

©Copyright 2010  
Dhileepan Sivam



**A Rule-Based Strategy for Accurately Describing Gene  
Content Similarities and Differences Across Multiple Genomes**

**Dhileepan Sivam**

**A dissertation submitted in partial fulfillment of the  
requirements for the degree of**

**Doctor of Philosophy**

**University of Washington**

**2010**

**Program Authorized to Offer Degree:  
Department of Medical Education and Biomedical  
Informatics**

UMI Number: 3422014

All rights reserved

**INFORMATION TO ALL USERS**

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI 3422014

Copyright 2010 by ProQuest LLC.

All rights reserved. This edition of the work is protected against unauthorized copying under Title 17, United States Code.



ProQuest LLC  
789 East Eisenhower Parkway  
P.O. Box 1346  
Ann Arbor, MI 48106-1346

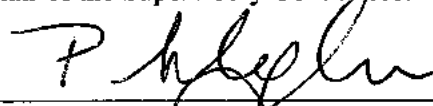
University of Washington  
Graduate School

This is to certify that I have examined this copy of a doctoral dissertation by

Dhileepan Sivam

and have found that it is complete and satisfactory in all respects,  
and that any and all revisions required by the final  
examining committee have been made.


Chair of the Supervisory Committee:

  
\_\_\_\_\_  
Peter J. Myler

Reading Committee:

  
\_\_\_\_\_  
Peter J. Myler

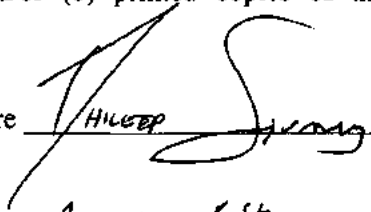
  
\_\_\_\_\_  
John H Gennari

  
\_\_\_\_\_  
Roger E. Bumgarner

Date: 7/1/10

In presenting this dissertation in partial fulfillment of the requirements for the doctoral degree at the University of Washington, I agree that the Library shall make its copies freely available for inspection. I further agree that extensive copying of the dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for copying or reproduction of this dissertation may be referred to ProQuest Information and Learning, 300 North Zeeb Road, Ann Arbor, MI 48106-1346, 1-800-521-0600, to whom the author has granted "the right to reproduce and sell (a) copies of the manuscript in microform and/or (b) printed copies of the manuscript made from microform."

Signature

A handwritten signature in black ink, appearing to read "Hilbert Spring", written over a horizontal line.

Date

April 1<sup>st</sup> 2010

University of Washington

**Abstract**

**A Rule-Based Strategy for Accurately Describing Gene  
Content Similarities and Differences Across Multiple Genomes**

Dhileepan Sivam

Chair of the Supervisory Committee:

Research Professor Peter J. Myler

Department of Medical Education & Biomedical Informatics,

and Department of Global Health

A fundamental task in genome research is comparing gene content between multiple genomes. In infectious disease research such comparisons are critical for determining the parasite genetic factors that are responsible for disease transmission, pathogenicity and clinical outcome. Though numerous technologies exist for comparing gene sequences and clustering similar genes, the genomics field lacks structured methods for describing the complicated evolutionary dynamics that caused the differences between compared species. This dissertation puts forth novel technologies for accurately and precisely describing differences in gene content across multiple genomes. A novel knowledge representation specification aggregates gene annotation and sequence comparison results from heterogeneous data sources. A newly developed ontology describes pairwise homology relationships between genes and a rule-based system applies those terms to sequence comparison results. Those ontologically annotated sequence comparison results serve as inputs to a novel method for grouping genes based on their homology relationships. Finally, this dissertation presents techniques for querying the gene groups to uncover interesting evolutionary trends across the compared genomes. These methods represent a significant advance in the clarity and detail with which large-scale comparative genomics can be described; furthermore, the novel techniques presented herein are amenable to integration with existing sequence comparison and clustering technologies.

# TABLE OF CONTENTS

<b>List of Figures .....</b>	<b>iv</b>
<b>List of Tables .....</b>	<b>v</b>
<b>Acknowledgements .....</b>	<b>vi</b>
<b>Chapter 1: Introduction.....</b>	<b>1</b>
1.1: Comparative Genomics	2
1.2: Statement of Problem	4
1.3: Proposed Solutions	8
1.4: Dissertation Outline	10
1.5: Significance And Impact	12
<b>Chapter 2: Background and Context .....</b>	<b>15</b>
2.1: Overview	16
2.2: Mechanisms of Evolution: Gene Duplication, Acquiring New Genes And Gene Loss	16
2.3: Relationships – The Many Types of Homology	18
2.4: Genome Sequencing and Annotation	20
2.5: Inferring Function Through Sequence Similarity	23
2.6 Sequence Comparison	24
2.9: Ontologies	24
2.10: Chado	27
2.11: Leishmania	28
2.12: Comparative Genomics For Drug Discovery	29
<b>Chapter 3: Ontology Creation.....</b>	<b>31</b>
3.1: Overview	32
3.2: Pairwise Genome Comparison Ontology (PGCO)	34



3.3: Gene Homology Ontology (GHO)	40
3.4: Using the Ontologies	47
<b>Chapter 4: Rule-Based Classification Of Homology Relationships .....</b>	<b>49</b>
4.1: Rule-Based Classification Overview	50
4.2: Knowledge Representation Schema	51
4.3: Rule-Based Classification Strategy	57
4.4: Rule-Based Application of PGCO Terms	61
4.5: Rule-Based Semantic Homology Annotation	63
4.6: Rule Base Classification Conclusion	72
<b>Chapter 5: Semantic Gene Grouping .....</b>	<b>75</b>
5.1: Overview	76
5.2: Moving From Pairwise Analysis To Multi-Genome Analysis	77
5.2: Creating Ortholog Groups	78
5.3: Adding Inparalogs To Ortholog Groups	81
5.4: Joining Outparalogs	84
5.5: Semantic Gene Grouping Summary	86
<b>Chapter 6: Logical Gene Group Querying .....</b>	<b>87</b>
6.1: Introduction To Logical Gene Group Querying	88
6.2: An Example of A Logical Gene Group Query	90
6.3: Posing a Logical Gene Group Query	92
6.4: Gene Presence Or Absence Rules	94
6.5: Gene Expansion Rules	98
6.6: Logical Gene-Group Querying Conclusions	103
<b>Chapter 7: Leishmania Comparative Genomics.....</b>	<b>105</b>
7.1: Overview of the Leishmania Comparative Genomics Project	106

7.2: The Four Leishmania Species	108
7.3: Methods Used In The Leishmania Comparisons	110
7.4: Advantages Of Employing Rule-Based Homology Annotation	111
7.5: Evolution of New Genes In Leishmania	116
7.6: Gene Loss in The Leishmanias	121
7.7: Inparalogous Expansion In Leishmania	126
7.8: Summary of the Leishmania comparative genomics study	135
<b>Chapter 8: Cross-Phyla Comparative Genomics .....</b>	<b>139</b>
8.1: Introduction	140
8.2: Pathogens Used in This Comparison	141
8.3: Genes Present In All Three Species	142
8.4: Genes Groups That Are Present In Two Species	146
8.5: Species-Specific Gene Groups	148
8.6: Unclassifiable Genes	148
8.7: Cross-Phyla Comparison Conclusions	149
<b>Chapter 9: Conclusions And Future Work.....</b>	<b>153</b>
9.1: Project Goals	154
9.2: Contributions Of This Work	156
9.3: Future Directions	161
9.5: Concluding Remarks	164
<b>Appendix A: Pairwise Genome Comparison Ontology Terms .....</b>	<b>165</b>
<b>Appendix B: Gene Homology Ontology Terms .....</b>	<b>169</b>
<b>Appendix C: Results of Pairwise Leishmania Comparisons .....</b>	<b>173</b>
<b>Bibliography .....</b>	<b>177</b>
<b>Vita.....</b>	<b>184</b>

## LIST OF FIGURES

<b>Figure 3-1: The Pairwise Genome Comparison Ontology .....</b>	<b>38</b>
<b>Figure 3-2: The Gene Homology Ontology .....</b>	<b>41</b>
<b>Figure 3-3: Outparalogs.....</b>	<b>44</b>
<b>Figure 3-4: Inparalogs .....</b>	<b>46</b>
<b>Figure 3-5: Pseudo-inparalog .....</b>	<b>47</b>
<b>Figure 4-1: Calculating the Positional Conservation Ratio .....</b>	<b>65</b>
<b>Figure 4-2: Fusion/Splice Genes .....</b>	<b>67</b>
<b>Figure 5-1: Ortholog Groups.....</b>	<b>79</b>
<b>Figure 5-2: Determining Uncertain Orthology Group Membership .....</b>	<b>83</b>
<b>Figure 5-3: Joining Outparalogs .....</b>	<b>85</b>
<b>Figure 6-1: Four Species Genome Comparison Cladogram .....</b>	<b>89</b>
<b>Figure 6-2: Binary Gene Absence and Presence .....</b>	<b>91</b>
<b>Figure 6-3: Gene Expansions .....</b>	<b>92</b>
<b>Figure 6-4: Gene Gain and Loss Profiles.....</b>	<b>93</b>
<b>Figure 6-5: Gene Expansion Profiles.....</b>	<b>101</b>
<b>Figure 7-1: <i>Leishmania</i> Cladogram.....</b>	<b>107</b>
<b>Figure 7-2: Gene loss and gain in <i>Leishmania</i> .....</b>	<b>118</b>
<b>Figure 7-3: Gene Expansion in <i>Leishmania</i> .....</b>	<b>129</b>
<b>Figure 7-4: Ancestral Duplication Results in Multiple Gene Groups.....</b>	<b>133</b>
<b>Figure 7-5: Ancestral Duplication Followed by Gene Loss .....</b>	<b>134</b>
<b>Figure 8-1: Gene Group Presence for the <i>Bps/Rpr/Mtb</i> Comparison .....</b>	<b>144</b>
<b>Figure 8-2: Splice genes in <i>Bps/Rpr/Mtb</i> .....</b>	<b>145</b>

## LIST OF TABLES

<b>Table 6-1:</b> GHO Assignments for the Patterns of Expansion Shown in Figure 6-3.....	<b>102</b>
<b>Table 7-1:</b> Gene Count and Gene Group Information by <i>Leishmania</i> Species.....	<b>111</b>
<b>Table A-1:</b> Pairwise Genome Comparison Ontology (PGCO) Terms.....	<b>166</b>
<b>Table B-1:</b> Gene Homology Ontology (GHO) Terms.....	<b>170</b>
<b>Table C-1:</b> Results of the <i>LmjF</i> / <i>LmxM</i> Comparison .....	<b>174</b>
<b>Table C-2:</b> Results of the <i>LinJ</i> / <i>LbrM</i> Comparison .....	<b>174</b>
<b>Table C-3:</b> Results of the <i>LbrM</i> / <i>LmxM</i> Comparison .....	<b>175</b>
<b>Table C-4:</b> Results of the <i>LmjF</i> / <i>LinJ</i> Comparison.....	<b>175</b>
<b>Table C-5:</b> Results of the <i>LinJ</i> / <i>LmxM</i> Comparison .....	<b>175</b>
<b>Table C-6:</b> Results of the <i>LmjF</i> / <i>LbrM</i> Comparison.....	<b>176</b>

## ACKNOWLEDGEMENTS

Foremost, I would like to sincerely thank the friends and family that have been so very supportive of me over my graduate years. They been there for me every time I needed them, and for that I cannot thank them enough.

I have been truly fortunate to spend the bulk of my graduate school career at the Seattle Biomedical Research Institute. Seattle BioMed, as it is now known, has furnished me with a fantastic work environment, superb coworkers and has facilitated my work in countless ways. I can also say that my years at Seattle BioMed have provided me with many friends who will continue to play a role in my life and career many years into the future.

I'd like to thank my advisor Peter Myler for his years of support and scientific guidance. Peter has a keen eye for finding practical applications for informatics research, and I've learned greatly the many projects that he and his lab have been involved with over the years. I've always had wonderful colleagues in the Myler Lab and the Bioinformatics Core; there have been more of them than I can possibly thank individually, and I value my years working with them.

My thesis committee have been both knowledgeable and supportive and have taught me how to refine an idea into a scientifically meaningful project; many thanks to John Gennari, Roger Bumgarner and Larry Ruzzo.

I attended my first class at the University of Washington campus many years ago as an 18 year old freshman; after some years away I returned for graduate school. It would appear that my academic career at the UW is nearing conclusion. GO HUSKIES, and thanks for the memories!

## **CHAPTER 1: INTRODUCTION**

### **1.1: COMPARATIVE GENOMICS**

The pace of genome sequencing is rapidly increasing; determining similarities and differences between these genomes is a fundamental first step in pinpointing answers to critical biological and medical questions. The process of elucidating these similarities and differences is known as comparative genomics. Although the term *comparative genomics* can encompass any number of analytic techniques involving genomic sequence, perhaps the most common and fundamental task in comparative genomics is assessing gene content across multiple genomes.

On the most basic level, comparison of gene content entails determining which genes are present or absent in one organism as compared to another. A refinement of such an analysis is to determine the level of similarity between those genes. A further refinement is the task of determining evolutionary forces that might have caused the differences in gene content. Finally, we can also move beyond comparing pairs of genomes to comparing genomes from multiple organisms. These analyses all fundamentally serve to provide insights into how and why a particular organism functions the way that it does.

Comparative genomics can serve as a particularly important tool in the fight against global infectious disease. According to the World Health Organization there are approximately 250 million cases of malaria per year, resulting in 1 million deaths; Tuberculosis is even more deadly, killing 1.6 million people per year. Even in the United States 36,000 people a year die of influenza related complications alone.

The advancing technologies for DNA sequencing have provided genome sequence for a wide spectrum of disease causing bacteria, viruses, fungi, protozoa and multi-cellular parasites. This wealth of knowledge runs not only wide, but deep; in many cases genomes have been

sequenced for numerous species or strains of closely related parasites. For example PlasmoDB <sup>1</sup>, the online resource for the malaria causing organism *Plasmodium falciparum*, lists DNA sequence data for ten different species.

The depth of DNA sequence data for similar organisms furnishes the opportunity to better explore species-specific adaptations that allow a certain pathogen to behave in a particular manner. For instance the protozoan pathogen *Leishmania infantum* causes the disease visceral leishmaniasis (also known as kala azar), which can carry a nearly 100% fatality rate within two years if left untreated; by contrast the closely related parasite *Leishmania major* causes the still serious, but considerably less fatal, disease cutaneous leishmaniasis.

*L. major* and *L. infantum* have 99% similar gene contents, indicating that the differences in clinical manifestation are likely due to subtle genetic differences. Furthermore, the research community has DNA sequence for several more closely related *Leishmania* species, with several more sequencing projects in the planning or early stages. Characteristics such as clinical manifestation, disease vector, disease host and geographic distribution vary amongst these species. Comparative studies between these highly similar species can provide valuable insights into the genomic causes for these differences.

At a broader level, comparison between more distantly related organisms can provide knowledge about more fundamental and generally applicable matters. Comparing large groups of pathogens can, for instance, can provide information on which genes are associated with general mechanisms of virulence; this provides lab scientists researching mechanisms of infection with insights as to which genes may be most critical for study <sup>2</sup>. Furthermore such broad comparisons between various types of pathogens supply insights into which genes are generally necessary to support the basic cellular functions of a parasite; such information can guide development of broad-spectrum drugs for the treatment of infectious disease <sup>3</sup>.



While the above examples are compelling rationales for comparative studies on infectious organisms, they represent only a few of the vast array of comparative studies that have been performed or will be performed in the future. The increasing pace of genome sequencing and advances in technologies for assessing gene expression ensure that comparative genomics will become an increasingly vital tool in the scientific community's efforts towards promoting health and well-being in humans, animals and agriculture.

Though comparative genomics holds much promise, it also generates many challenges. Of particular importance is the need to develop robust, but flexible, strategies for dealing with the vast amounts of data generated by these studies. The challenge of dealing with genomic data is many-fold: advances in computer hardware and software are necessary to accommodate ever more complicated analytic techniques; biological scientists increasingly need an understanding of mathematics and statistics in order to fully comprehend the results of gene expression studies; new reporting and visualization strategies are necessary to deal with data that potentially spans thousands of genes across numerous genomes.

This work addresses a very specific and fundamental problem in comparative genomics data analysis: accounting in an accurate and descriptive manner for the similarities and differences in gene content across genomes.

## **1.2: STATEMENT OF PROBLEM**

The field of comparative genomics would greatly benefit from a clear, generalized and systematic methodology for representing the results of genome comparisons. Genomic researchers typically possess a sound understanding of how best to approach most comparative genomic questions; however the increase in sequenced genomes – and the consequent growth

in the number of relevant cross-species, cross-strain and cross-version comparisons exacerbates the need for structured tools for improved comparative genomics.

The lack of structure in comparative genomics can be described in three succinct points:

1. There is no standardized way to describe relationships between pairs of genes
2. Most technologies that group related genes do not describe the relationships between genes in the group
3. Despite a lack of standards, comparative genomics studies are relatively easy to interpret at the gene-to-gene level, but more difficult to interpret at the genome scale.

As evidenced by the Gene Ontology, Sequence Ontology <sup>4-6</sup>, Chado <sup>7</sup>, as well as numerous other projects <sup>8-10</sup>, the genome community is progressing towards software that adheres to formal data representations. Such ontologies provide standardized vocabularies, clear semantics, the ability to query at varying degrees of similarity, and provide a common standard for the integration of disparate data sources. Although the inference of gene relationships through sequence comparison is perhaps the most fundamental of bioinformatics tasks, no such structured vocabularies have been created for describing the results of these analyses.

Gene clustering is a clear example of how the lack of semantics in comparative genomics leads to difficult-to-interpret results. Most clustering methodologies collect genes with some defined degree of sequence similarity into groups of genes; however, most clustering technologies output groups of genes without necessarily describing how those genes are related. Given the scale of data in multi-species comparisons of gene content, this represents a significant challenge to researchers attempting to understand clustering outputs.

A clear description of relationships between genes in a cluster would supply a researcher with a better understanding of how a group of genes evolved from a single ancestral gene. Such an understanding provides greater insight toward the functional similarities between the genes; furthermore, explicit relationship descriptions at the gene-level would allow researchers to pose structured queries that could result in a clearer view of genome-wide differences across the compared species.

Determining homologous relationships between genomes requires detailed investigation to fully catalog differences in gene content, composition, synteny, and copy number. An exhaustive comparative genome analysis entails much more than simple sequence comparison - exploration of paralogous groups within genomes and orthologous groups across genomes is often complicated by many-to-many relationships, varying degrees of similarity, syntenic breaks, as well as false-positive and false-negative gene predictions. This unfailingly requires considerable human curation. Although the automation of such comparisons can significantly lessen the tedious bookkeeping efforts often involved in such analyses, the above-mentioned obstacles present challenges in developing software capable of addressing those complexities and presenting the results in a comprehensible manner.

### ***Prior Attempts at Solutions***

This above issue has been addressed by prior works, such as the COG/KOGs <sup>11,12</sup> project and the INPARANOID <sup>13,14</sup> tool (both described below); however, these technologies rely on a strict set of rules for defining how clusters are formed. These rules do provide a certain *de-facto* description of pairwise relationships between genes in each cluster. While such a strategy is certainly useful, these tools lack the flexibility to truly accommodate the breadth of comparative genomics research questions.

The COG project has clustered orthologous genes from 66 prokaryotes – similarly the Eukaryotic Clusters of Orthologous Groups (KOG) project has clustered orthologs from seven eukaryotes. Both projects use mutual best hits as criteria for classifying a pair for genes as orthologous. Both projects also employ further refinement to ensure that common problems such as multi-domain proteins are appropriately resolved. The COG and KOG databases provide a framework for functional annotation by grouping together genes that likely have similar structure and function.

A particular weakness in the COG project, and indeed a weakness in most comparative genomics studies, is the lack of semantic annotation of results. The current COG database contains 4873 clusters, which consist of 128,458 genes. Given the sheer number of relationships contained within the clusters, manual assessment of the actual cluster content and full understanding of the relationships contained therein is extremely difficult. An ontology by which homologous genes could be categorized would allow for far richer and more understandable summaries of gene clusters from COG, KOG, and other projects.

The fundamental goal of INPARANOID is to cluster orthologs across genomes (in a conceptual manner similar to COG) and add to the clusters any genes that have duplicated from those orthologs. A strength of INPARANOID is that the addition of duplicated genes to the clusters provides distinction between one-to-one and one-to-many orthology scenarios, which is a functionally important difference.

While INPARANOID does build upon the relatively simple orthology assignments conducted by COGs, the two projects share many of the same weaknesses. They do not provide means for accurately describing a number of functionally important homology relationships and they do not employ any sort of defined semantics to concisely describe the relationships between genes in a cluster.

The sheer scale and complexity of homologous relationships in multiple genome comparisons exceeds the descriptive capabilities of existing comparative genomics methodologies. This issue is extensively explored later in this dissertation in the context of a four-way *Leishmania* comparative genomics study (Chapter 7) and a three-way bacterial pathogen comparative study (Chapter 8).

### **1.3: PROPOSED SOLUTIONS**

#### ***Rule-Based Comparative Genomics Pipeline***

This work presents a rule-based method for describing homologous relationships; this novel method is referred to as *homology annotation*. This work also presents a rule-based system, known as *semantic grouping*, for grouping genes based on their homology relationships. Finally, this work introduces a rule-based system, which we call *logical cluster querying*, for interrogating the content of semantic groups. The above rule-based steps collectively form a pipeline for concisely describing relationships between genes, grouping those genes and then assessing meaningful similarities and differences at the genome scale.

An important component of the rule-based system is the use of a novel knowledge representation schema that affords us a lightweight data integration platform for collecting information in a flexible manner from any number of sources. This schema allows users to input results from other analytic tools into the above pipeline and serves as the output format from all the individual steps in the pipeline.

#### ***Homology Annotation***

The solutions presented in this dissertation rely on a newly developed ontology for unambiguously describing homology relationships between pairs of genes; this ontology will serve as the foundation upon which the entire *rule-based comparative genomics pipeline* will

be built. The benefits of employing ontologies as a basis for software development are many-fold<sup>5</sup>. First, an-ontology serves as a standardized vocabulary of terms, thereby facilitating data sharing and unambiguous discussion of terms. Second, ontologies describe the relationships between vocabulary terms, hence allowing for intelligent queries and automated inference. Third, ontologies separate the semantic layer of software applications from the reasoning and control layers, which allows decentralized software systems to leverage common terminologies. Finally, semantic annotation of text and data provides a starting point for natural language processing of scientific literature.

The *homology annotation* strategy entails classifying homologous relationships (as elucidated by sequence comparison results) according to the newly developed ontology using a set of Prolog rules. Along with concisely describing sequence comparison results, the formal ontology of comparative genomic terms serves as a standardized, extensible template for developing comparative genomics software. This strategy has been used effectively in the genomics community, most notably for functional annotation by Gene Ontology terms<sup>15</sup>.

### ***Semantic Gene Grouping***

The *semantic gene grouping* methodology groups genes based on the *homology annotation* relationships. This strategy provides biologists with collections of related genes that have been assembled according to a set of readily understandable logical steps. This strategy is advantageous in that it is flexible (the rules can be changed), human comprehensible (the grouping relies on logic as opposed to complex mathematics), and is agnostic to which sequence comparison methodology was used to elucidate the *homology annotation* relationships.

### ***Logical Cluster Querying***

*Logical cluster querying* provides a means for posing directed, rule-based queries to the *semantic groups* in order to answer biologically meaningful questions, such as when in a lineage a gene evolved, or which genes are expanding in a particular genome; this functionality is possible because the groups are not represented simply as a list of genes, instead the evolutionary relationships between the group members are explicitly described.

#### **1.4: DISSERTATION OUTLINE**

##### ***Pairwise Genome Comparison Ontology (PGCO) & Gene Homology Ontology (GHO)***

The GHO provides a structured vocabulary for describing homologous relationships between pairs of genes. This ontology describes relationships that are computable by sequence comparison, but is also structured such that new types of relationships can be easily added. The GHO is discussed in depth in **Chapter 3**.

The process of assigning GHO terms to a pairwise relationship involves parsing high-throughput sequence comparison results. We have created an intermediate ontology, the PGCO, to assist in efficiently describing these results. Though the PGCO describes sequence comparison results independently of the GHO, in this work it serves as a bridge between raw sequence comparison results and the GHO assignments.

The PGCO has been created as a means of describing important characteristics of comparisons of gene content across two genomes; these descriptions serve as the underpinnings for the work described below, as well as potentially for any analysis that involves a similar comparison of one group of genes to another. The PGCO is described in **Chapter 3**.

##### ***Rule-Based Homology Classification System***

This work has resulted in an extensive set of rules, implemented using the Prolog<sup>16</sup> language, which classify sequence comparison results according to the PGCO and subsequently describe homologous relationships according to the GHO. The rule base is written such that it is readily extensible and amenable to changes in either or both ontologies. The homology classification system is discussed in **Chapter 4**.

#### ***Rule-Based Semantic Grouping And Querying System***

This work has created a methodology for grouping genes together according to their homologous relationships as described by the GHO. This *grouping* technology is similar in purpose to most clustering technologies, however it employs semantics rather than statistics or graph theory to create groups of evolutionarily related genes. Furthermore, this work has created a series of rule-based queries that can interrogate the semantically linked groups of genes. The grouping technology is described in **Chapter 5** and the querying technology is described in **Chapter 6**.

#### ***Leishmania comparative genomics analysis***

The above technologies have been employed to perform a comprehensive analysis of four species from the genus *Leishmania* (*L. major*, *L. infantum*, *L. braziliensis* and *L. mexicana*)<sup>17</sup>. The purpose of this study is to highlight the differences between four very closely related human parasites. The results of this comparison are presented in **Chapter 7**.

#### ***Cross-phyla bacterial pathogen comparative genomics analysis***

Our newly developed technologies have also been employed to perform a cross-phyla comparative analysis on three infectious organisms: *Burkholderia pseudomallei*, *Rickettsia prowazekii* and *Mycobacterium tuberculosis*. In contrast to the *Leishmania* comparative genomics study, this work aims to elucidate similarities and differences across a range of disease causing organisms. These results are presented in **Chapter 8**.



### **1.5: SIGNIFICANCE AND IMPACT**

The primary purpose of this work is that of creating a foundation which researchers can leverage to better represent sequence comparison results and group genes in semantically meaningful ways. This work will provide the type of structure and conciseness to sequence comparisons that the Gene Ontology project has provided to functional annotation and that the Sequence Ontology has provided to the representation of sequence data.

Bioinformatics researchers are increasingly leveraging a “pipelining” approach<sup>18</sup> whereby they create a workflow that involves numerous software tools and data repositories. A key problem in such a strategy is that of efficiently translating output from one tool into an appropriate input format for the next tool. The endeavor of translating between formats is greatly aided by standard ontologies that allow for communication and interoperability between the wide array of software tools developed by the genomics community.

Current sequence comparison tools such as BLAST and `cross_match` output results that are relatively clear at the pairwise sequence-level, but have no larger context for explaining the specific relationship between genes. For instance, orthology, inparalogous expansion and outparalogy all give similar BLAST results at a pairwise comparison level. In practice, most researchers develop *ad hoc* methodologies for viewing sequence comparison results in the larger genomic context, but a discipline-wide ontology would allow for the development of more meaningful standards-based sequence comparison tools.

Functional genomics studies, such as whole-genome expression analysis, are becoming less expensive, and hence the availability of data for related species is becoming increasingly common. This expansion provides many opportunities for comparative expression studies.

Unambiguously representing the relationships between genomes will allow researchers to construct complex comparative functional genomics queries in a simple manner.

Releases of new, presumably more complete and accurate, versions of a genome assembly pose challenges for researchers who have performed analysis on an earlier version of the genome. The current approach for solving this problem often involves individual researchers constructing their own mappings between versions. This creates problematic discrepancies between research groups studying the same organism. A discipline wide means for representing homology relationships would allow sequencing centers to better annotate the differences between draft versions of a genome, thereby allowing the scientific community to more easily transfer results from earlier versions of genomes to more current versions.

Project such as COG and KOG serve as central repositories for comparative genomic data that are widely leveraged by genome researchers. Although these projects are of tremendous value, interpreting the clusters requires a significant amount of parsing by the end-user. For instance, a user cannot simply pose questions such as “find a gene that is widely present in genus X, but absent in species Y” – instead a user has to develop *ad hoc* methods for parsing and representing the clusters before posing such questions. Semantic homology annotation of resources such as COG would allow for queries that are more meaningful and reduce the need for post-processing.



## **CHAPTER 2: BACKGROUND AND CONTEXT**

## **2.1: OVERVIEW**

This dissertation explores methods for elucidating and describing differences and similarities between groups of genomes using sequence comparison tools. The following chapter provides background information on a set of topics necessary relevant to the understanding of the dissertation's aims.

This chapter begins by discussing the evolutionary forces by which new genes are created by duplication of existing genes. Those duplication events result in new genes with shared ancestry; in this chapter we discuss patterns of duplications, how those patterns result in different types of relationships between duplicated genes, and the functional implications of those differences. Next, we discuss the process of sequencing genomes, predicting genes within those sequenced genomes, and the use of gene-clustering technologies to assign putative functional annotations to predicted genes. Afterward, we discuss some data-management and programming strategies that aid in the process of genomic research. Finally, we provide background for two comparative genomics experiments for which we employ the newly developed technologies developed as a result of this thesis work.

## **2.2: MECHANISMS OF EVOLUTION: GENE DUPLICATION, ACQUIRING NEW GENES AND GENE LOSS**

### ***Gene Duplication***<sup>19</sup>

Gene duplication events usually occur during DNA replication through a number of different mechanisms. Gene duplication by unequal cross-over results in what are known as tandemly repeated genes; in other words the "new" gene will be immediately adjacent on the chromosome to the gene from which it originated. Alternatively, gene duplication as a result

of some retrotransposon event will result in a “new” gene placed at some arbitrary chromosomal location in the genome. Finally, duplication of entire chromosomes, or large sections of a chromosome, results in large blocks of duplicated genes.

### ***Subfunctionalization and Neofunctionalization*** <sup>20</sup>

A newly duplicated gene can have several different fates subsequent to the duplication event. It is possible that maintaining multiple copies of a particular gene in a relatively static form could convey some advantage to the organism, in this case the newly duplicated gene is unlikely to evolve at a significant rate. Alternatively, one copy of the gene could be free to evolve a new function given that the other copy of the gene is ensuring that its original functional role is satisfied. This process of evolution of a new function is known as *neofunctionalization* and is thought to be the primary mechanism by which new genes evolve.

Certain genes may serve multiple different biological roles in many cellular processes. When such genes duplicate one or more times, the duplicate genes may specialize to more efficiently perform some subset of the roles that were once performed entirely by one gene. This process is known as *subfunctionalization*.

### ***Horizontal Gene Transfer***

The above definitions have all described evolutionary dynamics resulting from vertical gene transfer – the transfer of a gene from a parent to an offspring. Another evolutionary force, known as horizontal gene transfer (HGT), alternatively known as lateral gene transfer (LGT), involves the transfer of a genetic material from an organism to another organism in a non-parent-offspring manner. In prokaryotes, HGT can occur because of the uptake and expression of genetic material of one individual by another, or by a process known as transduction <sup>21</sup> whereby a bacterial phage transfers genetic material from one organism to

another. HGT can also occur in eukaryotes by the uptake of genetic material from prokaryotes in a process known as endosymbiosis <sup>22</sup>.

### ***Gene Loss by Pseudogenization*** <sup>23</sup>

Pseudogenes are genes that have lost their ability to encode for a functional protein as a result of some sort of mutation to the once-coding gene. Pseudogenes can occur as a result of some sort of change in selective pressure that renders a gene unnecessary. Pseudogenization is thought to be the primary mechanism by which a lineage loses a gene, however excision of sections of a chromosome during DNA replication is an alternative means by which genes are lost.

## **2.3: RELATIONSHIPS – THE MANY TYPES OF HOMOLOGY**

The term homolog carries numerous connotations; here we refer the homologs as two or more genes that show evidence of shared ancestry. The evolutionary relationships between related genes are a complex interplay between the forces of speciation, gene duplication, and gene loss, and horizontal gene transfer <sup>24</sup>. Understanding these relationships is more than a mere semantic exercise – each relationship carries with it a particular connotation in terms of evolutionary relatedness and functional similarity. The sum of relationships seen across two genomes provides clues as to key similarities and differences between the species.

Several types of homology have been identified and terms such as ‘ortholog’ and ‘paralog’ are commonly used in genomics literature to describe gene content patterns across and within genomes. However, the current state of homology description is notably lacking in two areas: first, there have been no formal semantic specifications (*e.g.* ontologies, controlled vocabularies) specified for the various types of homologous relationship. Second, since there has been no formal semantic specification, there are no established rules for actually

representing and summarizing the homologous relationships identified by sequence comparison experiments.

A 2005 review article <sup>24</sup> summarizes several types of homologous relationships: orthology, paralogy, inparalogy, outparalogy, pseudoorthology, pseudoparalogy and xenology. This summary will serve as a starting point for the ontology development phase of this work. Below is a summary of several broad categories of gene homology:

### ***Orthologs***

Perhaps the simplest type of homologous relationship is orthology – by definition, orthologs are genes that whose ancestor was a single gene. In other words, orthology is the result of a single gene diverging into two genes due to a speciation event.

### ***Paralogs***

The second type of homologous relationship is paralogy. Paralogs are defined as genes that are related by a gene duplication event.

### ***Inparalogs and Outparalogs***

Paralogs across two species can be related either by a gene duplication that occurred pre-speciation, or by a gene duplication that occurred post-speciation. In the event that the duplication occurred pre-speciation, the genes are known as outparalogs, or less commonly alloparalogs. In the event that the gene duplication event occurred post speciation, the paralogs are known as inparalogs, or symparalogs. Of note is that this definition makes no distinction between within-species paralogs and across-species paralogs.

### ***Co-orthologs***



A paralogous expansion in a given species can give rise to a type of relationship known as co-orthology. For instance, a particular gene can expand to numerous inparalogs in species A while remaining a single copy in species B. In this scenario, all of the inparalogs in species B are considered co-orthologous to the single copy gene in species A.

### ***Xenologs***

In practice, orthologs are usually defined as genes that are the mutual best sequence comparison match between genomes. Xenology occurs when one of the genes involved in a mutual best-hit scenario was acquired by horizontal gene transfer as opposed to common descent from its mutual best hit.

### ***Pseudo-paralogs***

A species may acquire a gene by horizontal gene transfer that bears sequence or functional similarity to a native gene. This scenario is known as pseudo-paralogy.

## **2.4: GENOME SEQUENCING AND ANNOTATION**

The genome sequencing and annotation process is a multi-stage endeavor, with opportunity for refinement and improvement at each stage. A difficulty associated with this model is that a change in one step can alter the results obtained in future steps. This would not pose a particularly complicated problem were the steps performed in a strict order; however, the process moves non-linearly, with the first versions of annotation often occurring prior to the publishing of the finished sequence.

The process of adding biological insight through explanatory text or positional specification to an assembled genome is known as annotation <sup>25</sup>. Protein coding genes are the most common form of annotation attached to a genome sequence, usually by an automated annotation software package. Once the gene coordinates on the parent sequence (usually a

contig or chromosome) are established further refinement occurs by the process of functional annotation, which is the assignment of putative function to the predicted gene, most often by sequence-based homology to well-studied genes of known function <sup>26</sup>. The functional annotation can take the form of free text, or, increasingly commonly, as structured Gene Ontology codes <sup>15</sup>. Though protein coding genes are the most often annotated feature, many other sorts of features are commonly annotated as well – including, but not limited to: promoters, regulatory regions, repeat regions, transposable elements, RNA genes, and telomeric features.

Gene annotation and functional annotation often occur well before the final release of a finished genome. For example, the two chromosome, 7.4 million base pair (Mbp) genome of the opportunistic pathogen *Burkholderia pseudomallei* 1106b is still in the stage of 241 contigs; nonetheless, 7738 protein-coding genes have already been computationally predicted <sup>27</sup>. Furthermore, determining the absolutely most accurate set of gene predictions for a genome is complicated by the wide variety of gene prediction software tools <sup>28</sup>, the inevitable rise of newer tools, the variability in interpreting results from those tools and errors in the sequencing or assembly.

The increasing speed and decreasing costs associated with genome sequencing suggest that soon multiple strains and variants of a particular species will be sequenced. In such cases it is likely that many of those genome sequences will never be truly finished, instead existing in a state of continuing refinement.

Widely studied parasite genomes such as *Leishmania major* <sup>17</sup> and *Plasmodium falciparum* undergo frequent updates in genome assembly and gene prediction after the initial release of a draft version to the research community. These updates can include reassembly of improperly assembled regions, sequencing of previously unsequenced regions, elucidation of

previously unannotated genes, and improvement (via refinement, addition or subtraction) of prior gene predictions. Scientific imperative dictates that initial stage analysis and publication of results be performed before a “finished” genome is elucidated. Furthermore, as mentioned above, a genome will never be truly complete, given the continual advancement of the technologies available for automating and improving the sequencing and annotation processes.

Significant high-throughput work is often performed on these draft genomes. Such work includes comparative genomics, microarray analysis and proteomics. The results of these types of analysis are then further propagated to increasingly more species. This propagation of result sets, which are prone to change and differential interpretation leads to a sort of pyramid of data. For instance, species such as *Leishmania major* are well curated and annotated, other *Leishmania* species such as *L. infantum*, *L. braziliensis*, *L. mexicana* are much less extensively curated and can benefit from detailed pairwise comparisons to *L. major*. As an example of the need for multi-genome comparisons, the closely related *Burkholderia* species *Burkholderia mallei*, *Burkholderia pseudomallei* and *Burkholderia thailandensis* have thirty-five sequenced strains between them, each in various stages of completion.

The above issues will only become increasingly salient and complicated as time passes: The genome online database lists 684 currently published complete genomes and over three times as many genomes projects (2312) in progress <sup>29</sup>.

While much valuable insight and information can be gained by comparison of highly annotated genomes to less annotated genomes, no methodology exists for propagating those changes in a structured manner across the genomes. The full process of describing and propagating the aforementioned relationships and data will likely require an extensive set of

software, and as such is beyond the scope of this work. Nevertheless, the base technology for implementing such a system is a well-structured understanding of the types of relationships among genes.

## **2.5: INFERRING FUNCTION THROUGH SEQUENCE SIMILARITY**

The overall field of functional genomics encompasses a great number of techniques, both wet bench and computational. However, the sheer amount of information gleaned from the genome projects is rapidly outpacing the ability of researchers to understand and fully leverage the data by conventional means<sup>30</sup>. As such, there is tremendous added value in widely applying information gained by in-depth study of a particular gene or set of genes to homologous genes that have yet to be functionally annotated. As an increasing number of similar genomes are sequenced, pairwise and group sequence-based comparisons can elucidate genes and other conserved sequence features that have not been predicted by *ab initio* methods<sup>31</sup>.

At the most basic level comparative genomics consists of the comparison of a particular protein or amino acid sequence to another sequence or group of sequences. Pairwise comparisons of sequence are performed using sequence comparisons algorithms such as BLAST<sup>32</sup>, Fasta, or 'cross\_match.' Assessment of similarity across more than two sequences typically involves multiple sequence alignment (MSA) tools such as ClustalW<sup>33</sup>. These types of analyses establish sequence-level of similarity between two or more genes and elucidate conserved sequence motifs or regions.

Sequence similarity tends to imply common descent and hence common function. Orthologous relationships across genomes tend to imply highly conserved function, whereas paralogous gene expansion tends to imply functional diversification. Theoretically, a full phylogenetic analysis is necessary in order to fully elucidate orthologous and paralogous

relationships; however, for pragmatic computational reasons, surrogates (notably, mutual best blast hit) can be used to approximate a full phylogenetic analysis <sup>11</sup>.

## **2.6 SEQUENCE COMPARISON**

In practice sequence comparison requires the use of a particular software algorithm, each with its own strengths and weaknesses <sup>32,34,35</sup>. In this work, we do not discuss specifics of those algorithms, instead focusing on general strategies and issues that are applicable to any methodology.

The underlying premise behind sequence comparison is the assessment of similarity between gene sequences, represented as a series of alphabetical characters <sup>36</sup>, and determination of a score that reflects the degree of similarity <sup>37,38</sup>. Though we typically refer to comparing one group of genes to a second group of genes, it is important to note that these types of comparisons represent a pairwise comparison of every gene in the first group to every gene in the second group. Typical sequence comparison programs will formulate a score that represents the level of similarity between each pair of sequences in the two groups, remove all comparisons that fall below a certain threshold and report the remaining matches.

The techniques presented in this paper employ sequence comparison as a surrogate for building phylogenetic tree representations of gene evolution. While such phylogenetic analysis is the gold standard for elucidating homologous relationship, these methods are not particularly practical in the context of comparing full gene content across multiple genomes; furthermore, in practice sequence comparisons can well approximate the results obtained from phylogenetic analysis <sup>13,14,39,40</sup>.

## **2.9: ONTOLOGIES**

In the last two decades ontologies have grown from an esoteric branch of philosophy to a key technology in fields as diverse as enterprise software, the semantic web and functional genome annotation <sup>41</sup>. Ontologies provide a structured and extensible means of annotating data, thus providing a standardized platform for communicating and sharing of data. Ontologies also contain semantically rich relationships between terms, thus providing software agents with a set of rules by which to query and make inferences from the ontology structure. Finally, ontologies provide a centralized vocabulary by which decentralized software tools may interoperate.

Data sharing is becoming progressively more complicated as the pace of data generation increases and the number of data sources multiply. Ontologies are increasingly serving as a basis on which data sharing tools are built <sup>42</sup> and as a means for rectifying information across information sources <sup>9,43</sup>.

The data-sharing problem in the biological sciences is many-fold:

1. In most scientific disciplines, many terms are used to describe the same concept.
2. Researchers tend to describe a given concept to varying degrees of granularity.
3. Software tools do not interact well with each other.

As to the first point, ontologies do not *per se* solve the problem of numerous terms for a given concept. However, ontology development does initiate the process of standardizing vocabularies. Large consortiums, like the Gene Ontology Consortium <sup>15</sup>, can form discipline-wide consensus on definitions. At an abstract level, projects like the Relationship Ontology <sup>44</sup> and the Basic Formal Ontology <sup>45</sup> are taking steps to form consensus on the basic meta-principles behind relationships in biological ontologies. Even biological ontologies that have arisen independently with heterogeneous terms have been mapped to each other, for example the mappings <sup>46</sup> of Enzyme Commission and Prosite terms to the Gene Ontology.

As to the varying degree of granularity to which researchers describe a given term, the very structure of ontologies allow for variable specificity of description. Take for example gene annotation using the Gene Ontology: due to lack of information, interest, or time, a researcher may simply annotate a particular gene as a 'helicase.' Another researcher may annotate the same gene with a more specific term such as 'ATP-dependant DNA helicase.' Ontologies specify relationships and their properties, so given the 'is\_a' relationships between the more specific and the less specific term, a human or software agent can easily identify that one researcher has simply provided a more granular term and that the two annotations, while different, are not contradictory.

Another facet of the granularity issue is that highly granular annotations do not lend themselves well to summary statistics. The structure of ontologies allows for higher-level aggregation of granular data, which can create a more manageable summary-level view. An example of this functionality are GO-Slims, which are simplified subsets of the full Gene Ontologies which provide a "Bird's Eye View" of the functional annotation.

The final issue, the inability of software tools to interact with each other, is a matter that is addressable at many levels. An ideal solution would involve software tools that could directly interface with each other with no intermediate communication layers. This solution, while attractive, is not likely to occur soon in genome research. A more realistic and reachable goal is that of software that semantically annotates its output according to some discipline-wide standard, with the notion that other software that adheres to the standard can understand the output. To a degree this has been achieved in the field of functional gene annotation by the Gene Ontology (GO.) While researchers are increasingly employing GO annotations, our informal analysis of available functional annotation reveals that they are not necessarily structuring the annotations in a uniform matter, thus requiring a certain amount of

text parsing. Although not optimal, text parsing of GO codes is tremendously less difficult than parsing natural language free text descriptions of gene function.

Researchers are increasingly adopting the Sequence Ontology <sup>5</sup> (SO) as a tool for standardizing genomic annotation. The SO provides a structured vocabulary for describing annotation terms such as ‘Chromosome’, ‘CDS’, ‘RNA’, etc. These terms are not inherently complicated, however SO ameliorates issues pertaining to the ambiguity of certain terms (*e.g.* ‘gene’ versus ‘CDS’) and variability in the use of terms across research groups. Furthermore, SO provides varying degrees of granularity, so for instance an RNA can be described in the most general terms as a ‘transcript’, or in slightly more specific terms as an ‘RNA’, or in any number of more specific terms such as ‘snoRNA.’

The above issues are largely unaddressed for genome comparisons. A suitable exchange format for such data is critically important given the number of comparative tools (BLAST, cross\_match, Clustalw), the number of sources of comparative data (organism-specific resources such as GeneDB and PlasmODB; omnibus resources such as GenBank), and the ubiquity of such analyses.

## **2.10: CHADO**

Chado is an ontology-driven, open source, extensible, generic database schema for the representation of biological knowledge <sup>7</sup>. Chado was created for the FlyBase project, which aimed to accrue and make public genotypic, phenotypic and molecular data from the well-studied model organism *Drosophila*. Despite the very specific aim of the project, the developers took pains to ensure that their work would remain flexible enough to apply to other organisms and new types of experimental data, while maintaining sufficient structure such that standard genomic software tools can be designed atop Chado.



The Chado schema was designed such that entities, attributes of entities, and relationships between entities could all be semantically typed using ontology terms. In this model, entities could refer to sequence (e.g. chromosomes), locatable feature on sequenced entities (e.g. genes, splice sites), sequence comparison matches, or quantitative results (e.g. microarray results).

The work described in this dissertation relies heavily on Chado for data storage and retrieval.

## **2.11: LEISHMANIA**

In Chapter 7 we will present the results of our comparative analysis of four *Leishmania* species. The analysis will be performed using the novel technologies we outlined in Chapter 1. Approximately 20 species of the protozoan parasite *Leishmania* are human pathogenic -- worldwide incidence of leishmaniasis is estimated at 2 million per year with approximately 250 million individuals at risk. A complex array of host and pathogen factors result in a wide range of clinical manifestations<sup>47</sup>. Symptomatic disease can manifest itself in one of three forms (listed in order of increasing severity): cutaneous, mucocutaneous, and visceral. Further complexity in disease manifestation is introduced by regional variation among *Leishmania* species.

Comparison of the genomes of the three extensively sequenced *Leishmania* genomes (*L. major*, *L. infantum*, *L. braziliensis*) reveals that the genomes are relatively highly conserved in terms of gene order and content<sup>48,49</sup>. Nonetheless, the same studies have elucidated approximately 200 differences at the gene or pseudogene level that are potentially responsible for the differences in clinical manifestation observed across the three species.

The *Leishmania* parasite has a complicated life cycle involving a series of morphological and functional changes as it undergoes differentiation from its insect (sand fly) vector stage to its

host stage <sup>50</sup>. Although differentiation is complicated by the presence of several distinct intermediate steps, the process fundamentally involves morphing from the flagellate promastigote stage in the vector to the amastigote stage in the mammalian host.

## **2.12: COMPARATIVE GENOMICS FOR DRUG DISCOVERY**

Whereas the comparative genomics study discussed in **Chapter 7** covered a comparison between four closely related species within the same genus, this comparative genomics study assesses three relatively distantly related human pathogens: *Burkholderia pseudomallei*, *Rickettsia prowazekii* and *Mycobacterium tuberculosis*.

The *Leishmania* study aims to find relatively minor differences between four similar parasites to uncover a (likely relatively small) group of genes responsible for differences in virulence and pathogenicity between the organisms. This broader study aims to form orthologous clusters for the Seattle Structural Genomics Center for Infectious Disease <sup>51</sup> (SSGCID) drug target discovery project <sup>52</sup>.

The aim of the SSGCID drug target discovery project is to elucidate the protein structure for potentially drugable targets in a list of emerging or weaponizable pathogens <sup>53</sup> identified by the National Institute for Allergies and Infectious Disease (NIAID). Although the project has numerous areas of focus and employs a number of discovery strategies, one particular area of research is that of finding closely related orthologs across a wide group of pathogens. Such ubiquitously present genes represent potential drug targets for wide-spectrum infectious disease therapies, and as such are particularly interesting to researchers.



## **CHAPTER 3: ONTOLOGY CREATION**

### 3.1: OVERVIEW

We have created two ontologies for the *semantic annotation* of comparative genomics results; the first is the Pairwise Genome Comparison Ontology (PGCO) for describing the results of sequence-based comparisons of gene sequences from pairs of genomes, and the second is the Gene Homology Ontology (GHO) for describing homologous relationships between pairs of genes.

Though the two ontologies both serve to describe some aspect of a relationship between two sequences, they operate on distinctly different levels and describe these relationships in fundamentally different ways. The PGCO is designed purely to describe a particular match (gene-to-gene similarity) in the context of all the matches generated by a group of sequence comparisons. *Context* refers to the quality of that match relative to the other matches and whether that match was generated by a comparison of a group of sequences to itself or some different group of sequences. The PGCO also describes the reciprocal context to any match; in other words a match between *Gene\_A* and *Gene\_B* might be the highest scoring match in a particular search, but the reciprocal match from *Gene\_B* to *Gene\_A* may not necessarily be the highest scoring match in the reciprocal comparison. These concepts are further enumerated in **Section 3.2**.

The GHO, on the other hand, describes homology relationships that were a result of some particular evolutionary dynamic. The GHO certainly covers a set of concepts that are similar to the PGCO; however the GHO factors in multiple gene-to-gene relationships to determine

the nature of the homology relationship between any two given genes. As such, despite some apparent similarities, the two ontologies operate at distinctly different levels. Furthermore, the PGCO can describe sequence comparison results between any set of DNA or protein sequences, the GHO is designed strictly for use with annotated genes as it contains terms that are associated with gene evolution.

The PGCO serves as an intermediate step between raw sequence comparison results and assigning GHO terms to those results; the relationships listed in the GHO are all computable by sequence comparison and furthermore are computable solely by the PGCO terms assigned to a particular sequence comparison match. While the GHO builds on the PGCO in this work, particularly in our rule-based classification system described in **Chapter 4**, it bears mentioning that the two are independent of each other and are structured such that they are amenable to separate use.

The GHO is not meant to exhaustively cover all topics associated with gene homology (or the topic of homology in general); instead it focuses on the types of relationships we can elucidate using pairwise sequence comparison. A more in-depth description of homology relationships is provided by the Homology Ontology <sup>54</sup>, which covers topics ranging from molecular level homology to functional similarity and phenotypic mimicry. However, as mentioned earlier, the structure of this work is such that a researcher may substitute other ontologies into our overall framework and still achieve the principles of *semantic homology annotation*, *semantic gene grouping* and *logical group querying* with no modification to the logic of the overall work and very minimal changes to the rule/code-base.

While the ontology encompasses a set of terms that are largely comprehensive in regard to most genome comparisons, some researchers may choose to add more terms; similarly some researchers may choose to employ an entirely different ontology all together. The GHO, as

with all well-structured ontologies, is fully amenable to the adding, removing and refining of concepts. However, a particular advantage to the overall strategy lies in the flexibility of assigning of ontology terms using the rule-based system. Additionally, the rule-based classification system allows for easy mapping of ontology terms; this permits researchers to potentially extend the rule-base to new ontologies.

### **3.2: PAIRWISE GENOME COMPARISON ONTOLOGY (PGCO)**

The phrase *pairwise genome comparison* typically refers to comparing the set of genes from one genome to the set of genes from another genome. Furthermore, researchers often run reciprocal comparisons and self-comparisons as well. In other words, for a comparison of *Genome A* to *Genome B* involves comparing all the genes from *Genome A* to the genes from *Genome B* and then the genes from *Genome B* to the genes from *Genome A*; the next step is a comparison of all the genes from *Genome A* to themselves and all the genes from *Genome B* to themselves.

The self comparisons (*A* to *A*, *B* to *B*) are performed to find inparalogs and elucidate gene families within a genome. While the *A* to *B* and *B* to *A* comparisons are in many senses the same comparison, in practice most genome researchers run both due to pragmatic issues related to querying and organizing the search results.

While the PGCO ontology is designed to handle the above-described pairwise genome comparisons, it is capable of handling any sort of sequence comparison. It can describe the results of a one-way genome comparison (e.g. *A* to *B*); an internal genome comparison (e.g. *A* to *A*); or any other type of sequence comparison strategy.

This ontology was originally developed to describe the results of comparisons using the Basic Local Alignment Search Tool (BLAST)<sup>32</sup>. However, the ontology terms are applicable to the

results of any type of sequence comparisons. Later this dissertation describes results obtained from the Fasta sequence comparison tool <sup>35</sup> (not to be confused with FASTA format), and we have effectively used this ontology for analysis of `cross_match` results as well.

The PGCO ontology is designed to describe three properties of sequence comparisons: *internality/externality*, *quality* and *reciprocity*.

### ***Internality Versus Externality***

The idea of *Internality* versus *externality* describes whether a match occurs between two genes from within the same *groups* or two genes from different *groups*. For the types of comparative genomics data described here *groups* tend represent genomes, the notions of internality and externality can apply to any grouping of genes.

### ***Quality***

*Quality* refers to whether a match represents the highest scoring match for a given gene in a given comparison. “Highest scoring” has numerous connotations, depending on the sequence comparison tool used and the selected scoring options; however, scores are presumed to serve as a directly proportionate surrogate for closeness of evolutionary relation. The PGCO ontology differentiates between the “best hit” and “secondary hits” (all other non-best hits).

Since the ontology is optimized for pairwise genome comparisons, it is likely that a gene will have more than one best hit, an internal best hit describing the best match from within the genome and an external best match from the other genome in the comparison. It is also possible for a gene to have no best matches (internally, externally or at all), *i.e.*, in the case of a truly novel gene that has no discernible sequence comparison match to any other gene in the comparison group.



**Reciprocity**

“Best matches” are not necessarily reciprocal. In other words, in a comparison of *Genome\_X* to *Genome\_Y*, the fact that *Gene\_X1* is the closest match to *Gene\_Y3* according to some scoring metric does not necessarily imply that *Gene\_Y3* is the closest match to *Gene\_X1*; there may be some other gene from *Genome\_Y* that is more similar to *Gene\_X1*. In this case the match between *Gene\_X1* and *Gene\_Y3* is a ‘unidirectional best match’.

Amongst the secondary matches (non-best-matches) there are two three types of reciprocal relationships: ‘proximate’, ‘intermediate’, and ‘distant’. Proximate secondary matches are matches such that the match is a non-best-match in the context of one gene, but the best match in the context of the other gene. This scenario is the inverse of the ‘unidirectional best hit’ scenario described above; a match that is a ‘proximate secondary’ match from the point of view of one gene in the match would be a ‘unidirectional best hit’ from the point of view of the other gene in the match. An ‘intermediate match’ refers to a match that is a non-best-match from the point of view of both genes involved in the comparison. Finally a ‘distant match’ refers the scenario in which a *Gene A1* from *Genome A* matches *Gene B2* from *Genome B* in a comparison of *Genome A* to *Genome B*; however, in a comparison of *Genome B* to *Genome A*, *Gene B2* does not match *Gene A1*. While this scenario may seem implausible, most sequence comparison tools have certain score cutoffs below which matches are not reported; furthermore, scores are often influenced by factors such as length of query sequences, size of the total number of query genes and other factors that are not necessarily equivalent in both directions of reciprocal pairwise genome comparison. These ‘distant matches’ represent very tenuous sequence similarity and as such will barely meet score cutoff under a certain conditions and fail to meet those cutoffs if the conditions slightly change. In practice we ignore these types of matches.

**PGCO Structure:**

As mentioned above, the PGCO is built to describe the concepts of quality, internality/externality and reciprocity in the context of sequence comparison matches. The PGCO can be separated into four conceptual levels (denoted by color in **Figure 3-1**). Each level describes a particular concept in cross-product with the concept(s) described in the level above it. In other words, each section successively refines the term in the ancestor level.

**The first level (shown in black, Figure 3-1)** describes very basic attributes about a match: the root term *PGCO:match* and terms to describe self and non-self matches. Note that the term *PGCO:self\_match* has no descendant terms. In a comparison of a group of sequences to itself each sequence will have a self-match; that match will be of perfect quality and there are no further useful refinements to describe such a match.

**The second level (shown in orange - Figure 3-1)** contain terms that describe the principles of internality/externality and best/secondary hit separately. Although in practice most users of this ontology would not annotate matches with terms at this level, these terms are included both for completeness; some users may find these term relevant for describing certain types of experiments. Furthermore, these higher-level terms serve as a means to aggregate the more specific lower-level terms. Since all the terms are related by *is\_a* relationships (in other words, any term satisfies all the conditions of its parent terms) a user could use the term *PGCO:internal\_match* to aggregate all the matches that have been described with a more specific descendant term.



**Third-level terms (shown in green - Figure 3-1)** combine the concepts of best/secondary match with the concepts of externality/internality. Furthermore, this level introduces the concept of reciprocity/unidirectionality along with the analogous concepts of proximate/intermediate/distant matches.

Finally, **fourth level terms (shown in blue - Figure 3-1)** combine all three concepts and are suitable for describing pairwise sequence comparison experiments of the type described earlier in this section (*Genome A* compared to *Genome B*).

This dissertation exclusively uses terms from the fourth level of the ontology to describe our sequence comparison results from our pairwise genome comparisons. **Appendix A** fully describes each term in detail. Furthermore, each PGCO term has a computable, formal definition.

#### ***Use of the PGCO***

The above-described principles apply to results generated by any sequence comparison algorithm. Furthermore, the structure of the ontology is such that a researcher can use the terms to describe the most simple or the most complicated sequence comparison experiments. A researcher running a simple comparison of a group of genes against a publicly available database (Genbank, for instance) could use the terms *PGCO:best\_match* and *PGCO:secondary\_match* to describe the results. Of course, such a simple comparison would likely not benefit much from the use of these two terms; however, sequence comparisons can quickly become significantly more complex. For instance in the aforementioned search against a public database, some of the database sequences might belong to the same species as some of the genes in the query group. In such a comparison, level two terms like *PGCO:external\_best\_match* and *PGCO:internal\_best\_match*, and their corresponding

secondary match terms, become useful. Species is not the only criterion by which we can describe internality and externality; the group of query genes might belong to some class of genes (say genes involved in some pathogenic process), a researcher could use internality and externality to describe matches within and across such categories.

The true power of the PGCO arises when moving into reciprocal comparisons. In other words, when comparing some set of genes ( $X$ ) to some other set of genes ( $Y$ ). In such a case level three terms like *PGCO:reciprocal\_best\_hit* or *PGCO:unidirectional\_best\_hit* provide context into how the match ranks from the perspective of both genes. Furthermore, reciprocal searches that have some sort of group membership criteria (like species) can employ level four terms such as *PGCO:external\_reciprocal\_best\_hit* to provide full insight into the context of the match.

### **3.3: GENE HOMOLOGY ONTOLOGY (GHO)**

The GHO (Figure 3-2) describes homology relationships between a pair of genes. A “pair of genes” refers to any two genes that have a discernible sequence comparison match; as such any gene will have as many GHO terms as it has sequence comparison matches. A researcher can employ any number of methodologies to describe relationships by GHO terms; however, in this dissertation we use PGCO terms to classify the sequence comparison matches and then subsequently compute (using Prolog rules) the GHO terms associated with each match. While the two ontologies are intentionally structured such that they function independently, we feel that the PGCO serves as a useful foundation by which to compute GHO terms. The GHO terms are fully described in **Appendix B**.



### **Structure and symmetry**

Every term in the Gene Homology Ontology is related by an 'is\_a' relationship to its parent. As a consequence every term has all the properties of its parent term, plus some further refinement of the parent term. Furthermore, a term can have more than one parent; a multi-parent term will have all the properties of all of its parent terms. The GHO describes relationships between pairs of genes; these relationships can be thought of as binary relation of the form:  $(x, R, y)$ . In this case the variables  $x$  and  $y$  are genes in either of the compared genomes and  $R$  is a term from the GHO. Many of the GHO terms are symmetric binary relations, in other words the relationship  $(x, R, y)$  necessarily implies the inverse relationship  $(y, R, x)$ . The rest of the GHO terms are asymmetric meaning that the relationship  $(x, R, y)$  implies the inverse relationship  $(y, R2, x)$  where  $R$  and  $R2$  are not equal. The asymmetric relationship terms in the GHO fall into two distinct categories: one in which the inverse relationship  $R2$  can be directly determined by the relationship  $R$  and another category in which  $R2$  can take on a number of possibilities based on  $R$ .

In **Figure 3-2** the blue boxes represent symmetric GHO terms; the grey boxes denote asymmetric terms that have a specific term, connected by a red arrow, that describes the inverse relationship; the brown boxes represent asymmetric terms that can have as any number of inverse relationships. The terms outlined in orange (*inparalog* and *outparalog*) represent terms whose formal computable definition is a union of the definitions of their descendant terms. For instance the term *inparalog* has a well-understood definition, however computing *inparalogs* entails separately computing *internal inparalogs* and *external inparalogs*.

**Figure 3-2** shows the complete Gene Homology Ontology. For organizational purposes the ontology has been partitioned into four sections: the *General Homology Section*, *Ortholog Section*, the *Outparalog Section*, and the *Inparalog Section*.

### **General Homology Section**

The *General Homology Section* contains broad types of relationships that are typically appropriate for summary-level analysis of genome comparisons. The root level term *GHO:homolog* describes any relationship detectable by sequence comparison. This section also contains terms to describe internal and external homologs, as well as closest (most similar in a given comparison) homologs.

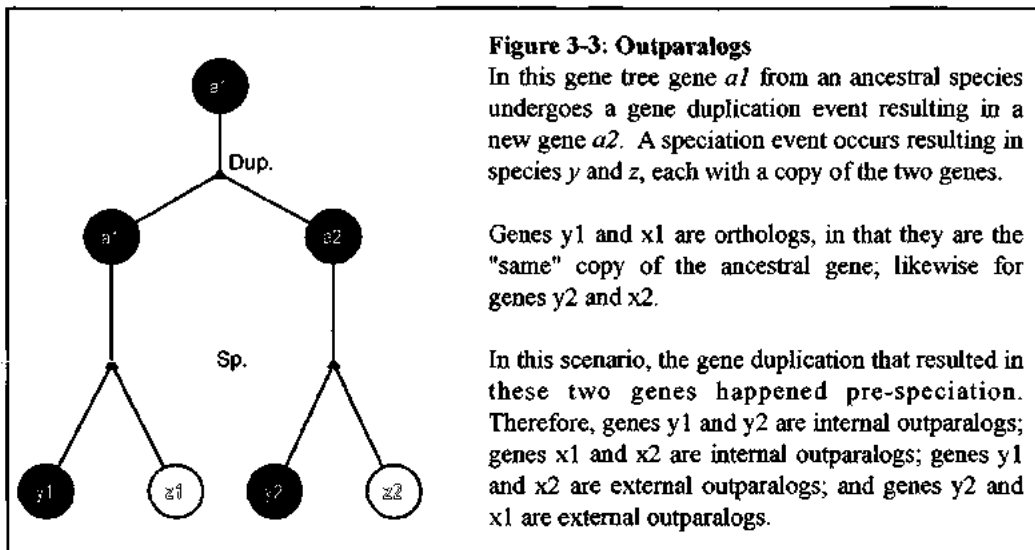
### **Ortholog Section**

The *Ortholog Section* describes orthologs and results of fusion/splice events. The term *GHO:ortholog* is used as a general term to describe orthologs (reciprocal best matches) that arose as a result of the speciation event that separated the two compared species. The terms *GHO:fusion* and *GHO:splice* have been placed in this section as well because a fusion gene does represent the functional equivalent of the two spliced genes that it spans. Though some fusion or splice event (by definition) occurred to create this relationship, at some point in the evolution of these genes they were separated by a speciation event. Given that the fusion/splice relationship maintains the evolutionary condition of an orthologs (separated by speciation) and the functional condition (encode for the same functional roles), these relationships do represent a type of orthology. Furthermore, the fusion/splice relationship is often the results of a mistake in gene predictions; two genes may mistakenly be annotated as one or one gene may be mistakenly annotated as two.

### **Outparalog Section**



The *Outparalog Section* describes homology relationships that are a result of gene duplications before the speciation event that separated the compared genomes. The term *GHO:internal\_outparalog* describes outparalogs in the same genome and *GHO:external\_outparalog* describes outparalog from different genomes while the general term *GHO:outparalog* includes both types. **Figure 3-3** details these relationships.



### ***Inparalog Section***

The *Inparalog Section* contains terms that describe relationships between genes that have duplicated subsequent to the speciation event that separated the compared genomes. The general term *GHO:inparalog* describes all such genes. However, in any inparalog relationship there is a gene that was present in the ancestral genome and one or more gene(s) that arose from the post-speciation duplication events. This section of the GHO is structured to describe the intricacies of such relationships.

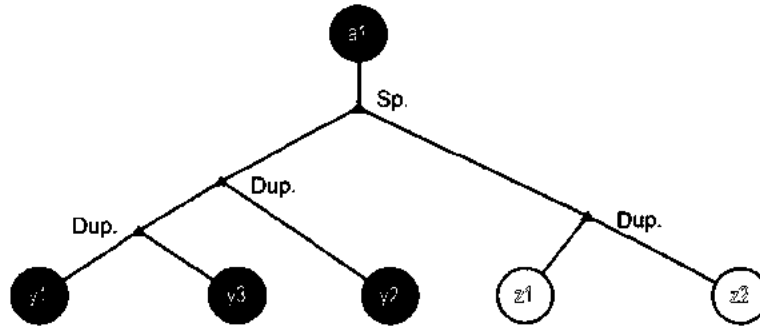
When a gene duplicates post-speciation and subsequently evolves, one copy of the gene typically maintains a greater degree of sequence similarity to its original form than the other gene. This phenomena holds true when a gene undergoes multiple duplications; one copy of

the gene still maintains a relatively fixed sequence, while the rest diverge to some degree. This divergence can be minimal and unlikely to affect the function, or it can be more drastic and result in neofunctionalization or subfunctionalization. Regardless of degree of divergence, the copy of the duplicated gene that maintains the most similarity is the most likely to have maintained the original ancestral (pre-speciation) function of the gene.

In the case of *GHO:internal\_inparalogs* - post-speciation duplications of a gene within a genome - the ancestrally present gene has a *GHO:internal\_parent\_inparalog* relationship to the genes that arose post-speciation, those genes in turn have a *GHO:internal\_child\_inparalog* relationship with the ancestrally present gene, and a *GHO:internal\_sibling\_inparalog* relationship with each other.

In the case of *GHO:external\_inparalogs* - gene duplications of an orthologous gene - the ancestrally present gene is the *GHO:external\_parent\_inparalog* of the resultant *GHO:external\_child\_inparalogs*. Furthermore, in the event of independent post-speciation gene duplications in both compared species the resultant *child inparalogs* have a *GHO:external\_sibling\_inparalog* relationship with each other. The various types of internal and external inparalogs are illustrated in **Figure 3-4**.

The *Inparalog section* also contains the term *GHO:species\_specific\_inparalog* to describe genes that were presumably not present in the ancestral species, however are in one of the compared species in more than one copy. It cannot conclusively be determined whether the gene was present in the ancestral species, however the absence of a gene in one of the compared species serves as an (albeit imperfect) indication that the gene was absent in the ancestor and arose post-speciation. Note that *GHO:species\_specific\_inparalog* refers to relationships between species-specific genes that have duplicated into a species-specific gene



**Figure 3-4: Inparalogs**

A speciation event creates two species (y & z) each with a copy of the ancestral gene *a1*. In species z the gene duplicates once, resulting in two copies of the gene (*z1* & *z2*). In species y the gene duplicates twice, leading to three copies of the gene (*y1*, *y2* & *y3*). In this example, genes *y1* and *z1* have maintained the greatest degree of sequence similarity to the ancestral gene (*a1*), while the other genes have all diverged at greater rates.

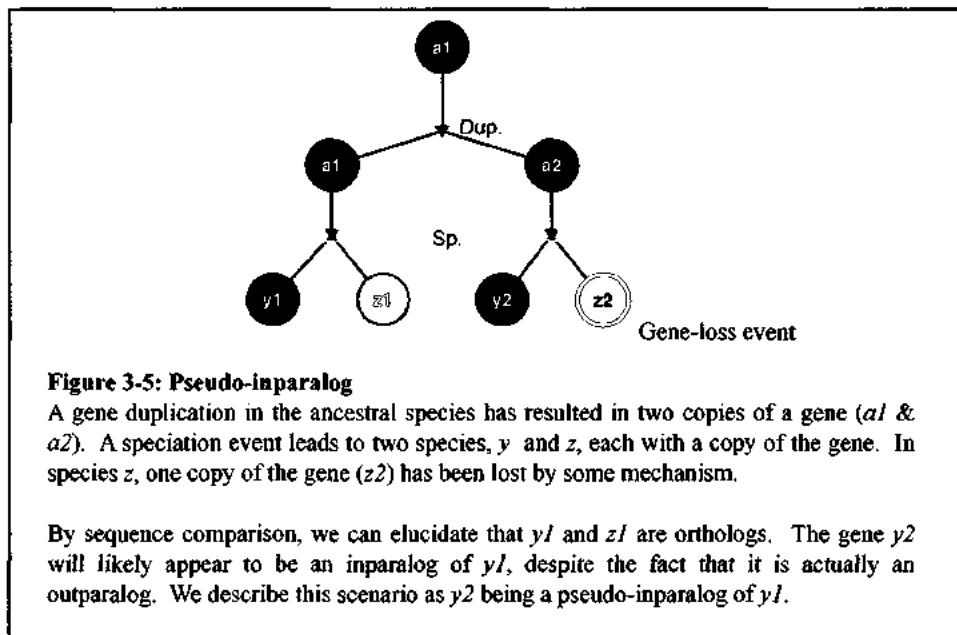
In this scenario genes *y1* and *z1* are orthologs. Gene *y1* is the 'internal parent inparalog' to genes *y2* and *y3* and is the 'external parent inparalog' to gene *z2*. Similarly, gene *z1* is the 'internal parent inparalog' to gene *z2* and the 'external parent inparalog' to genes *y2* and *y3*.

Genes *y3* and *y2* are 'external child inparalogs' to *z1*. Gene *z2* is the 'external child inparalog' of gene *y1*.

Genes *y3* and *y2* are 'internal sibling inparalogs' and gene *z2* is the 'external sibling inparalog' of *y2* and *y3*.

family. These genes differ from novel genes, which are species-specific genes that are truly novel and show no detectable homology either within their own genome or to genes from the comparison genome.

Finally, the *Inparalog Section* contains the term *GHO:pseudo\_inparalog*. This refers to the relationship between genes that appear to be inparalogs, but are actually outparalogs which have undergone a gene loss event. This scenario is described graphically in **Figure 3-5**. There are a number of ways of determining potential gene-loss events and those topics are further discussed in **Chapter 4**.



### 3.4: USING THE ONTOLOGIES

This chapter has described two ontologies that contain terms central to comparative genomics: sequence comparison matches and pairwise gene homology relationships. The next area of research, subsequent to the development of these ontologies, was exploring methodologies for applying these terms to actual sequence comparison results. We decided upon a rule-based system, as we felt that the logical nature of rule-based programming well approximates the logical steps that genome researchers employ to assess their genomic data. The work towards rule-based application of these ontology terms is discussed in **Chapter 4**.

A fundamental principle that directed the ontology development efforts was flexibility of use. That being said, the PGCO and GHO are particularly effective in dealing with assigning homology relationships between pairs of genes in a pairwise genome comparison. As **Chapter 4** will discuss in detail, the PGCO serves as a highly descriptive means of describing the types of sequence comparisons that researchers typically perform to compare gene content across pairs of genomes. Such comparisons typically involve self-*versus*-self comparisons of the genes from both genomes, comparisons of one genome's gene set to the

other's and the reciprocal comparison. The PGCO describes the results of these multi-component analyses in unambiguous terms.

An unambiguous description of the results aids in finding patterns in those results that indicate certain types of evolutionary relationships between a pair of genes; those relationships are then described using GHO terms. GHO terms move beyond the type of vague descriptions often employed in comparative genomics and describe in concise detail the exact relationship between two genes. For instance, when dealing with inparalogs, the GHO describes which inparalog has maintained sequence similarity to its ortholog and which inparalog is evolving. As another example, the GHO can describe complex orthology relationships that result from gene fusions and gene splicing. Just as the PGCO provides an unambiguous means for describing sequence comparisons, the GHO provides an equally unambiguous means for describing homology relationships.

**CHAPTER 4: RULE-BASED CLASSIFICATION OF  
HOMOLOGY RELATIONSHIPS**

#### **4.1: RULE-BASED CLASSIFICATION OVERVIEW**

This chapter describes in detail how to use the ontologies from **Chapter 3** to describe homology relationships between genes. The GHO ontology provides a semantically rich means for describing the various types of relationships; this chapter will discuss the rule-based methodologies that have been developed to apply those results to actual data.

Our rule-based methodology consists four distinct steps:

1. Collection of data from heterogenous sources
2. Applying Pairwise Genome Comparison Ontology (PGCO) terms
3. Removing low-quality sequence comparison results
4. Applying Gene Homology Ontology (GHO) terms.

The above four steps constitute a novel methodology is referred to as *Semantic Homology Annotation* (SHA). A particular benefit of SHA is the ability to easily modify, refine or extend any of the four steps. This versatility is critical both from the standpoint of accommodating diverse research goals, as well as allowing researchers to refine the results of the SHA step based on their own knowledge or the results of other experiments. Furthermore, the SHA process has been structured such that it can serve as a foundation for gene clustering and posing queries to those clusters (described in **Chapters 5 & 6**).

The results of the SHA, gene clustering and the subsequent queries are all structured according to a novel knowledge representation schema. This schema connect the three steps using structured inputs and outputs and provides structured results that are amenable for use as a foundation for future work.

This chapter describes our knowledge representation schema and the four step SHA process. **Chapter 5** describes how to use the SHA results for *Semantic Gene Grouping*, a novel form of gene clustering that we have developed. **Chapter 6** describes a rule-based strategy that has been implemented for querying the semantic gene groups, a process dubbed *Logical Gene Group Querying*. Finally, **Chapters 7 & 8** describe the application of these newly developed technologies towards actual comparative genomics experiments.

## **4.2: KNOWLEDGE REPRESENTATION SCHEMA**

### ***Overview***

This work introduces a knowledge representation schema that represents scientific data from diverse sources in a standard way; this project uses the schema to describe information about genes and sequence comparison results. This knowledge representation method is well suited both for integrating data from diverse types of analysis, and serves as a means of connecting steps in our pipeline process.

A primary goal when developing this schema was to provide a means of knowledge representation that is agnostic to the technologies used to generate the sequence comparison results, as well as the conventions used to store annotations and other information surrounding genes. Furthermore, personal experience has shown that complex file-format specifications are often restrictive in that they do not accommodate changing project needs particularly well owing to their inherent inflexibility. Another goal was to create a system that was well suited for representing data in a manner such that it could serve as inputs and outputs for a rule-based pipeline.

We decided upon a strategy that represents information as a series of fact statements. This method is both extremely simple and highly customizable. Information is specified in a series of assertions about the data and does not rely on a complex data model or file-format



specification. The method is flexible in that it only expects a certain minimal set of information, yet also allows users to also specify any additional properties surrounding our data.

This flexibility is leveraged in the rule-based homology annotation steps describes in this chapter as well as **Chapters 5 & 6**. The pipeline requires very basic input data (gene name and very straightforward sequence comparison information); however a user can specify more information to generate more refined results. Furthermore, in our own experience the data representation syntax has proven to be a useful lightweight data integration platform for a wide variety of scientific tasks, such as processing Gene Ontology annotations as well as microarray probe mapping <sup>55</sup>.

#### ***Knowledge representation using Nodes, Edges, Sets and Properties***

Although the knowledge representation strategy could conceivably describe any sort of information, the representation syntax is structured such that it would be particularly effective in representing genomics data. Our representation syntax consists of four main data types:

- 1 . **Nodes**, which represent some sort of concrete entity.
- 2 . **Edges**, which represent a piece of data that connects two nodes.
- 3 . **Sets**, which represent a collection of nodes or edges.
- 4 . **Properties**, which describe any attribute of a node, edge or set.

The above structure is particularly effective for genomic data for a number of reasons. Generally, the *node* and *property* structure is a flexible way of describing any sort of discrete entity, for instance a chromosome or a gene. Researchers often generate data that assesses the degree of similarity or connectedness between two pieces of genomic data; *edges* are particularly effective in describing these sorts of connections. Furthermore these connections can be further defined and described by adding *properties* to the *edges*. Finally, genome

research often involves aggregating data into groups, which can be described using *sets*. These groups might represent some sort of biological reality (genes belonging to a particular species), some sort of categorical similarity (genes from a number of species that all play a similar functional role) or they may represent some sort of commonality in how the data was generated (BLAST matches from the same search).

This work employs the knowledge representation schema to provide input data into the beginning of the pipeline. *Nodes* are used to describe genes, and the *node properties* can describe the functional annotation, gene location and any other relevant information. All the genes from a particular species are grouped together within a *set*, with *set properties* describing any additional information regarding the species. *Edges* represent sequence comparison matches between two genes (*nodes*) and the *edge properties* describe the statistics, scores, attributes associated with the match. *Sets* describe a BLAST search as a collection of the blast hits (*edges*) generated by that search. *Set properties* can describe attributes of that search, such as the substitution matrix or any user-specified options.

Later, this chapter describes the use of *properties* to attach PGCO assignments to edges and GHO assignments to *genes*. **Chapter 5** describes the use of *sets* to aggregate groups of genes related by particular types of homology relationships. Finally, **Chapter 6** describes how to query these *properties* and *sets* to generate meaningful statements regarding the evolutionary histories of genes.

The remainder of this section (4.2) is devoted to specific examples of how particular data types are represented using our knowledge representation schema.

***Describing a gene:***

```
node('LinJ36.7110').
```

```

prop('LinJ36.7110', name, 'LinJ36_V3.7110').
prop('LinJ36.7110', description, 'ATPase, putative').
prop('LinJ36.7110', srcfeature, 'LinJ_chromosome_36').
prop('LinJ36.7110', strand, -1).
prop('LinJ36.7110', fmax, 2615975).
prop('LinJ36.7110', fmin, 2614380).
prop('LinJ36.7110', member_of, 'LinJ_V3').
prop('LinJ36.7110', type, gene).

```

The above example describes a gene from *Leishmania infantum* as a node that has the identifier: 'LinJ36.7110'. In this particular example the node identifier is the same as the name that the genome project assigned the gene; however, any sort of unique identifier can be used, for example a unique number or text string. This example also describes the functional annotation ('ATPase, putative') and uses the Chado 'feature' table conventions to describe that this gene is located on 'LinJ\_chromosome\_36', on the negative strand from base pair 2,615,975 to base pair 2,614,380.

### **Describing a Genome**

```

set(LinJ_V3).
prop(LinJ_V3, version, 3.0).
prop(LinJ_V3, name, 'LinJ TriTryp, mRNAs').
prop(LinJ_V3, type, 'Genome mRNAs').
prop(LinJ_V3, description, 'L infantum genes,TriTrypDB ).

```

This example describes a genome, *Leishmania infantum* (Version 3.0). Every member of this set has a special 'member\_of' property that specifies that it belongs to the set:

```

prop('LinJ36.7110', member_of, LinJ_V3).
prop('LinJ01.0100', member_of, LinJ_V3).
prop('LinJ23.5020', member_of, LinJ_V3).
etc..

```

### **Describing a BLAST Hit**

```

edge(bhid_2, 'LmjF01.0010', 'LinJ20.0360').
prop(bhid_2, query_hit_stop, 330).
prop(bhid_2, frac_identical, 0.359).
prop(bhid_2, sbjct_strand, 1).

```

```

prop(bhid_2, sbjct_hit_start, 13).
prop(bhid_2, score, 119.7).
prop(bhid_2, evalue, 8e-28).
prop(bhid_2, frac_positive, 0.641).
prop(bhid_2, type, blast_hit).
prop(bhid_2, member_of, aid_60).
prop(bhid_2, query_hit_start, 330).
prop(bhid_2, sbjct_hit_stop, 312).

```

The above example describes a blast hit (to which we have assigned the unique identifier 'bhid\_2') that connects the query sequence 'LmjF01.0010' to the subject sequence 'LmjF20.0360'. Additional properties describing the quality and location of the BLAST hit have been also provided. Although BLAST was chosen for this particular example, the same principles would apply to representing the results of any sequence comparison experiment.

### ***Describing a BLAST search***

This example describes a BLAST search that compares all *Leishmania major* protein coding genes to all *Leishmania infantum* protein coding genes; this BLAST search has been assigned the unique identifier 'aid\_64'.

```

set(aid_64).
  prop(aid_64, algorithm, fasta).
  prop(aid_64, name, 'LinJ protein vs LmjF proteins').
  prop(aid_64, program, 'blastp').
  prop(aid_64, type, 'blast_search').
  prop(aid_64, sbjct, 'LmjF_v5').
  prop(aid_64, query, 'LinJ_v3').

```

Every member of this set has a special 'member\_of' property that specifies that it belongs to the set, e.g.:

```

edge(bhid_2, 'LmjF01.0010', 'LinJ20.0360')
prop(bhd_2, member_of, aid_64).

```

### ***Flexibility***

The knowledge representation schema consists of a set of facts about some entity. Each fact is represented as a single assertion; one benefit of this methodology is that researchers can represent their data without having to adhere to a standard for well-formed documents, such

as is the case with XML and most other file formats. The data requirements for the overall system described in this dissertation are quite minimal: a researcher only need describe every gene, specify which genes belong to which genomes and describe each sequence comparison match as an edge from the query gene to the subject gene and provide a score for the match.

In summary, the basic requirements for running the homology annotation pipeline are as follows:

Describe the genomes:

```
set(Genome_X).
```

Describe the genes:

```
node(Gene_A).
prop(Gene_A, member_of, Genome_ID).
```

Describe the sequence comparisons:

```
set(Blast_search_ID).
```

Describe each sequence comparison match

```
edge(Match_ID, Gene_A, Gene_B).
prop(Match_ID, score, 100).
```

The above is the minimal set of information that the user needs to supply, however the user can provide more information for more refined (and, presumably, more accurate) results. For example:

1. A researcher can provide information regarding the chromosomal location for each gene. This allows the system to determine positional conservation of genes (a surrogate for presumed conservation of a syntenic block) thereby allowing the system to detect weak homology among positionally conserved blocks of genes.
2. More sequence comparison match information, such as coverage statistics, degree of conservation and number of gaps. This information allows the system

to “break ties” between similarly scoring matches to determine the “best match”. Furthermore, one can use the coverage information to include low-scoring matches for short genes - matches that might be removed as too weak without such information.

3. Additional properties of genomes (e.g pathogenic or nonpathogenic) can be supplied to pose questions regarding patterns of gene presence/absence/loss/expansion between categories of species.
4. Further gene annotations can be included, such as labeling a gene a pseudogene, pre-specifying genes that a researcher knows to be homologs that might not be readily detected by sequence comparison, and conversely specifying that certain genes are not homologs despite some amount of sequence similarity.

### **4.3: RULE-BASED CLASSIFICATION STRATEGY**

#### ***Overview***

This chapter describes this implementation of a rule-based classification strategy for assigning Pairwise Genome Comparison Ontology (PGCO) to sequence comparison results and Gene Homology Ontology terms to genes. Three primary reasons compel the use of a rule-based strategy for this task:

1. Rule-Based programming operates at a higher level of abstraction; a programmer describes what should be done, as opposed to how to do it. This allows users to specify the logic behind the assignments, as opposed to having to write or modify procedural code for making the ontology assignments.
2. Production rules are typically easier to create, modify and extend than most types of computer code. Any sort of programming requires knowledge of variable types, syntax and control-structures; however, rule-based programming tends to have simpler syntax and consists mainly of variables, constants and Boolean operators.

While “difficulty” is highly subjective and often task-specific, for this project, rule-based programming represents the least difficult implementation as far as customization by the end-user is concerned. Rule-bases can become unmanageable as the number of rule increase, however our pipeline approach separates this process into several discrete steps. At each step there are only a certain number of applicable rules and as such this creates a *de facto* sub-setting of the rule-base; thereby keeping the code manageable.

3. Our data representation syntax describes data as discrete facts; this representation structure works well with rule-based programming, and therefore constitutes an efficient and simple means of drawing data from disparate sources into the overall framework.

The remainder of this section will discuss in detail the task of classifying “high-quality” BLAST hits versus “low-quality” BLAST hits. Though classifying BLAST hit quality is not a particularly compelling scientific goal, it does constitute a necessary and important aspect of any experiment that involves high-throughput scientific data. Furthermore, this task provides concrete examples of the three principles described above.

#### ***Classifying Sequence Comparison Hits***

This subsection describes the rule-based process of determining “high-quality” sequence comparison matches. This is not the first step in the pipeline; nonetheless, it serves as a discrete, relatively easy-to-understand process that will serve to illustrate in a more concrete manner how the overall rule-based strategy for applying ontology terms functions **Section 4.4** will cover the entire process in a step-by-step manner.

Any pairwise sequence comparison algorithm will provide a series of matches between the query sequences and the subject sequences; typically, in large sequence comparisons these

matches will lie on a continuum from extremely low-quality matches that likely represent no real shared evolutionary similarity, to short domain-level matches that likely do not represent true functional homology, to high-quality matches that likely do indicate true functional similarity and recently shared ancestry <sup>13</sup>.

Interpreting the results of a sequence comparison experiment involves separating the low-quality matches from the high-quality matches. This is a highly subjective matter, and the matches that one researcher might view as “high-quality” would constitute “low-quality” matches to another equally knowledgeable researcher; nevertheless, there are a set of heuristics that are widely used to perform this task. Furthermore, the rule-based strategy employed here allows relatively easy refinement and extension of the “high-quality” definitions.

Faced with the problem of selecting high-quality BLAST hits, most researchers choose some sort of score or coverage cutoff below which they disregard any matches, for example, the INPARANOID clustering algorithm disregards BLAST hits which have a score lower than 100 and a cover less than 50% of the query and subject sequences <sup>13</sup>. Although heuristics such as these are sound (and quite common), they lead to a number of false negatives. For instance, comparison of two short amino acid sequences might fail to meet a score cutoff (given that most sequence comparison scores are proportional to the length of match), even if the match is of high quality. Furthermore, there are subtle indications of shared ancestry that might be missed when only considering sequence comparison statistics as the arbiter of shared ancestry. For instance a long block of positionally conserved (syntenic <sup>56</sup>) genes across two genomes (*Genome\_A*, *Genome\_B*), might contain two genes (*Gene\_A1*, *Gene\_B1*), which have relatively low sequence similarity, but are reciprocal best matches to each other and are contained within the positionally conserved block. Given this scenario, a researcher would conclude that these genes are orthologous to each other; the *Leishmania*



comparative genomics study (**Chapter 7**) contains several examples of such orthology that fell below traditional scoring cutoffs. This illustrates a scenario in which using purely numerical score cutoffs serves as a substandard methodology for determining high-quality matches. Both of these issues could be solved by lowering the score and coverage cutoffs, but doing so increases the number of false-positives.

The above discussion provides a brief glimpse into the difficulties associated with choosing which sequence comparison results qualify as “good matches”. The rule-based methodology described here provides researchers with a means for applying human logic to the problem of selecting good matches and allows for a more sophisticated means of addressing the issue. Furthermore, choosing good matches is often context-specific, so the fact that rules are easily refined and edited further enhances their suitability for this particular task.

The first step in determining high-quality matches is a rule that states that any match above a certain numerically calculated threshold is automatically included. The example below uses a score of 100; in practice choosing a suitable score is predicated upon the type of comparison and what the researcher hopes to discover.

```
quality_match(MATCH) :-
    prop(MATCH, score, SC),
    SC > 100.
```

As discussed earlier, high-quality matches between short sequences often fail to meet score cutoffs. This rule states that a hit that covers greater than a certain percent of the query sequence or subject sequence is considered high quality:

```
quality_match(MATCH) :-
    prop(MATCH, query_coverage, QC),
```

```
QC > 0.50.
```

```
quality_match(MATCH) :-
    prop(MATCH, subjct_coverage, SC),
    SC > 0.50.
```

Furthermore, a rule encodes the logic of deciding that reciprocal best hits between positionally conserved genes likely implies syntenic orthologs. Note that this example presupposes that positionally conserved is already calculated. In practice the rule-base does indeed have a set of rules for determining positional conservation. The rule encoding this scenario is:

```
quality_match(MATCH) :-
    prop(MATCH, query_id, Q),
    prop(MATCH, subjct_id, S),
    prop(Q, positionally_conserved, SC),
    prop(MATCH, PGCO_term, reciprocal_best_match).
```

The above example shows how rule-based classification of BLAST hits allows for the creation of a highly refined definition of what constitutes a high-quality BLAST hit. Although, owing to the popularity of the algorithm, the examples refer to BLAST outputs, this methodology applies to any sequence comparison experiment. These techniques have been applied to results of the Fasta sequence comparison tool and the cross\_match sequence comparison tool.

#### **4.4: RULE-BASED APPLICATION OF PGCO TERMS**

The prior section (4.2) discussed in detail our rule-based methodology for determining “high-quality” matches. That is not the first step in the process, but was included in the prior section because it serves as a simple introduction to the rule-based strategy. The first step in our pipeline (described here) is assigning PGCO terms to the sequence comparison results.

We apply PGCO terms to all the matches, not just the “high-quality” matches. We do so for a number of reasons. First, the qualities that we look for to determine PGCO terms are present regardless of whether a match meets certain “high-quality” criteria; for instance a “best match” is a best match whether the score exceeds some floor value or not. Second, the types of comparative analyses described here are often iterative and a researcher may wish to change their definition of high-quality matches; the pipeline process is ordered such that they can do so without recalculating PGCO terms. Finally, PGCO terms help determine what constitutes a high-quality match based on reciprocity and positional conservation of genes (see the end of **Section 4.2** for further discussion).

The first step in assigning PGCO terms is determining the highest scoring hit for every query sequence. Ideally this simply involves selecting the hit with the highest score; however, multiple hits may have the same score. The rule-base handles this task by specifying multiple rules that first rank matches based on score, then coverage, and then fraction of identical amino acids (or nucleotides). The rule set for determining highest quality hit is another area that a researcher may edit, as there are many secondary criteria that make effective arbiters between two matches of identical score.

The PGCO assignments are performed subsequent to determining the highest quality hit. Each node in the PGCO corresponds to a rule that returns true or false for every sequence comparison match. The ontology employs ‘*is\_a*’ terms, which thereby mandate that every node is a refinement of its parent node; every rule begins with the provision that the rule(s) attached to the parent(s) must be satisfied.

The pipeline then outputs the PGCO assignments as facts structured in the data representation syntax discussed in **Chapter 4.2**. For example, the following fact asserts that a blast hit with

unique identifier (BH\_123) represents a *PGCO:external\_reciprocal\_best\_hit* between its query and subject gene:

```
prop(BH_123, pgco_term, external_reciprocal_best_hit)
```

After assigning the PGCO terms, the pipeline performs the quality match determination process mentioned in **Section 4.3**; subsequently GHO terms are assigned, as described in **Section 4.5**. Quality match determinations are performed before the GHO assignments because low-quality matches tend to indicate short stretches of very distant relationships in some region of the genes. These tenuous relationships are likely not evidence of any functional similarity and add noise to the already complicated homology data.

#### **4.5: RULE-BASED SEMANTIC HOMOLOGY ANNOTATION**

After completing the PGCO assignments and determining high-quality matches, Gene Homology Ontology terms are assigned to each pair of related genes. The rationale behind first assigning PGCO terms is that the GHO term assignment relies nearly exclusively on concepts such as best hit, reciprocity of hits, and relative quality of hits from a query gene. These concepts are encoded within PGCO assignments, so as a matter of efficiency it is advantageous to pre-compute them.

##### ***Determining positionally conserved genes***

The first step in GHO assignment is determining positionally conserved genes. In order for this process to continue the user must provide facts detailing the chromosome on which each gene is located (*srcfeature*), the base pair start position of that gene (*fmin*), the base pair stop position of that gene (*fmax*), and the strand on which the gene is located. Positional conservation cannot be calculated without location information. As an example, for the *Leishmania major* gene 'LmjF02.0040' the location information is structured as such:

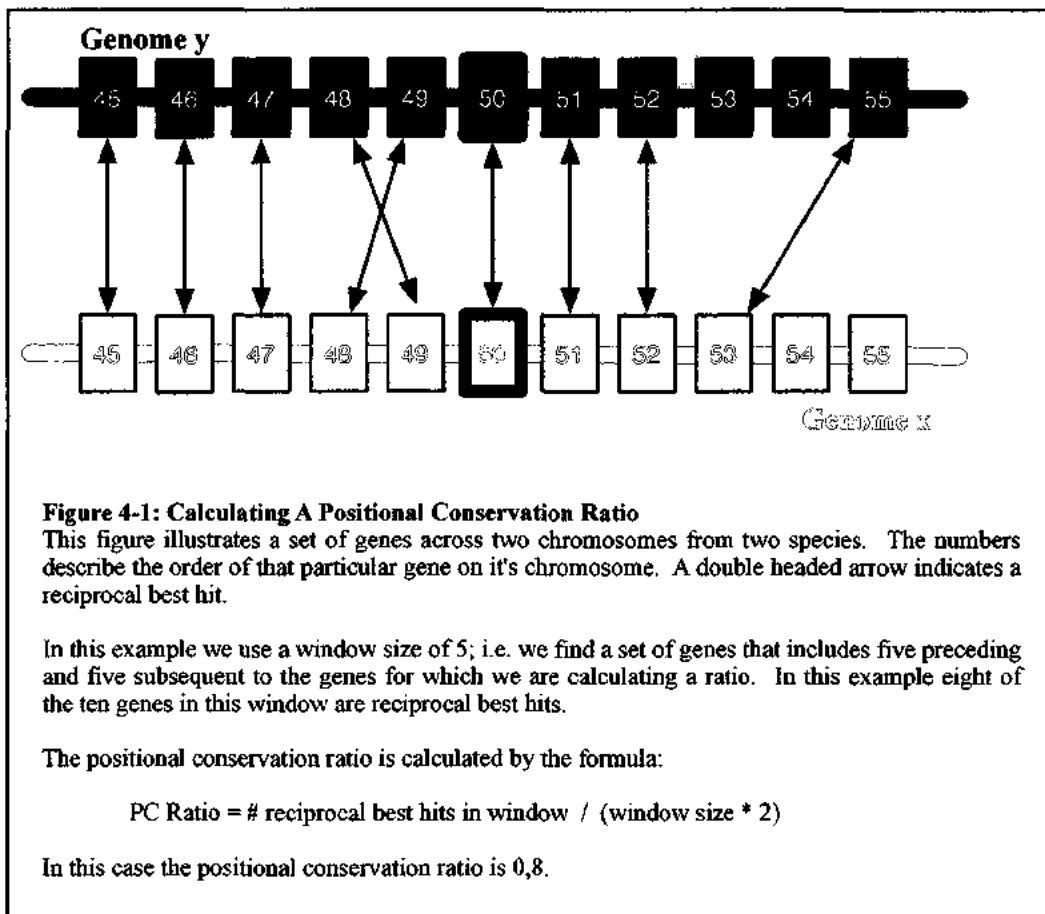
```
prop('LmjF02.0040', fmax, 13538).
prop('LmjF02.0040', fmin, 11679).
prop('LmjF02.0040', srcfeature, fid_8421).
```

Using the information above the rule-base calculates gene order for each gene along each chromosome. For example, the furthest 5-prime occurring gene is assigned position one, the next occurring gene is assigned position two, *etc.* The gene order facts are structured as such:

```
prop('LmjF02.0040', gene_order, 4).
```

The 'gene order' information calculated can be used to define a window of  $W$  genes surrounding a given gene (**Figure 4-1**). For example, consider the hypothetical *Gene\_y50* from *Genome y*, which has an external reciprocal best hit of *Gene z50* in *z*: employing a window size of five ( $W = 5$ ) provides a set of ten genes (*Set\_y50*) that are within five positions on either side of *Gene\_y50* and a set of ten genes (*Set\_z50*) that are on either side of *Gene\_b50*.

Next, the system determines the number of genes in *Set\_y50* that are external reciprocal best hits of genes in *Set\_z50*. That number divided by the size of *Set\_y50* (which is in this case twenty) constitutes a *positional conservation ratio*. Two genes are positionally conserved if their *positional conservation ratio* is above a threshold value. The results presented in **Chapter 7 & 8** employed a threshold value of 0.5.



### ***Determining High Quality Sequence Comparison Matches***

The next step in the GHO assignments is determining which sequence comparison matches have sufficient quality to imply homology; this is the process that we elaborated upon in **Chapter 4.2**. As mentioned earlier, low-quality matches are removed from the analysis because they typically indicate short stretches of distant homology that do not add insight into the comparison of two genomes.

### ***Initial Classification of Homology Relationships***

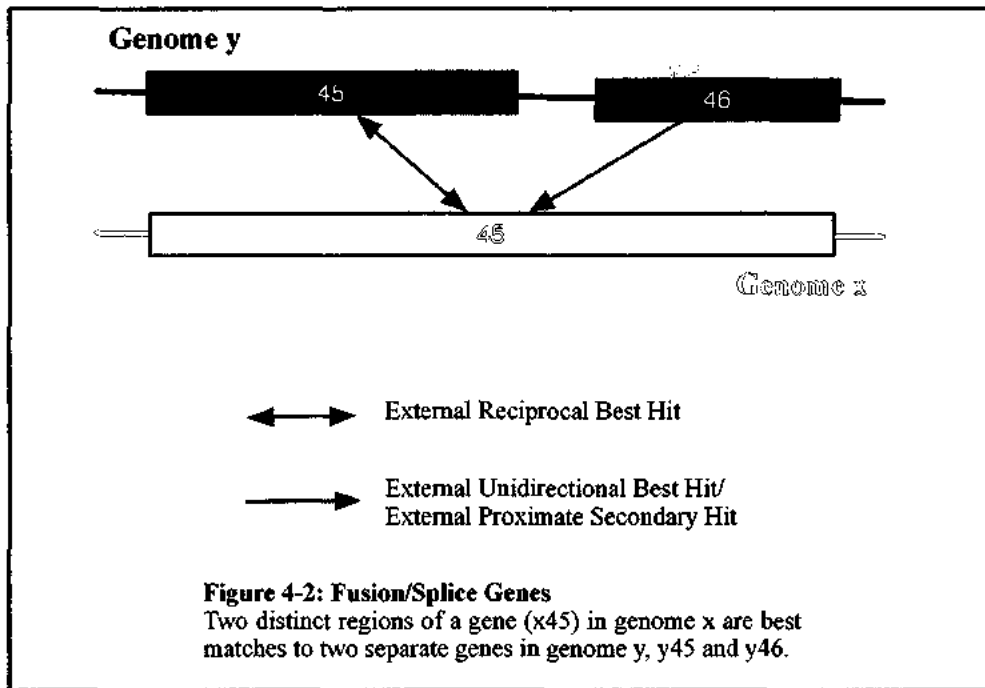
Next, preliminary Gene Homology Ontology term assignments are made using the terms ‘*GHO:external\_homolog*’, ‘*GHO:internal\_homolog*’, ‘*GHO:closest\_homolog*’, ‘*GHO:closest\_internal\_homolog*’ and ‘*GHO:closest\_external\_homolog*’. The rules for these assignments are straightforward and essentially involve a mapping of PGCO terms. For instance:

```
prop(Gene_A1, external_match, Gene_B1) :-
    edge(Match_ID1, Gene_A1, Gene_B1),
    prop(Match_ID1, pgco_term, external_match),
    edge(Match_ID2, Gene_B1, Gene_A1),
    prop(Match_ID2, pgco_term, external_match).
```

The reason for the mapping from a PGCO term to a high-level GHO term is PGCO terms apply to matches, GHO terms apply to pairs of genes. The mapping from a term that applies to a match to a term that applies to a pair of genes is straightforward because a match inherently consists of a pair of genes; nonetheless, this requires an explicit step.

### ***Classifying Novel Genes***

Novel genes are genes that have no homologs either inside or outside of their group (“group” implies species in most cases); in other words, these are genes that have no sequence comparison matches that meet or exceed our criteria for a ‘quality match’. There is no formal GHO term for novel genes, as they are genes lacking homology.



### ***Classifying Fusion and Splice Genes***

***(GHO:fusion, GHO:splice)***

The issue of gene fusions (or splices) is particularly difficult to deal with in comparative genomics studies, particularly any study that involves clustering of genes because homology to a hybrid (fusion) gene can cause two distinct orthologous groups to overlap <sup>57</sup>; **Figure 4-2** shows an example of this phenomena.

Identification of fusion/splice events is the first step in our *Semantic homology annotation* process, because the issues that fusion and splice genes can cause in ortholog assignments across genomes. The strategy for finding fusion/splice events is:

1. Find instances where two or more genes (the splice genes) in one genome are orthologous to a single fusion gene in another the genome.
2. Determine the longest splice gene.
3. Specify that the longest splice gene is the ortholog of the fusion gene.



4. Specify that the remaining splice genes (usually just one), is fused with the longest splice gene.

This strategy is advantageous in that it maintains the traditional one-to-one ortholog relationship; however, it also makes clear that one genome of the ortholog relationship contains two genes that are fused in the other genome. Explicitly describing this relationship type affords researchers using our system an easy way of querying (and thereby detecting) fusion/splice occurrences. Furthermore, in the *Semantic gene grouping*, *Logical gene group querying* steps, researchers can specify how to handle fusion/splice occurrences, avoiding the pitfalls of algorithms that decide in advance how to handle this scenario.

As an example we illustrate the detection and representation strategy of two genes, *GENE\_A1* and *GENE\_A2*, in *GENOME\_A* that are orthologous to a single fusion gene *GENE\_B1* in *GENOME\_B*. In this hypothetical example *GENE\_A1* is more similar to *GENE\_B1* than *GENE\_A2* is to *GENE\_B1*.

We detect the above scenario by finding a set of genes [FUSION, SPLICE\_1, SPLICE\_2] such that FUSION is the 'PGCO:external\_reciprocal\_best\_hit' of SPLICE\_1 and the 'PGCO:external\_unidirectional\_best\_hit' of SPLICE\_2, and SPLICE\_1 and SPLICE\_2 have no homology relationship. The rule for this scenario is:

```
prop(GENE_X, external_fusion_of, GENE_Y):-
    prop(GENE_X, has_closest_external_homolog, GENE_Y),
    prop(GENE_Y, has_closest_external_homolog, GENE_X),
    prop(GENE_X, has_external_homolog, GENE_Z),
    not(prop(GENE_Y, has_internal_homolog, GENE_Z)).
```

That scenario is represented in this manner:

```
prop(GENE_A1, has_ortholog, GENE_B1).
```

```
prop(GENE_A1, fused_with, GENE_A2).

prop(GENE_A1, external_splice_of, GENE_B1).
prop(GENE_A2, external_splice_of, GENE_B1).

prop(GENE_B1, external_fusion_of, GENE_A1).
prop(GENE_B1, external_fusion_of, GENE_A2).
```

### **Classifying Orthologs**

#### **(GHO:ortholog)**

The next step in the *Semantic Homology Annotation* process is determining orthologous genes. This step is accomplished by finding reciprocal best hits. The rule is:

```
prop(GENE_X, has_ortholog, GENE_Y) :-
  prop(GENE_X, has_closest_external_homolog, GENE_Y),
  prop(GENE_Y, has_closest_external_homolog, GENE_X).
```

This process is called 'seeding orthologs'; the next step is adding inparalogs to those ortholog seeds.

### **Classifying Inparalogs**

#### **(GHO:inparalog)**

After finding the ortholog pairs, the system adds inparalogs to those seeds; an inparalog is a gene which arises from a post-speciation duplication event. Inparalogs have the following properties:

1. An internal homolog (IH) with an ortholog (O) in the external genome.
2. The ortholog (O) is the closest external homolog to the inparalog (IP).
3. Is more similar to its internal homolog (IH) than it is to the ortholog (O) of (IH).

The rule for classifying a gene as an inparalog is:

```
prop(X, has_inparalog, X2) :-
```

```
prop(X, has_ortholog, Y),
prop(X, has_internal_homolog, X2),
prop(X2, has_closest_external_homolog, Y),
not(further_from(X, Y, X2)).
```

### ***Classifying Species-Specific Inparalogs***

#### ***(GHO:species\_specific\_inparalog)***

Species-specific inparalogs are groups of genes that are homologous and have no ortholog in the external genome. These groups of genes presumably arose as duplication events from some species-specific novel gene which evolved post-speciation; alternately, they arose as duplications from a gene that was lost in the external genome. These genes are different from genes that are truly novel and have no homolog internally or externally.

The rule for defining a species specific inparalog is as follows:

```
prop(X, has_species_specific_inparalog, X2):-
prop(X, has_internal_homolog, X2),
not(prop(X, has_external_homolog, _)),
not(prop(X2, has_external_homolog, _)).
```

### ***Classifying Pseudo-Inparalogs***

#### ***(GHO:pseudo\_inparalog)***

Some potential genome-expansions meet most of the standard inparalog definitions, except the species-specific expanded gene (EG) is more closely related to the Ortholog (OG) than the parent gene (PG). This suggests that the EG is not an expansion of the PG, but rather an ortholog of some gene that has been lost in the external genome. The rationale for this assumption is that were EG truly expanded from PG, it is highly unlikely that it would have a higher degree of sequence similarity to OG than PG. From a sequence similarity standpoint the loss of a gene in the external genome is a more likely explanation for this pattern. The rule for finding pseudo-inparalogs is:

```
prop(X, has_inparalog, X2) :-
    prop(X, has_ortholog, Y),
    prop(X, has_internal_homolog, X2),
    prop(X2, has_closest_external_homolog, Y),
    not(further_from(X, Y, X2)).
```

### ***Classifying Outparalogs***

#### ***(GHO:internal\_outparalog, GHO:external\_outparalog)***

Outparalogs are paralogs that branched from each other as a result of a duplication event before the speciation event between the two compared species. This version of the rule-base specifies that an outparalog must be an ortholog as well because the current metric for presence in the ancestral genome is orthologous relationship across the two compared genomes. More precise metrics are available for determining presence in the ancestral genome (detection of possible pseudogene degeneration in one genome, or presence in an earlier diverging outgroup<sup>58</sup>); however, this first version of the system does not employ such methods. The use of more complex means of determining ancestral presence is discussed in the future work section (**Chapter 9**).

The rule for determining external outparalogs is:

```
prop(XA, has_external_outparalog, YB) :-
    prop(XA, has_internal_homolog, XB),
    prop(XA, has_ortholog, YA),
    prop(XB, has_ortholog, YB),
    prop(YA, has_internal_homolog, YB).
```

Essentially the above rule states that for a gene (XA) with an ortholog (YA) and an internal homolog XB, such that XB has an ortholog YB, then XA will have an external outparalog YB.

The rule for determining an internal outparalog is similar:

```
prop(XA, has_internal_outparalog, XB) :-
    prop(XA, has_internal_homolog, XB),
    prop(XA, has_ortholog, YA),
```

```
prop(XB, has_ortholog, YB),  
prop(YA, has_internal_homolog, YB).
```

#### **4.6: RULE BASE CLASSIFICATION CONCLUSION**

This chapter described our novel knowledge representation syntax that is highly suitable for describing genomic data; this format is employed for aggregating heterogeneous data from a number of different sources into a common format. This representation strategy also serves as the input to and output from the various steps described in this work. The flexibility of this knowledge representation schema allows researchers to translate genomic information from any resource into a simple fact-based representation; furthermore, a researcher may employ the results of any type of sequence comparison algorithm as inputs into the overall workflow. This ensures that the pipeline can evolve to accommodate advances in sequencing and sequence comparison technologies.

Later, this chapter described the rule-based *Semantic Homology Annotation* process. This process entails determination of positional gene conservation, selection of high-quality sequence comparison matches, classification of those matches according to the Pairwise Genome Comparison Ontology and finally classification according to the Gene Homology Ontology.

The modular nature of this rule-based process lends itself to refinement and extension. While the Gene Homology Ontology accurately describes the types of relationships that vast majority of genome researchers use in their analyses, there certainly are other concepts associated with homology in which researchers may be interested. For instance, in the future the community may wish to extend the ontology to encapsulate terms associated with syntenic relationship between orthologs. This may be particularly interesting to researchers investigating conservation of operons across bacterial species. Ontologies are, by nature,

amenable to this type of extensions. The rules for ontology classification are similarly extensible; researchers wishing to refine this work can add new rule definitions for new ontology terms.

The next chapter describes the *Semantic gene grouping* strategy, a process that clusters genes based on GHO homology relationships. The result of this process is groups of evolutionarily related genes that have clearly defined relationships. Later, **Chapter 6**, describes methods for querying these gene groups using a rule-based strategy that clearly articulates these similarities and differences; this rule-based querying is known as *Logical Gene Group Querying*. Both of these technologies rely on our knowledge representation schema described earlier in this chapter for the input, output and processing of results.



## **CHAPTER 5: SEMANTIC GENE GROUPING**



## 5.1: OVERVIEW

**Chapter 4** detailed the process of attaching semantic homology information to relationships between pairs of genes. After establishing these pairwise homology relationships, the next step is grouping genes into sets of evolutionarily related genes based on semantic criteria. Furthermore, while the prior chapter discussed comparisons of two genomes, this chapter explores methodologies for combining a series of pairwise comparisons into a multi-genome comparison.

In this dissertation we use the term '*semantic gene grouping*' (SGG) rather than the more traditional term 'clustering'. While the formal definition of 'clustering' is simply separating a set of inputs or observations into subsets, in bioinformatics the term tends to imply certain mathematical, statistical or graph-theory methodologies for distinguishing between sets of genes. Since we are using a novel approach, predicated upon semantics, we have employed a new description.

The *semantic homology annotation* process described in **Chapter 4** described in an unambiguous way how pairs of genes are related. While this is highly useful from a gene-centric point of view, from a genome-centric point of view we need to aggregate these pairwise relationships into larger gene groups to understand trends in gene content differences across multiple genomes. One of the hallmarks of the work we present in this chapter is a lack of rigid grouping definitions. The evolutionary forces that act upon genes do not necessarily lead to groups of genes that conform to the idea of "clusters". We accommodate that fact by describing how related genes are grouped without imposing highly structured

grouping. Furthermore, the rule-based grouping system described in this chapter allows users to alter or expand the grouping definitions based on their own needs.

The overall workflow for grouping a multi-genome comparison is as follows:

1. Performing a series of pairwise comparison across all the genomes
2. Assigning Gene Homolog Ontology (GHO) terms to the relationships from the pairwise comparisons
3. Defining sets of orthologous genes
4. Adding inparalogs to the groups.
5. Querying the groups to determine interesting biological trends.

**Note:** Chapter 4 had inline examples of the rules to clarify the logic behind the GHO annotations, as well as to provide a primer on writing these rules in Prolog. This chapter does not provide the rules inline, as they are more complicated and require a more in-depth understanding of rule-based programming.

## **5.2: MOVING FROM PAIRWISE ANALYSIS TO MULTI-GENOME ANALYSIS**

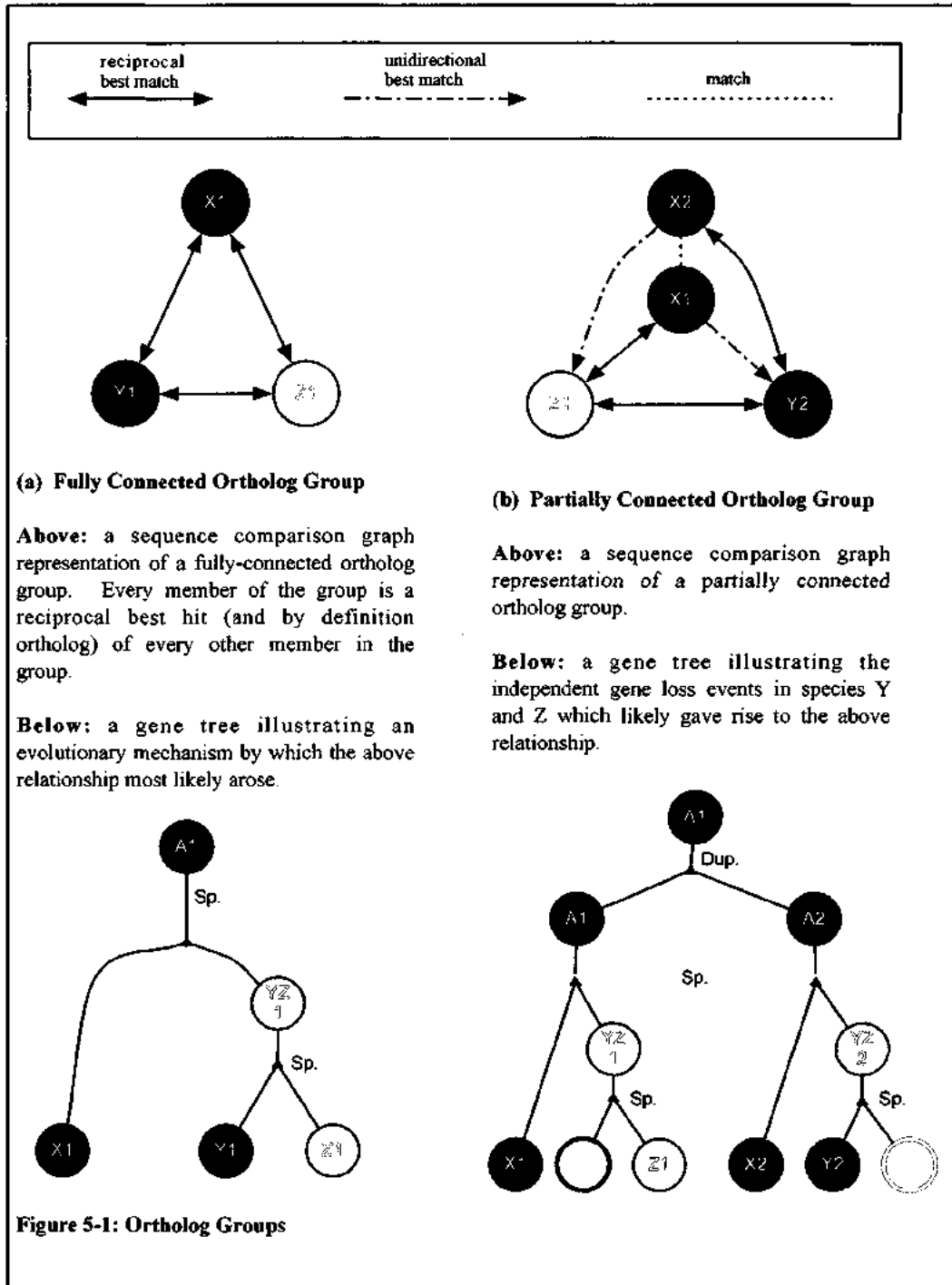
**Chapter 4** discussed gene homology from a pairwise genome perspective, in other words comparing some genome to some other genome. While most current sequence-based comparative studies are based on pairwise similarity as elucidated by BLAST or some other comparison algorithm, the power in comparative genomics arises from comparing multiple genomes.

The methodology presented here for performing multi-way genome comparisons involves aggregating the results of all possible pairwise comparisons. In other words a multi-way comparison of genomes *A*, *B* & *C* entails the three pairwise comparisons (*A* to *C*, *A* to *B* and *B* to *C*), each performed as described in **Chapter 4**.

The pairwise comparison of some number of genomes ( $N$ ) is an  $N^2$  operation since each individual genome is compared to every other genome. The strategy described in this work is useful for such comparisons because including an additional genome involves performing the individual pairwise comparisons of that genome to the existing genomes, as opposed to recalculating all the sequence comparison results. The semantic gene grouping, however, must be performed each time a genome is added to the comparison, and the complexity of the grouping will increase proportionally to the square of the number of genes in the comparison. Thus far, we have performed semantic gene groupings (as described in this chapter) of up to 32,000 genes in approximately 3-4 minutes on a desktop computer with 4 gigabytes of memory. Note, that this does not include generating the sequence comparisons, which can be time-consuming and is dependent on a host of factors. We have not yet tested the system on larger numbers of genes and cannot comment at this point on how well the system scales in terms of computational time or memory use.

## 5.2: CREATING ORTHOLOG GROUPS

The first step in the grouping process is creating orthologous groups. Establishing pairwise orthology based on a sequence comparison is relatively straightforward (reciprocal best matches across genomes); however determining groups of orthologs across multiple genomes involves establishing some concrete definition of orthologous groups. A simple definition (one used by the COG/KOG project <sup>12,11</sup>) defines an orthologous group as a group of genes in which each group member is an ortholog of every other group member **Figure 5-1 (a)**. This work refers to these types of ortholog groups as *fully-connected ortholog groups* (FCOG). While FCOGs represent an especially succinct definition of ortholog, particularly from the



point-of-view of asserting functional equivalence across genomes, ancestral gene duplication and subsequent loss in one or more species can lead to more complicated orthologous relationships. For example, **Figure 5-1 (b)** illustrates a scenario in which a gene duplication in an ancestor species, followed by independent gene loss in two species has led asymmetric orthology relationships. This work refers to this scenario as a *partially-connected ortholog group* (PCOG).

PCOGs do not necessarily carry the same implication of functional equivalence that FCOGs carry. For instance in **Figure 5-1 (b)** Genome *X* has maintained two copies of the gene (perhaps indicating some sort of neofunctionalization or subfunctionalization) and genomes *Y* and *Z* have lost different copies of that gene; perhaps indicating different selective pressures on those organisms. This asymmetry suggests functional difference between members of this PCOG. By contrast the FCOGs do not carry such asymmetry. Nonetheless, for purposes of efficiency we consider both PCOGs and FCOGS as orthologous groups as they both represent a collection of genes with closely related functions across a group of genomes.

Our knowledge representation schema and rule-based strategy (**Chapter 4**) allows us to flexibly define what constitutes a COG. In this work we've described PCOGS and FCOGS, but there are any number of ways that a researcher might define a COG. For instance, some researchers may have different opinions on grouping fusion and splice genes. The flexibility and descriptiveness that our strategy affords accommodates such varied definitions. This ameliorates a problem with most clustering methodologies; they define "groups" such all groups can only contain the same (or similar) types of relationships. The *semantic gene grouping* strategy operates on a fundamentally different philosophy; it describes gene groups in a flexible manner and then afford researchers a means of querying those groups (**Chapter 6**) in a way that highlights biologically relevant patterns.

The assumption of equivalence between PCOGs and FCOGs is not always accurate, however it illustrates an important facet of the rule-based semantic approach for gene grouping. The rule-base contains a rule to define PGOs and a rule to define FCOGs; we can easily query our gene groupings to differentiate between the two and describe the asymmetric relationships in an FCOG, this allows researchers to investigate orthologous groups that may carry some indication of functional nonequivalence. Such a methodology serves a researcher interested in a fine-grain analysis of functional differences between a group of species. On the other hand, a researcher performing a broad, summary-level analysis of genomic differences might prefer to consider all ortholog groups (whether FCOGs or PCOGs) equivalently. This situation illustrates one of the central thrusts of this work: instead of assuming how best to perform a grouping analysis, this work describe the data as facts that researchers can interrogate using custom rules.

### 5.3: ADDING INPARALOGS TO ORTHOLOG GROUPS

The next step in the *semantic gene grouping* process is assigning inparalogs to the gene groups. When comparing two genomes, determining the inparalog(s) of a given ortholog group is a relatively well-defined process, as illustrated in **Chapter 4.5**. However, multi-way comparisons carry some peculiarities regarding assigning inparalogs.

The first peculiarity is illustrated in **Figure 5-1 (b)**: in the gene group containing genes X1, X2, Z1 and Y2, gene X2 has an orthologous relationship with gene Y2 and an apparent pseudo-inparalogous relationship with Z1; similarly gene X1 has an orthologous relationship with gene Z1 and a pseudo-inparalogous relationship with gene Y2.

In practice this does not affect the results of our gene-grouping rules, as all of these genes are part of the same partially connected orthologous cluster as defined in **Section 5.2**. Nonetheless, this example illustrates a gene that has different relationships to different

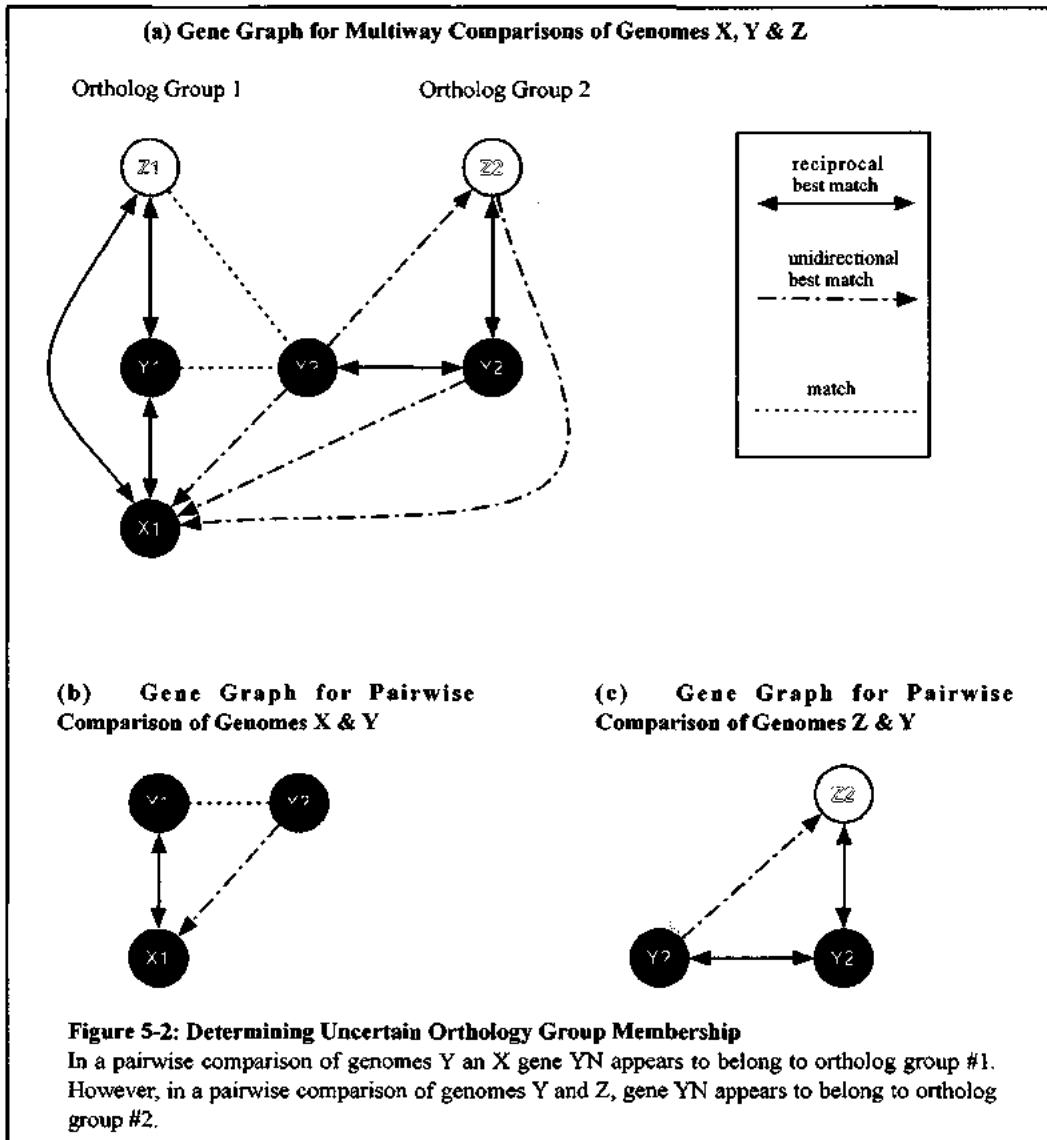
members of an ortholog group. This points out a critical issue when dealing with grouping genes in multi-genome comparisons. Failing to account for the multiple relationships that a gene has within an ortholog group may result in redundant groupings. For instance, the following fact asserts that a given gene belongs to a gene group:

```
prop(X2, member_of, some_gene_group).
```

When building the orthologous group such facts are created to specify membership of genes to gene groups. However, were a rule specifying that all inparalogs (or pseudo-inparalogs) of any gene in the ortholog group are to be joined to the gene group, the above fact would be duplicated, since gene X2 is also a pseudo-inparalog to gene Z1.

This scenario illustrates the general point that complex (and seemingly contradictory) relationships can exist within FCOGs. Creating accurate FCOGs and posing appropriate questions of those FCOGs (describe later in this chapter) requires an understanding of these complexities.

Adding inparalogs to ortholog groups can also require careful analysis in some cases, such as an example is illustrated in **Figure 5-2**. In this case two hypothetical ortholog groups are named *Ortholog Group #1* and *Ortholog Group #2*. Pairwise comparison of Genomes X and Y indicate that gene YN is an inparalog belonging to Ortholog Group #1, since its closest external homolog is gene X1. On the other hand, pairwise comparison of



genomes Y and Z, indicates that gene YN has gene Z2 as the closest external homolog and therefore gene YN is an inparalog of Ortholog Group #2.

The above problem is solved by joining the gene YN to whichever ortholog group contains the internal homolog to which it has a higher scoring sequence comparison match. Were the sequence comparison match between YN and Y1 stronger than the match between YN and



Y2, YN would be joined to Ortholog Group #1. Were the opposite true, YN would be joined to Ortholog Group #2.

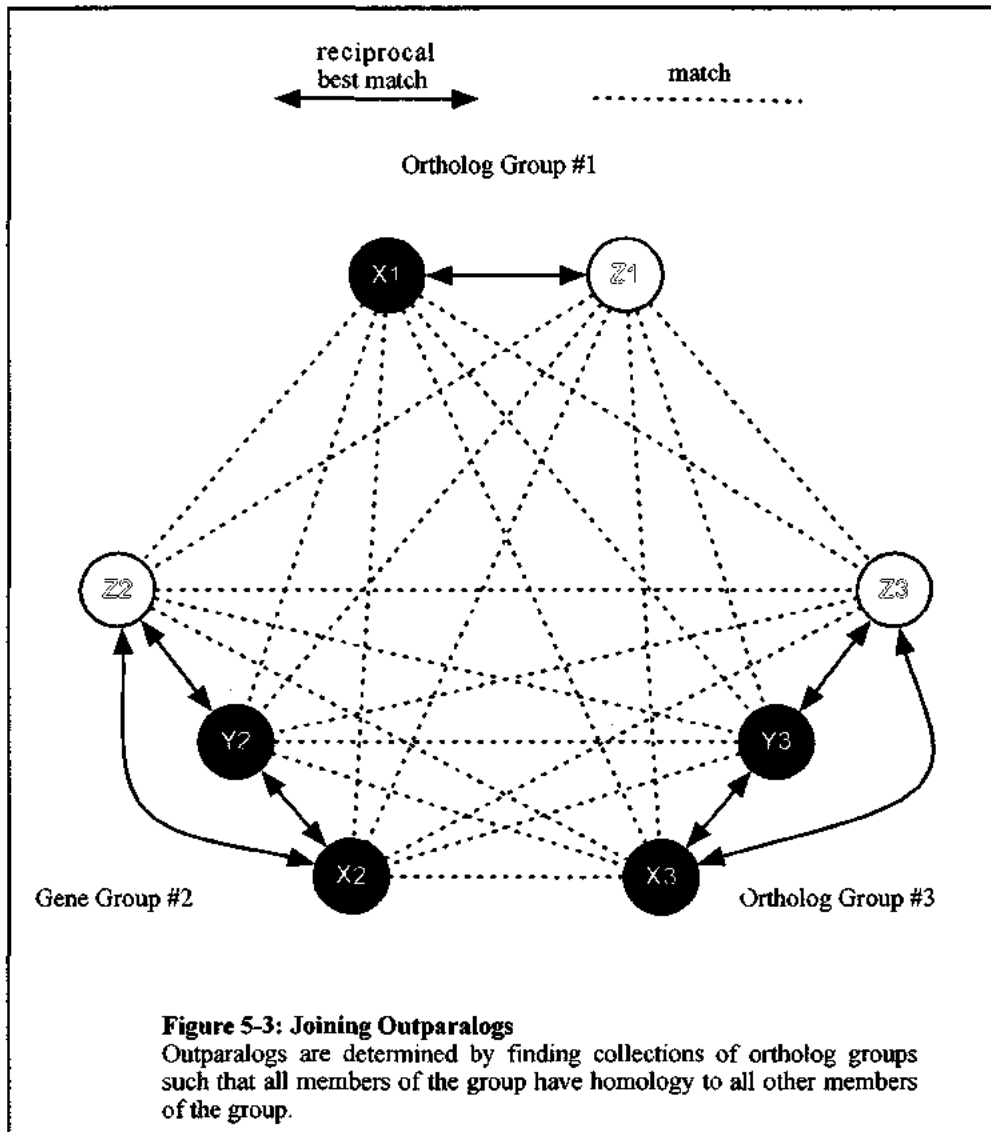
This “higher scoring” strategy for resolving such a scenario illustrates another benefit of the semantic rule-based approach. In this example a rule finds inparalogs that have potentially arisen from more than one ortholog groups. Using such a rule allows researchers to quickly query for gene groups that are worth further investigation. As with the ortholog group example in **Section 5.2**, this scenario illustrates the benefit of asserting homology relationships as facts that are queried by rules. This is in contrast to most clustering and grouping methodologies that make “set-in-stone” decisions as to group membership and mask these types of interesting or unusual gene relationships.

#### **5.4: JOINING OUTPARALOGS**

The last step in the semantic gene grouping process is connecting outparalog, paralogs that arose prior to the speciation event that separates the compared species. This step elucidates large gene families that have multiple copies that were inherited from the ancestor species.

This involves means joining ortholog groups that have homology to each other. This involves finding a number of ortholog groups such that every member of any one of those ortholog groups is a homolog to every member of the other ortholog groups. As shown in **Figure 5-3**, the joined ortholog groups do not necessarily have to have a member from each of the genomes.

In practice most researchers are more interested in ortholog groups and their subsequent post-speciation inparalogous expansion than they are in grouping together outparalogs. The reason for this preference is twofold: functional equivalence of ortholog groups and post-



speciation trends in gene loss and expansion. As to the first point, ortholog groups imply functional equivalence, and therefore provide a sort of accounting methodology for determining which functional roles are present or absent in a particular species. As to the second point, patterns of inparalog expansion give insight as to “recent” evolutionary trends in an organism and illustrate how that organism has adapted to its particular niche. By contrast groups of outparalogs are typically less interesting. Outparalogs do not carry the same connotation of functional equivalence that orthologs carry<sup>24</sup>, and furthermore their pre-

speciation divergence makes them less useful for contrasting recent evolutionary trends in the compared species.

This issue again illustrates the benefit of our semantic rule-based system. Instead of joining outparalogs or separating outparalogs this strategy specifies that genes are outparalogs. This provides a groundwork by which users of our system can query genes according to their needs and interests.

### **5.5: SEMANTIC GENE GROUPING SUMMARY**

This chapter presented a strategy for grouping evolutionarily related genes. In contrast to standard clustering methodologies we do not seek to categorize genes based on strict definitions of what constitutes a “group”. Instead, we chose a philosophy of using biological reality as a guide and describing how related genes fall into discernible sets.

The knowledge representation schema and the gene homology annotations described in **Chapter 4** serve as the foundation for the work described in this chapter. The technologies discussed in **Chapter 4** provided a clear and structured descriptions of how genes are related to their homologs. Those relationships allowed for the creation of groups of genes, and accommodated multiple relationship types in those groups. **Chapter 6** describes how to query the groups for interesting biological patterns.

Gene homology is a complex and often convoluted issue that does not lend itself well to concretely defined groupings of genes. The primary motivation of the work in this chapter was to accommodate that flexibility; the next chapter will discuss ways of delving into that complexity and finding patterns that provide insight into the evolutionary forces that shape the compared genomes.

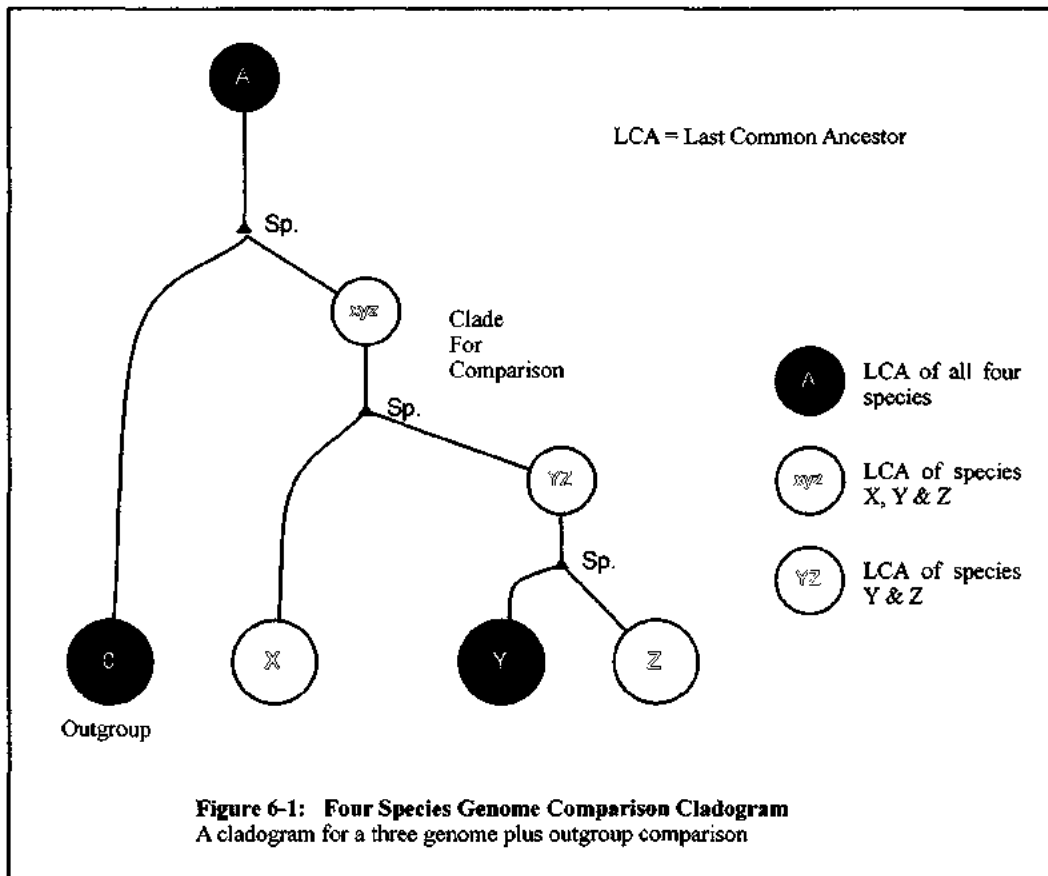
## **CHAPTER 6: LOGICAL GENE GROUP QUERYING**

## 6.1: INTRODUCTION TO LOGICAL GENE GROUP QUERYING

This chapter is devoted to a discussion of and examples of *logical gene group querying*, a rule-based system for finding biologically interesting patterns in gene groups. The work described in this chapter represents the culmination of the technologies described in **Chapters 3 - 5**. Those chapters described novel and accurate means of describing existing knowledge; this chapter discusses methodologies for generating new knowledge and furthering the understanding of evolutionary patterns across the compared species.

Most clustering technologies output a list of genes that share similarity as determined by some algorithm. These clustering tools make *a priori* assumptions as to what constitutes a “cluster” or “group” and do not add any semantic information specifying why the a gene is in a given cluster or how the genes in the cluster are related. By contrast, the *semantic gene grouping* technology described in the prior chapter explicitly describes the relationships between genes and allows for further rule-based querying.

The explicit specification of relationships between grouped genes allows for querying for particular patterns of gene expansion, gene gain and gene loss. Some of these patterns are quite obvious or simple. For instance, the comparative genomics study on *Leishmania spp.* described in **Chapter 7** yielded many gene groups that fit typical models of gene expansion, gain or loss. However, after posing queries for expected patterns, a number of gene groups were left unaccounted for, pointing to dynamics that are not particularly intuitive or well-understood. Analysis of these groups found numerous instances of gene expansion followed by differential gene loss. This dynamic paints a fundamentally more complicated picture of orthology than is typically imagined. Though the functional implications of this finding are



not yet known, this example shows how our gene grouping and querying strategy can uncover patterns of evolution that would go ignored by traditional clustering methodologies. This issue is discussed in more depth in **Section 7.7**.

**Section 6.2** discusses the basic principles behind *logical gene group queries* and **Section 6.3** explains the structure of a simple query. **Sections 6.4 & 6.5** provide a conceptual discussion of some of the *logical gene queries* that are posed against actual comparative genomics data in **Chapters 7 & 8**

The remainder of this chapter will refer to a hypothetical four-genome comparison; the comparison consists of three closely related species (species X, Y & Z), plus an outgroup (O). The cladogram for the four organisms is shown in **Figure 6-1**. The use of an outgroup that is

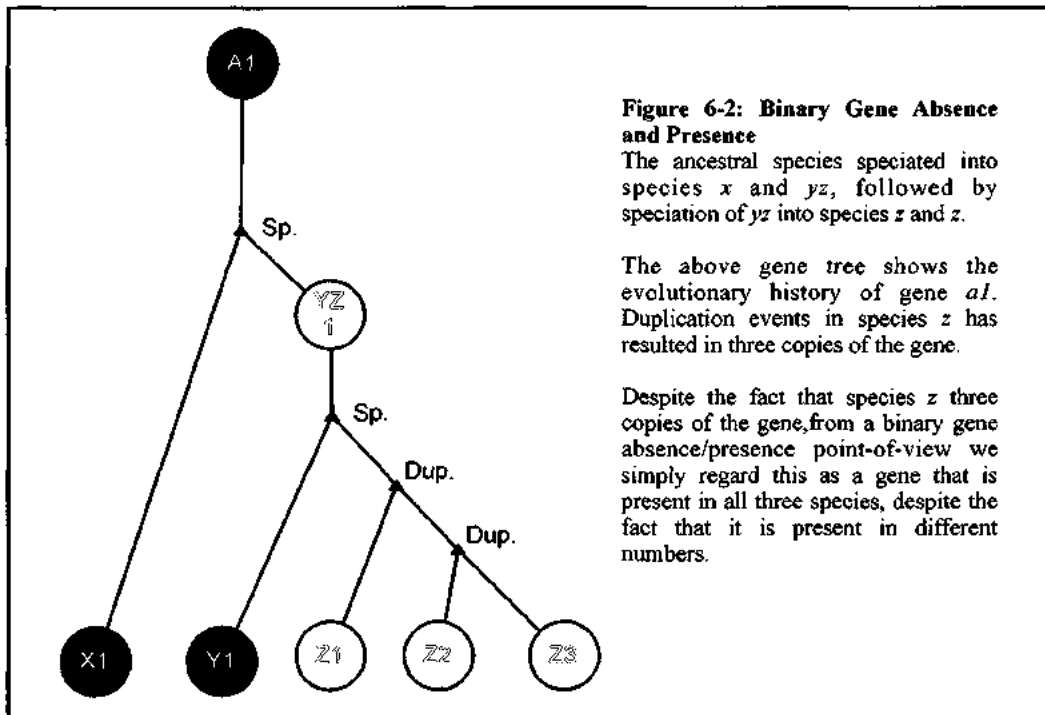
relatively closely related to a group of compared species, but further evolutionarily from any of the species than they are to each other, is a common design for understanding patterns of evolution. Sections 6.4 & 6.5 will further elaborate upon this strategy.

## 6.2: AN EXAMPLE OF A LOGICAL GENE GROUP QUERY

A simple query on the *semantic gene groups* is that of determining whether a particular genome is represented in a group of orthologs. An orthologous group represents a group of genes that encode for a proteins which perform the same (or similar) functional roles across genomes. The absence of a genome from a group of orthologs indicates absence (or at least significant divergence) of that functional role from that genome. Conversely, presence of a gene from a given genome indicates presence of a functional role. This idea is hereafter referred to as *gene presence* or *gene absence*.

The notion of *gene presence* bears some further clarification; consider **Figure 6-2**. This graph shows a gene that is present in all four compared species, but has two inparalogs in one species. These two inparalogs are in some senses *absent* in the three species for which the gene has not expanded; however, by the definitions used in this work, this scenario represents *presence* of the gene with a species-specific expansion. In other words, despite clear differences, the gene is *present* in all four species. However, other reasonable interpretations of this scenario certainly exist and can easily be encoded using the rule-based system.

Assessment of gene expansion is also an area of importance in *logical gene group querying*; gene expansion is interesting both from a standpoint of evolution of novel gene function<sup>59,60</sup>, host-pathogen defense mechanisms<sup>61</sup>, as well as species-specific adaptation in pathogens<sup>48</sup>.



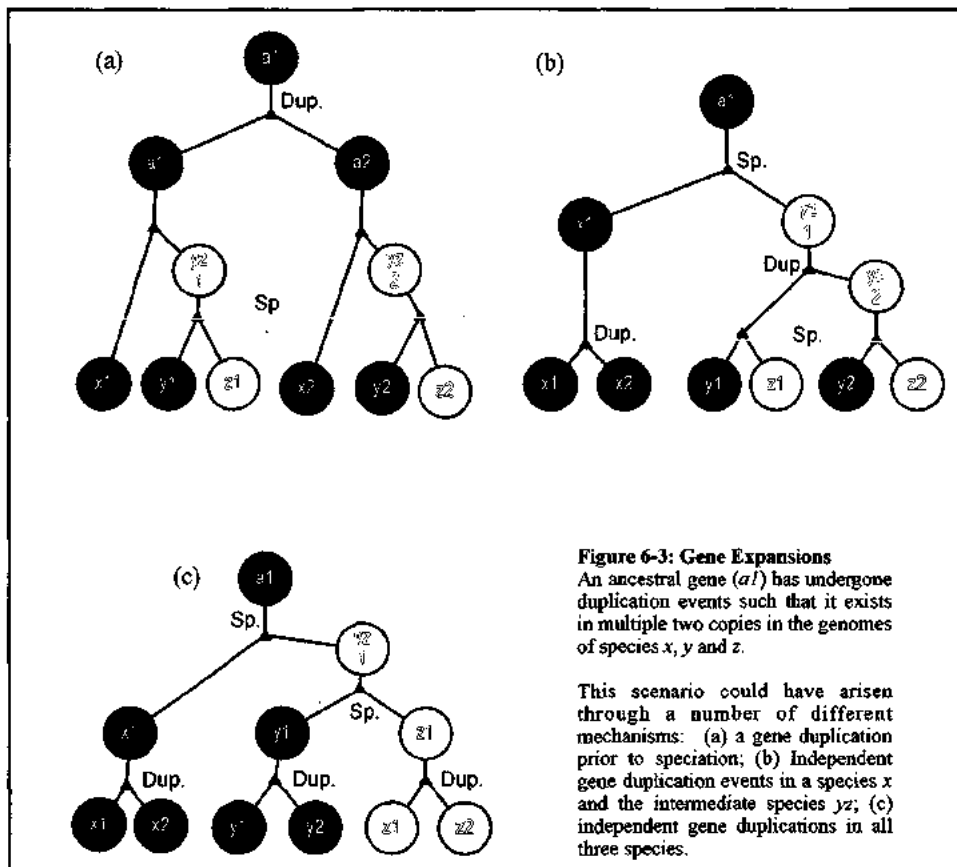
Gene expansion is defined in this work as the presence of multiple copies of the same gene in a given genome. For any given genome comparison this can result from outparalog expansion in the last common ancestor, by expansion in an intermediate ancestral species, or inparalogous expansion post-speciation of the compared genomes. These scenarios are illustrated in **Figure 6-3**; this figure is referred to again in **Section 6.6**, which discusses how our *semantic homology annotation* differentiates between these possibilities.

The rule-base has several recursive rules that generate gene profiles for multiway genomes comparison. For example, the recursive profiling rules may return the following gene presence/absence profile for the four-way (three genomes plus an outgroup) comparison shown in **Figure 6-1**:

```
> profile(Outgroup, GenomeX, GenomeY, GenomeZ, PR).
PR = [0, 1, 1, 1]
```

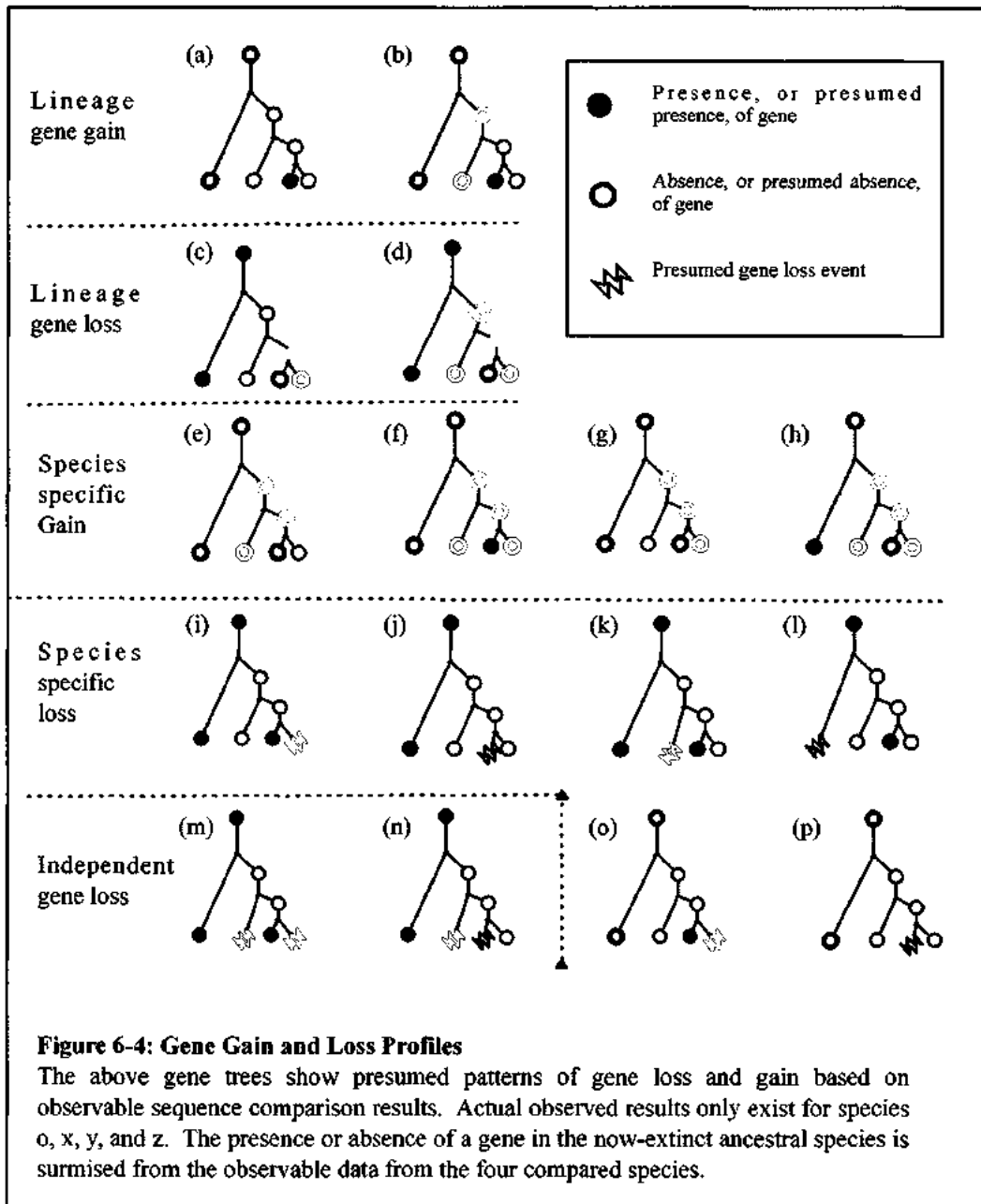


This profile indicates absence of the gene in the outgroup and presence of the gene in the three compared genomes. This profile points to the scenario in which the gene is novel to: *GenomeX*, *GenomeY* and *GenomeZ*. The most likely explanation is the gene tree shown in **Figure 6-4 (a)**, but alternatively could be a result of the scenario shown in **Figure 6-4 (l)**. Further investigation (such as comparisons to another, more distant outgroup) is needed to decide between those two scenarios.



### 6.3: POSING A LOGICAL GENE GROUP QUERY

Use of the utility rules (such as the profiling rule above) is the most efficient means for assessing gene absence and presence; nonetheless, a user can pose simple queries to their semantic gene groups. This section discusses the construction of some basic *logical gene group queries* to provide a simple example of how rule-based queries operate.



**Section 5.5** discussed the data representation syntax for joining a gene to a gene group:

```
prop(Gene_X1, member_of, Group1).
prop(Group1, type, Semantic_gene_cluster).
```

Further more, as discussed in **Section 4.2**, the genome to which a gene belongs is described:

```
prop(Gene_X1, member_of, GenomeX).
prop(GenomeX, type, genome).
```

Users wishing to implement their own rule-based queries without learning much Prolog could pose a relatively simple query such as:

```
genome_present_in_gene_group(Genome, Gene_Group) :-
    prop(Gene1, member_of, Gene_Group),
    prop(Gene1, member_of, Genome).

gene_absent_in_outgroup(Gene_group) :-
    prop(Gene_group, type, semantic_gene_cluster),
    genome_present_in_gene_group(genomeW, Gene_Group),
    genome_present_in_gene_group(genomeY, Gene_Group),
    genome_present_in_gene_group(genomeZ, Gene_Group),
    not(genome_present_in_gene_group(genomeX, Gene_Group)).
```

## 6.4: GENE PRESENCE OR ABSENCE RULES

This section discusses rules that cover the presence or absence of genes, using the definition of *gene presence* outlined in **Section 6.1**. For a multi-way genome comparison, which includes an outgroup, the following types of assessments can be made regarding gene presence or absence:

1. Instances of gene gain in a clade or lineage
2. Loss of gene in a clade or lineage
3. Species-specific gene gain

4. Species-specific gene loss
5. Independent gene gain (likely indicative of poor genome annotation)
6. Independent gene loss in more than one species.

In infectious disease research, assessing patterns of gene loss and gain can provide insights into the underlying genomic causes behind clinical manifestations or methods of pathogenicity of disease causing parasites.

Species-specific presence of certain genes can provide insights into unique behaviors of a specific species. These species-specific genes can serve to provide the pathogen with the ability to survive in certain hosts, vectors or geographic locales. Furthermore, these genes potentially encode for host-pathogen interactions that could lead to certain disease manifestations. Such species-specific genes are also likely to be highly divergent from any host genes and serve as attractive targets for drug development.

Species-specific gene loss can indicate that the environment or host that a species inhabits has rendered a particular gene unnecessary. This is a powerful piece of evidence in comparative genomics because it illustrates a scenario in which a gene that is critical in a group of species and their ancestors is not needed in a particular species. Species-specific gene loss events show indicate that an external factor is uniquely acting on a particular species; elucidating these unique factors can provide a greater understanding as to the lost gene's role in the life-cycle of the other organisms in which it is present. As with species-specific gene gains, species-specific loss may provide insights as to unique clinical manifestations of a particular parasite.

On a wider level, lineage-specific gene gain or loss can address how broad categories of species differ. Furthermore, particular lineages typically possess features (such as geographic

distribution, disease manifestation, *etc.*) that are more similar within the lineage than in outside species. Lineage-specific gene content characteristics can begin to explain why such differences occur.

Independent loss of a particular gene by two species or lineages is interesting because the co-occurrence of gene loss events indicates that two lineages are separately facing some selective pressure relative to the lineages that did not lose the gene. Again, this knowledge may provide insights into the life-cycle or disease manifestation of the organism.

The general goal behind the *logical gene group query* rules is finding patterns that indicate types of gene gain or loss events, such as the ones described above. The remainder of **Section 6.4** discusses the types of patterns that the rules can find. These patterns manifest themselves as a profile of binary presence or absence of a gene; for instance, the pattern of evolution shown in **Figure 6-4 (a)** manifests itself as a binary presence/absence profile of [0, 1, 1, 1]; similarly the pattern of evolution shown in **Figure 6-4 (b)** manifests itself as a profile of [0, 0, 1, 1]. While this exploration is limited to a comparison of an outgroup plus three clade member genomes, the principles discussed generalize to smaller or larger comparisons.

#### ***Instances of gene gain in a clade or lineage***

**Figure 6-4 (a, b)** shows patterns that indicate gene gain in a particular lineage. **Panel (a)** shows a gene that is present in the comparison clade and absent in the outgroup. This indicates either a gain of that gene in the last common ancestor of the comparison clade (species JYZ) or a loss in the outgroup. Scenario (a) is indistinguishable from scenario (l) in the comparison as it currently exists. One possible way of discerning between the two scenarios is searching for the gene in a somewhat distantly related species using sequence comparison to a large gene database, such as the *GenBank non-redundant database*<sup>62</sup>; the presence of the gene in a distantly related species would indicate that the gene was indeed

lost in the outgroup, conversely, absence of the gene in other species would suggest the gain of that gene in the comparison clade. Another potential means of discerning between loss in the outgroup and gain in the clade is determining if remnants of a degenerated gene, lacking coding potential, (*i.e.* pseudogene<sup>23</sup>) is present in the outgroup; such presence would indicate a gene loss event in the outgroup.

**Panel (b)** shows a gene that is present in the species Y and Z, and therefore most likely evolved in the ancestral species YZ. This gene's absence in the outgroup suggests (though doesn't guarantee) that the gene was gained in YZ, as opposed to lost in X.

#### ***Species-specific gene gain***

**Figure 6-4 (e-h)** illustrates a species-specific gene gain. **Panels (e-g)** show a gene gain in an individual species of the comparison clade, and **panel (h)** shows a gain in the outgroup. The scenario shown for **(h)** manifests itself identically in the sequence comparisons results to the scenario show in **(d)**; further analysis is required to differentiate between gain of the gene in the outgroup or a loss of the gene early in the evolution of the comparison clade. These two possibilities can be distinguished by comparison to an additional, more distant outgroup, or by searching for a pseudogene, as described above.

#### ***Loss of gene in a clade or lineage***

**Figure 6-4 (c-d)** shows patterns indicative of gene loss in the entire clade relative to the outgroup or within a particular lineage of the outgroup. In **panel (c)**, the absence of the gene in Y and Z, coupled with the presence of the gene in the outgroup and X indicates that the gene was most likely lost in the YZ ancestor species.

The complete absence of the gene in the comparison clade, show in **panel (d)** indicates that the gene was lost early in the evolution of the clade, perhaps in ancestral species XYZ. As

mentioned earlier, further analysis is needed to differentiate this pattern from the pattern indicating species specific gain in the outgroup, **panel (h)**.

### ***Species-specific gene loss***

**Figure 6-4 (c-d)** shows patterns indicative of species-specific gene loss events. As discussed above, the species-specific loss of a gene in the outgroup **panel (l)** is not discernible from the gain of a gene in the last common ancestor of the comparison clade (species XYZ), shown in **panel (a)**, without further analysis.

### ***Independent gene loss***

**Figure 6-4 (m-p)** show patterns indicative of independent gene loss, in other words loss of a particular gene in multiple species as a result of multiple gene loss events. **Panel (m)** suggests that the gene loss events were independent in species X and species Z because both the outgroup and species Y contain the gene. This indicates that the last common ancestor of the comparison clade (species XYZ) had the gene, and that the common ancestor of Y and Z (species YZ) had the gene. This means that a gene loss event must have occurred in species Z, post-speciation of YZ into Y and Z, and that an independent gene loss event occurred in species X post-speciation of XYZ into X and YZ.

Similar logic suggests that the gene loss events shown in panels **(n-p)** are examples of multiple independent events.

## **6.5: GENE EXPANSION RULES**

This work classifies genome expansions according to the following categories:

1. Lineage-specific expansions
2. Species specific expansions
3. Independent gene expansions

#### 4. Ubiquitous gene expansion

As with the prior section on gene gain and gene loss, expansion events can furnish insights into how a particular species has adapted to its geographic, vector, or host environments; furthermore they provide insights into how that species causes disease. As with gene loss and gene gain, expansion of genes can be assessed from the lens of species-specific events, lineage-specific events, and independent events.

In reality, gene expansion often involves a complex mixture of gene duplication, coupled with gene loss <sup>48</sup>. Often a gene will expand independently in multiple genomes, but to different extents (e.g., a gene might have expanded to two inparalogs in species X and seven inparalogs in species Y). Furthermore, complete understanding of the extent of expansion of a given gene family often involves analysis of both outparalog and inparalog expansion.

The semantic rule-based system is well-suited for handling complicated gene expansion scenarios; however, for clarity this chapter presents a generalizable model of gene expansion, referred to as binary expansion. For every gene in a multi-way genome comparison the rule-base determines if it has expanded relative to the last common ancestor. This model treats any number of expansions as a single expansion event; in other words a gene that has undergone one inparalog duplication event is treated as “expanded” and similarly a gene family that has undergone seven gene duplication events is also treated as “expanded”. **Chapters 6 & 7** present actual sequence comparison data and discuss gene expansion in more depth.

##### ***Lineage-specific expansions***

**Figure 6-5** illustrates expansion of a gene in the entire comparison clade, or a particular lineage within that clade. In **panel (a)** an expansion event in the ancestral species XYZ has



resulted in multiple copies of a gene that have remained viable in species X, species Y and species Z; similarly **panel (b)** illustrates an expansion event in ancestral species YZ that has resulted in multiple copies of a gene that has remained viable.

#### ***Species specific expansions***

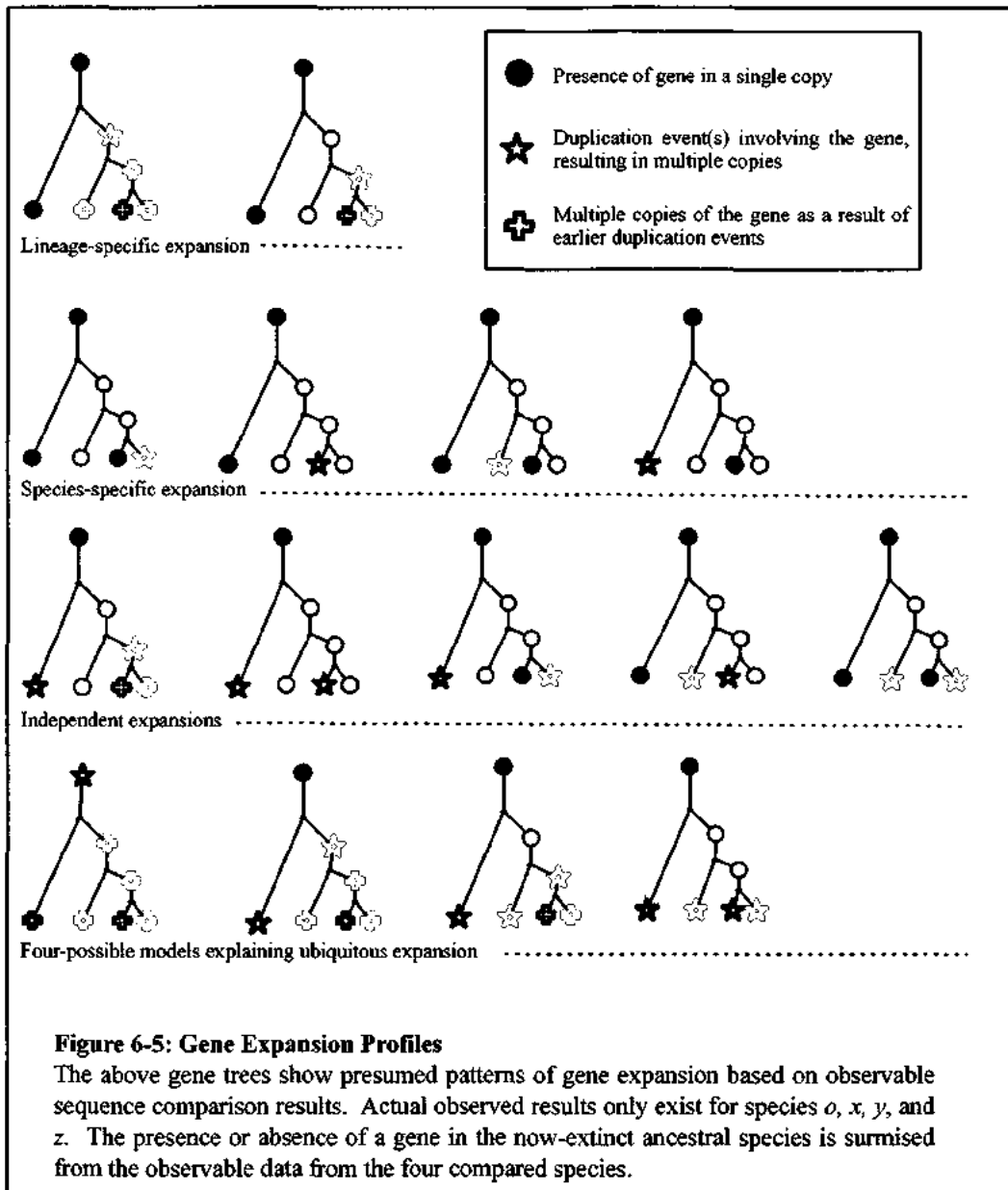
**Figure 6-5 (c-f)** shows gene expansions that are specific to a particular species.

#### ***Independent gene expansions***

**Figure 6-5 (g-k)** shows instances of independent gene expansions. In **panel (g)** the presence of multiple inparalogs in the outgroup, an apparent gene expansion in the YZ ancestral species and the absence of multiple copies of the gene in species X points to independent expansion events. Similar logic indicates that independent expansions have happened in the scenarios shown in **panels (h-k)**.

#### ***Ubiquitous gene expansion***

**Figure 6-5 (l-o)** illustrates four particular scenarios in which a gene could appear to be expanded in all the compared species. Although **panel (l)** represents the most parsimonious (and therefore most likely) model, the scenarios illustrated in **panels (m-o)** are all possible.



**Figure 6-3** examines three possible scenarios for a ubiquitous expansion across three compared genomes. This is a simplified version of the scenario illustrated in **Figure 6-5 (l-o)**; this figure illustrates only three genomes (as opposed to four) for visual simplicity; nonetheless the principle applies to any number of compared genomes.

<b>Table 6-1: GHO Assignments for the Patterns of Expansion Shown in Figure 6-3</b>			
	<b>Figure 6-3 (a)</b>	<b>Figure 6-3 (b)</b>	<b>Figure 6-3 (c)</b>
x1 y1	ortholog	ortholog	ortholog
x1 z1	ortholog	ortholog	ortholog
y1 z1	ortholog	ortholog	ortholog
x1 x2	internal_outparalog	internal_inparalog	internal_inparalog
x1 y2	external_outparalog	CEH / external_child_inparalog	CEH / external_child_inparalog
x1 z2	external_outparalog	CEH / external_child_inparalog	CEH / external_child_inparalog
y1 x2	external_outparalog	CEH / external_child_inparalog	CEH / external_child_inparalog
y1 y2	internal_outparalog	internal_outparalog	internal_inparalogs
y1 z2	external_outparalog	external_outparalog	CEH / external_child_inparalog
z1 x2	external_outparalog	CEH / external_child_inparalog	CEH / external_child_inparalog
z1 y2	external_outparalog	external_outparalog	CEH / external_child_inparalog
z1 z2	internal_outparalog	internal_outparalog	internal_inparalogs
x2 y2	ortholog	sibling_inparalogs	sibling_inparalogs
x2 z2	ortholog	sibling_inparalogs	sibling_inparalogs
y2 z2	ortholog	ortholog	sibling_inparalogs

In **Figure 6-3** a gene present in the last common ancestor of the compared species has undergone some number of duplication events such that there are two copies of the gene in each of the compared genomes. In the case of a three-way genome comparison this could occur because of a duplication in the last common ancestor, followed by two speciation events that created species Z, Y and Z (**panel a**). It could occur by a duplication in the intermediate species YZ, which then speciates to Y and Z; followed by an independent

duplication in species Z (**panel b**) . Finally it could occur by three independent, post-speciation duplications in species X, Y and Z.

The first proposed scenario (**panel a**) represents the most likely situation; however, as seen in the following chapter on *Leishmania* comparative genomics, gene expansion is often a complex mixture of pre-speciation and post-speciation duplication events. The ability to differentiate between the two is a highly useful tool for understanding evolutionary dynamics.

**Table 6-1** shows the different pairwise relationships between the six expanded genes that will result from each of the three possible scenarios. As the table shows, each of the scenarios results in a different set of relationships for each of the scenarios, thus allowing us to differentiate between the three. These relationships are all describable by the Gene Homology Ontology and calculable using our rule-based classification system.

## **6.6: LOGICAL GENE-GROUP QUERYING CONCLUSIONS**

The *logical gene group queries* discussed in this chapter find interesting or counterintuitive patterns in the gene groups that were generated using the strategy described in **Chapter 5**. The technologies presented in this chapter actualize the goal of providing an effective means of understanding the complex data generated by multi-species genome comparisons. The ontologies presented in **Chapter 3** form a solid foundation for the comparative genomics pipeline; the rule-based classification system in **Chapter 4** applied those ontology terms to pairwise gene relationships generated by sequence comparison results; finally, the *semantic gene grouping* procedure detailed in **Chapter 5** aggregated the pairwise relationships. Using queries discussed in this chapter we can leverage all the benefits of the above work to find patterns in families of genes that can illuminate the dynamics that shaped the evolution of our compared species.

*Logical gene group queries* can be both directed and exploratory. “Directed” in sense that there are certain well-known patterns of evolution for which they can search (*e.g.*, differential inparalogous expansion). These queries can be “exploratory” in the sense that after posing known (directed) queries a number of gene groups will still not be accounted for. These gene groups can represent patterns of evolution that are not widely known. Indeed, unusual patterns of gene expansion and differential gene loss were found in our *Leishmania* comparative genomics study (**Section 7.6**). Whereas the bulk of the work described before this chapter represent advances in representing existing knowledge, the strategies presented in this chapter facilitate the discovery of new knowledge as well.

**CHAPTER 7: *LEISHMANIA* COMPARATIVE GENOMICS**

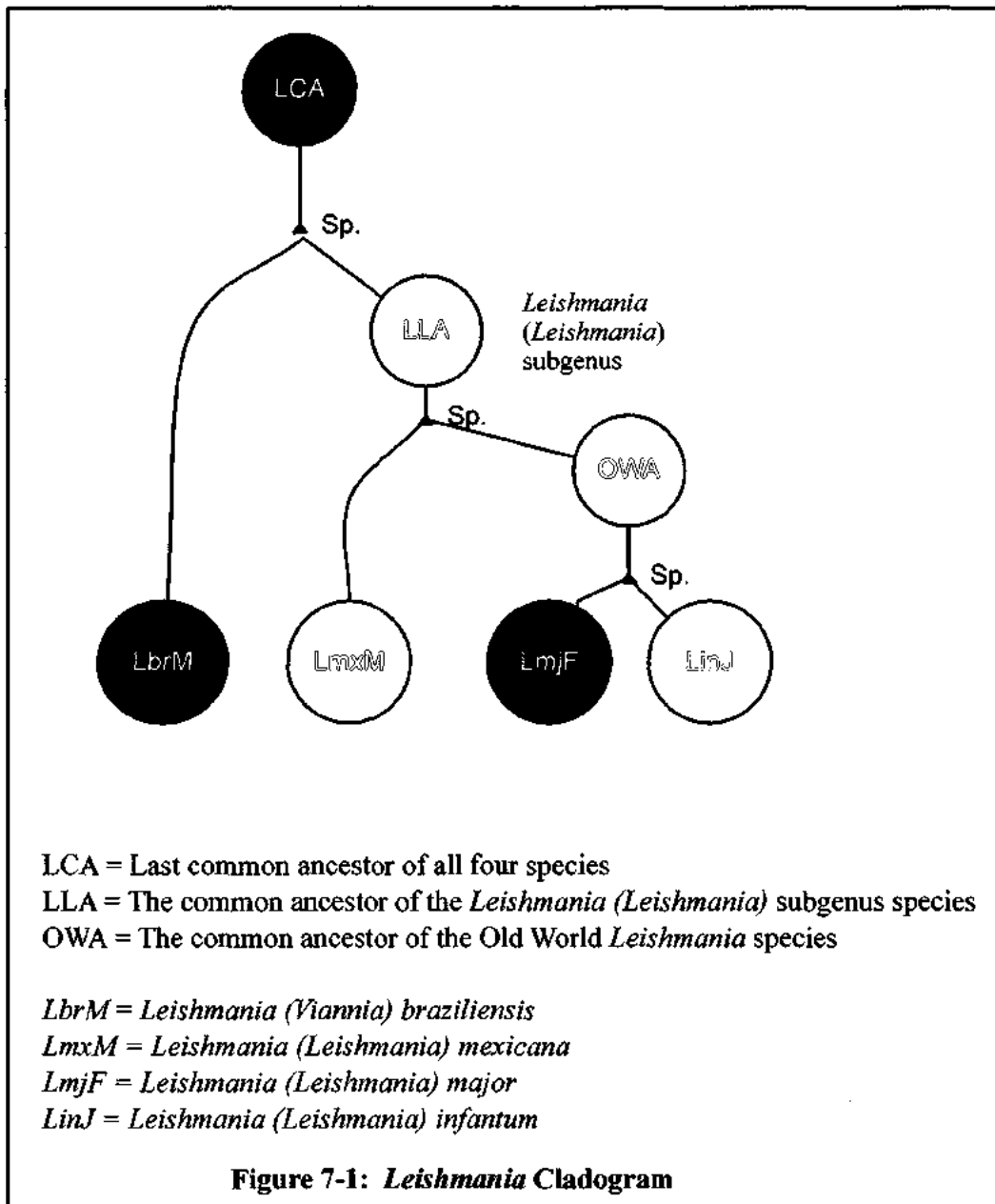
## 7.1: OVERVIEW OF THE *LEISHMANIA* COMPARATIVE GENOMICS PROJECT

This chapter discusses the results of a multi-way genome comparison of four human pathogenic species from the genus *Leishmania*: *L. braziliensis* (*LbrM*), *L. mexicana* (*LmxM*), *L. major* (*LmjF*) and *L. infantum* (*LinJ*). The latter three species are part of the sub-genus *Leishmania* (*Leishmania*), whereas *LbrM* belongs to the subgenus *Leishmania* (*Viannia*) and was the first to diverge from the last common ancestor (LCA) of the four species<sup>63</sup>. A cladogram that depicts the evolution of these species is shown in **Figure 7-1**. This tree is very similar to the tree shown in **Figure 6-1** of the previous chapter; therefore, the evolutionary patterns that are discussed in **Figure 6-4** (gene presence/absence) and **Figure 6-5** (gene expansion) are relevant to the results of this comparison as well.

**Section 7.2** provides background on the four *Leishmania* species that are compared in this experiment. **Section 7.3** gives a brief outline of the methods and technologies used for this experiment. **Section 7.4** discusses the benefits of our rule-based system in the context of the results from the subsequent three sections. **Sections 7.5, 7.6, 7.7** and **7.8** present various aspects of the results of the comparison.

### ***Leishmaniasis***<sup>64</sup>

Human Leishmaniasis is caused by approximately twenty protozoan species of the genus *Leishmania* and is transmitted by phlebotomine sandflies. Initially, the sandfly uptakes the parasite by biting an infected human host and subsequently, the parasite undergoes further development for four to twenty-five days within the sandfly. After this incubation stage is complete the sandfly is infective and is able to pass the disease on to a new host.



Leishmaniasis can result in several different disease manifestations, depending on the species or strain of *Leishmania*, as well as other host or environmental factors. The disease is classified into three broad categories: cutaneous leishmaniasis, mucocutaneous leishmaniasis, and visceral leishmaniasis. Cutaneous leishmaniasis (CL), typically the least serious form of the disease, causes serious skin ulcers on the face, arms and legs of infected individuals.



These ulcers can be numerous (up to 200 in number) and leave extensive, permanent scarring after the disease has passed. The mucocutaneous (ML) form of the disease causes severe lesions on the mucous membranes of the nose, mouth and throat and the surrounding tissues. These lesions can result in permanent disability. Finally, the most serious form of the disease is visceral leishmaniasis (VL). VL is characterized by swelling of the spleen and liver and causes fever, weight loss, and anemia. The fatality rate for persons infected with untreated VL is nearly 100% in the developing world.

Currently, over 350 million people in 88 countries are at risk for some form of leishmaniasis; the annual incidence is approximately 2 million cases. There are currently no vaccines for these diseases and existing drug therapies are highly toxic and are prone to development of resistance in the parasites <sup>65</sup>.

Given the serious nature of the disease and lack of currently available therapeutics and prophylactics, genomics analysis of species of the genus *Leishmania* is a potentially effective strategy for further understanding the parasite's mechanisms of pathogenicity and for eventually finding prospective drug targets. Furthermore, multi-genome comparison across the genus are particularly useful, given that disease manifestation and outcome is highly species-specific.

## **7.2: THE FOUR LEISHMANIA SPECIES**

Along with these four currently existing species, three relevant hypothetical extinct species existed (see **Figure 7-1**): the last common ancestor (LCA) of all four species, the common ancestor of the species in the *Leishmania* (*Leishmania*) subgenus (LLA) and the ancestor of the Old World *Leishmania* species (OWA). Later this chapter considers gene loss and gain in the context of these ancestor species.

***Leishmania (Viannia) Braziliensis (LbrM)***

*Leishmania braziliensis* is the leading cause of cutaneous leishmaniasis in Latin America and is known to cause a spectrum of diseases, including in some cases visceral leishmaniasis <sup>66</sup>.

*LbrM* is a particular human health risk since antimonials, the compounds typically used in the treatment of both cutaneous and mucocutaneous leishmaniasis, are typically less effective in the treatment of cutaneous leishmaniasis caused by *LbrM* <sup>67</sup>.

***Leishmania (Leishmania) mexicana (LmxM)***

*Leishmania mexicana* is endemic to South and Central America and causes cutaneous leishmaniasis along with the more severe diffuse cutaneous leishmaniasis.

***Leishmania (Leishmania) major (LmjF)***

*Leishmania major* is found in the subtropical and tropical regions in Africa, the Middle East and Asia; *LmjF* causes cutaneous leishmaniasis.

***Leishmania (Leishmania) infantum (LinJ)***

*Leishmania infantum* is the causative agent of visceral leishmaniasis, which if left untreated is fatal <sup>68</sup>. *LinJ* is most commonly found in the Mediterranean regions of Europe, Asia and the Middle East; however, cases have been found in Latin America <sup>69</sup>.

***Categorizing the four Leishmania species***

There several categories around which to organize the four *Leishmania* species. Two of the species (*LmxM* and *LbrM*) are endemic to Central and South America; these are known as the New World species. The other two species (*LinJ* and *LmjF*) are most commonly endemic to parts of Southern Europe, Northern Africa, the Middle East and South Asia and are known as the Old World species. Comparing the gene content of Old World and New World species can provide valuable insights regarding how the two groups have adapted to environmental,

host and vector selective pressures unique to their environments. One of the species, (*LinJ*) causes a considerably more serious disease than the other three; a comparison of the four species can illuminate possible genetic causes for this difference. Though the cutaneous disease-causing species (*LbrM*, *LmxM* & *LmjF*) result in somewhat similar clinical manifestations in humans, there are meaningful differences in severity and outcome between the three that can perhaps be explained by species-specific genomic differences.

### **7.3: METHODS USED IN THE LEISHMANIA COMPARISONS**

The *LbrM*, *LmjF* and *LinJ* genomes and gene predictions were obtained from the TriTrypDB ([www.tritrypdb.org](http://www.tritrypdb.org)) Kinetoplastid Genome Resource <sup>70</sup>, release 1.2. The *LmxM* data was obtained from GeneDB, a genomic sequence resource run by the Wellcome Trust Sanger Institute <sup>71</sup>. Sequence comparison was done using the Fasta sequence comparison program (not to be confused with the FASTA sequence format). The gene sequences and sequence comparison results were stored in a Chado database. The data were subsequently translated to Prolog facts (Section 4.2) and thereafter inputted into the rule-based system (Chapter 4) that assigns PGCO and GHO terms to the sequence comparison results. Next, the genes were grouped into clusters using the semantic gene grouping strategy described in Chapter 5. Finally, those groups were interrogated using the query strategies discussed in Chapter 6.

Statistics describing the number of genes in each genome and the number of gene groups generated are shown in Table 7-1. The results of each pairwise comparison are shown in Appendix C.

<b>Table 7-1: Gene Count and Gene Group Information by <i>Leishmania</i> Species</b>					
	<i>LbrM</i>	<i>LmxM</i>	<i>LmjF</i>	<i>LiuJ</i>	<i>TOTAL</i>
Total Genes + Pseudogenes	8133	8201	8406	8216	32956
Gene Groups	<i>7880 non-species-specific gene groups</i>				
Gene groups that contain at least one gene from this species	7426	7684	7849	7826	30785
% of total gene groups that contain at least one gene from this species	94.2%	97.5%	99.6%	99.3%	N/A
Number of genes in non-species-specific gene groups	8000	8169	8382	8191	32742
Percentage of genes in these gene groups	98.4%	99.6%	99.7%	99.7%	99.4%
Species Specific Gene Groups	<i>155 species-specific gene groups</i>				
Number of species-specific gene groups	77	32	21	25	155
Species-specific genes	133	32	24	25	214
Percentage of genes that are species-specific	1.64%	0.39%	0.29%	0.30%	0.65%

#### **7.4: ADVANTAGES OF EMPLOYING RULE-BASED HOMOLOGY ANNOTATION**

The *Leishmania* comparative genomics results presented in Sections 7.5 - 7.8 were all generated by our rule-based system. Before presenting the actual results of the comparison, this section discusses in some detail how the rule-based system bolstered this work and resulted in more precise and well-defined comparison of gene content.

##### ***Explicit specification of gene attributes***

The rule-based system explicitly specifies, *via* facts, attributes about each gene. The semantic gene grouping stet generates a fact that specifies group membership such as:

```
prop('LmjF07.1105', member_of, 'Gene Group 10')
```

In addition, each gene has already been annotated with properties describing various attributes, such as the genome to which the gene belongs:

```
prop('LmjF07.1105', member_of, 'LmjF V5.2').
```

and other attributes of the gene

```
prop('LmjF07.1105', pseudogene, TRUE).
```

Given the flexibility of the knowledge representation strategy, a user could add any number of annotations or supplemental information to the genes. Explicitly stating facts regarding the genes allows for querying the groups according to the attributes of the genes within that group. The remainder of **Section 7.4** discusses the types of queries that we posed on our gene groups.

#### ***Posing directed queries based on genome membership***

Since genome membership (and any other attribute) is explicitly stated as a fact, queries can be posed as to the evolutionary origin of genes. For example, in this particular *Leishmania* comparison, a gene most likely arose in the ancestor of the *Leishmania (Leishmania) (LLA)* subgenus if it is present in *LmxM* and one of the New World *Leishmania* species, but absent in *LbrM*. That rule can be encoded as such:

```
arose_in_leishmania_leishmania_ancestor(Gene_group) :-
  prop(Gene_LmxM, member_of, Gene_group),
  prop(Gene_LmxM, member_of, 'LmxM V3.0'),
  (
    (prop(Gene_LmjF, member_of, Gene_group),
     prop(Gene_LmjF, member_of, 'LmjF V5.2'))
    ;
    (prop(Gene_LinJ, member_of, Gene_group),
     prop(Gene_LinJ, member_of, 'LinJ V4.0'))
  )
),
not((
  prop(Gene_LinJ, member_of, Gene_group),
  prop(Gene_LinJ, member_of, 'LinJ V4.0')
)).
```

Rules such as this allow for queries on the gene groups that provide insights as to the probable phylogenetic timing of the gain of new genes and the loss of existing genes. Data obtained using this strategy is presented in **Section 7.6 & 7.7**.

### ***Easy editing of cluster information***

Because of the flexibility of fact- and rule-based representation of group membership, a user can encode their own knowledge seamlessly into the output of the semantic gene grouping. For instance, were a researcher to believe (based on some prior knowledge) that two groups should be merged into one group, that user could add a statement such as the following to the gene grouping output:

```
prop(Gene, member_of, 'Gene Group 11'):-
    prop(Gene, member_of, 'Gene Group 23').
```

The above statement says that all members of a given gene group (23) are members of another gene group (11). This is a fairly simple example, but in a detailed comparative genomics study a researcher may generate any number of study-specific observations and refinements of the data. The strategy of representing group membership using facts allows researchers to solidly encode those observations and ensure that they are accurately reflected in subsequent queries.

In the data presented in the following chapters, this strategy was employed to properly join genes that potentially had ambiguous group memberships. In the results presented hereafter the automatic gene grouping was able to place 99.9% of the genes (32,938 of 32,956) into groups without any user intervention, while 18 genes were assigned manually.

### ***Explicit relationships between genes in a group***

The output of the *semantic gene grouping* methodology, described in **Chapter 5**, explicitly describes the relationships between any two genes in a cluster. For instance, in this analysis a cluster of *cyclophilin* genes contains the following genes:

*LmjF33.1630, LinJ25\_V3.0940, LinJ33\_V3.1730, LbrM33\_V2.1900 and LmxM32.1630*

The Gene Homology Ontology assignments assert that gene *LinJ25\_V3.0940* is an ‘internal child inparalog’ of *LinJ33\_V3.1730*. The above group of genes is not simply a list, but rather a family tree of evolutionary relatedness.

One type of information that researchers typically seek from the results of a comparative genomics experiment is a summary accounting of how unique a gene is; does the gene have copies in other genomes, how many copies are in a particular genome, which genome has more copies, *etc.* The explicit Gene Homology Ontology assignments that this system provides, along with the group membership assignment, provides an accurate means for generating an accounting of such gene relationships.

#### ***Accurate cataloging of expansion events***

The explicit specification of relationships between genes allows for the better determination of when gene duplication events occurred. For instance, in the above example, it is specified that the gene *LinJ33\_V3.1730* is an inparalog of gene *LinJ25\_V3.0940*, because of the Gene Homology Ontology assignment:

`prop(LinJ33_V3.1730, internal_child_inparalog, LinJ25_V3.0940).`

The inparalog criteria satisfied by these two genes suggest that this gene expansion occurred after the speciation event that led to *Leishmania infantum*.

On the other hand, certain queries can uncover clusters that indicate expansion in ancestor species such as LLA. For instance, this analysis found a *hexokinase* gene group that apparently expanded in the LLA resulting in two copies in LmxM, LmjF and LinJ. This determination was made because the GHO assignments were more indicative of an expansion in the LLA ancestor than expansions post speciation in the three *Leishmania* (*Leishmania*) species. The issue of ancestral expansion and their implications is examined further in **Section 7.7**.

#### ***Accurately cataloging multi-copy gene families***

A common issue with clustering methodologies is that of dealing with closely related outparalogs. One school of thought suggest grouping them together (OrthoMCL <sup>72</sup>), other schools of thought suggest separating them (INPARANOID <sup>13</sup>). Both strategies have strengths and drawbacks. This system specifies explicitly that genes are outparalogs using fact statements such as:

```
prop('LmjF26.0900', internal_outparalog, 'LmjF30.2470').
```

This allows for differentiation between inparalogs and outparalogs in large families, thereby allowing for the better cataloging of patterns of evolution in large gene families.

This *Leishmania* analysis found a *heat shock 70* gene family with 28 genes across all four genomes. However, further analysis of the Gene Homology Ontology and gene grouping data established that these were two different genes in the LCA and that LLA had added an extra (third) copy of the gene. Furthermore, one of the copies of the gene had heavily duplicated in all four of the species to result in a total of 11 inparalogs. This represents a far more in-depth picture of the evolution of the heat shock gene family than simply stating that there are 28 copies across four genomes.



### ***Accurately placing fusion and splice genes***

A number of gene groups include fusion and splice genes. Most clustering methodologies make some sort of explicit “decisions” regarding how to best deal with fusion genes (*e.g.* put the splice genes in the same cluster with the fusion gene, join the fusion gene to the closest splice gene, *etc.*). This is similar to the earlier dilemma of dealing with outparalogs in that there are potential benefits and drawbacks to any predetermined strategy. This rule-based system explicitly states which genes are fusions of other genes; this analysis groups splices of fusion genes into the same cluster, however the rule-based system allows for the easy determination of which clusters contain splice genes using a rule such as:

```
group_has_splices(Gene_Group) :-
    prop(Gene, member_of, Gene_Group),
    prop(Gene, splice, Genes2).
```

Queries similar to the above uncovered 83 gene groups with potential splice events that warrant further analysis. Given the relatively small phylogenetic distances between the *Leishmania* species used in this comparison, it is highly probably that these fusion/splice events represent incorrectly called gene boundaries.

## **7.5: EVOLUTION OF NEW GENES IN *LEISHMANIA***

This section discusses when particular genes arose in the *Leishmania* species. Certain genes arose in the last common ancestor (LCA) of all four species. These genes may play a role in the basic processes of life common to a wide spectrum of *Leishmania* species. Certain genes likely arose in the *Leishmania (Leishmania)* subgenus (LLA), other genes arose in the ancestor of *LmjF* and *LinJ* (OWA); furthermore, each species has some number of gene that arose post-speciation and are unique to that species alone.

A potential complication in determining the origin of a gene is that certain patterns of evolution are difficult to differentiate. For instance, pairwise gene content comparison does

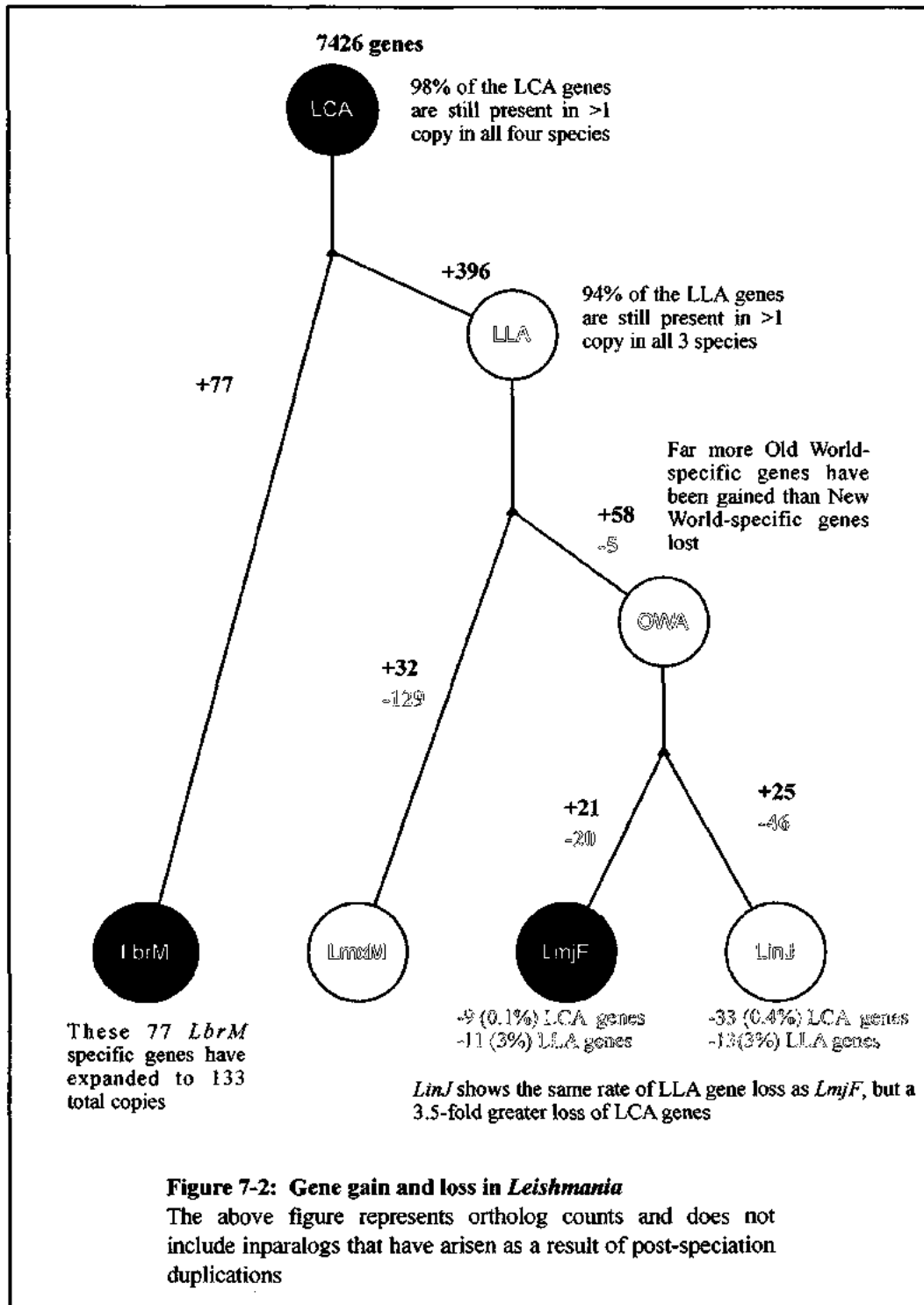
not definitively discern whether a pattern of gene loss/presence was a result of the evolutionary pattern shown in (for example) **Figure 6-4 (a)** or **Figure 6-4 (l)**; Similarly **Figure 6-4 d & h** represent patterns of evolution that are difficult to differentiate. There are some additional analyses that could distinguish between these patterns; the beginning of **Section 7.5** will discuss those techniques, and the concluding chapter (**Chapter 9**) will explore how those techniques could fit into our rule-based pipeline.

The above caveats aside, a number of definitive, or at least highly plausible, statements can be made regarding the origins of the genes in our comparison. The data in this section, as well that in **Section 7.6** are displayed graphically in **Figure 7-2**.

#### **Genes That Were Present In The Last Common Ancestor Of All Four Species**

Of the 8035 gene groups (155 species-specific groups and 7880 cross-species groups) that we generated using the strategy outlined in **Chapter 6**, 92% (7426) of them apparently arose in the LCA of the four species. These gene groups represent the evolutionary patterns shown in **Figure 6-4 (c, i, j, k & l)**. Of those ancestral gene groups 98%, (7241) are still present in some for in all four species, while 445 gene groups (6%) have undergone some sort of inparalog expansion in one or more of the species.

Another 77 gene groups are specific to *LbrM* and potentially represent genes that were either: 1) gained during the evolution of the *Leishmania (Viannia)* subgenus; or 2) present in the last



common ancestor, and lost in the *Leishmania (Leishmania)* subgenus. The subsequent section discusses these genes.

#### **Genes Specific To *Leishmania (Viannia) braziliensis***

There are 77 gene groups that are unique to *LbrM*, either due to species-specific gene gain, or loss in the *Leishmania (Leishmania)* lineage. Of these groups, 72 exist as single copies in the genome, and five exist as multi-gene families: one family of unknown function contains 34 members, a family of SLACS retrotransposable element <sup>73</sup> genes contains 15 members, and four more families of unknown function with five, four, two and two members each. A total of 133 individual genes across these 77 families are unique to *LbrM* in this comparison.

As mentioned earlier in **Section 7.2** *LbrM* displays some drug resistance characteristics not observed in the other cutaneous disease-causing species (*LmxM* and *LmjF*); furthermore, unlike those species, *LbrM* can cause mucocutaneous leishmaniasis as well. These 77 gene groups possibly play a role in understanding these clinically significant characteristics of *LbrM*.

The presence of large families of species-specific genes in *LbrM* is in marked contrast the other three *Leishmania (Leishmania)* species; the 78 species-specific genes in those species are all truly unique and none share any detectible sequence similarity. This is most certainly because *LbrM* diverged from the LCA far longer ago than any of the other species.

#### **Genes that arose in the *Leishmania (Leishmania)* lineage**

There are 396 gene groups that arose in the *Leishmania (Leishmania)* lineage, and of those 94% (372) are still present in all three species (*LmxM*, *LmjF*, *LinJ*). These subgenus-specific genes show a slightly higher rate of gene loss than the genes inherited from the LCA; 98% of the genes from the LCA are still present in all three species. The patterns of evolution that

likely produced these genes are shown in **Figure 6-4 (l, o & p)**. Of these 396, 9% (35) have undergone some sort of inparalogous expansion in one or more species. By contrast only 6% of the gene groups inherited from the LCA have undergone one or more expansions. While the LCA gene groups have existed much longer than the subgenus-specific gene groups, this definition of expansion only accounts for post-speciation expansion events; by this measure it appears that the subgenus-specific genes are more prone to expansion than the ancestrally obtained genes. In this case we are measuring expansions that have occurred subsequent to the four speciation events that created the four species. As such, the LLA genes and the LCA genes have had an equal amount of time to accumulate duplications. This, along with the knowledge that these subgenus-specific genes are also lost at a higher rate than ancestral genes, points to a greater degree of dynamism among these genes.

***Old World Genes - Genes that arose in the ancestor of LmjF and LinJ***

There are 58 genes that appear to have arisen in the ancestor of LmjF and LinJ; these are the two species in our comparison that are most prevalent in the Old World (Mediterranean, Africa, the Middle East and South Asia); these 58 genes may represent genes that are critical for surviving in the face of some pressure uniquely present in the Old World. These pressures likely take the form of some combination of climatic factors, genetic factors in the host (human) population in the area, and the particular species of sandfly vectors that propagates the disease in the Old World. Though horizontal gene transfer (HGT) is not explored in this work, it is possible that some amount of the 58 genes are a result of HGT events from some species (presumably bacterial) that is only present in the Old World.

Of these 58 genes, 8.6% (five genes) have undergone species-specific inparalogous expansions.

***Species-Specific genes in Leishmania (Leishmania)***

There are 78 genes that are specific to one of the three *Leishmania* (*Leishmania*) species (*LmxM*, *LmjF* & *LinJ*). These genes are truly novel with no detectable sequence similarity to any other gene in the comparison. Presumably these, genes were only recently gained (after the speciation events that divided these three species); none of them have yet duplicated into larger gene families, suggesting that perhaps the speciation events that separated these three species occurred relatively recently.

The 78 genes consist of 32 genes that are specific to *LmxM*, 21 genes are specific to *LmjF* and 25 genes are specific to *LinJ*. Some of these genes may be related to nuances of parasite survival in their particular niche. Additionally, some of these genes may be related to disease manifestation. For instance the 25 *LinJ* specific genes are perhaps related to the parasite's ability to cause visceral leishmaniasis. Similarly, analysis of the *LmxM* and *LmjF* specific genes may provide some insight into why the former causes diffuse cutaneous leishmaniasis.

## **7.6: GENE LOSS IN THE LEISHMANIAS**

Just as our methods cannot determine with absolute certainty when a particular gene arose (Section 7.5), genome comparison cannot absolutely determine when a gene was lost in a lineage. Nonetheless, the data and resulting analyses provide considerable insight into the most probable model of gene loss in *Leishmania* lineage. The gene loss data presented in this section are summarized in Figure 7-2.

### ***Gene loss in LbrM***

Given that *LbrM* was the first species to diverge from the last common ancestor of the compared species, the data cannot describe whether genes absent from this genome were lost in *LbrM* Figure 6-5 (l) or gained in the *Leishmania* (*Leishmania*) subgenus Figure 6-5 (a).

Typically, assessing the relative merits of multiple evolutionary possibilities that would result in the same observation entails invoking the principle of maximum parsimony, the notion that the explanation that requires the least number of evolutionary events is most likely true. This principle does not necessarily differentiate between these two scenarios, since the loss of genes in the *LbrM* is no more or less likely than the evolution of new genes in the *Leishmania* (*Leishmania*) subgenus. As such, some subset of the 372 genes were described as *Leishmania* (*Leishmania*)-specific are likely genes that arose in the last common ancestor of the four species, but were subsequently lost in *LbrM*.

Comparing the 372 lost-or-gained gene families to the genome of *LbrM* could provide some amount of insight regarding whether they were lost in *LbrM*. For example, a TBLASTN (a sequence comparison of protein sequence against a collection of nucleotide sequences) analysis could identify former genes that have degenerated and lost their coding potential in *LbrM*. This would not result in a comprehensive cataloging of lost genes; a gene could have deteriorated in *LbrM* to a point where it is unrecognizable through TBLASTN analysis, alternatively a gene loss in *LbrM* could have been lost as a result of some manner of excision of an entire chromosomal region in *LbrM*.

The above caveats aside, a TBLASTN analysis would certainly provide some additional insights into the question of gene loss or gene gain. Such an analysis has not been included in this discussion, but **Chapter 9** does discuss how one could integrate such an analysis into the rule-based framework.

#### ***Gene loss in the Leishmania (Leishmania) subgenus***

As discussed in the prior section the data cannot discriminate between the question of loss in *Leishmania* (*Leishmania*) **Figure 6-5 (d)** versus gain in *LbrM* **Figure 6-5 (h)**. As such, the

77 gene families that are novel to *LbrM* might constitute lost genes in *Leishmania* (*Leishmania*).

**Gene loss in the New World Leishmanias (*LmjF* and *LinJ*)**

Only five genes have been lost in the Old World species. Such a loss manifests itself as presence of a gene in *LbrM* and *LmxM* (indicating that the genes were present in the last common ancestor) and an absence of genes in *LinJ* and *LmjF*.

These five genes may represent genes that are critical to survival in the New World (owing to geographic, host or vector pressures), but were unnecessary or disadvantageous for survival in the Old World. This result is somewhat surprising; relatively few genes are specific to the New World species, especially since our analysis shows 58 genes that appear to be specific to the Old World species.

As of this publication, the gene prediction in the Old World species are more refined and complete than the gene prediction in the New World species; this somewhat confounds the analysis, and especially so when considering genes gained in the New World species. Nonetheless, the scale of difference between the “hemisphere-specific” genes (5 compared to 58) is extraordinary; furthermore, in all likelihood 95% of the genes in the New World species have been identified, meaning that there most probably is a true disparity.

**Gene Loss in *LmjF***

Nine of the gene groups that were apparently present in the last common ancestor and 11 gene groups that arose in the *Leishmania* (*Leishmania*) subgenus are absent in *LmjF*; these patterns are illustrated by **Figure 6-4 (i & p, respectively)**. Of the genes that were present in the last common ancestor only 0.1% were lost, by contrast 3% of the genes that were gained



in the *Leishmania* (*Leishmania*) subgenus were lost, again indicating greater dynamism in the newly acquired genes.

#### **Gene loss in *LmxM***

Of the ancestrally present gene groups, 1.7% (129) have been lost in *LmxM*; this is illustrated in **Figure 6-4 (k)**.

This analysis cannot accurately determine how many of the subgenus-specific genes have been lost in *LmxM*, as the data cannot distinguish between such a gene loss and a gene that arose in the ancestor of *LmjF* and *LinJ*; for simplicity, we are assuming that the most straightforward explanation for the pattern seen in **Figure 6-4 (b)** is the latter and not the former.

The rate of gene loss of LCA genes in *LmxM* is markedly higher (129) than in *LmjF* (20) indicating extremely high numbers of lost genes in *LmxM*; although the effect is likely exacerbated by a significant under-prediction of genes in the current release of the genome sequence.

#### **Gene loss in *LinJ***

Of the 7426 genes present in the LCA, 0.4% (33) of them have been lost in *LinJ*, and 3.5% (13) of the 372 *Leishmania* (*Leishmania*) subgenus-specific genes have been lost. *LinJ* exhibits a nearly identical rate of subgenus-specific gene loss to *LmjF*, but a 3.5-fold higher rate of loss of ancestrally present genes. While the *LinJ* gene predictions are not quite as well-curated as the *LmjF* gene predictions, they are of sufficient quality to indicate that there is a trend towards more rapid than expected loss of ancestral genes in *LinJ*; furthermore, that trend does not appear in the subgenus-specific genes.

The above evidence suggests that the genes related to cutaneous leishmaniasis appear to have originated in the LCA and are among the 33 LCA genes lost in *LinJ*. Given that both the *Leishmania (Leishmania)* subgenus and the *Leishmania (Viannia)* subgenus are capable of causing cutaneous leishmaniasis, logic suggests that the genes responsible for the disease were mostly present in the ancestor species of the two genera. The accelerated loss of LCA genes in *LinJ* (which causes visceral leishmaniasis), as compared to *LmjF* (which still causes cutaneous leishmaniasis) further strengthens the notion that the genes relevant to cutaneous leishmaniasis were gained in the LCA, and furthermore those have perhaps been rapidly lost in *LinJ*.

**Gene Loss in the cutaneous disease causing species (*LmxM* and *LmjF*).**

Six gene groups were apparently lost in both *LmjF* and *LmxM*. Owing to the topology of the phylogenetic tree, a loss of a gene family that was present in the LCA and is still present in *LinJ* would necessarily represent a loss in *LmxM* and an independent loss event in *LmjF*. This pattern is illustrated in **Figure 6-4 (n)**. In other words, there was evidently some selective pressure operating independently on both *LmjF* and *LmxM* that caused them to lose these genes.

A potential explanation for the above phenomena is that *LbrM* causes a relatively wide spectrum of human disease, including cutaneous and mucocutaneous leishmaniasis, whereas *LmxM* and *LmjF* only cause cutaneous leishmaniasis. The loss of these genes might represent the narrowing of the spectrum of disease that occurred in these lineages. The continued presence of these genes in *LinJ*; a potential explanation is that they are also somehow involved in the more serious visceral leishmaniasis caused by *LinJ*.

The presence of independently lost genes can point to genome assembly or annotation errors; in practice, further analysis is needed to determine if these genes were indeed independently lost or simply unannotated in a particular genome.

### **Genes independently lost in *LmxM* and *LinJ***

Three genes appear to have been independently lost in *LmxM* and *LinJ*. These genes follow the pattern shown in **Figure 6-4 (m)**. It is difficult to determine a clear story that describes why a New World, cutaneous-disease-causing pathogen (*LmxM*) would concurrently lose the same ancestral gene as an Old World, visceral-disease-causing pathogen (*LinJ*), while that same gene would be maintained in an Old World, cutaneous-disease-causing pathogen (*LmjF*). Indeed, this illustrates the point that while simple explanations are often attractive, the factors responsible for gene or lost are often elusive. Hypotheses such as “certain genes are necessary for survival in the New World, but lost in Old World species”, might appeal to a researcher’s sense of order, the truth is likely far more complicated. That is not to say that such statements are never true; however, as illustrated here, the factors that effect genome dynamism are often elusive or convoluted.

As with any genome comparison, this analysis is constrained by the quality of the underlying genome assemblies and gene predictions. These three genes, which have a somewhat counterintuitive species distribution, make an excellent starting point for genome annotators looking to find errors in the gene predictions.

### **7.7: INPARALOGOUS EXPANSION IN LEISHMANIA**

Recalling the homology definitions from **Section 2.2**, inparalogs refer to gene duplications that occurred after a speciation event that caused the divergence of the compared genomes. When comparing a pair of genomes this distinction is quite clear: for all practical purposes there is one point of evolutionary bifurcation that eventually resulted in the two species.

Comparing multiple species adds an element of complexity to the discussion; however, a discussion of the definitional complexities of inparalogs in multi-species comparisons should be preceded by noting the evolutionary implications of inparalogs. Inparalogs are typically interesting because they illustrate a functional role that is somehow “expanded” in a particular lineage. One gene was sufficient to fill some functional role in an ancestor species, but for some reason multiple copies of that gene appear to be advantageous in a particular descendant species. Analysis of such expansions can provide insight into particular selective pressures placed on a species or unique functional abilities of that species. As mentioned earlier in **Section 7.6**, there are 7241 genes in all four of the compared *Leishmania* species that were in all likelihood inherited from the LCA; these genes represent a sort of genetic “common ground” shared by the four species. However, among those common genes 421 have undergone one or more inparalogous duplications in one or more of the species. These 421 cases represent a refinement of the common genes and are just as important to the understanding of genomic content as the evolution of new genes and the loss of old ones.

In all likelihood most “new” genes are inparalogs of existing genes (though some are acquired by horizontal gene transfer) that have diverged to a point that they can no longer be recognized as similar using sequence comparison. Indeed, this is thought to be the primary mechanism by which new genes arise <sup>19</sup>. From a pragmatic view this distinction does not necessarily affect the following discussion of inparalog expansion; nonetheless, it is worthwhile noting that a novel gene in a species and an inparalog are usually generated by similar mechanisms.

Returning to the complexity of defining gene expansions in a multi-genome comparison, this type of multi-way comparison can discern several types of inparalogs:

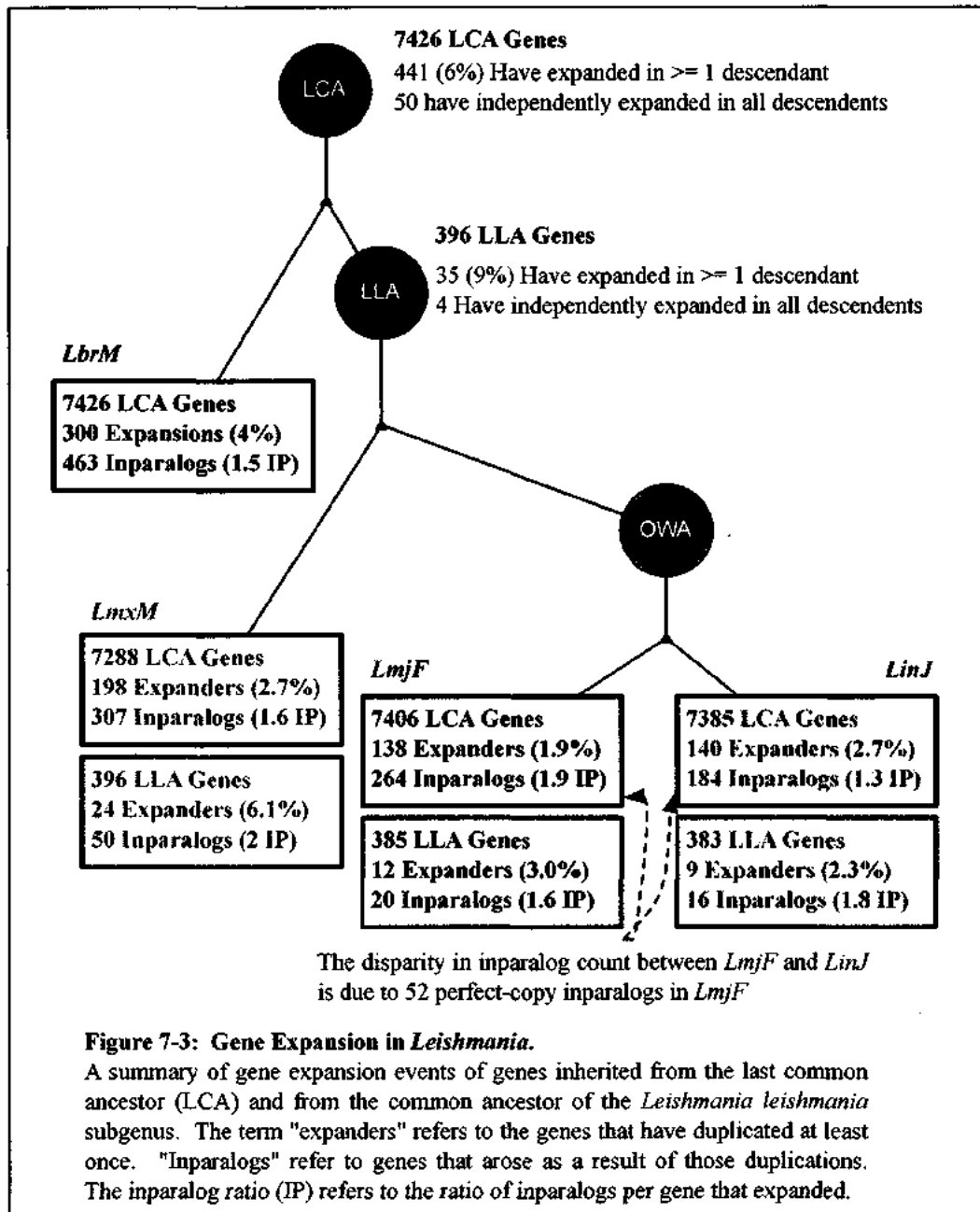
1. Genes that duplicated uniquely in a species following all the speciation events.

2. Genes that expanded in the *Leishmania (Leishmania)* subgenus after the speciation event that causes separation from the *Leishmania (Viannia)* subgenus.
3. Genes that expanded in the common ancestor of *LmjF* and *LinJ*.

There are two potential points of confusions that bear clarification at this point. First, while there are a number of ways of defining inparalogs in a multi-way comparison, the term “inparalog” is still very concrete when discussing a pair of genes in a comparison of two genomes. As such the Gene Homology Ontology definitions of various types of inparalogs are unambiguous. Second, the inparalogs discussed in **Section 7.6** were all a result of gene duplications that occurred post-speciation of all the compared species.

In reality there are nearly unlimited possible patterns of gene gains and losses followed by gene expansions; indeed this four species analysis contains more patterns than can possibly be cataloged here. Nonetheless, the remainder of this section serves to discuss some notable patterns and give some idea of the extent that gene duplications have played in the evolution of these species.

The results of the species-specific inparalog expansion analysis presented in this section are summarized in **Figure 7-3**.



***Inparalogs of LCA genes***

There are 7426 genes that are currently present in one or more species that were also apparently present in the LCA; of these 6% (445) have expanded in one or more species. *LbrM* has the most genes that have expanded one or more times among this group (300 genes), followed by *LmxM* (199 genes) and *LinJ* (140) and *LmjF* (138). This pattern is expected, given that *LbrM* diverged from the LCA considerably earlier than the other three, and as such has had more time to accumulate gene duplications. *LmxM* diverged next, and that is reflected in the number of duplications. Interestingly, *LmjF* and *LinJ* have had approximately the same number of genes diverge since they duplicated from each other. These numbers seem to confirm the admittedly intuitive notion that in this lineage the total number of duplicated genes is proportional to the time since the speciation event.

The above paragraph discusses the number of genes present in the LCA that have expanded, expansion can also be measured by the number of inparalogs that have been generated. In other words the duplication of some hypothetical gene (gene A1) twice (to gene A2 and gene A3), counts one expanded gene; however, it generates two new inparalogs.

By this definition *LbrM* has 463 inparalogs that arose from genes present in the LCA, *LmxM* has 309, *LmjF* has 264 and *LinJ* has 184. *LmjF* has 138 genes from the LCA that have led to 264 inparalogs, and *LinJ* has a similar number of genes (140) that have given rise to fewer inparalogs (184). Further analysis shows that 52 of these *LmjF* inparalogs are exact sequence copies of either another inparalog, so in that sense there are 212 (264 minus 52) *distinct* inparalogs from LCA genes in *LmjF*. By that logic there are only 7 exact copy inparalogs in *LinJ*, leaving 177 *distinct* inparalogs. It appears that this large discrepancy in inparalogs between *LmjF* and *LinJ* is largely due to the phenomena of exact-copy inparalogs in *LmjF*; this likely means that these are inparalogs that are a result of very recent gene duplications and hence have not had time to accumulate any sequence divergence. Similar

analysis of *LbrM* and *LmxM* show very few perfect copy inparalogs (3 and 10, respectively). It appears that *LmjF* has been subject to a number of recent gene duplications, and that this phenomenon is largely not present in the other three species. These exact-copy inparalogs are largely the result of duplications in 12 genes of varying functional annotation. An interesting aspect of these perfect-copy duplications in *LmjF* is that they are predominantly in tandemly repeated genes (closely related inparalogs that are adjacent on a chromosome). Such genes often pose sequence assembly difficulties such that tandemly repeated genes are often annotated as a single gene.

Of the genes that are present in the LCA, there are 50 that appear to have independently expanded in all the *Leishmania* species. The independent nature of these expansions suggest that these genes possibly play a role in the adaptation of a particular species to its unique ecological, vector and host niches. An alternate explanation is that the genes are surrounded by repeat elements that cause greater rates of duplication a given area of a chromosome. Of note among this group is an 'amastin-like surface protein' that has expanded 14 times in *LbrM*, 16 times in *LmjF* and once each in *LmxM* and *LinJ*. Pathogen surface proteins are known to interact with host immune defenses, so it is not surprising that there is a degree of dynamism in these genes. Notably, 23 of these 50 ubiquitously expanding genes are ribosomal proteins.

The previous few paragraphs present a great deal of information on the ways in which genes inherited from the LCA have expanded and possibly neofunctionalized or subfunctionalized. Nonetheless there is remarkable stability in the core genome that these four species have inherited from their ancestor: 90% of the genes that were present in the LCA have neither been lost nor have expanded in any of the four species.

#### ***Inparalogs of Leishmania (Leishmania) subgenus-specific genes***



There are 396 subgenus-specific genes in *Leishmania leishmania* and 9% (35) have expanded in one or more species. As mentioned in **Section 7.6**, this a greater percentage of these genes than that for genes that arose in the LCA (only 6%). Almost a third of the expanders are some sort of surface protein gene. In *LmxM* 24 genes have expanded, resulting in 49 new inparalogs. A significant percentage of these are the result of one gene (an amastin) that has expanded to 16 inparalogs in *LmxM*; by contrast this same gene has expanded once each in *LmjF* and *LinJ*. 12 genes from *LmjF* have expanded, resulting in 20 inparalogs. Nine genes from *LinJ* have expanded, resulting in 16 inparalogs.

Interestingly the phenomenon of perfect copy *LmjF* inparalogs is not seen here; all 20 of the inparalogs are distinct from each other. All the perfect copy inparalogs in *LmjF* are from genes inherited from the LCA. This could simply be an issue of small numbers: *LmjF* has only 12 expanded subgenus-specific genes (LLA) and 138 expanded LCA genes. Nonetheless, those 138 expanded LCA genes led to 55 perfect copy inparalogs and these 20 led to none. This reinforces the notion that some recent selective pressure is operating on *LmjF*, specifically in the core ancestral genes.

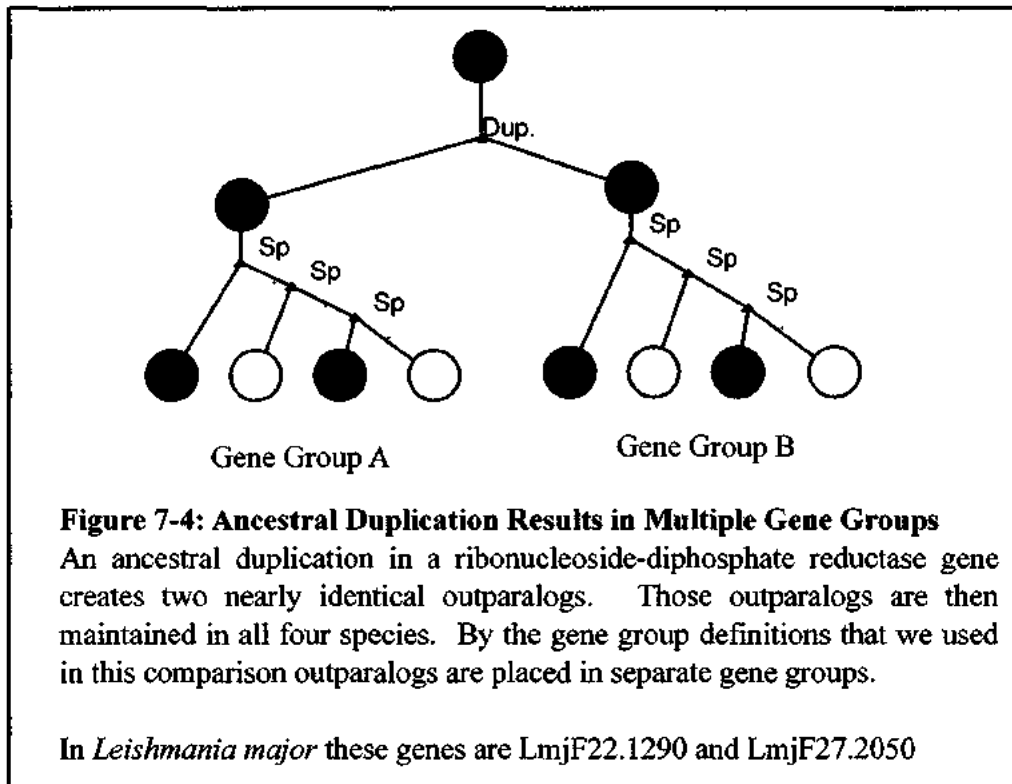
#### **Gene expansions in Old World specific genes**

Of the 58 genes that arose in the ancestor of *LmjF* and *LinJ*, there have only been five genes that have undergone any sort of expansion. Two genes have duplicated once in *LmjF*, two other genes have duplicated once in *LinJ* and an amastin has duplicated multiple times in both.

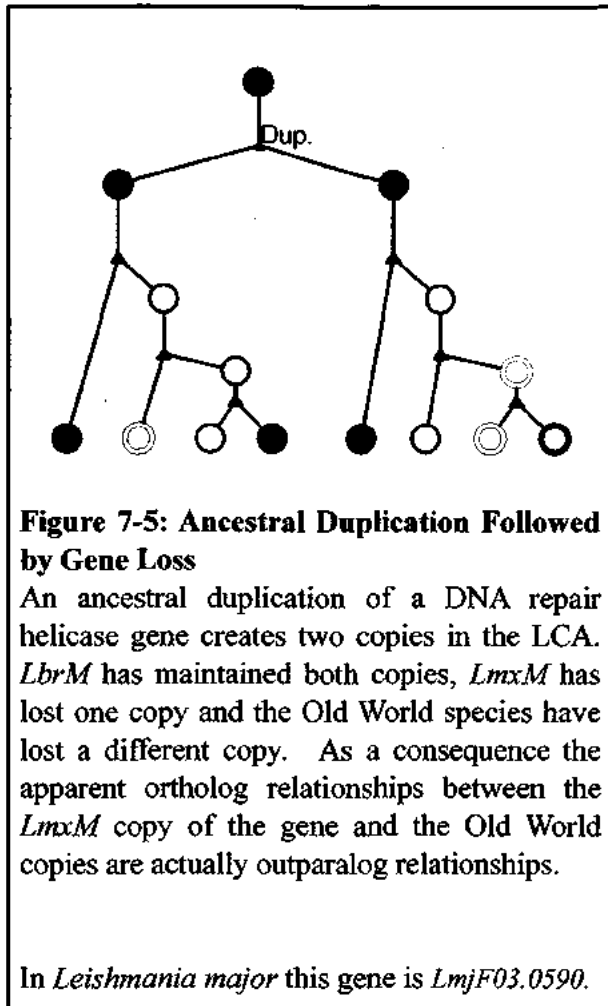
#### **Gene expansions in the intermediate species**

Thus far, the discussion has focused on gene expansions that have happened after the speciation event separated a particular species from all other species in the comparison. Gene

duplications can occur in ancestor species such as (LCA, LLA, OWA) as well, the remainder of this section focuses on those duplications.



Most gene duplications in the LCA manifest themselves as outparalogs in the descendant species; accordingly the genes that descended from each of the paralog LCA genes will end up in different gene groups. This is illustrated in **Figure 7-4**, which depicts the evolution of two nearly identical 'ribonucleoside-diphosphate reductase' genes (*LmjF22.1290* and *LmjF27.2050*) that duplicated in the LCA. Occasionally, however, a gene duplication will occur in the ancestor species, and a particular lineage will lose one copy of the gene, and another lineage will lose the other copy. This scenario is illustrated in **Figure 7-5**, which depicts an ancestral gene duplication in a DNA repair helicase (*LmjF03.0590*) that duplicated in an ancestral species, only to lose different copies of the gene in *LmxM* and the Old World Species.



That latter scenario seems to be a frequent dynamic in *Leishmania* evolution: gene duplication followed by species-specific or lineage-specific loss of an individual copy of the gene; 272 gene groups show some evidence to suggest that this has occurred. There are more minor and major variations on this scenario than we can catalog here, and indeed each requires some degree of individual inspection. Nonetheless, there are some easily described example scenarios that can further clarify this phenomena.

The chosen example is a scenario in which a gene duplicates in the LLA common ancestor, and *LmxM* retains both copies. The stipulation that both genes be retained in *LmxM* is needed because if one of the duplicated genes were lost in *LmxM* this analysis would not be able to conclusively say that the duplication occurred in LLA, as opposed to OWA. There are any number of ways that the duplicate copies of the gene could then evolve in *LmjF* and *LinJ*. This discussion will next focus on one of those possibilities: the case where *LmjF* maintains a copy of the gene that *LinJ* loses and *LinJ* maintains a copy of the gene that *LmjF* loses.

This course of events indicates that the apparent orthology between these genes in *LmjF* and *LinJ* is not true orthology. In a comparison of just *LinJ* and *LmjF* they would appear to be orthologs, and for most practical purposes they are, but this more refined analysis illustrates that they share a much more complicated relationship. In this example the gene has duplicated in LLA and presumably the two copies have diverged to some degree. Some selective pressure caused *LmxM* to maintain both copies of the gene. Similarly, selective pressures caused *LmjF* and *LinJ* to make the evolutionary choice between the two copies of the genes; in this case they chose differently. Although it is impossible to determine why exactly this “choice” was made, it is plausible that apparent orthologs resulting from such choices likely have some subtly different relationship to each other as compared to orthologs that arose by the common pattern of speciation.

The data suggest that there are 39 cases where a gene duplication has occurred in LLA and has been maintained in *LmxM*. In 18 of those cases, *LmjF* and *LinJ* have kept different copies of the expanded gene, indicating a more complicated relationship among the apparent orthologs.

The data surrounding gene duplications in intermediate species (such as the data presented above) are often difficult to analyze and do not really lend themselves to definitive conclusions; nonetheless the phenomena of gene duplication followed by gene loss illustrates what is likely to be an important dynamic in the evolution of these species.

## **7.8: SUMMARY OF THE LEISHMANIA COMPARATIVE GENOMICS STUDY**

### ***Comparison to existing studies***

An existing three-way comparison of *LmjF*, *LinJ* and *LbrM* was published in 2005<sup>48</sup>. This study employed a more standard means of comparison that involved clustering genes from the three species and finding commonalities and differences in the genomes, and thus did not

asses patterns of gene expansion and gene loss in the same manner that work presented here. As such, the earlier work does not necessarily present as detailed a picture of *Leishmania* evolution.

One major area of disagreement between the present and earlier work is the extent to which species-specific genes are present in the *Leishmania* lineage. The 2005 study found only 78 species-specific genes across the three lineages, 5 in *LmjF*, 26 in *LinJ* and 47 in *LbrM*. Our study found significantly more species-specific genes, 21 in *LmjF*, 25 in *LinJ* and 133 in *LbrM*. The difference in results between the two studies is probably due to improvements in the gene predictions (particularly in *LbrM*) over the last five years, as well as differences in clustering methodologies. A consequence of the *semantic gene grouping strategy* employed for this study is that distantly related genes will not be placed into the same cluster; this has the effect of calling more genes novel or species-specific.

Novel genes arise primarily from duplication of existing genes<sup>19</sup>, so the distinction between “novel” and “extremely diverged” is not concrete. Nevertheless, we feel that absent significant sequence similarity a gene should be designated as novel even if a clustering algorithm connects it to another gene. Considering the high level of sequence conservation in most genes across these three *Leishmania* species, highly diverged genes do represent a difference more than they represent a similarity, and we feel they should be represented as such.

### **Summary of results**

Through the use of Gene Homology Terms, semantic grouping methodology and subsequent rule-based querying this analysis was able to extract significantly more data from our four-way *Leishmania* comparative genomics project than would be available simply by assessing

the output of a standard clustering algorithm. Indeed, these results illuminate some interesting and potentially useful insights into *Leishmania* evolution.

There appears to be a remarkably well-conserved core genome that has remained present in all four of the species; of the genes that appear to have arisen in the last common ancestor, 98% are still present in all four species. The genes that appear to have arisen in the *Leishmania (Leishmania)* subgenus appear to be less stable, indeed > 6% of them have been lost in one of the three species in that subgenus. Furthermore, the subgenus-specific genes appear to have undergone 1.5 times as many inparalogous gene duplications as the genes from the LCA.

The species endemic to the Old World appear to have gained 58 genes, but lost relatively few genes, with *LinJ* losing LCA genes at a far faster rate than the closely related *LmjF*. Conversely, *LmjF* appears to have a number of recent gene duplications (as evidenced by copies identical sequence inparalogs), a trend that is not common any of the other four species.

The lone member of the *Leishmania (Viannia)* subgenus, *LbrM*, contains several multi-copy specie-specific gene families, presumably because it diverged from the last common ancestor much longer ago than the other three species.

Another trend that appears common among the *Leishmanias* is the tendency towards gene duplication followed by differential loss of a particular copy in the descendant species. This trend is difficult, even with our rule-based software, to detect; however, we have uncovered almost 300 cases where this dynamic appears to be in play. This paints a slightly more complicated picture of orthology than the commonly held notion that orthologs are simply the same gene from an ancestor species.



## **CHAPTER 8: CROSS-PHYLA COMPARATIVE GENOMICS**



## 8.1: INTRODUCTION

This chapter describes the application of comparative genomics pipeline to three widely diverged bacterial pathogens. Whereas the comparison of the four *Leishmania* juxtaposed very closely related species, this comparison assesses the utility of the pipeline in contrasting the gene content of vastly diverged species.

The results presented in **Chapter 7** relied upon the techniques of *Semantic homology annotation* (**Chapter 3 & 4**), *Semantic gene grouping* (**Chapter 5**) and *Logical gene group querying* (**Chapter 6**) to construct a detailed description of gene content across four genomes and postulate how the dynamics of gene gain, gene loss and gene duplication gave rise to currently observable differences in gene content. While that sort of investigation of closely related pathogens is arguably the most common sort of comparison performed in infectious disease genomics, comparisons of broadly divergent species is a relevant endeavor as well.

Comparing broadly different disease-causing species can provide insights into widely conserved fundamentals of survival as well as mechanisms of pathogenicity. Furthermore, such comparisons carry a degree of difficulty that is not present in the comparison of closely related species, because the homology relationships at these phylogenetic distances are often complicated and do not fit standard models. The rule-base presented in this work is an evolving body, and a primary motivation of the work described in this chapter is to serve as a guide for the expansion of the rule-base to better describe these difficult to interpret multi-genome cross-phyla comparisons.

Furthermore, this comparison consists of three species with genomes of vastly different size (in terms of number of genes); the compared species range from a relatively minimal 800 genes to a moderately sized 4000 gene genome to a larger 6300 gene genome. This variation

in size gives insight into the differences in how small and large genomes encode for the fundamentals of life.

Numerous dynamics can cause differences in genome size, and this comparison will assess those forces. Some issues for consideration are:

1. Inparalog expansions as a mechanism for generating larger genomes.
2. Subfunctionalization by domain shuffling or gene splicing.
3. Gene loss as a factor in the size of smaller genomes.
4. Degree of conservation across genomes of varying size.

## **8.2: PATHOGENS USED IN THIS COMPARISON**

All three of the bacterial genomes used in this comparison are human pathogens that pose significant threats to human health and livelihood worldwide. Furthermore, these organisms are all classified by the National Institute of Allergies and Infectious Disease as potential weaponizable bioterrorism agents. Moreover, these organisms are all subject of extensive research for new therapies because of lack of effective drugs, drug resistance or toxicity of existing drugs.

### ***Burkholderia pseudomallei*, strain 1710b<sup>74</sup> (Bps)**

The bacterium *B. pseudomallei* (*Bps*) is the causative agent of melioidosis, a disease which causes lung abscesses that can result in a range of outcomes from mild bronchitis to severe pneumonia in humans and other animals. The disease is spread as a result of direct contact between a potential host and contaminated water or soil. Melioidosis is most prevalent in Southeast Asia and northern Australia.

The *Bps* genome is approximately 7.2 million base pairs and the current gene prediction lists 6,347 genes organized in two chromosomes.

***Rickettsia prowazekii*, strain Madrid E<sup>75</sup> (*Rpr*)**

*R. prowazekii* (*Rpr*) is a gram-positive intracellular parasite that causes endemic typhus, a disease that results in chills, cough, severe head and muscle pain, as well as high fevers and general stupor. The disease is most commonly transmitted to humans *via* the bite of an infected louse. Though endemic typhus has been found world wide, it is most prevalent in Central and South America, Northern China and the Himalaya regions.

The genome of *Rpr* is approximately 1.1 million base pairs on a circular chromosome, with 834 currently annotated genes.

***Mycobacterium tuberculosis*, strain h37rv<sup>76</sup> (*Mtb*)**

*M. tuberculosis* (*Mtb*) is an intracellular parasite that is typically spread through the air when infected persons sneeze, cough or spit. The disease tuberculosis is characterized by fever, weight loss and chronic cough. Worldwide there are about 14 million chronic cases and 9 million new cases per year; about 2 million deaths per year can be attributed to tuberculosis. While there have been reported incidents of tuberculosis nearly world wide, the disease is most commonly found in the developing world. Resistance of the parasite to antibiotics is a growing problem, as is co-infection in persons who are HIV-positive or are otherwise immunosuppressed due to disease or medication.

The *Mtb* genome is approximately 4 million base pairs long, in one circular chromosome, and contains 3989 protein coding genes.

**8.3: GENES PRESENT IN ALL THREE SPECIES**

There are 391 gene groups present in all three species. Given the ancient divergence<sup>77</sup> of the *Actinobacteria* (*Mtb*) and *Proteobacteria* (*Bps* & *RprI*) phyla, these genes likely encode for

core functionalities present in most bacterial clades. **Figure 8-1** contains Venn diagrams illustrating the distribution of genes in gene groups across the three species.

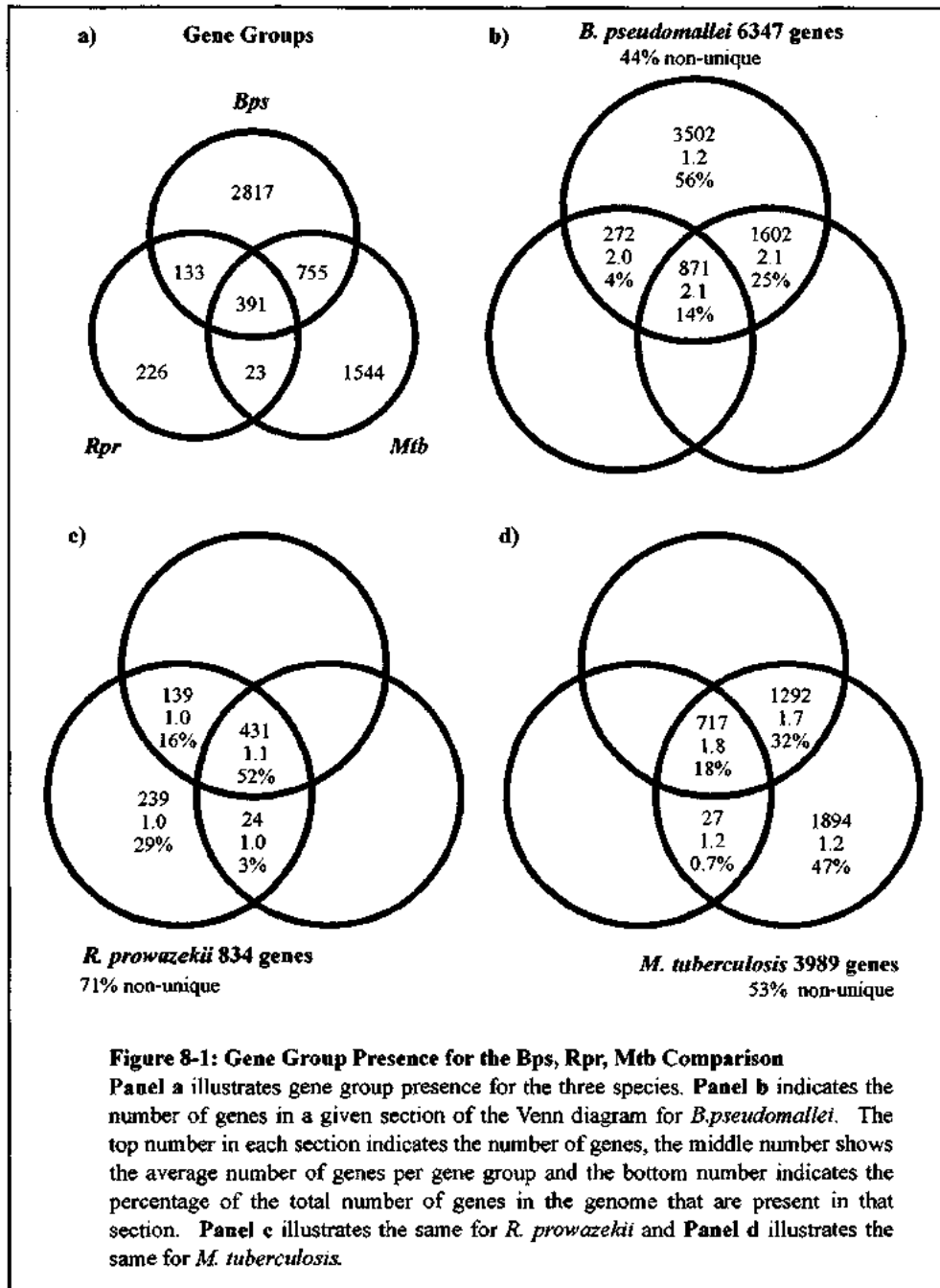
### **Static genes**

Approximately half of these genes (198) are present in one copy in all three of the species and have not undergone any detectable conserved inparalogous expansions; these represent genes that are not only highly conserved, but appear to be highly static across a great deal of evolutionary distance. Of these 198 static conserved genes, 45 encode for a ribosomal protein of some sort.

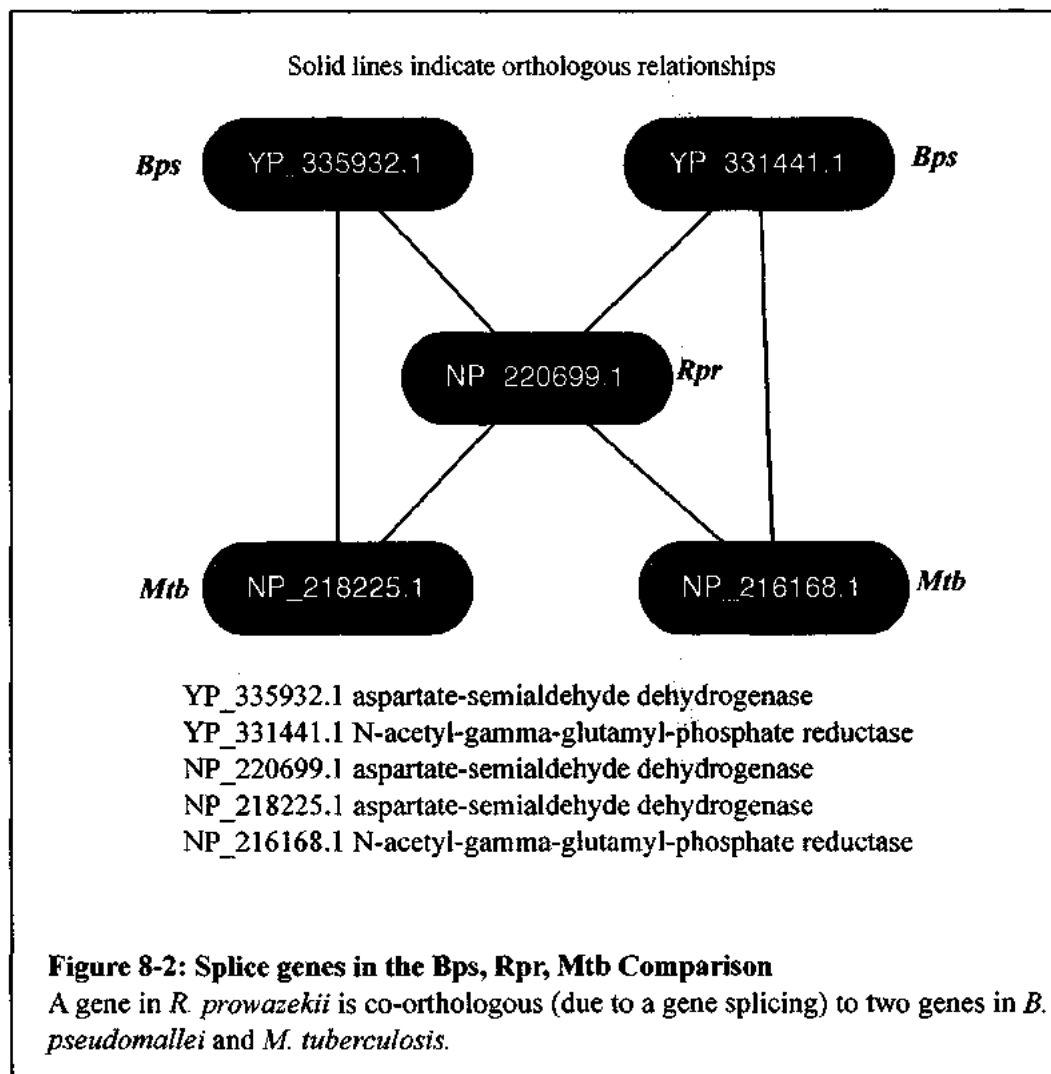
### **Subfunctionalization & Domain Shuffling in *Mtb* & *Bps***

A common theme in these omnipresent gene groups is the presence of fusion and splice events that result in a particular gene in *Rpr* having more than one ortholog in either *Mtb*, *Bps* or both (see **Figure 8-2**). This dynamic occurs in 30% (118) of the 391 fully present gene groups. These do not represent inparalogous post-speciation expansion, but rather gene splice or domain <sup>78</sup> shuffling events that result in more than one gene in either *Mtb* or *Bps* playing the same functional role that is encoded by a single gene in *Rpr*. Presumably the single gene in *Rpr* plays a role in multiple processes and the multiple splices of these genes in *Mtb* and *Bps* have subfunctionalized to serve narrower functional roles.

These 118 apparently subfunctionalized gene groups also appear to be duplicating at a faster rate than other gene groups. Overall, non-novel gene groups in *Bps* tend to contain



approximately two genes (*i.e.* one duplication) per group and non-novel gene groups in *Mtb* tend to contain 1.7 genes per group, by contrast these subfunctionalized groups contain 3.5 and 3.1 genes respectively. Even *Rpr*, which has undergone virtually no inparalog expansion (53 inparalogs, in all gene groups), had 35 inparalogs in the subfunctionalized gene groups. This suggests that even in the minimal content *Rpr* genome the functional roles played by these genes are prone to expansion and perhaps subfunctionalization.



The degree of gene splicing and inparalog expansion these ancient gene groups suggests that they may be particularly dynamic in terms of domain shuffling and gene duplication. This

carries a potential clinical ramification for elucidating drug targets. While these sort of ancient, omnipresent genes are useful for the development of broad-spectrum therapies, these results show that a significant percentage of such genes are subject to a great degree of dynamism and redundancy. This redundancy tends to make these genes poor drug targets as a cell potentially has “back-up” mechanisms for a particular process that a drug may seek to disrupt.

### **Gene Expansions**

Among the 391 fully present gene groups there are 273 that are not undergoing the domain-shuffling phenomena described above and are present as one orthologous copy in each genome; as mentioned earlier, 198 are static and have not undergone any expansions, leaving 75 that have expanded in one or more genome. Fifty-one of these gene groups have expanded in *Bps*, resulting in 158 copies and 35 have expanded in *Mtb* resulting in 111 copies. Whereas the prior section illustrated the effect of subfunctionalization by splicing of ancient gene families (and their subsequent expansion) on the evolution of the bacterial larger genomes, this data illustrates the effect of gene duplications on ancient genes. As mentioned in the prior subsection, the rate of inparalog expansion in *Bps* and *Mtb* for non-spliced fully present genes is markedly lower than the rate of inparalog expansion for spliced (relative to *Rpr*) fully present genes. This suggests that these non-spliced genes are relatively more stable than their spliced counterparts; again, this points to the possibility that gene splicing plays a critical role in the expansion of highly dynamic genes.

## **8.4: GENES GROUPS THAT ARE PRESENT IN TWO SPECIES**

### **Genes Present in *Bps* & *Mtb***

There are 1146 gene groups present in both *Bps* and *Mtb*, these represent 2473 genes (38% of total) in *Bps* and 2009 genes (50% of total genes) in *Mtb*. Of these 1146 shared gene groups

755 (65%) are absent in *Rpr* and hence specific to these two species. Given that *Rpr* diverged from *Bps* more recently than *Mtb* these 755 gene groups likely represent ancient gene groups that have been lost in *Rpr*. Given the relatively compact nature of the *Rpr* genome, it appears that extensive loss of ancient widely conserved genes has been a shaping factor in the evolution of this lineage. Furthermore, of those 755 gene groups, nearly 8% show two or more splice genes in *Bps* that show orthology to a single gene in *Mtb*. In contrast, the inverse situation (spliced genes in *Mtb*) occurs only half as often. This data further suggests that splice events play a significant role in the generation of new genes in larger bacterial genomes.

#### **Genes Present in *Bps* & *Rpr***

There are 524 gene groups present in both *Bps* and *Rpr*; these have only expanded to 540 genes in *Rpr* (65% of total genes) and have expanded extensively in *Bps* to 1143 (18% of total genes). Of these 524 gene groups 133 (25%) are absent in *Mtb*; these genes likely represent newly acquired genes in the *Proteobacteria* lineage or ancient genes lost in the *Actinobacteria* lineage.

#### **Genes Present in *Rpr* & *Mtb***

There are 412 gene groups present in *Rpr* and *Mtb*; these have expanded to only 455 genes (55% of total) in *Rpr* and have expanded to 743 (19% of total) genes in *Mtb*. Of these 412 gene groups only 24 (6%) are absent in *Bps*. These genes likely represent ancient genes that have been lost in *Bps*. Absent an outgroup for comparison, is difficult to discern whether the relatively small number of genes that have been lost in *Bps* connotes that *Bps* is not losing ancestral genes at very significant rate, or that *Rpr* has lost so many ancestrally present genes that we cannot accurately discern *Bps* gene loss. Further comparison to another *Proteobacteria* species (perhaps one with a larger number of genes) would further illuminate which explanation is more likely true.



### **8.5: SPECIES-SPECIFIC GENE GROUPS**

*Bps* has 3502 species-specific genes in 2817 gene groups; 2511 of those are truly novel genes that share no sequence similarity to any other gene in *Bps*. *Rpr* has 239 species-specific genes in 226 gene groups; 219 of those are completely novel. *Mtb* has 1894 species-specific genes in 1544 gene groups; 1364 of those are completely novel. See **Figure 8-1** for more details.

Quite notably, the species-specific genes are less prone to duplications than genes that are conserved across the species. The high degree of degree of duplication in genes present in two or more species (nearly 50% have duplicated once), stands in contrast to the relative scarcity of large species-specific gene families. Furthermore, the slower rate of duplication in species-specific genes is even more notable given the phylogenetic distance between the three species.

### **8.6: UNCLASSIFIABLE GENES**

There were 132 distinct genes in this comparison, plus 55 exact copies of those genes, for which we could not assign GHO terms using rule base. In nearly all the cases (126 of the 132) the inability of our system to classify these genes was a result of a domain-level match between the unclassifiable gene and a within-species gene (internal homolog); that internal homolog in turn had another domain-level match to gene from a different species (an external homolog). The unclassifiable gene did not qualify as an inparalog, because it did not share any sequence similarity to the external homolog (a prerequisite for the inparalog rules) and it did not qualify as part of a species-specific gene group because it had sequence similarity to an internal homolog that was not species specific.

From the point of view of rule-based classification, this scenario is not particularly difficult to classify. Indeed, the rule-base has already been amended and subsequent versions of the software will accommodate these types of relationships and place them into the appropriate gene groups. However, the ability to properly group a gene is somewhat distinct from the ability to accurately describe its relationship to other genes in its group by an ontology term; this issue does point to an opportunity for further development of the Gene Homology Ontology (GHO). The GHO does, to some degree, already accommodate this by describing fusion and splice orthologs; nonetheless, this comparison illustrates that further work is required to fully accommodate the types of domain-level matches that may occur.

The remaining unclassifiable six genes that did not fall into the above category were genes that contained a number of low-quality sequence comparison matches that lead to difficulty in interpreting relationships. As we discussed in Section 7.4 our knowledge representation schema and our rule-based methodologies facilitate user intervention in any of the steps. A researcher interested in performing a highly accurate comparison of these three species could manually assess these six genes and assign them appropriate GHO terms and/or place them into the appropriate gene group.

### **8.7: CROSS-PHYLA COMPARISON CONCLUSIONS**

The results in this chapter illustrate that the comparative genomics pipeline is capable of accurately cataloging sequence comparison experiments that compare distantly related species. The principles of inparalogy, outparalogy and orthology are as relevant in describing these types of distant comparisons as they are towards describing the results of comparisons among very similar species. This study did show that the rule-base and ontology require further refinement in order to accurately describe relationships among genes that contain domain-level matches. Although this does suggest additional work, the overall nature of the comparative genomics pipeline is highly amenable to refinement and extension. Without

question, a primary motivation behind choosing ontologies and rule-based programming was to accommodate lessons learned as the pipeline is used for more comparative genomics studies.

The conclusions were in some ways quite expected and in other ways very interesting. We found a relatively small number of genes that are conserved across all three species; this is not unexpected, due to the small number of genes in *Rpr* and the degree of divergence across the species. We also found a high number of species-specific gene groups and novel genes; which again is not particularly unusual given the phylogenetic distance spanned in our comparison.

An interesting outcome of our comparison is the degree to which universally conserved genes across the three species were prone to splice events. Nearly a third of the gene groups that were fully present across all three species had undergone splice events in *Bps*, *Mtb* or both; furthermore 10% of all multi-species gene groups contained at least one apparent splice event. This comparison suggests that bacterial genomes with fewer genes (in this case *Rpr*) often encode for multiple functional roles with one gene, whereas genomes with high gene counts encode for the same functions with multiple genes. Furthermore, the types of genes that are prone to this splice phenomenon are also more likely to undergo post-speciation inparalog expansion. This is true even in the relatively minimal *Rpr* genome. Further comparisons across disparate (in terms of gene count) genomes could further illuminate the degree to which this phenomenon generalizes.

Another particularly surprising result is that the rate of inparalog duplication is considerably higher in ancient, conserved genes than in the more recent, species-specific genes. Undoubtedly, most of these genes are not truly species-specific; given the phylogenetic distance at which this comparison was conducted these genes are more likely lineage-

specific. The earlier comparative study on four *Leishmania* species (**Chapter 7**) indicated that the lineage-specific genes were more dynamic than the more ancient genes. For unknown reasons this pattern is reversed in this comparison, suggesting that a degree of dissimilarity in genome dynamics across the two lineages.

Overall this study indicates that the rule-base and ontology act a strong foundation for a variety of comparative genomics experiments, including the assessment of evolutionary patterns across great phylogenetic distances. This study does point to the opportunity for further refinements to our comparative genomics pipeline; however, the system is well suited to accommodate those changes.



## **CHAPTER 9: CONCLUSIONS AND FUTURE WORK**

## **9.1: PROJECT GOALS**

### ***Driving factors behind this project***

The motivation behind this work is to make the results of comparative genomics experiments more relevant and understandable. The increasing availability of high-throughput biological data has been well documented and is unlikely to slow in the coming years. Researchers have responded (with varying degrees of adoption and effectiveness) to this growth by generating standards for representing quantitative data, ontologies for unambiguously describing biological features and database schemas for standardizing data management. Despite the rise and growing adoption of these technologies, one central question lacked a succinct answer: “How are these two genomes different?”

Numerous clustering and sequence analysis tools that can provide approximate answers regarding gene content differences; however, technologies typically provide general answers and mask the true complexity of the question. They excel at determining how many genes in one genome are related to how many other genes in another genome. While such analyses are useful, they typically fail to paint a full picture of how any particular gene is related to any other gene and do not address what evolutionary forces might have caused gene content differences across genomes.

This dissertation is an example of how relatively simple comparative genomics tasks can gain complexity and lead to larger, more generalizable results. Although there are many exotic and potentially world-changing implications of comparative genomic research, this project was initially inspired by a relatively modest question: how to compare gene content in a draft version of a sequenced genome to a finished version of a related sequenced genome. As most researchers know, no genome is ever “finished”, but the scientific community

typically comes to an understanding about which sequenced genomes are accurate and nearly complete. Such finished genomes provide a useful template for automated annotation of less-finished genomes. This process typically involves some manner of pairwise genome comparison, the results of which are not straightforward.

From that relatively modest problem, the Gene Homology Ontology was born. The original aim of the GHO was to attach relationship information to genes and then determine if the relationships were logical. Instead of simply clustering genes from a draft genome with genes from a related finished genome, the GHO would classify the relationships between the genomes and ask questions as to the content: were multiple copies of a gene in one genome the result of inparalog duplication or poor gene predictions? Did it make sense from a phylogenetic perspective that two genomes contained a given gene, whereas the third did not? Were there unexpected patterns, such as an abundance of fusion genes, across two genomes?

While this project was initiated to make comparisons on unfinished genomes, it became apparent that the principles associated with that task applied to numerous comparative genomic problems. The work presented here is highly valuable when applied towards comparison of finished genomes and constitutes a significant leap in the clarity of the description of comparative genomics results. This represents an important advance, considering the range of applications for comparative genomics studies.

### ***Evolution of the project***

The initial conception of the GHO quickly led to the question of how to assign the terms to the results of sequence comparison results. Rule-based programming seemed a natural fit for this task. By creating rules, instead of the procedures associated with most programming languages, the assignment process could mimic the evolutionary logic associated with a



particular GHO term. Specifying logic instead of procedure imitates the way a researcher might think about genome evolution, thereby making the assignments more meaningful.

Next, the question arose of how best to translate sequence comparison results into a form amenable to processing by production rules. Over several iterations the fact-based knowledge representation schema arose as an easy, yet powerful and extensible, means of representing both gene predictions and sequence comparison data.

Certain patterns presented themselves when parsing large-scale sequence comparison data and those patterns led to the creation of the Pairwise Genome Comparison Ontology (PGCO). The PGCO served as a key tool in simplifying the GHO assignments by encapsulating several aspects of sequence comparison matches into one term.

The need then arose to aggregate the numerous pairwise homology assignments into a form that was readily interpretable at the genome scale; this prompted the development of the *rule-based semantic grouping* technology. Finally, the development of *logical gene group querying* allowed us to interrogate the gene groups to extract maximum information from our comparisons.

## **9.2: CONTRIBUTIONS OF THIS WORK**

### ***A genomic knowledge representation syntax***

This dissertation presents a lightweight, fact-based knowledge representation schema for representing genomic and sequence comparison data. The representation does not depend on complex file formats, instead focusing on making assertions about entities such as genes and BLAST hits. This focus on facts instead of format allows for easy representation of any type of character or numeric data. Furthermore, the fact-based system is highly flexible; users can specify as much or as little about an entity as is appropriate or available.

### ***The Pairwise Genome Comparison Ontology (PGCO)***

This dissertation details the creation of a novel ontology to describe the qualities inherent to a sequence comparison match, based on the concepts of quality, reciprocity and internality *versus* externality. Researchers dealing with large amounts of sequence comparison results need to categorize sequence to sequence matches based on their context relative to the search as a whole. For instance, is the match the highest quality match between an individual sequence and the body of sequences to which it is compared? Is the match the highest quality match for both sequences involved, just one of them, or neither? Is the match internal to some subset of the sequences (*i.e.* a genes within a genome or group) or does the match join sequences from two subsets (*i.e.* genes across genomes or groups)?

The PGCO contains terms that unambiguously describe the answers to these issues. Assigning PGCO terms to a match facilitates the downstream GHO term assignments. For instance, a widely used definition for orthologs is reciprocal best hits across species. The PGCO encodes for the concept of reciprocal best hits, and therefore can serve as a starting point for creating orthology assignments. Similarly, patterns seen in sequence comparison can serve as starting points for more complicated homology definitions, as we have shown in this work.

### ***The Gene Homology Ontology (GHO)***

This dissertation presents an ontology that describes homology relationships between pairs of genes. Comparative genomics researchers typically use terms such as inparalog and outparalog to describe homology relationships, however the GHO expands those definitions to describe homology in more specific terms. For instance, the ontology describes whether genes are related within or across genomes; for inparalog relationships the ontology describes which gene has retained sequence similarity and which has drifted. Furthermore, it explores

concepts such as splice and fusion genes, as well as apparent (but not true) inparalogs that are created due to gene loss events.

### ***Rule-based classification***

The next contribution of this project is a rule-base that assigns PGCO terms to sequence comparison matches and then uses those terms to assign GHO terms to pairs of genes. The rule-based system attempts to model human logic and make assignments in a manner that is easily interpretable. Furthermore, the rule-base integrates gene positional conservation across genomes (as a surrogate measure that implies gene synteny) to make the GHO assignments. The use of positional conservation information illustrates a strategy by which we can write rules that integrate additional information (besides sequence comparison data) to make more accurate homology term assignments.

### ***Semantic gene grouping***

Using the GHO assignments the *semantic gene grouping* methodology creates clusters by joining genes that have certain homology relationship properties. This is similar to most existing gene clustering technologies, but does not employ statistical or graph-theory technologies to cluster, instead using gene-to-gene relationships as the joining criteria. Grouping genes by this strategy allows for the creation of groups of genes that have defined relationship types.

A primary benefit of this strategy is that it avoids the problem many clustering technologies have of simply outputting large groups of genes with no easy way to interpret logic behind why those genes were placed in the same cluster. This technology, *via* facts that can be queried by rules, specifies the GHO relationships between every gene in a group. Another benefit of this strategy is that clustering technologies often have to decide how to handle certain relationships, such as whether to join closely related outparalogs, or whether to join

spliced genes into the same cluster. *Semantic gene grouping* unambiguously describes those situations as opposed to deciding on one way or the other. By describing, instead of deciding, our technology allows users to pose queries on the gene groups to find these potentially ambiguous scenarios. Once found, the researchers can then apply their own knowledge to decode the meaning behind such situations.

#### ***Logical gene group querying***

The fact-based representations and semantic gene groupings allow for the complex querying of comparative genomics results. A user can pose simple queries to determine which genes are present in which genomes; however the true power of the system lies in the ability to pose more complicated and interesting questions. A researcher can, for instance, determine where in a lineage a particular gene arose, and whether that gene is expanding differentially in certain branches of the lineage. Such questions can provide insights into the underlying functioning of the compared organisms

*Logical gene group querying* can also serve as a knowledge discovery tool. While a user can pose obvious or intuitive queries, they can also search for gene groups that do not behave in a logical or expected manner. Such gene groups can provide insights into unexpected evolutionary dynamics or provide suggestions as to inconsistencies in gene prediction results.

#### ***A detailed understanding of genome dynamism in Leishmania spp.***

The above technologies aided greatly in better describing the results of the comparative genomics study of four *Leishmania* species. Prior work in this field has focused on presence or absence of a particular gene, but this work has expanded those analyses to a more exact accounting of instances of gene gain, gene loss and gene expansion across the four species.

This work has elucidated several novel insights into the evolution of the *Leishmania* genus. The rule-based queries have uncovered a greater dynamism (gene expansion and gene loss) in the *Leishmania* (*Leishmania*) sub-genus specific genes than in the more ancient genes that were inherited from the last common ancestor (LCA) of all four species. Furthermore, *L. infantum* is losing LCA genes at rate much higher than expected, perhaps explaining the genetic underpinnings of *L. infantum*'s unique clinical manifestations. The rule-based queries have shown that gene duplication and subsequent differential loss play a significant role in *Leishmania* evolution; this paints a more complicated picture of orthologous relationships across the species.

Certain unexpected gene patterns in the lineage also illustrate potential areas for improving the gene predictions. Unusual patterns of gene loss often indicate “missed” gene predictions, and the results that this system have generated provide clues to the *Leishmania* gene annotators who constantly strive to provide the research community with improved data.

#### ***Cross-phyla bacterial pathogen comparison***

This analysis uncovered evolutionary trends that explain differences across distantly related bacterial genomes. The results indicate that domain shuffling and gene mosaics are particularly prevalent; these phenomena most notably manifested themselves in the presence of apparently multifunctional genes in the small genome of *Rickettsia prowazekii* that are orthologous to two or more genes in the larger *Mycobacterium tuberculosis* and *Burkholderia pseudomallei* genomes. Furthermore, we this subset of genes was particularly prone to inparalogous expansion. This indicates that a relatively small group of genes are particularly dynamic and account for a disproportionate amount of the genes “added” to the two larger genomes in our comparison.

This comparison of distantly related species illuminated areas where the ontology and rule-base have opportunities to broaden to accommodate new relationship types. As mentioned above, this comparison resulted in many fusion/splice gene relationships, and those were well addressed by the system as it currently stands. However, there were about 132 genes (out of 11,1170) that had complex domain-level homology relationships that the system could not categorize. A primary goal throughout this entire project has been that of extensibility and the system as it now stands is well suited for accommodating these and other changes.

### **9.3: FUTURE DIRECTIONS**

The rule-based strategy presented in this dissertation is structured for expanded functionality for a number of reasons. First, the knowledge representation syntax is flexible and can accurately represent any sort of numerical or text string data. Second, the rule-based methodologies are extendable to accommodate new types of analyses. Third, the overall framework lends itself to flexible pipelining strategies because the inputs to and outputs from the various components of the system are represented as structured facts. Finally, the use of ontology terms to represent comparative genomics results facilitates sharing of data, allowing different research groups to extend and refine the results of analyses done by collaborators or colleagues.

The remainder of this section lists areas of opportunity for the expansion of this work:

#### ***Accommodating additional sequence comparison algorithms***

As mentioned in the discussion of the *Leishmania* comparative genomics data, it is often difficult to discern between gene gains in a particular lineage and gene losses in a parallel lineage. Additional types of sequence comparison experiments can yield clues that can help discern between the two scenarios. For instance, comparing a set of genes from one species to a set of genome sequences from another species can often yield evidence of regions that at

one time encoded for genes, but have degenerated to a point where they no longer do so; this is one type of evidence for gene loss. The system, as it currently stands, does not integrate this type of data, however the structure supports the addition of such analyses.

#### ***Using rule-based homology to refine gene prediction***

Determining genes in a newly assembled genome often involves integrating the results from a number of gene prediction methodologies to form a consensus which, presumably, is more accurate than the results of any individual gene prediction. Forming such consensus often requires developing some sort of weighting system to account for the various strengths and weaknesses of different gene prediction methodologies. A potential use of the homology annotation pipeline would be to compare the results of several different prediction methods. For example, the *Leishmania* comparative genomic study (Chapter 7) found three genes that are present in *LbrM* and *LmjF*, but absent in *LinJ* and *LmxM*. This pattern implies that *LinJ* and *LmxM* both lost the genes in independent events. While this is not impossible (and may have some important biological significance) such independent loss events are somewhat unusual. Feeding the original gene prediction data into our rule-based system could afford the ability to further assess whether such suspicious patterns are likely true or the result of some misinterpreted gene prediction results. The possibility exists that one of the gene prediction algorithms predicted that *LinJ* or *LbrM* contained the gene, but was overridden by other gene prediction algorithms. In such a situation a genome annotator might surmise that the gene does exist, but was not annotated due to the consensus nature of gene prediction methodologies. This is one example of the myriad difficulties with gene prediction that the work presented here can help solve.

#### ***Detection of horizontal gene transfer***

The first version of the rule-based system did not address the phenomenon of horizontal gene transfer - the acquisition of genetic material from a non-parent species, usually a bacterium.

Future extensions to the ontology and rule-base could allow a researcher to include sequence comparisons against groups of bacterial species to highlight potential horizontal gene transfer events. Furthermore, other types of data such as codon frequencies or GC content could be encoded in our fact-based knowledge representation syntax; such data would further refine the search for potential horizontal gene transfer events.

### ***Expansion of the Gene Homology Ontology***

The Gene Homology Ontology as it currently stands specifies a level of granularity greater than most researchers commonly use. Nonetheless, as comparative genomics moves forward and the research community better understands the complexities behind genome dynamism, the ontology could conceivably benefit from further, more granular terms. The ontology and the rule-base are evolving entities that are suited to expansion along with the community's inevitably expanding understanding of comparative genomics. The GHO can easily accommodate terms associated with horizontal gene transfer, syntenic orthology and any number of additional comparative genomics topics.

### ***Use of the knowledge representation syntax as a data exchange format***

As an increasing number of scientists leverage genomic sequence, annotation and functional data generated by individuals outside of their own research groups pragmatic issues surrounding data exchange are becoming increasingly relevant. A particularly salient issue is that of flexibility in representing data: quite simply, two labs rarely agree on how best to structure a given piece of data. Several file formats (for example GFF3 and Genbank format) have been developed to provide this sort of flexibility. Such file formats lend themselves to errors due to their complex structure and the presence of exceptional data types that they are not suited to describe. A fact-based representation ameliorates some of these issues. For instance, a fact-based approach does not have complicated structure; researchers can add new attributes to a piece of data without attending to such minutia as where exactly in the file the



data belongs and what sort of characters should separate data. Furthermore, should two related research groups choose to represent their data differently, they have a recourse (rule-based translation of the data) for rectifying those differences. Currently, rule-based programming knowledge is not common in the world of genomics and bioinformatics. Nevertheless, many of the advantages offered by rule-based programming and fact-based knowledge representation can provide great benefit to the genome research community.

### **9.5: CONCLUDING REMARKS**

The genomics community is awash with high-throughput data; most researchers who work in the field will agree that greater access to data and improved methodologies for integrating heterogeneous data is a critical step in the progression of the field. With that in consideration, the community must also acknowledge that data sets are growing to a size and level of complexity that makes human comprehension impossible. No doubt, most genome researchers would agree that many potentially groundbreaking advances are unrealized because of the inability to fully extract meaning from high-throughput experiments that have already been performed.

This semantically meaningful, rule-driven comparative genomics pipeline serves as a foundation for further integrating various data sources and analytic results. The pipeline is suited for growth and can adapt to changing research goals. However, apart from the technical aspects of this implementation, the fundamental strategies and principles presented in this work serve as a blueprint for future development of technologies that bring greater meaning to complex high-throughput data.

**APPENDIX A: PAIRWISE GENOME COMPARISON  
ONTOLOGY TERMS**

**Table A-1: Pairwise Genome Comparison Ontology (PGCO) Terms**

This table defines each PGCO term using a hypothetical sequence comparison match between two entities (X) and (Y). The "Best Hit" columns define whether the term refers to the highest quality match in a given direction (X to Y) or (Y to X); terms with the letter "O" in these columns can refer to either a highest quality match in that, or some lesser quality match. The "Internal/External" column denotes whether this term refers to a match internal to some genome or other group, or between members of different genomes or groups; the letter "O" denotes that the term can belong to either.

PGCO Term	Level	Definition	Best Hit X to Y	Best Hit Y to X	Internal/External
Match	1	Any match	O	O	O
Self match	1	Match such that X and Y are the same entity	N	N	O
Non-self match	1	Match such that X and Y are different entities	O	O	O
Best match	2	Match such that X and Y are different entities and there is no higher quality match for X	Y	O	O
External match	2	Match such that X and Y belong to different groups	O	O	Ext.
Internal match	2	Match such that X and Y belong to the same group	O	O	Int.
Secondary match	2	Match such that X has some other higher quality match	N	O	O
Reciprocal best match	3	Match such that neither X nor Y has a higher quality match	Y	Y	O
Unidirectional best match	3	Match such that X has no higher quality match and Y has a higher quality match	Y	N	O
External best match	3	Match such that X has no higher quality match and X and Y belong to different groups	Y	O	Ext.
Internal best match	3	Match such that X has no higher quality match and X and Y belong to the same group	Y	O	Int.
External secondary match	3	Match such that X has a higher quality match and X and Y belong to different groups	N	O	Ext.
Internal secondary match	3	Match such that X has a higher quality match and X and Y belong to the same group	N	O	Int.
Proximate secondary match	3	Match such that X has a higher quality match and Y has no higher quality match	N	Y	O
Intermediate secondary match	3	Match such that both X and Y have higher quality matches	N	N	O

<b>PGCO Term</b>	<b>Level</b>	<b>Definition</b>	<b>Best Hit X to Y</b>	<b>Best Hit Y to X</b>	<b>Internal/ External</b>
Outlying secondary match	3	Match such that X has a higher quality match and the quality is sufficiently low that Y does not match X in the reverse comparison	N	N	O
External reciprocal best match	4	Match such that neither X nor Y has a higher quality match and X and Y belong to different groups	Y	Y	Ext.
External unidirectional best match	4	Match such that X has no higher quality match and Y has a higher quality match and X and Y belong to different groups	Y	N	Ext.
Internal reciprocal best match	4	Match such that neither X nor Y has a higher quality match and X and Y belong the same group	Y	Y	Int.
Internal unidirectional best match	4	Match such that X has no higher quality match and Y has a higher quality match and X and Y belong to different groups	Y	N	Int.
External proximate secondary match	4	Match such that X has a higher quality match and Y has no higher quality match and X and Y belong to different groups	N	Y	Ext.
External intermediate secondary match	4	Match such that X has no higher quality match and Y has no higher quality match and X and Y belong to different groups	N	N	Ext.
External outlying secondary match	4	Match such that X has a higher quality match and the quality is sufficiently low that Y does not match X in the reverse comparison and X and Y are in different groups	N	N	Ext.
Internal proximate secondary match	4	Match such that X has a higher quality match and Y has no higher quality match and X and Y belong to the same group	N	Y	Int.
Internal intermediate secondary match	4	Match such that X has no higher quality match and Y has no higher quality match and X and Y belong the same group	N	N	Int.
Internal outlying secondary match	4	Match such that X has a higher quality match and the quality is sufficiently low that Y does not match X in the reverse comparison and X and Y are in the same group	N	N	Int.



**APPENDIX B: GENE HOMOLOGY ONTOLOGY TERMS**

**Table B-1: Gene Homology Ontology (GHO) Terms**

This table defines GHO terms relative to some hypothetical gene (X). The use of the word “relation” and “related” in the following definitions refers to “evolutionary relationship”. In practice the closeness of evolutionary relatedness is measured by some scoring metric associated with a sequence comparisons algorithm.

<b>Term</b>	<b>Definition</b>
Homolog	A gene that shares an evolutionary history with gene X.
Closest homolog	The gene in a particular comparison that is most closely related to gene X.
Internal homolog	A gene that shares an evolutionary history with gene X and is from the same genome as gene X.
External homolog	A gene that shares an evolutionary history with gene X and is from a different genome as gene X.
Closest internal homolog	The most closely related internal homolog to gene X.
Closest external homolog	The most closely related external homolog to gene X.
Ortholog	A gene that arose from the same ancestral gene as gene X and diverged from gene X as the result of a speciation event.
Fusion	An ortholog (Y) of gene X such that gene Y is also an ortholog to some other gene (X2) in the same genome as X. The region of similarity between gene Y and gene X is non-overlapping to the region of similarity between gene Y and gene X2.
Splice	An ortholog (Y) of gene X such that gene X is also an ortholog of some other gene (Y2) in the same genome as gene Y. The region of similarity between gene X and gene Y is non-overlapping to the region of similarity between gene X and gene Y2.
Inparalog	A gene related to gene X via a gene duplication event that occurred subsequent to the speciation of the compared species.
Pseudo-inparalog	
Internal inparalog	An inparalog that is in the same genome as gene X. An internal inparalog may have duplicated from gene X, or may have duplicated from the same gene from which X duplicated.
Internal parent inparalog	An internal inparalog (X2) to gene X that it has maintained greater functional similarity to the original ancestral gene than gene X or any of the other internal inparalogs that gene X and gene X2 share.
Internal child inparalog	An internal inparalog (X2) to gene X such that gene X has maintained greater functional similarity to the original ancestral gene than gene X2 or any of the other internal inparalogs that gene X and gene X2 share.

**Table B-1: Gene Homology Ontology (GHO) Terms (continued from previous page)**

<b>Term</b>	<b>Definition</b>
Internal sibling inparalog	An internal inparalog (X2) to gene X such that some third internal inparalog (X3) has maintained greater functional similarity to the original ancestral gene than gene X, gene X2, or any of the other internal inparalogs that X, X2 and X3 share.
External inparalog	An gene that is an internal inparalog of the ortholog of gene X; also a gene that is the ortholog of the internal parent inparalog of gene X.
External parent inparalog	A gene that is the ortholog of the internal parent inparalog of gene X.
External child inparalog	A gene that is the internal child inparalog of ortholog of gene X.
External sibling inparalog	A gene that is the internal child inparalog of the external parent inparalog of gene X.
Outparalog	A gene that is related to X via a gene duplication event that occurred prior the the speciation event that separated the compared genomes.
Internal outparalog	An outparalog that is from the same genome as gene X.
External outparalog	An outparalog that is from a different genome as gene X.





**APPENDIX C: RESULTS OF PAIRWISE LEISHMANIA  
COMPARISONS**

The following tables (C-1 to C-6) show the results of the individual pairwise *Leishmania* sequence comparisons that as an aggregate were used in the multiway comparison in **Chapter 7**. The results listed below are for *distinct* genes, meaning that multiple *identical* copies of a gene are represented as a single gene.

<b>Table C-1: Results of the <i>LmjF</i> / <i>LmxM</i> Comparison</b>		
	<b>LmjF</b>	<b>LmxM</b>
Orthologs	7686	7688
External Fusions	14	6
External Splices	13	28
Novel Genes	135	44
Species-specific Inparalogs	4	4
Inparalogs	300	269
Inparalogs/Possible pseudo-Inparalogs	70	135
<b>SUM</b>	<b>8195</b>	<b>8140</b>
<b>Table C-2: Results of the <i>LinJ</i> / <i>LbrM</i> Comparison</b>		
	<b>LinJ</b>	<b>LbrM</b>
Orthologs	7428	7412
External Fusions	26	36
External Splices	72	53
Novel Genes	304	99
Species-Specific Inparalogs	17	66
Inparalogs	260	368
Inparalogs/Possible Pseudo-Inparalogs	149	146
<b>SUM</b>	<b>8158</b>	<b>8091</b>

	LbrM	LmxM
Orthologs	7301	7287
External Fusions	13	33
External Splices	68	30
Novel Genes	165	304
Species-Specific Inparalogs	72	18
Inparalogs	388	345
Inparalogs/Possible Pseudo-Inparalogs	165	186
SUM	8091	8140

	LmjF	LinJ
Orthologs	7837	7860
External Fusions	27	2
External Splices	4	54
Novel Genes	53	39
Species-Specific Inparalogs	0	6
Inparalogs	236	138
Inparalogs/Possible Pseudo-Inparalogs	69	115
SUM	8195	8158

	LinJ	LmxM
Orthologs	7683	7663
External Fusions	13	33
External Splices	65	27
Novel Genes	141	72
Species-Specific Inparalogs	4	0
Inparalogs	223	278
Inparalogs/Possible Pseudo-Inparalogs	107	127
SUM	8158	8140

<b>Table C-6: Results of the <i>LmjF</i> / <i>LbrM</i> Comparison</b>		
	<b>LmjF</b>	<b>LbrM</b>
<b>Orthologs</b>	<b>7417</b>	<b>7430</b>
<b>External Fusions</b>	<b>32</b>	<b>13</b>
<b>External Splices</b>	<b>26</b>	<b>66</b>
<b>Novel Genes</b>	<b>300</b>	<b>80</b>
<b>Species-Specific Inparalogs</b>	<b>13</b>	<b>64</b>
<b>Inparalogs</b>	<b>337</b>	<b>362</b>
<b>Inparalogs/Possible Pseudo-Inparalogs</b>	<b>128</b>	<b>155</b>
<b>SUM</b>	<b>8195</b>	<b>8091</b>

## BIBLIOGRAPHY

1. Stoeckert, C. J. J. et al. PlasmoDB v5: new looks, new genomes. *Trends Parasitol* **22**, 543-546 (2006).
2. Cadag, E. Automated learning of protein involvement in pathogenesis using integrated queries. Doctoral dissertation, Biomedical and Health Informatics, University of Washington. (2009).
3. Zhang, R. & Zhang, C. T. The impact of comparative genomics on infectious disease research. *Microbes Infect* **8**, 1613-1622 (2006).
4. The Sequence Ontology Project. (2008). March 15, 2009. <[www.sequenceontology.org](http://www.sequenceontology.org)>.
5. Eilbeck, K. et al. The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol* **6**, R44 (2005).
6. Karen Eibeck, S. E. L. Sequence Ontology Annotation Guide. *Comp Funct Genom* **2004**, 642-647 (2004).
7. Mungall, C. J. & Emmert, D. B. A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics* **23**, i337-46 (2007).
8. Bahl, A. et al. PlasmoDB: the Plasmodium genome resource. A database integrating experimental and computational data. *Nucleic Acids Res* **31**, 212-215 (2003).
9. Donelson, L. et al. The BioMediator system as a data integration tool to answer diverse biologic queries. *Medinfo* **11**, 768-772 (2004).
10. Li, L. et al. ApiEST-DB: analyzing clustered EST data of the apicomplexan parasites. *Nucleic Acids Res* **32**, D326-8 (2004).
11. Tatusov, R. L. et al. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003).

12. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science* **278**, 631-637 (1997).
13. Alexeyenko, A., Tamas, I., Liu, G. & Sonnhammer, E. L. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics* **22**, e9-15 (2006).
14. O'Brien, K. P., Remm, M. & Sonnhammer, E. L. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res* **33**, D476-80 (2005).
15. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29 (2000).
16. *Logic Programming with Prolog* (Springer, 2005).
17. Ivens, A. C. et al. The genome of the kinetoplastid parasite, *Leishmania major*. *Science* **309**, 436-442 (2005).
18. Oinn, T. et al. Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* **20**, 3045-3054 (2004).
19. Zhang, J. Evolution by gene duplication: an update. *Trends in Ecology & Evolution* Volume **18**, Issue 6, 292-298 (2003).
20. He, X. & Zhang, J. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* **169**, 1157-1164 (2005).
21. de la Cruz, F. & Davies, J. Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends Microbiol* **8**, 128-133 (2000).
22. Koonin, E. V., Makarova, K. S. & Aravind, L. Horizontal gene transfer in prokaryotes: quantification and classification. *Annu Rev Microbiol* **55**, 709-742 (2001).
23. Li, W. H., Gojobori, T. & Nei, M. Pseudogenes as a paradigm of neutral evolution. *Nature* **292**, 237-239 (1981).
24. Koonin, E. V. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet* **39**, 309-338 (2005).
25. Lewis, S. E. et al. Apollo: a sequence annotation editor. *Genome Biol* **3**, RESEARCH0082 (2002).

26. Lee, D., Redfern, O. & Orengo, C. Predicting protein function from sequence and structure. *Nat Rev Mol Cell Biol* **8**, 995-1005 (2007).
27. Burkholderia pseudomallei 1106b. (2008).
28. Wang, Z., Chen, Y. & Li, Y. A brief review of computational gene prediction methods. *Genomics Proteomics Bioinformatics* **2**, 216-221 (2004).
29. Genomes On-Line Database. Nov 16, 2008. <[www.genomesonline.org](http://www.genomesonline.org)>
30. Teufel, A., Krupp, M., Weinmann, A. & Galle, P. R. Current bioinformatics tools in genomic biomedical research (Review). *Int J Mol Med* **17**, 967-973 (2006).
31. Windsor AJ, & T, M.-O. Comparative genomics as a tool for gene discovery. *Curr Opin Biotechnol.* **17**, 161-167 (2006).
32. Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer & Jinghui Zhang, Z., Webb Miller, and David J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402 (1997).
33. Li, K. B. ClustalW-MPI: ClustalW analysis using distributed and parallel computing. *Bioinformatics* **19**, 1585-1586 (2003).
34. Moreno-Hagelsieb, G. & Latimer, K. Choosing BLAST options for better detection of orthologs as reciprocal best hits. *Bioinformatics* **24**, 319-324 (2008).
35. Pearson, W. R. & Lipman, D. J. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A* **85**, 2444-2448 (1988).
36. Fasta Format Description. March 26, 2008. <[www.ncbi.nlm.nih.gov/blast/fasta.shtml](http://www.ncbi.nlm.nih.gov/blast/fasta.shtml)>
37. Edgar, R. C. Optimizing substitution matrix choice and gap parameters for sequence alignment. *BMC Bioinformatics* **10**, 396 (2009).
38. Frommlet, F., Futschik, A. & Bogdan, M. On the significance of sequence alignments when using multiple scoring matrices. *Bioinformatics* **20**, 881-887 (2004).
39. O'Brien, K. P., Westerlund, I. & Sonnhammer, E. L. OrthoDisease: a database of human disease orthologs. *Hum Mutat* **24**, 112-119 (2004).



40. Wall, D. P. & Deluca, T. Ortholog detection using the reciprocal smallest distance algorithm. *Methods Mol Biol* **396**, 95-110 (2007).
41. Michael Gruninger, J. L. Ontology Applications and Design. *Communications of the ACM* **45**, 39-41 (2002).
42. Brinkley, J. F., Suci, D., Detwiler, L. T., Gennari, J. H. & Rosse, C. A framework for using reference ontologies as a foundation for the semantic web. *AMIA Annu Symp Proc* 96-100 (2006).
43. Weng, C., Gennari, J. H. & Fridsma, D. B. User-centered semantic harmonization: a case study. *J Biomed Inform* **40**, 353-364 (2007).
44. Smith, B. et al. Relations in biomedical ontologies. *Genome Biol* **6**, R46 (2005).
45. Barry Smith, A. K., Thomas Bittner. Basic Formal Ontology for Bioinformatics. *Journal of Information Systems* **2005**,
46. Mappings of External Classification Systems to GO. March 11, 2008. <[www.geneontology.org/GO.indices.shtml](http://www.geneontology.org/GO.indices.shtml)>
47. Murray, H. W., Berman, J. D., Davies, C. R. & Saravia, N. G. Advances in leishmaniasis. *Lancet* **366**, 1561-1577 (2005).
48. Peacock, C. S. et al. Comparative genomic analysis of three *Leishmania* species that cause diverse human disease. *Nat Genet* **39**, 839-847 (2007).
49. Smith, D. F., Peacock, C. S. & Cruz, A. K. Comparative genomics: from genotype to disease phenotype in the leishmaniasis. *Int J Parasitol* **37**, 1173-1186 (2007).
50. Besteiro, S., Williams, R. A., Coombs, G. H. & Mottram, J. C. Protein turnover and differentiation in *Leishmania*. *Int J Parasitol* **37**, 1063-1075 (2007).
51. Van Voorhis, W. C., Hol, W. G., Myler, P. J. & Stewart, L. J. The role of medical structural genomics in discovering new drugs for infectious diseases. *PLoS Comput Biol* **5**, e1000530 (2009).
52. (SSGCID), S. S. G. C. f. I. D. Seattle Structural Genomics Center for Infectious Disease - Home Page. Jan 10, 2010. <[www.ssgcid.org](http://www.ssgcid.org)>

53. NIAID Category A, B & C Priority Pathogens. Dec 6, 2007. <[pathema.tigr.org/pathema/AbcGenomes.shtml](http://pathema.tigr.org/pathema/AbcGenomes.shtml)>.
54. The Homology Ontology. Feb 12, 2010. <[http://bgee.unil.ch/download/homology\\_ontology.obo](http://bgee.unil.ch/download/homology_ontology.obo)>
55. Jensen, B. C., Sivam, D., Kifer, C. T., Myler, P. J. & Parsons, M. Widespread variation in transcript abundance within and across developmental stages of *Trypanosoma brucei*. *BMC Genomics* **10**, 482 (2009).
56. Kikuta, H. et al. Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Res* **17**, 545-555 (2007).
57. Kuzniar, A., van Ham, R. C., Pongor, S. & Leunissen, J. A. The quest for orthologs: finding the corresponding gene across genomes. *Trends Genet* **24**, 539-551 (2008).
58. Wayne P. Maddison, M. J. D. a. D. R. M. Outgroup Analysis and Parsimony. *Systematic Zoology* Vol. **33**, No. 1, 83-103 (1984).
59. Hurles, M. Gene duplication: the genomic trade in spare parts. *PLoS Biol* **2**, E206 (2004).
60. Lespinet, O., Wolf, Y. I., Koonin, E. V. & Aravind, L. The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res* **12**, 1048-1059 (2002).
61. Cannon, S. B., Mitra, A., Baumgarten, A., Young, N. D. & May, G. The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. *BMC Plant Biol* **4**, 10 (2004).
62. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. *Nucleic Acids Res* **37**, D26-31 (2009).
63. Croan, D. G., Morrison, D. A. & Ellis, J. T. Evolution of the genus *Leishmania* revealed by comparison of DNA and RNA polymerase gene sequences. *Mol Biochem Parasitol* **89**, 149-159 (1997).
64. World Health Organization: Leishmaniasis. Jan 10, 2010. <[www.who.int/leishmaniasis/disease\\_epidemiology/en/index.html](http://www.who.int/leishmaniasis/disease_epidemiology/en/index.html)>

65. El-Sayed, N. M. et al. Comparative genomics of trypanosomatid parasitic protozoa. *Science* **309**, 404-409 (2005).
66. Sousa, A. Q. & Pearson, R. Drought, smallpox, and emergence of *Leishmania braziliensis* in northeastern Brazil. *Emerg Infect Dis* **15**, 916-921 (2009).
67. Sanchez-Canete, M. P., Carvalho, L., Perez-Victoria, F. J., Gamarro, F. & Castanys, S. Low plasma membrane expression of the miltefosine transport complex renders *Leishmania braziliensis* refractory to the drug. *Antimicrob Agents Chemother* **53**, 1305-1313 (2009).
68. le Fichoux, Y. et al. Occurrence of *Leishmania infantum* parasitemia in asymptomatic blood donors living in an area of endemicity in southern France. *J Clin Microbiol* **37**, 1953-1957 (1999).
69. Martin-Sanchez, J., Navarro-Mari, J. M., Pasquau-Liano, J., Salomon, O. D. & Morillas-Marquez, F. Visceral leishmaniasis caused by *Leishmania infantum* in a Spanish patient in Argentina: What is the origin of the infection? Case report. *BMC Infect Dis* **4**, 20 (2004).
70. Aslett, M. et al. TriTrypDB: a functional genomic resource for the Trypanosomatidae. *Nucleic Acids Res* **38**, D457-62 (2010).
71. Hertz-Fowler, C. et al. GeneDB: a resource for prokaryotic and eukaryotic organisms. *Nucleic Acids Res* **32**, D339-43 (2004).
72. Li, L., Stoeckert, C. J. J. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**, 2178-2189 (2003).
73. Yang, J., Malik, H. S. & Eickbush, T. H. Identification of the endonuclease domain encoded by R2 and other site-specific, non-long terminal repeat retrotransposable elements. *Proc Natl Acad Sci U S A* **96**, 7847-7852 (1999).
74. Holden, M. T. et al. Genomic plasticity of the causative agent of melioidosis, *Burkholderia pseudomallei*. *Proc Natl Acad Sci U S A* **101**, 14240-14245 (2004).

75. Andersson, S. G. et al. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* **396**, 133-140 (1998).
76. Camus, J. C., Pryor, M. J., Medigue, C. & Cole, S. T. Re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv. *Microbiology* **148**, 2967-2973 (2002).
77. Hugenholtz, P., Goebel, B. M. & Pace, N. R. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J Bacteriol* **180**, 4765-4774 (1998).
78. Gouzy, J., Corpet, F. & Kahn, D. Whole genome protein domain analysis using a new method for domain clustering. *Comput Chem* **23**, 333-340 (1999).

## VITA

Dhileep Sivam completed his Bachelor of Science in Cell and Molecular Biology at the University of Washington. After working at the Corixa Corporation in Seattle for several years he returned to the university, completing a Doctorate in Biomedical and Health Informatics while working in the lab of Dr. Peter Myler at the Seattle Biomedical Research Institute. His research activities have included integration of genomic knowledge from heterogenous data sources, statistical analysis of gene expression data and the representation of large data sets in a manner comprehensible to and usable by bench scientists. His primary research interest and motivation is that of bringing sense and order to the ever-expanding world of genomic data.