Visualizing the Prevalence of Gene Duplication in Bacterial Genomes

Jer-Yee (John) Chuang – UC Berkeley
School of Information Management and Systems
Roadmap

• Molecular Biology 101

• Why study gene duplication?

• Prevalence of duplication in bacterial genomes
  – Method/Data
  – Challenges
  – Visualization
Molecular Biology 101

- Multiple levels of complexity

Genome

\[ \text{... gacggtagccaagtgcggtagcca} \text{gcg ...} \]

Genes

Proteins

Pathways

\[
\begin{align*}
\text{L-rhamnose} \\
\text{rhamnulose} \\
\text{rhamnulose-1-phosphate} \\
\text{lactaldehyde} \\
\text{lactate} \\
\text{propane-1,2-diol}
\end{align*}
\]
Why study gene duplication?

- Gene duplication:
  - What is it?
    - gacggtagccagtgcggtagccagcg ...
  - Is a driving force for creating new genes and functions
  - Is a mechanism for evolving complexity

Example: Feed Forward Motif
Duplication in Bacterial Genomes

• Method:
  – Given a particular genome, find all the duplicated genes in it (e.g., … gac\textbf{ggtagcca}gtgc\textbf{ggtagcca}gcg … )
  – Look for those duplicated gene pairs in other genomes

• Outcomes (5 possible):

```
  g  g
  g  g
  g  g
  g  g
  g  g
```
• What does it mean to say that a gene exists?
  – There is some statistical measure associated with that prediction (e.g., percent identity)

  e.g., ... gacggtagccagtgcggtagcgagcg ... → 6/8 bases → 75%

Percent identity of each gene with reference gene

Also need to include percent identity with each other!
Challenges (cont.)

• What about the evolutionary relationships among the genomes?
  – Knowing how related two genomes are tells us something about our results
Conservation of gene ordering within genomes:

Gene Pair: (rhaR,rhaS)
Genome: E.Coli K12
Perliden: (23%, -,-)

% Identity

0% 100%
Questions

Comments

Thank you!
Visualization (1st Attempt)