2015

Working Together: Using protein networks of bacterial species to compare essentiality, centrality, and conservation in Escherichia coli.

Christopher Wimble
Virginia Commonwealth University, wimblecf@mymail.vcu.edu

Working Together: Using protein networks of bacterial species to compare essentiality, centrality, and conservation in *Escherichia coli*.

This thesis is submitted in partial fulfillment of the requirements for the degree of Master of Science in Bioinformatics at Virginia Commonwealth University.

Christopher Wimble
VCU Life Sciences
Center for the Study of Biological Complexity
Grace E. Harris Hall
PO Box 842030
Richmond, VA 23284-2030
(804) 921-7256
wimblecf@vcu.edu

Dr. Peter Uetz
Center for the Study of Biological Complexity
Harris Hall, room 3136
(804) 827-4573 phuetz@vcu.edu
http://www.people.vcu.edu/~phuetz

Delivered: April 10th, 2015
Approved: April 28th 2015

Dedication

I dedicate this thesis to my loving wife, Melissa. Without her support and patience this thesis would not have been possible. Her advice and instruction in both computer software packages and in the written word have been of immeasurable value. The love and emotional support have been an inspiration driving the work.

Acknowledgements

There are many I would like to thank for the completion of this project. My appreciation goes to Dr. Tarynn Witten for introducing me to scholar's lunch and for our weekly meetings where she shared her vision. This vision provided the direction for my undergraduate work which, thanks to her, culminated in a textbook chapter. Many thanks go to Dr. Danail Bonchev, for his patience in explaining difficult concepts in centrality and graph theory, as well as the use of his office for the use of a Pathway Studio database that he helped build. I'd also like to thank Dr. Peter Uetz and the entire Uetz lab for helping me with both direction and useful input for my graduate work. Special thanks go to Harry Caufield for frequent help, answering of questions, and for sharing data. Thanks also go to Dr. Paul Fawcett, Dr. Fernando Tenho, and Jeff Elhai Ph.D. for their help and inspiration throughout my academic career. Last but certainly, not least, I would also like to thank Nihar Sheth for his patience in being one of the first people to introduce me to programming.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

Abstract

WORKING TOGETHER: USING PROTEIN NETWORKS OF BACTERIAL SPECIES TO COMPARE ESSENTIALITY, CENTRALITY, AND CONSERVATION IN *ESCHERICHIA COLI*.

Christopher Wimble, Masters of Science in Bioinformatics

This thesis is submitted in partial fulfillment of the requirements for the degree of Master of Science in Bioinformatics at Virginia Commonwealth University.

Virginia Commonwealth University, 2015

Dr.Peter Uetz, associate professor, Center for the Study of Biological Complexity

Proteins in *Escherichia coli* were compared in terms of essentiality, centrality, and conservation. The hypotheses of this study are: for proteins in *Escherichia coli*, (**1**) there is a positive, measureable correlation between protein conservation and essentiality, (**2**) there is a positive relationship between conservation and degree centrality, and (**3**) essentiality and centrality also have a positive correlation. The third hypothesis was supported by a moderate correlation, the first with a weak correlation, and the second hypotheis was not supported. When proteins that did not map to orthologous groups and proteins that had no interactions were removed, the relationship between essentality and conservation increased to a strong relationship. This was due to the effect of proteins that did not map to orthologus groups and suggests that protein orthology represented by clusters of orthologus groups does not accurately dipict protein conservation among the species studied.

Keywords: Essentiality, Protein Conservation, Centrality, Graph Theory, Protein-Protein Interaction Network, PPI, *Escherichia coli*, *Saccharomyces cerevisiae*, Baker's Yeast, Network Biology, Aging, Replicative Aging, Target of Rapamycin, TOR, Interactomics

Introduction & Background

Hypotheses and reasons for study:

The hypotheses of this study are: for proteins in *Escherichia coli* (**1**) there is a positive correlation between protein conservation and essentiality and that this relationship is measurable, (**2**) that there is also a positive relationship between conservation and degree centrality, and (**3**) that essentiality and degree centrality are positively correlated.

These questions will be addressed using bioinformatic and systems biology approaches which will look at the topology of protein-protein interaction networks, protein essentiality in *Escherichia coli* and conservation of those proteins in ten other species. Understanding how model organisms operate is an important first step towards understanding how human systems operate. Proteins are of particular interest because of their importance to how cells operate. The hypotheses will be tested by comparing data representing protein conservation, centrality, and essentiality. Comparisions will be evaluated using correlations which establish acceptance or rejection of the hypotheses. The goal is to establish if there are relationships between these aspects of proteins to improve understanding of how they operate.

Bioinformatics and systems biology:

The reductionist perspective has widely been used to study phenomena and has worked well due to the fact that complicated problems can be broken up into a collection of simpler problems. However, there are weaknesses with this method which have been addressed. Considering parts of a biological system individually misses how those parts are interrelated which is very important to understanding how such systems function. Now that large amounts of biological information are available, a systems biology perspective has gained popularity. Such a perspective incorporates the use of a holistic approach with reductionist techniques to better understand how these pieces are interconnected.

To make sense of large amounts of data, bioinformatics has established itself as a discipline and as a set of techniques to assist with this task. Bioinformatics is a term used to describe the application of computers to address biological questions. It is an interdisciplinary approach incorporating aspects of statistics, mathematics, organic chemistry, engineering, computer science, and biology. The term was coined in the early 1970's by Pauline Hogeweg and Ben Hesper who originally defined it as the study of informatics processes in biological systems <19>. With the flood of DNA sequence data, the term is more commonly used to describe the analysis of genomic data. However, the approach is used to study a wide variety of different

types of biological data from proteins to tree populations. An area of study that uses bioinformatic techniques is interactomics, which is the study of molecular interactions in cells. Interactomics involves the construction of biological networks and principles of graph theory are used for their analysis.

Graph theory:

Graph Theory is a branch of mathematics that utilizes interaction networks to study how pieces of a system are interrelated and as a tool is used to study how parts of systems interrelate. In these networks each object is represented by a node and each connection between nodes is an edge. Nodes might represent genes, proteins, people, or any other type of subject under study. The edges could be a wide variety of associations, such as physical interaction between proteins. It takes a top-down, rather than bottom-up perspective which can yield insights, such as identifying the best candidates for further study <15, 34>. As early as 1736 graph theory has been used to solve a wide variety of problems and since the 1980's pioneering scientists have used networks to study biological processes <1, 3, 5, 14, 15, 46, 47>. Some of the first applications were in the study of social networks. It is also now widely used in the study of protein-protein interaction networks (PPI or PIN) <30>.

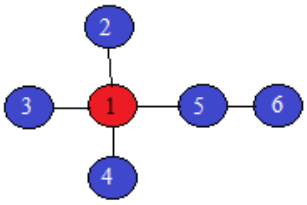
Protein-protein interaction networks:

The analysis of protein-protein interaction networks is one of the more frequently used methods in interactomics. In these networks the nodes are generally proteins, either specific to the organism or a member of an orthologous group and the edges of such networks can represent different types of interactions, such as: physical, genetic, or regulation <11, 41>. Understanding how proteins interrelate is critical to understanding how a cell operates and PPI networks are used in the study of such connections <30>. Proteins often function in complexes and or in concert with other proteins and because of this it is frequently very difficult to understand the function of a protein in isolation and without an understanding of its relationship to other proteins <11>. Such networks are often used as a starting point to understand how a cell operates <30>. Some uses of PPI networks are to predict protein function, suggest relative importance of interacting members, and assist in choosing the best targets for drug therapies. These studies are not without challenges. Just because two proteins have been shown to interact does not necessarily mean that they actually interact within the cell. Graph theory can be used in the study of and in the improvement of the quality of PPI networks <34>.

Centrality and network topology:

One of the techniques used in graph theory to study PPI networks is the concept of centrality. Centrality is a measurement that helps to establish the relative importance of nodes. There are several ways to calculate this. The most basic is degree centrality, which is the sum of connections for a node. Closeness centrality is a measure of how close a node is to every other node. It is determined for a node by taking the inverse of all of the shortest distances between it and any other node. The shortest distance between any two nodes is also called geodesic distance and is the smallest number of edges between two nodes. Betweenness measures how often a node

acts as a bridge between two other nodes. It is the ratio of the sum of all shortest paths between any two nodes in a network that contain a node divided by the total sum. See Figure 1 for a demonstration of how these values are calculated. Each of these measurements has a large number of applications including but not limited to: establishing essentiality, determining the robustness of a network, predicting protein function, and minimizing drug side-effects <1>.



Node 1 (red)

Degree: 4
This is number of nodes it connects to.
Closeness: 0.833
= 5(number of other nodes)/6(total of the number of edges in between it and the other nodes)
Betweeness: 0.9
= 10(number of paths between each two nodes, 2-3, 2-4, ect)/9(number of paths that include node 1)

Figure 1: Centrality measurements. This figure demonstrates how degree, closeness, and betweenness centrality are calculated. Each of the measurements is calculated for node 1. It is connected to 4 other nodes and so has a degree of 4. Node 1 has a total distance, measured in number of edges of 6 between it and each of the 5 different nodes. The number of other nodes divided by the total distance gives the closeness centrality value. For each non-redundant path through the network between any two nodes, node 1 is part of that path 9 out of 10 times. The number of paths divided by the number of times node 1 is part of that path gives the betweenness centrality score.

There are a number of other ways to analyze networks. Node density, also referred to as network density, is the ratio of actual connections by the number of potential connections in the network. It is calculated by doubling the number of edges and dividing that by the number of nodes multiplied by the number of nodes minus one. Vertex degree range shows the range of values for how many edges nodes in the network possess and the mean vertex degree is the average of these values. Network diameter is the longest of the shortest paths between any two nodes. The average number of edges of the shortest paths between any two nodes is the mean node distance <3>.

Networks often have nodes that are highly connected which are referred to as Hubs. In some cases components of a network will be unconnected to any other part, which are called Islands. Many networks, especially biological networks, are considered to follow a power-law distribution, have small-world properties, and are scale free. Power-law describes a statistic relationship where one value varies as a power of the other. A network can be said to have small world properties if the distance, in terms of edges, between nodes is smaller than one would expect by chance (5). Networks that have these characteristics are often referred to as scale-free <14>.

Protein essentiality and conservation:

The networks in this study are networks of protein-protein interactions. Proteins are so essential, so central to how life operates that the primary function of DNA, the language of life, is to code for their assembly. As a result of their importance, there are several ways that they can be studied. One of those ways is the concept of essentiality. A protein is considered to be essential if when the gene that codes for that protein is removed, the organism survives. This is not to say that if a protein is non-essential that its function is not necessary for the organism's survival. Often processes performed by a protein are also performed by other proteins. However essentiality is a good method for establishing the relative importance of proteins.

One of the attributes of a protein considered in this paper is protein conservation. Orthologous groups are a way to compare proteins across species classifying them in terms of similar structure, such as clusters of orthologous groups (COG), which is a way of categorizing proteins. Each COG is identified by a number representing a protein that has orthologs in at least three different species <31>. By using these categories, proteins can be compared among species to determine how often similar proteins are found. If similar proteins are found in several species it can be said to be well conserved.

Other studies comparing essentiality, centrality, and conservation:

The relationships between protein essentiality, centrality, and conservation are well studied. Several studies have discovered a positive relationship between a protein's essentiality and the number of interactions it has with other proteins <22, 25>. However other studies found this not to be true for binary interactions in model organisms <48, 50>. This is further complicated by other factors, such as the effect of non-essential promiscuous proteins <48>. Other studies suggest that it is not essentiality and protein connectivity but instead essentiality and pleiotropy that are related <50, 19>. It has been suggested that essential genes would be more conserved and would be under more evolutionary constraint. This has not been found to be true for several model organisms but may be for bacterial species <19, 23>. It is possible that this was not found due to conserved non-essential genes <16>. The relationship between these attributes of proteins is complex and depends on the function of proteins as well as differences in environment between organisms <27, 36, 49>.

Organisms studied:

In order to compare protein essentiality, centrality, and conservation *Escherichia coli* was chosen as the basis for the study. As a model organism it is a good candidate because it is so heavily studied. There are a large number of databases dedicated to collecting and storing information on this one organism alone. Two examples of such databases are EcoGene, which was used in the construction of the list of proteins used as a key, and EcoCyc. EcoGene provides nearly extensive lists of genes and proteins, along with aliases for those genes and proteins while EcoCyc contains metabolic and signal-transduction pathways for *Escherichia coli* <24, 35>.

There are many ways to identify a protein in *Escherichia coli,* some of which include B-number, JW number, Uniprot ID, and locus ID. When the genome of *Escherichia coli* was first sequenced the genes were assigned B-numbers in the order that they were found <4>. Uniprot ID is from UniProt, a database of protein sequence and functional information <13>. Both B-number and UniprotID are frequently used to identify proteins in *Escherichia coli.*

Ten other bacterial species were chosen to determine protein conservation through orthology. Like *Escherichia coli*, *Campylobacter jejuni* and *Helicobacter pylori* are gram-negative bacteria that can live in the digestive track. The species *Bacillus subtilis* also can live in the digestive track but is gram-postive. Although these four species can exist in the gut, they each have complicated life-cycles and are not restricted to that environment. Two of the other species, *Caulobacter crescentus* and *Synechocystis* live in water with the first being gram-negative and the second is gram-positive. The next five species are pathogens. *Streptococcus sanguinis* and *Streptococcus pneumoniae* are gram-positive. One is a blood-borne disease agent; the other can colonize the nose and is a cause of pneumonia. *Mycoplasma pneumoniae* has more than half as many genes as *Streptococcus*. The last two species *Treponema pallidum* and *Mycoplasma genitalium* are causes of sexually-transmitted diseases and *Mycoplasma* is often used to study the minimal genome. *Escherichia coli, Bacillus subtilis, Caulobacter crescentus,* and *Synechocystis* have similar genome sizes, each being about 4 million base pairs. While *Streptococcus sanguinis, Streptococcus pneumoniae, Campylobacter jejuni, Helicobacter pylori,* and *Treponema pallidum* have about half that amount, between 2.4 and 1.4 million base pairs. *Mycoplasma pneumoniae* and *Mycoplasma genitalium* both have fewer than 1 million base pairs, 800 kilobases and 600 kilobases respectively. The choices are intended to provide a range of genome sizes, large, medium, and small. See Table 1 for more information on the species studied.

Table 1: Organisms studied. This contains information for 11 species studied including the name of the organism, substrain, reference number in the eggNOG database, genome size in base pairs, number of genes, and number of COGs (clusters of orthologous groups, see Protein conservation section above).

OrganismName	Substrain	GenomeSize(BP)	Genes	COGs
Escherichia coli	K-12 (MG1655)	4.6 million	4288	917
Bacillus subtilis	168	4.2 million	4100	765
Caulobacter crescentus	CB15	4 million	3767	754
Synechocystis	PCC 6803	3.6 million	3618	628
Streptococcus sanguinis	ATCC 49296	2.4 million	2,274	177
Campylobacter jejuni	NCTC 11168	1.6 million	1654	451
Streptococcus pneumoniae	ATCC 700669	2 million	1553	335
Helicobacter pylori	26695	1.7 million	1550	374
Treponema pallidum	Nichols	1.4 million	1041	268
Mycoplasma pneumoniae	M129	0.8 million	687	124
Mycoplasma genitalium	G37	0.6 million	390	149

Related Works

There are three bodies of work that relate to the thesis, one of which was a study of aging using *Saccharomyces cerevisiae* as a model organism, the Aging Yeast Network study (AYN). In another the PPI networks of eight bacterial species were compared to determine the conservation of proteins and their interactions, the Bacterial protein-protein Interaction Conservation study (BIC). Finally a PPI meta-network was constructed from the networks of six bacterial species in order to illustrate and study protein conservation, which will be referred to as the Bacterial Meta-interactome Network study (BMN).

The Aging Yeast Network Study (Introduction and Aims):

The goal of the aging yeast network study was to study the interplay of networks of two cellular processes with a network of genes associated with aging. One of the networks, the Cellular Response to Heat (CRH), was chosen because it was suspected to have a link to aging and is a good model for how an organism deals with environmental stressors, an aspect of aging. Aging is often described as a limitation in the ability to handle biological stresses. The Target of Rapamycin (TOR) was chosen because of its well established link to aging and thus could be used as a control.

Aging is a complex phenomenon and there are several ways to measure its effects. In unicellular species there are two primary ways of measuring aging: replicative and chronological. Chronological aging measures the length of time individual cells can live while in a non-replicating state, while Replicative aging measures how many times a cell can replicate before dying. Replicative life span is generally considered to be a more applicable model to study aging as it relates to more complicated organisms <6>.

The Aging Yeast Network Study (Methods):

First a set of proteins associated with aging was collected from the literature, and supplemented with information from databases. The core of this list of genes is from the paper “Shortest-Path Network Analysis Is A Useful Approach Toward Identifying Genetic Determinants of Longevity” <23>. This core was expanded using yeast gene databases including; The Saccharomyces Genome Database (SGD), YEASTRACT, The Comprehensive Yeast Genome Database (CYGD), The NetAge Database, Sageweb, and AmiGO <7, 10, 18, 31, 33, 40, 42>. This list of proteins was reduced to only proteins associated with replicative aging. It was

further reduced to those in which the organism has an increase in replicative life span when not present. This was done because the genes that code for these proteins represent druggable targets. Since lifespan increases when these proteins are not present it is thought that they, or the genes that code for them, can be targeted to increase lifespan. The Replicative Life Span (RLS) list was the basis for comparison for the other two lists, the Target of Rapamycin (TOR), and the Cellular Response to Heat (CRH) lists. Lists of proteins for the TOR and CRH were chosen based on Gene Ontology (GO) terms from the AmiGO database. Proteins chosen for the TOR list were involved in the Target of Rapamycin pathway, while members of the CRH list were involved in the heat-shock response pathway. A list of all proteins from the RLS, TOR, and CRH was also complied, the Total protein list (TOT). The TOT list was constructed as a means of comparison for the other three lists.

For each of the four lists of proteins, networks were constructed from a database of interacting proteins from yeast using the Pathway Studio software <28>. See Figure 3 for an example of the RLS network. In each case two networks were constructed, a Direct Connect (DC), and a Shortest Path (SP) network. The Direct Connect (DC) networks contained the interactions between proteins within their respective list, while the Shortest Path (SP) networks also included the interactions of the proteins in the list with their nearest neighbors. In other words the SP networks also included all the proteins that each protein on the DC list had an interaction with. In all there were eight networks; the Replicative Life Span Direct Connect (RLS DC), Replicative Life Span Shortest Path (RLS SP), Target of Rapamycin Direct Connect (TOR DC), Target of Rapamycin Shortest Path (TOR SP), Cellular Response to Heat Direct Connect (CRH DC), Cellular Response to Heat Shortest Path (CRH SP), Total Direct Connect (TOT DC), and Total Shortest Path (TOT SP).

Once the interactions were collected for each of the lists, PPI networks were constructed using Cytoscape and Pajek <38>. Using these two software packages the networks were both illustrated and analyzed. Degree, Betweenness, and Closeness Centrality was calculated using Network Analyzer, an add-on for Cytoscape, while Eigenvector Centrality was calculated using Pajek <2>. See Table 2 and 3 for example values from the RLS SP and TOT DC networks respectively. Other network calculations were also performed using Excel. These measurements, along with centrality values were used to study these networks as well as their relationships to the other networks.

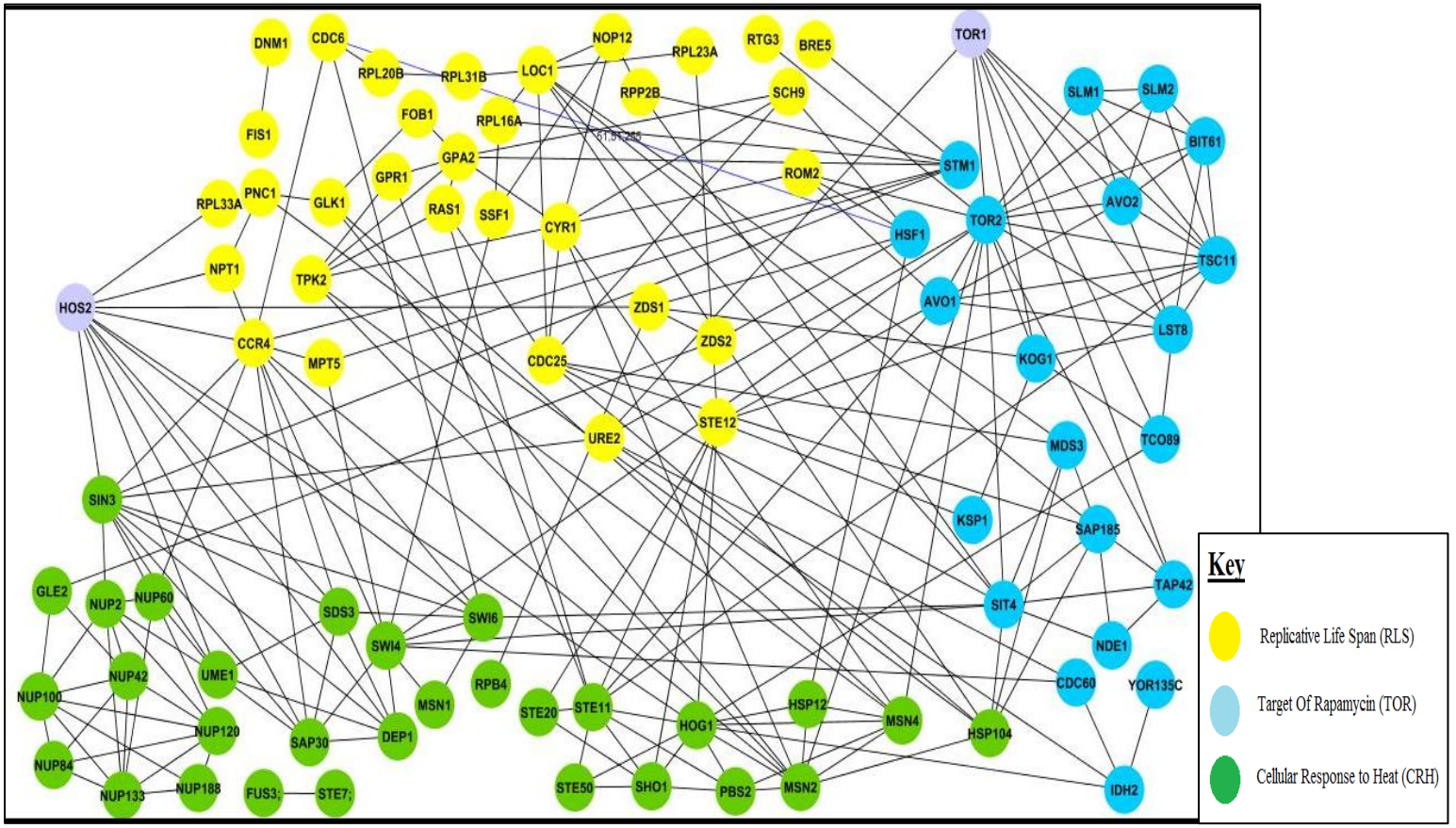


Figure 2: TOT network. The total (TOT) PPI network includes the proteins of the Replicative Life Span (RLS) network (in Yellow), the Target of Rapamycin (TOR) network (in Blue), and the Cellular Response to Heat (CRH) network (Green). Each node in the network is a protein and is identified by name. Protein interactions are represented by lines connecting proteins. The protein TOR1 is shared between the RLS and TOR networks, while HOS2 is shared between the CRH and RLS networks. This figure was generated using Cytoscape.

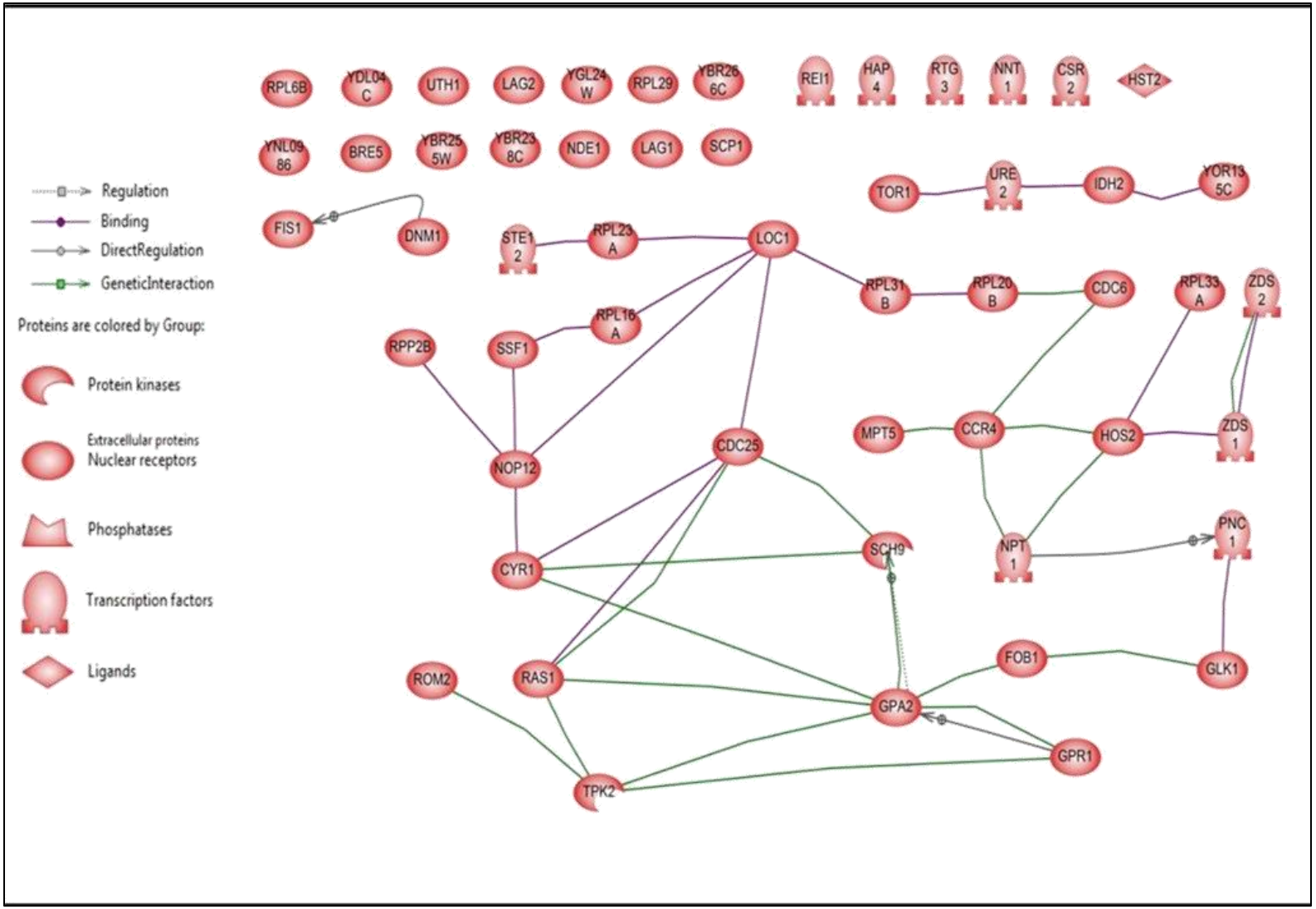


Figure 3: RLS network. The Replicative Life Span (RLS) PPI network is composed of a collection of proteins that when knocked-out have the effect of increasing replicative life span in yeast. Proteins are classified by function and interactions are classified by type. This was generated using Pathway Studio.

Table 2: RLS SP centrality measurements samples. This table is a sample of the centrality measurements from the Replicative Life Span (RLS) network for yeast. For each listing, protein name is given followed by degree, betweenness, closeness, and eigenvector centrality measurements. The measurements for this table were calculated using Network Analyzer, an add-on for Cytoscape.

	TOT Centrality Measurements			
Name	Degree	Betweenness	Closeness	Eigenvector
ASM4	0	0.00	0.00	0.00
AVO1	5	0.01	0.22	0.11
AVO2	6	0.00	0.20	0.12
BIT61	5	0.00	0.20	0.11
BOI2	5	0.03	0.19	0.01
BRE5	1	0.00	0.17	0.00
CDC25	8	0.02	0.20	0.32
CDC6	2	2.24E-03	0.16	0.01
CDC60	3	0.01	0.20	0.01
CSR2	2	0.00	0.18	0.01
CYR1	2	0.00	0.17	0.10
DEP1	5	0.00	0.21	0.02
ELP4	1	0.00	0.19	0.01
FOB1	0	0.00	0.00	0.00
FUS3	0	0.00	0.00	0.00
GCN4	2	0.00	0.20	0.01
GLE2	3	0.01	0.15	9.27E-04
GLK1	0	0.00	0.00	0.00
GPA2	1	0.00	0.02	1.65E-49
GPR1	0	0.00	0.00	0.00
HOG1	14	0.19	0.28	0.12
HOS2	7	0.07	0.24	0.04
HSF1	2	0.01	0.18	0.01
HSP104	8	0.05	0.23	0.32

Table 3: TOT DC centrality measurements samples. This table is a sample of the centrality measurements from the Total (TOT) network for yeast. For each listing, protein name is given followed by degree, betweenness, closeness, and eigenvector centrality measurements. These measurements were calculated using Network Analyzer, an add-on for Cytoscape.

	RLS SP Centrality Measurements			
Name	Degree	Betweenness	Closeness	Eigenvector
ACE2	4	2.38E-03	0.31	0.02
APL2	3	3.26E-03	0.38	0.05
ARG5,6	3	1.14E-03	0.35	0.05
ARO1	9	1.92E-02	0.40	0.13
ARP2	24	4.87E-02	0.41	0.29
ARR4	7	0.01	0.37	0.07
ARX1	4	0.00	0.31	0.03
BOI1	4	4.70E-04	0.29	0.02
BOI2	10	0.01	0.36	0.05
BRE5	7	4.93E-03	0.38	0.09
BSD2	3	3.83E-03	0.36	0.04
CAF17	2	9.69E-04	0.29	0.01
CBR1	3	0.00	0.32	0.03
CDC25	12	0.02	0.39	0.08
CDC6	9	0.01	0.37	0.06
CFT1	2	1.29E-03	0.35	0.03
CHS1	4	0.00	0.30	0.02
CLA4	15	0.04	0.42	0.19
COG5	4	0.00	0.30	0.02
CSR2	5	0.01	0.38	0.05
CTK2	2	0.00	0.34	0.03
CYR1	14	0.03	0.39	0.08
CYS4	10	0.02	0.39	0.10
CYT1	2	5.81E-04	0.37	0.05

The Aging Yeast Network Study (Conclusions):

The AYN study found that the TOR and RLS networks were well connected. It was determined that the TOR and CRH networks were densely connected to the RLS network. However there were relatively few connections between the CRH and TOR networks. It was discovered that there was a protein shared between the TOR and RLS networks,TOR1, and one shared between the CRH and RLS networks, HOS2 (Histone deacetylase). See Figure 2 for more information on the relationships between the RLS, TOR, and CRH networks.

Mean Vertex Degree (MVD) was noticeably different between the shortest-path and direct connect networks. It was more pronounced in the TOR network, the shortest-path MVD being nearly double what it was in the direct connect. This was even more noticeable between the total networks, which was more than double. There was a large difference in node densities with the shortest-path networks having ones lower than the direct connect. Between the RLS networks this was far less pronounced with the shortest-path having half the node density of the direct connect. The TOR shortest-path node density was a tenth of the direct connect. The network diameter was similar for the TOR direct connect and the shortest-path yet for the RLS and total networks the shortest-path diameter was half what it was in the direct connect. See Table 4 for more information on the measurements for each of the networks.

It was expected that the TOR and RLS networks would be highly connected but it is interesting to note that the CRH and RLS networks were also well-connected. This suggests that cellular response to heat is related to the aging process and would make a good model for studying how a cell's response to thermal stresses relates to aging. It was known that TOR1 has a relationship to aging but HOS2 would be a further candidate for study on its relationship to replicative aging.

Table 4: Network property calculations. This table contains calculations of several network properties from the yeast networks. RLS corresponds to the replicative life span network, CRH to the cellular response to heat, TOR to the target of rapamycin network, TOT refers to the combined networks, DC refers to the direct connect version of that network, and SP to the network that includes the nearest neighbors of each member.

Network Property Calculations							
Network	Number of Nodes	Number of Edges	Vertex Degree Range	Node Density	Mean Vertex Degree	Mean Node Distance	Network Diameter
CRH DC	36	100	0 to 16	0.159	5.778	3.146	7
CRH SP	123	524	0 to 35	0.070	106.750	2.821	6
RLS DC	45	52	0 to 8	0.053	2.311	4.087	8
RLS SP	168	540	1 to 46	0.039	86.378	2.893	6
TOR DC	21	51	0 to 13	0.243	4.857	2.892	7
TOR SP	332	1542	2 to 110	0.028	85.667	2.855	5
TOT DC	100	171	0 to 14	0.035	5.780	3.804	8
TOT SP	470	2483	0 to 110	0.023	92.550	3.175	6

Fragmentation (Direct Connection)	0.630
Fragmentation (Shortest Path)	0.306
Clustering Coefficient (Direct Connection)	0.151
Clustering Coefficient (Shortest Path)	0.046

The Bacterial protein and protein Interaction Conservation Study (Introduction):

The goal of the Bacterial protein and protein Interaction Conservation (BIC) study was to determine the degree to which proteins and the interactions between them are conserved between bacterial species. There were two separate comparisons as part of the study. In the first study four bacterial species were compared in terms of their protein content as well as their interactions. In the other eight, bacterial species were compared in terms of conservation of proteins. In both studies, bacterial species were chosen for which there was protein and PPI data. The data was mined from the literature <9, 20, 29, 32, 39, 43, 44>. Both studies found far less conservation than expected.

The Bacterial protein and protein Interaction Conservation Study (Methods):

The four species that were compared for the first study were: *Escherichia coli*, *Helicobacter pylori*, *Treponema pallidum*, and *Streptococcus pneumoniae.* Each of the bacterial species was compared to *E. coli*, which was used as a basis of comparison. Statistics were generated using Excel and a network was constructed using Cytoscape (see Figure 4). Proteins were compared using orthologous groups (OG).

In the other study the species compared were: *M. pneumoniae*, *M. genitalium*, *B. subtilis*, *S. sanguinis*, *H. pylori*, *C. crescentus*, *P. aeruginosa*, and *E. coli*. As before, they were compared by assigning them to OGs. As before a network of the combined interactions of the species was constructed using Cytoscape (see Figure 5). They were compared using Excel and tables were generated showing conservation in terms of numbers and percentages. The comparison was performed two ways. In one, paralogous proteins were included; in the other, they were removed by only including unique OGs.

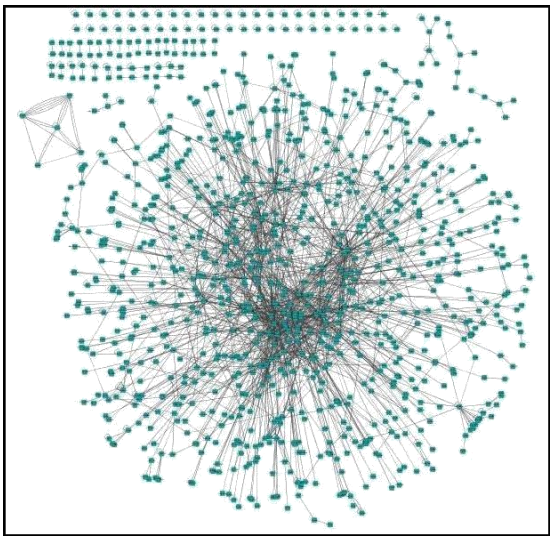


Figure 4: Four species PPI. An interaction network was constructed using Cytoscape for the interaction data from 4 species: *E. coli*, *H. pylori*, *T. pallidum*, and *M. tuberculosis*.

Table 5: Shared clusters of orthologous groups (COG) among four species. PPI interaction data was compared for four organisms using Excel. Each organism was compared to *E. coli*. Table A indicates number of shared orthologous groups represented by COGs, while Table B shows number of shared protein-protein interactions.

A: Shared COGs				
Organism Name	*E. coli*	*H. pylori*	*T. pallidum*	*M. tuberculosis*
E. coli	1269	226	68	386
H. pylori	226	917	69	290
T. pallidum	68	69	576	78
M. tuberculosis	386	290	78	2907

B: Shared Interactions				
Organism Name	*E. coli*	*H. pylori*	*T. pallidum*	*M. tuberculosis*
E. coli	2231	3	1	16
H. pylori	3	2154	0	6
T. pallidum	1	0	992	2
M. tuberculosis	16	6	2	8042

Table 6: Summary statistics. Four species; *E. coli*, *H. pylori*, *T. pallidum*, and *M. tuberculosis* were compared in terms of shared interactions and shared COGs to study conservation. These values were calculated using Excel formulas.

Statistics Compared to *E. coli*	Number	Percentage
Total Number of COGs in *E. coli*	1269	
Number of COGs Shared by all Groups	6	0.50%
Number of COGs Shared by Three Groups	38	3%
Number of COGs Shared by Two Groups	123	9.70%
Number of COGs Unique to *E. coli*	1146	90.30%
Total Number of Interactions Present in *E. coli*	2231	
Number of Interactions Shared by all Groups	0	0%
Number of Interactions Shared by Three Groups	2	0%
Number of Interactions Shared by Two Groups	40	1.80%
Number of Interactions Unique to *E. coli*	2191	98.20%

Table 7: Proteins shared among eight species. Section 7A shows the percentages of OGs shared between two of eight species, along with the total number of OGs for that species. Table 7B shows the counts of the OG that are shared between two species. Tables 7C and 7D are the same except that paralogy has been removed. One organism compared to its self shows the total number of proteins for that species in the dataset. The above table 7A shows percentage of proteins in common. The organism name to the left is the one started with and the name in the top row is the one it is compared to. The bottom table, 7B, shows the numbers of proteins that the species share. Tables 7C and 7D are the same except that paralogy has been removed.

Table 7A: Percentages	*M. pneumoniae*	*M. genitalium*	*B. subtilis*	*S. sanguinis*	*H. pylori*	*C. crescentus*	*P. aeruginosa*	*E. coli*
M. pneumoniae		86.02%	71.55%	71.05%	55.91%	63.06%	65.72%	67.89%
M. genitalium	96.68%		80.50%	79.05%	60.17%	69.09%	73.24%	73.65%
B. subtilis	15.88%	15.46%		48.67%	33.88%	51.80%	60.77%	59.86%
S. sanguinis	26.37%	25.44%	74.22%		39.41%	58.14%	65.74%	67.60%
H. pylori	21.43%	20.75%	57.61%	43.07%		62.53%	66.83%	67.71%
C. crescentus	12.58%	12.20%	57.21%	41.21%	38.83%		72.56%	66.91%
P. aeruginosa	11.74%	11.34%	56.72%	40.75%	34.00%	62.61%		69.40%
E. coli	14.33%	13.68%	58.34%	43.21%	34.11%	58.12%	73.03%	

Table 7B: Values	*M. pneumoniae*	*M. genitalium*	*B. subtilis*	*S. sanguinis*	*H. pylori*	*C. crescentus*	*P. aeruginosa*	*E. coli*
M. pneumoniae	601	517	430	427	336	379	395	408
M. genitalium	466	482	388	381	290	333	353	355
B. subtilis	644	627	4056	1974	1374	2101	2465	2428
S. sanguinis	538	519	1514	2040	804	1186	1341	1379
H. pylori	314	304	844	631	1465	916	979	992
C. crescentus	454	440	2064	1487	1401	3608	2618	2414
P. aeruginosa	678	655	3275	2353	1963	3615	5774	4007
E. coli	594	567	2418	1791	1414	2409	3027	4145

Table 7C: Percentages	*M. pneumoniae*	*M. genitalium*	*B. subtillis*	*S. sanguinis*	*H. pylori*	*C. crescentus*	*P. aeruginosa*	*E. coli*
M. pneumoniae		69.72%	0.16%	57.07%	43.26%	49.08%	53.24%	54.08%
M. genitalium	86.93%		71.16%	69.92%	53.32%	61.00%	65.35%	65.56%
B. subtillis	8.56%	8.46%		24.21%	17.14%	25.64%	30.52%	30.23%
S. sanguinis	16.81%	16.52%	48.14%		25.49%	36.52%	42.55%	43.38%
H. pylori	17.75%	17.54%	47.44%	35.49%		48.94%	53.79%	53.65%
C. crescentus	8.18%	8.15%	28.82%	20.65%	19.87%		38.53%	34.89%
P. aeruginosa	5.54%	5.46%	21.44%	15.03%	13.65%	24.07%		28.84%
E. coli	7.84%	7.62%	29.58%	21.35%	18.96%	30.37%	40.17%	

Table 7D: Values	*M. pneumoniae*	*M. genitalium*	*B. subtilis*	*S. sanguinis*	*H. pylori*	*C. crescentus*	*P. aeruginosa*	*E. coli*
M. pneumoniae	461	419	347	343	260	295	320	325
M. genitalium		435	343	337	257	294	315	316
B. subtilis			2582	982	695	1040	1238	1226
S. sanguinis				1437	520	745	868	885
H. pylori					1180	717	788	786
C. crescentus						2231	1390	1259
P. aeruginosa							3112	1665
E. coli								2593

The Bacterial protein and protein Interaction Conservation Study (Results):

Only 6 proteins out of the 1,269 OGs in *E. coli* (0.50%) were present in all four groups, 38 were shared by three (3%), and 123 (9.7%) were shared by two or more. Over 90% of the proteins were unique to *E. coli*. Only interactions for which there was data was considered, and it is unlikely that these represent all the interactions present in each species. Thus it is also possible that there might be false positives, as well as false negatives. No interactions were shared between all four, only two were shared out of the 2231, and 40 (1.80%) were shared by two or more. Over 98% of the interactions were unique to *E. coli*. See Table 5 for shared COGs and Table 6 for shared interactions. It was concluded for this comparison that (1) the interaction networks are vastly incomplete and that (2) protein interactions in bacteria are less well conserved than previously thought.

In the study of 8 species, it was found that species that were more similar in terms of evolutionary hierarchy shared more proteins. When paralogy was included *M. genitalium* and *M. pneumoniae* were the most similar in terms of protein content (96.68%), while *P. aeruginosa* and *M. genitalium* were the least similar (11.34%). When paralogy was removed this was less evident. *M. genitalium* and *M. pneumoniae* shared 86.93% of their proteins, while *P. aeruginosa* and *M. genitalium* shared only 5.46%. The results were different when paralogy was removed. For example, when paralogy was included *M. pneumoniae* and *B. subtillis* were similar (71.55%), however when it was removed they were far less so (0.16%). In fact when paralogy was removed *M. pneumoniae* and *B. subtillis* were in fact quite different. See Table 7 for more information on shared proteins and interactions.

The Bacterial Meta-interactome Network study:

The goal of the Bacterial Meta-interactome Network (BMN) study is to identify and illustrate conservation between six bacterial species. For this study, the number of species was chosen for which there were 1,000 or more known proteins. Those species were: *B. subtilis*, *S. sanguinis*, *H. pylori*, *C.crescentus*, *P. aeruginosa*, and *E. coli*. A network was constructed using Cytoscape (see Figure 6). This study used the same source of data so the results were similar, however this network did not use *E. coli* as a basis for comparison so that conservation could be studied in more depth. Of the proteins in the network only 58 proteins were found to be present in all 6 species.

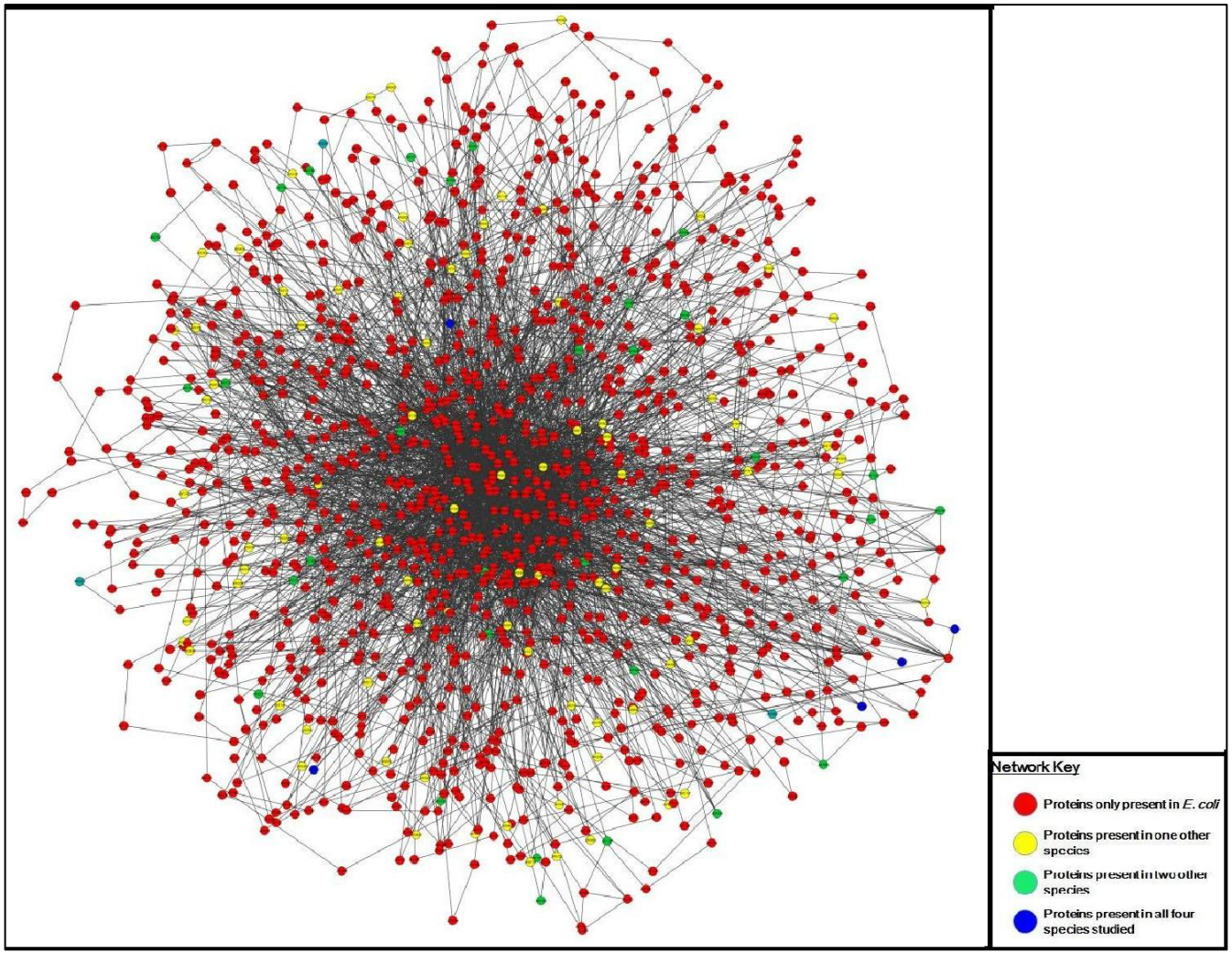


Figure 5: Eight species protein conservation network.

The PPI networks for eight species were integrated into a meta-interaction network. Each node represents a protein and each edge represents an interaction between proteins. The nodes are color-coded to represent their presence in a species or their conservation among species.The interaction network was constructed from the PPI data for eight organisms: *M. pneumoniae*, *M. genitalium*, *B. subtilis*, *S. sanguinis*, *H. pylori*, *C. crescentus*, *P. aeruginosa*, and *E. coli*. The network was constructed using Cytoscape.

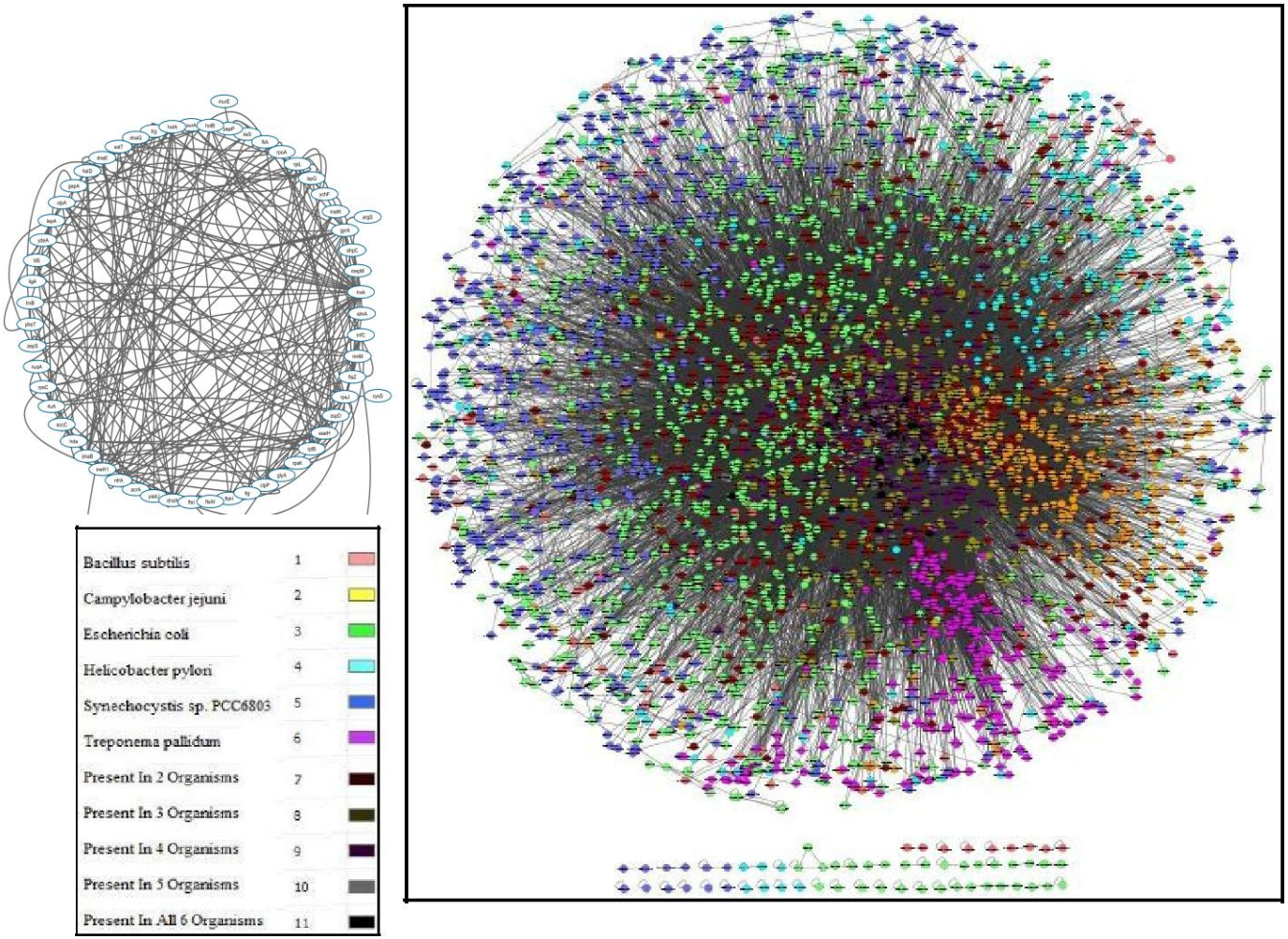


Figure 6: Meta-interactome network. This is a meta-PPI network constructed from the PPI networks of 6 species: *B. subtilis*, *S. sanguinis*, *H. pylori*, *C. crescentus*, *P. aeruginosa*, and *E. coli*. Each node represents a protein and is color-coded to indicate which organism it is in as well as the degree of conservation. From this network a sub-network of proteins conserved in all 6 species was generated.

Methods

Data Collection:

The list of proteins for *Escherichia coli* K-12 was collected from the EcoGene database <32>. The protein list contained a number of ways to identify a protein including: UniProt ID, B number, and gene name. A second list was generated containing all COGs in *Escherichia coli* and their corresponding B number from the eggNOG database using the flat files COG members and NOG members <31>. The protein list was used as a key for identifying a protein by name and the second was used as a key to map COGs to proteins.

To represent protein conservation, information from the eggNOG database was used to select species from a candidate list of bacterial model organisms. Species for which there were no COGs were removed from the list and a table comparing protein conservation among the remaining species was generated. The species chosen included: *Bacillus subtilis* (subspecies 168), *Campylobacter jejuni* (subspecies NCTC 11168), *Caulobacter crescentus* (subspecies CB15), *Helicobacter pylori* (subspecies 26695), *Mycoplasma genitalium* (subspecies G37), *Mycoplasma pneumoniae* (subspecies M129), *Streptococcus pneumoniae* (subspecies ATCC 700669), *Streptococcus sanguinis* (subspecies ATCC 49296), *Synechocystis* (PCC 6803), and finally *Treponema pallidum* (subspecies Nichols).

Each of the species chosen had an identifying number in the eggNOG database (for example *Escherichia coli*, subspecies K-12 is 511145). COGs were collected from the flat files "COG members" and "NOG members" using the species identifying numbers. Conservation was assigned to the proteins in the list from the EcoGene database using the list of COG to B-Number mappings from the eggNOG data base. A number was assigned based on how many of the 11 organisms a COG mapping to that gene was found. If a B-Number didn't map to a COG or if the COG it mapped to was not found in any of the 10 other species, a conservation value of 1 was assigned. Otherwise a number from 2-11 was assigned based on how many organisms the COG that a protein mapped to was found in. If a protein mapped to multiple COGs the highest value was assigned to that protein.

Protein essentiality was collected from the Online Gene Essentiality (OGEE) database from a flat file identifying proteins by B number, which was used to map to the protein list from the EcoGene database <8>. Protein-protein interactions were taken from Supplementary Table 5 from 2014 Rajagopala, et. al. which contained both interactions discovered from Y2H studies as

well as from a search of the literature <32>. From these interactions a Cytoscape network was built and the Cytoscape add on, Network Analyzer was used to calculate centrality values for each protein <38>. There were 27 proteins that were found in the list from 2014 Rajagopala that were not found in the list of proteins from EcoGene. The information for each of these proteins was looked up using the UniProt database and they were added to the list, along with their respective conservation, centrality, and essentiality values.

Statistical tests:

The list of proteins in *Escherichia coli* which contained the essentiality, conservation, and centrality values for each protein was used for the analysis. To avoid the complication of N/A values, the proteins that had not been tested for essentially were removed from the list used for the analysis. Each of the hypotheses was tested using correlations performed using SPSS. A Goodman and Kruskal's gamma measure was used in each case <17>. Each two of the three attributes of proteins were compared: essentiality & conservation (1), conservation & degree centrality (2), as well as essentiality & degree (3). To investigate the complication of proteins that did not map to COGs and proteins that had no known interactions a second list was prepared that had such proteins removed and the correlations were performed again. So the effect of proteins that didn't map to COGs and proteins that had no known interactions could be isolated, two more lists were made and tested. In one only the proteins that didn't map to COGs were removed and in the other only proteins that had no known interactions were removed. For each of these tests a 90% level of confidence was chosen to establish the acceptance of the hypotheses because the relationships between the values are complex.

Escherichia coli PPI network:

To visually represent the relationship between protein essentiality, conservation, and centrality a second Cytoscape network was built (See Figure 7). First the interactions were mapped to gene name using the list from EcoGene as a key. Next a list was constructed mapping conservation to gene name using B number as a key. This list was imported into Cytoscape and the essentiality values were mapped to color, red if non-essential, blue if essential, and green if there was no data. Another list was constructed mapping conservation to gene name and was mapped to node size in the network. Centrality was represented by the position of the node. The network was subjected to an edge-weighted spring-embedded layout. This layout uses an algorithm that puts nodes with more connections closer to the center <12>. Thus nodes with higher centrality are closer to the center of the network.

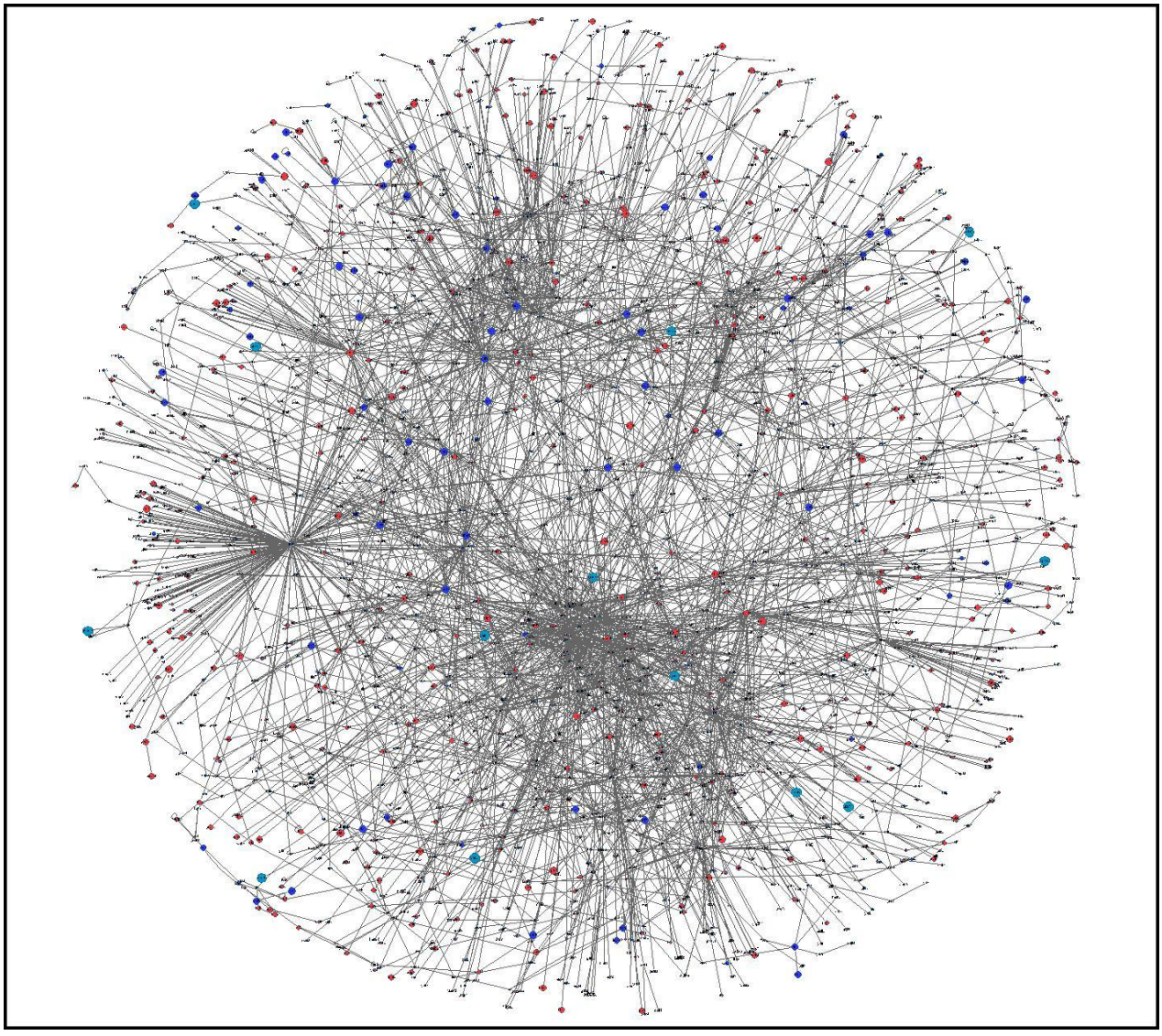


Figure 7: *Escherichia coli* network figure. This PPI network contains all of the proteins in *Escherichia coli* as well as known interactions. Essential proteins are dark blue, non-essential are red and proteins not tested for essentiality are teal. Conservation is mapped to node size and the layout is such that proteins with high centrality values congregate towards the center.

Results

Data collected:

The list of genes in *Escherichia coli*, originally from the EcoGene database but supplemented with values from the UniProt database, contained 4,529 distinct B-Numbers. There were 917 unique COGs for *Escherichia coli* from the eggNOG database. For the number of COGs in the other 10 species see table 8. Of the 4,529 proteins in the list for *Escherichia coli* 34.4% of them mapped to a COG and 30 B-Numbers mapped to multiple COGs. Essentiality data was available for 4,203 of the proteins in the list which represents 92.8% of the B-Numbers. See Table 8 for more information.

Statistical Analysis:

For the initial list of proteins for which there was essentialiy data, there was a moderate, positive relationship beween essentiality & conservation (1), a strong, positive relationsip for essentiality & degree (3), and the relationship between conservation & degree was inconclusive (2). For (1) the confidence was level 99.9% and was higher for (3). This was considerably more than the 90% chosen. However the approximate significance for (2) was 0.702. See table 9A for the values.

In order to isolate the effect of proteins that didn't map to COGs from the effect of proteins for which there was no known interactions, two more lists were made. When proteins that didn't map to COGs were removed essentiality & conservation (1) and essentiality & degree (3) showed a strong, positive relationship, while conservation & degree (2) had a weak, positive relationship. See table 9C for the values for the list where only proteins that did not map to COGs were removed. There was a moderate, positive relationship between essentiality & conservation (1) and for essentiality & degree (3). The relationship between conservation & degree (2) was not established because the significance was 0.828. See table 9D for the values for the list of proteins which had only proteins for which there were no known interactions for were removed.

Table 8: Protein content and conservation. Table 8A shows the number of COGs for each of the species under study diagonally, for example *Campylobacter jejuni* has 451 known COGs. It also compares number of COGs in common with each other species. Table 8B shows the percentange of COGs shared between any two of the eleven species.

Table 8A	*Escherichia coli*	*Bacillus subtilis*	*Caulobacter crescentus*	*Synechocystis*	*Campylobacter jejuni*	*Helicobacter pylori*	*Streptococcus pneumoniae*	*Treponema pallidum*	*Streptococcus sanguinis*	*Mycoplasma genitalium*	*Mycoplasma pneumoniae*
Escherichia coli	917	551	531	459	380	319	255	221	136	130	104
Bacillus subtilis		765	443	421	336	286	273	221	155	142	116
Caulobacter crescentus			754	416	339	286	201	189	123	115	93
Synechocystis				628	310	269	201	175	112	122	97
Campylobacter jejuni					451	326	164	178	95	109	83
Helicobacter pylori						374	150	170	77	104	83
Streptococcus pneumoniae							335	130	47	106	114
Treponema pallidum								268	66	103	80
Streptococcus sanguinis									177	48	16
Mycoplasma genitalium										149	115
Mycoplasma pneumoniae											124

Table 8B	*Escherichia coli*	*Bacillus subtilis*	*Caulobacter crescentus*	*Synechocystis*	*Campylobacter jejuni*	*Helicobacter pylori*	*Streptococcus pneumoniae*	*Treponema pallidum*	*Streptococcus sanguinis*	*Mycoplasma genitalium*	*Mycoplasma pneumoniae*
Escherichia coli		60%	58%	50%	41%	35%	28%	24%	15%	14%	11%
Bacillus subtilis	72%		58%	55%	44%	37%	36%	29%	20%	19%	15%
Caulobacter crescentus	70%	59%		55%	45%	38%	27%	25%	16%	15%	12%
Synechocystis	73%	67%	66%		49%	43%	32%	28%	18%	19%	15%
Campylobacter jejuni	84%	75%	75%	69%		72%	36%	39%	21%	24%	18%
Helicobacter pylori	85%	76%	76%	72%	87%		40%	45%	21%	28%	22%
Streptococcus pneumoniae	76%	81%	60%	60%	49%	45%		39%	14%	32%	34%
Treponema pallidum	82%	82%	71%	65%	66%	63%	49%		25%	38%	30%
Streptococcus sanguinis	77%	88%	69%	63%	54%	44%	27%	37%		27%	9%
Mycoplasma genitalium	87%	95%	77%	82%	73%	70%	71%	69%	32%		77%
Mycoplasma pneumoniae	84%	94%	75%	78%	67%	67%	92%	65%	13%	93%	

Table 9: Correlations. The correlations between essentiality & conservation (1), essentiality & degree centrality (3), as well as conservation & degree (2) were calculated using SPSS. Section 9A shows the values for the list of proteins containing all proteins from *Escherichia coli* for which essentiality had been tested. For 9B the proteins which didn't map to COGs and proteins for which there were no known interactions were removed. In 9C only proteins that didn't map to COGs were removed and in 9D only proteins with no known interactions were removed. In each Goodman and Kruskal's gamma was calculated and the value for the correlation given. The approximate significance, standard error, and relationship are shown for each.

9A

Essentiality& Conservation	Relationship	**Moderate** +
Symmetric Measures		
	Value	Asymp. Std. Error	Approx. Sig.
Gamma	.174	.048	.001

Essentiality& Degree	Relationship	**Strong** +
Symmetric Measures		
	Value	Asymp. Std. Error	Approx. Sig.
Gamma	.337	.034	.000

Conservation& Degree	Relationship	**Inconclusive**
Symmetric Measures		
	Value	Asymp. Std. Error	Approx. Sig.
Gamma	.008	.022	**.702**

9B

Essentiality& Conservation	Relationship	**Strong** +
Symmetric Measures		
	Value	Asymp. Std. Error	Approx. Sig.
Gamma	.585	.060	.000

Essentiality& Degree	Relationship	**Moderate** +
Symmetric Measures		
	Value	Asymp. Std. Error	Approx. Sig.
Gamma	.268	.069	.000

Conservation& Degree	Relationship	**Weak** +
Symmetric Measures		
	Value	Asymp. Std. Error	Approx. Sig.
Gamma	.083	.034	.015

9C

Essentiality& Conservation	Relationship	**Strong +**

Symmetric Measures

	Value	Asymp. Std. Error	Approx. Sig.
Gamma	.423	.054	.000

Essentiality& Degree	Relationship	**Strong +**

Symmetric Measures

	Value	Asymp. Std. Error	Approx. Sig.
Gamma	.330	.055	.000

Conservation& Degree	Relationship	**Weak +**

Symmetric Measures

	Value	Asymp. Std. Error	Approx. Sig.
Gamma	.071	.026	.007

9D

Essentiality& Conservation	Relationship	**Moderate +**

Symmetric Measures

	Value	Asymp. Std. Error	Approx. Sig.
Gamma	.222	.060	.001

Essentiality& Degree	Relationship	**Moderate +**

Symmetric Measures

	Value	Asymp. Std. Error	Approx. Sig.
Gamma	.285	.043	.000

Conservation& Degree	Relationship	**Inconclusive**

Symmetric Measures

	Value	Asymp. Std. Error	Approx. Sig.
Gamma	.006	.029	**.828**

Discussion & Conclusion

Conclusions:

The hypotheses are that there is a measurable, positive correlation between conservation & essentiality (1) as well as that there is a positive correlation between essentiality & degree centrality (3). However a correlation between protein conservation & degree centrality (2) was not established. This was due to the influence of proteins that did not map to COGs. When those proteins were removed a weak, positive correlation was established with significance values within the accepted range. Although well within the acceptable range, the confidence was slightly higher for the correlation between essentiality & conservation (1) when proteins that did not map to COGs were removed. An explination for this influence would be that COGs do not accurately represent protein conservation between species. When proteins for which there were no known interactions were removed, the significance values were not within the accepted range which suggests that it was not due to the influence of proteins for which there are no known interactions.

The link between essentiality and conservation (1) found here was expected because it is thought that proteins that are necessary for survival evolve slower <16>. Proteins with more connections are thought to be more likely to be essential so the correlation between protein essentiality and degree centrality (3) was also expected. Due to the relationships between (1) and (3), it was thought that there should also be a relationship between protein conservation & degree (2), however this was not found. As mentioned previously, this may be due to the methods used. It might also be due to the fact that the relationship between centrality and conservation is complex <27>. The essentiality of proteins often flips between species due to differences between the environments they inhabit and is dependent on function <36, 49>.

Reasoning for the Hypotheses:

It was originally speculated that proteins with many interactions would be more likely to be essential. Several studies have found such a relationship yet others did not and this has been investigated in more detail <19, 48, 50, 51>. Proteins with many interactions would be expected to have a larger effect on an organism's metabolism than ones with fewer interactions. This increased importance might mean that it would be under selective pressure and thus it would be more conserved. However this might instead be due to a protein's role in a complex or its pleiotropy <48, 50>. Not counting differences in environment, a protein with an essential function might be expected to be retained, thus there should also be a link between essentiality

and conservation. This relationship has been discovered between proteins in *Saccharomyces cerevisiae* and *Caenorhabditis elegans* however it should be studied across a wider range of organisms for confirmation and for better understanding of the phenomenon <23>.

Limitations:

There were aspects of the data that may have confounded the results. The most obvious is that not all of the proteins in *Escherichia coli* were tested for essentiality. A less intuitive complication is that the definition of essentiality defines a protein as either wholly essential or wholly non-essential. This is further compounded by the fact that just because a protein is non-essential does not necessarily mean that it does not perform an essential function. The functions of proteins are often redundant, being performed by other proteins in the organism. It is likely that there are functions that might be essential in an organism's environment yet non-essential in rich medium.

When determining protein-protein interactions it is possible that proteins do not exist in enough abundance to detect an interaction which could lead to genuine interactions being missed. It is also possible for two proteins that would not normally come in contact in the organism to interact leading to false positives. Not all of the proteins in *Escherichia coli* have been tested for interactions, thus it is unlikely that the known interactions for *Escherichia coli* are completely and accurately represented. When establishing protein conservation, proteins in *Escherichia coli* were compared using COGs. Only a relatively small portion of *Escherichia coli*'s proteins map to COGs. Ten organisms were chosen as a comparison, and relative to the large amount of bacterial species, this represents an incredibly small portion. As more information becomes available more organisms will be represented and those who are represented will be more accurately represented.

Bibliography

Bibliography

1. Barabasi, A., & Oltvai, Z. (2004). Network Biology: Understanding the cell's functional organization. *Nature Reviews: Genetics , 5* (2), 101-113.

2. Batagelj, V., & Mrvar, A. (1999, January 3). Pajek – Program for Large Network Analysis.

3. Biggs, N. L., Lloyd, E., & Wilson, R. (1986). *Graph Theory, 1736-1936.* New York: Clarendon Press.

4. Blattner, F. R., III, G. P., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., et al. (1997). The complete genome sequence of Escherichia coli K-12. *Science , 277* (5331), 1453-1462.

5. Bonchev, D., & Buck, G. A. (2007). From Molecular to Biological Structure and Back. *Journal of Chemical Information and Modeling , 47* (3), 909-917.

6. Breitenbach, M., Jazwinski, S. M., & Laun, P. (2012). *Aging Research in Yeast.* Springer.

7. Carbon, S., Ireland, A., Mungall, C. J., Shu, S., Marshall, B., Lewis, S., et al. (2008). AmiGO: online access to ontology and annotation data. *Bioinformatics , 25* (2), 288-289.

8. Chen, W.-H., Minguez, P., Lercher, M. J., & Bork, a. P. (2012). OGEE: an online gene essentiality database. *Nucleic Acids Research , 40* (Database Issue), D901-906.

9. Cherkasov, A., Hsing, M., Zoraghi, R., Foster, L. J., See, R. H., Stoynov, N., et al. (2011). Mapping the protein interaction network in methicillin-resistant Staphylococcus aureus. *Journal of proteome research , 10* (3), 1139-1150.

10. Cherry, J. M., Adler, C., Ball, C., Chervitz, S. A., Dwight, S. S., Hester, E. T., et al. (1998). SGD: Saccharomyces Genome Database. *Nucleic Acids Research , 26* (1), 73-79.

11. Collura, V., & Boissy, G. (2007). From protein-protein complexes to interactomics. *Sub-cellular biochemistry , 43*, 135-183.

12. Conn, P. M. (Ed.). (2011). *Methods in ENZYMOLOGY* (Vol. 498).

13. Consortium, T. U. (2008). The Universal Protein Resource. *Nucleic Acids Research , 36* (Database Issue), D190-195.

14. Diestel, R. (2010). *Graph Theory* (4th ed.). Heidelberg: Springer-Verlag.

15. Estrada, E. (2011). *The Structure of Complex Networks: Theory and Applications.* Oxford University Press.

16. Fang, G., Rocha, E., & Danchin, A. (2005). How Essential Are Nonessential Genes? *Molecular Biology and Evolution , 22* (11), 2147-2156.

17. Goodman, L. A., & Kruskal, W. H. (1954). Measures of Association for Cross Classifications. *Journal of the American Statistical Association , 49* (268), 732-764.

18. Guldener, U., Munsterkotter, M., Kastenmuller, G., Strack, N., Helden, J. v., Lemer, C., et al. (2005). CYGD: the Comprehensive Yeast Genome Database. *Nucleic Acids Research , 33* (Database Issue), D364-368.

19. Hahn, M. W., & Kern, A. D. (2004). Comparative Genomics of Centrality and Essentiality in Three Eukaryotic Protein-Interaction Networks. *Molecular Biology and Evolution , 22* (4), 803-806.

20. Häuser, R., Ceol, A., Rajagopala, S. V., Mosca, R., Siszler, G., Wermke, N., et al. (2014). A second-generation protein-protein interaction network of Helicobacter pylori. *Molecular & cellular proteomics , 13* (5), 1318-1329.

21. Hogeweg, P. (2011). The Roots of Bioinformatics in Theoretical Biology. *PloS Computational Biology , 7* (3), e1002021.

22. Jeong, H., Mason, S. P., Barabási, A.-L., & Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature , 411*, 41-42.

23. Jordan, I. K., Rogozin, I. B., Wolf, Y. I., & Koonin, E. V. (2002). Essential Genes Are More Evolutionarily Conserved Than Are Nonessential Genes in Bacteria. *Genome Research , 12*, 962-968.

24. Karp, P. D., Riley, M., Saier, M., Paulsen, I. T., Collado-Vides, J., Paley, S. M., et al. (2002). The EcoCyc Database. *Nucleic Acids Research , 30* (1), 56-58.

25. Khuri, S., & Wuchty, S. (2015). Essentiality and centrality in protein interaction. *BMC Bioinformatics , 16* (109), 1-17.

26. Managbanag, J. R., Witten, T. M., Bonchev, D., Fox, L. A., Tsuchiya, M., Kennedy, B. K., et al. (2008). Shortest-Path Network Analysis Is a Useful Approach toward Indentifying Genetic Determinants of Longevity. *PLoS One , 3* (11), e3802.

27. Mazurie, A., Bonchev, D., Schwikowski, B., & Buck, G. A. (2010). Evolution of metabolic network organization. *BMC Systems Biology , 4* (59), 1-10.

28. Nikitin, A., Egorov, S., Daraselia, N., & Mazo, I. (2003). Pathway studio--the analysis and navigation of molecular networks. *Bioinformatics , 19* (16), 2155-2157.

29. Parrish, J. R., Yu, J., Liu, G., Hines, J. A., Chan, J. E., Mangiola, B. A., et al. (2007). A proteome-wide protein interaction map for Campylobacter jejuni. *Genome Biology , 8* (7), R130.

30. Pavlopoulos, G. A., Secrier, M., Moschopoulos, C. N., Soldatos, T. G., Kossida, S., Aerts, J., et al. (2011). Using graph theory to analyze biological networks. *BioDataMining , 4*, 10.

31. Powell, S., Szklarczyk, D., Trachana, K., Roth, A., Kuhn, M., Muller, J., et al. (2012). eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Research , 40* (Database Issue), D284-289.

32. Rajagopala, S. V., Sikorski, P., Kumar, A., Mosca, R., Vlasblom, J., Arnold, R., et al. (2014). The binary protein–protein interaction landscape of Escherichia coli. *Nature biotechnology , 32* (3), 285-290.

33. Retrieved from Sageweb: http://sageweb.org/

34. Rivas, J. D., & Fontanillo, C. (2010). Protein–Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks. *PLoS Computational Biology , 6* (6), e1000807.

35. Rudd, K. E. (2000). EcoGene: a genome sequence database for Escherichia coli K-12. *Nucleic Acids Research , 28* (1), 60-64.

36. Ryan, C. J., Krogan, N. J., Cunningham, P., & Cagney, G. (2013). All or Nothing: Protein Complexes Flip Essentiality between Distantly Related Eukaryotes. *Genome Biology and Evolution , 5* (6), 1049-1059.

37. Sato, S., Shimoda, Y., Muraki, A., Kohara, M., Nakamura, Y., & Tabata, S. (2007). A large-scale protein protein interaction analysis in Synechocystis sp. PCC6803. *DNA Research , 14* (5), 207-216.

38. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research , 13* (11), 2498-2504.

39. Shimoda, Y., Shinpo, S., Kohara, M., Nakamura, Y., Tabata, S., & Sato, S. (2008). A Large Scale Analysis of Protein–Protein Interactions in the Nitrogen-fixing Bacterium Mesorhizobium loti. *DNA Research , 15* (1), 13-23.

40. Tacutu, R., Budovsky, A., & Fraifeld, V. E. (2010). The NetAge database: a compendium of networks for longevity, age-related diseases and associated processes. *Biogerontology , 11* (4), 513-522.

41. Tatusov, R. L., Koonin, E. V., & Lipman, D. J. (1997). A genomic perspective on protein families. *Science , 278* (5338), 631-637.

42. Teixeira, M. C., Monteiro, P., Jain, P., Tenreiro, S., Fernandes, A. R., Mira, N. P., et al. (2006). The YEASTRACT database: a tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae. *Nucleic Acids Research , 34* (Database Issue), D446-451.

43. Titz, B., Rajagopala, S. V., Goll, J., Häuser, R., McKevitt, M. T., Palzkill, T., et al. (2008). The binary protein interactome of Treponema pallidum--the syphilis spirochete. *PLoS One , 3* (5), e2292.

44. Wang, Y., Cui, T., Zhang, C., Yang, M., Huang, Y., Li, W., et al. (2010). Global protein-protein interaction network in the human pathogen Mycobacterium tuberculosis H37Rv. *Journal of proteome research , 9* (12), 6665-6677.

45. Whitt, W. (1983). The Queueing Network Analyzer. *The Bell System Technical Journal , 62* (9), 2779-2815.

46. Witten, T. M. (1984). A return to time, cells, systems and aging: II. Relational and reliability theoretic approaches to the study of senescence in living systems. *Mechanisms of ageing and development , 27* (3), 323-340.

47. Witten, T. M. (1985). A return to time, cells, systems and aging: III. Critical elements, hierarchies, and Gompertzian dynamics. *Mechanisms of ageing and development , 32* (2-3), 141-177.

48. Wuchty, S., & Uetz, P. (2014). Protein-protein Interaction Networks of E.coli and S. cerevisiae are similar. *Scientific Reports , 4* (7187), 1-7.

49. Xu, P., Ge, X., Chen, L., Wang, X., Dou, Y., Xu, J. Z., et al. (2011). Genome-wide essential gene identification in Streptococcus sanguinis. *Scientific Reports , 1* (125), 1-9.

50. Yu, H., Braun, P., Yıldırım, M. A., Lemmens, I., Venkatesan, K., Sahalie, J., et al. (2008). High-Quality Binary Protein Interaction Map of the Yeast Interactome Network. *Science , 322*, 104-109.

51. Zotenko, E., Mestre, J., O'Leary, D. P., & Przytycka, T. M. (2008). Why Do Hubs in the Yeast Protein Interaction Network Tend To Be Essential: Reexamining the Connection between the Network Topology and Essentiality. *PLoS Computational Biology , 4* (8), e1000140.

Vita

Christopher Frederick Wimble was born December 13th, 1976 in Henrico County, Virginia and is a citizen of the United States of America. He graduated from Mills Godwin High School, Henrico, Virginia in 1995. He received his Associates of the Sciences from John Tyler Community College in 2009, and then Bachelor of Bioinformatics with a minor in chemistry from Virginia Commonwealth University in 2013. He has published several works including "Modeling Aging Networks: A Systems Biology Approach – Applications", chapter 2 of the textbook Aging and Health – A Systems Biology Perspective which he was first author, "Protein Complexes in Bacteria", a paper published in PLoS Computational Biology, and is currently working on another paper, "Comparision of Protein Interaction Networks", which he is first author. In addition he is working as a graduate assistant for Margaret Henderson, the director of research data management at MCV's Tompkis-McCaw library for health sciences, and works in the lab of Dr. Peter Uetz, the founder of the reptile database and associate professor at VCU's center for biological complexity.