Measuring Similarity Within Species

Rajat Sirohi

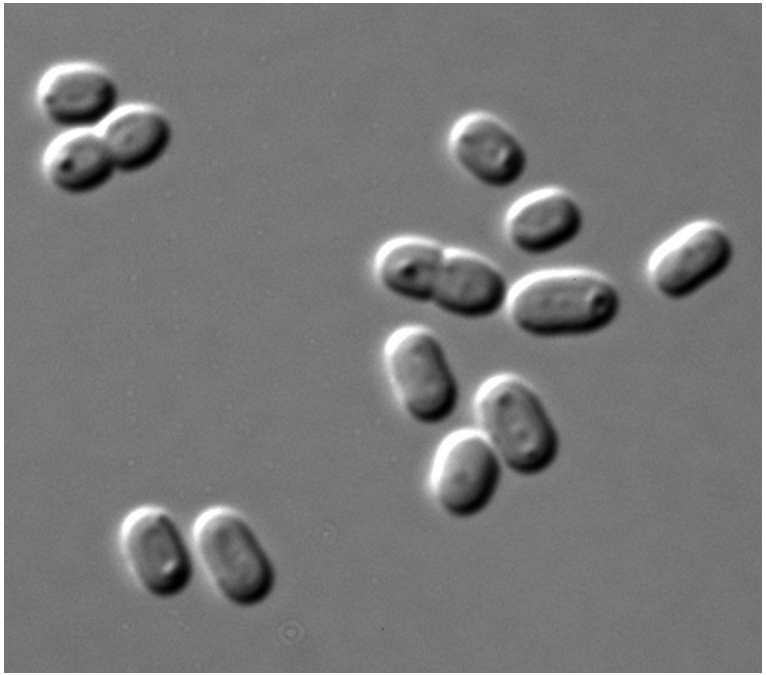
Overview and Introduction

“How similar are organisms within the same species compared to any random pair of organisms?”

- Goals
  - Devise measure of similarity
  - Calculate intraspecies similarity
- Hypotheses
  - H_0: the two organisms are no more similar than random chance would predict
  - H_a: the two organisms belong to the same species

Data Collection

- NCBI Genome Database: GenBank
- *Synechococcus elongatus*
- 100 random chosen samples
- Conditions:
  - Reputable database ensures non-biased samples
  - Random selection process
  - 10% condition
  - Success/failure condition



Expected Distribution

N = length of genome strand

n = number of nucleotide options = 4

S = number of simulations

ψ = measure of similarity

μ = mean

σ = standard deviation

$P(a_i == b_i) = 1/n$

$E(\psi) = N * P(a_i == b_i) = N/n$

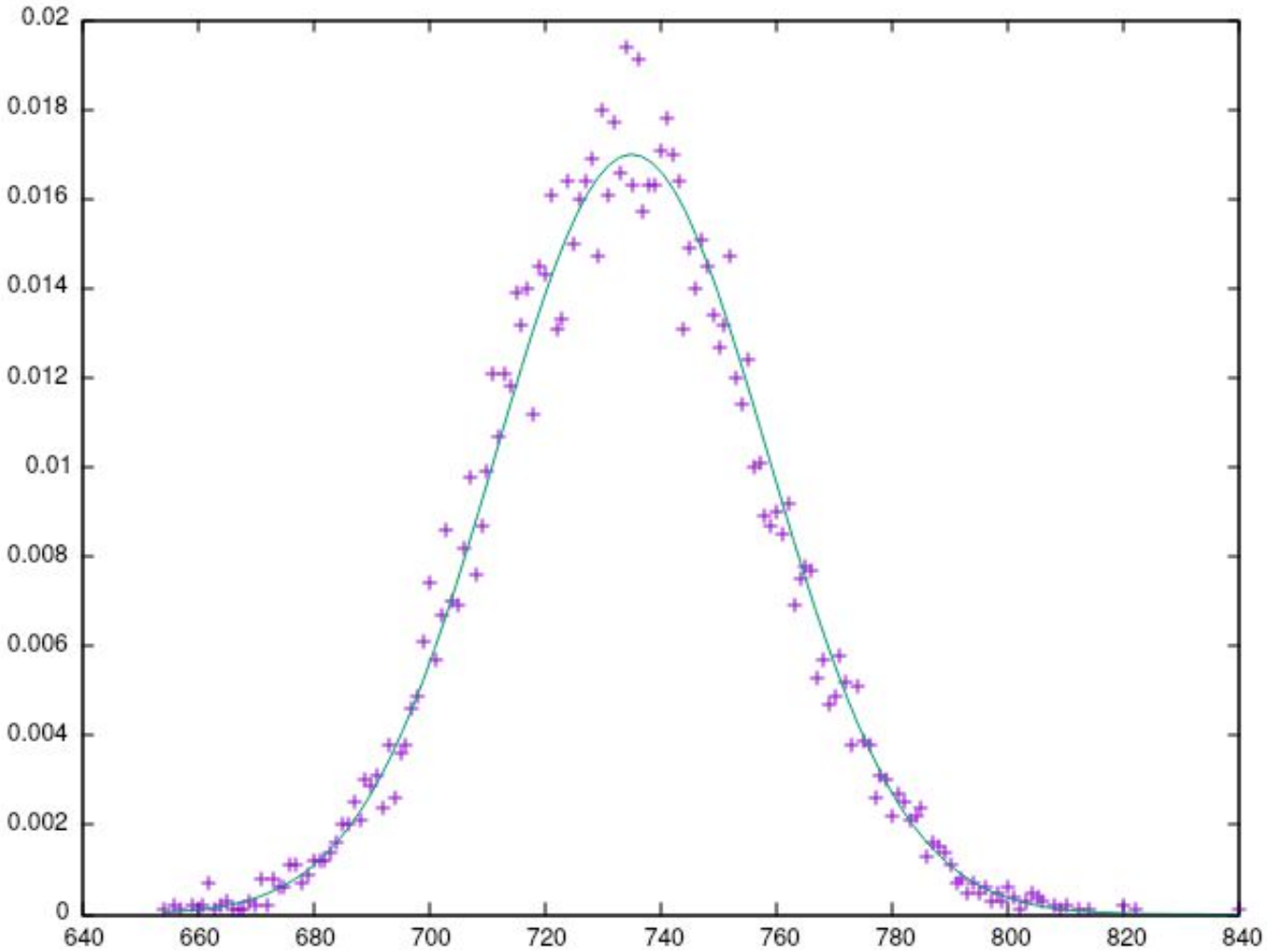
$Freq(\psi) = (N \text{ choose } \psi)\ (1/n)^{\psi}\ ((n-1)/n)^{N-\psi}$

$\mu = Np = N/n$

$\sigma = sqrt(Npq) = sqrt(N * (n-1)/n/n)$

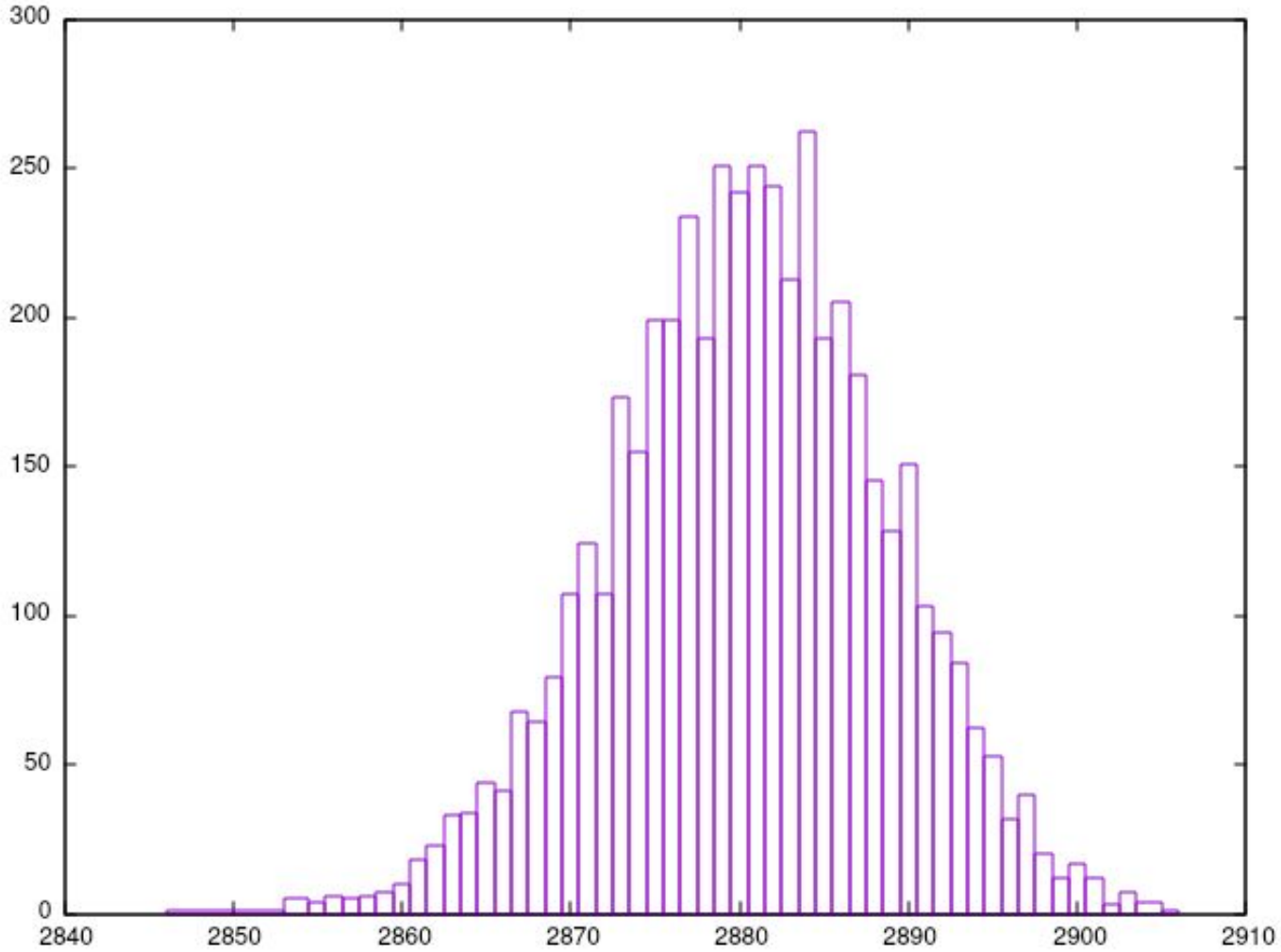
$A = a_1a_2a_3...a_N$

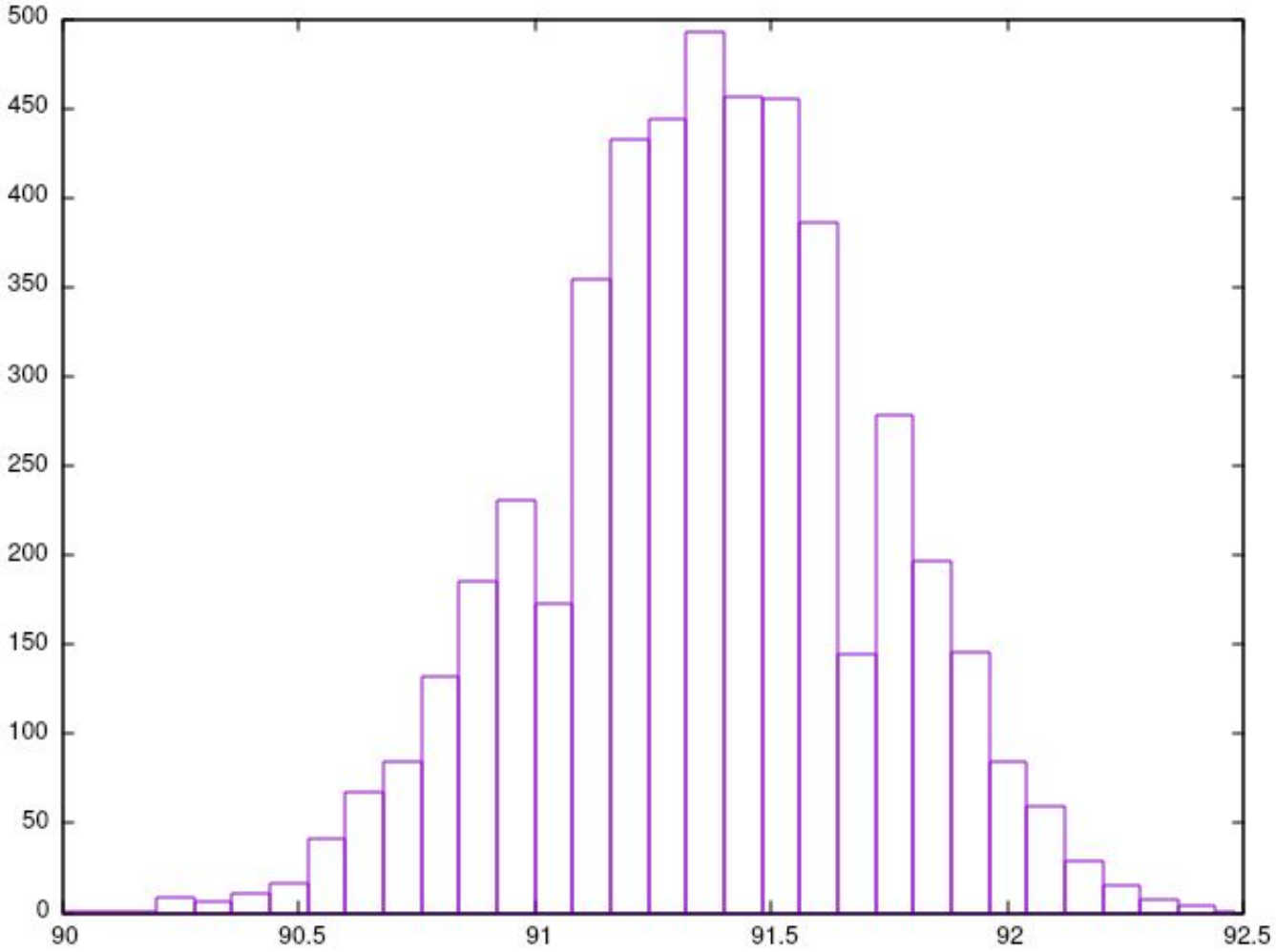
$B = b_1b_2b_3...b_N$


0.02
0.018
0.016
0.014
0.012
0.01
0.008
0.006
0.004
0.002
0
640
660
680
700
720
740
760
780
800
820
840

Data Analysis

- Analyzed 100 samples, resulting in (100 choose 2) = 4950 comparisons
- Results
  - ψ = 735
  - Sigma ~ 23.479
  - ψ_hat ~ 2880.503
  - Z_score ~ 91.380
  - P_val ~ 0


300
250
200
150
100
50
0
2840
2850
2860
2870
2880
2890
2900
2910


500
450
400
350
300
250
200
150
100
50
0
90
90.5
91
91.5
92
92.5

Conclusions

- Given the extremely low p-value, we can reasonably **reject the null hypothesis** in favor of the alternative hypothesis with near 100% confidence

- Confidence interval: (-1410.503, 2880.503)

Reflection

- Extension opportunities
  - Use different measure of similarity
    - Cosine similarity
    - Compare pairs rather than singular nucleotides
    - Weighted element-wise or cosine similarity
  - Restrict random distribution to viable genomes
- Difficulties collecting data
- Unsurprising results, stepping stone for further research

Thank you