2009

Inference, Orthology, and Inundation: Addressing Current Challenges in the Field of Metagenomics

Gregory Detlev Alexander Vey
Wilfrid Laurier University

Your file Votre référence
ISBN: 978-0-494-54249-1
Our file Notre référence
ISBN: 978-0-494-54249-1

Inference, Orthology, and Inundation: Addressing Current Challenges in the Field of Metagenomics

by

Gregory Detlev Alexander Vey

Bachelor of Arts, University of Western Ontario, 1997

Bachelor of Science, Wilfrid Laurier University, 2007

Thesis

Submitted to the Department of Biology

Faculty of Science

in partial fulfillment of the requirements for the

Master of Science in Integrative Biology

Wilfrid Laurier University

2009

Abstract

The vast increase in the number of sequenced genomes has irreversibly changed the landscape of the biological sciences and has spawned the current post-genomic era of research. Genomic data have illuminated many adaptation and survival strategies between species and their habitats. Moreover, the analysis of prokaryotic genomic sequences is indispensible for understanding the mechanisms of bacterial pathogens and for subsequently developing effective diagnostics, drugs, and vaccines. Computational strategies for the annotation of genomic sequences are driven by the inference of function from reference genomes. However, the effectiveness of such methods is bounded by the fractional diversity of known genomes. Although metagenomes can reconcile this limitation by offering access to previously intangible organisms, harnessing metagenomic data comes with its own collection of challenges. Since the sequenced environmental fragments of metagenomes do not equate to discrete and fully intact genomes, this prevents the conventional establishment of orthologous relationships that are required for functional inference. Furthermore, the current surge in metagenomic data sets requires the development of compression strategies that can effectively accommodate large data sets that are comprised of multiple sequences and a greater proportion of auxiliary data, such as sequence headers. While modern hardware can provide vast amounts of inexpensive storage for biological databases, the compression of nucleotide sequence data is still of paramount importance in order to facilitate fast search and retrieval operations through a reduction in disk traffic. To address the issues of inference and orthology a novel protocol was developed for the prediction of functional interactions

that supports data sources that lack information about orthologous relationships. To address the issue of database inundation, a compression protocol was designed that can differentiate between sequence data and auxiliary data, thereby offering reconciliation between sequence specific and general-purpose compression strategies. By resolving these and other challenges, it becomes possible to extend the potential utility of the emerging field of metagenomics.

Co-authorship

- Vey G, Moreno-Hagelsieb G: **Beyond the bounds of orthology: functional inference from metagenomic context.** *Molecular BioSystems*, under review.

  Gregory Vey contributed to the development of this manuscript by performing the following;

  1. development of computational tools.
  2. data processing and analysis.
  3. co-creation of the text, figures, and tables.

- Vey G: **Differential direct coding: a compression algorithm for nucleotide sequence data.** *Database: The Journal of Biological Databases and Curation* 2009, Vol. 2009:bap013; doi:10.1093/database/bap013.

  Gregory Vey developed all components of this manuscript including;

  1. conception of the manuscript topic.
  2. background research and literature review.
  3. development and implementation of algorithm.
  4. data processing and analysis.
  5. creation of the text, figures, and tables.

Acknowledgements

First and foremost I thank my supervisor Dr. Gabriel Moreno-Hagelsieb for guidance and mentorship during my graduate studies, and for his contributions toward the development of this thesis. I also thank the members of my advisory committee, Dr. Angèle Hamel and Dr. Matthew Smith, for their guidance and feedback on my research efforts. I thank Dr. Juan Javier Díaz-Mejía for ongoing discussions about various topics related to my research. I thank Dr. Frédérique Guinel for her assistance with administrative issues and Mélanie Lafrance for her help with word processing portions of the manuscripts. Finally, I thank the Department of Biology, my course instructors, and my fellow graduate students for their support and encouragement during the course of my graduate studies. Additional specific acknowledgements are included in their respective manuscripts. This work was funded by a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant to Dr. Gabriel Moreno-Hagelsieb and by scholarships and other funding from Wilfrid Laurier University.

Table of Contents

Abstract ii

Co-authorship iv

Acknowledgements v

Table of Contents vi

List of Tables ix

List of Figures & Illustrations x

Chapter 1 General Introduction 1

1.1 The Rise of Functional Genomics 1

1.2 Beyond the Limitations of Genomic Data 2

1.3 Current Challenges in the Field of Metagenomics 4

1.3.1 Phylogenetic classification of sequence fragments 4

1.3.2 Functional inference in the absence of orthology 5

1.3.3 Compression for large heterogeneous data sets 6

1.4 Figures 8

Chapter 2 Beyond the Bounds of Orthology: Functional Inference from Metagenomic Context 10

2.1 Abstract 11

2.2 Introduction 12

2.3 Results and Discussion 15

2.3.1 Baseline predictions network 15

2.3.2 Prediction reliability metrics 16

2.3.3 Filtered predictions network 18

2.3.4 Contribution of the metagenome 19
2.4 Conclusions 20
2.4.1 Beyond orthology 20
2.4.2 Metagenomic functional inference 21
2.5 Methods 21
2.5.1 Data sources 21
2.5.2 Prediction generation phase 22
2.5.3 Prediction mapping phase 22
2.5.4 Prediction reduction phase 23
2.6 Acknowledgements 24
2.7 References 24
2.8 Figure Legends 31
2.9 Tables 32
2.10 Additional Files 36
2.11 Figures 37
Chapter 3 Differential Direct Coding: A Compression Algorithm for Nucleotide Sequence Data 43
3.1 Abstract 44
3.2 Introduction 45
3.3 Nucleotide Sequence Compression Strategies 47
3.3.1 Evolving models 47
3.3.2 Direct coding 48
3.4 Differential Direct Coding (2D) 49

3.4.1 Objectives 49
3.4.2 Model 51
3.4.3 Coding 52
3.4.4 Algorithm 53
3.4.5 Compression ratio 55
3.4.6 Benchmarking 56
3.5 Conclusion 59
3.6 Funding 60
3.7 Acknowledgments 60
3.8 References 60
3.9 Figures 63
3.10 Tables 63
3.11 Additional Files 66
Chapter 4 General Discussion 67
4.1 Contributions to the Field of Metagenomics 67
4.1.1 Functional inference from metagenomic context 67
4.1.2 Differential direct coding 68
4.2 Future Research Directions 69
4.3 Toward a Post-metagenomic Era 71
Literature Cited 73

List of Tables

Table 2-1 Baseline functional interaction networks 32

Table 2-2 Effects of data preparation variables 33

Table 2-3 Filtered functional interaction networks 34

Table 2-4 Gain in functional interactions from combined sets 35

Table 3-1 The 2D data model 63

Table 3-2 The 2D encoding process 63

Table 3-3 Genomic compression benchmarking 64

Table 3-4 Genomic decompression benchmarking 65

Table 3-5 Metagenomic compression benchmarking 65

List of Figures & Illustrations

Figure 1-1 Growth rate of the GenBank sequence database 8

Figure 1-2 Genome projects versus metagenome projects 9

Figure 2-1 The problem of paralogy 37

Figure 2-2 Relative frequencies of correlation of expression values 38

Figure 2-3 Target intergenic distance versus positive predictive value 39

Figure 2-4 Source interaction count versus positive predictive value 40

Figure 2-5 Functional interaction network for the *E. coli* K12 MG1655 genome 41

Figure 2-6 Relative frequencies of correlation of expression values 42

Figure 3-1 The 2D byte coding schema 63

Chapter 1

General Introduction

The vast increase in the number of sequenced genomes has irreversibly changed the landscape of the biological sciences and has spawned the current post-genomic era of research. Genomic data have illuminated many adaptation and survival strategies between species and their habitats [1.1]. Moreover, the analysis of prokaryotic genomic sequences is indispensible for understanding the mechanisms of bacterial pathogens and for subsequently developing effective diagnostics, drugs, and vaccines [1.1]. With the advent of techniques to capture various microbial communities including freshwater, marine, subterranean, intestinal, and many other previously uncharacterized environments, the field of genomics rests at the forefront of a new generation of computationally driven biological sciences.

1.1 The Rise of Functional Genomics

Toward the end of the last millennium, a large increase in the number of sequenced genomes began to emerge with the total amount of sequenced DNA doubling at a rate of roughly every 18 months [1.2] (see Figure 1.1). However, this influx of data did not initially equate to an immediate increase in knowledge about proteins and their respective functions [1.3]. Researchers were faced with the task of transforming this vast repository of sequences into meaningful interpretations, thereby giving rise to the field of functional genomics [1.3].

Traditionally, knowledge about proteins has been acquired experimentally on the basis of biochemical, genetic, or structural properties [1.3]. However, conducting such

approaches on a genomic scale poses high costs combined with difficult and time intensive procedures [1.4]. This is compounded by the fact that different experimental methods provide minimal agreement in a comparison of their determinations of function [1.4]. In an effort to overcome these limitations, the field of functional genomics relies on computational procedures that attempt to infer functional relationships among the complete set of proteins encoded by a given organism [1.5]. As a result, computational approaches to inference have evolved as powerful tools to aid in the classification of hypothetical proteins and the assignment of functional annotations to newly sequenced genomes.

1.2 Beyond the Limitations of Genomic Data

A fundamental aspect of functional inference is that it relies on the current body of sequence information as a primary data source [1.5]. Therefore, its efficacy is largely constrained by the quality and representativeness of available sequence databases. Until recently, much of what had been deposited in sequence databases was data from microorganisms that are amenable to culturing [1.6-1.8]. However, it has been estimated that more than 99% of microorganisms are not culturable [1.6-1.8]. Furthermore, even among the culturable microorganisms there may exist additional biases that have resulted from the potential applications gained by studying certain categories of microbes, as illustrated in Figure 1.2, panel A. Consequently, the degree of database completion combined with compositional biases can impact both functional assignments and taxonomic classifications [1.9]. In fact, it has been recently demonstrated that the resulting taxonomic assignments for a set of open reading frames have clearly changed

over time and in conjunction with the growth of the GenBank non-redundant protein database [1.9]. Thus, capturing a greater sample of biodiversity has the capability of reducing the biases contained in existing sequence databases by extending the repertoire of known genes and known functions [1.6]. This will subsequently benefit both functional assignments and taxonomic classifications.

Metagenomics can be regarded as stemming from conventional microbial genomics but without requiring pure cultures for sequencing [1.10]. Instead, it involves the sequencing of heterogeneous samples of DNA that contain a variety of genomic sources, rather than a single target organism [1.7]. The benefit of this approach is that it provides access to previously intangible organisms and environments [1.7] (see Figure 1.2, panel B). For example, environmental microbes are typically not able to grow in pure culture and symbionts and obligate pathogens cannot survive outside of their hosts [1.7]. Therefore, DNA from such organisms can be extracted directly from them while in their natural habitats as a heterogeneous mixture of DNA that can be fragmented into a library of sequence data [1.7]. In turn, this data can provide insight into various systems, like the species dynamics among the organisms of particular environments [1.7]. Perhaps most importantly, the availability of metagenomic data sets offer a means to reconcile the current limitations of functional genomics by vastly extending the amount of usable sequence data. However, the effective harnessing of metagenomic data comes with its own collection of challenges.

1.3 Current Challenges in the Field of Metagenomics

1.3.1 Phylogenetic classification of sequence fragments

Understanding the taxonomic composition of the microbial community that comprises a particular metagenomic data set is essential for studying individual populations and their respective interactions [1.11, 1.12]. Sequence reads generated from metagenomic samples are assembled into scaffolds where the average length is affected by factors like the number of distinct populations present and their relative abundance in the sample, and also the size and architecture of the individual genomes [1.11, 1.12]. Thus, scaffold length typically decreases with increasing community complexity [1.11, 1.12]. Since this greatly reduces the likelihood of recovering complete genomic entities, methods have been developed to assign individual sequence fragments to populations or higher-level clades [1.11, 1.12].

Universally present markers such as rRNA can be used to construct phylogenies that can be subsequently applied to make taxonomic assignments to individual sequence fragments [1.12]. Another approach is to use homologs retrieved from database searches for the assignment of fragments [1.12]. However, the previously discussed bias in the databases toward cultivable organisms raises concerns about the effectiveness of this approach, particularly with respect to assignments for novel organisms [1.12]. An alternative method is to use oligomer frequencies to classify sequence fragments based on their genome sequence composition [1.12]. While all of these methods can be used to make reliable phylogenetic classifications for sufficiently long sequence fragments, none of them can provide confident assignments for fragments shorter than 1000 base pairs

[1.12]. Therefore, the development of a method that would increase the proportion of assignments for short fragments, such as pyrosequencing reads, would represent a major breakthrough for the phylogenetic classification of sequence fragments [1.12].

1.3.2 Functional inference in the absence of orthology

As soon as the publication of a sufficient number of genomes first allowed for testing, methods were proposed to infer functional interactions by genomic context [1.3, 1.13]. These methods are dependent on the establishment of orthology, which is the condition of homology resulting from a speciation event [1.14, 1.15], which differs from paralogy, the condition of homology resulting from a duplication event [1.14, 1.15]. The three main methods used to infer functional interactions involve finding: (a) Gene fusions [1.16, 1.17], where two genes are assumed to interact if their orthologs are fused into a gene coding for a multidomain protein in another genome; (b) Conservation of gene order [1.18], where the conservation of adjacent orthologs, beyond expectations by chance, provides a clue for a functional interaction; and (c) Phylogenetic profiles [1.19-1.21], where the orthologs to genes coding for functionally interacting proteins are expected to co-occur; in other words, be both present or both absent across genomes. An additional method of functional inference exploits methods to predict operons. A functional interaction is inferred if the genes themselves, or their orthologs, are found to be in the same operon [1.22-1.25].

In order to extend the computational inference of functional associations by genomic context, the use of metagenomic sequences could be included to complement the functional associations predicted for a genome of interest. Although functional

inference from metagenomic context offers an invaluable means to exceed the current limitations of functional genomics, it poses an inherent challenge with respect to orthology. The inability to distinguish the particular types of homologies within a metagenome stems from the fact that the environmental sequence fragments do not equate to complete and discrete genomic entities. However, conventional approaches to functional inference are dependent on the detection of orthology. Since comparisons between genes in a metagenome are confined to a consideration of only the general case of homology, rather than specific orthology, many spurious functional inferences can arise due to the presence of paralogs because they possess homologous sequences but potentially divergent functions [1.26]. These extraneous inferences add noise to the predictions and are evident in the form of false positives upon validation of the overall set of predicted functional interactions. Although the previously discussed methods to classify sequence fragments may shed light on establishing orthology by way of phylogenetic classification, the creation of a protocol that is not limited to using known orthologous data would represent a major step toward permitting functional inference from metagenomic context.

1.3.3 Compression for large heterogeneous data sets

In recent years, metagenomic data sets derived from environmental shotgun sequence data have gained a position of increased prominence in many sequence repositories. In fact, the sheer volume of metagenomic sequence data has exceeded the combined total of the microbial genomes [1.27]. While modern hardware can provide vast amounts of inexpensive storage for biological databases, the compression of

metagenomic sequence data is still of paramount importance in order to facilitate fast search and retrieval operations through a reduction in disk traffic.

To accommodate this surge in volume, compression strategies must be developed to accommodate large-scale data sets that are comprised of multiple sequences and a greater proportion of auxiliary data, such as sequence headers. Compression protocols developed specifically for sequence data offer good compression ratios but may perform poorly on large data sets or data sets that contain a significant amount of auxiliary data. In comparison, general-purpose compression utilities can easily compress large heterogeneous data files but cannot take advantage of the predominantly limited range of symbols that occur in sequence data. Thus, the development of a protocol that could offer reconciliation between sequence-specific and general-purpose compression strategies would have a beneficial impact on the management and processing of large heterogeneous data sets, such as metagenomes.

1.4 Figures

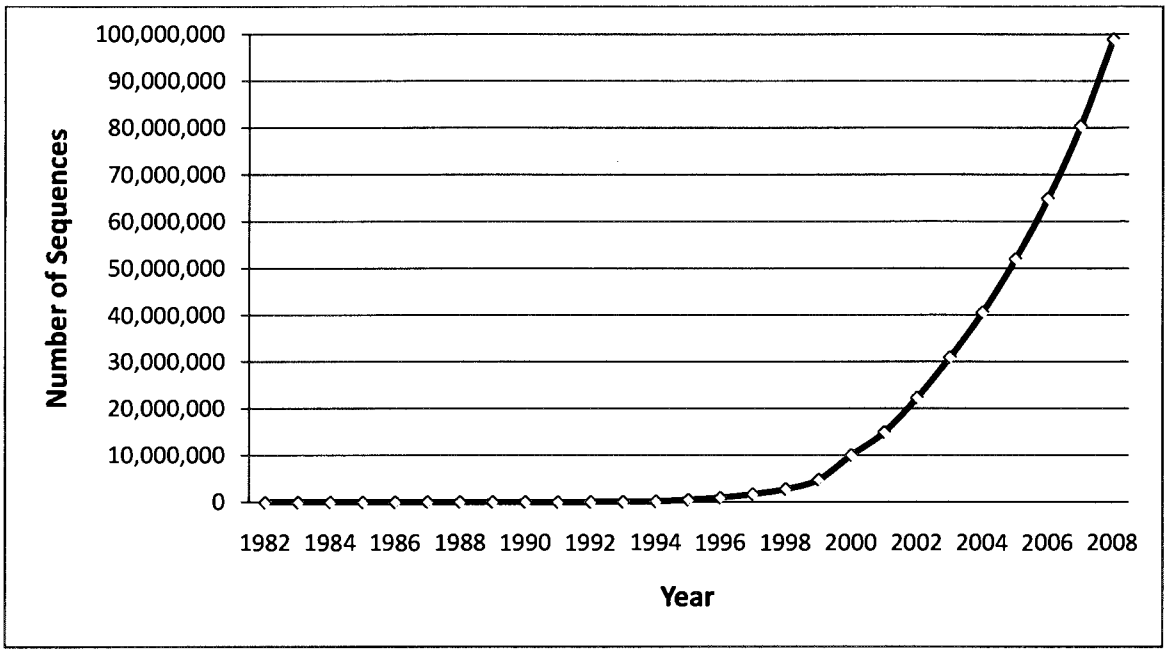


Figure 1-1 Growth rate of the GenBank sequence database

A graph of number of sequences (Number of Sequences) contained in the GenBank sequence database versus the year (Year). Beginning in 1982 GenBank contained 606 sequences and by 2008 it contained 98,868,465 sequences. GenBank growth statistics are provided by the National Center for Biotechnology Information (NCBI) [1.28] which maintains the GenBank sequence database.

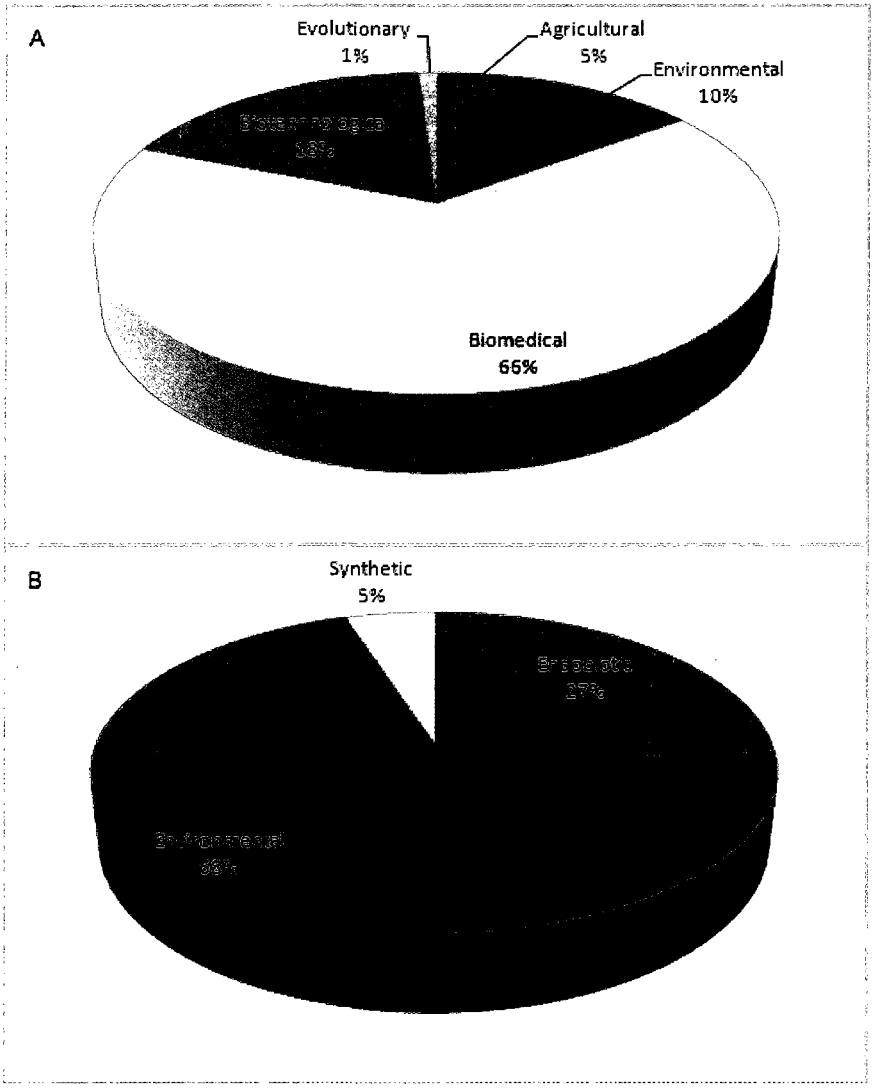


Figure 1-2 Genome projects versus metagenome projects

Panel A - Funding relevance of bacterial genome projects: The relative allocation of funding for bacterial genome projects with respect to project category. Funding allocation data was provided by the Genomes OnLine Database (GOLD) [1.29]. Panel B - Metagenome project categories: The relative allocation of metagenome projects with respect to project category. Metagenome project data was provided by GOLD [1.29].

Chapter 2

Beyond the Bounds of Orthology: Functional Inference from Metagenomic Context

Gregory Vey and Gabriel Moreno-Hagelsieb*

Department of Biology, Wilfrid Laurier University, 75 University Avenue West, Waterloo ON, Canada, N2L 3C5

*Corresponding author

Email addresses:

GV: veyx9970@wlu.ca

GM-H: gmoreno@wlu.ca

Submitted to: *Molecular BioSystems*.

2.1 Abstract

The effectiveness of the computational inference of function by genomic context is bounded by the diversity of known microbial genomes. Although metagenomes offer access to previously intangible organisms, these sequenced environmental fragments prevent the conventional establishment of orthologous relationships required for reliably predicting functional interactions. We introduce a novel protocol for the prediction of functional interactions using data sources without information about orthologous relationships. To illustrate this process, we use the Sargasso Sea metagenome to construct a functional interaction network for the *Escherichia coli* K12 genome. We identify two reliability metrics, target intergenic distance and source interaction count, and apply them to selectively filter the predictions retained to construct the network of functional interactions. The resulting network contains 2,297 nodes with 10,072 edges with a positive predictive value of 0.80. The metagenome yielded 8,423 functional interactions beyond those found using only the genomic orthologs as a data source. This amounted to a 134% increase in the total number of functional interactions that are predicted by combining the metagenome and the genomic orthologs versus the genomic orthologs alone. In the absence of detectable orthologous relationships it remains feasible to derive a reliable set of predicted functional interactions. This offers a strategy for harnessing other metagenomes and homologs in general. Because metagenomes allow access to previously unreachable microorganisms, this will result in expanding the universe of known functional interactions thus furthering our understanding of functional organization.

2.2 Introduction

The main objective of the present work is to provide a method to extend the computational inference of functional associations by genomic context, to include the use of metagenomic sequences to complement the functional associations predicted for a genome of interest (Figure 1).

Almost as soon as there were sufficient genomes available for a test, researchers proposed methods to infer functional interactions by genomic context [1, 2]. The three main methods, which we call the Three Musketeers of Genomic Context [3], rely on finding orthologs, homologs diverging after a speciation event [4], and inferring a functional interaction by finding: (a) Gene fusions [5, 6], where two genes are assumed to interact if their orthologs are fused into a gene coding for a multidomain protein in another genome; (b) Conservation of gene order [7], where the conservation of orthologs next to each other, beyond expectations by chance, is used as a clue for a functional interaction; and (c) Phylogenetic profiles [8-10], where the orthologs to genes coding for functionally interacting proteins are expected to co-occur, be both present or both absent across genomes. The D'Artagnan of functional inference builds on top of methods to predict operons. A functional interaction is inferred if the genes themselves, or their orthologs, are found to be in the same operon [3, 11-13].

While the above-mentioned methods provide many high-quality predictions of functional interactions, their coverage might be limited by the biases determining which genomes have been sequenced. In recent years, metagenomic data sets derived from environmental shotgun sequencing have gained a position of increased prominence in

biological databases. Large-scale metagenomic projects have been completed that depict various viral and microbial communities including freshwater, marine, subterranean, intestinal, and many other environments [14-27]. Furthermore, the sheer volume of metagenomic sequence data has exceeded the combined total of the microbial genomes [28]. Thus, metagenomes offer the prospect of providing novel insights into the dynamics of microorganisms with populations that are neither clonal, nor single species, such as symbionts and obligate pathogens [29-31]. Such an increase in the accessibility of microbial biodiversity has the potential to further our understanding of fundamental biological functions and processes [28]. It also has the capability of reducing the biases contained in existing sequence databases by extending the repertoire of known genes and functions [30, 32].

The fields of comparative metagenomics and functional metagenomics have emerged in an effort to compare microbial communities in terms of their relative biodiversity and respective functional activities [33-35]. While uncovering novel functions is an integral aspect of these fields [33-35], functional metagenomics remains in its infancy and little effort has been directed toward treating the metagenomes as sources of functional interactions useful at complementing the information of fully sequenced genomes. This is a paramount consideration since prior to the introduction of metagenomic data the information that had been deposited had been principally derived from the genomic sequences of microorganisms that are amenable to culturing [29-31, 36]. However, it has been estimated that more than 99% of microorganisms are uncultivable [29-31, 36]. Capturing a greater sample of the biodiversity of microorganisms and their known functional

interactions would ensure a more accurate representation of existing proteins and potentially help to assign function to the larger number of currently uncharacterized proteins [30]. Therefore, metagenomic data sets can provide an opportunity to extend the universe of known functional interactions and subsequently facilitate pursuits such as classifying hypothetical proteins and assigning functional annotations.

Functional inference from metagenomic context offers an invaluable means to reconcile the current limitations of functional genomics through the expansion of usable data sources. However, reliable functional inference is dependent on the detection of orthology. The particular types of homologies are hard to identify in metagenomes because of the fragmented nature of the environmental sequences. Therefore, comparison between genes is confined to a consideration of only the general case of homology. However, using homology rather than orthology generates many spurious predictions that arise from paralogs because they possess homologous sequences but potentially divergent functions [37]. For instance, if each member of a family of proteins interacts with a specific member of another family of proteins, the problem of solving for orthology would result in predictions for all members of the first family interacting with all members of the second, thus generating a high number of false positives (see Figure 1). Therefore, the development of a protocol that is not limited to using known orthologous data, yet solves the problem of correct assignment of interactions, would represent a major step toward furthering many different pursuits in functional genomics and metagenomics.

In the case of functional inference, we propose that the use of indiscriminate homology results in a superset of functional interaction predictions. A reliable set of

predictions should lie within the prediction superset that has been inflated by paralog families producing many extraneous functional interactions. If this is indeed the case, then it should be possible to demonstrate an improvement in validation measures through the removal of these spurious predictions. To explore this possibility, we present a three-part protocol that extends on the use of rearranged operons [3] into metagenomes. First, we predict operons in the metagenome sequences based on intergenic distances [38, 39]. Next, these predictions are mapped using BLASTP [40] results against a target genome. Lastly, spurious predictions are reduced through filtering with a set of prediction reliability metrics. To illustrate the feasibility of this process, we used the Sargasso Sea metagenome [15] to construct a functional interaction network (FIN) for the *Escherichia coli* K12 MG1655 [41] target genome (NCBI Version: NC_000913.2).

2.3 Results and Discussion

2.3.1 Baseline predictions network

To assess the construction of FIN from homologs in metagenomes, we developed three contrasting FINs (see Methods and materials), one using the genomic orthologs, another using genomic homologs, and a final one using metagenomic homologs (Table 1). As expected, the genomic-ortholog FIN provides a better positive predictive value (PPV) than either of the FINs derived from all homologs. The homologs also performed poorly when assessed using correlation of expression data to validate the predictions that are not captured by the measure of PPV. Figure 2 shows the distribution of the correlation of expression values for the metagenomic-homologs FIN. The metagenome exhibits a trend that is only marginally better than that of gold negatives (GN) derived from the

EcoCyc database [42, 43]. Overall, the indiscriminate homologs, whether genomic or metagenomic, appeared to be a poor data source for the development of a FIN.

To rule out effects of the data preparation process, we considered how the PPV of the metagenomic homologs was affected by various preparation variables. Specifically, we examined two levels for each of the following variables: source interaction prediction threshold, minimum sequence coverage in the alignment, and maximum E-value threshold (see Methods and materials). Table 2 shows the results of the different combinations for these variables. In general, increasing the stringency (using the High treatment level) of any variable resulted in an increased PPV and the best PPV was achieved by increasing the stringency for all variables. However, increasing the PPV through increased filtering markedly decreased the proportion of recovered gold positives (GP) derived from EcoCyc [42, 43]. Therefore, adjusting the values of the preparation variables does not satisfactorily reconcile the poor PPV of the metagenomic homologs. Instead, we elected to use the largest FIN as a baseline and explore the feasibility of discarding a portion of the predictions according to some other measure of reliability.

2.3.2 Prediction reliability metrics

We attempted to identify properties of predicted functional interactions that could serve as metrics to determine their reliability on an individual case basis. This approach was intended to provide a protocol to selectively filter the full data set and remove spurious predictions, thereby improving the overall quality of the remaining set of functional interactions. To accomplish this we selected two specific metrics; target intergenic distance and source interaction count.

Target intergenic distance was defined as the distance in base pairs between two target genes according to the following formula:

$$D = gene2\_start - (gene1\_end + 1)$$

For example, if the functional interaction M_1-M_2 was predicted in the source metagenome and M_1 and M_2 map to T_1 and T_2 in the target genome (see Methods), then the target intergenic distance would be defined by the distance between T_1 and T_2, regardless of these target genes being adjacent or not. To accommodate the circularity of the *E. coli* K12 genome, distances were calculated in each direction and the lesser of the two values was defined as the target intergenic distance. We experimented with the use of a maximum value for target intergenic distance as a metric for determining the reliability of individual predictions (see Additional file 1). Figure 3 shows the relationship between PPV and maximum target intergenic distance. As expected, this metric was particularly useful for recovering genes belonging to the same experimentally verified operons in the GP dataset.

Source interaction count was defined as the number of predicted interactions in the data source that equated to a given target interaction. For example, the interaction T_1-T_2 in the target genome must have been mapped from at least one observed source interaction, such as M_1-M_2 in the metagenome. However, as a consequence of the mapping process (See Materials and methods) it is possible that multiple interactions observed in the metagenome all translate into the same interaction in the target genome. As a result, any target interaction must be instantiated from one or more predictions from the source interactions. We experimented with the use of a minimum value for source

interaction count as a metric for determining the reliability of individual predictions (see Additional file 1). Figure 4 shows the relationship between PPV and minimum source interaction count. This metric was useful for increasing the number of non-operonic GPs that would otherwise not be recovered by using target intergenic distance alone.

2.3.3 Filtered predictions network

We applied the prediction reliability metrics to filter the previously constructed baseline FINs. It was possible to achieve a range of improved PPVs (see Additional file 1). As a result, it was possible to construct a FIN for the metagenome that yielded a reliable PPV value (0.80), despite the absence of any information about orthologous relationships. Figure 5 shows the FIN that was obtained for the *E. coli* K12 genome using the Sargasso Sea metagenome [15], as viewed using Cytoscape [44]. Next, we investigated whether the prediction reduction protocol could be used for other data where only homology is determined, not orthology. This was demonstrated by generating a reliable FIN (0.80) for the genomic homologs. Finally, we verified that the prediction reduction protocol was also suitable for filtering orthologous data by constructing a reliable FIN (0.80) for the genomic orthologs. Table 3 shows the results for filtering each of the previously constructed FINs to achieve a reliable PPV value. It was observed that each of the filtered FINs retained large proportions (77% to 94%) of their original nodes but had undergone a substantial reduction in their numbers of edges. This reduction in edges corresponded to the removal of spurious predictions of functional interactions and facilitated the improved PPV.

In addition to the improved PPV, we were also interested in the correlation of expression data for the metagenomic FIN. This was essential since 8755 (87%) of the predicted functional interactions are neither GPs nor GNs. Therefore, examining the distribution of the correlation of expression values provided an indication of the reliability of those unknown predictions. Figure 6 shows the difference between the distributions of the correlation of expression values for the filtered metagenomic FIN versus the unfiltered metagenomic FIN. A distinct improvement can be seen for the filtered FIN versus the unfiltered FIN.

2.3.4 Contribution of the metagenome

Having demonstrated the ability to construct a reliable FIN from a metagenomic source data source, we investigated the contribution of this FIN with respect to expanding the universe of known functional interactions for the *E. coli* K12 genome. First, the filtered set of metagenomic predictions was compared against the filtered set of predictions for the genomic orthologs. An intersection of 1,649 predictions showed that a common core of functional interactions existed. Furthermore, the combination of these two sets yielded 8,423 more functional interactions than using only the genomic orthologs, resulting in a 134% increase. To determine the impact of filtering the predictions from the genomic orthologs, the filtered set of metagenomic predictions was compared against the unfiltered set of predictions for the genomic orthologs. The metagenome still donated 8,161 that were not found using the full set orthologs for an increase of 51% for the total number of functional interactions. To determine whether the level of homology was a factor, the filtered set of metagenomic predictions was

compared against the filtered set of predictions for the genomic homologs. In this case, the metagenome contributed 1,232 functional interactions for a 7% gain in the total number of interactions. Compared to the genomic orthologs, the metagenome exhibited a smaller relative union and a larger relative intersection with the genomic homologs, suggesting that there was a greater mutual component given a common level of homology, likely due to the robust coverage of the genomic homologs versus the genomic orthologs. Finally, to explore whether the genomic homologs could extend the genomic orthologs the filtered set of predictions for the genomic homologs was compared against both sets of predictions for the genomic orthologs. While the homologs clearly added a large proportion of functional interactions, the orthologs, whether filtered or unfiltered, contained their own unique contribution of functional interactions. Table 4 summarizes the results for comparing and combining the various FINs.

2.4 Conclusions

2.4.1 Beyond orthology

The prediction reliability metrics used in the present work to filter homolog-based predictions have demonstrated that in the absence of known orthologous relationships it remains possible to derive a reliable set of predicted functional interactions. This is noteworthy because it offers a strategy for harnessing other metagenomes and homologs in general. Not only does this offer the opportunity to utilize novel data sources, it also provides a means to use homologs and orthologs together, thereby yielding an addendum to results achieved by the conventional use of only the genomic orthologs. Future works should be aimed at determining more and better prediction reliability metrics and to

examine their portability between different target genomes. Techniques such as binary logistic regression could be used to develop a general predictive model that could potentially eliminate the constraint of orthology.

2.4.2 Metagenomic functional inference

The ability to infer functional interactions from metagenomic data sources creates the opportunity to further functional metagenomics. Because the metagenomes allow access to previously intangible microorganisms, this will result in expanding the universe of known functional interactions, especially as the number of deposited environmental data sets continues to grow. In turn, increasing the existing collection of functional interactions will have a cascading effect on our understanding of functional organization while improving our accuracy in identifying hypothetical proteins and assigning functional annotations. Future works should be aimed at extending the present proof of concept through the incorporation of multiple metagenomes. Thus, the omissions and biases that have arisen from the prevalence of clonal microbial organisms could be eventually rectified through capturing a greater breadth of the true microbial biodiversity. Ultimately, the recovery of novelty from the metagenomes will propel applications across a wide spectrum of other fields thereby allowing the advancement of countless interests.

2.5 Methods

2.5.1 Data sources

The Sorcerer II data package available online from the Sorcerer II Expedition website [21] was used as the metagenomic data source for this work. This is an annotated data set of 811,372 contiguous environmental fragments (contigs) that include 1,001,987

different genes obtained from the Sargasso Sea [15]. The website of this metagenome provides FASTA format files for both the nucleotide and the peptide sequences. Additionally, a gene feature format (gff) file is included that maps individual genes to their corresponding peptide sequences.

The gff file was used to identify a total of 1,001,987 annotated genes. There were 403,051 contigs containing a single gene each. The remaining 598,936 genes were distributed across 251,638 contigs. These data provided 347,298 pairs of adjacent genes that could be used to predict operons by intergenic distances.

2.5.2 Prediction generation phase

To generate predictions of functional interactions, we used an existing method to infer functional relationships from the recombination of predicted operons [3]. We predicted operons within the data set of metagenomic adjacent, same-strand, gene pairs explained above, by the methods described previously [38, 39]. Distances were determined using start and end coordinates contained in the gff Sargasso Sea files. A minimum log-likelihood [38, 39] (LLH) threshold of 0.01 was used. The final result was a prediction set that included a total of 197,678 predicted interactions derived on the basis of co-occurrence within a mutual operon.

2.5.3 Prediction mapping phase

The proteins that corresponded to the genes in the metagenomic prediction set were compared against the set of *E. coli* K12 proteins using NCBI's BLASTP [40]. The results were filtered to remove hits with less than 60% alignment (target or query), or

with E-values greater than $1x10^{-6}$. The 1,231,909 remaining hits were used as the mapping set.

The mapping phase was designed to generate all possible functional interactions, without regard to spurious interactions that result from the combinatorial use of homologs. We generated an interaction superset by using the prediction set in conjunction with the mapping set, in the following manner. First, the individual elements of the mapping set were aggregated into mapping lists that were sorted according to a prediction key. Next, an individual prediction was obtained from the prediction set. For each of the two members of the predicted interaction, a list of target proteins was created. This involved searching the mapping lists to retrieve the list of proteins from the target genome that mapped onto the given interaction member. Finally, if the lists for each interaction member were non-empty, we generated a set of target functional interactions using the complete bipartite graph of the two lists. The resulting set was added to the overall superset of interactions and contributed *mn* interactions, where *m* and *n* were the respective sizes of the non-empty lists. This process was repeated until all entries from the prediction set had been exhausted, resulting in the final interaction superset.

2.5.4 Prediction reduction phase

The prediction reduction phase was designed to reduce the number of spurious predictions that were produced by the previous phase, thereby improving the overall quality of the functional interaction network. Each element of the interaction superset was tested according to the values of the prediction reliability metrics (see Results and Discussion) that were selected to generate the particular reduced functional interaction

network. Specifically, if an individual predicted functional interaction did not exceed the maximum target intergenic distance, then it was retained as part of the reduced set of functional interactions. Otherwise, if it was not below the minimum source interaction count, then it was retained as part of the reduced set of functional interactions. Otherwise, no further metrics were applied and the interaction was rejected from the reduced set of functional interactions.

2.6 Acknowledgements

The authors thank Juan Javier Díaz-Mejía and Sarath Chandra Janga for helpful discussions. This work was supported by an NSERC Discovery Grant to GM-H. We acknowledge computational power supplied by SHARCNET (Shared Hierarchical Academic Research Computing Network), and computer-equipment obtained through a grant from Wilfrid Laurier University.

2.7 References

1. D. Eisenberg, E. M. Marcotte, I. Xenarios and T. O. Yeates, *Nature*, 2000, **405**, 823-826.
2. M. Huynen, B. Snel, W. Lathe, 3rd and P. Bork, *Genome Res*, 2000, **10**, 1204-1210.
3. S. C. Janga, J. Collado-Vides and G. Moreno-Hagelsieb, *Nucleic Acids Res*, 2005, **33**, 2521-2530.
4. W. M. Fitch, *Trends Genet*, 2000, **16**, 227-231.
5. A. J. Enright, I. Iliopoulos, N. C. Kyrpides and C. A. Ouzounis, *Nature*, 1999, **402**, 86-90.

6. E. M. Marcotte, M. Pellegrini, M. J. Thompson, T. O. Yeates and D. Eisenberg, *Nature*, 1999, **402**, 83-86.

7. R. Overbeek, M. Fonstein, M. D'Souza, G. D. Pusch and N. Maltsev, *Proc Natl Acad Sci U S A*, 1999, **96**, 2896-2901.

8. R. L. Tatusov, E. V. Koonin and D. J. Lipman, *Science*, 1997, **278**, 631-637.

9. T. Gaasterland and M. A. Ragan, *Microb Comp Genomics*, 1998, **3**, 199-217.

10. M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg and T. O. Yeates, *Proc Natl Acad Sci U S A*, 1999, **96**, 4285-4288.

11. I. B. Rogozin, K. S. Makarova, J. Murvai, E. Czabarka, Y. I. Wolf, R. L. Tatusov, L. A. Szekely and E. V. Koonin, *Nucleic Acids Res*, 2002, **30**, 2212-2223.

12. B. Snel, P. Bork and M. A. Huynen, *Proc Natl Acad Sci U S A*, 2002, **99**, 5890-5895.

13. P. Hu, S. C. Janga, M. Babu, J. J. Diaz-Mejia, G. Butland, W. Yang, O. Pogoutse, X. Guo, S. Phanse, P. Wong, S. Chandran, C. Christopoulos, A. Nazarians-Armavil, N. K. Nasseri, G. Musso, M. Ali, N. Nazemof, V. Eroukova, A. Golshani, A. Paccanaro, J. F. Greenblatt, G. Moreno-Hagelsieb and A. Emili, *PLoS Biol*, 2009, **7**, e96.

14. C. Schmeisser, C. Stockigt, C. Raasch, J. Wingender, K. N. Timmis, D. F. Wenderoth, H. C. Flemming, H. Liesegang, R. A. Schmitz, K. E. Jaeger and W. R. Streit, *Appl Environ Microbiol*, 2003, **69**, 7298-7309.

15. J. C. Venter, K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H.

Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y. H. Rogers and H. O. Smith, *Science*, 2004, **304**, 66-74.

16. F. E. Angly, B. Felts, M. Breitbart, P. Salamon, R. A. Edwards, C. Carlson, A. M. Chan, M. Haynes, S. Kelley, H. Liu, J. M. Mahaffy, J. E. Mueller, J. Nulton, R. Olson, R. Parsons, S. Rayhawk, C. A. Suttle and F. Rohwer, *PLoS Biol*, 2006, **4**, e368.

17. S. R. Gill, M. Pop, R. T. Deboy, P. B. Eckburg, P. J. Turnbaugh, B. S. Samuel, J. I. Gordon, D. A. Relman, C. M. Fraser-Liggett and K. E. Nelson, *Science*, 2006, **312**, 1355-1359.

18. H. N. Poinar, C. Schwarz, J. Qi, B. Shapiro, R. D. Macphee, B. Buigues, A. Tikhonov, D. H. Huson, L. P. Tomsho, A. Auch, M. Rampp, W. Miller and S. C. Schuster, *Science*, 2006, **311**, 392-394.

19. T. Woyke, H. Teeling, N. N. Ivanova, M. Huntemann, M. Richter, F. O. Gloeckner, D. Boffelli, I. J. Anderson, K. W. Barry, H. J. Shapiro, E. Szeto, N. C. Kyrpides, M. Mussmann, R. Amann, C. Bergin, C. Ruehland, E. M. Rubin and N. Dubilier, *Nature*, 2006, **443**, 950-955.

20. A. B. Martin-Cuadrado, P. Lopez-Garcia, J. C. Alba, D. Moreira, L. Monticelli, A. Strittmatter, G. Gottschalk and F. Rodriguez-Valera, *PLoS One*, 2007, **2**, e914.

21. D. B. Rusch, A. L. Halpern, G. Sutton, K. B. Heidelberg, S. Williamson, S. Yooseph, D. Wu, J. A. Eisen, J. M. Hoffman, K. Remington, K. Beeson, B. Tran, H. Smith, H. Baden-Tillson, C. Stewart, J. Thorpe, J. Freeman, C. Andrews-

Pfannkoch, J. E. Venter, K. Li, S. Kravitz, J. F. Heidelberg, T. Utterback, Y. H. Rogers, L. I. Falcon, V. Souza, G. Bonilla-Rosso, L. E. Eguiarte, D. M. Karl, S. Sathyendranath, T. Platt, E. Bermingham, V. Gallardo, G. Tamayo-Castillo, M. R. Ferrari, R. L. Strausberg, K. Nealson, R. Friedman, M. Frazier and J. C. Venter, *PLoS Biol*, 2007, **5**, e77.

22. F. Warnecke, P. Luginbuhl, N. Ivanova, M. Ghassemian, T. H. Richardson, J. T. Stege, M. Cayouette, A. C. McHardy, G. Djordjevic, N. Aboushadi, R. Sorek, S. G. Tringe, M. Podar, H. G. Martin, V. Kunin, D. Dalevi, J. Madejska, E. Kirton, D. Platt, E. Szeto, A. Salamov, K. Barry, N. Mikhailova, N. C. Kyrpides, E. G. Matson, E. A. Ottesen, X. Zhang, M. Hernandez, C. Murillo, L. G. Acosta, I. Rigoutsos, G. Tamayo, B. D. Green, C. Chang, E. M. Rubin, E. J. Mathur, D. E. Robertson, P. Hugenholtz and J. R. Leadbetter, *Nature*, 2007, **450**, 560-565.

23. L. Wegley, R. Edwards, B. Rodriguez-Brito, H. Liu and F. Rohwer, *Environ Microbiol*, 2007, **9**, 2707-2719.

24. J. J. Grzymski, A. E. Murray, B. J. Campbell, M. Kaplarevic, G. R. Gao, C. Lee, R. Daniel, A. Ghadiri, R. A. Feldman and S. C. Cary, *Proc Natl Acad Sci U S A*, 2008, **105**, 17516-17521.

25. A. Schluter, T. Bekel, N. N. Diaz, M. Dondrup, R. Eichenlaub, K. H. Gartemann, I. Krahn, L. Krause, H. Kromeke, O. Kruse, J. H. Mussgnug, H. Neuweger, K. Niehaus, A. Puhler, K. J. Runte, R. Szczepanowski, A. Tauch, A. Tilker, P. Viehover and A. Goesmann, *J Biotechnol*, 2008, **136**, 77-90.

26. A. Schluter, L. Krause, R. Szczepanowski, A. Goesmann and A. Puhler, *J Biotechnol*, 2008, **136**, 65-76.

27. S. G. Tringe, T. Zhang, X. Liu, Y. Yu, W. H. Lee, J. Yap, F. Yao, S. T. Suan, S. K. Ing, M. Haynes, F. Rohwer, C. L. Wei, P. Tan, J. Bristow, E. M. Rubin and Y. Ruan, *PLoS One*, 2008, **3**, e1862.

28. E. D. Harrington, A. H. Singh, T. Doerks, I. Letunic, C. von Mering, L. J. Jensen, J. Raes and P. Bork, *Proc Natl Acad Sci U S A*, 2007, **104**, 13913-13918.

29. C. S. Riesenfeld, P. D. Schloss and J. Handelsman, *Annu Rev Genet*, 2004, **38**, 525-552.

30. M. Ferrer, F. Martinez-Abarca and P. N. Golyshin, *Curr Opin Biotechnol*, 2005, **16**, 588-593.

31. S. G. Tringe and E. M. Rubin, *Nat Rev Genet*, 2005, **6**, 805-814.

32. S. Yooseph, G. Sutton, D. B. Rusch, A. L. Halpern, S. J. Williamson, K. Remington, J. A. Eisen, K. B. Heidelberg, G. Manning, W. Li, L. Jaroszewski, P. Cieplak, C. S. Miller, H. Li, S. T. Mashiyama, M. P. Joachimiak, C. van Belle, J. M. Chandonia, D. A. Soergel, Y. Zhai, K. Natarajan, S. Lee, B. J. Raphael, V. Bafna, R. Friedman, S. E. Brenner, A. Godzik, D. Eisenberg, J. E. Dixon, S. S. Taylor, R. L. Strausberg, M. Frazier and J. C. Venter, *PLoS Biol*, 2007, **5**, e16.

33. S. G. Tringe, C. von Mering, A. Kobayashi, A. A. Salamov, K. Chen, H. W. Chang, M. Podar, J. M. Short, E. J. Mathur, J. C. Detter, P. Bork, P. Hugenholtz and E. M. Rubin, *Science*, 2005, **308**, 554-557.

34. E. A. Dinsdale, R. A. Edwards, D. Hall, F. Angly, M. Breitbart, J. M. Brulc, M. Furlan, C. Desnues, M. Haynes, L. Li, L. McDaniel, M. A. Moran, K. E. Nelson, C. Nilsson, R. Olson, J. Paul, B. R. Brito, Y. Ruan, B. K. Swan, R. Stevens, D. L. Valentine, R. V. Thurber, L. Wegley, B. A. White and F. Rohwer, *Nature*, 2008, **452**, 629-632.

35. D. H. Huson, D. C. Richter, S. Mitra, A. F. Auch and S. C. Schuster, *BMC Bioinformatics*, 2009, **10 Suppl 1**, S12.

36. M. S. Rappe and S. J. Giovannoni, *Annu Rev Microbiol*, 2003, **57**, 369-394.

37. A. H. Singh, T. Doerks, I. Letunic, J. Raes and P. Bork, *J Bacteriol*, 2009, **191**, 32-41.

38. H. Salgado, G. Moreno-Hagelsieb, T. F. Smith and J. Collado-Vides, *Proc Natl Acad Sci U S A*, 2000, **97**, 6652-6657.

39. G. Moreno-Hagelsieb and J. Collado-Vides, *Bioinformatics*, 2002, **18 Suppl 1**, S329-336.

40. S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman, *Nucleic Acids Res*, 1997, **25**, 3389-3402.

41. F. R. Blattner, G. Plunkett, 3rd, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau and Y. Shao, *Science*, 1997, **277**, 1453-1474.

42. P. D. Karp, M. Riley, S. M. Paley and A. Pelligrini-Toole, *Nucleic Acids Res*, 1996, **24**, 32-39.

43. I. M. Keseler, C. Bonavides-Martinez, J. Collado-Vides, S. Gama-Castro, R. P. Gunsalus, D. A. Johnson, M. Krummenacker, L. M. Nolan, S. Paley, I. T. Paulsen, M. Peralta-Gil, A. Santos-Zavaleta, A. G. Shearer and P. D. Karp, *Nucleic Acids Res*, 2009, **37**, D464-470.

44. S. Killcoyne, G. W. Carter, J. Smith and J. Boyle, *Methods Mol Biol*, 2009, **563**, 219-239.

2.8 Figure Legends

Figure 1 – The problem of paralogy

Two genes, A and B, might be separated in a target genome. Yet, their orthologs, A_o and B_o, within an informative genome might be in the same operon, indicating that genes A and B might functionally interact in the target genome. In metagenome fragments, orthology cannot be inferred. Genes homologs to A and B, A_h and B_h, might indicate a functional interaction. However, if genes A and B belong to protein families with several paralogs, where each member of Family A interacts with a specific member of Family B (solid lines), there is a potential for a large number of false positives. In the example, we would infer three true positives (solid lines) and 9 false positives (dashed lines).

Figure 2 – Relative frequencies of correlation of expression values

A graph of the relative frequencies of correlation of expression values for the EcoCyc gold negative functional interactions (GNs), the Ecocyc gold positive functional interactions (GPs), and the full set of predicted functional interactions from the Sargasso Sea metagenome (S-Full).

Figure 3 – Target intergenic distance versus positive predictive value

A graph of positive predictive value (PPV) scores and proportions of EcoCyc gold positive functional interactions (GPs) versus maximum target intergenic distances (Distance) that are used as thresholds to reject predictions that exceed these maximum values.

Figure 4 – Source interaction count versus positive predictive value

A graph of positive predictive value (PPV) scores and proportions of EcoCyc gold positive functional interactions (GPs) versus minimum source interaction counts (Count) that are used as thresholds to reject predictions that do not meet these minimum values.

Figure 5 – Functional interaction network for the E. coli K12 MG1655 genome

A functional interaction network for the *E. coli* K12 MG1655 genome derived from the prediction reduced Sargasso Sea metagenome, as viewed through Cytoscape [44].

Figure 6 – Relative frequencies of correlation of expression values

A graph of the relative frequencies of correlation of expression values for the EcoCyc gold negative functional interactions (GNs), the Ecocyc gold positive functional interactions (GPs), and the reduced set of predicted functional interactions from the Sargasso Sea metagenome (S-Red).

2.9 Tables

Table 2-1 Baseline functional interaction networks

FIN	Nodes	Edges	GNs	GPs	PPV
Sargasso	2,991	53,126	3,439	1,535	0.309
Homologs	3,837	217,701	10,948	3,040	0.217
Orthologs	3,672	15,959	1,415	1,879	0.570

A summary of the functional interaction networks constructed by using all generated predictions from each respective data source. For each functional interaction network the number of network nodes (Nodes) is listed along with the number of network edges (Edges), the number of recovered EcoCyc gold negative interactions (GNs), the number

of recovered EcoCyc gold positive interactions (GPs), and the positive predictive value (PPV).

Table 2-2 Effects of data preparation variables

Factors			Interactions	GPs	GNs	PPV	Coverage
Low LLH	Low % Align	Low E Value	53,126	1,535	3,439	0.309	100.00%
		High E Value	36,373	1,354	2,495	0.352	88.21%
	High % Align	Low E Value	40,324	1,389	2,700	0.340	90.49%
		High E Value	29,208	1,244	2,088	0.373	81.04%
High LLH	Low % Align	Low E Value	31,085	1,108	1,906	0.368	72.18%
		High E Value	21,432	967	1,331	0.421	63.00%
	High % Align	Low E Value	24,005	1,006	1,488	0.403	65.54%
		High E Value	17,402	893	1,102	0.448	58.18%

A summary of the functional interaction networks constructed by manipulating three different data preparation variables. Source interaction prediction threshold was tested using two values for log likelihood (0.01 (Low LLH) and 1.00 (High LLH), combined with minimum sequence (target or query) alignment percentage using two values (60%

(Low % Align) and 80% (High % Align), combined with the maximum allowable E value using two values (1e-6 (Low E Value) and 1e-10 (High E Value). It should be noted that here Low E Value refers to the factor level (level of stringency), not the magnitude of the value itself. For each functional interaction network the number of interactions (Interactions) obtained is listed (this is synonymous with network edges) along with the number of recovered EcoCyc gold negative interactions (GNs), the number of recovered EcoCyc gold positive interactions (GPs), the positive predictive value (PPV), and the proportion of recovered GPs (Coverage) versus the GPs found in the functional interaction network derived from the least stringent levels for the preparation variables (Low LLH, Low % Align, Low E Value).

Table 2-3 Filtered functional interaction networks

FIN	Nodes	Edges	GNs	GPs	Operons	Coverage	PPV
Sargasso	2,297	10,072	263	1,054	781	76.80%	0.800
Homologs	3,380	17,740	443	1,776	1,267	88.09%	0.800
Orthologs	3,437	6,267	387	1,550	1,057	93.60%	0.800

A summary of the functional interaction networks constructed by using filtered predictions from each respective data source. For each functional interaction network the number of network nodes (Nodes) is listed along with the number of network edges (Edges), the number of recovered EcoCyc gold negative interactions (GNs), the number of recovered EcoCyc gold positive interactions (GPs), the number of EcoCyc gold positive interactions that contained string "operons" in the keywords list (Operons), the

proportion of nodes contained from the corresponding unfiltered network (Coverage), and the positive predictive value (PPV).

Table 2-4 Gain in functional interactions from combined sets

S1	S2	|S1|	|S2|	|S1 ∪ S2|	|S1 • S2|	|S1 - (S1 • S2)|	|S2 - (S1 • S2)|	S2 % Gain
Sarg	Orth	10,072	6,267	14,690	1,649	8,423	4,618	134.40%
Sarg	Orth*	10,072	15,959	24,120	1,911	8,161	14,048	51.14%
Sarg	Hom	10,072	17,740	18,972	8,840	1,232	8,900	6.94%
Hom	Orth	17,740	6,267	20,743	3,264	14,476	3,003	230.99%
Hom	Orth*	17,740	15,959	30,091	3,608	14,132	12,351	88.55%

A set theoretical summary of the functional interactions contained in the various networks and their combinations with one another. Set One (S1) and Set Two (S2) are given for each combined superset of predicted functional interactions. The size of Set One (|S1|) and the size of Set Two (|S2|) are listed along with the size of their union (|S1∪ S2|), the size of their intersection (|S1 • S2|), the number of unique predictions found only in Set One (|S1 – (S1 • S2)|), the number of unique predictions found only in Set Two (|S2 – (S1 • S2)|), and proportional increase in the number total number of predictions from the combined sets (S2 % Gain) versus Set Two alone. Comparisons were performed for the filtered Sargasso Sea metagenome functional interaction network (Sarg), the filtered genomic homologs functional interaction network (Hom), the filtered genomic orthologs functional interaction network (Orth), and the unfiltered genomic orthologs functional interaction network (Orth*).

2.10 Additional Files

File 2-1 Metrics.xls – Spreadsheet of results from prediction reliability metrics tests

Minimum source interaction count (Min Count) was tested ranging from 0 to 200 predictions at intervals of 10 predictions, in combination with maximum target intergenic distance (Max Dist) ranging from 0 to 150,000 base pairs at intervals of 500 base pairs. For each combination of prediction reliability metrics, the number of interactions (Interacts) obtained is listed (this is synonymous with network edges) along with the number of recovered EcoCyc gold negative interactions (GNs), the number of recovered EcoCyc gold positive interactions (GPs), the number of EcoCyc gold positive interactions that contained string "operons" in the keywords list (Operons), and the positive predictive value (PPV) obtained using the given values of the prediction reliability metrics. This was performed on the genomic orthologs (Orthologs), the genomic homologs (Homologs), and the Sargasso Sea metagenome (Sargasso). An additional set of experiments was carried out on the genomic orthologs with source interaction count held at 10 (Orthologs 10) to obtain values that yielded a positive predictive value that was comparable to the other two sets of values. The sets of values that were used to construct the respective functional interaction networks have been highlighted.

2.11 Figures

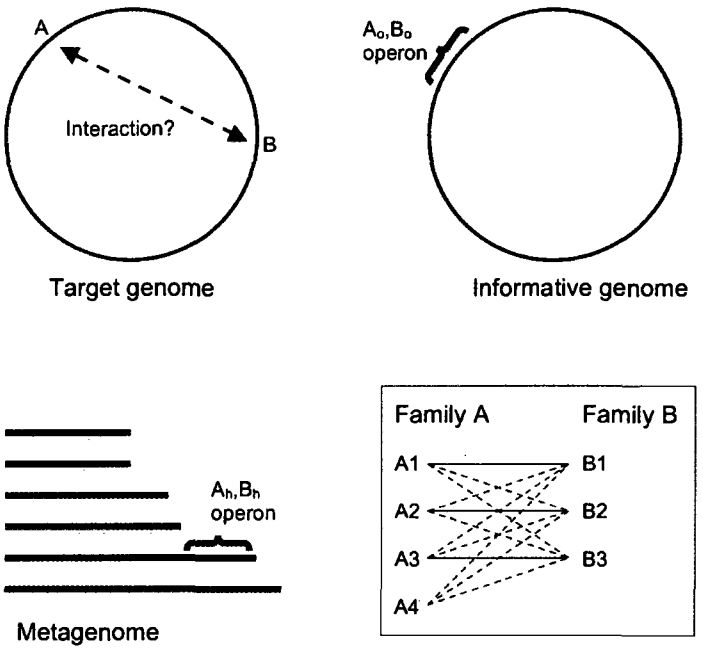


Figure 2-1 The problem of paralogy

Two genes, A and B, might be separated in a target genome. Yet, their orthologs, A_o and B_o, within an informative genome might be in the same operon, indicating that genes A and B might functionally interact in the target genome. In metagenome fragments, orthology cannot be inferred. Genes homologs to A and B, A_h and B_h, might indicate a functional interaction. However, if genes A and B belong to protein families with several paralogs, where each member of Family A interacts with a specific member of Family B (solid lines), there is a potential for a large number of false positives. In the example, we would infer three true positives (solid lines) and 9 false positives (dashed lines).

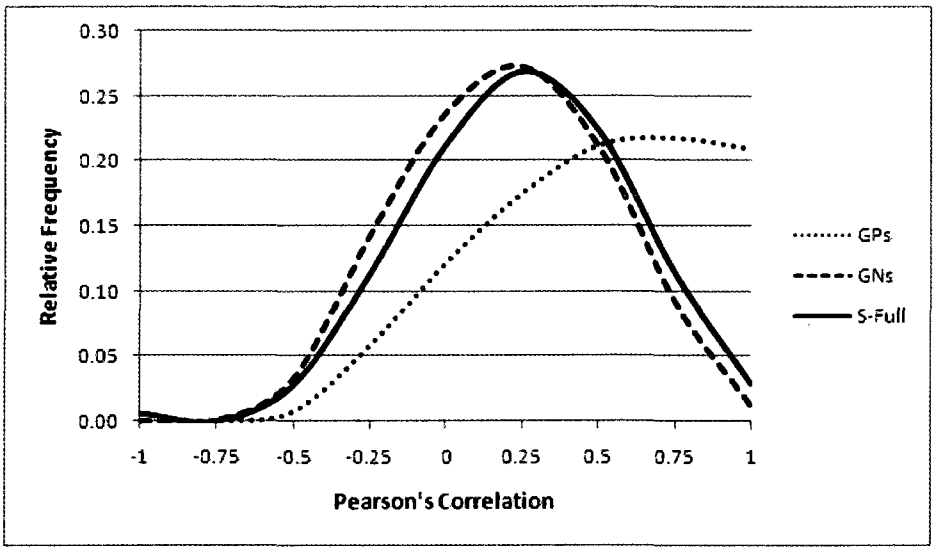


Figure 2-2 Relative frequencies of correlation of expression values

A graph of the relative frequencies of correlation of expression values for the EcoCyc gold negative functional interactions (GNs), the Ecocyc gold positive functional interactions (GPs), and the full set of predicted functional interactions from the Sargasso Sea metagenome (S-Full).

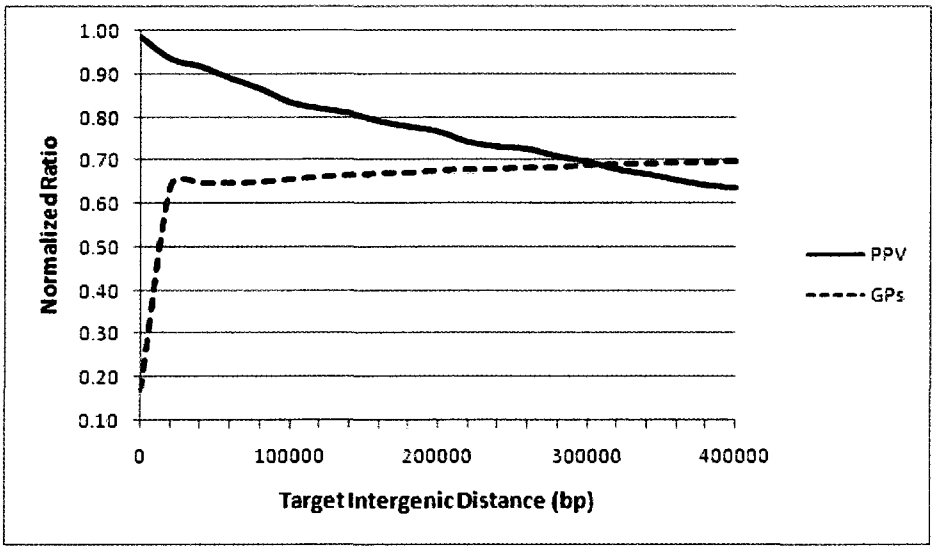


Figure 2-3 Target intergenic distance versus positive predictive value

A graph of positive predictive value (PPV) scores and proportions of EcoCyc gold positive functional interactions (GPs) versus maximum target intergenic distances (Distance) that are used as thresholds to reject predictions that exceed these maximum values.

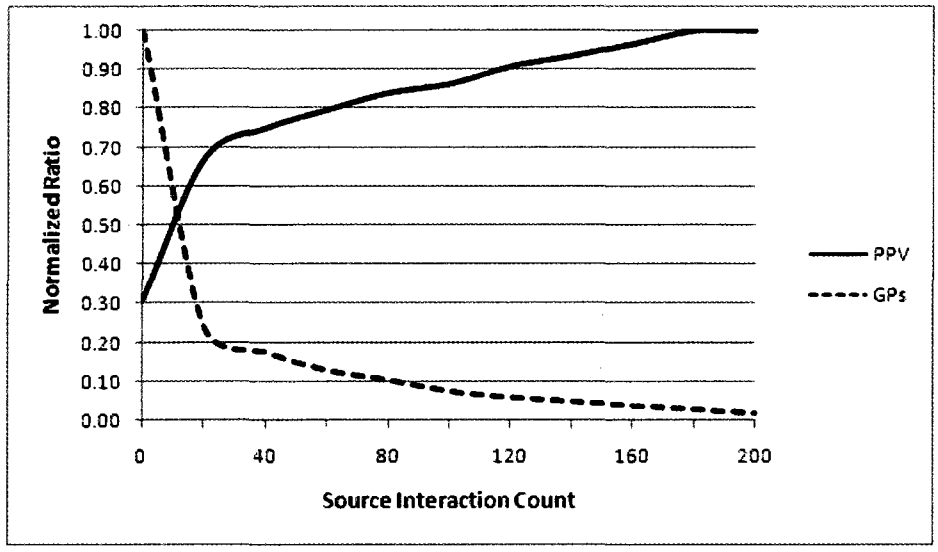


Figure 2-4 Source interaction count versus positive predictive value

A graph of positive predictive value (PPV) scores and proportions of EcoCyc gold positive functional interactions (GPs) versus minimum source interaction counts (Count) that are used as thresholds to reject predictions that do not meet these minimum values.

Figure 2-5 Functional interaction network for the *E. coli* K12 MG1655 genome

A functional interaction network for the *E. coli* K12 MG1655 genome derived from the prediction reduced Sargasso Sea metagenome, as viewed through Cytoscape [44].

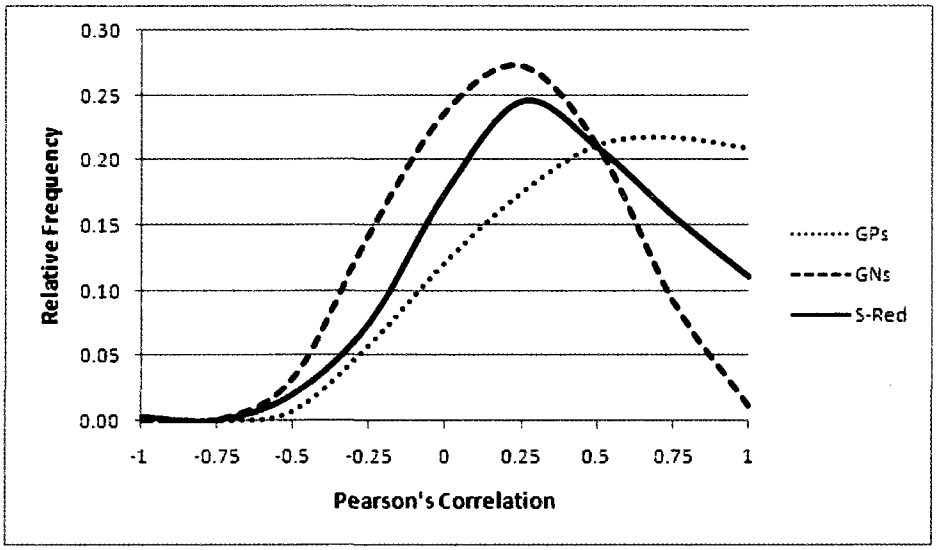


Figure 2-6 Relative frequencies of correlation of expression values

A graph of the relative frequencies of correlation of expression values for the EcoCyc gold negative functional interactions (GNs), the Ecocyc gold positive functional interactions (GPs), and the reduced set of predicted functional interactions from the Sargasso Sea metagenome (S-Red).

Chapter 3

Differential Direct Coding: A Compression Algorithm for Nucleotide Sequence Data

Gregory Vey[1§]

[1]Department of Biology, Wilfrid Laurier University, 75 University Avenue West, Waterloo ON, Canada, N2L 3C5

[§]Corresponding author

Email address:

GV: veyx9970@wlu.ca

Running Head: Differential direct coding

Keywords: nucleotide sequence, compression algorithm, direct coding, auxiliary data, metagenomes

Published as: Vey G: **Differential direct coding: a compression algorithm for nucleotide sequence data.** *Database: The Journal of Biological Databases and Curation* 2009, Vol. 2009:bap013; doi:10.1093/database/bap013.

3.1 Abstract

While modern hardware can provide vast amounts of inexpensive storage for biological databases, the compression of nucleotide sequence data is still of paramount importance in order to facilitate fast search and retrieval operations through a reduction in disk traffic. This issue becomes even more important in light of the recent increase of very large data sets, such as metagenomes. In this paper, I propose the Differential Direct Coding algorithm, a general-purpose nucleotide compression protocol that can differentiate between sequence data and auxiliary data by supporting the inclusion of supplementary symbols that are not members of the set of expected nucleotide bases, thereby offering reconciliation between sequence specific and general-purpose compression strategies. This algorithm permits a sequence to contain a rich lexicon of auxiliary symbols that can represent wildcards, annotation data, and special subsequences, such as functional domains or special repeats. In particular, the representation of special subsequences can be incorporated to provide structure-based coding that increases the overall degree of compression. Moreover, supporting a robust set of symbols removes the requirement of wildcard elimination and restoration phases, resulting in a complexity of $O(n)$ for execution time, making this algorithm suitable for very large data sets. Because this algorithm compresses data on the basis of triplets, it is highly amenable to interpretation as a polypeptide at decompression time. Also, an encoded sequence may be further compressed using other existing algorithms, like gzip, thereby maximizing the final degree of compression. Overall, the Differential Direct

Coding algorithm can offer a beneficial impact on disk traffic for database queries and other disk intensive operations.

3.2 Introduction

The field of bioinformatics necessitates a particular set of considerations, with respect to database management systems. A fundamental requirement is the capacity to warehouse large amounts of biological sequence data that are currently inundating the publicly available database resources. As of January 2009, the Nucleic Acids Research online Molecular Biology Database Collection listed 1170 publicly available biological databases [1]. GenBank, a major sequence database and a component of the International Nucleotide Sequence Databases (INSD), doubles in size roughly every 18 months [2]. Furthermore, biological data is distinct in that it requires accompanying annotation data in order for it to be useful [3]. While modern hardware can provide vast amounts of inexpensive storage, the compression of biological sequence data is still of paramount concern in order to facilitate fast search and retrieval operations, primarily by reducing the number of required I/O operations. Therefore, the effective management and compression of both sequence data and corresponding annotation data are indispensable considerations for biological database management systems.

Data compression requires two fundamental processes, modeling and coding [4]. Modeling involves constructing a representation of the distinct symbols in the data, along with any associated data, like the relative frequencies of the symbols [4]. Coding involves applying the model to each symbol in the data to produce a compressed representation of the data, preferably by assigning short codes to frequently occurring symbols and long

codes to infrequently occurring symbols [4]. A variety of dictionary methods, such as the Ziv-Lampel algorithms [5, 6], can be employed to achieve this [7]. Likewise, the Huffman algorithm [8] or some form of arithmetic coding could also be applied to yield a compaction in data [7]. However, methods that rely on evolving models may not perform adequately for sequences of genomic proportions. Such limitations will certainly be exacerbated by the recent surge in large-scale metagenomic data sets.

In the case of DNA sequences, the finite set of nucleotide symbols {A, C, G, T} can be efficiently modeled as a corresponding set of binary values {00, 01, 10, 11} [9]. This model constitutes an effective binary representation where each nucleotide base is directly coded by two bits. This assumes that sequence data is indeed composed solely from the four symbols of the nucleotide set. However, this assumption is not guaranteed to be met and a nucleotide sequence may include additional wildcard symbols, like N or S [4]. Therefore, to reconcile the potential occurrence of symbols other than the expected four nucleotide bases, any unexpected symbol is randomly converted into one of the valid symbols that it represents [4]. Eliminated wildcards are subsequently restored during sequence decompression [4].

The study of the compression of sequence data began with the work of Grumbach & Tahi [10, 11] and separately with the work of Milosavijevic [12] and the work of Rivals *et al.* [13]. Since then several major compression tools have been developed. While a variety of different underlying approaches have been employed, all of these efforts draw on the large body of existing work on general data compression, particularly text compression algorithms. In this work, I present the Differential Direct Coding

algorithm, a general-purpose nucleotide compression protocol that can differentiate between sequence data and auxiliary data by supporting the inclusion of supplementary symbols that are not members of the set of expected nucleotide bases, thereby offering reconciliation between sequence specific and general-purpose compression strategies.

3.3 Nucleotide Sequence Compression Strategies

3.3.1 Evolving models

Most previous approaches to nucleotide sequence compression consider a sequence as a finite length string of symbols where each nucleotide base corresponds to an individual symbol. On this basis, information content can be assessed and repeating patterns can be exploited using dictionary methods that progressively evolve models for data by encoding selected strings of symbols as tokens [7]. In general, dictionary-based compression protocols, such as the Ziv-Lampel algorithms [5, 6], are entropy encoders and will compress a string of n symbols to nE bits, where E is the entropy of the string [7].

While some sequence compression tools, like DNASequitur [14] and RNACompress [15], use grammar-based compression algorithms, most use some form of evolving model driven by a dictionary-based algorithm, typically derived from the Ziv-Lampel algorithms [5, 6]. Both biocompress [10] and biocompress-2 [11], along with GenCompress [16], DNACompress [17], DNAPack [9], and CASToRe [18 - 20] all involve the detection of approximate repeats to evolve a model for the encoding of a given sequence. While dictionary-based algorithms are often applied to string-like data to achieve general purpose compression, their effective use is contingent on having a

sufficiently large input file [7]. However, as input size increases, the running time of some algorithms becomes unmanageable, especially those that use greedy approaches for the selection of repeat segments [9]. Moreover, nucleotide sequences often need to be subdivided into discretely accessible records and this reduces the effectiveness of compression strategies that rely on evolving data models [4]. Arithmetic coding can be used to overcome this limitation but does not typically offer the speed required for modern database applications [4].

3.3.2 Direct coding

Williams and Zobel [4] developed a direct coding strategy for nucleotide sequence compression, including wildcard symbols. The first stage involves replacing each wildcard symbol with a random nucleotide from the set of nucleotides represented by the given wildcard [4]. Eliminated wildcards are maintained in a separate structure, rather than deleting them which would alter the semantics of the sequence [4]. After wildcard elimination, the resulting sequence is composed of only four different symbols corresponding to the four expected nucleotide bases and each base can be coded using two bits [4]. Instead of a space inefficient fixed-length integer representation, a variable-byte representation is used where seven bits are used to code an integer and the least significant bit indicates whether or not the current byte is followed by another byte [4]. Decompression requires two steps, the first of which involves mapping the two bit codes back to their nucleotide bases [4]. This is followed by decoding the wildcard tuples and overwriting nucleotide bases at the appropriate locations with the proper wildcard symbol [4].

Direct coding offers a rapid and uniform method of compression that is not affected by the size of the input file. However, wildcard elimination and restoration require at least a two-phase process for either compression or decompression operations. Furthermore, eliminated data requires storage in a secondary structure and that structure must include additional information about the location of its data for use at restoration time. Finally, sequences that have been compressed by direct coding are not readily re-compressible by alternative compression strategies that might increase the overall factor of compression.

3.4 Differential Direct Coding (2D)

3.4.1 Objectives

With the current surge in metagenomic data sets compression strategies must be developed to accommodate large data sets that are comprised of multiple sequences and a greater proportion of auxiliary data, such as sequence headers. Compression protocols developed specifically for sequence data offer good compression ratios but may perform poorly on large data sets or data sets that contain a significant amount of auxiliary data. In comparison, general-purpose compression utilities can easily compress large heterogeneous data files but cannot take advantage of the predominantly limited range of symbols that occur in sequence data. Therefore, the 2D algorithm is designed to provide a general-purpose nucleotide compression protocol that can differentiate between sequence data and auxiliary data, thereby offering reconciliation between the specific and general extremes of data compression. The following list enumerates the specific objectives of 2D:

- Linear execution time to support large data sets: Both compression and decompression operations must support implementations with a complexity of *O(n)* for execution time.
- Support for the inclusion of supplementary symbols that are not members of the set of expected nucleotide bases: Auxiliary symbols can be used to represent wildcards, annotation data, or special subsequences, such as functional domains or special repeats.
- Single phase direct coding: The compression phase must require only a single pass with no wildcard elimination phase and no storage of data in secondary structures or temporary intermediate files. Likewise, the absence of secondary data storage must permit a single pass restoration process for the decompression phase.
- Lossless compression: The original sequence must be obtained following decompression. This can be implemented either with respect to sheer sequence only, that is regardless of line breaks and formatting, or optionally with respect to the verbatim line-by-line layout of the original sequence data.
- Sequence type indifference: It must not be necessary to specify whether a given sequence is DNA or mRNA prior to compression or decompression.
- Polypeptide decompression: It must be possible to optionally restore a compressed nucleotide sequence directly to a polypeptide chain of amino acids based on an indicated reading frame.
- Amenable to further compression: A 2D encoded sequence must be readily compressible by other compression utilities to optionally provide potential further compression of the original sequence.

3.4.2 Model

To provide linear execution time, 2D uses a static model to encode sequence data along with any other content that may be contained within the input. For DNA 2D expects {A, C, G, T} and for mRNA 2D expects {A, C, G, U}. By taking the union of these sets, the set of expected symbols for the 2D model becomes {A, C, G, T, U}. This removes the burden of explicit declaration of sequence type. In the event of non-nucleotide symbols, 2D supports the set of traditional ASCII values, from 0 to 127, inclusive. The motivation for such a rich lexicon of symbols is not merely to accommodate the handful of wildcards. In addition to wildcards, the other ASCII symbols could be used to support the direct inclusion of annotation data or to denote special subsequences, such as functional domains or special repeats. The representation of domains and repeats through additional symbols can be optionally applied to add a degree of structure-based coding within the 2D protocol, thereby increasing the overall efficacy of the compression method. The values for the non-printing ASCII characters are particularly good candidates for reassignment since supporting them does not offer utility for wildcards or annotation data. Finally, 2D needs to support a single general-purpose value for occurrences of symbols that are not categorized by the two previously defined sets.

To achieve compression, it is necessary to represent multiple bases with a single byte, as in the two-bits-per-base schema. 2D uses direct coding on a triplet (three consecutive nucleotide bases) basis for the following reasons. First, this allows for three nucleotide bases to be consolidated into a single byte, rather than multiple bytes. Second,

by compressing on a triplet basis, rather than a two bit basis, unexpected symbols can be coded directly. This removes the need for a wildcard elimination phase and for storage of wildcard data in a secondary structure. This is beneficial both at compression time and decompression time. Last, representation in terms of triplets makes 2D highly amenable to decompression as a polypeptide sequence of amino acids by interpreting the triplets as codons.

The 2D model accommodates a total of 125 different triplets according to any of the nucleotide bases at any of the three triplet positions, such that the set of codons is {AAA, AAC, . . ., UUT, UUU}. Although some combinations should never occur because they violate the nucleotide base subsets for DNA and mRNA, such as UUT, these instances are accommodated in order to provide simplified arithmetic translation. Also, 128 different ASCII symbols are supported as extra symbols and a single unknown flag is included to denote a symbol that belongs to neither set. Table 1 shows the 2D model for representing symbols as either aggregate groups (triplets), wildcards or special data (single characters), or as unknown.

3.4.3 Coding

At the lowest level, 2D uses a signed byte that can range in value from -128 to 127 inclusive. Conceptually, the low seven bits of each byte are used for coding and the most significant bit is used as a compression flag. This schema is shown in Figure 1. Symbols are sequentially parsed into triplets if each member is a valid nucleotide base. A valid triplet is assigned a single value ranging from 1 to 125 inclusive and the compression flag is set, equating to assigning a value between -1 and -125 inclusive. 2D

will attempt to differentiate between sequence data and other symbols and if an unexpected value occurs that is interpretable as an ASCII value ranging from 0 to 127 inclusive, then this value is stored verbatim and the compression flag is not set, equating to assigning a value from 0 to 127 inclusive. In the event of an unexpected value, the other members of the current triplet must also be encoded individually and uncompressed, whether nucleotide bases or not, in order to maintain the current reading frame to support interpretation as an accurate polypeptide. By default, implementations can assume that the desired reading frame begins with the start of the sequence. However, multiple reading frames are easily supported by encoding the first symbol or the first two symbols as uncompressed data and then commencing the 2D process. Finally, in the event of an unknown symbol 2D denotes this by storing it uncompressed as the minimum possible signed byte value, -128. The values -126 and -127 are currently unused. Table 2 illustrates the 2D encoding steps to produce a compressed nucleotide sequence from an input string of symbols that includes an auxiliary symbol.

3.4.4 Algorithm

The following psuedocode describes the core 2D compression algorithm that takes an input string and returns a 2D encoding of the input sequence as a byte array. A more complete demonstration tool has been implemented using Java to support the Windows-1252 character set for Windows platforms and the MacRoman character set for Apple Macintosh platforms. This tool is available as an accompanying JAR file that will compress and decompress sequence data on the basis of entire files rather than individual strings. It should be noted that this particular implementation defines lossless in terms of

file sequence rather than specific line formatting. Decompressed data is restored into lines with lengths of mod 3. For example, if the source file's sequence was parsed into lines of 70 symbols each, then the restored file's sequence will have line lengths of 69, 69, 72, 69, 69, 72, etc. This was done in an effort to increase overall compression while maintaining readability. However, if required, a completely faithful line-by-line version can be easily implemented at the cost of a minor reduction in overall compression. Future efforts could include a purely byte based implementation, rather than character based, to maximize the degree of compression, particularly if file layout and formatting are not requisites. The use of blocked I/O should also be considered.

```
begin
        byte list = new List
        char triplet = new Array
        int baseCount = 0
        int nonCompressCount = 0
        foreach character c in input string
                if nonCompressCount = 0 then
                        if c is a nucleotide base then
                                triplet at position baseCount = c
                                baseCount = baseCount + 1
                                if baseCount = 3 then
                                        convert triplet to byte b and add b to list
                                        reset triplet
```

```
                    baseCount = 0
            else
                    foreach character t in triplet
                            convert t to byte b and add b to list
                    endfor
                    convert c to byte b and add b to list
                    reset triplet
                    nonCompressCount = 2 - baseCount
                    baseCount = 0
        else
                convert c to byte b and add b to list
                nonCompressCount = nonCompressCount - 1
    endfor
    return list as byte Array
end
```

3.4.5 Compression ratio

Because 2D uses a direct coding schema, its compression ratio, as defined by original size divided by encoded size, can be approximated by a general formula. Assuming a requirement of one byte to represent an uncompressed symbol as a character, the following considerations can be used to derive a predictive formula. If the sequence is assumed to be composed only of nucleotide bases and has a length of L symbols and therefore a size of L bytes, then its encoded size will be (L / 3 + L mod 3) bytes which is

the sum of all triplets plus any remaining symbols. However, it is likely that auxiliary symbols will occur at some approximate frequency. Since the occurrence of one or more of such symbols within a given triplet will cause all of the triplet members to be encoded at a cost of one byte each, there is an added cost of two bytes to each triplet (this triplet now requires three bytes instead of one) that contains one or more auxiliary symbols. Therefore, two bytes must be added to the encoded size for each occurrence of an auxiliary symbol and there will be ⌊aL⌋ such symbols, where a is the frequency of auxiliary symbols and the auxiliary symbols are randomly distributed, rather than packed together. Thus, the size of a 2D encoded sequence can be approximated by the following general formula:

$$Encoded\ size \bullet (L / 3 + L \bmod 3 + 2\lfloor aL \rfloor)\ bytes$$

This formula can be substituted into the original definition for compression ratio to provide a general formula for the 2D compression ratio:

$$Compression\ ratio \bullet L\ bytes / (L / 3 + L \bmod 3 + 2\lfloor aL \rfloor)\ bytes$$

3.4.6 Benchmarking

In order to test 2D, it was used to compress several bacterial genomes and its performance was compared against several other compression utilities. The *Bacillus subtilis* and *Escherichia coli* K12 MG1655 genomes were selected because they are commonly used model genomes and the *Mycoplasma genitalium* genome was selected because of its small size and the expectation that some of the compression utilities may perform poorly with sequence data of genomic proportions. All genomes were downloaded from the NCBI FTP server and the files were not modified in any way,

thereby conserving the header data as well as the actual genomic sequence. Except for GenCompress, all compression utilities were run on an iMac5,1 with 3GB of memory. The MS-DOS executable for GenCompress was run on a Gateway laptop with comparable hardware and 1GB of memory. It should be noted that the benchmarking process itself incurs a certain amount of computational overhead and therefore may introduce an artifact of inflated execution times. However, this effect can be minimized by using sufficiently long sequences.

The results show that gzip provided the best compression ratios while 2D had the fastest execution times. If 2D was applied and then followed immediately with gzip, this provided the best compression ratios and at execution times that were still faster than gzip alone. The MS-DOS executable for GenCompress failed before completion after a considerable execution time, even for the smallest genome. Despite the similarity in compression ratios for the 2D compressed genomes the frequencies of the auxiliary symbols were 2.1E-05 (89 out of 4214719) for *Bacillus subtilis*, 1.9E-05 (88 out of 4639763) for *Escherichia coli* K12 MG1655, and 1.3E-04 (73 out of 580149) for *Mycoplasma genitalium*. However, in all cases the auxiliary symbols were contained only in the sequence header, a single line FASTA identifier at the beginning of each file. Therefore, the actual sequences were compressed uniformly and the overall compression ratios were similarly impacted by the condensed occurrence of a similar number of auxiliary symbols at the start of each file. Table 3 summarizes the compression results.

Decompression for 2D was also tested by restoring the 2D compressed genomes. A consistent file size increase of one byte was observed in all cases along with an

increase in file length of one line. Unless a sequence has a last line length that is divisible by three when combined with any symbols that may already be cached in the compression buffer, then there will be either one or two remainder symbols. The current implementation will treat any remainder symbols as uncompressible symbols and deposit them on their own line at the end of the compressed sequence. In the case of the test genomes, the compressed files became one line longer than their source files because they each had remainder symbols that were uncompressible. This resulted in the creation of one new line for each compressed file and this increase was propagated during decompression. To verify this, the last line of symbols from each decompressed file was merged with the previous line and both the original line count and original file size were restored. Table 4 shows the decompression results.

To test its robustness for use with very large data sets, 2D was used to compress the Sargasso Sea metagenome, a 918.1MB FASTA format file available from the Sorcerer II Expedition website [21]. This file is interesting because it contains a very large ratio of auxiliary data to sequence data since the metagenome is broken into a vast number of individual FASTA records rather than having a single header at the beginning. 2D performance was measured against gzip, bzip2, and against 2D in combination with gzip. As with the genomes, 2D had a faster execution time than gzip, while gzip had a better compression ratio. Moreover, bzip2 yielded an even better compression ratio in slightly less time than gzip but was considerably slower than 2D. However, the combination of both 2D and gzip produced the best compression ratio in less time than gzip alone or bzip2. Table 5 summarizes the results for compression of the metagenome.

It was observed that 2D read 11,418,321 lines from the source file but wrote 11,959,572 lines to the compressed file resulting in a gain of 541,251 lines and a definite decrease in the compression ratio that was obtained for the metagenome. The Sargasso Sea metagenome is composed of 811,372 sequence fragments. Since each sequence begins with a header of auxiliary symbols, any remaining symbols from a previous sequence are written to their own line before processing the upcoming header. The current implementation does this in an effort to maintain human readability between sequences. Future implementations should abandon this behaviour to improve the overall compression ratio.

3.5 Conclusion

2D provides a general-purpose nucleotide compression protocol that can differentiate between sequence data and auxiliary data thereby offering reconciliation between sequence specific and general-purpose compression strategies. This makes 2D suitable for any type of sequence data, including very large data sets, such as metagenomes. Because it supports the inclusion of auxiliary symbols that are not members of the set of expected nucleotide bases, the source sequence can contain a rich lexicon of added symbols that can represent wildcard symbols, annotation data, or special subsequences, such as functional domains or special repeats. The representation of domains and repeats through additional symbols can be applied to add a degree of structure-based coding within the 2D protocol, thereby providing a means to increase the overall degree of compression. Also, the encapsulation of unexpected symbols within the primary representation removes the need for a wildcard elimination phase and storage of

wildcard data in a secondary structure. This is also a benefit at decompression time when unexpected symbols must be restored. 2D employs compression by triplets making the compressed representation immediately amenable to interpretation as a polypeptide. 2D encoded sequences may be subsequently compressed by other compression protocols to further the overall degree of compression as demonstrated by its combination with gzip. 2D has the potential to have a beneficial impact on disk traffic for database queries and other disk intensive operations.

3.6 Funding

Gregory Vey and this work were supported by funds from a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant to Gabriel Moreno-Hagelsieb.

3.7 Acknowledgments

I thank Gabriel Moreno-Hagelsieb for reviewing the manuscript and Andre Masella for testing the accompanying JAR file.

3.8 References

[1] Galperin MY, Cochrane GR: **Nucleic Acids Research annual Database Issue and the NAR online Molecular Biology Database Collection in 2009.** *Nucleic Acids Res.* 2009, **37**(Database issue):D1-4.

[2] Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: **GenBank.** *Nucleic Acids Res.* 2008, **36**(Database issue):D25-30.

[3] Hoebeke M, Chiapello H, Gibrat JF, Bessieres P, Garnier J: **Annotation and databases: status and prospects.** In *Database Annotation in Molecular Biology.* Edited by Lesk AM. West Sussex, England: John Wiley & Sons; 2005:1-21.

[4] Williams HE, Zobel J: **Practical compression of nucleotide databases.** In *Proc. Australian Computer Science Conference: January 31-February 2 1996; Melbourne, Australia.* 1996:184-193.

[5] Ziv J, Lempel A: **A universal algorithm for sequential data compression.** *IEEE Trans. Inform. Theory* 1977, **23**(3):337-342.

[6] Ziv J, Lempel A: **Compression of individual sequences via variable-rate coding.** *IEEE Trans. Inform. Theory* 1978, **24**(5):530-536.

[7] Salomon D: *A Concise Introduction to Data Compression.* London: Springer; 2008.

[8] Huffman DA: **A method for the construction of minimum-redundancy codes.** *Proc. IRE* 1952, **40**:1098-1101.

[9] Behzadi B, LeFessant F: **DNA compression challenge revisited.** In *Symposium on Combinatorial Pattern Matching: June 19-22 2005; Jeju Island, Korea.* 2005:190-200.

[10] Grumbach S, Tahi F: **Compression of DNA sequences.** In *Proc. IEEE Symposium on Data Compression: March 30-April 1 1993; Snowbird, Utah.* 1993:340-350.

[11] Grumbach S, Tahi F: **A new challenge for compression algorithms: genetic sequences.** *J. Information Processing and Management* 1994, **30**(6):875-866.

[12] Milosavljevi• A: **Discovering sequence similarity by the algorithmic significance method.** *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 1993, **1**:284-291.

[13] Rivals E, Dauchet M, Delahaye JP, *et al*: **Compression and genetic sequence analysis.** *Biochimie* 1996, **78**(5):315-322.

[14] Cherniavski N, Lander R: **Grammar-based compression of DNA sequences.** *University of Washington Computer Science & Engineering Technical Report* 2004, **2007-05-02**:1-21.

[15] Liu Q, Yang Y, Chen C, *et al*: **RNACompress: grammar-based compression and informational complexity measurement of RNA secondary structure.** *BMC Bioinformatics* 2008, **9**:176.

[16] Chen X, Kwong S, Li M: **A compression algorithm for DNA sequences.** *IEEE Engineering in Medicine and Biology Magazine* 2001, **20**(4):61-66.

[17] Chen X, Li M, Ma B, *et al*: **DNACompress: fast and effective DNA sequence compression.** *Bioinformatics* 2002, **18**:1696-1698.

[18] Bonanno C, Galatolo S, Menconi G: **Information of sequences and applications.** *Physica A* 2002, **305**:196-199.

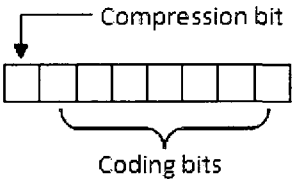
[19] Menconi G: **Sublinear growth of information in DNA sequences.** *Bulletin of Mathematical Biology* 2005, **67**(4):737-759.

[20] Menconi G, Benci V, Buiatti M: **Data compression and genomes: a two dimensional life domain map.** *Journal of Theoretical Biology* 2008, **253**(2):281-288.

[21] Venter JC, Remington K, Heidelberg JF, *et al*: **Environmental shotgun sequencing of the Sargasso Sea.** *Science* 2004, **304**:66-74.

3.9 Figures

Figure 3-1 The 2D byte coding schema



The seven least significant bits are used to encode data. The most significant bit is used as a flag to indicate the context of the byte as either compressed data or uncompressed data.

3.10 Tables

Table 3-1 The 2D data model

Type	Description	Range	Compressible
Auxiliary	ASCII	0 to 127	No
Sequence	Triplet	-1 to -125	Yes
Unknown	?	-128	No

For sequence data, auxiliary data, and unknown values the range of byte values is listed as well as whether the data will be compressed or uncompressed.

Table 3-2 The 2D encoding process

Step	Input Sequence	Triplet	Uncompress Count	Encoded Sequence
0	ACTCNTGAGA	empty	0	empty
1	CTCNTGAGA	A	0	empty
2	TCNTGAGA	AC	0	empty
3a	CNTGAGA	ACT	0	empty
3b	CNTGAGA	empty	0	~

4	NTGAGA	C	0	~
5a	TGAGA	empty	0	~C
5b	TGAGA	empty	0	~CN
6	GAGA	empty	1	~CNT
7	AGA	G	0	~CNT
8	GA	GA	0	~CNT
9a	A	GAG	0	~CNT
9b	A	empty	0	~CNTÀ
10	empty	A	0	~CNTÀA

An example of encoding process is given for the sequence ACTCNTGAGA that contains the auxiliary symbol N. The remaining input symbols, any symbols cached in the triplet structure, the value of the uncompress count (a variable to offset compression after the occurrence of an auxiliary symbol), and the encoded sequence are shown for each step in the process.

Table 3-3 Genomic compression benchmarking

Compression Method	**Source Genome**								
	Bacillus subtilis			*Escherichia coli K12 MG1655*			*Mycoplasma genitalium*		
	Size (bytes)	Ratio	Time (ms)	Size (bytes)	Ratio	Time (ms)	Size (bytes)	Ratio	Time (ms)
None	4,274,929	1.000	N/A	4,706,046	1.000	N/A	588,437	1.000	N/A
GenCompress	0	Ø	58,363,756	0	Ø	27,887,599	0	Ø	8,127,438
2D	1,465,177	2.918	717.5	1,612,930	2.918	788.9	201,721	2.917	100.5
gzip	1,300,308	3.288	1,671.3	1,431,844	3.287	1,819.4	174,398	3.374	254.5
2D + gzip	1,093,657	3.909	824.9	1,214,444	3.875	891.3	145,727	4.038	182.8

Compression data for GenCompress, 2D, gzip, and 2D + gzip was obtained using three bacterial genomes. File size, compression ratio, and execution time are given for each algorithm with respect to each genome. Execution time is the average result from 100

trials with the exception of GenCompress which is the shortest execution time obtained after three consecutive failures.

Table 3-4 Genomic decompression benchmarking

Source Genome	File Size (bytes)			File Inflation (bytes)		Decomp Time (ms)
	Normal	2D Comp	2D Decomp	bytes	lines	
Bacillus subtilis	4,274,929	1,465,177	4,274,930	1	1	923.9
Escherichia coli K12 MG1655	4,706,046	1,612,930	4,706,047	1	1	1,042.3
Mycoplasma genitalium	588,437	201,721	588,438	1	1	116.2

Decompression data was obtained using the 2D compressed genomes. File sizes are given for the original source file, the compressed file, and the decompressed file, with respect to each genome. The differences between the original sizes and the restored sizes are also given along with the respective execution times. Execution time is the average result from 100 trials.

Table 3-5 Metagenomic compression benchmarking

Compression Method	Sargasso Sea Metagenome		
	Size (bytes)	Ratio	Time (ms)
None	962,651,334	1.000	N/A
2D	419,368,931	2.295	145,115.0
gzip	261,995,558	3.674	315,564.6
bzip2	238,973,241	4.028	301,924.0
2D + gzip	220,487,270	4.366	153,175.8

Compression data for 2D, gzip, bzip2, and 2D + gzip was obtained using the Sargasso Sea metagenome. File size, compression ratio, and execution time are given for each algorithm. Execution time is the average result from 5 trials.

3.11 Additional Files

File 3-1 2D.jar

A java implementation of the 2D algorithm was developed and compiled using the JDK version 1.5.0_19. This demonstration tool provides compression and decompression of sequence data using the Windows-1252 character set for Windows platforms or the MacRoman character set for Apple Macintosh platforms. The demonstration tool represents a simplified implementation and is not intended to be a robust and exhaustively tested software tool.

Chapter 4

General Discussion

4.1 Contributions to the Field of Metagenomics

4.1.1 Functional inference from metagenomic context

Because metagenomes can reach previously inaccessible microbes, the discovery of novel enzymes and novel functionalities can have tremendous impact on a variety of applied fields such as medicine, agriculture, and industry [4.1]. Likewise, a process to harness metagenomes as a data source for functional inference has the potential to benefit these same fields by revealing novel functional associations for genomes of interest. By furthering the characterization of metabolic pathways countless ventures can be facilitated, including drug design and engineering pathogen resistance.

The first manuscript demonstrated that in the absence of detectable orthologous relationships it remains possible to make high quality functional inferences. This offers a strategy for harnessing other metagenomes and homologs in general. Because metagenomes allow access to previously unreachable microorganisms, this will result in expanding the universe of known functional interactions thus furthering our understanding of functional organization and enhancing our effectiveness at assigning functional annotations.

Although a functional interaction network was derived for the *Escherichia coli* K12 MG1655 genome using the Sargasso Sea metagenome, this result primarily represents a proof of the viability of the proposed process. Future work should use multiple metagenomes as a data source to make functional inferences across multiple

target genomes. Of particular interest is the relationship between the volume and type of source data versus the number of predicted functional interactions. This could provide an indication of to what extent the metagenomes actually extend microbial biodiversity and the repertoire of novel genes and novel functional interactions. Further attention should be devoted to exploring whether or not orthology should remain a necessary requisite for in conventional microbial genomics. Perhaps prediction viability metrics or a predictive formula derived from binary logistic regression could all together eliminate the need for establishing orthology.

4.1.2 Differential direct coding

As the prominence of the field of metagenomics continues to grow, there will be an intensification in research that relies on the efficient storage and retrieval of very large data sets. The development of the general-purpose nucleotide compression protocol can potentially have a beneficial impact on disk traffic for database queries and other disk intensive operations that involve sequence data. Moreover, it is possible that such compressed sequence representations might have future utility for pursuits such as the detection of novel patterns and subsequences.

The second manuscript presented an algorithm that uses a general-purpose nucleotide compression protocol that can differentiate between sequence data and auxiliary data. This provides reconciliation between sequence specific and general-purpose compression strategies thus making the algorithm suitable for very large data sets, such as metagenomes.

Future implementations should use a byte stream implementation, rather than a character stream implementation to explore the potential gain in compression ratio. Also, certain common and fixed sequences, like stop codons, could be encoded using some the non-printing ASCII characters that are currently allocated to represent auxiliary symbols. This should be explored with the goal of further increasing the compression ratio by using some amount of structure-based coding.

4.2 Future Research Directions

Metagenomics, like other areas of computational biology, is driven by user-friendly software [4.2]. A variety of generic tools could potentially benefit the research community. Therefore, any future versions of the research presented in this work should be formulated from the perspective of useful and extensible software. Effective implementation is a crucial aspect in bringing any proposed computational techniques into actual usage [4.2]. Applying fundamental principles from software engineering could greatly facilitate the design and maintenance of such projects.

Horizontal gene transfer (HGT) has been extensively studied in the completed genomes and a similar undertaking could be performed using the metagenomes [4.3]. Although this would require an adaptation of methods from the genomic approach, metagenomic studies of HGT could reveal patterns of prokaryotic evolution [4.3]. Metagenomic studies could be used to complement existing genomic studies [4.4] of codon usage and codon richness index to compare the relationships between recipient genomes and donated genes. Moreover, the connection between environmental factors and species composition versus the frequency of HGT events could also be an important

relationship that can only be characterized through metagenomic data [4.4]. A better understanding of HGT is indispensible to furthering our knowledge about the evolution of natural microbial communities [4.4].

As well the previously discussed bias in the databases toward cultivable organisms, there may be similar bias in validation metrics, such as the genomic correlation of expression data [4.5]. It is arguable how applicable this data is for benchmarking metagenomic functional interactions since the metagenomes are likely to contain novel proteins that necessarily exhibit novel functional interactions, as well as instances of novel functional interactions among previously characterized proteins. Therefore, appraising the validity of this and other validation metrics is essential in order to properly assess the results of future metagenomic research.

The metagenomes offers a perspective where functional modules form the atomic units of conceptualization, rather than the organisms that encapsulate them. This provokes a consideration of the validity of many traditional constructs in the biological sciences. For example, the accepted relationship between gene and protein has always mandated a one-to-one cardinality. However, this may be an artifact of conceptual convenience that has propagated to every corner of biological thought, rather than a rigidly understood stochastic rule. There is a growing body of evidence in functional genomics, proteomics, and epigenetics that points to organization greater than an encapsulated unit of inheritance resting at a fixed chromosomal locus. Genomes are not flat files; they exhibit robust topologies that defy the simplicity of a one-to-one cardinality. Perhaps our entire perspective on genomes has been skewed by the

tremendous impact of the one-gene-one-protein model. Exploring the validity of current ontology represents a colossal yet essential undertaking toward achieving a truly integrated biology.

4.3 Toward a Post-metagenomic Era

Addressing challenges to the field of metagenomics requires development in the areas of computation, technology, methodology, and conceptual perspectives [4.6]. Several major opportunities have been identified for metagenomics in relation to various application areas [4.6]. From a life sciences perspective, metagenomes have the potential to advance theory and predictive power in microbiology and evolution, while from an earth sciences perspective, genome-based microbial models of ecosystems could be used to predict global environmental processes [4.6]. Better understanding the biosynthetic and biocatalytic potential of microbes has immediate utility for a variety of pursuits in biotechnology, while understanding how the human microbiome contributes to health and disease will facilitate biomedical research [4.6]. Microbial communities also have the potential to drive environmental remediation by providing restoration to various ecosystems, and also to maximize the efficiency of agricultural practices that involve both plants and animals [4.6]. Even the need for economical and sustainable energy can potentially be addressed with microbes by harnessing of bioenergy resources [4.6].

Microbial communities are a major component of the biosphere, yet little is known about these communities and their dynamics [4.6]. By addressing current constraints and exploiting current opportunities, the scope of metagenomics can be extended thus paving the way for a post-metagenomic era of research. Harnessing the

metagenomes represents one of the remaining great frontiers in the biological sciences. Ultimately, the metagenomes will provide a definitive rendering of microbial biodiversity that will cascade into many facets of biology and address questions about the diversity of life, the ecological and evolutionary roles of viruses, and even what defines a species [4.6] With the current surge in biotechnological techniques and computational resources it is at last possible to propel biology into the forefront of the sciences and metagenomics will play a key role in achieving this goal.

Literature Cited

[1.1] Ahmed N: **A flood of microbial genomes–do we need more?** *PLoS ONE* 2009, **4**(6):e5831. doi:10.1371/journal.pone.0005831.

[1.2] Overbeek R: **Genomics: what is realistically achievable?** *Genome Biol.* 2000, **1**(2):COMMENT2002.

[1.3] Eisenberg D, Marcotte EM, Xenarios I, Yeates TO: **Protein function in the post-genomic era.** *Nature* 2000, **405**(6788):823-6.

[1.4] Yellaboina S, Goyal K, Mande SC: **Inferring genome-wide functional linkages in E. coli by combining improved genome context methods: comparison with high-throughput experimental data.** *Genome Res.* 2007, **17**(4):527-35.

[1.5] Fields S, Kohara Y, Lockhart DJ: **Functional genomics.** *Proc Natl Acad Sci U S A.* 1999, **96**(16):8825-6.

[1.6] Ferrer M, Martínez-Abarca F, Golyshin PN: **Mining genomes and 'metagenomes' for novel catalysts.** *Curr Opin Biotechnol.* 2005, **16**(6):588-93.

[1.7] Tringe SG, Rubin EM: **Metagenomics: DNA sequencing of environmental samples.** *Nat Rev Genet* 2005, **6**:805-814.

[1.8] Riesenfeld CS, Schloss PD, Handelsman J: **Metagenomics: Genomic analysis of microbial communities.** *Annu Rev Genet* 2004, **38**:525-552.

[1.9] Pignatelli M, Aparicio G, Blanquer I, Hernández V, Moya A, Tamames J: **Metagenomics reveals our incomplete knowledge of global diversity.** *Bioinformatics* 2008, **24**(18):2124-5.

[1.10] Kunin V, Copeland A, Lapidus A, Mavromatis K, Hugenholtz P: **A bioinformatician's guide to metagenomics.** *Microbiol Mol Biol Rev.* 2008, **72**(4):557-78.

[1.11] McHardy AC, Martín HG, Tsirigos A, Hugenholtz P, Rigoutsos I: **Accurate phylogenetic classification of variable-length DNA fragments.** *Nat Methods* 2007, **4**(1):63-72.

[1.12] McHardy AC, Rigoutsos I: **What's in the mix: phylogenetic classification of metagenome sequence samples.** *Curr Opin Microbiol.* 2007, **10**(5):499-503.

[1.13] Huynen M, Snel B, Lathe W 3rd, Bork P: **Predicting protein function by genomic context: quantitative evaluation and qualitative inferences.** *Genome Res.* 2000, **10**(8): 1204-10.

[1.14] Fitch WM: **Homology a personal view on some of the problems.** *Trends Genet.* 2000, **16**(5):227-31.

[1.15] Fitch WM: **Distinguishing homologous from analogous proteins.** *Syst. Zool.* 1970, **19**:99-113.

[1.16] Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA: **Protein interaction maps for complete genomes based on gene fusion events.** *Nature* 1999, **402**(6757):86-90.

[1.17] Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D: **A combined algorithm for genome-wide prediction of protein function.** *Nature* 1999, **402**(6757):83-6.

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

[1.19] Tatusov RL, Koonin EV, Lipman DJ: **A genomic perspective on protein families.** *Science* 1997, **278**(5338):631-7.

[1.20] Gaasterland T, Ragan MA: **Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes.** *Microb Comp Genomics* 1998, **3**(4):199-217.

[1.21] Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO: **Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.** *Proc Natl Acad Sci U S A* 1999, **96**(8):4285-8.

[1.22] Rogozin IB, Makarova KS, Murvai J, Czabarka E, Wolf YI, Tatusov RL, Szekely LA, Koonin EV: **Connected gene neighborhoods in prokaryotic genomes.** *Nucleic Acids Res.* 2002, **30**(10):2212-23.

[1.23] Snel B, Bork P, Huynen MA: **The identification of functional modules from the genomic association of genes.** *Proc Natl Acad Sci U S A* 2002, **99**(9):5890-5.

[1.24] Hu P, Janga SC, Babu M, Díaz-Mejía JJ, Butland G, Yang W, Pogoutse O, Guo X, Phanse S, Wong P, Chandran S, Christopoulos C, Nazarians-Armavil A, Nasseri NK, Musso G, Ali M, Nazemof N, Eroukova V, Golshani A, Paccanaro A, Greenblatt JF, Moreno-Hagelsieb G, Emili A: **Global functional atlas of Escherichia coli encompassing previously uncharacterized proteins.** *PLoS Biol.* 2009, **7**(4):e96.

[1.25] Janga SC, Collado-Vides J, Moreno-Hagelsieb G: **Nebulon: a system for the inference of functional relationships of gene products from the rearrangement of predicted operons.** *Nucleic Acids Res.* 2005, **33**(8):2521-30.

[1.26] Singh AH, Doerks T, Letunic I, Raes J, Bork P: **Discovering functional novelty in metagenomes: examples from light-mediated processes.** *J Bacteriol.* 2009, **191**(1):32-41.

[1.27] Harrington ED, Singh AH, Doerks T, Letunic I, von Mering C, Jensen LJ, Raes J, Bork P: **Quantitative assessment of protein function prediction from metagenomics shotgun sequences.** *Proc Natl Acad Sci U S A* 2007, **104**(35):13913-8.

[1.28] National Center for Biotechnology Information: http://www.ncbi.nlm.nih.gov/.

[1.29] Genomes OnLine Database: http://genomesonline.org/.

[2.1] through [2.43]: See Chapter 2 Reference section (page 24).

[3.1] through [3.21]: See Chapter 3 Reference section (page 60).

[4.1] Ferrer M, Beloqui A, Timmis KN, Golyshin PN: **Metagenomics for mining new genetic resources of microbial communities.** *J Mol Microbiol Biotechnol.* 2009, **16**(1-2):109-23.

[4.2] Editorial: **Sofware by any name.** *Nature Methods* 2009, **6**:547-8.

[4.3] Tamames J, Moya A: **Estimating the extent of horizontal gene transfer in metagenomic sequences.** *BMC Genomics* 2008, **9**:136.

[4.4] Medrano-Soto A, Moreno-Hagelsieb G, Vinuesa P, Christen JA, Collado-Vides J: **Successful lateral transfer requires codon usage compatibility between foreign genes and recipient genomes.** *Mol Biol Evol.* 2004, **21**(10):1884-94.

[4.5] Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS: **Large-scale mapping and validation of Escherichia coli**

transcriptional regulation from a compendium of expression profiles. *PLoS Biol.* 2007, **5**(1):e8.

[4.6] Committee on Metagenomics, Board on Life Sciences, Division on Earth and Life Studies, National Research Council of the National Academies: *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet.* Washington, DC, United States of America: The National Academies Press; 2007.