Linguistic Analysis of Endosymbiosis

Ken Ober and Judith Klein-Seetharaman
Linguistic Characterization Methods: N-gram Frequencies

- $P_{\text{observed}}(ABCD)$
  - 4-gram / 4-gram$_{\text{total}}$
- Language Analogy
  - Common Bigram: “I had”
  - Uncommon Bigram: “mordant ire”
Escherichia coli K12

n=1

n=2

n=3

n=4
Linguistic Characterization Methods: Difference Values

- $P_{\text{observed}}(ABCD) - P_{\text{expected}}(ABCD)$
- Language Analogy
  - Overrepresentation: “bless you”
  - Underrepresentation: “map tree”
Linguistic Characterization Methods: Difference Values

- $P_{\text{observed}}(A) \times P_{\text{observed}}(B) = P_{\text{expected}}(AB)$
  - We say, “The expected probability of finding $A$ next to $B$, based upon the observed distributions of $A$ and $B$.”

- $P_{\text{observed}}(AB) = P_{\text{expected}}(AB)$ ?
  - We say, “If the observed probability of finding $A$ next to $B$ is not equal to the expected probability, then the distribution of $AB$ is not random.”
  - A non-random distribution means conserved units!
An Application

Endosymbiosis
The Test Subjects

- http://www.bio.mtu.edu/~jkoyadom/algae_webpage/charophyceans/Spirogyra_jason4_chassel2_20125.jpg
Difference Distribution Comparison

Cyanobacteria

Average Values

Cyanobacteria

Average Values

Cyanobacteria

Plastid

Average Values

Plant

Plastid

Plant

Plastid

Plant

Negative Controls...
Observation: Data follow the trend of the average.
Observation: Data follow the trend of the average.
Observation: Data follow the trend of the average.
Difference Distribution Comparison

Cyanobacteria

Average Values

Average Values

Plant

Plant

Plant

Plastid

Plastid

Plastid

Negative Controls...
Observation: Plant averages are "closest" to those of the cyanobacteria. Cyano-Plant, Cyano----Plastid
Observation: Plastid averages are “closest” to those of the plant. There is less variation here than in all other graphs. Plant-Plastid, Plant----Cyano
Observation: Plant averages and cyanobacteria averages are both more similar to those of the plastid than the controls. Plastid—Plant, Plastid—Cyano
Conclusions

- The average plastid difference distribution is *very* similar to that of the plants.
- This suggests the existence of a mechanism for maintaining non-random distributions of n-grams within organisms.
Next Steps

- Threshold for average distributions
- More species/sequence data
- Examine more closely the negative controls vs. the “closest” average
- Statistical analysis of the averages
Acknowledgements

- Judith Klein-Seetharaman & Co.
  - Vijayalaxmi Manoharan
  - Danny Lam
  - Deborah Weisser
- The BBSI Crew
- NIH & NSF
- All the rest of my peeps