GRAPH BASED ANALYSIS FOR GENE SEGMENT ORGANIZATION IN A SCRAMBLED GENOME

MUSTAFA HAJIJ, NATAŠA JONOSKA, DENYS KUKUSHKIN, AND MASAHICO SAITO

ABSTRACT. DNA rearrangement processes recombine gene segments that are organized on the chromosome in a variety of ways. The segments can overlap, interleave or one may be a subsegment of another. We use directed graphs to represent segment organizations on a given locus where contigs containing rearranged segments represent vertices and the edges correspond to the segment relationships. Using graph properties we associate a point in a higher dimensional Euclidean space to each graph such that cluster formations and analysis can be performed with methods from topological data analysis. The method is applied to a recently sequenced model organism *Oxytricha trifallax*, a species of ciliate with highly scrambled genome that undergoes massive rearrangement process after conjugation. The analysis shows some emerging star-like graph structures indicating that segments of a single gene can interleave, or even contain all of the segments from fifteen or more other genes in between its segments. We also observe that as many as six genes can have their segments mutually interleaving or overlapping.

1. INTRODUCTION

It has long been observed that genome rearrangement processes on an evolutionary scale can lead to speciation [15], while on developmental scale they often involve DNA deletions [3, 35] as well as wholescale programmed rearrangements [31, 36]. For example, the highly diverse collection of antibodies often is attributed to somatic DNA recombination [38], and rearrangements on a chromosomal levels can be observed during homologous recombination [33]. In recent years there are numerous observations of alternative splicing where rearranging patterns of exons and introns of a single gene can produce different protein variants from a single mRNA (e.g. [25]). Rearranging segments of nucleotide sequences can be organized in a variety of arrangements on the locus, for example, they can be overlapping or interleaving [5]. *Oxytricha trifallax* is a single cell organism that is often taken as a model organism to study DNA rearrangement processes. This, and similar species of ciliates undergo massive restructuring of a germline micronuclear DNA during development of a somatic macronucleous. Recent sequencing and annotation of the whole *O. trifallax* genome allow genome level studies [8, 13]. Scrambling patterns within thousands of genes were observed revealing hidden structures among the scrambled gene/nanochromosome segments that explain over 95% of the scrambled genome [7]. While those studies were focused on scrambled recurrent patterns within a single gene, in this paper we study inter-gene segment arrangements. Through clustering techniques we identify patterns in which segments of different genes interleave or overlap throughout the genome. We represent a micronuclear locus with the interleaving and overlapping gene segments by a directed graph. Such obtained graph data is then converted to a set of points (point cloud) in a Euclidean space and we apply topological data analysis techniques to obtain clusters of similar graphs. This method was applied to the whole genome data of *Oxytricha trifallax* [8, 13].

In the last decade, Topological Data Analysis (TDA) has shown to be another tool for data analysis and data mining that can be used to extract topological information from various types of data [9]. TDA originated from computational topology with ideas inspired from statistics and computer science. One of the notable tools in TDA is *persistence homology*. Persistence homology, or for brevity PH, was introduced by Edelsbrunner et al. [17] and later studied further by Zomorodian and Carlsson [41]. The theory has been studied extensively since then and many theoretical advances have been made [10–12]. Persistence

NJ was partially supported by National Science Foundation CCF-1526485 and National Institutes of Health R01GM109459. MS was partially supported by National Institutes of Health R01GM109459.

Homology is defined in all dimensions ≥ 0, but clustering analysis, as used in this paper, uses only dimension 0. More details on TDA and persistence homology can be found in [9, 16, 22].

Due to advances in bioscience and biotechnology, the growth of biomolecular data has exploded and many data analysis algorithms have been developed aiming to better understand the generated data [14, 19, 28]. Data analysis using topological methods has proven to be useful in showing general patterns that were difficult to observe with other techniques. TDA methods have been used recently in many applications including protein structure identification [20, 40], aggregation models for animal behavior [1], and fullerene stability [39].

Our data set consists of directed graphs $\mathscr{G} = \{G_1, \ldots, G_n\}$ which we convert to a set of points in a Euclidean space, point cloud. $S = \{P(G_1), \ldots, P(G_n)\} \subset \mathbb{R}^n$. For the clustering analysis we applied TDA with PH at dimension zero to the obtained point cloud in $\mathbb{R}^n$. This process is described in Section 2.

Although this paper is focused on analysis of a specific data of interleaving/overlapping gene segments, the method that we propose for converting a graph data to a point cloud data is novel and general, and can be applied to analyze similarities in various graph data.

2. Method and Data Construction

2.1. **Gene Segment Maps in *Oxytricha trifallax*.** *Oxytricha trifallax* is a species of ciliate used as a model organism to study genome rearrangements. It undergoes massive genome rearrangements during the development of a somatic macronucleus (MAC) specializing in gene expression from an archival germline micronucleus (MIC) [13]. Within this process, over 16,000 macronuclear nanochromosomes assemble through DNA processing events involving global deletion of 90-95% of the germline DNA, effectively eliminating nearly all so-called "junk" DNA, including intervening DNA segments (internally eliminated sequences, IESs). Because these IES segments interrupt the coding regions of the precursor macronuclear gene loci in the micronucleus, each macronuclear gene may appear as several nonconsecutive segments (macronuclear destined sequences, MDSs) in the micronucleus. Moreover, the precursor order of these MDS segments for thousands of genes can be permuted or inverted in the micronucleus such that during the macronuclear development, all IESs are deleted and the MDSs are rearranged to form thousands of gene-sized chromosomes.

In [13] and later in [5], it was observed that an IES between consecutive MDSs of one gene can contain MDS segments from other genes, and that this process can be nested. Furthermore MDSs from different MAC genes can overlap or one MDS can be a subsegment of an MDS of another gene.

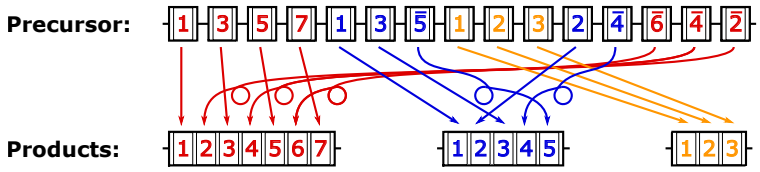

FIGURE 1. Rearrangement of gene segments in *Oxytricha trifallax*

The interleaving situation is schematically depicted in Figure 1. MDSs are represented by colored boxes with numbers. This example illustrates a MIC contig that has MDSs of three MAC contigs, each indicated with a different color. The numbers within the boxes indicate the order of the MDSs in the corresponding MAC contig. The barred numbers indicate MDSs in a reverse orientation (inverted) in the MIC contig relative to the ordering of the MDSs in the corresponding MAC contig. The thin black lines between the squares indicate IESs.

The MDS interrelationship analyzed in this paper uses the genome sequencing data in [13], and can be downloaded from the Supplemental Information in [13] and also in [7]. The data used for analysis in this paper is the processed data reported in [7]. This data was filtered so that consecutive MDSs of a single MAC contig that overlap or have no nucleotide gap (are adjacent) in the MIC are merged into a

singe MDS (correcting to possible previous annotation artifacts). In addition, we excluded MAC contigs that contain segments that are distant in the MAC contig but overlap in a MIC contig, as well as MAC contigs that are alternative fragmentations of longer MAC contigs. Both of these cases could be considered artifacts of the MIC-MAC maps although we cannot exclude the possibility that these cases are genuine to the data. We refer to such processed data as data $\mathscr{D}$ available at `http://knot.math.usf.edu/data/scrambled_patterns/processed_annotation_of_oxy_tri.gff`.

2.2. **Graphs Corresponding to MIC Contings.** As MDS from different MAC contigs can overlap or interleave in a MIC contig we define the following types of relationships between MAC contigs located on a single MIC contig.

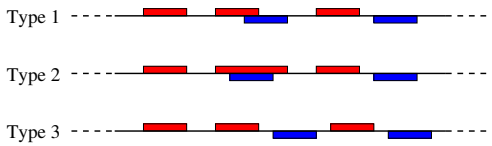

FIGURE 2. Three types of MDS segment organization of different MAC contigs. MDSs of the same MAC contig are colored the same.

- (Type 1 : Overlapping) If an MDS of a MAC contig g_1 overlaps with an MDS of another, distinct MAC contig g_2, then it is said that g_1 and g_2 *overlap*, or they are *overlapping*. We also say that g_1 has type 1 interaction with g_2, or g_1 has interaction of type 1 with g_2.

  Two MAC contigs are considered to be overlapping if they have at least one pair of MDSs that overlap with at least 20bp in common. This is because two consecutive MDSs of the same MAC contig usually share sequences at their ends (pointers) that guide the rearrangement process [31], and two MDSs from distinct MAC contigs can share the same pointer sequence. The length of these pointer sequences usually ranges between 2 to 20 nucleotides.

  The overlapping relation is symmetric, if g_1 overlaps with g_2, then g_2 overlaps with g_1. The situation is depicted in Figure 2 (Type 1). In the figure, MDSs of g_1 and g_2 are represented by blue and red rectangles, respectively. This case excludes the case when one MDS is completely included in another, even though being a subsequence is a particular type of "overlapping". Such situation is included in Type 2 case (below).
- (Type 2 : Containment) If an MDS of a MAC contig g_1 is contained in (is a subsegment of) an MDS of another distinct MAC contig g_2, then it is said that an MDS of g_1 is contained in an MDS of g_2, and we say g_1 has type 2 interaction with g_2.

  For this interaction when an MDS M of g_2 contains an MDS M' of g_1, we require that both ends of M have at least 5 bps that are not in common with M'. That is, we require a complete inclusion such that there are no pointer sequences in common. In Figure 2 (Type 2), MDSs of g_1 are depicted in blue, and those for g_2 in red. This relation is not symmetric. We distinguish this situation from the one in Type 1 because the unscrambling of an MDS that is next to at least one IES (Type 1) may use a different biological process involving Piwi-interacting RNA [18] rather than one that does not neighbor an IES (g_1 in Type 2).
- (Type 3 : Interleaving) If an IES of a MAC contig g_1 contains an MDS of another, distinct MAC contig g_2, then it is said that an MDS of g_1 interleaves (or is interleaving) with g_2, or g_1 has Type 3 interaction with g_2. We allow that the ends of an interleaving MDS of g_1 and the MDSs of g_2 to intersect (overlap) up to (including) 5 bases. This requirement distinguishes type 3 case from the 'overlapping' case where the requirement is at least 20 bases.

We consider pairs (g_1, g_2) of MAC contigs g_1 and g_2 that belong to the same MIC contig. To each pair of MAC contigs (g_1, g_2) we associate a triple $c(g_1, g_2) = (b_1, b_2, b_3)$ where each entry b_i $(i = 1, 2, 3)$ indicates

whether g_1 is in relationship of Type i with g_2. The value of b_i is either 0 (there is no relationship of Type i) or 1 (g_1 is related to g_2 of with Type i).

To investigate the situations of these three types of interactions of MDSs, we associate a directed graph with labeled edges to each MIC contig in data $\mathscr{D}$ as follows.

Each graph $G = G_M = (V(G_M), E(G_M))$, which may be disconnected and have multiple connected components, corresponds to a MIC contig M. Each vertex $g \in V(G_M)$ corresponds to a MAC contig g whose MDSs are segments of the MIC contig M.

A labeled directed edge (g_1, c, g_2) is in $E(G_M)$ if $c = c(g_1, g_2) \neq (0,0,0)$.

In the figures below we use colors on the edges to indicate the labels of the edges: red= $(1,1,1)$, green= $(1,1,0)$, blue= $(1,0,1)$, orange= $(0,1,1)$, purple= $(1,0,0)$, cyan= $(0,1,0)$, and black= $(0,0,1)$.

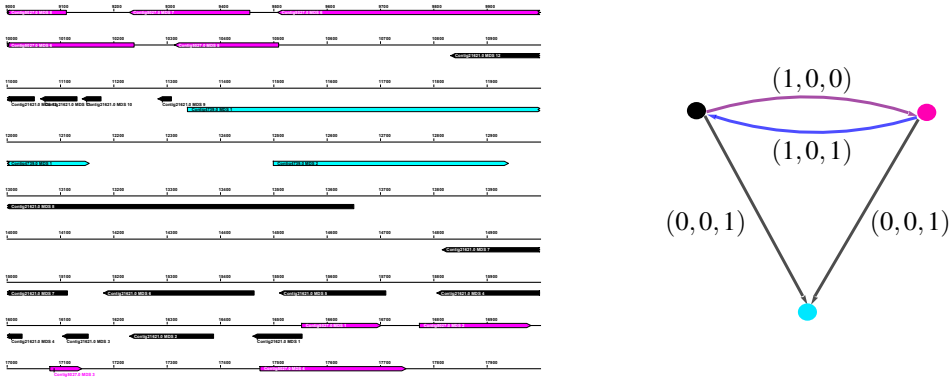

FIGURE 3. A segment of MIC contig ctg7180000069854 (left) and its corresponding graph (right). There is a relatively short overlap of MDSs of the black and purple contigs and the purple contig interleaves with the black (hence there is a blue edge indicating label $(1,0,1)$ from the purple to the black vertex) but there is no MDS segment of the black contig that interleaves with the purple contig, so there is a purple edge indicating $(1,0,0)$ in the opposite direction. IESs of both black and purple contigs contain MDSs of the cyan contig, so there are two black edges indicating labels $(0,0,1)$ ending at the cyan vertex.

Figure 3 shows a locus of the MIC contig ctg7180000069854 containing MDSs of three MAC contigs 5027.0 (purple), 21621.0 (black), and 4739.0 (cyan). Figures 15 and 28 depict some other examples of graphs for MIC contigs in the data.

The set of graphs corresponding to the data $\mathscr{D}$ that is analyzed here is denoted $\mathscr{G}_{\mathscr{D}}$ such that $\mathscr{G}_{\mathscr{D}} = \{G_M \mid M$ is a MIC contig in $\mathscr{D}\}$. There are 629 distinct colored graph isomorphism classes and 288 isomorphism classes of colored connected components of the graphs in $\mathscr{D}$, and they can be found at: `http://knot.math.usf.edu/data/Colored_Components/index.html`.

2.3. **Graph Features Selection.** We describe a method of converting graph data set to a set of points in the Euclidean space $\mathbb{R}^n$, i.e., the point cloud.

To each (directed and colored) graph G in our data set we associate a vector $P(G) \in \mathbb{R}^n$ obtained by using relevant numerical graph invariants. This vector is then considered as a point in $\mathbb{R}^n$ (Figure 4).

The vector $P(G)$ is obtained by using local, vertex specific, and global, graph specific, features of G. In this first global analysis of the genome we cluster the data according to general graph structure properties, therefore for each $G \in \mathscr{G}_{\mathscr{D}}$ we also consider a corresponding undirected graph $U(G)$. The undirected, uncolored graph $U(G)$ is obtained from G by replacing each pair of parallel edges with opposite directions in G with an undirected edge, and by ignoring the direction and the colors of the edges as shown in Figure 5.

Global Vector. A vector $P_{gl}(G)$ with three entries, called the *global vector*, is associated to each graph $G \in \mathscr{G}_{\mathscr{D}}$. This vector $P_{gl}(G)$ consists of three features $\langle |V(G)|, |E(G)|, CN(G)\rangle$ where $|V(G)|$ and $|E(G)|$ are the numbers of vertices and edges in G, respectively, and $CN(G)$ is the size of the largest clique in $U(G)$.

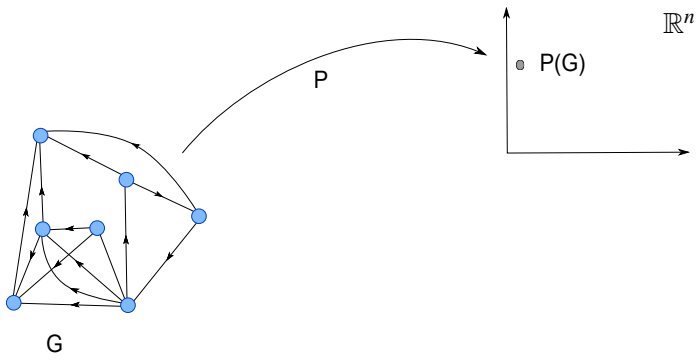

FIGURE 4. Every graph G is associated to a feature vector $P(G)$, a point in the Euclidean space $\mathbb{R}^n$.

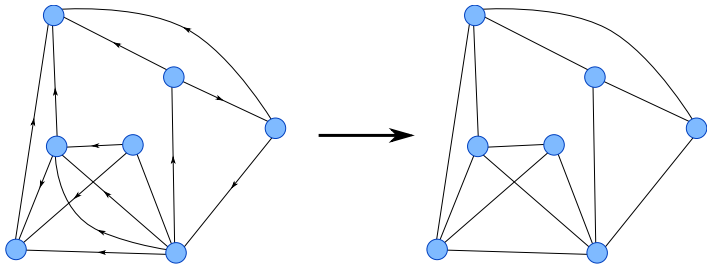

FIGURE 5. An undirected graph associated to a directed one.

The isolated vertices are not counted in $|V(G)|$ as they represent MAC contigs that have no interrelation with any other MAC contig present in the MIC contig represented by the graph. In our data the maximum number of vertices is 43, the maximum number of edges is 74, and the largest clique size is 6 (appears twice in the data).

Local Vector. Vectors that use local properties of the vertices are associated to each $G \in \mathscr{G}_{\mathscr{D}}$. For each vertex v_i we consider two numbers, its valency $val(v_i)$, and the clique number $cq(v_i)$. The valency $val(v_i)$ is a summation of its out-degree and its in-degree (including the parallel oppositely oriented edges) and the number of cliques of given sizes. The clique number $cq(v_i)$ is the number of cliques (induced subgraphs of $U(G)$ isomorphic to the complete graph K_k for some k) that contain this vertex.

The vertices in G, $v_1, v_2, \ldots, v_{|V(G)|}$ are ordered such that their valences are non-increasing. Vertices that have the same valency are further ordered such that their clique numbers are non-increasing. This order remains fixed for the graph G. The *valency vector*, denoted $P_{val}(G)$, consists of a list of valencies of the preordered vertices $P_{val}(G) = \langle \text{valence}(v_i) \rangle_{v_i \in V(G)}$ of the graph G. The maximum valency of a vertex in our data is 29 achieved by contig 67157 with 25 outgoing edges and 4 incoming edges. The 23 of the outgoing edges have $(0,0,1)$, one edge has label $(0,1,0)$ and the other is labeled $(0,1,1)$. The maximum outgoing valency is 25 and it is achieved by the contig 67157. The maximum incoming valency is 6 and it is achieved by contig 67223.

The vertex order of the clique vector follows the same predetermined order of vertices for G. We denote this vector by $P_{cq}(G) = \langle cq(v_i) \rangle_{v_i \in V(G)}$. An example of construction of $P_{cq}(G)$ is depicted in Figure 7.

The Graph Vector. The *graph feature vector* $P(G)$ is defined by concatenating the vectors $P_{gl}(G)$, $P_{val}(G)$ and $P_{cq}(G)$. For a graph G, the number of entries of the vectors $P_{val}(G)$ and $P_{cq}(G)$ are the same and therefore $P(G)$ is a vector in $\mathbb{R}^{2|V(G)|+3}$. We denote the set of vectors associated to $\mathscr{G}_{\mathscr{D}}$ with $S_{\mathscr{D}}$, or simply S .

The Point Cloud. Observe that the number of entries of the vectors in S is not uniform, because this number depends on the number of vertices in the corresponding graph. In order to work in the common Euclidean

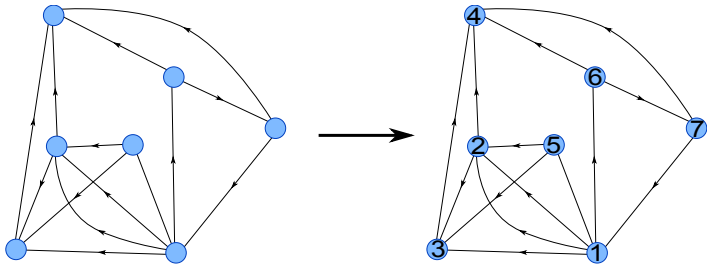

FIGURE 6. Vertices of a graph G are ordered by their valances. Here, $P_{val}(G) = \langle 6,5,4,4,3,3,3\rangle$.

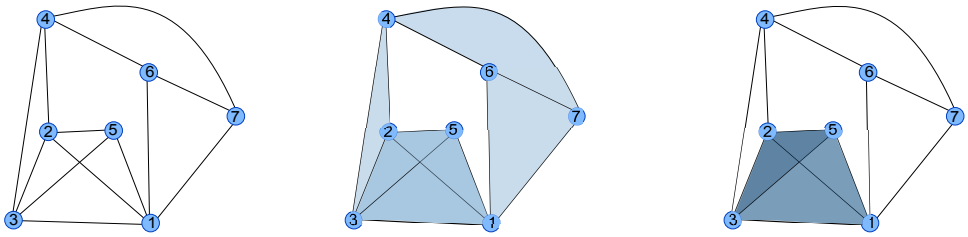

FIGURE 7. The number of cliques associated with the vertex v_i, vertices ordered as in Figure 6. Left: $k = 1$ vertices and $k = 2$ edges. Middle: $k = 2$. Right: $k = 3$. This graph has no cliques of size higher than 4. The clique vector in this example is $P_{cq}(G) = \langle 11,10,10,7,8,6,6\rangle$ which is the sum of the 4 vectors.

space, we expand some of the vectors (by appending 0's) to obtain a consistent number of entries in all vectors. This modification is obtained in the following way.

Let $d = \max\{\ |V(G)|\ |\ G \in \mathscr{G}_{\mathscr{D}}\ \}$. If the valence vector of G is

$$P_{val}(G) = \langle v_1, v_2, \ldots, v_{|V(G)|}\rangle,$$

then we construct an auxiliary valence vector for G with

$$\hat{P}_{val}(G) = \langle v_1, v_2, \ldots, v_{|V(G)|}, 0, \ldots, 0\rangle$$

increasing the number of entries of $P_{val}(G)$ to d such that $d - |V(G)|$ entries of zeros are added at the end. Similarly we construct auxiliary clique vector $\hat{P}_{cq}(G)$ by adding $d - |V(G)|$ zeros at the end of $P_{cq}(G)$. The graph vector $\hat{P}(G)$ is redefined with the concatenation $\langle P_{gl}(G), \hat{P}_{val}(G), \hat{P}_{cq}(G)\rangle$. For our graph data the $\mathscr{G}_{\mathscr{D}}$ the maximum number of vertices in a graph is $d = 43$.

We abuse the notation and use $P(G)$ instead of $\hat{P}(G)$ to refer to the zero-augmented feature vector associated with a graph G. The final point cloud set $S = \{P(G)|G \in \mathscr{G}_{\mathscr{D}}\}$ forms a subset of $\mathbb{R}^{2d+3} = \mathbb{R}^{89}$.

For comparison, we also consider the point cloud S_{gl} from the global features vector $P_{gl}(G)$. The point cloud S_{gl} is in $\mathbb{R}^3$. There are 283 vectors obtained with the above adjustments.

It is important to notice that the entries of the vectors P, and P_{gl} for a graph G are graph isomorphism invariants.

Lemma 2.1. If the graphs G and G' are graph isomorphic (but not necessarily label preserving) then $P(G) = P(G')$ and $P_{gl}(G) = P_{gl}(G')$.

Proof. Let $\phi : G \to G'$ be a graph isomorphism. Then G and G' have the same number of vertices, edges and the size of the maximal cliques in their undirected versions $U(G)$ and $U(G')$. Therefore $P_{gl}(G) = P_{gl}(G')$ and the first three entries of $P(G)$ and $P(G')$ are the same. Also the number of non-zero entries in $P(G)$ and $P(G')$ are the same. A graph isomorphism maps vertices of G to vertices of G' with the same number of

outgoing and incoming edges. Similarly, the number of cliques incident to a vertex in $U(G)$ is the same to the number of cliques of the corresponding vertex in $U(G')$. Let $V_1,\dots,V_s$ be a partition of $V(G)$ such that

i for all $v,w \in V_i$, for all $i = 1,\dots,s$ $val(v) = val(w)$ and $cq(v) = cq(w)$, and
ii for all $v \in V_i$ and $w \in V_j$ with $i < j$ $(i,j = 1,\dots,s)$, either $val(v) > val(v')$ or $val(v) = val(v')$ and $cq(v) > cq(v')$.

Then $\{\phi(V_1),\dots,\phi(V_s)\}$ is a partition of the vertices of G' satisfying the properties [i] and [ii]. Any order of vertices of $V(G)$ (resp. $V(G')$) that has non-increasing valencies and non-increasing clique numbers must list vertices of V_i (resp. $\phi(V_i)$) before vertices of V_j (resp. $V(G')$) whenever $i < j$. Therefore it must be that $P_{val}(G) = P_{val}(G')$ and $P_{cq}(G) = P_{cq}(G')$. □

In our analysis there are three reasons that reduced the data from 688 graphs to 283. Many of these graphs are isomorphic if the edge color is ignored, and directed graphs often reduce to isomorphic undirected graphs. Of course there are graphs G and G' that are non isomorphic but $P(G) = P(G')$. Consider attaching two edges to a 4-cycle to obtain a 6-vertex graph. They can be attached to neighboring or to diagonally opposite vertices of the cycle. In both cases the associated vectors will be the same.

3. Clustering Analysis with TDA

For a data set $S \subset \mathbb{R}^n$, in our case corresponding to a set of directed graphs, a TDA analysis of persistent 0-dimensional homology gives rise to a hierarchy of connected components of (clustered) graphs as described below.

To understand the distribution of the points of S in $\mathbb{R}^n$ we use the notion of the neighborhood graph, as defined below, to construct a hierarchy of undirected graphs whose vertices are S. The neighborhood graph of S depends on a chosen distance function. In our case the distance d is the Euclidean distance between two points $\sqrt{\sum_i (x_i - y_i)^2}$.

Definition 3.1. Let S be a set of points in $\mathbb{R}^n$ and let $\varepsilon \geq 0$ be a non-negative number. The *ε-neighborhood graph* is an undirected graph $G_\varepsilon(S)$, where $G_\varepsilon(S) = (S, E_\varepsilon(S))$ and $E_\varepsilon(S) = \{[u,v] \mid d(u,v) \leq \varepsilon, u,v \in S, u \neq v\}$.

The clustering analysis is done by considering a sequence of neighborhood graphs $G_{\varepsilon_1}(S), G_{\varepsilon_2}(S), \dots$ for $S \subset \mathbb{R}^n$ obtained by a sequence of incrementally increasing values $\varepsilon_1 < \varepsilon_2 < \cdots$.

Definition 3.2. A *cluster* of S at level ε is a connected component in the neighborhood graph $G_\varepsilon(S)$.

We observe some facts about the graphs vectors P and P_{gl}. Suppose $\mathscr{G}$ is a family of graphs and $S = S(\mathscr{G})$ and $S_{gl} = S_{gl}(\mathscr{G})$ are points in $\mathbb{R}^n$ obtained as described above. The vectors in the set S and S_{gl} are all part of the integer lattice of $\mathbb{R}^n$ and $\mathbb{R}^3$ respectively, therefore any distance between two distinct vectors is at least 1. The observation below indicates that small changes in the graphs represented by the vectors induce larger distances of the vectors in S.

Lemma 3.3. Let $G, G' \in \mathscr{G}$. Then the following hold.

(a) If G' is obtained from G by addition of one vertex and one edge incident to that vertex. Then $d(P(G),P(G')) \geq 3$ and $d(P_{gl}(G),P_{gl}(G')) \geq \sqrt{2}$.
(b) If G' is obtained from G by addition of one directed edge without changing the total number of vertices, nor the number of cliques, then $d(P(G),P(G')) \geq 2$.
(c) If G' is obtained from G by addition of one edge that adds a clique to the graph $U(G')$ without changing the number of vertices, then $d(P(G),P(G')) \geq \sqrt{5}$.
(d) If G' is obtained from G by changing the target of one edge from vertex v to vertex v' without changing the number of the cliques, either $d(P(G),P(G')) \geq \sqrt{2}$ or $d(P(G),P(G')) = 0$.

Proof. (a) The addition of a vertex in G' changes the number of non-zero entries in $P(G')$ in two places, once at $P_{val}(G')$ and again at $P_{cq}(G')$. Let w be the new vertex in G' added to $V(G)$ and let $[v,w]$ be the new edge in G' connecting $v \in V(G)$ with the new vertex w. Then w can be taken to be the last vertex

in $V(G')$ in the order of the vertices, while the order of v in $V(G')$ might be either the same as its order in $V(G)$ or different. In both cases the entries in $P(G')$ corresponding to $|V(G')|$, $|E(G')|$, $val(v)$, $cq(v)$, $val(w)$ are at least one more than the corresponding entries in $P(G)$ and the entry of $cq(w)$ is at least two more (a 1-clique vertex w and a 2-clique the new edge) than the corresponding entry in $P(G)$ which is 0. So $d(P(G),P(G')) = \sqrt{\sum_i (x_i - y_i)^2} \geq \sqrt{5+2^2} \geq 3$, and $d(P_{gl}(G),P_{gl}(G')) \geq \sqrt{2}$.

The proofs of (b) and (c) follow a similar argument. Note that in case of (b), if the new directed edge is incident to vertices v and w, then because the number of cliques in $U(G')$ is not changed from the number of cliques in $U(G)$, there is an edge in G incident to v and w in opposite direction. So $P(G')$ has at least one more in the entries $|E(G)|$, $val(v)$ and $val(w)$. Observe that this may imply a change in the order of the vertices, in which case there may be a difference in the entries corresponding to the $cq(v)$ and $cq(w)$ which would increase the distance between the vectors. Therefore $d(P(G),P(G')) \geq \sqrt{3}$.

For the case of (c), the entries of $|E(G')|$, $val(v)$, $val(w)$, $cq(v)$, $cq(w)$ in vector $P(G')$ have a change of at least one and therefore the distance $d(P(G),P(G')) \geq \sqrt{5}$ and $d(P_{gl}(G),P_{gl}(G')) \geq \sqrt{1} = 1$. The case (d) follows the argument of (b) if the valencies change or, if valencies don't change, the graphs are represented by the same vectors and the distance is 0. □

3.1. **Analyzing The Data Using Neighborhood Graphs.** A *filtration of a graph G* is a sequence of nested graphs $G_1 \subseteq G_2 \subseteq \cdots \subseteq G_k = G$ where each G_i is a subgraph of G_{i+1}. The definition of the neighborhood graph on a point cloud S naturally induces a filtration for a connected graph with vertices S. Namely, given a point cloud $S \in \mathbb{R}^n$ and a finite sequence of non-negative numbers $0 = \varepsilon_1 < \varepsilon_2 < \cdots < \varepsilon_k$ we obtain a filtration $G_{\varepsilon_1}(S) \subseteq G_{\varepsilon_2}(S) \subseteq \cdots \subseteq G_{\varepsilon_k}(S)$. We assume that $\varepsilon_1 = 0$, which implies that $E(G_{\varepsilon_1}) = \emptyset$. This filtration also helps to extract the connected components (clusters) of S at various spatial resolutions. For a given ε, each connected component of $G_\varepsilon(S)$ corresponds to a cluster of directed graphs whose corresponding points in $\mathbb{R}^n$ are connected by edges that are of lengths less than ε. This means that the vectors associated with the graphs are at most ε apart, i.e., the graph properties indicated in the vectors are similar and are within ε neighborhood from each other. To have a better information about the topological properties encoded in a filtration one usually considers *the persistence diagram* of the filtration. For our purpose, the persistence diagram describes a way the connected components of the neighborhood graph merge together as we increase the value of ε. The persistence diagram is also equivalently described by the persistence *barcode* [23]. The barcode construction is described as follows.

Let $S = S_{\mathscr{D}} \subset \mathbb{R}^n$, where $n = 2d+3$ (in our case $n = 89$). In Figures 8 and 12, the vertical axis enumerates points of S, and ε-values are listed on the horizontal axis. At $\varepsilon_1 = 0$, $E(G_{\varepsilon_1}) = \emptyset$, and each point of $S \subset \mathbb{R}^n$ forms a single connected component. There are $|S|$ connected components, and hence the number of bars in the barcode at value 0 is equal to the number of data points in S corresponding to the birth of all connected components. With appropriate increments of ε new edges are added to the neighborhood graph and the connected components start joining each other forming larger clusters. The merging event of connected components is represented by a termination of all but one of the corresponding bars of the barcode. The choice of the bar that does not terminate in a merge of components is arbitrary, and we use the established convention (see [23]) where bars are vertically ordered by their length from the shortest at the bottom of the diagram to the longest on the top.

The number of connected components of the graph G_ε is the number of horizontal bars intersecting the vertical line at distance equal to ε. For instance, from Figure 8 we deduce that the number of connected components in $G_\varepsilon(S)$ is 2 for $\varepsilon = 15$ indicating two clusters at that distance. Typically, the filtration ends with a neighborhood graph that has a single connected component. That is, the sequence of ε values increase from 0 to the value that gives rise to a single component graph. In the case of data $\mathscr{D}$ for the set of global vectors and the point clouds S and S_{gl}, the ε values range from 0 to 22 and 0 to 15 respectively.

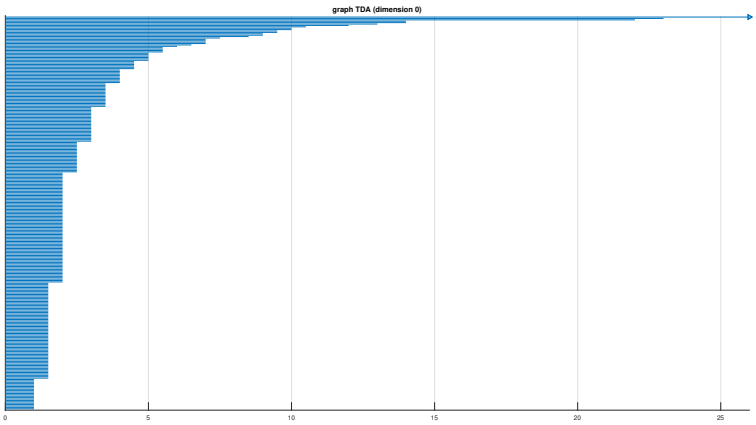

FIGURE 8. The barcode diagram describing the birth and death of the connected components of the neighborhood graph of the dataset S.

3.2. **Tree Diagrams Representing Merging Components.** The merging events of connected components described in the persistence diagram can be encoded using a tree diagram called a *dendrogram* [29]. The bottom points of the tree diagram correspond to the points of S (resp. S_{gl}), that also correspond to the connected components of $G_0(S)$. The vertical direction of the tree diagram represents values of ε.

At each level ε the connected components (clusters) are enumerated and each vertex in the tree is labeled by (i,ε) where i is an index that corresponds to the ith cluster of the graph at level ε. At each level ε, the number of nodes corresponds to the number of clusters of $G_\varepsilon(S)$. For a node (vertex) v at level ε_i, the children of v correspond to the clusters at level ε_{i-1} (i.e., the graph $G_{\varepsilon_{i-1}}(S)$) that have joined to a single connected component represented by v in G_{ε_i}.

For a large enough value of ε_k, $G_{\varepsilon_k}(S)$ is connected, and it corresponds to the single node (root) of the tree. The dendrograms corresponding to the persistent diagrams for S and S_{gl} are shown in Figures 9 and Figure 13 in the supplementary documentation, respectively.

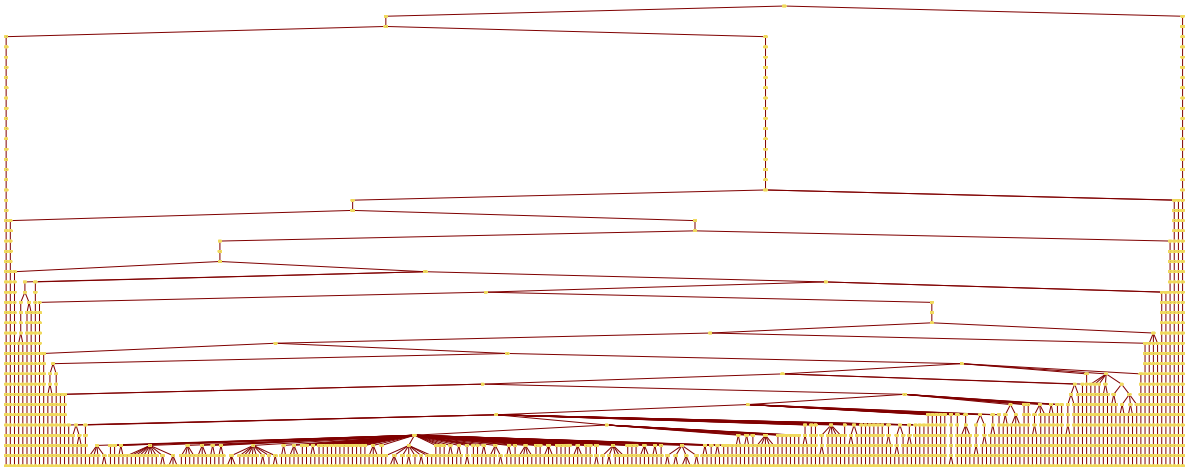

FIGURE 9. The dendrogram clustering tree of dataset S.

3.3. **Implementation.** The point cloud generated from the data $\mathscr{D}$ was generated using a custom Python script. The persistence diagrams were generated using Javaplex [37] and the dendrogram tree diagrams

were generated using Mathematica [26]. The sequence data, the graph data and the scripts are available at `http://knot.math.usf.edu`.

4. RESULTS

The analyzed data $\mathscr{D}$ consists of processed [7] micronuclear contigs obtained after sequencing of *O. trifallax* [13] and is available at [7]. The directed graphs that correspond to the contigs in $\mathscr{D}$ can be found at `http://knot.math.usf.edu/data/Colored_Graphs/index.html`.

As mentioned, the data $\mathscr{D}$ produced 273 distinct vector entries in $\mathscr{S} = \mathscr{S}_{\mathscr{D}}$ that correspond to the same number of isomorphism classes of graphs ranging from 2 to 43 vertices. Each MIC contig corresponds to a vector in $\mathscr{S}$ while the MAC contigs whose MDS segments do not have any of the types 1, 2 or 3 interactions with MDSs of other contigs represent isolated vertices in the graphs and are not taken in consideration for the construction of $S_{\mathscr{D}}$.

We constructed filtration with ε increments of .5 in order to detect small neighborhood changes in the neighborhood graph, these sometimes are reflected by reorienting a directed edge.

4.1. Output of Hierarchal Clustering.

The bar code diagram and the dendrogram for the filtration and clustering of the neighborhood graph of $\mathscr{S}$ are depicted in Figure 8 and Figure 9. As expected by Lemma 3.3, the neighborhood graph consists of isolated vertices for $\varepsilon \leq 1$ and the fist edges appear at $\varepsilon = 1.5$ when there are 14 two point and 4 three point clusters. The two or three graphs joined at this distance differ from each other by small changes such as a single directed edge addition that does not change the cliques.

At $\varepsilon = 2$, as noted in Lemma 3.3, most points remain distant from each other and only those representing graphs with small changes in their structure are joined by en edge. In addition two, three and in one instance four of the previously formed clusters join in (also with some additional points) to form new clusters, and there are 25 new small two or three point clusters. Most of the points in $\mathscr{S}$ remain as isolated vertices.

At $\varepsilon = 2.5$ a dramatic change occurs and one large cluster of 155 elements is formed with a second cluster of 5 points, and several small (two or three point) clusters. All other points stay as isolated vertices. At this point the feature of the point-cloud becomes clear, it consists of a single large cluster, singletons, and some small two or three element components.

At $\varepsilon = 9.5$, there is one large cluster of 269 points while the second largest cluster is of 4 elements, and there are 10 isolated points.

In the last 5 digits of contig numbers, the second largest cluster consists of:

$$88928, \quad 88096, \quad 67742, \quad 67187.$$

Figure 10 shows the graphs contained in this cluster.

All four of these graphs contain a 'star' vertex that is of high valency having multiple black (label $(0,0,1)$) outgoing edges. This indicates that there is one MAC contig whose MDSs interleave with MDSs of multiple other MAC contigs, and we observed that in most of these cases it is one IES of the central 'star' MAC contig that contains most or all of the MDSs of the other contigs. This coincides with the observation in [5] where the depth of these embeddings were considered.

The isolated points belong to 10 contigs

$$67761, \quad 87162, \quad 87484, \quad 67363, \quad 67280, \quad 67243, \quad 67157, \quad 67223, \quad 67417, \quad 67411.$$

These are depicted in the corresponding figures in Supplementary Material Section. We note that some of these graphs have multiple 'star' vertices, or the component that contains a 'star' vertex also has additional cycles and cliques. In particular, the two graphs with 6-cliques (contigs 6742 and 67223) and the one with a 5-clique (contig 67411) are part of these isolated points. Furthermore, the graph with the longest path of 5 vertices (contig 87484) indicated in [5] as one of the most in-depth embedding of genes within a single IES is also on this list. In all these cases we observe that the majority of the edges are black and purple, meaning that the prevailing inter-gene MDS organization is interleaving.

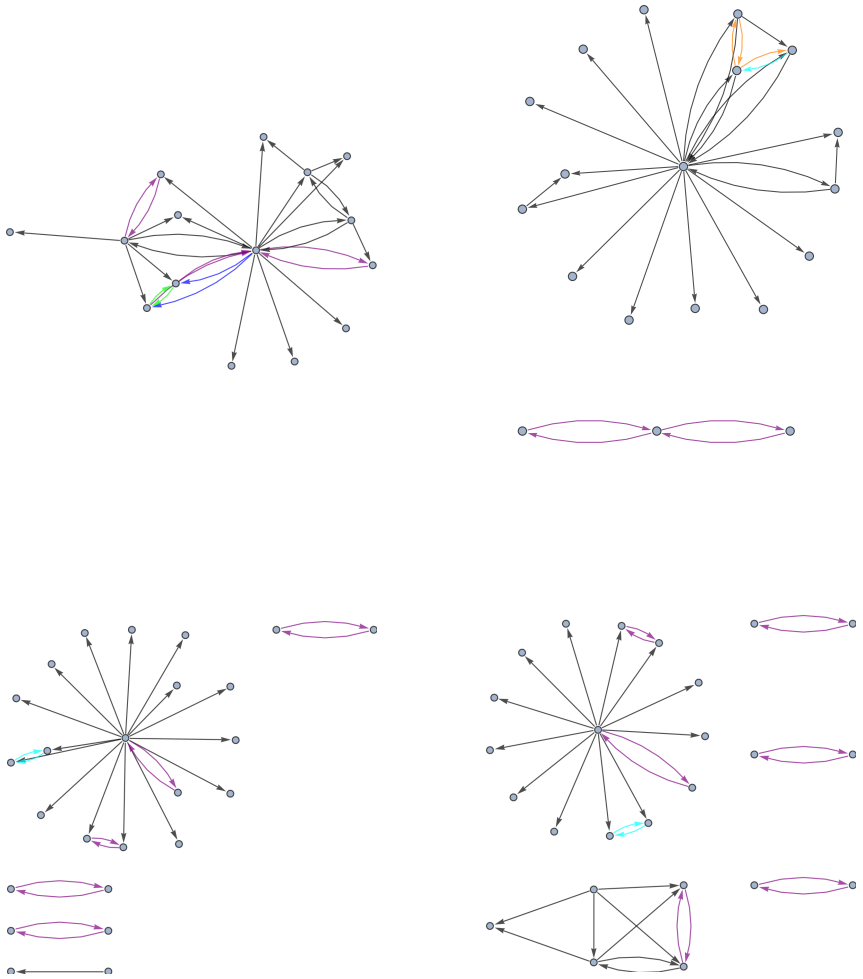

FIGURE 10. Top ctg7180000088928 and ctg7180000088096. Down ctg7180000067742 and ctg7180000067187.

As ε increases, the four-element cluster becomes part of the large cluster at $\varepsilon = 10.5$ and the isolated singleton points join the large cluster one or two at the time until $\varepsilon = 14.5$ when the two, most distant contigs 67517 and 67223 remain isolated until $\varepsilon = 22$ and $\varepsilon = 23$ respectively.

The pattern of clusters for S_{gl} is similar to that of S. A large single cluster is formed at value $\varepsilon = 1.5$, with 2 clusters of 5 elements, 3 clusters of 2 elements, and 23 singleton clusters.

At $\varepsilon = 4.5$, the clusters consist of a large single cluster, the second largest of 9 elements, the two clusters of two elements, and 5 singleton clusters. The size two clusters are $\{67417, 67243\}$ and $\{67187, 67228\}$. The elements of the former cluster appear as isolated points in the neighborhood $\varepsilon = 9.5$ of S, while 67187 of the latter cluster, appears in the 4-elements cluster of S, and 67228 is in the largest cluster of S.

The isolated points for $\varepsilon = 4.5$ are 67223, 67363, 67157, 67280, 87484. We note that all also appear as isolated points of the neighborhood graph of S for $\varepsilon = 9.5$.

In this case, as in the case of S, the two most distant graphs correspond to the contigs 67517 and 67223 that join the large cluster at $\varepsilon = 14.5$ and $\varepsilon = 18.5$ respectfully.

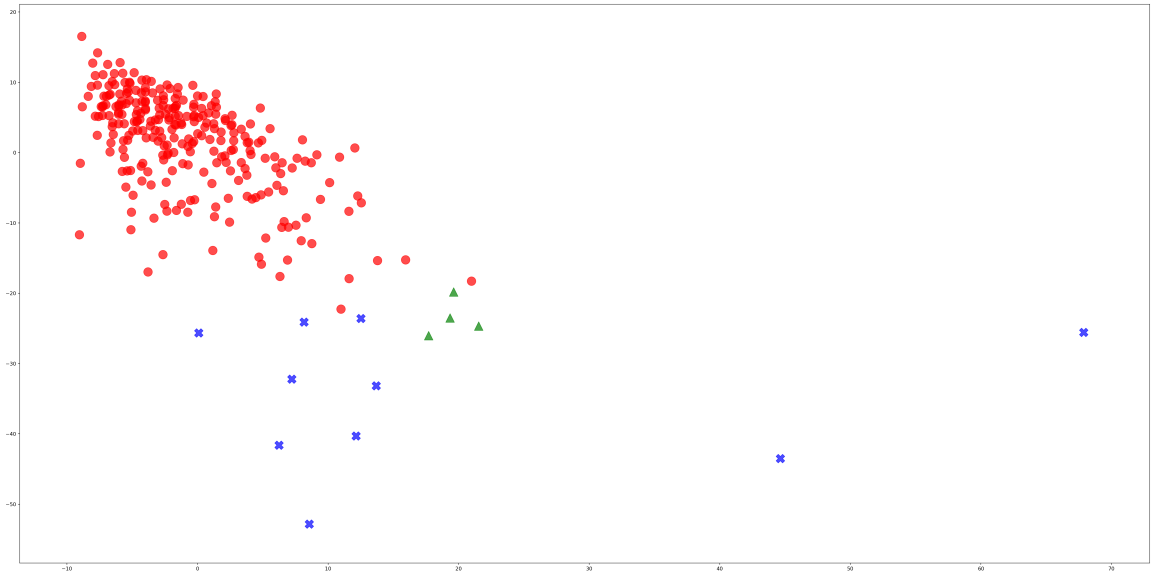

FIGURE 11. A 2d multidimensional scaling projection for S. The points of S are colored according to clustering at $\varepsilon = 9.5$. At this level we have 12 clusters, the largest cluster is colored red in the figure, the second largest cluster consists of 4 elements and is colored green and the singletons are all colored blue.

Figures 11 and 14 represent the 2d multidimensional scaling projections[27] of the point clouds S and S_{gl}, respectively.

5. DISCUSSION

In this paper we initiated a mathematical method of representing and analyzing gene segment relationship in a scrambled genome of *Oxytricha trifallax* and up to our knowledge, such genome wide study for inter-gene segment arrangement has not been done before. The inter-gene segment arrangements are represented by graphs representing their segment relationship. We analyzed the graph data by converting these graphs to a point cloud in a higher dimensional Euclidean space and applied clustering methods from topological data analysis to identify patterns in the graph structures.

The big majority of interactions within a single MIC contig are represented with small graphs up to five vertices (corresponding to the large cluster at $\varepsilon = 2.5$) and one can 'move' from one graph to another by small vertex/edge changes. This suggests that genes with complex interaction patterns are unique and often not found among macronuclear genes. The majority of the inter-gene segment arrangement within micronuclear chromosomes involve two or three genes and are pairwise interleaving and sometimes overlapping.

The most prevalent multi-gene segment arrangements in the *Oxytricha*'s genome are interleaving and often this appears as one gene (or one IES in a MAC contig) interleaving with multiple other MAC contigs (as seen through the 'star' like vertices). In the cluster consisting of four graphs, a single IES of the 'star' MAC contig interleaves with multiple MAC contigs. All star contigs are scrambled, which follows the analysis in [5] where it was observed that contigs whose IESs interleave with other MAC contigs are mostly scrambled.

The graph representation of the inter-gene segment relationship introduced here is novel, and we hope that a similar approach can be used in studies of the scrambled genomes of other species. Comparisons among orthologous genes in other species with scrambled genomes may reveal whether patterns in these graph structures are conserved or embellished over evolutionary time. Furthermore, studies whether genes with interleaved gene segments are co-expressed may indicate whether the rearrangement of these MAC segments are in parallel or sequential. We suggest that models of gene rearrangement should also focus on

operations that can be applied to these frequent interleaving gene segments, which in some cases resemble the odd-even patterns detected within scrambled genes [7].

The construction of the graph data into a point cloud in this paper is by a vector whose entries are common graph invariant properties, such as the number of vertices, edges and cliques. We used two vectors, one that had more local vertex properties and the other in $\mathbb{R}^3$ which included only the number of vertices, edges and the maximal clique. It is interesting that in both cases the isolated points are the same, and much distant from the rest of the points. The rearrangement process of the MIC contigs corresponding to these isolated points may indicate specific biological process that include multiple genes simultaneously. The graphs with large cliques (5 and 6) imply that segments of up to 5 or 6 genes all mutually interleave and we suggest further rearrangement gene analysis for these situations. In our study we did not consider the length of overlapping segments, nor the number of interleaving gene segments. Further methods that include edge weights may be suitable for more detailed analysis.

Our representation of graphs relied mainly on representation of a graph via a feature vector in $\mathbb{R}^n$. Similar attempts in this direction have been made for graph similarities [24, 34]. Their work focus on undirected graphs and do not consider the local properties that we used. There are other venues that rely on developing a similarity measure between graph [6, 30] or graph kernels [2, 21] that we have not explored here. These are methods that often relay on the structural properties of the graph sometimes identified through topological methods and they may also reveal other properties in the genome. For example, such methods have been successfully applied in protein function prediction [4] and chemical informatics [32]. Comparison of such graph analysis methods will be subject of another study.

REFERENCES

[1] Michele Ballerini, Nicola Cabibbo, Raphael Candelier, Andrea Cavagna, Evaristo Cisbani, et al., *Interaction ruling animal collective behavior depends on topological rather than metric distance: Evidence from a field study*, Proceedings of the national Academy of Sciences **105** (2008), no. 4, 1232–1237.

[2] Michael Baur and Marc Benkert, *Network comparison*, Network analysis, 2005, pp. 318–340.

[3] Sigrid Beermann, *The diminution of heterochromatic chromosomal segments in cyclops (crustacea, copepoda)*, Chromosoma **60** (1977), no. 4, 297–344.

[4] Karsten M Borgwardt, Cheng Soon Ong, Stefan Schönauer, SVN Vishwanathan, Alex J Smola, and Hans-Peter Kriegel, *Protein function prediction via graph kernels*, Bioinformatics **21** (2005), no. suppl 1, i47–i56.

[5] Jasper Braun, Lukas Nabergall, Rafik Neme, Laura F Landweber, Masahico Saito, and Nataša Jonoska, *Russian doll genes and complex chromosome rearrangements in Oxytricha trifallax* (In Preparation).

[6] Horst Bunke and Kaspar Riesen, *Recent advances in graph-based pattern recognition with applications in document analysis*, Pattern Recognition **44** (2011), no. 5, 1057–1067.

[7] Jonathan Burns, Denys Kukushkin, Xiao Chen, Laura F Landweber, Masahico Saito, and Nataša Jonoska, *Recurring patterns among scrambled genes in the encrypted genome of the ciliate oxytricha trifallax*, Journal of theoretical biology **410** (2016), 171–180.

[8] Jonathan Burns, Denys Kukushkin, Kelsi Lindblad, Xiao Chen, Nataša Jonoska, and Laura F Landweber, *<mds_ies_db>: a database of ciliate genome rearrangements*, Nucleic Acids Research **44** (2015), no. D1, D703–D709.

[9] Gunnar Carlsson, *Topology and data*, Bulletin of the American Mathematical Society **46** (2009), no. 2, 255–308.

[10] Gunnar Carlsson and Vin De Silva, *Zigzag persistence*, Foundations of Computational Mathematics **10** (2010), no. 4, 367–405.

[11] Gunnar Carlsson, Vin De Silva, and Dmitriy Morozov, *Zigzag persistent homology and real-valued functions*, Proceedings of the twenty-fifth annual symposium on computational geometry, 2009, pp. 247–256.

[12] Frédéric Chazal, Leonidas J Guibas, Steve Y Oudot, and Primoz Skraba, *Persistence-based clustering in riemannian manifolds*, Journal of the ACM (JACM) **60** (2013), no. 6, 41.

[13] Xiao Chen, John R. Bracht, Aaron David Goldman, Egor Dolzhenko, Derek M. Clay, Estienne C. Swart, David H. Perlman, Thomas G. Doak, Andrew Stuart, Chris T. Amemiya, Robert P. Sebra, and Laura F. Landweber, *The architecture of a scrambled genome reveals massive levels of genomic rearrangement during development*, Cell **158** (2014), no. 5, 1187–1198.

[14] Jianlin Cheng, Michael J Sweredoski, and Pierre Baldi, *Dompro: protein domain prediction using profiles, secondary structure, relative solvent accessibility, and recursive neural networks*, Data Mining and Knowledge Discovery **13** (2006), no. 1, 1–10.

[15] Theodosius Dobzhansky, *On the sterility of the interracial hybrids in drosophila pseudoobscura*, Proceedings of the National Academy of Sciences **19** (1933), no. 4, 397–403.

[16] Herbert Edelsbrunner and John Harer, *Persistent homology – a survey*, Contemporary Mathematics **453** (2008), 257–282.

[17] Herbert Edelsbrunner, David Letscher, and Afra Zomorodian, *Topological persistence and simplification*, Foundations of computer science, 2000. proceedings. 41st annual symposium on foundations of computer science, 2000, pp. 454–463.
[18] Wenwen Fang, Xing Wang, John R. Bracht, Mariusz Nowacki, and Laura F. Landweber, *Piwi-interacting RNAs protect DNA against loss during Oxytricha genome rearrangement*, Cell **151** (2012), no. 6, 1243–1255.
[19] Carlos Fernandez-Lozano, Enrique Fernández-Blanco, Kirtan Dave, Nieves Pedreira, Marcos Gestal, Julián Dorado, and Cristian R Munteanu, *Improving enzyme regulatory protein classification by means of SVM-RFE feature selection*, Molecular Biosystems **10** (2014), no. 5, 1063–1071.
[20] Marcio Gameiro, Yasuaki Hiraoka, Shunsuke Izumi, Miroslav Kramar, Konstantin Mischaikow, and Vidit Nanda, *A topological measurement of protein compressibility*, Japan Journal of Industrial and Applied Mathematics **32** (2015), no. 1, 1–17.
[21] Thomas Gärtner, Peter Flach, and Stefan Wrobel, *On graph kernels: Hardness results and efficient alternatives*, Learning theory and kernel machines, 2003, pp. 129–143.
[22] Robert Ghrist, *Barcodes: the persistent topology of data*, Bulletin of the American Mathematical Society **45** (2008), no. 1, 61–75.
[23] _____, *Barcodes: The persistent topology of data*, Bullentin of the American Mathematical Society **45** (2008), 61–75.
[24] Jaume Gibert, Ernest Valveny, and Horst Bunke, *Graph embedding in vector spaces by node attribute statistics*, Pattern Recognition **45** (2012), no. 9, 3072–3083.
[25] Irmgard U Haussmann, Zsuzsanna Bodi, Eugenio Sanchez-Moran, Nigel P Mongan, Nathan Archer, Rupert G Fray, and Matthias Soller, *m6a potentiates Sxl alternative pre-mrna splicing for robust drosophila sex determination*, Nature **540** (2016), no. 7632, 301–304.
[26] Wolfram Research, Inc., *Mathematica, Version 11.1*. Champaign, IL, 2017.
[27] Joseph B Kruskal and Myron Wish, *Multidimensional scaling*, Vol. 11, Sage, 1978.
[28] Peter Meinicke, *Uproc: tools for ultra-fast protein domain classification*, Bioinformatics **31** (2015), no. 9, 1382–1388.
[29] Fionn Murtagh, *A survey of recent advances in hierarchical clustering algorithms*, The Computer Journal **26** (1983), no. 4, 354–359.
[30] Panagiotis Papadimitriou, Ali Dasdan, and Hector Garcia-Molina, *Web graph similarity for anomaly detection*, Journal of Internet Services and Applications **1** (2010), no. 1, 19–30.
[31] David M Prescott, *The dna of ciliated protozoa*, Microbiological Reviews **58** (1994), no. 2, 233–267.
[32] Liva Ralaivola, Sanjay J Swamidass, Hiroto Saigo, and Pierre Baldi, *Graph kernels for chemical informatics*, Neural Networks **18** (2005), no. 8, 1093–1110.
[33] Loren H Rieseberg, *Chromosomal rearrangements and speciation*, Trends in Ecology & Evolution **16** (2001), no. 7, 351–358.
[34] Kaspar Riesen and Horst Bunke, *Graph classification and clustering based on vector space embedding*, World Scientific, 2010.
[35] Yoshiyuki Shibata, Pankaj Kumar, Ryan Layer, Smaranda Willcox, Jeffrey R Gagan, Jack D Griffith, and Anindya Dutta, *Extrachromosomal microDNAs and chromosomal microdeletions in normal tissues*, Science **336** (2012), no. 6077, 82–86.
[36] Jeramiah J Smith, Carl Baker, Evan E Eichler, and Chris T Amemiya, *Genetic consequences of programmed genome rearrangement*, Current Biology **22** (2012), no. 16, 1524–1529.
[37] Andrew Tausz, Mikael Vejdemo-Johansson, and Henry Adams, *JavaPlex: A research software package for persistent (co)homology*, Proceedings of ICMS 2014, 2014, pp. 129–136. Software available at `http://appliedtopology.github.io/javaplex/`.
[38] Susumu Tonegawa, *Somatic generation of antibody diversity*, Nature **302** (1983), no. 5909, 575–581.
[39] Kelin Xia, Xin Feng, Yiying Tong, and Guo Wei Wei, *Persistent homology for the quantitative prediction of fullerene stability*, Journal of Computational Chemistry **36** (2015), no. 6, 408–422.
[40] Kelin Xia and Guo-Wei Wei, *Persistent homology analysis of protein structure, flexibility, and folding*, International Journal for Numerical Methods in Biomedical Engineering **30** (2014), no. 8, 814–844.
[41] Afra Zomorodian and Gunnar Carlsson, *Computing persistent homology*, Discrete & Computational Geometry **33** (2005), no. 2, 249–274.

SUPPLEMENTARY MATERIAL

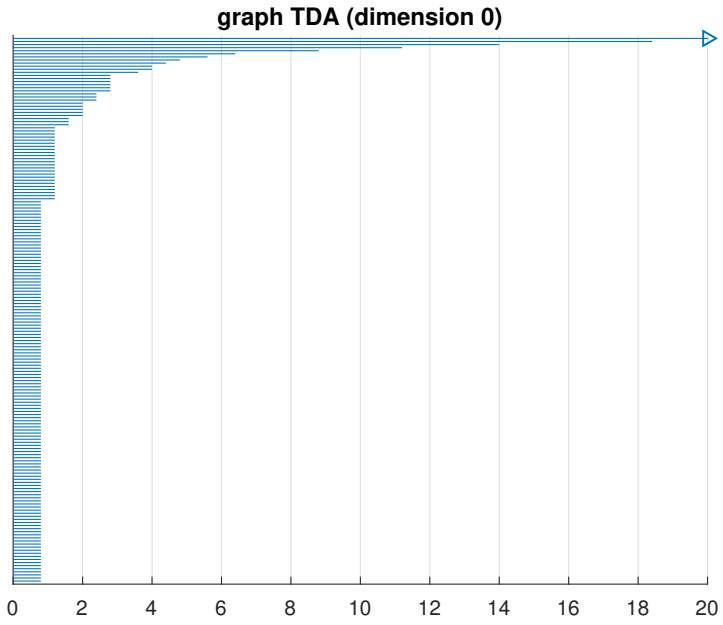

FIGURE 12. The persistence diagram of the data set S_{gl} obtained from the global features.

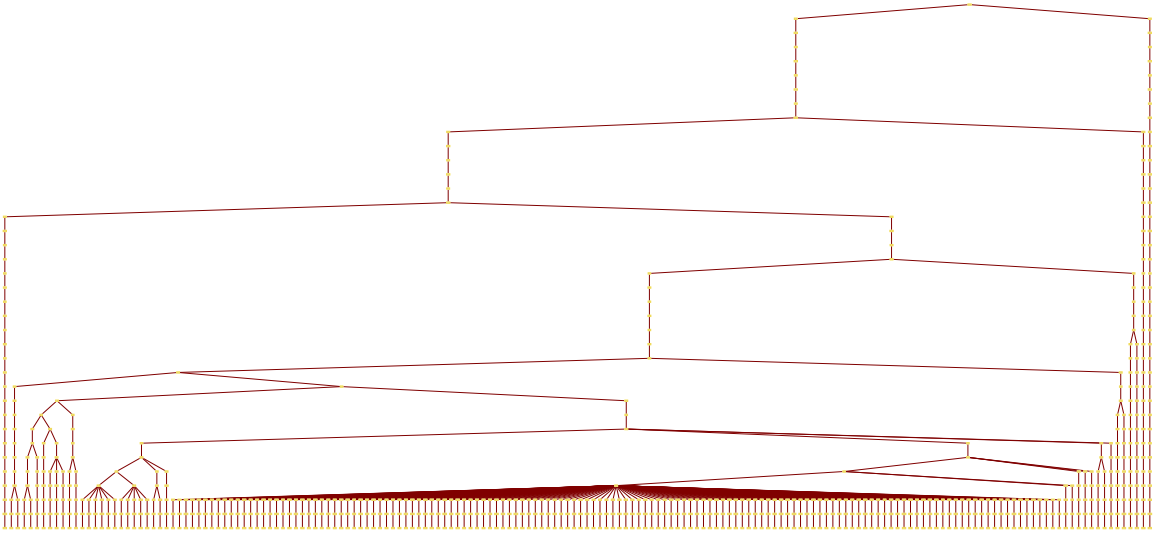

FIGURE 13. The clustering tree of the data S_{gl} set obtained from the global features.

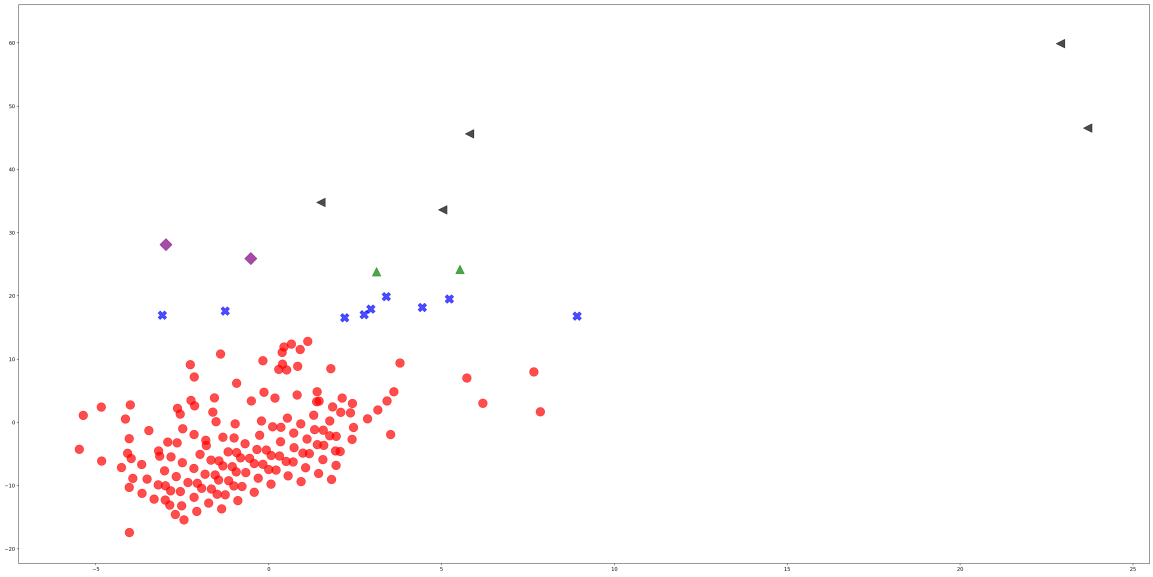

FIGURE 14. A 2d multidimensional scaling projection for S_{gl}. The points of S_{gl} are colored according to clustering at $\varepsilon = 4.5$. At this level we have 9 clusters, the largest cluster is colored red, the second two largest clusters consists of 2 elements and they are colored magenta and green, and the singletons are all colored black.

The graphs in the cluster of S at $\varepsilon = 9.5$ consisting of four elements depicted in Figures 15,16,17 and 18.

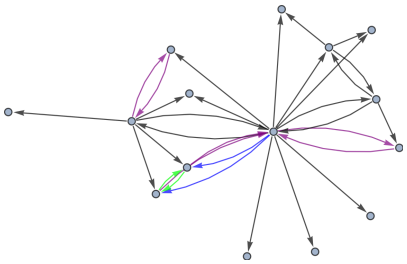

FIGURE 15. ctg7180000088928

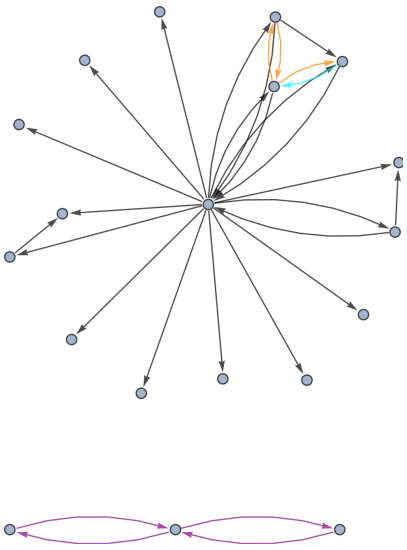

FIGURE 16. ctg7180000088096

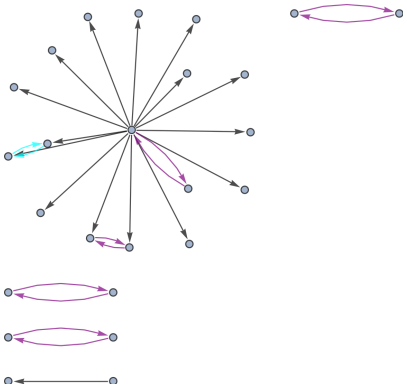

FIGURE 17. ctg7180000067742

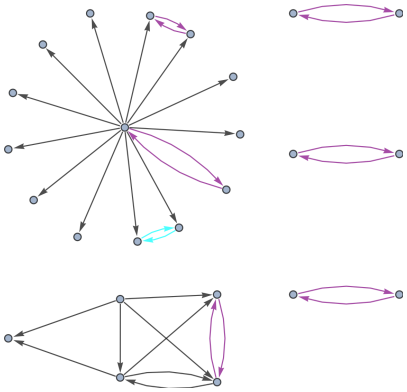

FIGURE 18. ctg7180000067187

The graphs that form singleton clusters (isolated points) in S ($\varepsilon = 9.5$) are depicted in the remaining figures.

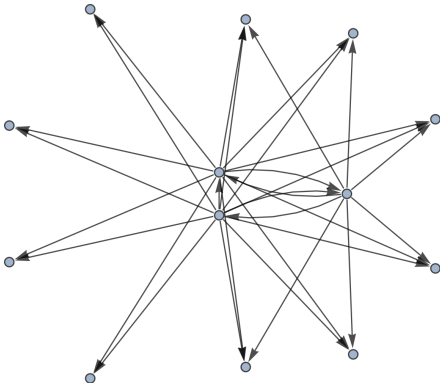

FIGURE 19. ctg7180000067761

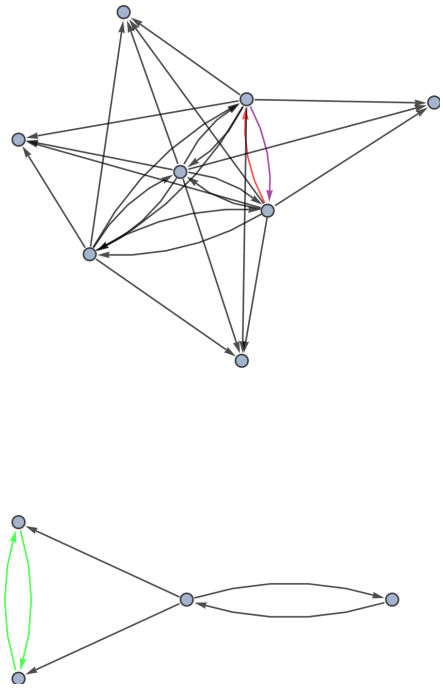

FIGURE 20. ctg7180000087162

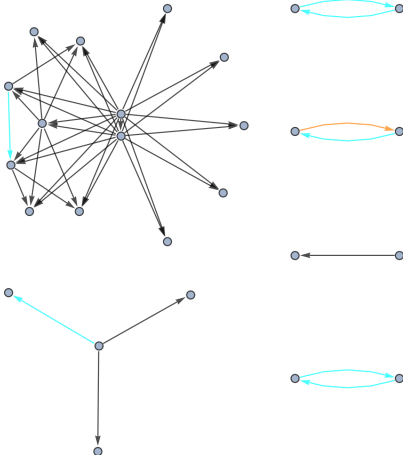

FIGURE 21. ctg7180000087484

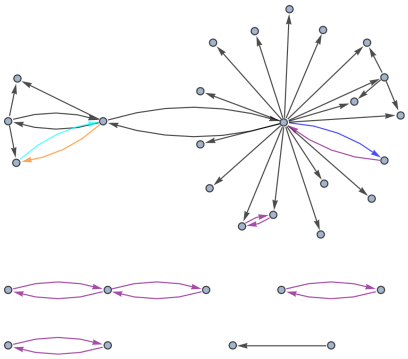

FIGURE 22. ctg7180000067363

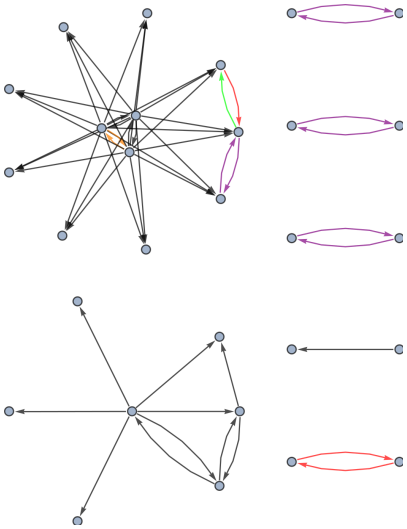

FIGURE 23. ctg7180000067280

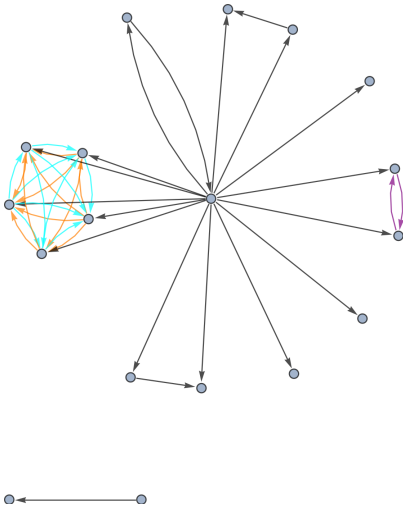

FIGURE 24. ctg7180000067243

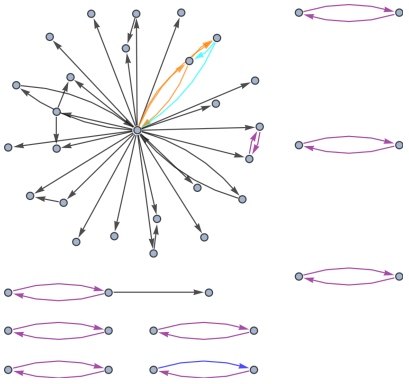

FIGURE 25. ctg7180000067157

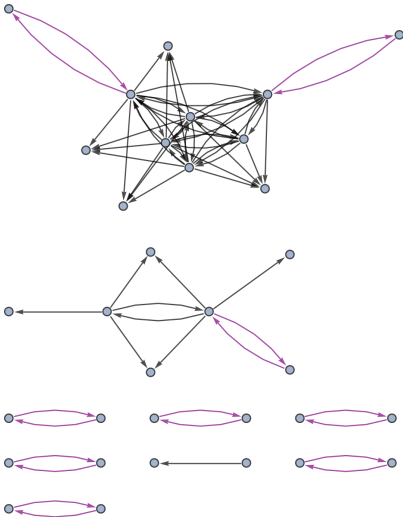

FIGURE 26. ctg7180000067223

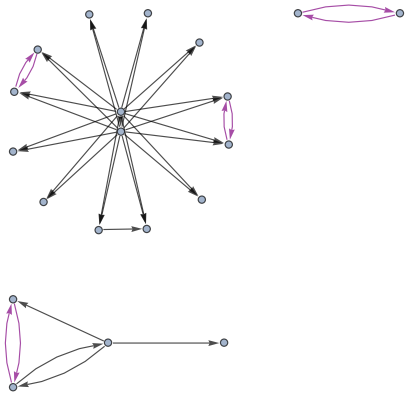

FIGURE 27. ctg7180000067417

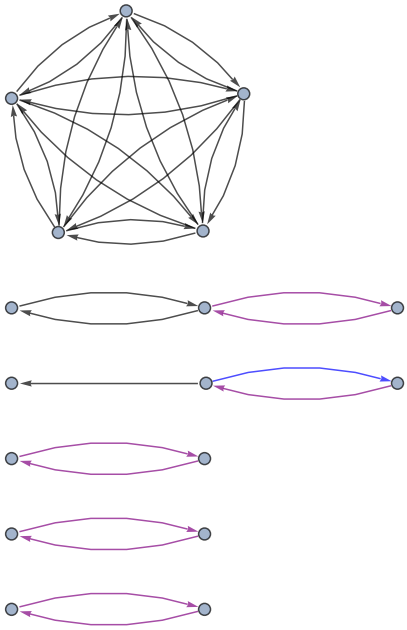

FIGURE 28. ctg7180000067411

(1) DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, UNIVERSITY OF SOUTH FLORIDA, TAMPA, FL 33612, USA

(2) DEPARTMENT OF MATHEMATICS AND STATISTICS, UNIVERSITY OF SOUTH FLORIDA, TAMPA, FL 33612, USA