An Improved Model for Gene Cluster Inference

Xiao Yang Srinivas Aluru
Department of Electrical and Computer Engineering
Iowa State University, USA, 50011
Email: xyang, aluru@iastate.edu

Abstract

Inferring functionally related genes in microbial genomes is an important problem, which has been addressed by various gene cluster detection methods. Existing models well capture genomic inversions, gene duplications and insertions across a pair of genomic regions. Clusters involving multiple regions are indirectly inferred through constituent pairs. In this paper, we improve upon current work by a) formally modeling gene losses, and b) directly inferring gene clusters comprising of multiple genomic regions. Our preliminary results show a better sensitivity and specificity in identifying verified operons when compared with HomologyTeams model.

I. Introduction

The spatial proximity among a set of genes in microbial genomes indicates their functional correlation [8]. These gene sets are known as gene clusters and have been explored by numerous models.

We choose to improve on *gene teams* (or max-gap) model due to its desired properties and biological relevance [5]. The first gene teams model is proposed by Bergeron *et al.* [1], which requires the pre-assignment of orthologous gene pairs and constrains that all genes be present in every genomic region. The HomologyTeams model [4] removes the first requirement, and is one of the first models to consider gene duplications, using a similar divide and conquer approach as described in [1]. Kim *et al.* [6] used a breadth-first search algorithm to find gene clusters in multiple genomes, which relaxes the required spatial proximity, also, the median gene cluster model [7], [2] formulated a less restrictive optimization criterion intended for computational efficiency, resulted in a very high sensitivity. Conversely, Zhu *et al.* [11] used a more stringent proximity criterion than gene teams model that yielded comparable results in multiple yeast genomes. Different from above attempts, GCQuery [9] predicts new gene clusters with known operons using p-value statistics. Missing from all of these studies is a model that formally captures gene losses, which is important for studying the evolutionary history of gene clusters.

In this paper, we extend the existing work and formally capture the gene loss event by modeling the (unknown) ancestor - (extant) descendants relationship. Then, we formally define a *multi-gene-cluster* (abbreviated MGC) – a set of related extant genomic regions conforming to our evolutionary model. MGC model also improves the state-of-the-art by inferring evolutionary relationships among multiple genomic regions without relying on pairwise comparisons. (Due to space limitation, we foucs on bacterial genomes only, and a generalization to synteny detection is addressed in [10]). With these improvements, we can infer many obvious strong MGCs that would likely be missed by prior methods. We show in Fig. 1 three such example cases, each with 6 homologous gene pairs, where we regard each example as a single MGC. In (a), prior methods would likely find 3 gene clusters of size 2 (although such small clusters will generally be disregarded in practice). In (b), differential gene loss has occurred: G_1 and G_3 belong to the same MGC through shared homologous gene pairs with a common region of G_2; and in (c), pairwise genomic regions share only one homologous gene pair. Although the MGCs in these two cases could be captured by median gene cluster models, the corresponding parameters would be too relax such that even random gene clusters will be considered as valid.

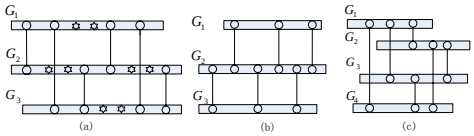


Figure 1. MGC examples. Genomic region G_i ($1 \leq i \leq 4$) is represented by a rectangle. Genes with and without homologous partners are denoted by circles and stars. Each line connects a pair of homologous genes.

We model a genomic sequence as a sequence of genes, and use homology between gene pairs to indicate shared evolutionary history. We present a polynomial approximation algorithm that takes a set of genomic sequences as input and infers all MGCs present in them. For our preliminary study, we test our model on two microbial genomes and present a detailed comparison with HomologyTeams model in identifying biologically verified operons, where our method shows a better sensitivity and specificity.

II. A Formal Model for Multi-gene-cluster

The existing gene cluster detection methods use gene content, proximity, and order constraints to infer related genomic regions. We formalize these empirical constraints in an evolutionary model that defines ancestor-descendant relationships, which is critical to formally define gene loss events although infering the ancestor is not necessary. Based on this model, a *multi-gene-cluster* (MGC) is defined to describe a set of evolutionarily related extant genomic regions.

A. An Evolutionary Model

Let $G = (g_1, g_2, \ldots, g_{|G|})$ denote a genomic sequence, where g_i is the i^{th} gene in G. Each g_i belongs to a *gene family*, defined as the transitive closure relation on gene homology. We label gene families with integers and write a genomic sequence using gene family labels when giving examples. We define the *gene distance* between two genes $g_i, g_j \in G$ as the number of intermediate genes: $d(g_i, g_j) = |i - j| - 1$, $(i \neq j)$. We use $F(\mathcal{S})$ to denote the set of gene family labels in $\mathcal{S}$, and overload F to let $\mathcal{S}$ be a gene or a set of genes, a sequence or a set of sequences of genes. We use $\mathcal{A} = (a_1, a_2, \ldots, a_{|\mathcal{A}|})$ to denote an unknown ancestral genomic region of G. A gene g is termed an *ancestral gene* if $F(g) \subset F(\mathcal{A})$. In order to make homologous relationships obvious in examples, for a set of genomic regions sharing an ancestor $\mathcal{A}$, we show labels only for ancestral genes, and denote the rest by "*". We formally define four basic evolutionary events on $\mathcal{A}$, where the first three have been considered in previous models:

1) *Inversion:* An inversion of a subregion $(a_j, a_{j+1}, \ldots, a_k)$ results in $(a_1, a_2, \ldots, a_{j-1}, a_k, a_{k-1}, \ldots, a_j, a_{k+1}, a_{k+2}, \ldots, a_{|\mathcal{A}|})$.
2) *Local gene duplication:* Ancestral gene a_j is duplicated and inserted into another position.
3) *Gene insertion:* A non-ancestral gene is inserted.
4) *Ancestral gene loss:* An ancestral gene is removed.

During evolution, $\mathcal{A}$ may undergo a sequence of events that result in an extant genomic region G. However, inferring this evolutionary relationship is not possible without knowing $\mathcal{A}$ and assuming parsimony. Therefore, we have to approximate the evolutionary distance between $\mathcal{A}$ and G, and we choose to limit this distance using the following definition with the assumption of parsimony.

Definition 1: An extant genomic region G is an (ϵ_i, ϵ_l)-descendant of an unknown ancestor $\mathcal{A}$ if G *could* be obtained by applying the following evolutionary events to $\mathcal{A}$.

1) Any number of inversions. This is because a mechanism for limiting the number of inversions is not obvious without knowing $\mathcal{A}$.
2) Any number of local gene duplications. Because limiting the number of paralogs is not necessary.
3) Up to ϵ_i gene insertions between every pair of consecutive genes in the genomic region $\mathcal{A}'$ resulting from inversions and local gene duplications.
4) Up to ϵ_l consecutive gene losses in any position of $\mathcal{A}'$ with the first and the last genes retained.

Note that the above definition relies only on the discovery of $F(\mathcal{A})$ rather than $\mathcal{A}$, and any extant genomic region G, resulting from some arbitrary sequence of evolutionary events, will adhere to the above definition if the gene insertion and loss rates are limited to ϵ_i and ϵ_l. For example, let $\epsilon_i = 1$ and $\epsilon_l = 1$. Let $\mathcal{A} = (1, 2, 3, 4)$ be an ancestral genomic region. $\mathcal{A}$ is transformed to $\mathcal{A}' = (1, 2, 1, 3, 3, 4)$ after two local gene duplications, and to $G = (1, *, 2, 3, *, *, 4)$ after 3 gene insertions and 2 ancestral gene losses (3rd and 5th genes in $\mathcal{A}'$). G is a $(1, 1)$-descendant of $\mathcal{A}$.

B. Multi-gene-cluster

We define a set of extant genomic regions $\mathcal{G}$ to be a MGC if there exists an ancestral genomic region $\mathcal{A}$ with gene families $F(\mathcal{A})$ such that each $G \in \mathcal{G}$ is a subregion of an (ϵ_i, ϵ_l)-descendant of $\mathcal{A}$. We regard this concept of gene cluster as biologically meaningful, an ideal that we seek to find.

In practice, the input consists of whole or partial extant genomes and homology relationships between the constituent genes. The task is to identify all the MGCs present in the input as per the above concept. However, even in the simplified problem of determining whether a given set $\mathcal{G}$ of genomic regions is a MGC, it is sometimes impossible to answer this with certainty. For example, $\mathcal{G} = \{(1, 3, 5, 7), (2, 4, 6)\}$ could be a MGC with $\mathcal{A} = (1, 2, 3, 4, 5, 6, 7)$ if $\epsilon_l = 1$; however, having no homologous genes in common, there is no evidence in the extant genomic regions to conclude they are in a MGC without knowing $\mathcal{A}$.

Our goal is to propose a formal definition of MGC that scouts for minimal credible evidence of an ancestral genomic footprint on the extant genomic regions. By analyzing multiple extant regions concurrently, the odds of inferring such an evolutionary connection are superior to relying on pairwise comparisons. However, this is challenging due to various uncertainties: a homology between two genes from different extant genomic regions could arise from speciation or horizontal gene transfer, causing a false positive; if two consecutive ancestral genes on an extant genomic region are separated by a distance d such that $\epsilon_i < d \leq \epsilon_i(\epsilon_l + 1)$, it is unclear whether this is due to $> \epsilon_i$ consecutive gene insertions (hence not in MGC), or due to ancestral gene losses (which is in MGC). In addition, genomic inversions disrupt the gene collinearity in different related genomic regions. We attempt to resolve such uncertainties by looking for homologous gene pairs due to speciation or segmental duplication and finding evidence for potential gene losses through connections with other extant genomic regions within the same MGC.

In what follows, we develop a formal definition of MGC that approximates $F(\mathcal{A})$ as the set of gene families shared among extant genomic regions in $\mathcal{G}$. Enroute, we define a few concepts:

Definition 2: Given two genomic sequences G and G', the G'-induced subsequence of G is the subsequence obtained from G (or the reverse of G) by retaining only the genes that have a homologous counterpart in G'.

Definition 3: An *α-run* in G w.r.t. G' is a substring $(g_{j_1}, g_{j_2}, \ldots, g_{j_k})$ of the G'-induced subsequence of G such that $d(g_{j_i}, g_{j_{i+1}}) \leq \alpha$ $(1 \leq i < k)$.

Definition 5: Given parameters ϵ_i, ϵ_l, β, let $\alpha = \epsilon_i(\epsilon_l + 1)$. A set $\mathcal{G}$ of extant genomic regions is termed a *multi-gene-cluster* if $\exists$ an ancestral genomic region $\mathcal{A}$ such that:

1) **[α-run Coverage]** For each $G \in \mathcal{G}$ there exists an α-run $G_\alpha = (\gamma_1, \gamma_2, \ldots, \gamma_{|G_\alpha|})$ in G w.r.t. $\mathcal{A}$ such that $\gamma_1 = g_1$ and $\gamma_{|G_\alpha|} = g_{|G|}$.
2) **[Conjugation]** For each such G_α, there exists a set of α-paths w.r.t $\mathcal{A}$, $\mathcal{P}_G = \{P_1, P_2, \ldots, P_k\}$, with each P_j consisting of α-runs taken from genomic regions in $\mathcal{G} \setminus \{G\}$, such that:
   a) $\mathcal{P}_G$ covers G_α: $F(G_\alpha) \subseteq F(\mathcal{P}_G)$.
   b) Each $P_j \in \mathcal{P}_G$ satisfies:
      i) *Projection:* $\exists$ an α-run Γ_j in G w.r.t. $F(P_j)$ with $|F(\Gamma_j)| \geq \beta$.
      ii) *Simplicity:* For any gene p_i in P_j such that $F(p_i) \subseteq F(G_\alpha)$, p_i occurs exactly once in P_j, and is not present in $\mathcal{P}_G \setminus \{P_j\}$.
      iii) *Gene Loss:* Let p_i and p_{i+1} be two consecutive genes in G_α-induced subsequence of P_j, the number of gene families between p_i and p_{i+1} is no more than ϵ_l (gene losses in G w.r.t. P_j).

      We call Γ_j and P_j a *conjugate pair*.
   c) Gene families in $\mathcal{P}_G$ account for ancestral gene losses in G: $|F(\mathcal{P}_G) - F(G_\alpha)| \geq \sum_j \max(\lceil \frac{d(\gamma_j, \gamma_{j+1})}{\epsilon_i} \rceil - 1, 0)$.
3) **[Reciprocity]** For each $P_j \in \mathcal{P}_G$ and $G' \neq G$, $\exists P'_j \in P_{G'}$ such that: p'_i is in P'_j if, 1) p'_i is not in G', 2) p'_i is in P_j or G, and 3) p'_i is homologous with some p_i in both P_j and G'.
4) **[Connectivity]** For any bipartition of $\mathcal{G}$ into sets $\mathcal{G}_1$ and $\mathcal{G}_2$, $\exists G \in \mathcal{G}_1$ and $P_j \in P_G$ such that P_j contains an α-run taken from $\mathcal{G}_2$.

Definition 4: Given $\mathcal{A}$ and $\mathcal{G}$, an *α-path* is a concatenation of α-runs, where each α-run occurs in some $G \in \mathcal{G}$ w.r.t. $\mathcal{A}$, and the two genes that mark the transition between two consecutive α-runs are the same or form a homologous pair.

Example: Let $F(\mathcal{A}) = \{1, 2, 3, 4, 5\}$, $\mathcal{G} = \{(1, *, *, 2, 4, 5)_1, (1, 3, *, 4)_2, (3, *, 2)_3\}$, and $\alpha = 1$. $(2, 4)_1$, $(1, 3, 4)_2, (2, 3)_3$ are three α-runs w.r.t. $\mathcal{A}$, and $(1_2, 3_2, 4_2, 3_2, 3_3,\ 2_3, 2_1, 4_1)$ is an α-path. The subscripts specify the respective genomic regions.

With these definitions, we provide a formal multi-gene-cluster definition (5). Though our definition still enforces that each $G \in \mathcal{G}$ is a subregion of an (ϵ_i, ϵ_l)-descendant of an unknown $\mathcal{A}$, we require a number of additional criteria to ensure that $\mathcal{G}$ is sufficient to discover $F(\mathcal{A})$; when ancestral gene losses in $G \in \mathcal{G}$ are implied by the model, we require strong evidence for the lost genes in $\mathcal{G} \setminus \{G\}$.

We present the biological justifications for definition 5 in the following. Parameter β specifies the minimum number of ancestral genes required in each genomic region for it to be included in the MGC. Therefore, it will be less likely for an unrelated genomic region to be identified as a member of a gene cluster if a larger value of β is used.

If there were no gene losses, we would wish to identify an ϵ_i-run w.r.t. $\mathcal{A}$ in each G, where ϵ_i is the gene insertion rate. To account for potential gene losses, we search for an

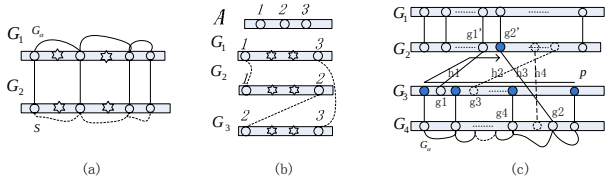


Figure 2. The elements in the figure follow the convention of Fig 1. (a) A gene cluster with collinear gene support; (b) Choice of α-path versus ϵ_i-path; (c) Two gene clusters connected by spurious homologous links.

α-run G_α, where $\alpha = \epsilon_i(\epsilon_l + 1)$ and ϵ_l is the gene loss rate. G_α specifies the inferable ancestral genes in G and justifies G to be a member of the MGC $\mathcal{G}$ (criterion 1). However, G_α is biologically meaningful only if there exists a set $\mathcal{S}$ of proximate genes in $\mathcal{G} \setminus \{G\}$ that consists of the homologous counterpart of every gene in G_α. A collinear relationship between genes in G_α and $\mathcal{S}$ would be an ideal support that has been targeted by many existing approaches. For example, in Fig. 2 (a), the genes in G_α (1-run, shown as the solid curved lines) and in its counterpart $\mathcal{S}$ (dotted curved lines) are collinear, providing mutual support for G_1 and G_2 to be related.

However, differential gene losses cause $\mathcal{S}$ to be divided into multiple non-intersecting gene sets $\{S_1, S_2, \ldots, S_{|\mathcal{S}|}\}$ occurring in different genomic regions that share no homologous gene pairs, e.g. in Fig. 1 (b), w.r.t. G_α for G_2 (6 genes), we could identify an $\mathcal{S}$ that consists of two gene sets with one set in G_1 and another in G_3. The projection, Γ_j, of each gene set onto G_α forms an α-run, where Γ_j and S_j provide mutual support for each other (criteria 2a and 2(b)i). In a more degenerate case, $S_j \in \mathcal{S}$ may occur in different genomic regions due to extensive gene losses and inversions that interrupt gene collinearity, making it difficult to check for gene proximities in S_j. This is handled by finding an α-path P_j that goes through every gene of S_j, therefore, insuring the "collinear" relationship between S_j and Γ_j. Gene order constraint is relaxed implicitly by only requiring $F(S_j) = F(\Gamma_j)$. Without the constraint of "simplicity", genomic region G_i could be regarded as a part of a gene cluster erroneously, as long as the homologous counterpart genes of G_i are in proximity to some common genes.

The evidence of ancestral gene losses is formally captured in criteria 2(b)iii and 2c. For any two adjacent genes g_j and g_k in an α-run, $d(g_j, g_k) > \epsilon_i$ indicates a minimum of $\lceil \frac{d(g_j, g_k)}{\epsilon_i} \rceil - 1$ ancestral gene losses that need to be inferred from the data. If there were no inversions, we would expect any gene loss to be manifested through the existence of a homologous gene pair $\{g, g'\}$ such that g occurs near g'_j while g' is proximate to g'_k, with g'_j and g'_k being the homologous counterparts of g_j and g_k, respectively. However, due to inversions, the actual gene losses may not necessarily be in between g_j and g_k. For example, let $\mathcal{A} = (1, 2, 3, 4)$, $\epsilon_i = 1$ and $\epsilon_l = 1$, and $\mathcal{A}$ evolved to $\mathcal{A}' = (1, *, 2, *, 3, 4)$ after insertions, to $\mathcal{A}'' = (1, *, *, 3, 4)$

after an ancestral gene loss, and to $G_1 = (1, *, *, 4, 3)$ after an inversion. Without knowing $\mathcal{A}$, but with the presence of two other descendants $G_2 = (1, 2, 3)$ and $G_3 = (2, 3, 4)$, we could identify a 2-run $(1, 4, 3)$ in G_1 and a corresponding 2-path $(1_2, 2_2, 2_3, 3_3, 4_3)$, where the homologous gene pair $(2_2, 2_3)$ marking the transition from G_2 and G_3 justifies the large distance ($> \epsilon_i$) between gene 1_1 and 4_1 to be an ancestral gene loss of 2_1, though the real loss is between 1_1 and 3_1. Criterion 2c specifies the minimum number of total ancestral gene losses that need to be inferred in G, if any, as discussed above. Criterion 2(b)iii identifies and restricts the maximum permissible ancestral gene losses as prescribed by the evolutionary model.

To justify criterion 3, we first use an example to explain the counterintuitive choice of an α-path in lieu of an ϵ_i-path for gene loss recovery. As shown in Fig. 2 (b), let $\epsilon_i = 1$ and $\epsilon_l = 1$, there exists a 2-run $(1_1, 3_1)$ in G_1, and by identifying a 2-path $P = (1_2, 2_2, 2_3, 3_3)$, we could identify a gene loss between 1_1 and 3_1. However, P consists of two α-runs, which in turn, need to be justified. By observing the symmetric property in this example, the two α-runs $(1_2, 2_2)$ and $(2_3, 3_3)$ of P could be reasoned with a gene loss of 3 and 1, respectively. As could be observed from the above example, while we are looking for gene losses of G in another genomic region G', genes in G may account for the gene losses in G'. Criterion 3 formally captures this "reciprocity" property and guarantees the validity of each α-path. In addition, it prevents the consideration of spurious homologous pairs that arose through accidental accumulation. Fig. 2 (c) shows such an example: $\mathcal{G}_1 = \{G_1, G_2\}$ and $\mathcal{G}_2 = \{G_3, G_4\}$ are two MGCs between long genomic regions. Genes $h2$ and $h4$ are hypothetical homologous gene pairs. Inevitably, there will be homologous links between $\mathcal{G}_1$ and $\mathcal{G}_2$, e.g. $h1$ and $h3$, which are spurious since g_1 and g_2 are far apart. G_α is an α-run in G_4, and genes visited exactly once by the α-path P are denoted by solid circles. However, P is invalid according to criterion 3 since reciprocally, for its constituent α-run $(g_1^{'}, g_2^{'})$ in G_2, there exists no α-path that visits both g_1 and g_2 to justify their proximity. However, with the presence of additional homologous links, e.g. $h2$ and $h4$, g_1 and g_2 could be shown to fall onto the same path as intermediated by g_3 and g_4. This, in addition, reasonably indicates a possible breakpoint (inversion) between g_1 and g_3, followed by gene losses in G_3. Following above statements, criterion 4 prevents merging of two different MGCs.

III. An Algorithm for Multi-gene-cluster Inference

Let $\mathcal{D}$ denote the set of input genomic sequences, which may be whole or partial extant genomes or chromosomes from one or more species. Our goal is to infer MGCs present among these sequences. For each MGC, we infer the set of genomic regions engaged in it and recover the corresponding ancestral gene families $F(\mathcal{A})$. Since our definition of MGC allows for at most α interlaying genes, two genomic regions on the same genomic sequence can be two distinct regions in the same MGC only if their separation is more than α.

We represent genomic sequences and homologous gene pairs in $\mathcal{D}$ using a graph. The algorithm performs a series of graph reductions, driven by violation of the MGC criteria. The final result is a reduced graph where each connected component corresponds to a MGC.

Let $\mathbb{G} = (V, E)$ denote the graph at some stage of the algorithm. Each vertex is a gene from one of the sequences in $\mathcal{D}$. Each edge is either one of the two types: an l-edge (linkage edge) exists between two genes lying on the same genomic region, and an h-edge (homologous edge) between two homologous genes. The *length* of an l-edge is equal to the distance between the corresponding genes. A maximal sequence of l-edges is considered a *partition*, and represents the corresponding genomic region. The initial graph consists of one partition for each genomic sequence in $\mathcal{D}$, l-edges between every consecutive pair of genes in every genomic sequence, and h-edges for every homologous gene pair.

Let a partition $\mathbb{P} = (v_1, v_2 \ldots, v_{|\mathbb{P}|})$. A subsequence $(v_{j_1}, v_{j_2}, \ldots, v_{j_l})$ of $\mathbb{P}$, where $1 \leq j_1 < j_2 < \ldots < j_l \leq |\mathbb{P}|$, is defined as an *$\alpha$-sequence* if $d(v_{j_k}, v_{j_{k+1}}) \leq \alpha$ for $1 \leq k < l$. Given a set $\mathcal{U}$ of vertices, a *$\mathcal{U}$-induced α-set* $\mathcal{S}$ is defined as a set of maximal α-sequences consisting of only and all of the vertices in $\mathcal{U}$.

We define the following operations for graph reduction:

- *Vertex Removal (VR)*. Let v be the vertex to be removed.
  - If v is incident to two l-edges (u, v) and (v, w), create an l-edge (u, w) if $d(u, w) \leq \alpha$. Otherwise, v's removal breaks its partition into two. Invoke *Partition Removal* on each of the resulting partition $\mathbb{P}$ if $|F(\mathbb{P})| < \beta$.
  - Invoke *Edge Removal* for all h-edges and delete all l-edges incident to v. Finally, remove v.
- *Edge Removal (ER)*. Remove the specified edge $e = (u, v)$, and
  - If e is an l-edge, it breaks a partition into two. Invoke *Partition Removal* on each of the resulting partition $\mathbb{P}$ if $|F(\mathbb{P})| < \beta$.
  - If e is the only h-edge incident to u (respectively, v), then invoke *Vertex Removal* on u (respectively, v).
- *Partition Removal (PR)*. Invoke *VR* for every vertex in the partition.

In the following, we provide a brief sketch of the algorithm and analysis, omitting details and the proofs due to space limitation.

A. Algorithm

Algorithm 1 begins with the initial graph $\mathcal{D}$ and removes unique genes and transforms $\mathcal{D}$ to $\mathbb{G}$ using the α distance constraint. For each connected component $\mathbb{C}$ in $\mathbb{G}$, we look

Algorithm 1 MGC Detection

1: Apply *ER* to each vertex that has no h-edges on initial graph, and obtain the resulting graph $\mathbb{G}$.
2: Divide $\mathbb{G}$ into a set of Connected Components.
3: **for** each connected component $\mathbb{C}$ **do**
4: **for** each partition $\mathbb{P} = (v_1, v_2, \ldots, v_{|\mathbb{P}|})$ in $\mathbb{C}$ **do**
5: Compute *anchor set* $\mathcal{U}_a = \{u \mid u \text{ in } \mathbb{C},\ u \text{ not in } \mathbb{P},\ F(u) \in F(\mathbb{P})\}$.
6: **for** every $u_j \in \mathcal{U}_a$ **do**
7: Enumerate paths starting at u_j in $\mathbb{C}$, where each path: (1) has no intersection with $\mathbb{P}$, (2) traverses at most ϵ_l h-edges, (3) contains no sub-path on l-edges with length $> \alpha$.
8: **if** During traversal, $u_k \in \mathcal{U}_a$ is encountered in another partition, and $F(u_k) \neq F(u_j)$ **then**
9: **if** $\exists v_l, v_m$ in $\mathbb{P}$ s.t. $d(v_l, v_m) \leq \beta\alpha$, $F(v_l) = F(u_j)$, and $F(v_m) = F(u_k)$ **then**
10: Mark the endpoints of all h-edges on path u_j to u_k as *co-anchors*.
11: **end for**
Let $\mathcal{U}_c$ denote the set of all co-anchors.
12: Identify a $(\mathcal{U}_a \cup \mathcal{U}_c)$-induced α-set $\mathcal{S} = \{S_1, S_1, \ldots, S_{|\mathcal{S}|}\}$. Let $\mathcal{U} = \emptyset$.
13: **for** each $S_j \in \mathcal{S}$ **do**
14: **if** $|F(S_j)| \geq \beta$, **then** $\mathcal{U} = \mathcal{U} \cup \{u \mid u \text{ is in } S_j\}$.
15: **else** Mark all h-edges between S_j and $\mathbb{P}$ as $\bar{r}$ (to be removed).
16: **end for**
17: Define a subgraph $\mathbb{C}'$ to be a $\mathcal{U}$-induced graph.
18: **for** each vertex $u_j \in \mathcal{U}_a$ **do**
19: Identify a path P in $\mathbb{C}'$ that visits u_j and forms a conjugate pair (as in definition 5) with its projection on $\mathbb{P}$.
20: **if** no such a path exists, **then** mark h-edges between u_j and $\mathbb{P}$ as $\bar{r}$.
21: **end for**
22: **end for**
23: Apply *ER* to h edges marked $\bar{r}$ and output all connected components.
24: **end for**

in every partition $\mathbb{P}$ for an extant genomic region that is part of a MGC. For any MGC that contains a genomic region in $\mathbb{P}$, we need to identify a conjugate pair: an α-run in $\mathbb{P}$ and an α-path that visits other partitions in $\mathbb{C}$. Lines 5–11 identify all subpartitions that could be involved in such an α-path, and lines 12–17 create a new graph $\mathbb{C}'$ that results in the former steps, where we search for the α-path in lines 17-21. Finally, we output resulting connected components in line 23, each denoting a MGC, and the ancestral gene families are inferred from gene pairs connected by h-edges.

In overall execution of the algorithm, lines 6–11 take $O(nf^{\epsilon_l})$ time, where n is the number of genes and f denotes the cardinality of the largest gene family in the input. In overall execution, lines 17–20 take $O(n^{\beta})$ time, which are the most time consuming steps in the algorithm, making the overall run time $O(n^{\beta})$. Although we permit shuffling of ancestral genes as a legitimate MGC, it can be shown that this leads the problem in line 19 to be NP-complete. Therefore, our algorithm limits the extent of rearrangement by constraining two adjacent ancestral genes in one extant genomic region to be at most $\beta\alpha$ distance apart in another. In practice, the total run time is mainly affected by two factors: ϵ_l and f, which control the density of graph $\mathbb{C}'$, as we could show that the probability for two vertices to be connected by an edge in $\mathbb{C}'$ is polynomial in f and exponential in ϵ_l. As ϵ_l is small in practice, the gene family size becomes a dominant factor. In current implementation, we limit the value of f and plan to relax this constraint in future work. Our algorithm runs fast in practice since a reasonable choice of parameters results in the graph $\mathbb{C}'$ being sparse. For all of our experiments in the following section, the algorithm terminates within 10 seconds on an Intel P4 desktop.

IV. Experimental Results

Although our method is applicable to multiple genomes (as we applied it to some specific studies between bacterial genomes, yeast, and some plant genomes, ranging from two to thirty), for comparison purpose, we demonstrate it using *E. coli* K12 and *B. subtilis*, where biologically verified operons are documented for *E. coli* [3], and compare with HomologyTeams model [4] as it has been shown to produce meaningful results on par with any other new methods. Note that although some models, such as median cluster [2], have higher sensitivity, the specificity which we also consider as very important is unknown. Too many false positives waste experimental resources and pose a significant problem. In addition, some newly developed methods tend to ignore fairness while making comparisons with the prior methods, which use very different parameters and distance metrics. This is likely due to the difficulties in finding suitable parametric conversions between different models. However, ignoring this forfeits the ability to accurately judge if a model has better biological significance. This issue is addressed in detail in our technical report [10], and for our comparison below, we choose a comparable setting between our model and the HomologyTeams model.

The HomologyTeams method (and many other gene cluster detection methods) uses a physical distance metric [4], [6] (the number of bases between genes) when detecting gene clusters, while our MGC model relies on the gene distance metric [1] (defined in section II-A). Our choice stems from our evolutionary model, where physical distance holds no obvious meaning. We limit the distance allowed between two genes using ϵ_i and ϵ_l in our model. The physical distance is limited by parameter δ. For both methods, β holds the same meaning. In addition, very large gene families are not used for path identification in MGC model as addressed previously. Note that, this does not introduce the bias to the comparison. Since any gene cluster that is captured by one model is compared with the results from the other, regardless of whether they contain any large gene families.

For the current comparison, we use $\delta = 3000$ for HomologyTeams model, since it gives a good result for HomologyTeams model (a good number of biologically valid operons and a limited number of unverified operons). This choice translates to $1 \leq \epsilon_i \leq 3$ empirically for our model, and we choose $\epsilon_i = 1$ such that the comparison is

Table I
COMPARISONS BETWEEN HOMOLOGYTEAMS AND MGC MODEL.

Model	predictions not made by the other method	True Positives	False Positives
HomologyTeams	13	5	8
MGC	22	16 + (5)	1

in favor of HomologyTeams model. In the conversion, we can ignore ϵ_l as it captures the gene loss rate, which is not captured in the HomologyTeams model. We set $\epsilon_l = 1$ as a reasonable choice. For both methods, we set $\beta = 3$, since a lower value results in biologically uninteresting clusters.

We applied both models on the same dataset that was previously used to validate the HomologyTeams software [4]. Running HomologyTeams software (with $\delta = 3000$ and $\beta = 3$) resulted in a total of 73 unique gene clusters, where each consists of two genomic regions. Running our software (implemented in C++ and Perl), using parameters $\epsilon_i = 1$, $\epsilon_l = 1$, and $\beta = 3$ resulted in a total of 53 unique MGCs. Note that although the number of MGCs is smaller than HomologyTeams cluster, the total number of genomic regions involved in MGCs is much bigger since each MGC could consist of more than 2 genomic regions. *Since currently, there exist no better ways of judging the specificity and sensitivity of a model, we will rely on the biologically verified operons in RegulonDB database [3] for evaluation*. We only focus on the different gene clusters resulted from the two models since only these are meaningful for the comparison purpose.

Due to space limitations, the full results are given in [10], and we only provide a brief summary in Table I: the second column refers to the number of gene clusters discovered by the corresponding model but missed by the other. These gene clusters are further compared with RegulonDB, where a gene cluster is considered as a true positive if it is found in the database, otherwise, a false positive.

As shown in the table, 8 HomologyTeams clusters are false positives, which unanimously contain large gene families. Although the remaining 5 HomologyTeams clusters are missed by MGC model, 3 of them contain large gene families that are not captured in the current implementation, but the rest are missed due to parameter constraints of MGC model, where two genes in the cluster are spaced by more than $\alpha = 2$ distance apart. Note that although by increasing the parameter ϵ_i to 2, these 2 operons will be detected by MGC model, we still consider these 2 operons as the false negatives for a fair comparison purpose.

On the other hand, among 22 unique MGCs, 5 (indicated in the parentheses) of them contains genomic regions only present in *B. subtilis*. Hence, we verify only the remaining 17 MGCs since RegulonDB only contains operon information for *E. coli*. The high true positive rate (16/17) indicates the 5 *B. subtilis* specific MGCs are potentially true operons. These sum up to 21 false negatives of HomologyTeams model.

As shown from above results, our method identified many more gene clusters with higher specificity. It is noteworthy to mention that the MGC model identified quite a few species specific clusters, which could not be identified by HomologyTeams model, in both *E. coli* and *B. subtilis*. These clusters correspond to functional units that were likely duplicated during the organisms' evolution.

V. CONCLUSIONS

We presented an improved model for gene cluster detection that formally captures the gene loss event, and directly infers multiple genomic regions engaged in a MGC even if pairwise signals are weak or absent, and presented an algorithm to do so. Our model limits the number of false positives and increases the chances of finding biologically relevant MGCs as shown by the preliminary results. We are working on improving the algorithmic efficiency and validating our model for synteny detection in plant genomes.

REFERENCES

[1] A. Bergeron, S. Corteel, and M. Raffinot. The algorithmic of gene teams. volume 2452 of *Lect N Bioinformat*, pages 464–476, 2002.

[2] S. Böcker, K. Jahn, J. Mixtacki, and J. Stoye. Computation of median gene clusters. In *RECOMB*, volume 4955 of *Lect Notes Comput Sc*, pages 331–345. Springer, 2008.

[3] S. Gama-Castro *et al.* RegulonDB (version 6.0): gene regulation model of *Escherichia coli K-12* beyond transcription, active (experimental) annotated promoters and textpresso navigation. *Nucleic Acids Res*, 36:D120–D124, 2008.

[4] X. He and M. H. Goldwasser. Identifying conserved gene clusters in the presence of homology families. *J Comput Biol*, 12(6):638–656, 2005.

[5] R. Hoberman and D. Durand. The incompatible desiderata of gene cluster properties. In *Comparative Genomics*, Lecture Notes in Bioinformatics, Dublin, Ireland, 2005. Springer-Verlag.

[6] S. Kim, J. H. Choi, and J. Yang. Gene teams with relaxed proximity constraint. In *Proc IEEE Comput Syst Bioinform Conf*, pages 44–55, 2005.

[7] Sven Rahmann and Gunnar W. Klau. Integer linear programs for discovering approximate gene clusters. In *WABI*, pages 298–309, 2006.

[8] B. Snel, P. Bork, and M. Huynen. The identification of functional modules from the genomic association of genes. *Proc Natl Acad Sci*, 99(9):5890–5895, Apr 2002.

[9] Q. Yang and S. Sze. Large-scale analysis of gene clustering in bacteria. *Genome Res*, 18(6):949–956, Jun 2008.

[10] X. Yang and S. Aluru. A unified model for multi-genome synteny and gene cluster inference (url: http://archives.ece.iastate.edu/archive/00000496/). Technical report, Iowa State university, 2009.

[11] Q. Zhu, Z. Adam, V. Choi, and D. Sankoff. Generalized gene adjacencies, graph bandwidth and clusters in yeast evolution. In *ISBRA*, pages 134–145, 2008.