Search for a Star: Approximate Gene Cluster Discovery Problem (AGCDP) as Minimization Problem on Graph

Jeffrey A. Aborot, Henry Adorna, Jhoirene B. Clemente,
Brian Kenneth de Jesus and Geoffrey Solano

Algorithms and Complexity Laboratory
Department of Computer Science
College of Engineering
University of the Philippines Diliman

{jeffrey.aborot, jbclemente, badejesu, gasolano}@up.edu.ph, ha@dcs.upd.edu.ph

ABSTRACT

Finding gene clusters in genomes is an essential process in establishing relationship among organisms. Gene clusters may express functional dependencies among genes and may give insight into expression of specific traits. The problem of finding gene clusters among several genomes is referred to as Gene Cluster Discovery and several models has already been formulated for its definition. One formulation of this problem is the Approximate Gene Cluster Discovery Problem (AGCDP) which is modelled as a combinatorial optimization problem in some works. In this paper we propose an approach which produces a transformation of AGCDP into a minimum-weight star finding problem in graph. Detailed examples are also presented to further clarify the notion of the transformation. Proof of equivalence is also presented in the paper to show the equivalence of input parameters of AGCDP and the construction of the graph representing the input parameters to the problem.

Keywords

Gene, genome, gene cluster, gene content, linear interval, genome, minimization, minimum-weight star, combinatorial optimization

1. INTRODUCTION

Gene clusters are set of genes that are closely related to each other. Genes belonging to a cluster may share functional dependencies and may be involved in the expression of a specific trait. Identifying gene clusters is also an essential step in establishing relationships between organisms as well as discovery of drug and treatments for diseases. The problem of identifying this set of genes is called *Gene Cluster Discovery*. This problem has been modelled several times, examples of which are presented in [3], [4], [5], where genes are modelled as integers and genomes are either permutations or sequences defined over the set of all genes. Models in [3] also takes into account gene clusters with (max-gap clusters) and without gaps (exact clusters).

The focus of this work is on the model presented in [5], where they define *Approximate Gene Cluster Discovery Problem (AGCDP)* as a combinatorial problem which identifies the set of genes that are kept "more or less" together across genome sequences. An Integer Linear Programming (ILP) formulation is also presented in [5]. Several modifications of the model, specifically on the objective function, is also presented to take into account characteristics of real biological data. Among these includes, absence of gene cluster occurrence in some of the input genomes, identification of valid gene clusters, and use of certain reference genome.

In this paper we will represent AGCDP as a graph problem. We will define how we transformed the set of inputs to a specific graph called G_{ACGDP}. Then we will discussed how the problem is reduced to finding minimum weight $star(u)$ in a graph. Two cases were both modelled in this paper. We consider scenarios with and without a given reference genome.

This paper is organized as follows. Section 2 presents a brief discussion of AGCDP as well as the naive and ILP formulation of the problem. Section 3 contains the detailed discussion of how AGCDP is represented as a graph problem. Proof of equivalence of the two representations is discussed in Section 4. Finally Section 5 concludes the paper.

2. APPROXIMATE GENE CLUSTER DISCOVERY PROBLEM

Necessary for our understanding of the problem are the following definitions.

1. **Gene** A *gene* is represented by an integer $g \in \mathcal{Z}^0$. Special genes represented by the integer 0 are genes with non existing homologs, with which we are not interested of in this problem.
2. **Gene Universe** The set of all unique genes is called the *gene universe* and is denoted by $\mathcal{U} = \{0, 1, 2, \ldots, N\}$.

3. **Genome** For simplicity, the genome of a certain individual can be represented as a single sequence of genes from a chromosome. For instance, $g^i = (g^i_1, g^i_2, \ldots, g^i_{n_i})$, where each g^i_j is the jth gene in the ith genome. In general, a set of genomes is represented by $\mathcal{G} = \{g^1, g^2, \ldots, g^t\}$ for some $t \in \mathcal{Z}^+$, where each g^i has length $n_i \in \mathcal{Z}^+$.

4. **Linear Interval** A *linear interval* J^i in a genome $g^i = (g^i_1, g^i_2, \ldots, g^i_{n_i})$ is an index set which can either be empty $J^i = \emptyset$ or $J^i = \{j, j+1, \ldots, k\}$, which can also be denoted as $J^i_{j_i,k_i}$ where $1 \leq j_i \leq k_i \leq n_i$. A set of linear intervals from all genomes is denoted by J.

5. **Gene Content** The *gene content* $G(J^i_{j,k})$ of a linear interval $J^i_{j,k}$ in genome g^i is the set of unique genes contained in that interval.

6. **Set Difference** The set difference between two distinct gene contents $G(J^i_{j_i,k_i})$ and $G(J^p_{j_p,k_p})$, $1 \leq i \leq t$, $1 \leq p \leq t$ and $i \neq p$, is defined as

$$G(J^i_{j_i,k_i}) \backslash G(J^p_{j_p,k_p}) = \{g | g \in G(J^i_{j_i,k_i}) \text{ and } g \notin G(J^p_{j_p,k_p})\}.$$

Given the definitions listed above, AGCDP aims to find a gene set $X \subset \mathcal{U}$, where $0 \notin X$ and a set of linear intervals $J = J^i_{j_i,k_i}$ where the gene content $G(J^i_{j_i,k_i})$ for each genome g^i is roughly equal to X. To define formally how close X is to the gene contents $G(J^i_{j_i,k_i})$, the number of missing genes $|X \backslash G(J^i_{j_i,k_i})|$ and additional genes $|G(J^i_{j_i,k_i}) \backslash X|$ are computed.

Let us formally define the Basic Approximate Gene Cluster Discovery Problem (AGCDP).

DEFINITION 1 (AGCDP [5]). *Given the gene universe $\mathcal{U} = \{0, 1, \ldots, N\}$, set of genomes $\mathcal{G} = \{g^1, g^2, \ldots, g^t\}$, size range $[D^-, D^+]$ or positive constant D, and integer weights w^- and w^+ corresponding to cost of missing and additional genes in an interval, identify $X \subset \mathcal{U}$, $0 \notin X$, $D^- \leq |X| \leq D^+$ or $|X| = D$ and a set of linear intervals $J = \{J^i_{j_i,k_i}\}$, $\forall\, i$ such that the cost function*

$$cost(X, J) = \sum_{i=1}^{t} [(w^- \cdot |X \backslash G(J^i_{j_i,k_i})|) + (w^+ \cdot |G(J^i_{j_i,k_i}) \backslash X|)]$$

*is **minimum**.*

AGCDP is a double minimization problem. In order to identify the cost of X, we identify the set of linear intervals J which minimizes the value of the objective function $cost(X, J)$. The naive way of identifying the set X is to check the cost of all possible $X \subset \mathcal{U}$, which is $\binom{N}{D}$ if $|X| = D$, otherwise

$$\sum_{\forall d} d \binom{N}{d}, \text{ where } D^- \leq d \leq D^+$$

if we have $D^- \leq |X| \leq D^+$. Also note that we have to identify the set of linear intervals for each genome which minimizes the cost given X. Naively this can be done by checking all possible linear intervals in each genome. The total running of time the naive AGCDP solver is the number of $X \subset \mathcal{U}$ satisfying the constraint $|X| = D$, times the running time of identifying the best linear interval given X. Thus, naive AGCDP solver takes $O(N!(n^{2t}))$.

EXAMPLE 1. *To illustrate Definition 1, suppose we have the set of genomes $\mathcal{G} = \{g^1, g^2, g^3, g^4\}$ defined over the gene universe $\mathcal{U} = \{0, 1, 2, 3, 4, 5, 6, 7\}$. The aim of AGCDP is to discover a set of genes $X \subset \mathcal{U}$, with cardinality $|X| = D = 3$ such $cost(X, J)$ is minimum. Let $\mathcal{G}$ be equal to the following set. In this example, gene 0 does not exist in at least one of the genomes.*

$$\begin{array}{lccccc} g^1: & 1 & 1 & 3 & 2 & 4 \\ g^2: & 3 & 2 & 1 & 4 & \\ g^3: & 5 & 6 & 1 & 4 & 2 \\ g^4: & 1 & 3 & 7 & & \end{array}$$

The naive approach of finding minimum X is to evaluate each X that satisfy the constraints $X \subset \mathcal{U}$ and $|X| = D$. In this case we have to evaluate

$$X = \{\{1, 3, 2\}, \{2, 3, 4\}, \{3, 4, 5\}, \ldots\},$$

where the total number of gene sets to be evaluated is $\binom{7}{3}$. Note that to evaluate the score of a certain gene set, another minimization procedure is needed, i.e. identification of the best linear interval for each genome.

After checking all possible linear interval and all possible gene cluster X, we see that $cost(X^*, J^*)$ is minimum where $X^* = \{1, 2, 3\}$ and $J^* = \{J^1_{1,4}, J^2_{1,3}, J^3_{3,5}, J^4_{1,2}\}$ as shown in the figure below.

$$\begin{array}{lccccc} g^1: & \mathbf{1} & \mathbf{1} & \mathbf{3} & \mathbf{2} & 4 \\ g^2: & \mathbf{3} & \mathbf{2} & \mathbf{1} & 4 & \\ g^3: & 5 & 6 & \mathbf{1} & \mathbf{4} & \mathbf{2} \\ g^4: & \mathbf{1} & \mathbf{3} & 7 & & \end{array}$$

Figure 1: The set of gene contents $\{\{1, 3, 2\}, \{3, 2, 1\}, \{1, 4, 2\}, \{1, 3\}\}$ corresponding to J^*.

Given that the penalties for missing and additional genes are the same, i.e. $w^+ = w^- = 1$, the computation for $cost(X^*, J^*)$ is as follows.

$$\begin{array}{lcl} cost(X^*, J^*) & = & \sum_{i=1}^{4} |X^* \setminus G^i_{J^*_i}| + |G^i_{J^*_i} \setminus X^*| \\ cost(X^*, J^*) & = & (0+0) + (0+0) + (1+1) + (1+0) \\ cost(X^*, J^*) & = & 3 \end{array}$$

The integer linear programming (ILP) formulation of AGCDP is presented in [5]. They defined the set of missing and additional genes in Equation 1 and 2 respectively.

$$|X \setminus G^i_{J_i}| = \sum_{q=0}^{N} (x_q - l^i_q) \tag{1}$$

$$|G^i_{J_i} \setminus X| = \sum_{q=0}^{N} (\mathcal{X}^i_q - l^i_q) \quad (2)$$

where x_q, l^q_i, and $\mathcal{X}^i_q$ are binary indicator vectors which pertains to the reference gene set X, the intersection of reference gene set and gene content $X \bigcap G^i_{J_i}$, and the gene content ${}^{i}_{J_i}$ respectively. Further details of their integer linear programming formulation is presented in [5].

A simpler approach to solve ACGDP is to make use of a *reference genome*. Note that in the objective function, we are comparing the gene set of each linear interval to a reference gene set X. With the use of a reference genome sequence, instead of evaluating all possible $X \subset \mathcal{U}$ which we can generate from the universal set, we only evaluate Xs which are present in the reference genome. This reduces the solution space of AGCDP. However, it is important to note that a global minimum X identified using the naive solution (considering all possible clusters in all the input genomes) may not be the X identified when we use a reference genome, i.e. X may not occur in at least one of the genomes in the naive solution. Since in some cases a reference genome is available, one may opt to use other specialized algorithms simpler than the ILP formulation presented in [5]. For instances in which reference genome sequences are available through comparison and alignment of genomes of several species, these reference genome sequences are used in the process of identifying new genome sequences.

3. AGCDP AS MINIMIZATION PROBLEM ON GRAPH

Given the definition of AGCDP we look into two different cases of the problem which we discuss in the succeeding subsections. The first case is when we consider one of the input genomes as a reference genome from which we generate gene contents which we treat as reference gene sets without loss of generality. The second case is when we do not assume a reference genome from the set of input genomes. Instead, we consider each gene content of a linear interval in any of the parts of the graph satisfying the constraint on $|X|$ as a possible solution gene set X. This lessens the number of possible reference gene set subsets of $\mathcal{U}$ which we need to evaluate.

3.0.1 Construction of Graph

Given gene universe $\mathcal{U} = \{0, 1, \ldots, N\}$, a set of genomes $\mathcal{G} = \{g^1, g^2, \ldots, g^t\}$, size range $|X| = [D^-, D^+]$ or $|X| = D$, and integer penalty weights w^- and w^+ for missing and additional genes respectively in a linear interval, we propose the following transformation of AGCDP input parameters into a t-partite edge-weighted undirected graph representation.

DEFINITION 2. *Define a t-partite edge-weighted undirected graph $G_{AGCDP} = (V, E)$, where*

$$V = \bigcup_{i=1}^{t} V_i.$$

1. *A part $V_i \subset V$ represents a genome $g^i \in \mathcal{G}$ where $1 \leq i \leq t$.*
2. *A function $v(.)$ is a bijection $v : G^i \longrightarrow V_i$ from the set of all gene contents $G^i = \{G(J^i_{j_i,k_i})\}$ in genome g^i to the set of all vertices in V_i $\forall\, i, 1 \leq i \leq t$. An evaluation of $v(G(J^i_{j_i,k_i}))$ is a vertex in graph G_{AGCDP} and an evaluation of $v(x_i)$ is a gene content of a linear interval in genome g^i.*
3. *An edge $e \in E$ is incident to vertices $x \in V_x$ and $y \in V_y$ if and only if $x \neq y$.*
4. *The weight $w_{x,y}$ assigned to an edge incident to vertices x and y is equal to*

$$w_{x,y} = \; w^- \cdot |v(x) \backslash v(y)| + w^+ \cdot |v(y) \backslash v(x)| \quad (3)$$

where the "\" is the set difference operator.

Given G_{AGCDP} as the graph representation of the input parameters of AGCDP, let us now define a collection of vertices $star(x)$ rooted at vertex $x \in V_j$ defined in [10]. Basically, $star(x)$ is a set of vertices from each part where the weight between the root and each vertex is minimum among all vertices in a part incident to the root. Below shows the formal definition of $star(x)$.

DEFINITION 3. *Given a vertex x in G_{AGCDP}, let $star(x)$ be a set of vertices of size $(t-1)$ such that x_i is rooted at $x \in V_u$, where*

$$x_i = arg\ min_{x_i}\ w_{x_i,x}\ \forall\ x_i \in V_i,\ i \neq u$$

.

We can evaluate a $star(x)$ by computing its weight as

$$weight(star(x)) = \sum_{i=1}^{t} w_{x_i,x},\ i \neq u \quad (4)$$

3.0.2 Minimization on Graph

Given the proposed transformation of the input parameters of AGCDP into its graph representation, we define AGCDP as a minimization problem on graph as follows.

DEFINITION 4. *Given a graph $G_{AGCDP} = (V, E)$ and a size range $[D^-, D^+]$ or positive constant D, find a vertex $x \in V_u$ and $star(x)$ such that $|v(x)| = D$ or $|v(x)|$ is in the range $[D^-, D^+]$ and*

$$weight(star(x)) = \sum_{i=1}^{t} w_{x_i,x},\ i \neq u$$

*is **minimum**.*

As discussed in section 2, AGCDP is a double minimization problem. This is also true for the graph problem transformation of AGCDP in which a vertex x mapped to a gene content (possible solution gene set X) and the set of vertices

x_i, $star(x)$, which minimizes the weight of each edge incident to pairs x, x_i is searched for in the graph. From among the found $star(x)$, the minimum-weight star is identified. From the resulting minimum-weight $star(x)$ we identify the gene content $G^u_{j_u,k_u}$ to which the root vertex x maps into. $G^u_{j_u,k_u}$ is the solution gene cluster. Note that there may be one or more solution gene cluster and so we may find one or more root x. Also, we identify the set $J^u_{j_u,k_u} \cup \{J^i_{j_i,k_i}\}$ as the set of linear intervals J where $\{J^i_{j_i,k_i}\}$ is the set of linear intervals to which $star(x)$ maps into and $J^u_{j_u,k_u}$ is the linear interval associated with the gene content $G^u_{j_u,k_u}$.

To give a more intuitive notion of the transformation discussed, we present examples of the two cases aforementioned with input parameters as specified in Example 1.

3.0.3 AGCDP with Reference Genome

Without loss of generality we assume genome g^1 to be the reference genome. All gene contents of linear intervals in g^1 which satisfy the size constraint $|X| = D$ will be considered as possible solution gene set X to the problem. These gene contents are mapped to vertices in part V_1 of the graph. These vertices are the roots of $star(x)$ which will be evaluated for weight. Assume that the penalty weight $w^- = w^+ = 1$ and the size constraint $|X| = 3$ as specified in Example 1.

EXAMPLE 2. *Given the set of genomes $\mathcal{G}$, we construct the graph G_{AGCDP}.*

1. *Identify the linear intervals $J^i_{j_i,k_i}$ for each genome g^i in $\mathcal{G}$.*

2. *Identify gene content $G(J^i_{j_i,k_i})$ for each linear interval $J^i_{j_i,k_i}$.*

3. *To each identified gene content $G(J^i_{j_i,k_i})$ a vertex $v(G(J^i_{j_i,k_i}))$ in G_{AGCDP} is mapped into.*

4. *For all pairs x_1, x_i, $x_1 \in V_1$, $x_i \in V_i$ and $i \neq 1$ define an edge $e_{x_1,x_i} \in E$. Define the weight assigned to each edge e_{x_1,x_i} as defined in (3) where $x = x_1$ and $y = x_i$.*

5. *For each $x \in V_1$ such that $|v(x)| = |G(J^1_{j_1,k_1})| = 3$ determine $star(x)$.*

$$\begin{aligned} star(v(G(J^1_{1,4}))) &= \{v(G(J^2_{1,3})), \\ &\quad v(G(J^3_{3,3})) \mid v(G(J^3_{5,5})) \mid \\ &\quad v(G(J^3_{3,5})), v(G(J^4_{1,2}))\} \\ star(v(G(J^1_{2,4}))) &= star(v(G(J^1_{1,4}))) \\ star(v(G(J^1_{3,5}))) &= \{v(G(J^2_{1,2})) \mid v(G(J^2_{1,4})), \\ &\quad v(G(J^3_{4,5})), v(G(J^4_{2,2}))\} \end{aligned}$$

   where the "|" symbol means "or".

6. *Determine minimum-weight $star(x)$ from among those*

g^i	$J_{j,k}$	$\{j\}$	$G(J_{j,k})$	$\{1,2,3\}$	$\{2,3,4\}$
g^2	$J_{1,1}$	(3)	{3}	2	2
	$J_{2,2}$	(2)	{2}	2	2
	$J_{3,3}$	(1)	{1}	2	4
	$J_{4,4}$	(4)	{4}	4	2
	$J_{1,2}$	(3, 2)	{2, 3}	1	1
	$J_{1,3}$	(3, 2, 1)	{1, 2, 3}	0	2
	$J_{1,4}$	(3, 2, 1, 4)	{1, 2, 3, 4}	1	1
	$J_{2,3}$	(2, 1)	{1, 2}	1	3
	$J_{2,4}$	(2, 1, 4)	{1, 2, 4}	2	2
	$J_{3,4}$	(1, 4)	{1, 4}	3	3
g^3	$J_{1,1}$	(5)	{5}	4	4
	$J_{2,2}$	(6)	{6}	4	4
	$J_{3,3}$	(1)	{1}	2	4
	$J_{4,4}$	(4)	{4}	4	2
	$J_{5,5}$	(2)	{2}	2	2
	$J_{1,2}$	(5, 6)	{5, 6}	5	5
	$J_{1,3}$	(5, 6, 1)	{1, 5, 6}	4	6
	$J_{1,4}$	(5, 6, 1, 4)	{1, 4, 5, 6}	5	5
	$J_{1,5}$	(5, 6, 1, 4, 2)	{1, 2, 4, 5, 6}	4	4
	$J_{2,3}$	(6, 1)	{1, 6}	3	5
	$J_{2,4}$	(6, 1, 4)	{1, 4, 6}	4	4
	$J_{2,5}$	(6, 1, 4, 2)	{1, 2, 4, 6}	3	3
	$J_{3,4}$	(1, 4)	{1, 4}	3	3
	$J_{3,5}$	(1, 4, 2)	{1, 2, 4}	2	2
	$J_{4,5}$	(4, 2)	{2, 4}	3	1
g^4	$J_{1,1}$	(1)	{1}	2	4
	$J_{2,2}$	(3)	{3}	2	2
	$J_{3,3}$	(7)	{7}	4	4
	$J_{1,2}$	(1, 3)	{1, 3}	1	3
	$J_{1,3}$	(1, 3, 7)	{1, 3, 7}	2	4
	$J_{2,3}$	(3, 7)	{3, 7}	3	3

Table 1: Identification of edge weights with respect to the gene contents (candidate solution gene set) $\{1,2,3\}$ and $\{2,3,4\}$. These are the gene contents of linear intervals in genome g^1 which satisfies the size constraint $|X| = 3$. Highlighted in the table are minimum weights for each reference gene content in relation to the gene contents of other genomes.

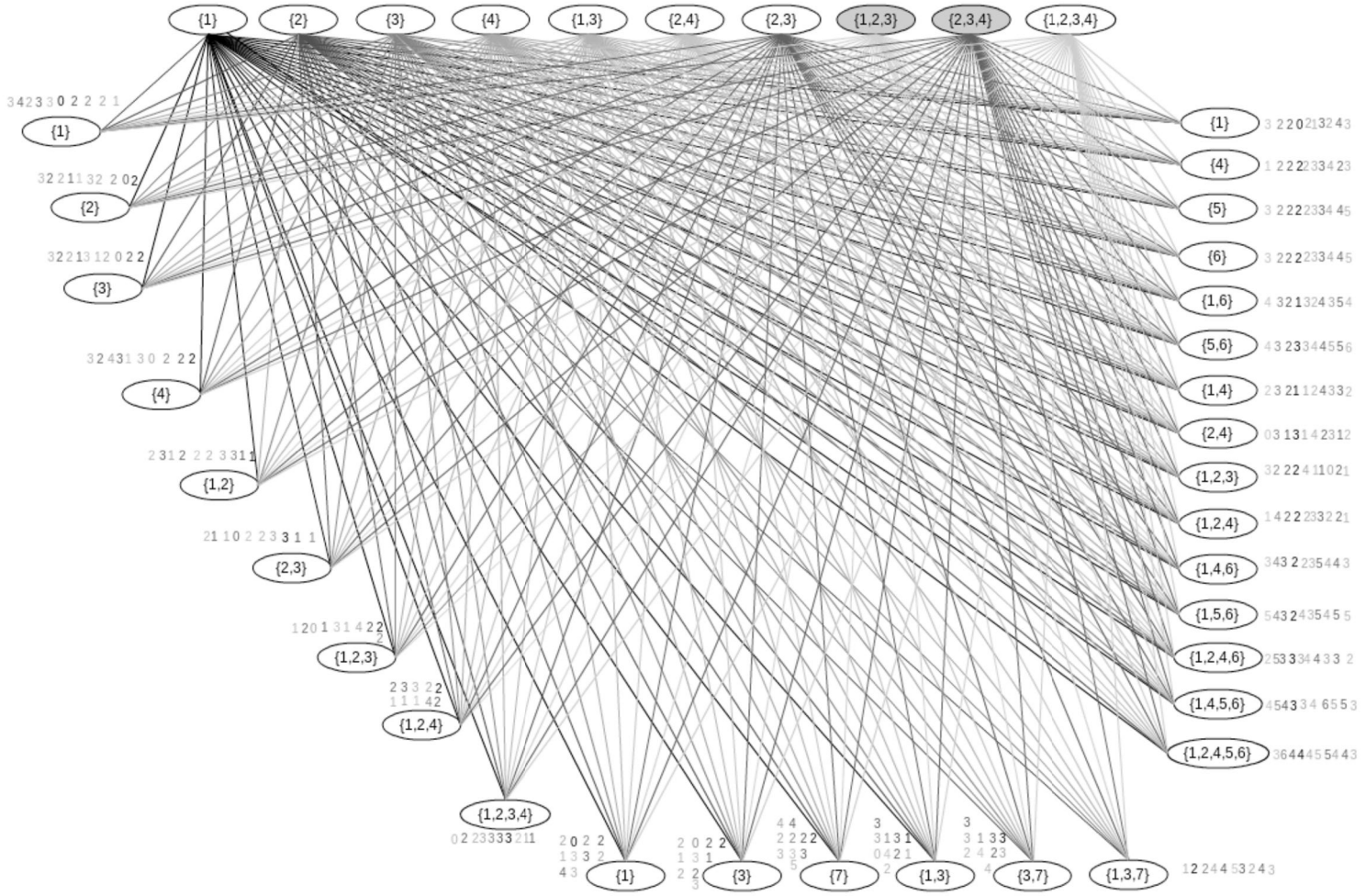

Figure 2: The constructed graph G_{AGCDP}. The vertices on top are elements of V_1, on the left are elements of V_2, on the right are elements of V_3 and on the bottom are elements of V_4. The vertices which are colored in yellow are the only vertices which satisfy the constraint $|X| = 3$ and therefore are the only valid roots for $star(x)$. The weight of each edge are indicated on the side of each vertex with color corresponding to the color of the edge.

identified in the previous step.

$$
\begin{aligned}
weight(star(v(G(J^1_{1,4})))) &= weight((v(G(J^1_{1,4})), \\
&\quad v(G(J^2_{1,3})))) + \\
&\quad weight((v(G(J^1_{1,4})), \\
&\quad v(G(J^3_{3,3})))) + \\
&\quad weight((v(G(J^1_{1,4})), \\
&\quad v(G(J^4_{1,2})))) \\
&= 0 + 2 + 1 \\
&= \mathbf{3} \\
weight(star(v(G(J^1_{2,4})))) &= star(v(G(J^1_{1,4}))) \\
weight(star(v(G(J^1_{3,5})))) &= weight((v(G(J^1_{3,5})), \\
&\quad v(G(J^2_{1,2})))) + \\
&\quad weight((v(G(J^1_{3,5})), \\
&\quad v(G(J^3_{4,5})))) + \\
&\quad weight((v(G(J^1_{3,5})), \\
&\quad v(G(J^4_{2,2})))) \\
&= 1 + 1 + 2 \\
&= 4
\end{aligned}
$$

In this example, we see that the minimum-weight star(x), with weight of 3, in graph G_{AGCDP} are the following star(x)

$$
\begin{aligned}
star(v(G(J^1_{1,4}))) &= \{v(G(J^2_{1,3})), \\
&\quad v(G(J^3_{3,3})), v(G(J^4_{1,2}))\} \\
&= \{v(G(J^2_{1,3})), \\
&\quad v(G(J^3_{5,5})), v(G(J^4_{1,2}))\} \\
&= \{v(G(J^2_{1,3})), \\
&\quad v(G(J^3_{3,5})), v(G(J^4_{1,2}))\} \\
star(v(G(J^1_{2,4}))) &= \{v(G(J^2_{1,3})), \\
&\quad v(G(J^3_{3,3})), v(G(J^4_{1,2}))\} \\
&= \{v(G(J^2_{1,3})), \\
&\quad v(G(J^3_{5,5})), v(G(J^4_{1,2}))\} \\
&= \{v(G(J^2_{1,3})), \\
&\quad v(G(J^3_{3,5})), v(G(J^4_{1,2}))\}
\end{aligned}
$$

The vertices $v(G(J^1_{1,4}))$ and $v(G(J^1_{2,4}))$ are then the roots of star(x) in G_{AGCDP} which has the minimum weight. The solution gene set X we are searching for must then be either of the gene contents to which vertices $v(G(J^1_{1,4}))$ and $v(G(J^1_{2,4}))$ are mapped into. Note that the gene content to which the vertex $v(G(J^1_{1,4}))$ is mapped into is $\{1,2,3\}$ and the gene content to which the vertex $v(G(J^1_{2,4}))$ is mapped into is also $\{1,2,3\}$. Given the minimum-weight star$(v(G(J^1_{1,4})))$ and star$(v(G(J^1_{2,4})))$ we see that the reference gene set

$$X = \{1,2,3\}$$

together with either of the sets of linear intervals

$$
\begin{aligned}
J_1 &= \{J^1_{1,4}, J^2_{1,3}, J^3_{3,3}, J^4_{1,2}\} \\
J_2 &= \{J^1_{1,4}, J^2_{1,3}, J^3_{5,5}, J^4_{1,2}\} \\
J_3 &= \{J^1_{1,4}, J^2_{1,3}, J^3_{3,5}, J^4_{1,2}\} \\
J_4 &= \{J^2_{1,3}, J^2_{1,3}, J^3_{3,3}, J^4_{1,2}\} \\
J_5 &= \{J^2_{1,3}, J^2_{1,3}, J^3_{5,5}, J^4_{1,2}\} \\
J_6 &= \{J^2_{1,3}, J^2_{1,3}, J^3_{3,5}, J^4_{1,2}\}
\end{aligned}
$$

form a solution to our example problem as what is concluded in Example 1.

3.0.4 AGCDP with no Reference Genome

In this case we do not choose a specific genome from the set of input genomes and instead consider all vertices mapped to gene contents which satisfy the input constraint $|X| = D$. As a result, all vertices from each part of G_{AGCDP} are adjacent to other vertices from other parts. This largely increases the number of edges in graph G_{AGCDP} which needs to be evaluated for weight. All gene contents of linear intervals in all genomes g^i which satisfy the size constraint $|X| = D$ or $|X|$ in the range $[D^-, D^+]$ will be considered as possible solution gene cluster X to the problem. These vertices are the roots of stars which will be evaluated for finding the solution to the problem. We also assume that the penalty weight $w^- = w^+ = 1$ and the size constraint $|X| = 3$ as defined in Example 1.

EXAMPLE 3. *Given the set of genomes $\mathcal{G}$, we construct the graph G_{AGCDP}.*

1. *Identify the linear intervals $J^i_{j_i,k_i}$ for each genome g^i in $\mathcal{G}$.*

2. *Identify gene content $G(J^i_{j_i,k_i})$ for each linear interval $J^i_{j_i,k_i}$.*

3. *To each identified gene content $G(J^i_{j_i,k_i})$ a vertex $v(G(J^i_{j_i,k_i}))$ in G_{AGCDP} is mapped.*

4. *For all pairs x, x_i, $x \in V_u$, $x_i \in V_i$ and $i \neq u$ define an edge $e_{x,x_i} \in E$. Define the weight assigned to each edge e_{x,x_i} as defined in (3) where $x = x$ and $y = x_i$.*

5. *For each vertex $x \in V_u$, $1 \leq u \leq t$, such that $|v(x)| = 3$ determine star(x). As what was done for Example 2, we determine the star(x) for the identified gene contents $v(x)$ of linear intervals from all genomes satisfy-*

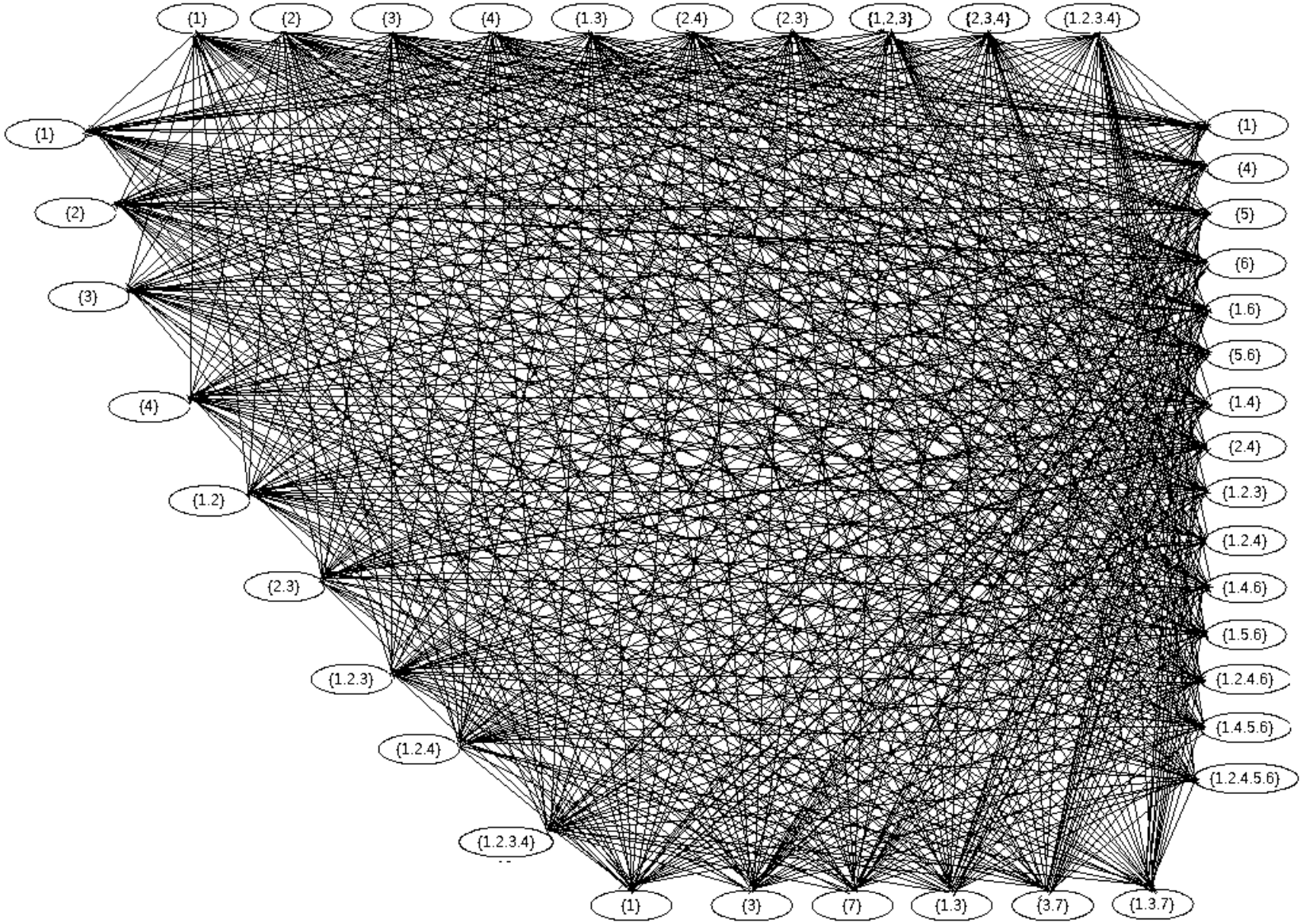

Figure 3: The constructed graph G_{AGCDP} for case 2. The set of vertices on the top are elements of V_1, on the left are elements of V_2, on the right are elements of V_3 and on the bottom are elements of V_4. All vertices are connected to all other vertices in the other parts of the graph. The weights of edges are omitted for clarity of the graph. The weights are identified in Table 2.

ing the size constraint $|X| = 3$.

$$
\begin{aligned}
star(v(G(J^1_{1,4}))) &= \{v(G(J^2_{1,3})),\\
&\quad v(G(J^3_{3,3})) \mid v(G(J^3_{5,5})) \mid\\
&\quad v(G(J^3_{3,5})), v(G(J^4_{1,2}))\}\\
star(v(G(J^1_{2,4}))) &= star(v(G(J^1_{1,4})))\\
star(v(G(J^1_{3,5}))) &= \{v(G(J^2_{1,2})) \mid v(G(J^2_{1,4})),\\
&\quad v(G(J^3_{4,5})), v(G(J^4_{2,2}))\}\\
star(v(G(J^2_{1,3}))) &= \{v(G(J^1_{1,4})) \mid v(G(J^1_{2,4})),\\
&\quad v(G(J^2_{3,3})) \mid v(G(J^2_{5,5})) \mid\\
&\quad v(G(J^2_{3,5})), v(G(J^4_{1,2}))\}\\
star(v(G(J^2_{2,4}))) &= \{v(G(J^1_{1,4})) \mid v(G(J^1_{2,5})) \mid\\
&\quad v(G(J^1_{4,5})), v(G(J^3_{3,5})),\\
&\quad v(G(J^4_{1,1}))\}\\
star(v(G(J^3_{1,3}))) &= \{v(G(J^1_{1,1})) \mid v(G(J^1_{2,2})) \mid\\
&\quad v(G(J^1_{1,2})), v(G(J^2_{3,3})),\\
&\quad v(G(J^4_{1,1}))\}\\
star(v(G(J^3_{2,4}))) &= \{v(G(J^1_{1,1})) \mid v(G(J^1_{2,2})) \mid\\
&\quad v(G(J^1_{1,2})) \mid v(G(J^1_{1,2})),\\
&\quad v(G(J^2_{3,4})), v(G(J^4_{1,1}))\}\\
star(v(G(J^3_{3,5}))) &= \{v(G(J^1_{1,5})) \mid v(G(J^1_{2,5})) \mid\\
&\quad v(G(J^1_{4,5})), v(G(J^2_{2,4})),\\
&\quad v(G(J^4_{1,1}))\}\\
star(v(G(J^4_{1,3}))) &= \{v(G(J^1_{1,3})) \mid v(G(J^1_{2,3})),\\
&\quad v(G(J^2_{1,1})) \mid v(G(J^2_{3,3})) \mid\\
&\quad v(G(J^2_{1,3})), v(G(J^3_{3,3}))\}
\end{aligned}
$$

where the $\mid$ *symbol means "or".*

6. *Determine minimum-weight* $star(x)$ *from among those identified in the previous step.*

$$
\begin{aligned}
weight(star(v(G(J^1_{1,4})))) &= 0+2+1\\
&= \mathbf{3}\\
weight(star(v(G(J^1_{2,4})))) &= 0+2+1\\
&= \mathbf{3}\\
weight(star(v(G(J^1_{3,5})))) &= 1+1+2\\
&= 4\\
weight(star(v(G(J^2_{1,3})))) &= 0+2+1\\
&= \mathbf{3}\\
weight(star(v(G(J^2_{2,4})))) &= 1+0+2\\
&= \mathbf{3}\\
weight(star(v(G(J^3_{1,3})))) &= 2+2+2\\
&= 6\\
weight(star(v(G(J^3_{2,4})))) &= 2+1+2\\
&= 5\\
weight(star(v(G(J^3_{3,5})))) &= 1+0+2\\
&= \mathbf{3}\\
weight(star(v(G(J^4_{1,3})))) &= 1+2+2\\
&= 5
\end{aligned}
$$

From the evaluation of weight of the identified $star(x)$ *the ones with minimum weight are* $star(v(G(J^1_{1,4})))$, $star(v(G(J^1_{2,4})))$, $star(v(G(J^2_{1,3})))$, $star(v(G(J^2_{2,4})))$ *and* $star(v(G(J^3_{3,5})))$. *From these* $star(x)$ *we identify the solution gene cluster* $\{1,2,3\}$ *from* $v(G(J^1_{1,4}))$, $v(G(J^1_{2,4}))$ *and* $v(G(J^2_{1,3}))$, *and gene cluster* $\{1,2,4\}$ *from* $v(G(J^2_{2,4}))$ *and* $v(G(J^3_{3,5}))$. *Notice that the solution gene cluster* $\{1,2,4\}$ *was not identified in the case wherein a reference genome was assumed. This is a consequence of assuming a reference genome since the possible solution gene clusters which can be found are limited to the available gene contents in the chosen reference genome.*

4. PROOF OF EQUIVALENCE

We show the equivalence of AGCDP with the proposed transformation into graph problem of finding minimum-weight $star(x)$ in an edge-weighted undirected t-partite graph. We show the equivalence for the presented two cases of AGCDP wherein the first case we identify a reference genome from among the input genomes and the second case wherein no reference genome is identified. We refer to Definition 1, 2, 3, and 4 as basis for the following proofs.

With Reference Genome

Without loss of generality we identify genome g^1 as the reference genome. Recall that for any vertex $x \in V_1, x = v(G(J^1_{j_1,k_1}))$ such that $1 \leq j \leq k \leq n_1$. Equation 3 can then be written as

$$w_{\mathrm{x,y}} = w^- \cdot |G(J^1_{j_1,k_1}) \backslash G(J^i_{j_i,k_i})| + w^+ \cdot |G(J^i_{j_i,k_i}) \backslash G(J^1_{j_1,k_1})|$$

and the total weight of a $star(x)$ defined in (4) can then be written as

$$\sum_{i \neq 1} w^- \cdot |G(J^1_{j_1,k_1}) \backslash G(J^i_{j_i,k_i})| \; + \; w^+ \cdot |G(J^i_{j_i,k_i}) \backslash G(J^1_{j_1,k_1})|$$

If we let X represent any possible solution gene cluster $G^1_{j_1,k_1}$, the previous equation is equal to

$$\sum_{i \neq 1} w^- \cdot |X \backslash G(J^i_{j_i,k_i})| \; + \; w^+ \cdot |G(J^i_{j_i,k_i}) \backslash X|$$

Note that this is just the objective function $cost(X, J)$ to be minimized for AGCDP with the first genome, g^1, defined as reference genome. It is easy to see that a unique pair of gene cluster X and set of linear intervals J which minimizes the objective function $cost(X, J)$ in AGCDP with reference genome, (X^*, J^*), is mapped to a unique pair of vertex x and a set of vertices $star$(x), $(x, star(x))$, which also minimizes the objective function $weight(star(\mathrm{x}))$ for the minimum-weight $star$(x) finding problem in an edge-weighted undirected t-partite graph G_{AGCDP}.

g^i	$J_{j,k}$	$\{j\}$	$G(J_{j,k})$	$g^1:\{1,2,3\}$	$g^1:\{2,3,4\}$	$g^2:\{1,2,3\}$	$g^2:\{1,2,4\}$	$g^3:\{1,5,6\}$	$g^3\{1,4,6\}$	$g^3:\{1,2,4\}$	$g^4:\{1,3,7\}$
g^1	$J_{1,1}$	(1)	$\{1\}$			2	2	2	2	2	2
	$J_{2,2}$	(1)	$\{1\}$			2	2	2	2	2	2
	$J_{3,3}$	(3)	$\{3\}$			2	4	4	4	4	2
	$J_{4,4}$	(2)	$\{2\}$			2	2	4	4	2	4
	$J_{5,5}$	(4)	$\{4\}$			4	2	4	2	2	4
	$J_{1,2}$	(1,1)	$\{1\}$			2	2	2	2	2	2
	$J_{1,3}$	(1,1,3)	$\{1,3\}$			1	3	3	3	3	1
	$J_{1,4}$	(1,1,3,2)	**{1,2,3}**			0	2	4	4	2	2
	$J_{1,5}$	(1,1,3,2,4)	$\{1,2,3,4\}$			1	1	5	3	1	3
	$J_{2,3}$	(1,3)	$\{1,3\}$			1	3	3	3	3	1
	$J_{2,4}$	(1,3,2)	**{1,2,3}**			0	2	4	4	2	2
	$J_{2,5}$	(1,3,2,4)	$\{1,2,3,4\}$			1	1	5	3	1	3
	$J_{3,4}$	(3,2)	$\{2,3\}$			1	3	6	5	3	3
	$J_{3,5}$	(3,2,4)	**{2,3,4}**			2	2	6	4	2	4
	$J_{4,5}$	(2,4)	$\{2,4\}$			3	1	5	3	1	5
g^2	$J_{1,1}$	(3)	$\{3\}$	2	2			4	4	4	2
	$J_{2,2}$	(2)	$\{2\}$	2	2			4	4	2	4
	$J_{3,3}$	(1)	$\{1\}$	2	4			2	2	2	2
	$J_{4,4}$	(4)	$\{4\}$	4	2			4	2	2	4
	$J_{1,2}$	(3,2)	$\{2,3\}$	1	1			5	5	3	3
	$J_{1,3}$	(3,2,1)	**{1,2,3}**	0	2			4	4	2	2
	$J_{1,4}$	(3,2,1,4)	$\{1,2,3,4\}$	1	1			5	3	1	3
	$J_{2,3}$	(2,1)	$\{1,2\}$	1	3			3	3	1	3
	$J_{2,4}$	(2,1,4)	**{1,2,4}**	2	2			4	2	0	4
	$J_{3,4}$	(1,4)	$\{1,4\}$	3	3			3	1	1	3
g^3	$J_{1,1}$	(5)	$\{5\}$	4	4	4	4				4
	$J_{2,2}$	(6)	$\{6\}$	4	4	4	4				4
	$J_{3,3}$	(1)	$\{1\}$	2	4	2	2				2
	$J_{4,4}$	(4)	$\{4\}$	4	2	4	2				4
	$J_{5,5}$	(2)	$\{2\}$	2	2	2	2				4
	$J_{1,2}$	(5,6)	$\{5,6\}$	5	5	5	5				5
	$J_{1,3}$	(5,6,1)	**{1,5,6}**	4	6	4	4				4
	$J_{1,4}$	(5,6,1,4)	$\{1,4,5,6\}$	5	5	5	3				5
	$J_{1,5}$	(5,6,1,4,2)	$\{1,2,4,5,6\}$	4	4	4	2				6
	$J_{2,3}$	(6,1)	$\{1,6\}$	3	5	3	3				3
	$J_{2,4}$	(6,1,4)	**{1,4,6}**	4	4	4	2				4
	$J_{2,5}$	(6,1,4,2)	$\{1,2,4,6\}$	3	3	3	1				5
	$J_{3,4}$	(1,4)	$\{1,4\}$	3	3	3	1				3
	$J_{3,5}$	(1,4,2)	**{1,2,4}**	2	2	2	0				4
	$J_{4,5}$	(4,2)	$\{2,4\}$	3	1	3	1				5
g^4	$J_{1,1}$	(1)	$\{1\}$	2	4	2	2	2	2	2	
	$J_{2,2}$	(3)	$\{3\}$	2	2	2	4	4	4	4	
	$J_{3,3}$	(7)	$\{7\}$	4	4	4	4	4	4	4	
	$J_{1,2}$	(1,3)	$\{1,3\}$	1	3	1	3	3	3	3	
	$J_{1,3}$	(1,3,7)	**{1,3,7}**	2	4	2	4	4	4	4	
	$J_{2,3}$	(3,7)	$\{3,7\}$	3	3	3	5	5	5	5	

Table 2: Evaluation of edge weights with respect to the possible solution gene contents identified from all the input genomes. These gene contents are emphasized under the column $G^i_{j_i,k_i}$. The minimum weights are also highlighted in the table.

We have shown that the objective function $weight(star(\mathrm{x}))$ to be minimized for the minimum-weight $star(\mathrm{x})$ finding problem in graph G_{AGCDP} is just equal to the objective function $cost(X, J)$ to be minimized for AGCDP with genome g^1 set as the reference genome. The equivalence from objective function of AGCDP to the objective function of the minimum-weight $star(\mathrm{x})$ finding problem can be shown by constructing graph G_{AGCDP} from the input parameters of AGCDP. □

With No Reference Genome
For the case wherein a reference genome is not chosen in AGCDP we can extend the proof for AGCDP with a reference genome to look also into the inter-genome similarity of gene contents of linear intervals. We consider each gene content in each genome satisfying the size constraint as a possible solution gene cluster for AGCDP instead of enumerating all the gene set $X \subset \mathcal{U}$ such that $|X| = D$ or $|X|$ is in $[D^-, D^+]$.

Since we are looking into the inter-genome relationship of gene contents, an edge in $G_{AGCDP} = (V, E)$ is defined such that for all pairs $x, y,\ x \in V_u, y \in V_i,$ and $i \neq u, (x, y) \in E$.

Recall that for any vertex $x \in V, x = v(G(J^i_{j_i,k_i}))$ for $1 \leq i \leq t$. Equation 3 can then be written as

$$w_{\mathrm{x,y}} = w^- \cdot |G(J^u_{j_u,k_u}) \backslash G(J^i_{j_i,k_i})| + w^+ \cdot |G(J^i_{j_i,k_i}) \backslash G(J^u_{j_u,k_u})|$$

and the total weight of a $star(x)$ defined in 4 can then be written as

$$\sum_{i \neq u} w^- \cdot |G(J^u_{j_u,k_u}) \backslash G(J^i_{j_i,k_i})| \ + \ w^+ \cdot |G(J^i_{j_i,k_i}) \backslash G(J^u_{j_u,k_u})|$$

Suppose that vertex $x = v(G(J^u_{j_u,k_u}))$ in the definition of $star(x)$ in Definition 3. $G(J^u_{j_u,k_u})$ then is a possible solution gene cluster. If we denote $G(J^u_{j_u,k_u})$ as X in the above equation, the total weight of a $star(x)$ can then be written as

$$\sum_{i \neq u} w^- \cdot |X \backslash G(J^i_{j_i,k_i})| \ + \ w^+ \cdot |G(J^i_{j_i,k_i}) \backslash X|$$

Notice that this is also just the objective function $cost(X, J)$ to be minimized for AGCDP. A unique pair of gene cluster X and set of linear intervals J which minimizes the objective function $cost(X, J)$ in AGCDP, (X^*, J^*), is mapped to a unique pair of vertex x and a set of vertices $star(\mathrm{x})$, $(x, star(x))$, which also minimizes the objective function $weight(star(\mathrm{x}))$ for the minimum-weight $star(\mathrm{x})$ finding problem in an edge-weighted undirected t-partite graph G_{AGCDP}. The equivalence of the transformation from AGCDP to the minimum-weight star finding problem in graph G_{AGCDP} can be shown by constructing G_AGCDP from the input paramaters of AGCDP or by reversing the presented proof. □

5. CONCLUSION

In this paper the authors presented the Approximate Gene Cluster Problem (AGCDP) as a combinatorial optimization problem. A proposed transformation of AGCDP into a minimum-weight $star(x)$ finding problem in an edge-weighted undirected t-partite graph was presented both for the cases wherein a reference genome is assumed from the set of input genomes and the case wherein no reference genome is identified. Also, proof of equivalence of the transformation from AGCDP to the grap problem were presented for both cases and it was shown that the objective function in AGCDP subject to minimization is equivalent to the objective function in the graph problem. It was observed that there is a possibility that some solution gene clusters within the input genomes may not be found when a reference genome from the input genomes is assumed. That is the case for the solution gene cluster $\{1, 2, 4\}$ in Example 3. It is worth to further investigate other approaches to identifying solution gene clusters given gene contents from all genomes, *i.e.* consensus gene cluster. The authors hope that by presenting the proposed transformation of AGCDP into a graph problem further results on AGCDP will materialize based on what was presented by the authors in this paper.

6. ADDITIONAL NOTES

Finding Nearby Stars
We could widen the range of the solution space to finding minimum-weight $star(x)$ in a graph to include $star(x)$ which are within the neighborhood of the best solution. Same with the idea of neighborhood in AGCDP as discussed in [5], a threshold γ can be identified such that given the weight $weight(star(x))$ of the $star(x)$ with the minimum weight,

$$weight(star(x_i)) \leq (1 + \gamma)(weight(star(x)))$$

such that $\gamma > 0$.

7. FUTURE WORKS

For future works on this topic, the authors aim to achieve the following:

- Classification of the complexity of the graph problem
- Enhancements on the proposed transformation from AGCDP to minimum-weight $star(x)$ finding problem
- Proposal of algorithm and optimization on finding minimum-weight star in a graph, both for cases with reference genome and without reference genome
- Evaluation of the proposed transformation based on variations of AGCDP
- Investigate the idea of Consensus Gene Cluster in identifying solution gene clusters given gene contents from all input genomes

8. REFERENCES

[1] R. Karp, Reducibility Among Combinatorial Problems, Complexity of Computer Computations, 1972

[2] P. Pevzner, S. Sze, Combinatorial Approaches to Finding Subtle Signals in DNA Sequences, American Association for Artificial Intelligence, 2000

[3] Bergeron A, Corteel S, Raffinot M., The Algorithmic of Gene Teams, Workshop on Algorithms in Bioinformatics (WABI), Vol. 2452 of LNCS, pp. 464-476, 2002

[4] Hoberman R, Durand D. The Incompatible Desiderata of Gene Cluster Properties, In: McLysaght A, Huson DH, editors. Comparative Genomics: RECOMB 2005 International Workshop, Vol. 3678 of LNCS, pp. 73-87, 2005

[5] S. Rahmann, G. Klau, Integer Linear Programming Techniques for Discovering Approximate Gene Clusters, Bioinformatics Algorithms, Techniques and Applications, Wiley-Interscience, 2008

[6] N. Deo, Graph Theory with Applications to Engineering and Computer Science, Prentice Hall, 1974

[7] T. Cormen, C. Leiserson, R. Rivest, C. Stein, Introduction to Algorithms, 3rd Edition, The MIT Press, 2009

[8] M. Garrey, D. Johnson, Computers and Intractability, Bell Telephone Laboratories, Inc., 1979

[9] A. Gibbons, Algorithmic Graph Theory, Cambridge University Press, 1985

[10] E. Zaslavsky, M. Singh, A combinatorial Optimization Approach for Diverse Motif Finding Applicationss, Algorithms for Molecular Biology, 2006