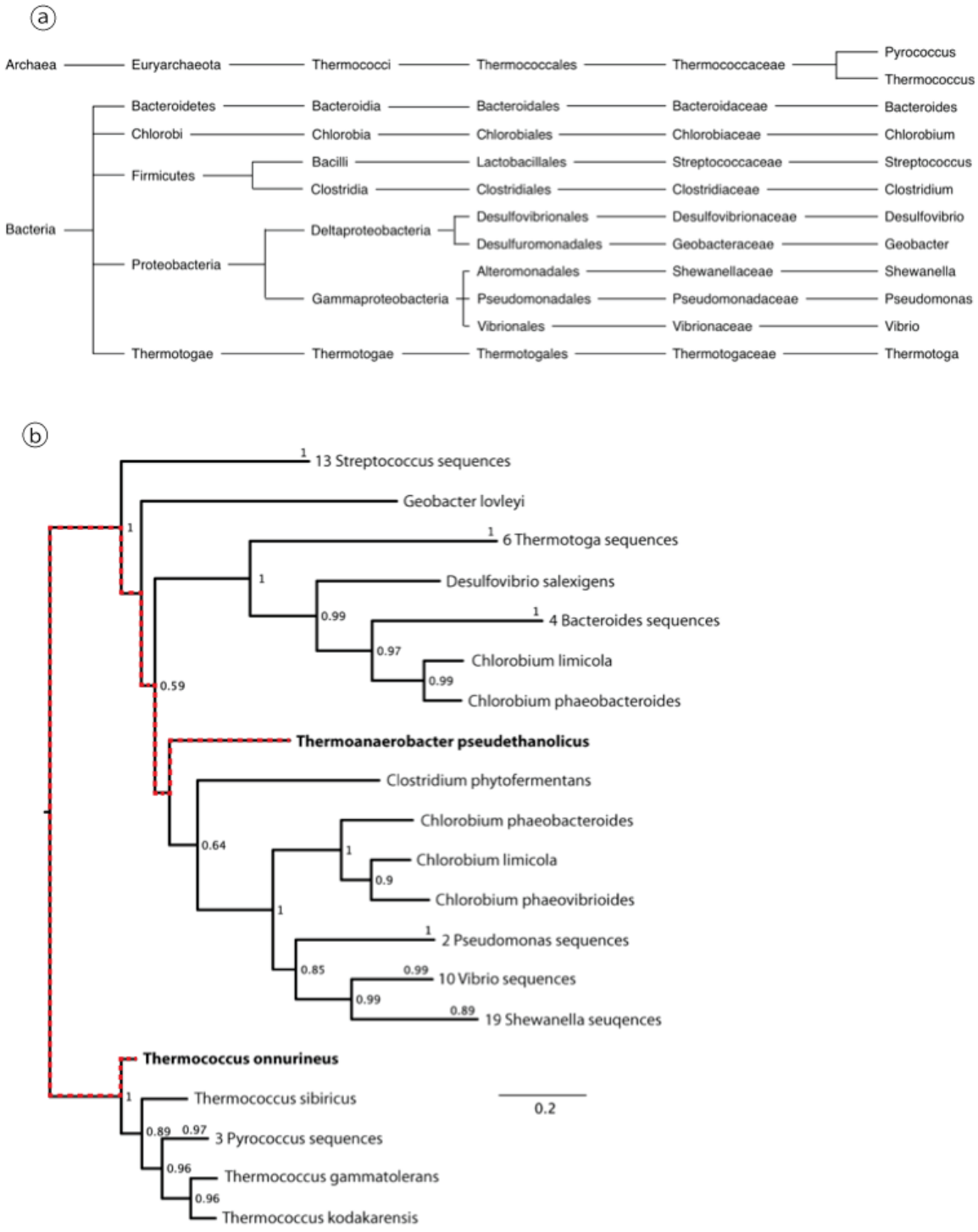
Figure 2.7a shows analysis at a species level of novelty (where assignments are made to the rank of genus or higher). Best BLAST assigned proteins to the rank of genus and had on average 0.1 ranks incorrect for fragment lengths of 1000. LCA on the same fragments had an average taxonomic precision between family and genus, with between 0.04 and 0.06 incorrectly assigned ranks. SPANNER at y=0.65 had similar accuracy as LCA at the same *p*, and increased both precision and incorrect ranks as *y* increased. The most accurate SPANNER parameters were p=0.95 and y=0.95 with taxonomic precision 5.97 and 0.075 incorrect ranks, 0.005 ranks less accurate than best BLAST.

At a genus level of novelty (Figure 2.7b) for 1000 bp fragments, best BLAST had an average of 1.36 ranks incorrect, while LCA assigned proteins on average between class and family with between 0.57 and 0.83 incorrect ranks. For both fragment lengths 200 bp and 1000 bp SPANNER accuracy was similar to LCA at the same *p* and y=0.65; as *y* increased SPANNER approached the accuracy of best BLAST, surpassing it when p=0.95 and y=0.85 or 0.95. Similar trends were seen at a phylum level of novelty (Figure 2.7c), with SPANNER surpassing best BLAST at p=0.95 and y=0.95 for both fragment lengths.

2.4 Classification of Laterally Transferred Genes

Evidence of lateral gene transfer was found in the leave-one-out results by searching for LCA Profiles (at p=0.85) where the taxonomic precision of LCA was phylum or higher, while SPANNER assigned to the rank of family or lower. One example is a sodium/glutamate symporter from *Bacillus pseudofirmus* OF4. This example was taken from analysis at a genus level of novelty, where all other genera within the same family are excluded and the taxonomic assignment was at the family level (since it was impossible to correctly assign to the genus level). After removing matches to its own genus from its LCA Profile and removing matches below the threshold of p=0.85, the two remaining matches are to the archaeon *Methanosarcina mazei* and the bacterium *Geobacillus* sp. Y412MC10. LCA assigned a rank of "cellular organisms," while SPANNER correctly detected the similarities between this LCA Profile and other Bacillaceae LCA Profiles, making the correct assignment of Bacillaceae at the family level. This assignment is 5 ranks more precise than LCA.

Figure 2.8 Classification of a simulated metagenomic read from Thermoanaerobacter pseudethanolicus. **A**: Taxonomic tree of all genera in a Thermoanaerobacter pseudethanolicus LCA Profile at p=0.85 (after removing all matches to its own species). The LCA Profile contained 48 species total with e-values ranging from 10^{-109} to 10^{-93}. LCA classified this as a "cellular organism" while SPANNER correctly identified the genus Thermoanaerobacter by matching it to a similar LCA Profile from Thermoanaerobacter sp. X514. The best BLAST match was to Thermococcus onnurineus, an archaeon. **B**: Phylogenetic tree of all sequences in the LCA Profile from **A**. Although the best BLAST match was to Thermococcus onnurineus, the Thermoanaerobacter pseudethanolicus sequence is placed more closely to its taxonomic neighbours in the Clostridia.


a
Archaea
Euryarchaeota
Thermococci
Thermococcales
Thermococcaceae
Pyrococcus
Thermococcus
Bacteria
Bacteroidetes
Bacteroidia
Bacteroidales
Bacteroidaceae
Bacteroides
Chlorobi
Chlorobia
Chlorobiales
Chlorobiaceae
Chlorobium
Firmicutes
Bacilli
Lactobacillales
Streptococcaceae
Streptococcus
Clostridia
Clostridiales
Clostridiaceae
Clostridium
Proteobacteria
Deltaproteobacteria
Desulfovibrionales
Desulfovibrionaceae
Desulfovibrio
Desulfuromonadales
Geobacteraceae
Geobacter
Gammaproteobacteria
Alteromonadales
Shewanellaceae
Shewanella
Pseudomonadales
Pseudomonadaceae
Pseudomonas
Vibrionales
Vibrionaceae
Vibrio
Thermotogae
Thermotogae
Thermotogales
Thermotogaceae
Thermotoga
b
13 Streptococcus sequences
Geobacter lovleyi
6 Thermotoga sequences
Desulfovibrio salexigens
4 Bacteroides sequences
Chlorobium limicola
Chlorobium phaeobacteroides
Thermoanaerobacter pseudethanolicus
Clostridium phytofermentans
Chlorobium phaeobacteroides
Chlorobium limicola
Chlorobium phaeovibrioides
2 Pseudomonas sequences
10 Vibrio sequences
19 Shewanella seuqences
Thermococcus onnurineus
Thermococcus sibiricus
0.2
3 Pyrococcus sequences
Thermococcus gammatolerans
Thermococcus kodakarensis

An example from the species level of novelty (assigning proteins to the genus level) is a protein from *Thermoanaerobacter pseudethanolicus*, whose LCA assignment at p=0.85 is again "cellular organisms" (Figure 2.8a). Only one reference LCA Profile scored above the y threshold, from *Thermoanaerobacter sp. X514*, making a correct assignment to the genus level. This assignment is 6 ranks more precise than LCA. The best BLAST match for this protein is *Thermococcus onnurineus*, which is 6 ranks incorrect (domain to genus). SPANNER was able to match the LCA Profile from *Thermoanaerobacter pseudethanolicus* with a similar profile from *Thermoanaerobacter sp. X514*, matching the same lateral gene transfer events in both profiles. A phylogenetic tree (created using the UPGMA algorithm [67] on a Smith-Waterman distance matrix of all sequences) of the query sequence as well as the sequences in its LCA Profile is shown in Figure 2.8b, placing the query sequence closer to its taxonomic neighbours instead of with *Thermococcus onnurineus*. The dotted line shows the phylogenetic distance between the query sequence and its best BLAST match.

2.5 Conclusions

Homology comparisons of a query sequence can identify the closest proxy (best match) in a reference database. This is a rank-specific assignment, identifying the query sequence as being most similar to a target sequence in the reference genome. This assignment can only be correct if genomes that share the taxonomic label of the query genome at the target rank are represented in the reference database. Otherwise, the best match will be overly specific, choosing the species of a relative of the query sequence's genome. Rank-flexible classifiers do not suffer from this limitation, since they can assign to any taxonomic rank based on the weight of evidence. Forcing assignments to ranks below the novelty of the query genomes caused best BLAST to have more incorrectly assigned ranks than LCA and SPANNER (at most parameters) in the results. An alternative would be to perform rank-specific classifications at a higher rank, for example Hess *et al.* [68] who classified all sequences to the rank of order. However, using a higher rank forces fragments which could otherwise be assigned precisely to a less-precise category. LCA avoids the over-specific problem by providing a rank-flexible approach, using the top few matches instead of the single best match. However, overly broad

taxonomic matches can unnecessarily reduce the taxonomic precision of a fragment's assignment, and a single LGT event from a distant group can raise the lowest common ancestor to taxonomic ranks as high as phylum, domain, or even "cellular organisms," making the assignment under-specific. In the results LCA had fewer incorrect ranks than best BLAST or SPANNER (at the same *p*) but at the cost of less precision.

SPANNER exploits the situation where many genes from the same genome and from closely related genomes are expected to have similar patterns of homology matches (i.e., LCA Profiles). In this manner, a gene with a broad set of BLAST matches need not generate an overly broad taxonomic prediction, if the overall similarity pattern is characteristic of the correct taxonomic group. Instead of forcing all these matches to contribute to the assignment (causing the under-specific problem in LCA) they can be used as features in an LCA Profile for comparison. The SPANNER approach considers the same set of matches, but does not necessarily need to apply them to the final taxonomic assignment.

There is a trade-off between precision and incorrect ranks among LCA, SPANNER, and best BLAST. LCA is the most conservative classifier, with the fewest incorrectly assigned ranks but the least precision: for example, LCA assigned 25 times more sequences to the "cellular organisms" level than did SPANNER on the KB-1 metagenome, but had fewer assignments to "other" taxa (i.e., those not expected to occur in KB-1). Using best BLAST scores in a rank-specific manner yields high taxonomic precision but many incorrectly assigned ranks. SPANNER spans these two approaches based on the parameters p and y. The best choice of approach in a given situation might depend on the expected degree of taxonomic novelty: for example, metagenomes with many taxa novel at high ranks might be better classified with a conservative approach such as LCA, while metagenomes with close proxies could be confidently assigned using SPANNER. Perhaps alarmingly, Figure 2.3 shows that the three approaches assigned "expected" taxonomic labels at all ranks more precise than domain in only ~50% (SPANNER and best BLAST) and < 50% (LCA) of cases. Although this problem is partially due to the presence of taxa that are novel at high ranks, and the presence in the

dataset of short fragments of predicted coding sequences, it is nonetheless clear that improvements are needed if reliable predictions are to be made by any of the three approaches.

Several types of improvement to SPANNER can be envisioned. Many approaches such as MetaPhyler [69] use lineage-specific scoring thresholds to achieve higher precision; such an approach could also be used here to distinguish lineages with highly variable affinities from those whose LCA Profiles are more similar to one another. Hybrid classifiers [66; 36] currently use the top-scoring homology assignment (e.g., best BLAST matches) in combination with compositional information. The distributional approach of SPANNER would likely decrease the number of false positive predictions, particularly if both SPANNER and the compositional approach were used to define "probable sets" from which intersecting information could be extracted. SPANNER is also complementary to other refinements of LCA such as SOrt-ITEMS [38] which use the degree of orthology in the BLAST and PMK matches to determine the appropriate rank to make an assignment.

CHAPTER 3 KB-1 METABOLIC RECONSTRUCTION AND TAXONOMIC DEPENDENCY ANALYSIS

3.1 INTRODUCTION

Once functional annotation and taxonomic assignments are complete, the pathway of reactions performing PCE degradation shown in Figure 1.1 can be identified, as well as the taxon (or taxa) responsible for them. Metabolic networks can be modeled for each of the expected KB-1 taxa to reveal their functional capabilities. These are referred to as the "member networks." To identify connections between the metabolism of each KB-1 taxa (hand-off points, see Chapter 1) a metabolic network can be constructed that represents the entire KB-1 community, containing all metabolites and edges from the network of each member. This network is referred to as the "community network" or "KB-1 network." To distinguish the functions of each member all edges are labeled with the taxon its protein sequence was assigned to. This network $CN=(V,E,T)$ contains a set of nodes and edges (V and E) as defined in Chapter 1 and a set of taxa (T). For a taxon $t \in T$ the edge $e_{abn} \in t$ if the n^{th} edge from v_a to v_b has been labeled with t. The following pseudocode shows a proposed method to construct a community network CN from the union of $\hat{G}$, the set of all member networks; to avoid ambiguity V_X and E_X are nodes V and edges E of the network X.

```
V, E, T = set
for G in Ĝ
     V = V ∪ V_G
     t = set
     for e_ab in E_G
          i = |e_ab| in E
          i++
          E ← e_abi
          t ← e_abi
     T ← t
return (V,E,T) as CN
```

In the pseudocode, all nodes (metabolites) for all networks are added to the community network. Each node is added only once. All edges (reactions) in each member's network are then added to the union of nodes. A metabolic network representing multiple

organisms is expected to contain multiple proteins performing the same reaction, more so for closely related organisms whose overall metabolism is similar. For this reason all edges are added even if an edge connecting the same nodes was already created from another taxon, making the community network a directed multi-graph (see section 1.6). The following pseudocode defines detecting hand-off points v_t in a community network of metabolite v for taxon t. In the pseudocode, the producers of each metabolite is subtracted from the consumers. Any metabolite consumed by a taxon that cannot produce that metabolite is considered a hand-off point.

```
HP, producers, consumers = set
for v in V
     for t in T
          if e_av in t
               producers ← t
          if e_vb in t
               consumers ← t
     for t in consumers − producers
          HP ← v_t
return HP
```

Rank-flexible classification does not always identify a single organism at the species level, for this reason the KB-1 "members" and their reconstructed networks may represent either more than one organism that classification was not able to differentiate or only a subset of reactions for one organism. Furthermore, incorrect functional annotation can identify reactions that are not actually present or miss reactions that are. Another caveat is that a community metabolic network presumes that all metabolites in all organisms can be shared with all other organisms. Small lipophilic (dissolvable in lipids, the primary molecule that composes the cell membrane) and uncharged metabolites permeate cell membranes naturally; transporters (special proteins on cell membranes that selectively acquire or excrete metabolites) are needed for large or charged metabolites [70].

In this chapter a community metabolic model for the KB-1 metagenome is described. This network is compared to the member networks and network reconstructions from other studies. The union of multiple metabolic networks for multiple organisms is

expected to affect topology although common characteristics should remain. For example, the bow tie topology should still be detectable since some members should be able to use substrates that others cannot and produce products that others cannot. The degree distribution is still expected to show hubs of highly connected metabolites, although no assumptions are made for which distribution it matches and if the network is scale-free or hierarchal. Clustering is performed on the community network. Clusters have been shown in previous studies on the networks of single organisms to form based on shared biological function (i.e. grouping reactions by pathway). Hand-off points and clusters in the KB-1 network are identified.

3.2 METHODS

16S profiling of KB-1 provided a list of 13 taxa expected to be present in the metagenome. Since it is assumed that genes in KB-1 can come only from one of these taxa, the reference database was restricted to only contain the closest proxies to these taxa to reduce the number of incorrect assignments. All protein-coding sequences from all organisms belonging to each taxon in KB-1 were added to the database (Table 3.1). For example the organism identified in KB-1 as a member of the genus *Dehalococcoides* could be any of the 31 already described *Dehalococcoides* species or a novel one, so all sequences from all 31 species were used. Each KB-1 taxon therefore represents a set of reference organisms and a sequence assigned to any organism in a set is considered an assignment to its representative. The sequences were obtained from Genbank (http://www.ncbi.nlm.nih.gov/protein) in May 2012. Reference protein-coding sequences need to represent at least one complete genome per KB-1 taxon, since the contigs could contain any protein from the KB-1 genomes. If no complete genomes were available for a taxon, the rank of that taxon was increased until a complete genome was available (Table 3.1 "Taxon representing reference set"). If two reference sets were merged and represented by a single taxon if they overlapped. The novel phylum OP5 would require including all species under the domain Bacteria which would overlap with seven other taxa, it was not included and OP5 sequences will be incorrectly assigned to other taxa. This is not expected to cause substantial errors as OP5 represents an estimated 1% of the

metagenome. In total ten taxa represent the sets of protein-coding sequences in the reference database. SPANNER was used to assign KB-1: the same KB-1 protein sequences predicted in Chapter 2 were compared using BLASTP against all sequences in the new reference database at an e-value threshold of 10^{-3}, producing a set of KB-1 LCA Profiles. Reference sequences were compared (also with BLASTP at the same threshold) against each other to produce reference LCA Profiles. As in Chapter 2, only the highest (by bitscore) match to a given taxon was kept in an LCA Profile, matches with lesser bitscores to the same taxon were ignored. The KB-1 and reference LCA Profiles were compared (p=0.9, y=0.9) to produce rank-flexible assignments.

Table 3.1 The 13 KB-1 taxa predicted using 16S profiling and the corresponding ten taxa used in the reduced reference database to represent each set of reference sequences. The rank of each representative taxon is given along with the number of complete genomes in each set.

Taxon in KB-1	Taxon representing reference set	Representative taxon rank	Number of complete genomes in set
Dehalococcoides	*Dehalococcoides*	Genus	5
Geobacter	*Geobacter*	Genus	10
Methanomethylovorans	Methanosarcinaceae	Family	11
Methanomicrobiales	Methanomicrobiales	Order	7
Methanosarcina	**Merged with Methanosarcinaceae**		
Methanosaeta	*Methanosaeta*	Genus	3
Sporomusa	Veillonellaceae	Family	35
Acetobacterium	Eubacteriaceae	Family	14
Spirochaeta SA-8	*Spirochaeta*	Genus	5
Spirochaeta SA-8 2	**Merged with *Spirochaeta***		
Syntrophus	*Syntrophus*	Genus	2
Chlorobi SJA-28	Chlorobi	Phylum	14
OP5	**Not represented**		

Functional annotation of KB-1 genes used the results of the BLASTP comparison to the reduced reference database: For each KB-1 protein sequence the Refseq GI number (a unique database key for proteins in Genbank) of the best BLASTP match was converted into a gene id (another database key) using gene2accession, a mapping file provided by Genbank. Each gene id was then converted into Enzyme Commission (EC) numbers using the Genbank records for that gene. An EC number is a hierarchical designation representing different types of reactions [71]. A gene could have zero, one, or multiple EC numbers. Each EC number was converted into a set of reactions in the KEGG database and the substrates and products for each reaction were retrieved. Currency metabolites in the KB-1 network (Table 3.2; these metabolites were chosen from the relevant currency metabolites used in [42]) were filtered from the results. Ten metabolite-centric networks, one for each of the ten representative taxa, were created as described in Chapter 1. Edges were labeled with the SPANNER taxonomic predictions. Edges assigned to a higher rank were included in any of the ten networks if any of the PMK matches that contributed to the classification were represented by the taxon for that network. For example, a sequence with PMK matches to *Dehalococcoides* (phylum Chloroflexi) and *Geobacter* (phylum Proteobacteria) above y=0.9 would be classified as Bacteria at the rank of domain. The edge(s) representing the sequence would be added to the *Dehalococcoides* and *Geobacter* networks, but not the other Bacteria networks. A KB-1 community network was created from the union of the ten member networks as described in section 3.1. The networks were visualized in Cytoscape [72] and analyzed with the Cytoscape plug-ins Network Analyzer [73] for topology metrics, BiNoM [74] for network component analysis, and the GLay [54] implementation in clusterMaker [76] for network clustering. Node degree distributions were analyzed and plotted using the netZ package [77] in R to test their fit to a power-law distribution model and to other models. Hand-off points were identified in the KB-1 network by searching for any metabolite that is a source of one taxon and is produced by a different taxon.

Table 3.2 The currency metabolites removed from the KB-1 metabolic network prior to analysis.

Currency metabolites
ADP
ATP
NAD
NADP
NADPH
Phosphate
H_2O
H^+
Diphosphate
Cytidine monophosphate
CO_2
O_2

Equations 2 and 3 were developed to quantify pathway and taxonomic cohesion in the network clusters generated by ClusterMaker. Equation 2 expresses the number of pathways in a cluster while Equation 3 expresses the number of taxa, both as a score from zero to one. Clusters with many pathways/taxa receive a lower cohesion score than clusters whose reactions all belong to a single pathway. The equations favour dominant pathways/taxa, for example in a cluster of two pathways and eight reactions, a higher cohesion score will be given if seven reactions belong to one pathway and one to the other than if the reactions are evenly split between the two pathways (four reactions each). In the equations p is the set of each pathway p_i in a cluster, t is the set of each taxa t_i in a cluster, e is the edges within a cluster, and $count(p_i)$ or $count(t_i)$ is the number of edges from pathway p_i or taxon t_i.

$$pathway\ cohesion = \frac{\sum_{i=1}^{|p|}\left(\frac{count(p_i)^2}{|e|}\right) - 1}{|e| - 1} \quad (2)$$

$$taxonomic\ cohesion = \frac{\sum_{i=1}^{|t|}\left(\frac{count(t_i)^2}{|e|}\right) - 1}{|e| - 1} \quad (3)$$

3.3 RESULTS

SPANNER assigned 85.9% of the KB-1 genes at the lowest possible rank for each clade and only 1.33% to cellular organisms (Figure 3.1). The taxonomic affinities (represented in the LCA Profile) of KB-1 genes assigned at higher ranks were not distinct enough for SPANNER to differentiate between clades. For example, 5.12% of the KB-1 genes were assigned as Firmicutes (the rank of phylum). SPANNER identified these were either *Sporomusa* or *Acetobacterium* but could not distinguish between them. All of the 13 expected KB-1 taxa belong to two domains: Bacteria and Archaea. SPANNER assigned 6.7% of the genes to the domain Bacteria, suggesting there were only enough differences in the LCA Profiles of those genes to determine they were not Archaea.

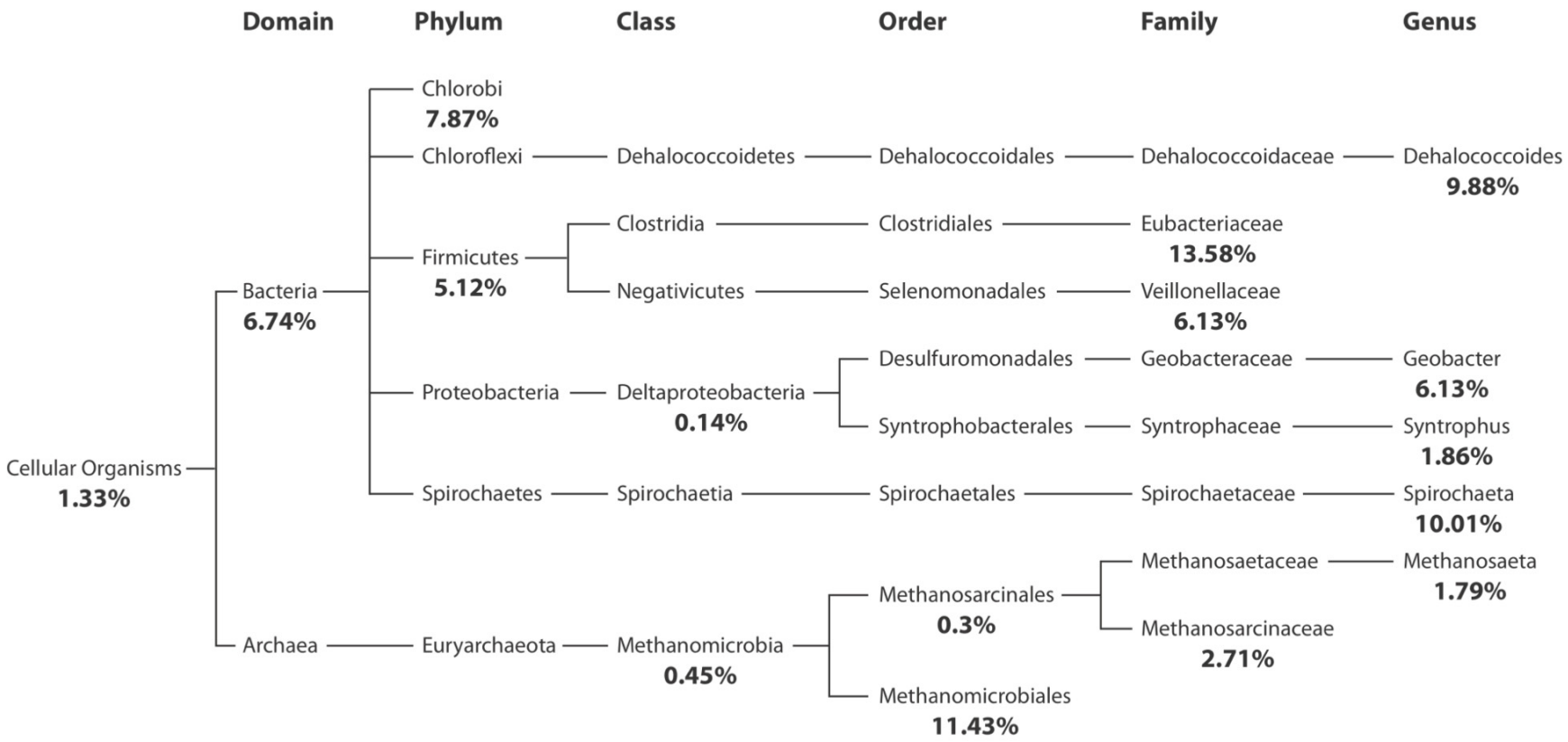


Figure 3.1 SPANNER assignment of KB-1 proteins using a reduced reference database. Taxa are ordered by rank from domain to genus. The percentage of KB-1 proteins classified at a taxon is given under its name.

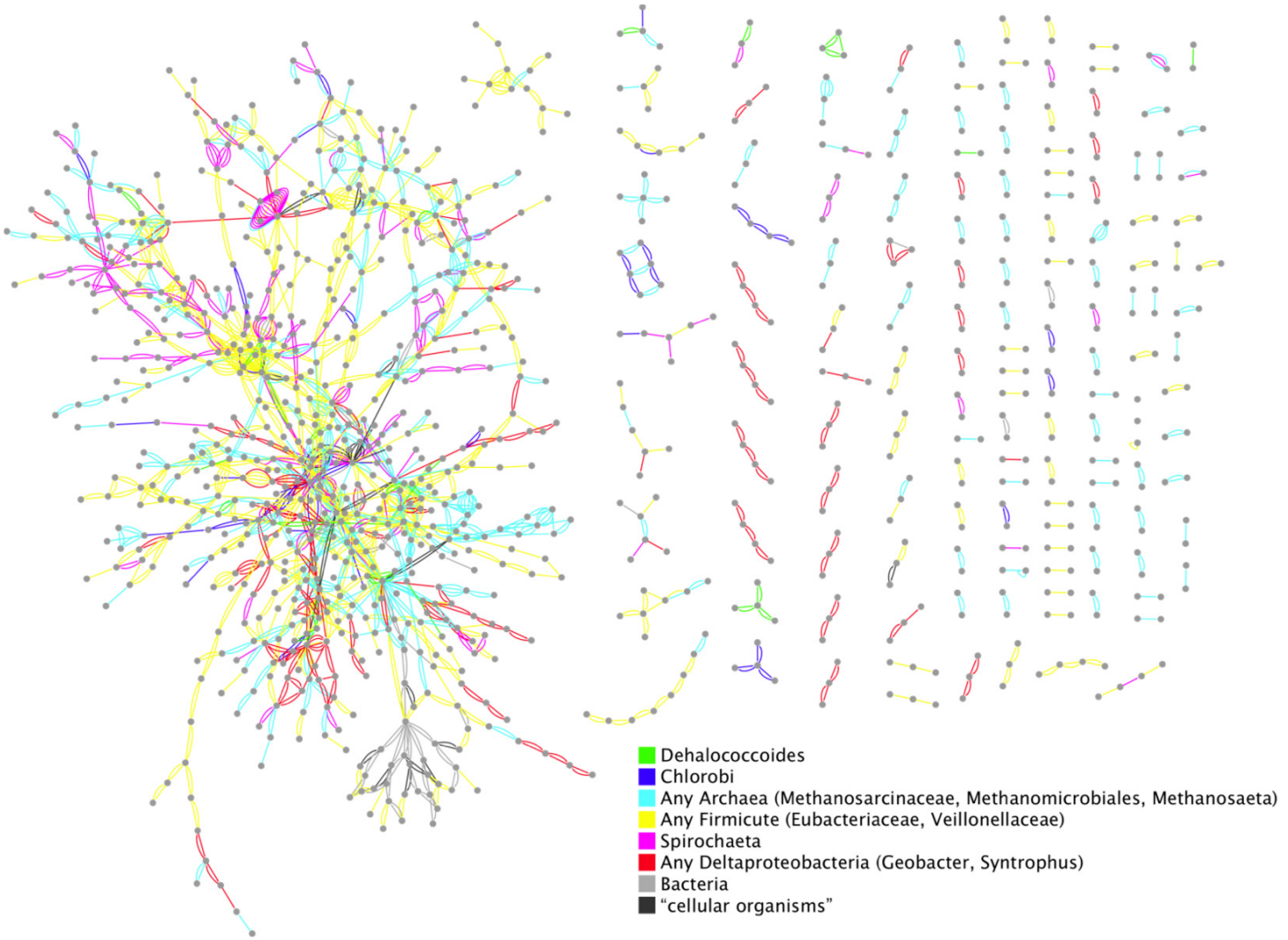


Figure 3.2 Metabolite-centric KB-1 community metabolic network. Reactions are coloured by rank-flexible SPANNER taxonomic predictions as shown in the legend. The network visualization was generated using Cytoscape [68].

The KB-1 metabolic network consists of 1065 metabolites (nodes) and 2259 reactions (edges) in 148 connected components (Figure 3.2). Edges in Figure 3.2 are coloured based on the taxa assigned to them. Since protein sequences from different taxa could have the same function a colour scheme was chosen to visualize reactions that multiple taxa can perform (see figure legend).

One of the network components is the PCE degradation pathway in Figure 1.1, however the final "VC reductive dehalogenase" reaction was not identified. All of the reactions were correctly assigned to *Dehalococcoides*, although *Geobacter* is known to encode proteins for the first two reactions [62] but these were not identified. This highlights the possibility of erroneous taxonomic assignments and functional annotations, as well as incomplete sequencing of KB-1. To estimate incomplete sequencing, all known genome sizes of the taxa in the reduced reference database were averaged to estimate the sizes of each KB-1 genome (OP5 was assumed to be the average of the other KB-1 genome sizes). At 28.5 M the KB-1 metagenome is 74% of the estimated 38.69 M nucleotides.

Table 3.3 compares the ten member networks with the KB-1 community network. The largest member network is Eubacteriaceae with 1075 edges. This network has a larger diameter than the KB-1 network, which shows the networks of other organisms would connect distant metabolites in Eubacteriaceae and create new shortest paths between them. The *Syntrophus* network has only 33 nodes and 36 edges, a strong indication of problems annotating its 619 assigned genes. The shortest path length distribution is for the KB-1 community network is shown in Figure 3.3.

Table 3.3 Properties of the ten reconstructed KB-1 metabolic networks and the community KB-1 network.

Network	Nodes (metabolites)	Edges (reactions)	Connected components	Average number of neighbours	Average shortest path length	Diameter
Dehalococcoides	84	112	25	1.476	2.389	5
Geobacter	199	246	71	1.397	2.154	7
Methanosarcinaceae	148	172	61	1.243	1.261	3
Methanomicrobiales	223	278	85	1.3	3.249	10
Methanosaeta	149	209	44	1.544	2.915	8
Veillonellaceae	164	185	63	1.256	2.243	5
Eubacteriaceae	636	1075	125	1.855	9.944	29
Spirochaeta	213	313	55	1.568	3.268	10
Syntrophus	33	36	15	1.091	1.059	2
Chlorobi	134	173	43	1.373	3.11	6
KB-1 Community (union of all other networks)	1065	2259	148	2.188	8.732	25

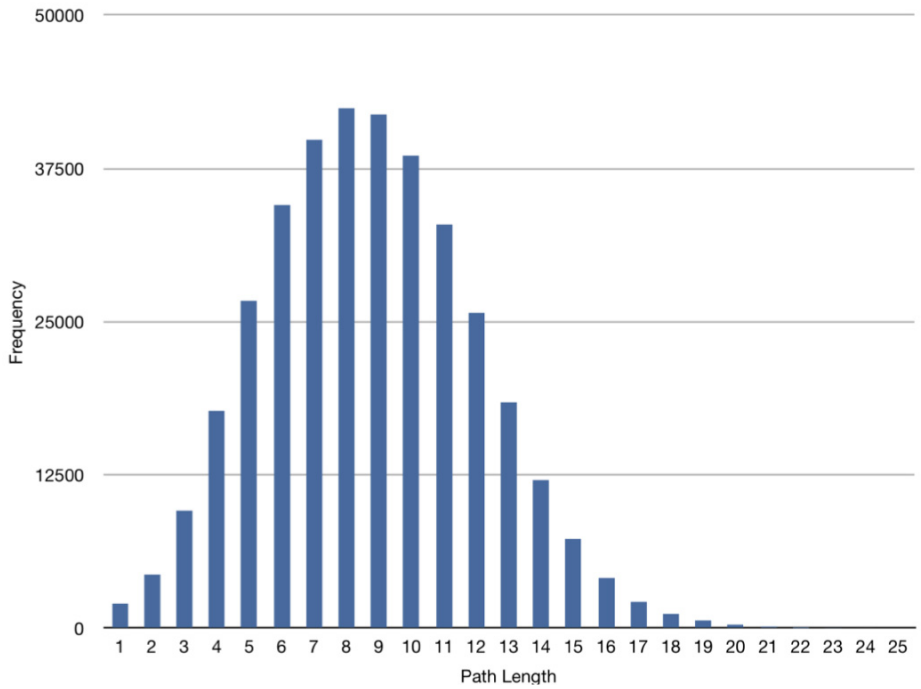


Figure 3.3 KB-1 metabolic network path length distribution.

In the KB-1 community network, the largest connected component contains 679 nodes and 1841 edges, the remaining 147 connected components contained 13 or fewer nodes. Focusing on the largest connected component, ten strongly connected components were found. Similar to the results of [43] one strong connected component was much larger (562 nodes) than the others (5 or fewer nodes), this component is referred to as the "giant strong component" (GSC). Sources and sinks were identified in the largest connected component, 52 and 50 were found respectively. By definition of strongly connected component none of the sources or sinks are part of the GSC. Besides the GSC, [43] characterized three other structures that form a bow-tie topology: The "substrates subset" includes all sources and carrier metabolites entering the GSC, 53 metabolites comprise this set. The "products subset" includes all carrier metabolites exiting the GSC and all sinks, 59 of these were found. The "isolated subset" is the set of nodes that are not in the other three bow-tie groups; 5 of these were found. The KB-1 metabolic network GSC is larger than any found by [43], who analyzed metabolic networks for 65 organisms and found each GSC was less than 300 nodes. Comparing relative sizes of the KB-1 subsets to the subsets [43] found in *E. coli*, the GSC in KB-1 is much larger (79% versus 34% in

E. coli), the substrate subset is smaller (8.5% versus 11% in *E. coli*), the product subset is smaller (8.5% versus 20%), and the isolated subset is smaller (3% versus 35%). Previous studies have suggested the GSC includes the most important pathways of an organism [43]. The KB-1 GSC has similar functional pathway groups as Zhao *et al.* [78] and Ding *et al.* [79] found in *E. coli* and *B. thuringiensis*, such as Carbohydrate Metabolism, Amino Acid Metabolism, Nucleotide Metabolism, and Energy Metabolism. Pathways involved in DNA Translation were only found in the product subset and GSC. The pathway groups Biosynthesis of Other Secondary Metabolites, Xenobiotics Biodegradation and Metabolism, and Glycan Biosynthesis and Metabolism were only found in the GSC. The isolated subset contained a reaction from the pathway Methane metabolism assigned to *Dehalococcoides*, and pathways Glutathione metabolism, Porphyrin and chlorophyll metabolism, Selenoamino acid metabolism, and Glycerophospholipid metabolism, mostly assigned to Archaea.

The GSC of each KB-1 member network was contained in the community network GSC. The union of multiple member networks increased the connectivity of nodes in each member's substrate, product, and isolated subsets potentially adding them to the community GSC. For example, since the GSCs of each member are part of the community GSC, any metabolite in the product subset of one member and the substrate subset of another are both produced and consumed by the GSC in the community network and therefore are a part of it. While the exclusion of the 148 smaller connected components could explain the disproportionately small sizes of the substrate, product, and isolated subsets, the KB-1 GSC would still be comparatively larger than in *E. coli* (51% of the KB-1 network if all connected components were included).

The average clustering coefficient for the KB-1 network is 0.055, substantially less than previously reported for the bacterium *E. coli* at 0.48 and other microbial networks with similar values [43]. Three of the ten member networks had an average clustering coefficient of zero (Chlorobi, Methanomicrobiales, and *Syntrophus*), the remainder ranged from 0.005 (*Spirochaeta*) to 0.1 (*Dehalococcoides*) with a mean of 0.039. The low clustering is likely due to missing edges from absent annotations, however the

metabolism of a microbial community does have higher clustering due to members increasing each other's connectivity. Figure 3.4 shows the KB-1 average clustering coefficient for nodes by degree on a log-log plot with a straight line of fit. The average clustering coefficient $C(k)$ scales with k^{-1} showing that as the number of neighbours for a node increases the connections between that node's neighbours decreases, implying the network topology has a hierarchical structure.

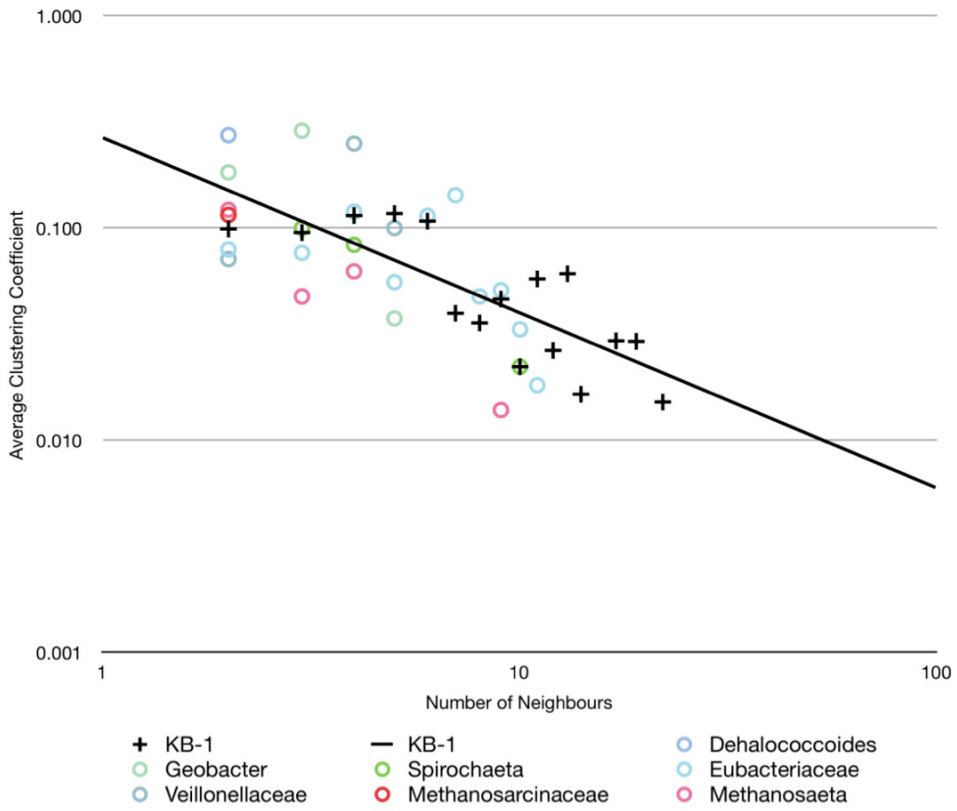


Figure 3.4 Distribution of the average clustering coefficient categorized by the number of neighbours for the KB-1 metabolic network (black). The straight line on this log-log plot is an indicator of a hierarchical structure in the network. The average clustering coefficients for seven other member networks are shown.

The in-degree and out-degree distributions for the community network are shown in Figure 3.5, along with the trend lines for the distributions of the ten member networks. These distributions are an indicator of network topology: The number of nodes decreases with increasing degree and a straight line of fit on a log-log scale suggests the distribution

follows a power law, this would indicate the network topology is scale-free in addition to being hierarchical. The ten member networks follow a similar slope as the community network but with fewer nodes and lower degree.

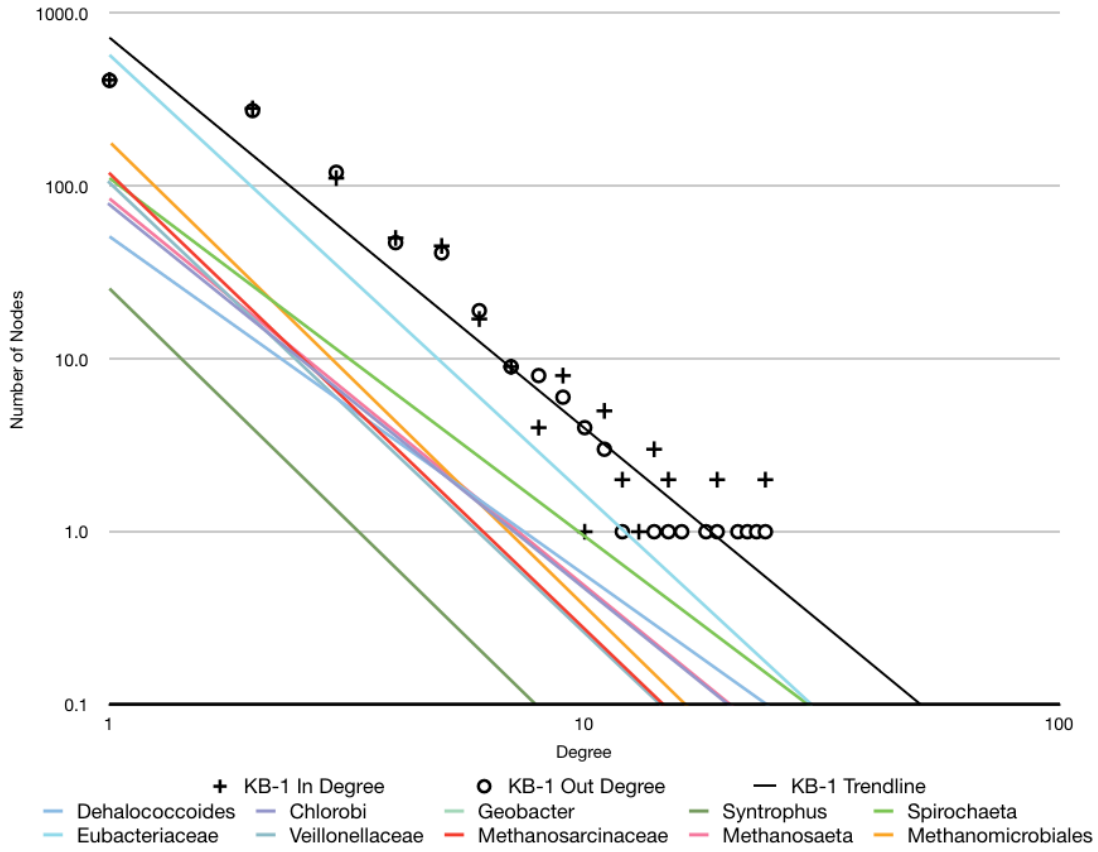


Figure 3.5 The in-degree and out-degree distributions of the KB-1 metabolic network shown on a log-log plot. The straight trend line suggests the data follows a power law distribution and the network has a scale free topology. The trend lines for the metabolic networks for each of the ten different taxa predicted in KB-1 are also shown for comparison.

To test if another model better fits the degree distribution, the netZ R package was used to calculate the maximum likelihood of five model distributions: power law, stretched exponential, Poisson, exponential, and log-normal; the results are in Table 3.4. netZ was then used to apply the Akaike-information criterion (AIC) to each likelihood to determine the best model. AIC chooses the model that best fits the distribution while favouring models with fewer parameters [52]. The AIC tests show the log-normal model best fits the data. Exponential and stretched-exponential distributions fit the KB-1 degree

distribution better than the power law. The ten member networks also best fit a log-normal distribution.

Table 3.4 Each model the KB-1 network degree distribution was fitted to, the maximum likelihood of the fit, and the Akaike weights generated by the netZ R package [77]. The AIC test uses the maximum likelihoods to choose a model that best fits the degree distribution (in this case log normal).

Model	Likelihood	log(Akaike Weight)
Power law	-1743.438	-152.30
Stretched Exponential	-1603.598	-13.46
Poisson	-1759.513	-168.38
Exponential	-1605.915	-14.78
Log normal	-1590.136	-1.80

The hierarchical topology of the KB-1 network was assessed using GLay to see if community structures within the hierarchy correlate to taxonomy or biological function. Previous studies on single-organism networks have shown that networks cluster by related biological function, identifying pathways [78; 79] and suggesting that metabolites within a pathway have a higher connectivity to each other than to metabolites in other pathways. GLay is a software package that decomposes a network into clusters based on shortest path edge betweenness removal, where at each iteration the edge involved in the greatest number of shortest paths is removed; this is repeated until a modularity score is maximized [53]. This identified 20 clusters. The reactions in each cluster were mapped to the pathway(s) they belong to. Each cluster contained between one and four dominant pathways that covered most reactions, the remaining reactions belonging to other pathways. The clusters occasionally showed preference for a particular taxonomic group, such as a mostly Firmicutes cluster containing pathways “Folate biosynthesis” and “Valine, leucine and isoleucine biosynthesis,” however some clusters were represented by a scattered collection of taxa. A cluster containing multiple taxa contained within it a subcluster of Archaeal reactions also from the “Folate biosynthesis” pathway. GLay splitting the same pathway into two clusters based on taxonomy suggests that the network hierarchy is driven by both pathway and taxonomy together. The pathways within some

clusters did not show this taxonomic cohesion. Table 3.5 shows the pathway and taxonomic cohesion for the 20 KB-1 clusters. Nine of the 20 clusters are more related by taxonomy than by pathway. Clusters 8 and 9 were described above: Cluster 8 contains the Folate biosynthesis and Valine, leucine and isoleucine biosynthesis pathways assigned mostly to Firmicutes and Cluster 9 contains mostly Folate biosynthesis reactions assigned to Archaea. The Folate biosynthesis reactions of each taxonomic group created more connectivity to other pathways in the same taxonomy than to each other, even though they are two versions of the same pathway. Cohesion therefore identifies a lack of Folate biosynthesis interconnectivity between the two taxa and may serve as an indicator of pathway interaction. Pathways in Table 3.5 are also summarized by a higher level category provided by the KEGG database (Pathway Group Cohesion) that groups related pathways. Clusters containing multiple pathways from the same group will have high pathway cohesion but low pathway group cohesion. Some clusters, such as 12, have less pathway cohesion than taxonomic cohesion, however grouping pathways into common categories increases pathway cohesion above taxonomy. These clusters suggest that while taxonomic connectivity can be greater than pathway connectivity, the connected metabolites may be still be related by a higher level of biological function.

Table 3.5 The 20 clusters identified by GLay and their pathway and taxonomic cohesion as measured by Equations 2 and 3.

Cluster	Pathway cohesion	Pathway group cohesion	Taxonomic cohesion	Number of edges (reactions)
1	0.156	0.209	0.405	92
2	0.450	0.473	0.264	42
3	0.122	0.296	0.245	97
4	0.142	0.271	0.192	187
5	0.291	0.672	0.611	9
6	0.136	0.288	0.287	124
7	0.319	0.499	0.262	175
8	0.213	0.304	0.607	53
9	0.293	0.422	0.767	71
10	0.325	0.490	0.510	26
11	0.111	0.180	0.267	127
12	0.110	0.372	0.369	233
13	0.152	0.243	0.264	66
14	0.221	0.314	0.466	31
15	0.217	0.359	0.380	19
16	0.126	0.286	0.289	132
17	0.148	0.281	0.326	61
18	0.310	0.493	0.324	37
19	0.523	0.608	0.241	18
20	0.207	0.380	0.251	19

A total of 69 hand-off points using 53 different metabolites were detected in the KB-1 network. False positives were eliminated from this list: The protein-coding sequence for all reactions producing the handoff point were retrieved from KEGG and compared to KB-1 using TBLASTN (a variant of BLAST). The BLAST bitscores of matches to KEGG were compared to the BLAST bitscore that lead to the original functional annotation for that KB-1 gene; if the KEGG-matching bitscore was equal or greater the gene function was re-assigned to that of the KEGG protein. This re-assignment meant the KB-1 protein now produces a hand-off point metabolite; if the SPANNER assignment for that gene was the same taxon that received the hand-off point, the hand-off point was considered a false positive. This reduced the set to 30 hand-off points using 18 different metabolites, shown in Table 3.6. None of these metabolites are used in the cobalamin synthesis or methionine pathways (where it is suspected *Dehalococcoides* receives hand-offs) and no hand-off points were found with *Dehalococcoides* as the recipient.

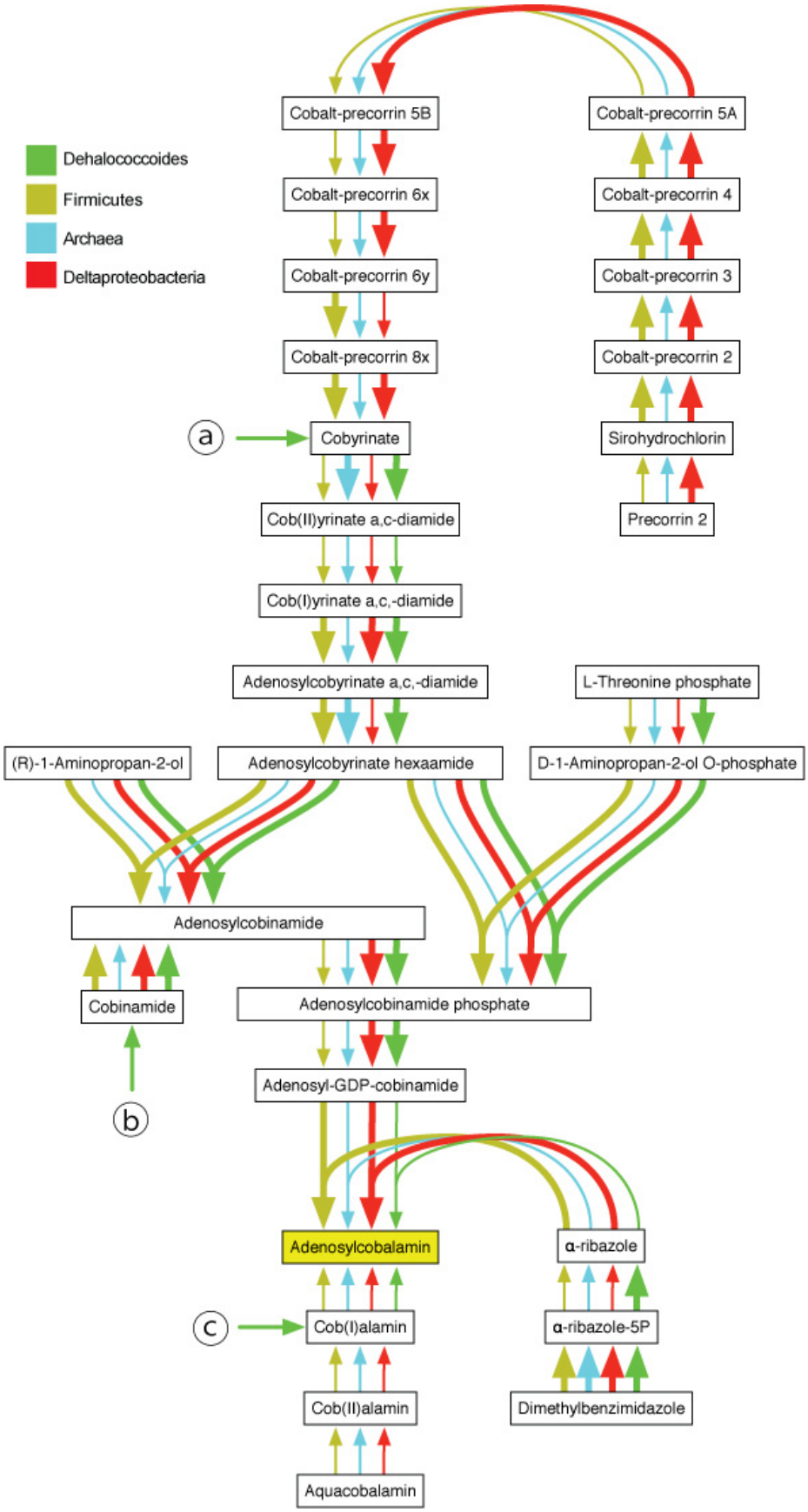
Table 3.6 Hand-off points in the KB-1 community network not ruled out as false positives.

Hand-off point	Recipient taxon
3'-AMP	Firmicutes
3'-UMP	Firmicutes
4-Aminobutanoate	Methanomicrobia
beta-Alanine	Methanomicrobia
Cytidine 3'-phosphate	Firmicutes
D-Mannitol 1-phosphate	Firmicutes
Deoxyuridine	Methanomicrobia
Folinic acid	Firmicutes
Glycerophosphocholine	Firmicutes
Glycerophosphoethanolamine	Firmicutes
Guanosine 3'-phosphate	Firmicutes
L-Histidine	Chlorobi
Manninotriose	Firmicutes
Melibiose	Firmicutes

Hand-off point	Recipient taxon
N-Acetyl-D-glucosamine 6-phosphate	Firmicutes
Selenite	Chlorobi
Trehalose 6-phosphate	Firmicutes
XTP	Methanomicrobia

To determine if hand-off point detection failed due to inaccurate functional annotation, the cobalamin pathway was reconstructed as a metabolic network. All genes for all cobalamin reactions identified in KB-1 by [62] were retrieved from KEGG and compared to the KB-1 metagenome using TBLASTN. A matched subsection of a KB-1 contig was annotated with the function of the KEGG gene if it matched at least 70% of the nucleotides. If multiple matches overlapped on a contig, the match with the highest bitscore was used. The same SPANNER taxonomic classifications as before were used. This targeted search for the cobalamin pathway found 44 reactions (shown as thick arrows in Figure 3.6); the initial functional annotation only found 18. The three cobalamin hand-off points suggested by Hug *et al.* [62] allow *Dehalococcoides* to synthesize Adenosylcobalamin from three precursors produced by other organisms and transported into the *Dehalococcoides* cell; the transport of these precursors are labeled in Figure 3.6 as **A**, **B**, and **C**. The targeted annotation of the cobalamin pathway identified hand-off points **A** and **B**, however **C** was not identified as a hand-off point because no reaction from Cob(I)alamin to Adenosylcobalamin was assigned to *Dehalococcoides*.

Figure 3.6 The cobalamin pathway, ending with the synthesis of Adenosylcobalamin (shown in yellow). The coloured edges are taxon-labeled reactions suggested by [62] to be in KB-1, these edges are thick if the were identified in the TBLASTN analysis described above. **A**, **B**, and **C** are known *Dehalococcoides* transporters.


Dehalococcoides
Firmicutes
Archaea
Deltaproteobacteria
Cobalt-precorrin 5B
Cobalt-precorrin 5A
Cobalt-precorrin 6x
Cobalt-precorrin 4
Cobalt-precorrin 6y
Cobalt-precorrin 3
Cobalt-precorrin 8x
Cobalt-precorrin 2
a
Cobyrinate
Sirohydrochlorin
Cob(II)yrinate a,c-diamide
Precorrin 2
Cob(I)yrinate a,c,-diamide
Adenosylcobyrinate a,c,-diamide
L-Threonine phosphate
(R)-1-Aminopropan-2-ol
Adenosylcobyrinate hexaamide
D-1-Aminopropan-2-ol O-phosphate
Adenosylcobinamide
Cobinamide
Adenosylcobinamide phosphate
Adenosyl-GDP-cobinamide
b
Adenosylcobalamin
α-ribazole
c
Cob(I)alamin
α-ribazole-5P
Cob(II)alamin
Dimethylbenzimidazole
Aquacobalamin

3.4 Conclusion

The KB-1 community network follows a similar structure to metabolic networks of microorganisms described previously. The KB-1 network retains the bow tie topology albeit with a larger GSC and smaller substrate and product subsets and an almost nonexistent isolated subset. This is due to the merger of all members' GSCs which connect via the hub metabolites that each member has in common; this increased pool of strongly connected metabolites increases the chances that members of an individual member's substrate, product, or isolated subsets will be added to the community GSC. The GSC represents core metabolism in the network of a single organism [43]; in a community network the larger GSC represents every member's core metabolism as well as all metabolites produced by one member's core metabolism and consumed by another's. The dominant pathways found in the KB-1 GSC were similar to those found in previous studies [78; 79]. The increased size of the GSC demonstrates that if community members shared all metabolites via hand-off points they would have access to a far wider variety of substrates.

Many metabolic networks that were first shown to be scale-free (via a power law degree distribution and hypothesized to be generated by preferential attachment) have since been shown to follow a Yule or log-normal distribution. The KB-1 metabolic network agrees with this finding and shows the need to rigorously identify the correct model of a distribution. Unfortunately no evolutionary models have yet been described that produce a log-normal distribution, although models such as Big Bang [50] offer an alternative to preferential attachment. Joining the networks of multiple organisms increases the degree but does not change the overall distribution. Clusters can still be detected within a metagenomic metabolic network, which largely correlate with biological function and to a lesser degree taxonomy. Further work is required to understand why some clusters are based on pathway and taxonomy, some on pathway only, and some on taxonomy only.

While a community-wide metabolic model can give insights to the role each member plays in metabolism, detecting details like hand-off points requires sequencing with good coverage, accurate taxonomic assignments, and accurate functional annotations. The

annotations of the KB-1 network failed to reconstruct member networks of sufficient size and detail, from these hand-off points could be detected but showed no relevance to known KB-1 interactions. Annotating function by searching the metagenome for homologs of genes in a specific pathway is more likely to identify putative reactions; a search for the cobalamin pathway allowed a better network reconstruction than a search using a broad database like Genbank. With improved pathway reconstruction expected hand-off points were detected.

CHAPTER 4 CONCLUSIONS

4.1 COMMUNITY ANALYSIS THROUGH METAGENOMICS

This thesis describes common techniques and presents new approaches for metagenomics, the exploratory analysis of environmental samples of microorganisms through DNA sequencing. This analysis typically involves taxonomic classification to determine the constituent taxa of the sample and functional annotation to describe the chemical reactions of the proteins the metagenome encodes. Each member's metabolism can be reconstructed as a network of metabolites and the reactions that process them. The KB-1 metagenome demonstrates the practical applications of this analysis: this work attempts to identify its bioremediation pathway, which taxa perform key bioremediation reactions, and the taxonomic dependencies of these community members.

4.1.1 Taxonomic Classification

Rank-flexible classification estimates the taxonomic novelty of a sequence, since a sequence can only be correctly assigned to ranks above its level of novelty; ranks at and below this level are not present in the reference database and assignment to these ranks will be incorrect. Rank-specific classification does not estimate novelty and tends to be over-specific, having high precision and many incorrect assignments. LCA tends to be under-specific, having low precision and few incorrect assignments by defining novelty as the lowest common ancestor of all taxonomies that show strong affinity to the query.

In Chapter 2 the algorithm SPANNER was proposed that uses these affinities (the "LCA Profile") as a single unit in comparisons to estimate novelty and assign taxonomy. SPANNER uses a "global" approach to taxonomic affinities by comparing each query LCA Profile against all reference Profiles. This allows SPANNER to assign a query to a taxon not actually present in the LCA Profile of a query. The classifiers described in Chapter 1 use a "local" approach to affinity, comparing the affinities in a LCA Profile to

one another. This limits query assignment to only one of the taxa contained within the LCA Profile.

Genes from closely related genomes are expected to have greater similarity in their LCA Profiles enabling classification using Profile similarity. Events such as LGT between distant organisms would push classification using LCA to higher ranks, however multiple events to the same genome would create a detectable pattern in those LCA Profiles. If this pattern is greater than the natural differences between LCA Profiles on a genome that arise due to mutation, SPANNER will detect it and make an assignment using more information than just the LCA Profile's contents, as other classifiers do. By not relying on the taxa within a LCA Profile to make an assignment, SPANNER can classify a query to a genome that is not part of its LCA Profile, an impossibility for the other classifiers. The PMK is a suitable measure for the similarity between two LCA Profiles, comparing them on their two dimensions: taxonomy and strength. By evaluating the similarity at decreasing levels of granularity the PMK is robust to minor dissimilarities between two otherwise similar LCA Profiles and naturally accommodates the hierarchical nature of taxonomic data.

4.1.2 Metabolic Networks

Combining taxonomic classifications and functional annotation allows the reconstruction of metabolic networks for each member and describes their metabolism: what chemicals a taxon processes and in what order. Network topology has been used to describe the robustness and high connectivity of metabolism [40]. Topology also guides theories of metabolic network evolution; evolutionary models have been proposed that recreate the observed topologies [49-51]. Nutritional requirements have been described via the "seed set" of a network, which is the set of substrates an organism cannot produce itself and must acquire from the environment [61].

A common approach to describing a metagenome is comparative analysis of the organisms it represents. The topologies of metabolic networks have been compared in

previous studies, finding attributes common to all microorganisms such as scale-freeness and a hierarchical modularity that forms clusters correlating to biological function [43]. If a metagenome represents an interacting community of microbes these interactions should be detectable in the reconstructed networks. Comparing the seed set of one organism to the products of another identifies possible hand-off points in which one organism provides nutrients to another; prior work showed a correlation between the seed set of parasitic microbes and the metabolic production of the hosts they feed off [61].

In Chapter 3 a method for reconstructing a metabolic network that represents an entire microbial community was proposed. This network is constructed by combining the metabolic networks of each member. This network allows topological comparisons of the community to each individual member, since the networks constructed for each member exclude community interactions, therefore representing it in isolation. Edges in the community network can be labeled with the taxon capable of performing that reaction, from this putative hand-off points between community members can be identified.

4.2 Summary of Results and Conclusions

The results of Chapter 2 show SPANNER spans the precision and incorrect ranks of LCA and best BLAST. SPANNER classified more KB-1 genes to the rank of genus than LCA. At this rank SPANNER assigned more genes to the 13 expected taxa than did LCA, however SPANNER had many more assignments than LCA to unexpected taxa. An increase in the number of incorrect assignments is a trade off for increased precision; the rank-specific classifier best BLAST has maximal precision but assigned the most genes (63%) to unexpected taxa (taxa other than the 13 expected in KB-1). Many of these unexpected taxa assignments were correct at higher ranks, highlighting the advantage of rank-flexible classifiers to choose the appropriate rank for each gene sequence. This is shown in the pseudometagenome where at the highest values for p and y SPANNER outperforms best BLAST: SPANNER chose higher ranks than best BLAST to classify some proteins which resulted in an overall decrease in incorrectly assigned ranks, and this was greater than the decrease in precision caused by choosing those higher ranks.

The accuracy of classifying a sequence depends on its taxonomic novelty. The most precise rank possible for an assignment is the rank above the novel rank; assignments to more precise ranks will be incorrect. On average SPANNER assignments reached this maximally precise rank in some organisms (*Dehalococcoides*, *Chlorobaculum*), while others averaged assignments two (*Geobacter*, *Methanoregula*) or three (*Veillonella*, *Moorella*) ranks above their maximal rank. Leave-one-out analysis shows that SPANNER spans LCA and best BLAST at all levels of taxonomic novelty, based on the parameters *p* and *y*. The optimal parameter settings for a metagenome are determined by the expected novelty of the query sequences. For example metagenomes with high novelty have no close relative in the reference database, this prevents accurate classification at lower ranks and low values for *p* and *y* should be used for conservative assignments. Metagenomes who have close relatives in the reference database are best classified at high values of *p* and *y*.

Chapter 3 reconstructs ten metabolic networks and a community network representing KB-1. The degree distribution and clustering coefficient distribution of the community network resembles other microbial networks from previous studies: The network is hierarchical and a goodness of fit test shows it is not scale-free; this is also seen in the ten member networks and networks from previous studies [52]. Previous studies have assumed network topology is an indication of the evolutionary forces acting on an organism, and while no evolutionary model has been described that generates networks with these distributions, the similar topology seen in the interacting community could mean the forces acting on it are similar. In previous studies organism networks form clusters based on biological function [54]. The community network forms clusters based on biological function or taxonomy, or a combination of the two.

Chapter 3 also proposes a method to detect hand-off points. While putative hand-off points were detected many were identified as false positives and the remainder did not correspond to any suspected KB-1 interaction. This was due to poor functional assignments which resulted in incomplete organism networks and missing reactions.

Targeted searching for the cobalamin pathway in the metagenome did find enough reactions to identify some hand-off points that were suggested previously [62], showing that with sufficient functional annotations community metabolic interactions can be detected.

4.3 Future Work

4.3.1 Taxonomic Classification

The use of LCA Profiles and their additional comparison step in SPANNER could be integrated into other rank-flexible classification algorithms. SOrt-ITEMS uses a reciprocal BLAST of the best BLAST sequence against the query and the remaining LCA Profile sequences. The lowest common ancestor is taken of all sequences with a stronger homology match than the query. SPANNER could be extended by SOrt-ITEMS by defining LCA Profiles as all reciprocal BLAST matches with a stronger homology match than the query. These LCA Profiles would then be used in the SPANNER algorithm. Both CARMA3 and SOrt-ITEMS could use taxonomic affinity (instead of homology) to make assignments by using LCA Profiles compared using the PMK instead of sequences compared using BLAST. The structure of these algorithms would otherwise be the same and this would eliminate the parameter y from SPANNER. SPANNER has shown that assignments using taxonomic affinity are more tolerant to events such as LGT and it should be tested if this improvement applies to other rank-flexible classifiers.

4.3.2 Metabolic Networks

The KB-1 topology showed similar degree and clustering coefficient distributions to the networks of each KB-1 member and networks from prior studies [43]. The objective was to show whether the interactions and dependencies within a community integrate the community's metabolism enough that evolution will treat it as a single organism or treat the member organisms differently. Such inferences were limited due to the incomplete

reconstruction of each member's network. The networks of communities known to interact should be compared to pseudo-communities of randomly chosen microbes to determine if topology correlates to interaction, and evolutionary models that produce the observed topology need to be described. The observed differences between the KB-1 community network and organism networks from previous studies, such as the larger GSC and reduced isolated subset, should provide a basis for comparing communities against other community networks as well as networks for individual organisms.

Metabolic networks are naturally hypergraphs, where hyperedges of multiple substrates connect multiple products. Metabolic networks have been built using hypergraphs but the appropriate definitions of properties such as shortest path and clustering coefficient are the subject of debate [80]. Zhou and Nakhleh [80] analyzed different versions of the same network using the proposed hypergraph definitions. They found that different definitions produce different degree and clustering coefficient distributions, and that these distributions are different between multigraphs and hypergraphs of the same network, making them products of graph representation and casting doubt their ability to describe fundamental metabolic evolution attributes. Before claims can be made on the evolution of a metabolic network, work is needed to develop a suitable network representation.

The composition of community network clusters could provide insights into the differences and connectivity of pathways between members and needs further analysis. Single organism networks tend to cluster based on related biological function; in community networks clusters are based on related function, taxonomy, or a combination of the two. Many pathways are found in more than one organism in the community; it is possible that different versions of the same pathway will exist as each organism implements the pathway through a different set of reactions. The composition of a cluster measures pathway redundancy or complementarity between organisms: Versions of the same pathway in different organisms are redundant if their reactions use the same substrate and product (they "overlap" in the community network forming multi-edges) and are complementary if they have a high degree of interconnectivity (each reaction uses different substrates and products but the same substrates and products are found in each

version; each version increasing connectivity within the pathway). Both redundancy and complementarity increase the likelihood versions from different organisms will be clustered together, causing high pathway cohesion. If however a pathway's connectivity to other pathways from the same organism is greater than the connectivity to versions of the same pathway from other organisms, pathways will cluster based on taxonomy and have high taxonomic cohesion. This measure of pathway redundancy/synergy between organisms should be explored.

The reconstructed networks for each KB-1 member suffered from incomplete functional annotations. This weakened the confidence in the analysis performed in Chapter 3, although expected topological features were still detected. The process of annotating function using best BLAST is error-prone and needs development. Hand-off point interactions between community members were detected although on further inspection many were ruled out as false positives and none of the remaining hand-off points were those suggested in KB-1 by previous work. By definition a hand-off point could be erroneously detected for any missing reaction, whether due to poor sequencing coverage or incorrect annotations: if a single reaction is missing in an otherwise complete pathway it will be indentified as a hand-off point. Methods have been published that parsimoniously identify unannotated reactions by searching for "gaps" in pathways [81]; this approach could potentially decrease the misidentification rate without requiring improvements to functional annotation.

In a similar fashion to how the reduced reference database improved SPANNER results, identifying function using a reduced reference database (e.g. the cobalamin pathway in Chapter 3) improves the chance of finding homologous reactions. This could be because there are many similar sequences in Genbank that confuse annotation. More complete reference annotation databases would help correct this problem, although annotation methods beyond the standard "best BLAST" are needed for reliable hand-off point detection. Hug *et al.* [62] used the online analysis pipeline MG-RAST [82] to annotate KB-1. MG-RAST annotates metagenomes by identifying subsystems (manually curated sets of related biological function, i.e. a pathway) likely to be found in the metagenome,

this guides construction of a taxonomically reduced reference database. All subsystems known to exist in the genomes in this reduced database are searched for to annotate the metagenome, any remaining unannotated protein sequences are searched against all subsystems from all reference genomes. This subsystem approach to annotation would provide better functional assignments and reduce the number of false positive hand-off points.

Hand-off point identification has the assumption that any organism's products can be released from its cell into the environment and acquired by another organism. False positive hand-off points could be filtered with membrane permeability information and identifying transporters in the metagenome, since a hand-off point must be acquired by a cell any metabolites that are not permeable or transported are unlikely candidates.

REFERENCES

1. Scott CS, Cogliano VJ: **Trichloroethylene health risks--state of the science.** *Environ Health Perspect.* 2000, **108**(Suppl 2):159-160.

2. Major DW, McMaster ML, Cox EE, Edwards EA, Dworatzek SM, Hendrickson ER, Starr MG, Payne JA, Buonamici LW: **Field demonstration of successful bioaugmentation to achieve dechlorination of tetrachloroethene to ethene.** *Environmental Science & Technology*. 2002, **36**(23):5106-16.

3. Steffan R, Schaefer C, Lippencott D: **Bioaugmentation for Groundwater Remediation.** Prepared by Shaw Environmental and Infrastructure, Inc. for ESTCP. 2010, **Project ER-200515**, Final Report.

4. Ahsanul Islam M, Edwards EA, Mahadevan R: **Characterizing the Metabolism of Dehalococcoides with a Constraint-Based Model.** *PLoS Comput Biol.* 2010, **6**(8):e1000887.

5. Shendure J, Mitra RD, Varma C, Church GM: **Advanced sequencing technologies: methods and goals.** *Nature Reviews Genetics.* 2004, **5**:335-344.

6. Flicek P, Birney E: **Sense from sequence reads: methods for alignment and assembly.** *Nature Methods.* 2009, 6:S6-S12.

7. Hooper SD, Dalevi D, Pati A, Mavromatis K, Ivanova NN, Kyrpides NC: **Estimating DNA coverage and abundance in metagenomes using a gamma approximation**. *Bioinformatics*. 2010, **26**(3):295-301.

8. Zhu W, Lomsadze A, Borodovsky M: ***Ab initio* gene identification in metagenomic sequences.** *Nucleic Acids Research* 2010, **38**: e132.

9. Wooley JC, Godzik A, Friedberg I: **A Primer on Metagenomics.** *PLoS Comput Biol.* 2010, **6**(2):e1000667.

10. Chou H-H, Holmes MH: **DNA sequence quality trimming and vector removal.** *Bioinformatics.* 2001, **17**: 1093-1104.

11. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BØ: **A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information.** *Mol Syst Biol*. 2007, 3:121.

12. Schnoes AM, Brown SD, Dodevski I, Babbitt PC: **Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies.** *PLoS Comput Biol* 2009, 5(12):e1000605.

13. Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucleic Acids Res.* 2000, 28:27-30.

14. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y: **KEGG for linking genomes to life and the environment.** *Nucl. Acids Res.* 2008, 36(suppl 1):D480-D484

15. Ross O, Tadhg B, Ralph MB, Jomuna VC, Han-Yu C, Matthew C, Valérie dC, Naryttza D, Terry D, Robert E, Michael F, Ed DF, Svetlana G, Elizabeth MG, Alexander G, Andrew H, Dirk I, Roy J, Neema J, Lutz K, Michael K, Niels L, Burkhard L, Alice CM, Folker M, Heiko N, Gary O, Robert O, Andrei O, Vasiliy P, Gordon DP, Dmitry AR, Christian R, Jason S, Rick S, Ines T, Olga V, Yuzhen Y, Olga Z, Veronika V: **The Subsystems Approach to Genome Annotation and its Use in the Project to Annotate 1000 Genomes.** *Nucl. Acids Res.* 2005, 33(17):5691-5702.

16. Schellenberger J, Park JO, Conrad TC, Palsson BØ: **BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions.** *BMC Bioinformatics.* 2010, 11:213.

17. UniProt Consortium, The: **Reorganizing the protein space at the Universal Protein Resource (UniProt).** *Nucleic Acids Res.* 2012, 40:D71-D75.

18. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank.** *Nucleic Acids Res.* 2011, 39(Database issue):D32-7.

19. Smith TF, Waterman MS: **Identification of common molecular subsequences.** *J Mol Biol.* 1981, **147**(1):195-7.

20. Altschul SF, Thomas LM, Alejandro AS, Jinghui Z, Zheng Z, Webb M, David JL: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res.* 1997, **25:**3389-3402.

21. Ng PC and Henikoff S: **Predicting the effects of amino acid substitutions on protein function**. *Annu Rev Genomics Hum Genet*. 2006, **7**:61-80.

22. Petrosino JF, Highlander S, Luna RA, Gibbs RA, Versalovic J: **Metagenomic Pyrosequencing and Microbial Identification.** *Clinical Chemistry.* 2009, **55**(5): 856-866.

23. Mayr E: **Systematics and Origin of Species.** New York: Columbia University Press. 1942.

24. Belkum A van, Struelens M, Visser A de, Verbrugh H, Tibayrenc M: **Role of Genomic Typing in Taxonomy, Evolutionary Genetics, and Microbial Epidemiology.** *Clin. Microbiol. Rev.* 2001, **14**(3): 547-560.

25. Doolittle WF: **Phylogenetic Classification and the Universal Tree.** *Science*. 1999, **284**(5423): 2124-2128.

26. Vandamme P, Pot B, Gillis M, de Vos P, Kersters K, Swings J: **Polyphasic taxonomy, a consensus approach to bacterial systematics.** *Microbiol Rev.* 1996, **60**(2): 407-38.

27. Thomas CM, Nielsen KM: **Mechanisms of, and Barriers to, Horizontal Gene Transfer between Bacteria.** *Nature Reviews Microbiology.* 2005, **3**: 711-721.

28. Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris BJ, Kulam-Syed-Mohideen AS, McGarrell DM, Marsh T, Garrity GM, Tiedje JM. **The Ribosomal Database Project: improved alignments and new tools for rRNA analysis.** *Nucleic Acids Res.* 2008, 37 (Database issue):D141-145.

29. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL: **Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB**. *Appl Environ Microbiol.* 2006, **72**(7): 5069-72.

30. Gevers D, Cohan FM, Lawrence JG, Spratt BG, Coenye T, Feil EJ, Stackebrandt E, Peer YV de, Vandamme P, Thompson FL, Swings J: **Re-evaluating prokaryotic species.** *Nature Reviews Microbiology.* 2005, **3**:733-739.

31. Duhamel M, Edwards EA: **Microbial composition of chlorinated ethene-degrading cultures dominated by Dehalococcoides.** *FEMS Microbiology Ecology* 2006. **58**:538-549.

32. Kumar MR and Saravanancorresponding VS: **Candidate OP Phyla: Importance, Ecology and Cultivation Prospects**. *Indian J Microbiol*. 2010, **50**(4): 474–477.

33. Parks D, MacDonald N, Beiko R: **Classifying short genomic fragments from novel lineages using composition and homology.** *BMC Bioinformatics* 2011, **12**: 328.

34. Perry SC and Beiko RG: **Distinguishing Microbial Genome Fragments Based on Their Composition: Evolutionary and Comparative Genomic Perspectives**. *Genome Biol Evol*. 2010, **2**:117–131.

35. Rosen G, Garbarine E,Caseiro D, Polikar R, Sokhansanj B: **Metagenome Fragment Classification Using N-Mer Frequency Profiles.** *Advances in Bioinformatics.* 2008, **2008**:205969.

36. MacDonald NJ, Parks DH, Beiko RG: **RITA: Rapid identification of high-confidence taxonomic assignments for metagenomic data**. *Nucleic Acids Res*. 2012.

37. Huson H, Auch A, Qi J, Schuster SC: **MEGAN Analysis of Metagenomic Data.** *Genome Research* 2007, **17**(3): 377-386.

38. Monzoorul HM, Ghosh TS, Komanduri D, Mande SS: **SOrt-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences.** *Bioinformatics* 2009, **25**:1722-1730.

39. Gerlach W and Stoye J: **Taxonomic classification of metagenomic shotgun sequences with CARMA3**. *Nucleic Acids Res*. 2011, **39**(14):e91.

40. Yamada T, Bork P: **Evolution of biomolecular networks: lessons from metabolic and protein interactions.** *Nat Rev Mol Cell Biol.* 2009, **10**(11):791-803.

41. Horne AB, Hodgman TC, Spence HD, Dalby AR: **Constructing an enzyme-centric view of metabolism**. *Bioinformatics*. 2004, **20**(13):2050-5.

42. Cottret L, Jourdan F: **Graph methods for the investigation of metabolic networks in parasitology**. *Parasitology*. 2010, **137**(9):1393-407.

43. Ma HW, Zeng A-P: **The connectivity structure, giant strong component and centrality of metabolic networks**. *Bioinformatics.* 2003, **19**(11):1423-1430.

44. Mithani A, Preston GM, Hein J: **Rahnuma: hypergraph-based tool for metabolic pathway prediction and network comparison**. *Bioinformatics*. 2009, **25**(14):1831-1832.

45. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási AL: **Hierarchical organization of modularity in metabolic networks.** *Science.* 2002, **297**(5586):1551-5.

46. Barabási A-L , Albert R. **Emergence of scaling in random networks**. *Science.* 1999, 286:509-12.

47. Albert, R., Jeong, H. & Barabási, A.-L. **The Internet's Achilles' heel: Error and attack tolerance of complex networks.** *Nature*. 2000, **406**(2000):200-0.

48. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási AL: **The large-scale organization of metabolic networks**. *Nature*. 2000, **407**(6804):651-4.

49. Light S, Kraulis P, Elofsson A: **Preferential attachment in the evolution of metabolic networks**. *BMC Genomics*. 2005, **6**:159.

50. Dokholyan NV, DeLisi C, Shakhnovich B, Shakhnovich EI: **Expanding protein universe and its origin from the biological Big Bang**. *Proceedings of National Academy of Science*. 2002, **99**:14132-14136.

51. Przytycka TM, Yu YK: **Scale-free networks versus evolutionary drift**. *Comput Biol Chem*. 2004, **28**(4):257-64.

52. Stumpf M, Ingram P, Nouvel I, Wiuf C: **Statistical model selection methods applied to biological networks**. *Trans Comp Sys Biol*. 2005, **3**:65–77.

53. Newman MEJ, Girvan M: **Finding and evaluating community structure in networks**. *Physical Review E*. 2003, **69**(2).

54. Su G, Kuchinsky A, Morris JH, States DJ, Meng F: **GLay: community structure analysis of biological networks**. *Bioinformatics*. 2010, **26**(24):3135-7.

55. Holme P, Huss M, Jeong H: **Subnetwork hierarchies of biochemical pathways**. *Bioinformatics*. 2003, **19**(4):532-8.

56. Ma HW, Zhao XM, Yuan YJ, Zeng AP: **Decomposition of metabolic network into functional modules based on the global connectivity structure of reaction graph**. *Bioinformatics*. 2004, **20**(12):1870-6.

57. Konopka A: **What is microbial community ecology?** *ISME J*. 2009, **3**(11):1223-30.

58. Morris JJ, Lenski RE, Zinser ER: **The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss.** *MBio.* 2012, **3**(2): e00036-12.

59. Wu D, Daugherty SC, Van Aken SE, Pai GH, Watkins KL, Khouri H, Tallon LJ, Zaborsky JM, Dunbar HE, Tran PL, Moran NA, Eisen JA: **Metabolic Complementarity and Genomics of the Dual Bacterial Symbiosis of Sharpshooters.** *PLoS Biol.* 2006, **4**(6): e188.

60. Raymond J, Segrè D: **The effect of oxygen on biochemical networks and the evolution of complex life**. *Science*. 2006, **311**(5768):1764-7.

61. Borenstein E, Feldman MW: **Topological Signatures of Species Interactions in Metabolic Networks**. *J Comput Biol*. 2009, **16**(2):191–200.

62. Hug LA, Beiko RG, Rowe AR, Richardson RE, Edwards EA: **Comparative metagenomics of three Dehalococcoides-containing enrichment cultures: the role of the non-dechlorinating community**. *In submission*. 2012.

63. Amos BK, Ritalahti KM, Cruz-Garcia C, Padilla-Crespo E, Löffler FE: **Oxygen effect on Dehalococcoides viability and biomarker quantification**. *Environ Sci Technol*. 2008, **42**(15):5718-26.

64. Ahsanul Islam M, Edwards EA, Mahadevan R: **Characterizing the metabolism of Dehalococcoides with a constraint-based model**. *PLoS Comput Biol*. 2010, **6**(8).

65. Grauman K, Darrell T: **The Pyramid Match Kernel: Efficient Learning with Sets of Image Features.** *Journal of Machine Learning Research* 2007, **8**:725-760.

66. Brady A, Salzberg SL: **Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models.** *Nature Methods* 2009, **6:**673-676.

67. Sokal R and Michener C: **A statistical method for evaluating systematic relationships.** *University of Kansas Science Bulletin.* 1958, **38**:1409-1438.

68. Hess M, Sczyrba A, Egan R, Kim TW, Chokhawala H, Schroth G, Luo S, Clark DS, Chen F, Zhang T, Mackie RI, Pennacchio LA, Tringe SG, Visel A, Woyke T, Wang Z, Rubin EM: **Metagenomic discovery of biomass-degrading genes and genomes from cow rumen.** *Science* 2011, **331**: 463-467.

69. Liu B, Gibbons T, Ghodsi M, Pop M: **MetaPhyler: Taxonomic profiling for metagenomic sequences**. *Proceedings of 2010 IEEE Bioinformatics and Biomedicine*. 2010, 95-100.

70. Al-Awqati Q: **One hundred years of membrane permeability: does Overton still rule?** *Nat Cell Biol*. 1999, **1**(8):E201-2.

71. Webb EC: **Enzyme nomenclature 1992: recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes**. San Diego: Academic Press. Published for the International Union of Biochemistry and Molecular Biology. 1992.

72. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T: **Cytoscape 2.8: new features for data integration and network visualization**. *Bioinformatics*. 2011, **27**(3):431-2.

73. Assenov Y, Ramírez F, Schelhorn SE, Lengauer T, Albrecht M: **Computing topological parameters of biological networks**. *Bioinformatics*. 2008, **24**(2):282-4.

74. Zinovyev A, Viara E, Calzone L, Barillot E: **BiNoM: a Cytoscape plugin for manipulating and analyzing biological networks**. *Bioinformatics*. 2008, **24**(6):876-877.

75. Scardoni G, Petterlini M, Laudanna C: **Analyzing biological network parameters with CentiScaPe**. *Bioinformatics*. 2009, **25**(21):2857–2859.

76. Morris JH, Apeltsin L, Newman AM, Baumbach J, Wittkop T, Su G, Bader GD, Ferrin TE: **clusterMaker: a multi-algorithm clustering plugin for Cytoscape**. *BMC Bioinformatics*. 2011, **12**:436.

77. Kelly WP, Ingram PJ, Stumpf MP: **The degree distribution of networks: statistical model selection**. *Methods Mol Biol.* 2012, **804**:245-62.

78. Zhao J, Yu H, Luo JH, Cao ZW, Li YX: **Hierarchical modularity of nested bow-ties in metabolic networks**. *BMC Bioinformatics*. 2006, **18**(7):386.

79. Ding DW, Ding YR, Li LN, Cai YJ, Xu WB: **Structural and Functional Analysis of Giant Strong Component of *Bacillus thuringiensis* Metabolic Network**. *Braz. j. microbiol*. 2009, **40**(2):411-416.

80. Zhou W and Nakhleh L: **Properties of metabolic graphs: biological organization or representation artifacts?** *BMC Bioinformatics*. 2011, **12**:132.

81. Ye Y, Doak TG: **A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes**. *PLoS Comput Biol*. 2009, **5**(8):e1000465.

82. Mitra S, Rupek P, Richter DC, Urich T, Gilbert JA, Meyer F, Wilke A, Huson DH: **Functional analysis of metagenomes and metatranscriptomes using SEED and KEGG**. *BMC Bioinformatics*. 2011, **12**(Suppl 1):S21.