Gene Team Tree
A Compact Representation of All Gene Teams

Melvin Zhang Hon Wai Leong

School of Computing,
National University of Singapore

October 14, 2008

This work as supported in part by NUS under
Grant R252-000-289-112 and R252-000-361-112

Outline

Outline

Genome Evolution

- Genomes evolve via local and global changes
- Local changes act at the nucleotide level
- Global changes affects the *gene order*

Rearrangement type	**Effect on gene order**		
Reversal	a b c d e	$\Rightarrow$	a b e d c
Transposition	a b c d e	$\Rightarrow$	a c d e b
Inverted transposition	a b c d e	$\Rightarrow$	a e d c b
Insertion	a b _ c d e	$\Rightarrow$	a b f c d e
Duplication	a b _ c d e	$\Rightarrow$	a b a c d e
Deletion	a b c d e	$\Rightarrow$	a b d e

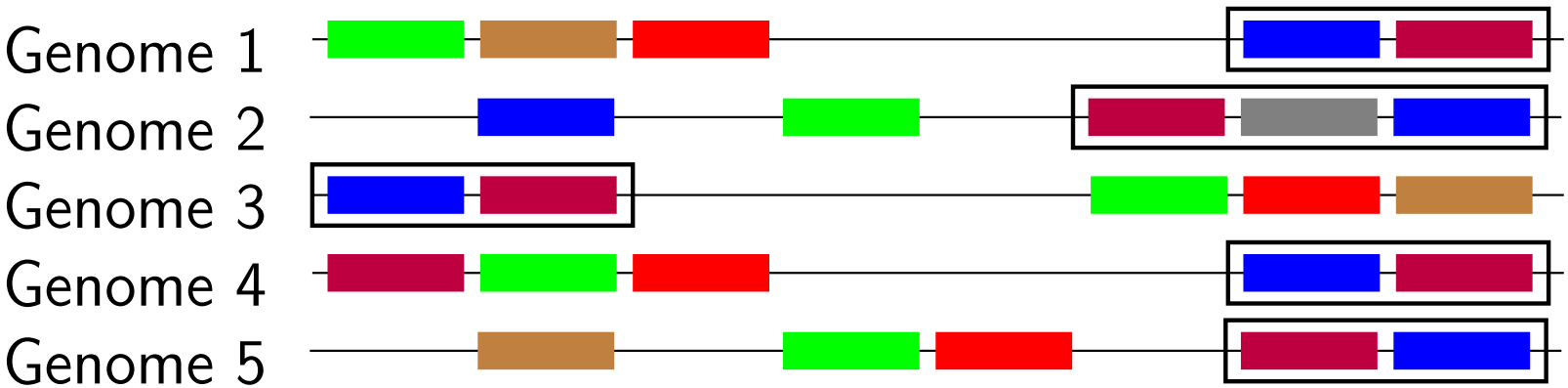
Table: Effect of rearrangements

Conserved Gene Clusters

Intuitively, sets of genes which are located close to one another in multiple genomes. Order is usually not conserved due to rearrangements.

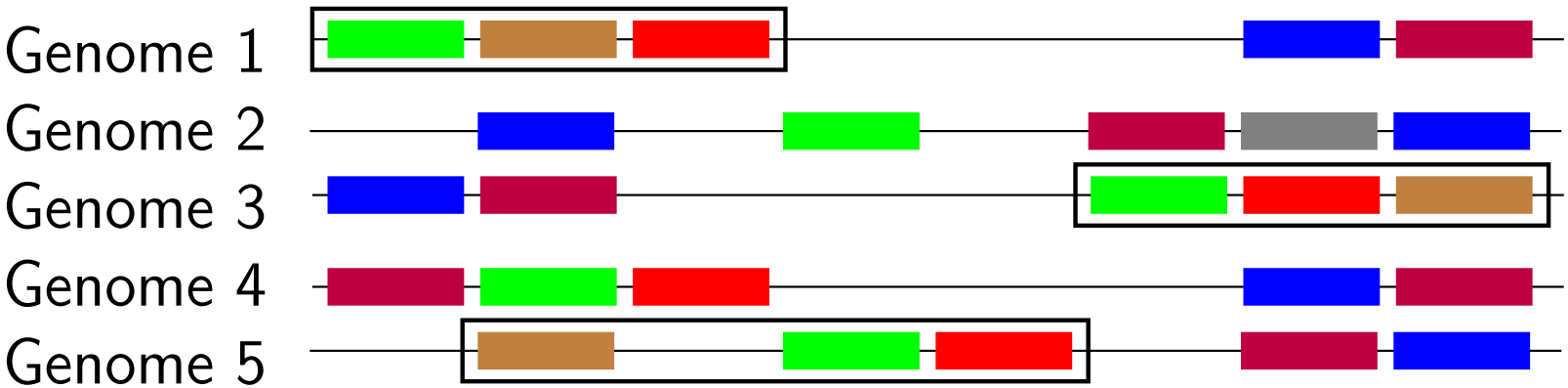
Conserved Gene Clusters

Intuitively, sets of genes which are located close to one another in multiple genomes. Order is usually not conserved due to rearrangements.



Conserved Gene Clusters

Intuitively, sets of genes which are located close to one another in multiple genomes. Order is usually not conserved due to rearrangements.



Applications

- Identify functional dependency/gene modules
- Identify large scale duplications/horizontal gene transfer

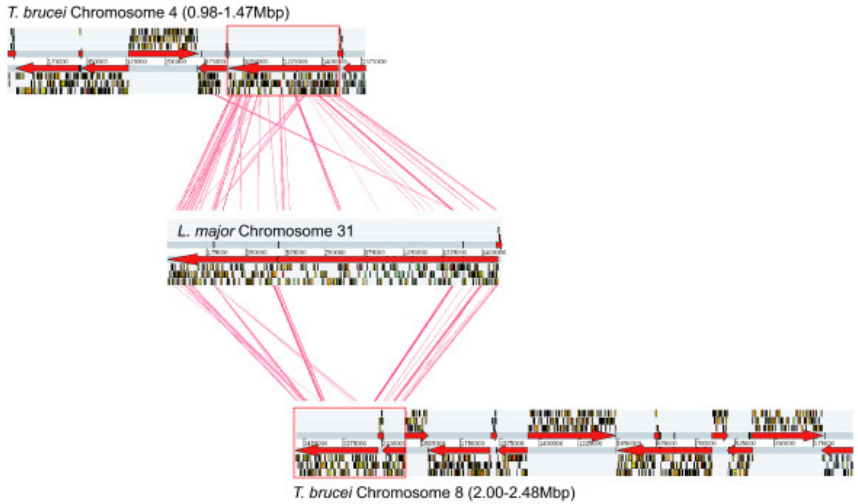


Figure: Trypanosoma brucei chromosomes 4 and 8 are partial duplicons [1]

- Measure of evolutionary distance

[1] Jackson, BMC Genomics 2007

Gene Teams: A model for conserved gene clusters

Gene Team

A maximal set of genes such that the distance between consecutive genes in all gene orders is at most δ [a].

[a] He and Goldwasser, JCB 2005

Example

$\delta = 2$

G : a b * * * a * c b

H : c * * c b a * b

Gene teams are

Gene Teams: A model for conserved gene clusters

Gene Team

A maximal set of genes such that the distance between consecutive genes in all gene orders is at most δ [a].

[a] He and Goldwasser, JCB 2005

Example

$\delta = 2$

G : a b * * * a * c b

H : c * * c b a * b

Gene teams are $\{a, b\}$,

Gene Teams: A model for conserved gene clusters

Gene Team

A maximal set of genes such that the distance between consecutive genes in all gene orders is at most δ [a].

[a] He and Goldwasser, JCB 2005

Example

$\delta = 2$

G : a b * * * a * c b

H : c * * c b a * b

Gene teams are $\{a, b\}, \{c\},$

Gene Teams: A model for conserved gene clusters

Gene Team

A maximal set of genes such that the distance between consecutive genes in all gene orders is at most δ [a].

[a] He and Goldwasser, JCB 2005

Example

$\delta = 2$

G : a b * * * a * c b

H : c * * c b a * b

Gene teams are $\{a, b\}, \{c\}, \{a, b, c\}$

FindTeams algorithms for Gene Teams

Observation

A gene order can be divided across gaps which are longer than δ [a].

[a] Bergeron et al. WABI 2002, B'eal et al., TCS 2004

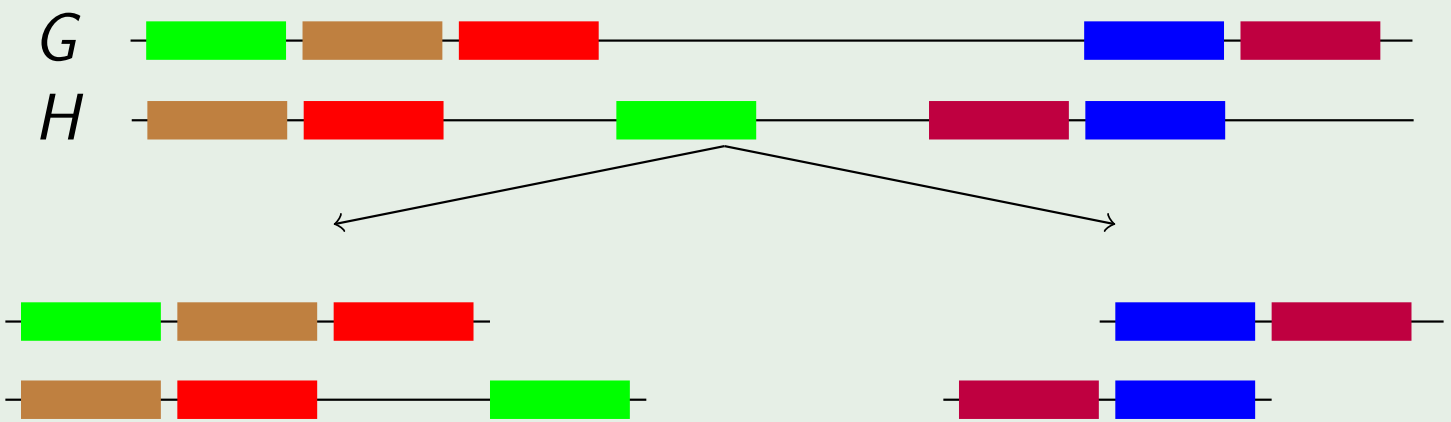
FindTeams algorithms for Gene Teams

Observation

A gene order can be divided across gaps which are longer than δ [a].

[a] Bergeron et al. WABI 2002, B'eal et al., TCS 2004

Example



Outline

Gene Teams in Practice

Problem: How to determine the "right" value of δ ?

δ is affected by:

- arrangement and distance between the genes
- rate of rearrangements
- type of conserved gene cluster, e.g. pathways, operons, etc.

Existing Approach

Benchmark against known conserved structures

1. Select a small number of known operons
2. Pick minimum δ so that selected operons were reconstructed. [a]

[a] He and Goldwasser, JCB 2005

Existing Approach

Benchmark against known conserved structures

1. Select a small number of known operons
2. Pick minimum δ so that selected operons were reconstructed. [a]

[a] He and Goldwasser, JCB 2005

Drawbacks

- May not have prior knowledge of conserved structures.
- How to select a representative set of known operons?

Our Approach

Key idea

- Find gene teams for ALL values of δ (this paper).
- Apply application specific tests to determine significance (ongoing work).

Key Observations

Observation 1

Decreasing δ causes existing teams to split into smaller teams.

Key Observations

Observation 1

Decreasing δ causes existing teams to split into smaller teams.

Example

G : a b * * * a * c b

H : c * * c b a * b

$\delta = 2 : \{a, b\}, \{c\}, \{a, b, c\}$

Key Observations

Observation 1

Decreasing δ causes existing teams to split into smaller teams.

Example

G : a b * * * a * c b

H : c * * c b a * b

$\delta = 2 : \{a, b\}, \{c\}, \{a, b, c\}$

$\delta = 1 : \{a, b\}, \{b\}, \{c\}, \{a\}, \{b\}, \{b, c\}$

Key Observations

Observation 1

Decreasing δ causes existing teams to split into smaller teams.

Example

G : a b * * * a * c b

H : c * * c b a * b

$\delta = 2 : \{a, b\}, \{c\}, \{a, b, c\}$

$\delta = 1 : \{a, b\}, \{b\}, \{c\}, \{a\}, \{b\}, \{b, c\}$

Implication

Set of all gene teams can be compactly represented as a tree.

Key Observations

Observation 2

Not all values of δ leads to new gene teams.

Key Observations

Observation 2

Not all values of δ leads to new gene teams.

Example

G : a b * * * a * c b

H : c * * c b a * b

$\delta = 3 : \{a, b\}, \{a, b, c\}$

Key Observations

Observation 2

Not all values of δ leads to new gene teams.

Example

G : a b * * * a * c b

H : c * * c b a * b

$\delta = 3 : \{a, b\}, \{a, b, c\}$

$\delta = 2 : \{a, b\}, \{c\}, \{a, b, c\}$

Key Observations

Observation 2

Not all values of δ leads to new gene teams.

Example

G : a b * * * a * c b

H : c * * c b a * b

$\delta = 3 : \{a, b\}, \{a, b, c\}$

$\delta = 2 : \{a, b\}, \{c\}, \{a, b, c\}$

Implication

We can do better than brute force.

Our Gene Team Tree

G : a b * * * a * c b

H : c * * c b a * b

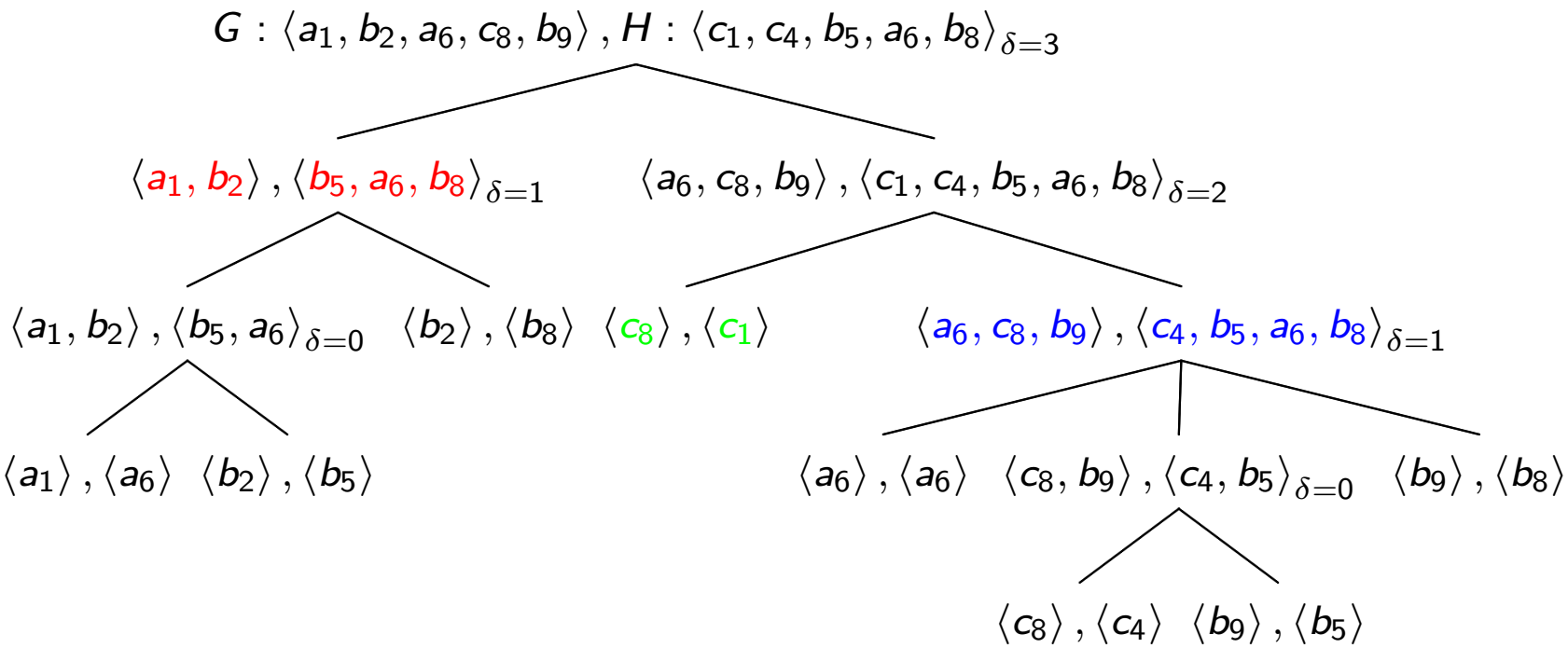


Figure: Gene team tree of G and H, value of δ used to split each node is show in subscripts. Colored nodes represent the gene teams for $\delta = 2$.

Our Main Results

Gene Team Tree Model

A compact representation of ALL gene teams.

Our Main Results

Gene Team Tree Model

A compact representation of ALL gene teams.

Algorithms to compute the GTT

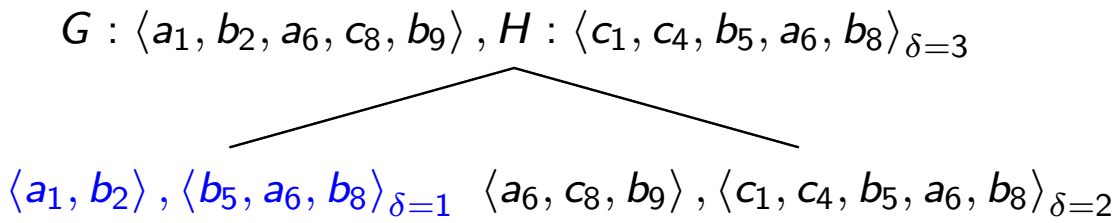
Generalize existing gene team mining algorithms with the *same worst case time complexity*.
$O(mn \lg^2 n)$ for m permutations of length n and
$O(\prod_{i=1}^{m} n_i)$ for m sequences of lengths $n_1, n_2, \ldots, n_m$.

Basic Algorithm

$$G : \langle a_1, b_2, a_6, c_8, b_9 \rangle, H : \langle c_1, c_4, b_5, a_6, b_8 \rangle_{\delta=3}$$

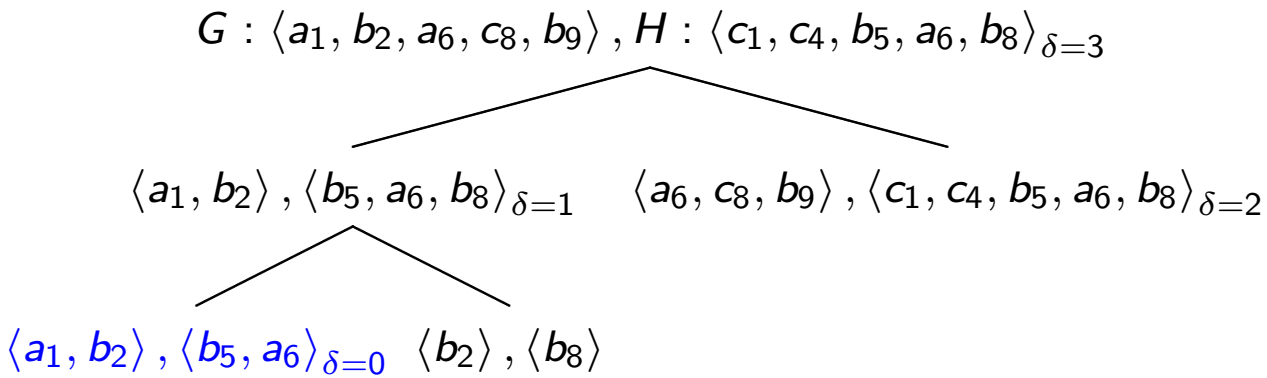
- Set $\delta = \textsc{MaxGap}(\text{team}) - 1$
- Run existing $\textsc{FindTeams}$ algorithms to determine children.

Basic Algorithm



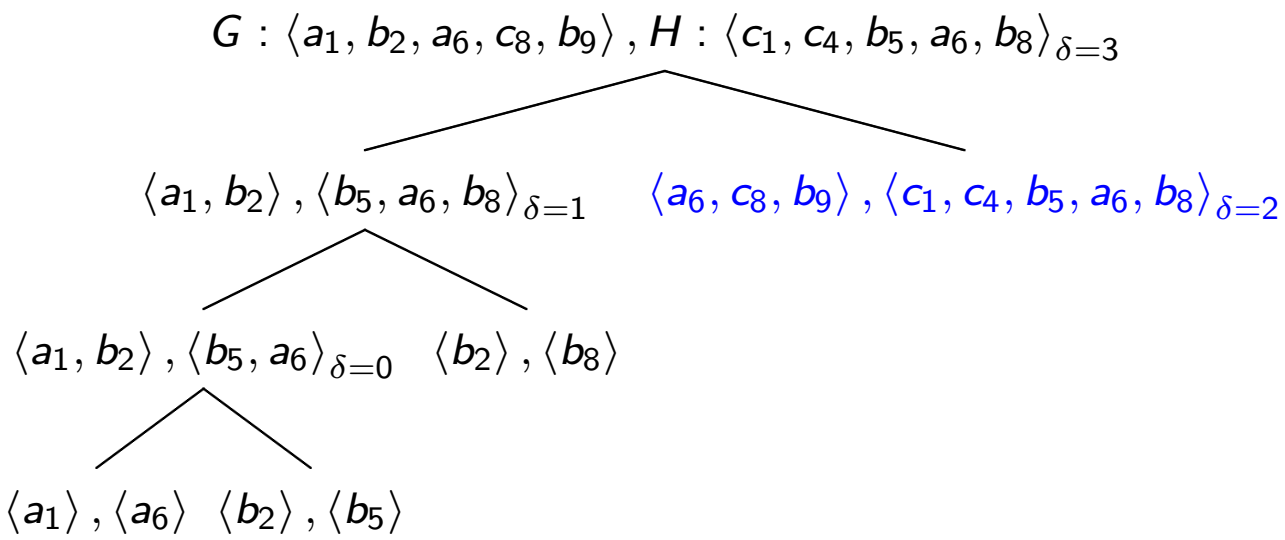
- Set $\delta = \textsc{MaxGap}(\text{team}) - 1$
- Run existing $\textsc{FindTeams}$ algorithms to determine children.

Basic Algorithm



- Set $\delta = \textsc{MaxGap}(\text{team}) - 1$
- Run existing $\textsc{FindTeams}$ algorithms to determine children.

Basic Algorithm



- Set $\delta = \textsc{MaxGap}(\text{team}) - 1$
- Run existing $\textsc{FindTeams}$ algorithms to determine children.

Speeding Up GTT Computation

Key insight

Instead of treating FindTeams as a black box, integrate computation of GTT into FindTeams.

Speeding Up GTT Computation

Key insight

Instead of treating FindTeams as a black box, integrate computation of GTT into FindTeams.

Details

- Modify base case of FindTeams reduce δ and continue.
- Implement MaxGap efficiently using heap.

Outline

Summary of Datasets

Prokaryote dataset (uni-chromosomal)

#homology families	1137 from He and Goldwasser, JCB 2005
#genes in *E. coli* K-12	2332
#genes in *B. subtilis*	2339
running time	4s [a]

[a] on a 2.33GHz processor

Eukaryote dataset (multi-chromosomal)

#homology families	12662 from Fu et al., RECOMB 2006
#genes in human	14193
#genes in mouse	14442
running time	5s [a]

[a] on a 2.33GHz processor

Gene Team Tree for *E. coli* K-12 and *B. subtilis*

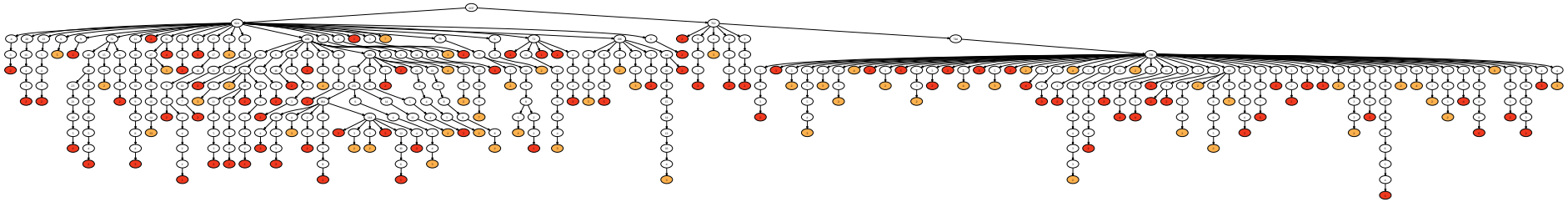


Figure: GTT for *E. coli* K-12(2332 genes) and *B. subtilis*(2339 genes) with 1137 homology families

Gene Team Tree for *E. coli* K-12 and *B. subtilis*

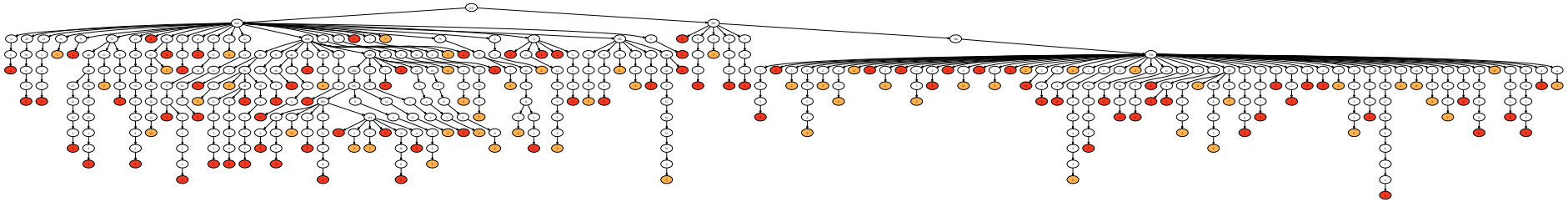


Figure: GTT for *E. coli* K-12(2332 genes) and *B. subtilis*(2339 genes) with 1137 homology families

Question

Does our GTT allows us to identify more operons as compared to using a single fixed value of δ?

Compare GTT against operons of *E. coli* K-12

1. Download set of *E. coli* K-12 operons from RegulonDB
2. For each operon, find best matching gene team in GTT based on Jaccard index

Compare GTT against operons of *E. coli* K-12

1. Download set of *E. coli* K-12 operons from RegulonDB
2. For each operon, find best matching gene team in GTT based on Jaccard index

Focus on 138 (out of 250) operons with Jaccard index $> 2/3$.
Recall: Gene team exists within a range of δ values, $[\delta_{\text{min}}, \delta_{\text{max}}]$

$$G : \langle a_1, b_2, a_6, c_8, b_9 \rangle, H : \langle c_1, c_4, b_5, a_6, b_8 \rangle_{\delta=3}$$

$$\langle a_1, b_2 \rangle, \langle b_5, a_6, b_8 \rangle_{\delta=1} \quad \langle a_6, c_8, b_9 \rangle, \langle c_1, c_4, b_5, a_6, b_8 \rangle_{\delta=2}$$

Range of δ for recovered operons

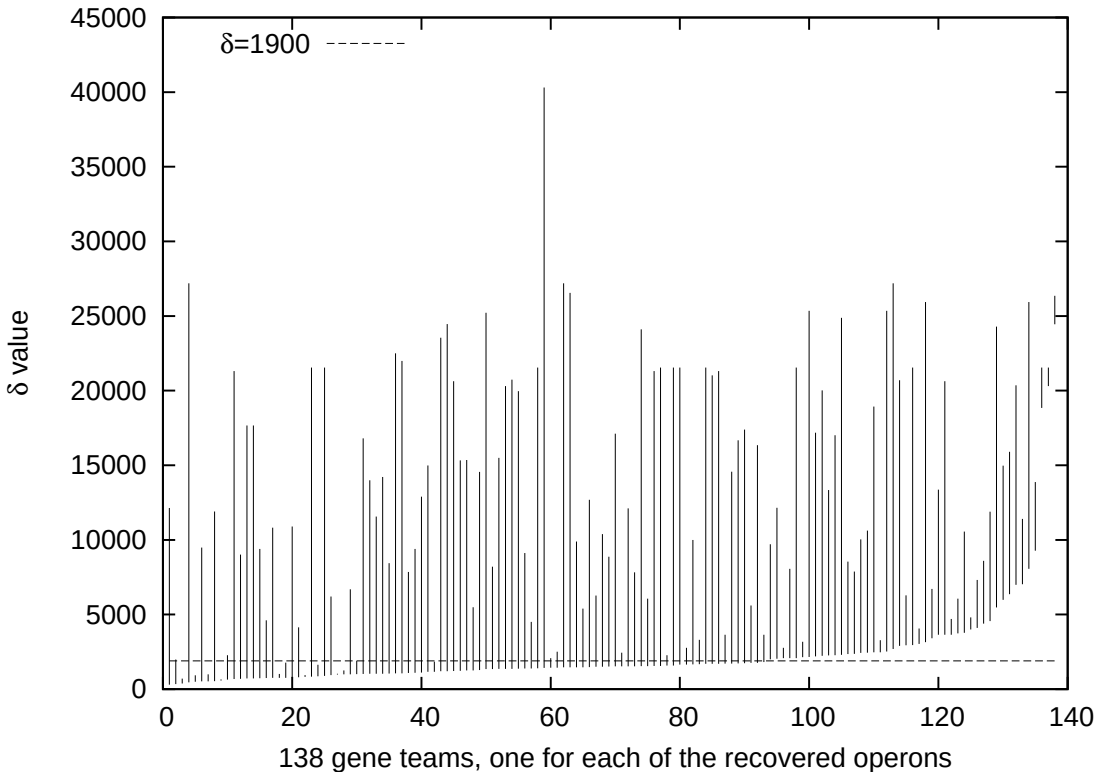


Figure: $[\delta_{\text{min}}, \delta_{\text{max}}]$ for best match gene teams, arranged in increasing δ_{min}

Wide variation in range of δ, more investigations are ongoing.

Number of recovered operons versus δ

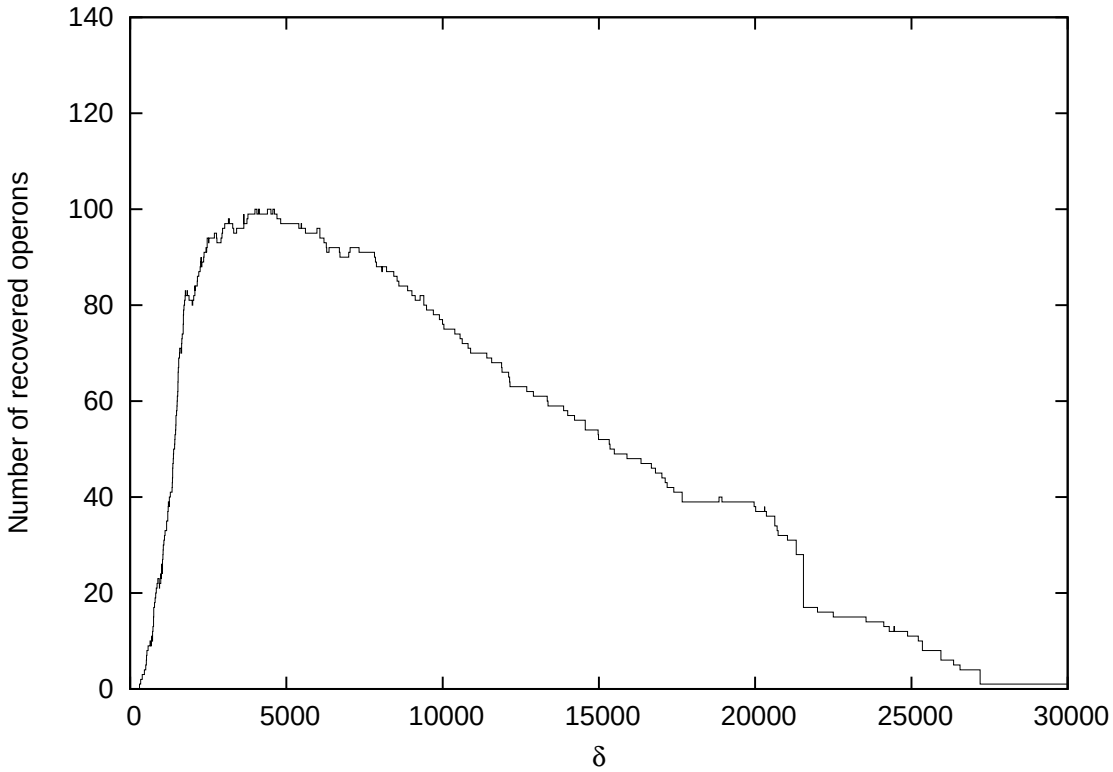


Figure: Number of recovered operons for different values of δ

$\delta = 1900$ recovers 81 operons. No single value of δ can obtain all the best matched gene teams.

Outline

Summary

- Gene Team Tree is a compact representation of all gene teams which makes explicit the structure of gene teams.

Summary

- Gene Team Tree is a compact representation of all gene teams which makes explicit the structure of gene teams.
- GTT can be computed efficiently by modifying existing parametrized algorithms.

Summary

- Gene Team Tree is a compact representation of all gene teams which makes explicit the structure of gene teams.
- GTT can be computed efficiently by modifying existing parametrized algorithms.
- Analysis of *E. coli* K-12 operons confirms that no single value of δ can produce all gene teams which is best match to known operons.

Future Work

- More datasets and compare against higher order mechanisms.
- Find segmental duplications via self comparison.
- Close the gap between performance of algorithms for permutations and sequences.

Future Work

- More datasets and compare against higher order mechanisms.
- Find segmental duplications via self comparison.
- Close the gap between performance of algorithms for permutations and sequences.

For downloads and updates on this project, please visit
`http://wwww.comp.nus.edu.sg/~melvin/GTT`