Identification of *Leptospira* strains and modeling of interaction between animal and human hosts and the environment using pairwise sequence alignment

M. Mason, Division of Epidemiology and Community Health, School of Public Health, University of Minnesota, Minneapolis, Minnesota

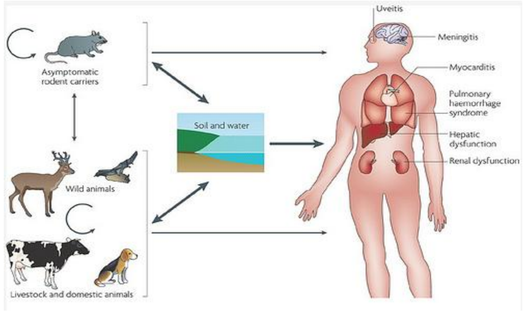
Summary

- Leptospirosis is a bacterial zoonotic pathogen that exists at endemic levels throughout the world and has been classified by the World Health Organization as a reemerging infectious disease.
- The *Leptospira* bacteria are known to be maintained in small mammals, namely rodents, who are asymptomatic carriers. However, the ecology and transmission of the disease remains under-studied.
- One means of mapping the distribution of the pathogen is to determine which serovars are present in various hosts and the environment, and examine similarities in those serovars.
- Genomic sequences of 17 *Leptospira* samples (7 from mice, 9 from water, and 1 reference strain *Borgpetersenii Ballum)* were evaluated using the Clustal W2 tool[1] to determine similarity scores for each pairwise comparison. A dynamic programming algorithm was run to determine the optimal alignment based on mismatches and gaps.
- Single linkage, average linkage, and complete linkage analyses conducted in R,[2] calculated distances between clusters and produced dendrograms. Results were compared to reference *Leptospira* strains in the BLAST database[3] to determine the name of the detected serovars.
- 16 of the 17 samples had matches in the BLAST database. 9 of the samples were of the *Borgpetersenii Javanica* serovar and 7 of the samples were of the *Interrogans Copenhageni* serovar demonstrating similarities of strains across species and the environment.
- Combining sequence alignment with geographic information systems in future research will likely provide a better understanding of the transmission cycle of the *Leptospira* pathogen.

Background

Leptospirosis is a worldwide public health problem caused by pathogenic bacteria of the genus *Leptospira*. This zoonotic pathogen is transmitted directly or indirectly from animals (wild and domestic) to humans. The transmission cycle and environmental factors contributing to human infection with the *Leptospira* pathogen are not fully understood. One aim of the "Eco-epidemiology of leptospirosis in Latin America" project is to formulate a model of transmission of *Leptospira* that incorporates all hosts of the bacteria, as well as interactions with the environment that result in human infection. Animal, human, and water samples are being evaluated for *Leptospira* presence through PCR analysis, and the similarities in *Leptospira* strains across samples highlight possible transmission pathways.

Figure 1. Leptospirosis Transmission and Human Disease [4]



Methods

Pairwise sequence alignment

- Pairwise sequence alignment was performed using dynamic programming in Clustal W2
- Dynamic programming determines the most likely alignment of two sequences through penalizing mismatches and gaps. The fewer mismatches and gaps between the two sequences, the higher the alignment score.
- Similarity scores (percent alignment of the nucleotides) were produced for each possible pairwise comparison through the dynamic programming algorithm, yielding a 17x17 matrix of similarity scores for cluster analysis

Hierarchical cluster analysis[5]

- Cluster analysis is most widely used for informing the creation of dendrograms that visually depict the similarities of genomic sequences
- Similarity scores such as those developed through the Clustal W2 procedures are converted into theoretical distances to further evaluate which samples have the most similar genomic sequences
- Three main hierarchical cluster analysis algorithms exist:
  - Single linkage: examines the minimum theoretical distance between clusters, separating the most dissimilar pairs until no more separations are viable
  - Complete linkage: examines the maximum distance between clusters, indicating high alignment, and begins the dendrogram with the two most similar sequences
  - Average linkage: separates the two most dissimilar sequences first, but then builds up from those using the most similar sequences

Strain Identification

- The BLAST tool of the National Center for Biotechnology Information maintains a database of all genomic sequences from its funded research
- Using the "megablast" procedure, the *Leptospira* species in the database with the highest alignment score to the queried strain was recorded
- Strain names were then matched with their corresponding query identification numbers on the various dendrograms from the hierarchical cluster analysis to determine whether the dedrogram clusters were, in fact, clusters of the same *Leptospira* strain

Acknowledgements: I would like to recognize Dr. Claudia Muñoz-Zanzi for allowing me to use data from the "Eco-Epidemiology of Leptospirosis in Latin America" project, funded by the Ecology of Infectious Disease program. Award # 0913570. I also thank Dr. Cavan Reilly of the Division of Biostatistics at the University of Minnesota School of Public Health for his guidance on using the pairwise sequence alignment techniques.

References

1. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ and Higgins DG Bioinformatics 2007 23(21): 2947-2948. http://www.ebi.ac.uk/Tools/msa/clustalw2/
2. R: A Language and Environment for Statistical Computing. R Development Core Team. Vienna, Austria. 2011. http://www.R-project.org
3. BLAST: Basic Local Alignment Search Tool. National Center for Biotechnology Information. Bethesda, Maryland. 2011. http://blast.ncbi.nlm.nih.gov/Blast.cgi
4. Ko A.I., Goarant, C, and M. Picardeau 2009: Leptospria: the dawn of the molecular genetics era for an emerging zoonotic pathogen. Nature Reviews Microbiology 7, 736-747.
5. Reilly, Cavan. Statistics in Human Genetics and Molecular Biology. 2009. CRC Press. Boca Raton, Florida. P. 216-226.

Results

Hierarchical Cluster Analysis

Figure 2: Single linkage dendrogram

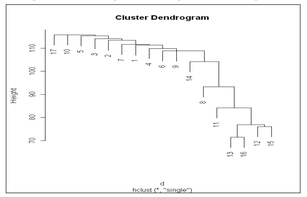


Figure 3: Complete linkage dendrogram

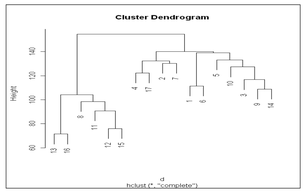
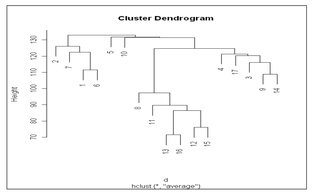


Figure 4: Average linkage dendrogram



Strain Identification

Table 1: BLAST search matches of *Leptospira* samples based on percent alignment scores

Sample Number	Sample Type	BLAST Species	BLAST Strain	% Alignment
1	Mouse	Interrogans	Copenhageni	97
2	Mouse	Interrogans	Copenhageni	97
3	Mouse	Borgpetersenii	Javanica	95
4	Mouse	Borgpetersenii	Javanica	93
5	Mouse	Interrogans	Copenhageni	97
6	Mouse	Borgpetersenii	Javanica	93
7	Mouse	Borgpetersenii	Javanica	96
8	Water	Borgpetersenii	Javanica	96
9	Water	Borgpetersenii	Javanica	94
10	Water	Interrogans	Copenhageni	99
11	Water	Interrogans	Copenhageni	81
12	Water	Interrogans	Copenhageni	94
13	Water	Interrogans	Copenhageni	94
14	Water	Alexanderi	Manzhuang	85
15	Water	Interrogans	Copenhageni	88
16	Water	none	none	n/a
17	Reference	Borgpetersenii	Javanica	97

Results (continued)

- All linkage techniques separated samples 8, 11,13,12, 15 and 16 from the other samples and put them far down in the dedrogram, indicating that they were not similar to any other samples analyzed.
- The single linkage dendrogram produce d a result consistent with a pattern of separating strains two-by-two, giving a wide chain of pairwise similarities.
- Both the complete linkage and average linkage techniques identified similarities among samples 3,4,9,14, and 17. Based on the matches found in BLAST, all five samples were of the *Borgpetersenii Javanica* strain.
- Samples 5 and 10 were consistently grouped together across all linkage techniques, often also considered to be very similar to sample 2. This is also consistent with the BLAST results of the three samples all being of the *Interrogans Copenhageni* strain.
- It is of note that the reference strain did not match with the *Borgpetersenii Ballum* serovar in the BLAST database. However, it was of the same, more broadly defined, *Leptospira* species.
- Additionally, one water sample was not able to be aligned with any *Leptospira* strain in the BLAST database.

Conclusions

- The objective of this project was to determine whether sequence alignment techniques would meaningfully contribute to the construction of a mathematical model of transmission of *Leptospira* that incorporates multi-host and environmental components. The repeated presence of two serovars (*Javanica and Copenhageni*) indicate that mice and water samples in the same area share *Leptospira* strains, presenting evidence for a shared rodent-water reservoir for infection.
- There were also distinct clusters identified by the three linkage techniques, suggesting that strains in the mice and water were so similar that they likely did not originate from different sources.
- All of the samples presented in this project were from one community in the Los Rio Region of Chile. Therefore, it is expected that little variation in the prevalent *Leptspira* serovars will exist. However, as more households and more communities become enrolled in the study, it is anticipated that increasingly more serovars will be detected. In conjunction with information about the geographic location at which the sample was obtained, sequence alignment can help establish the patterns of different serovars of the pathogen as the bacteria moves through a community.
- Given that there are over 200 serovars of *Leptospira* with varying human and animal pathogenicity, determining genomic similarities across strains might help inform treatment guidelines by allowing for conjecture as to the severity of the infection.

Limitations

- There were difficulties in matching the *Leptospira* sequences with reference strains in the BLAST database, which may indicate a geographic component to the *Leptospira* pathogen. Since this data is being collected in Chile, the strains may not match those found in the United States, which are the samples most likely to be contained in the BLAST database.
- Upon completion of this analysis, the primers for the sequencing of the *Leptospira* were re-evaluated. It appears that the primers were not sensitive enough to detect the bacteria, and therefore the samples presented here, namely those with poor alignment scores, may represent an insufficient portion of the genome for comparison to other samples.