An Algorithm for DNA Rearrangement Pathways

Maja Milošević
Mentors: Nataša Jonoska, PhD
Masahiko Saito, PhD

Spring 2014

Abstract

Internal DNA rearrangement, involving some DNA elimination, occurs during ciliate conjugation. The mechanism of the rearrangement is unknown and a great topic of interest in many scientific fields. If understood, this mechanism could have many applications, especially in the field of medicine. Some types of cancer cells are known to undergo dramatic DNA rearrangements. We develop our algorithm for predicting the DNA rearrangement pathways in ciliates. Our algorithm allows to predict whether a certain portion of a micronuclear gene, or the whole micronuclear gene, is able to align pointers, and unscramble, based on two different criteria: the sequence length and the chemical properties of the nucleotides, including the intrinsic bendability of the DNA caused by specific nucleotide sequences. The role of enzymes is also examined but not taken into account in the algorithm. The algorithm considers these criteria individually, changing it's output each time. It assigns a probability value to each portion of the gene, predicting how likely that portion is to bend. If the probability value of a certain portion of the gene is above a certain threshold that is scientifically determined, that portion is determined as likely to bend first. Then the remaining genetic sequence is run through the algorithm again, until the whole gene is unscrambled. These sequences of steps are presumed to correspond to the most energetically favorable unscrambling pathways of the gene.

1 Introduction

Excessive DNA rearrangements are known to occur in some cancer cells. Understanding the mechanisms by which the DNA rearrangements occur can provide insight into eliminating unwanted rearrangements or consequences thereof. Model organisms for studying DNA rearrangements are certain species of ciliates like *Oxytricha trifallax*. Ciliates are unicellular eukaryotes known for undergoing a massive DNA rearrangement. They possess two different types of nuclei: a germline micronucleus and a somatic macronucleus. The macronuclear DNA sequence is obtained from the micronuclear DNA sequence via vast DNA rearrangement. The micronuclear sequences contains *internally eliminated sequences (IESs)*, which refer to the sequences of DNA that are spliced out during the recombination, and *macronuclear destined sequences (MDSs)*, which refer to the sequences of DNA that appear in the macronucleus after the rearrangement. The MDSs do not necessarily appear in the same order in the micronucleus as they do in the macronucleus. They can also appear inverted in the micronucleus with respect to the macronucleus [1].

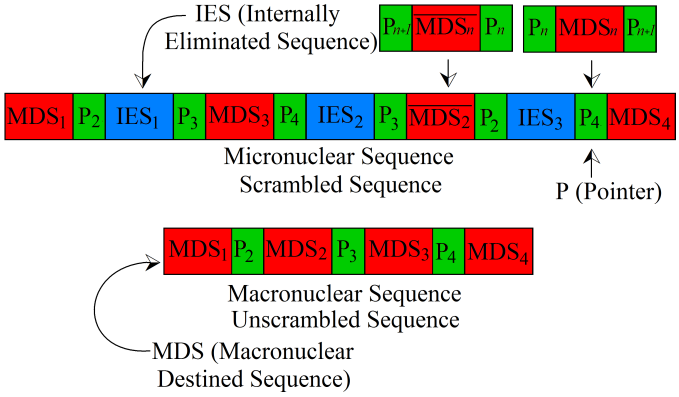


Figure 1: Representations of a micronuclear and a macronuclear DNA sequence of a ciliate gene, including MDSs, IESs, and pointers. MDS_2 is inverted, which is denoted by the bar above it.

The DNA rearrangement is presumed to be guided by *pointers*, which are specific sequences that occur twice in the micronuclear sequence. The pointers correspond to recombination sites of the gene. The macronuclear sequence only contains one copy of each pointer. The development of the macronucleus in these organisms involves the unscrambling of the micronuclear DNA sequence, which is composed of the excision of IESs and the alignment of pointers. Each MDS contains two pointers, one located at the beginning and the other located at the end. The pointers are part of the MDS sequence and they determine the position of the MDS in the macronuclear sequence. The beginning of the n^{th} MDS and the end of the $(n-1)^{st}$ MDS share the same pointer, which we call pointer n [See Figure 1].

2 Biological Parameters

We explore the likelihood of plausible pathways occurring in the unscrambling of ciliate micronuclei. We propose that the scrambled linear DNA sequence must bend for the pointers to align and the gene to unscramble. The cyclization parameters obtained in the algorithm are based on the biological properties of the macromolecular DNA, including the lengths of the sequences involved and the chemical properties of the nitrogenous bases, including the induced flexibility on the DNA by specific nucleotide sequences.

2.1 Sequence Length

Experimental data suggests that the beginning steps in unscrambling the macronuclear DNA sequence involve the alignment of the pointers [2]. The DNA sequence likely begins by bending as far as it can based solely on it's length. This may or may not be enough for the molecule to cyclize. If it is not, additional parameters are considered. We attempt to answer the question: How many nucleotides are necessary for a DNA sequence to be able to form a stable cycle? By stable cycle, we mean a sequence which can form a cycle without force and remain in it's cyclic form for extended periods of time.

One experiment found that 116 nucleotides were needed for a sequence of double-stranded DNA to be able to fully bend to form a cycle. In this particular experiment, 14 different DNA sequences, all of different lengths, were examined. Each sequence was combined with the enzyme ligase, which is known to have the specific function of ligating DNA nicks together. Then an "equilibrium constant" for the cyclization of the DNA sequence was calculated. This value was used to calculate the ability of the DNA sequence to cyclize. The more the equilibrium shifted towards the formation of the cyclized DNA, the more likely a DNA sequence of that length is able to bend into a cycle. The shortest DNA sequence considered in the experiment is a fragment obtained by of the restriction enzyme *Hae III*, which is 126 base pairs long [3]. *Hae III* is an endonuclease, which is an enzyme that cuts the phosphodiester bonds of both strands in double stranded DNA at specific nucleotide sequences known as recognition sites. The recognition sites are 4 to 8 base pairs long. Though the recognition sites are usually palindromic sequences, in type III restriction enzymes, such as *Hae III*, the recognition site is two non-palindromic sequences that are inversely oriented. These enzymes are only found in bacteria and archae and not in eukaryota [4]. The longest DNA sequence examined was the plasmid *pBR322*, which is 4361 base pairs long [3]. A plasmid is a usually circular double stranded DNA segment that can be found in eukaryotes,

as well as bacteria. It is separate from the chromosomal DNA sequence and it has the ability to replicate separately from the chromosomal DNA [5]. Of the 4361 base pairs, 983 are adenines, 1034 are thymines, 1210 are cytosines, and 1134 are guanines. The percent composition of adenine is 22.5%, the percent composition of thymine is 23.7%, the percent composition of cytosine is 27.7%, and the percent composition of guanine is 26%. This makes the percent composition of A-T base pairs 46.3% and that of C-G base pairs 53.7%, which is almost evenly divided [6]. The experiment found that bending between sequences of length 242 base pairs to 4361 base pairs is more favorable by about 100-fold than bending sequence of length 126 base pairs to 242 base pairs. However, the differences among the 242 base pair sequences to the 4361 base pair sequences were negligible [3]. A later experiment found that a slightly smaller sequence length of double-stranded DNA could also form a cycle. This experiment also depended on adding ligase to a mixture of DNA sequences of different lengths [7].

Another experiment, recently done by Vafabakhsh and Ha on protein-free segments of DNA, suggests that segments as small as 67 base pairs can form a cycle with no protein interactions. This experiment suggested that previous experiments may have produced misleading results due to high protein and ion presence as well as usage of sequences with extreme bendability. The DNA segments used were protein-free, except for their use of the enzyme ligase. A-tracts and other motifs that induce significant bendability were also avoided to ensure more accurate results. The DNA segments started in a buffer which did not include added ions because ions such as Na^{+} and Mg^{2+} are known to stabilize the cyclized state in DNA. The experiment used many different sequences and the time these segments were left in the buffer solution also ranged from less than a minute to four hours. One of three outcomes were observed: either the DNA did not cyclize, or it cyclized and stayed in this form, or it randomly cyclized and uncyclized in the alloted time period. This suggests that protein induced DNA cyclization may also be aided by the intrinsic bendability of DNA and that bending energy for short DNA cycles is independent of the sequence length. It also reinforces that DNA bendability is not only governed by sequence length and protein interactions, but also by intrinsic properties of the sequence. By these authors, 67 base pairs is believed to be able to create a cycle if DNA, independent of the sequence. However, nicks or kinks in the double stranded DNA may play a role in making the DNA more bendable[8].

A review of Vafabakhsh and Ha's experiment by Vologodskii suggests that 67 base pairs is not enough for DNA to make a cycle. He examines the experiment in detail and suggests a possible flaw which may have lead to the incorrect conclusion. Ligase is an enzyme which acts like a glue, it sticks DNA segments together. Adding ligase to a sample of linear DNA will increase the chances that the DNA will cyclize because it will simply glue the ends together. If this occurs, it does not suggest that the DNA prefers the cyclized state or that it can even cyclize without the help of ligase. Because of this, results from experiments which are done with ligase must be taken with caution. Vologodskii claims that Vafahakhsh and Ha did not use sufficient caution and their results are skewed. He suggests that, though it may be possible to cyclize a 67 base pair segment of DNA, it is not very likely that the DNA will favor this conformation. There is no experimental data to suggest that DNA below 100 base pairs can cyclize on it's own and remain relatively stable as a cycle. This is why 67 base pairs may be taken as a lower bound for cyclizing DNA, but 100 or more base pairs should be taken as an experimentally verified average value for cyclizing DNA [9]. However, folklore knowledge accepts 200 base pairs as the most favorable for cyclization.

Based on these experiments, here we assume that approximately 100 nucleotides are necessary to fully bend a sequence of double-stranded DNA and form a cycle. Under this assumption, we will consider both possibilities: that the enzyme DNA ligase may or may not be involved in the rearrangement process.

2.2 Nucleotides

If the double-stranded DNA sequence cannot bend into a complete cycle based on its length alone, certain chemical reactions likely take place to lower the energy barrier for bending. These reactions involve the nucleotides found in DNA. A nucleotide has three components: a nitrogenous base, a pentose, and a phosphate [5]. Four nitrogenous bases are found in DNA: adenine (A), thymine (T), cytosine (C), and guanine(G). Adenine and guanine are known as purines, due to their bicyclic shape, and cytosine and thymine are known as pyrimidines, due to their monocyclic shape [see Figure 2].

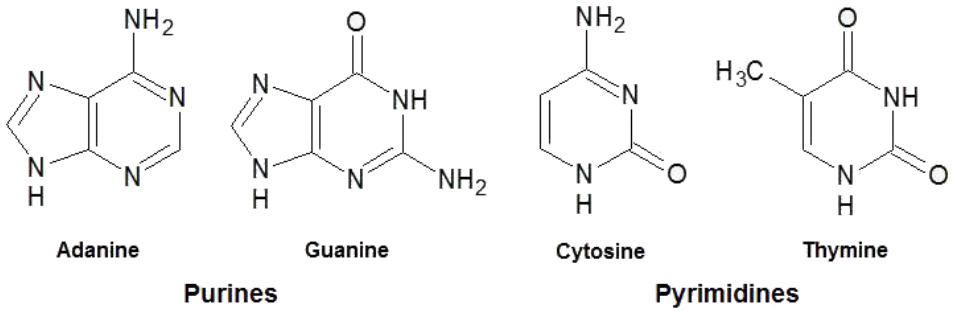


Figure 2: The four nitrogenous bases found in DNA.

The chemical interactions of DNA depend on the solvent in which the DNA is in. In ciliate cells, the solvent is mainly water, so that is what we base our analysis on. Water is famously polar, meaning it is very attracted to hydrogens. The water bonds with certain hydrogens in the DNA structure, which destabilizes the DNA double helix. The DNA accounts for this creating hydrogen bonds with itself. The typical length of one hydrogen bond in water is 197 pm [10]. In DNA, the adenine binds to the thymine via two hydrogen bonds and the guanine binds to the cytosine via three hydrogen bonds [see Figure 3]. One of the hydrogen bonds between adenine and thymine is between an NH and an O, and the other hydrogen bond is between and NH and an N. Two of the hydrogen bonds between guanine and cytosine are between an NH and an O and the last hydrogen bond is between an NH and an N [5, 11]. Hydrogen bonds can vary in strength. An average strength of a hydrogen bond between an NH and an O is 8 kiloJoules per mole and an average strength of a hydrogen bond between an NH and an N is 13 kiloJoules per mole [12]. This implies that the guanine-cytosine base pairs are an average of 8 kiloJoules per mole stronger than adenine-thymine base pairs. This extra strain makes it more difficult for a DNA sequence of 70 guanine-cytosine base pairs to form a cycle than for a DNA sequence of 70 adenine-thymine base pairs [see Figure 4].

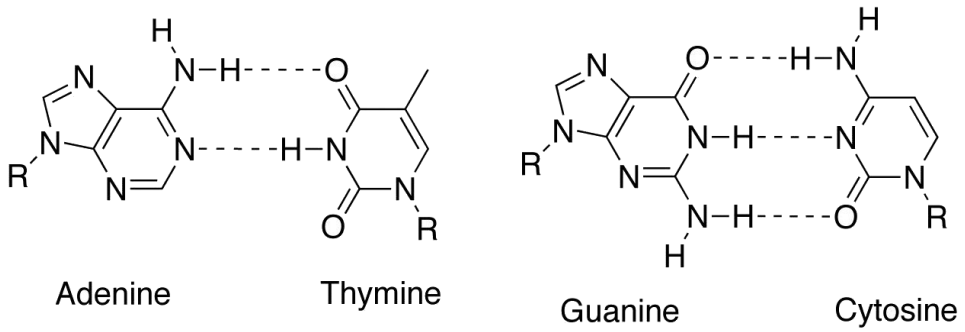


Figure 3: A-T and C-G base pairs.

An interesting observation concerning ciliate micronuclei is that the IESs are more densely packed with A-T base pairs than G-C base pairs. Most IESs have an average of 80% or more A-T base pairs. This likely plays a role in the bending of the DNA. Most IESs are smaller than the macronuclear destines sequences (MDSs), those sequences that make up the macronucleus. The length of the IES sequence alone is not enough to drive the sequence to bend to align the pointers [13]. The IESs are also known to be rich in transposons, which are mostly A-T rich sequences. These reasons are likely contributors to the fact that most IESs have a significantly higher content of A-T base pairs than C-G base pairs.

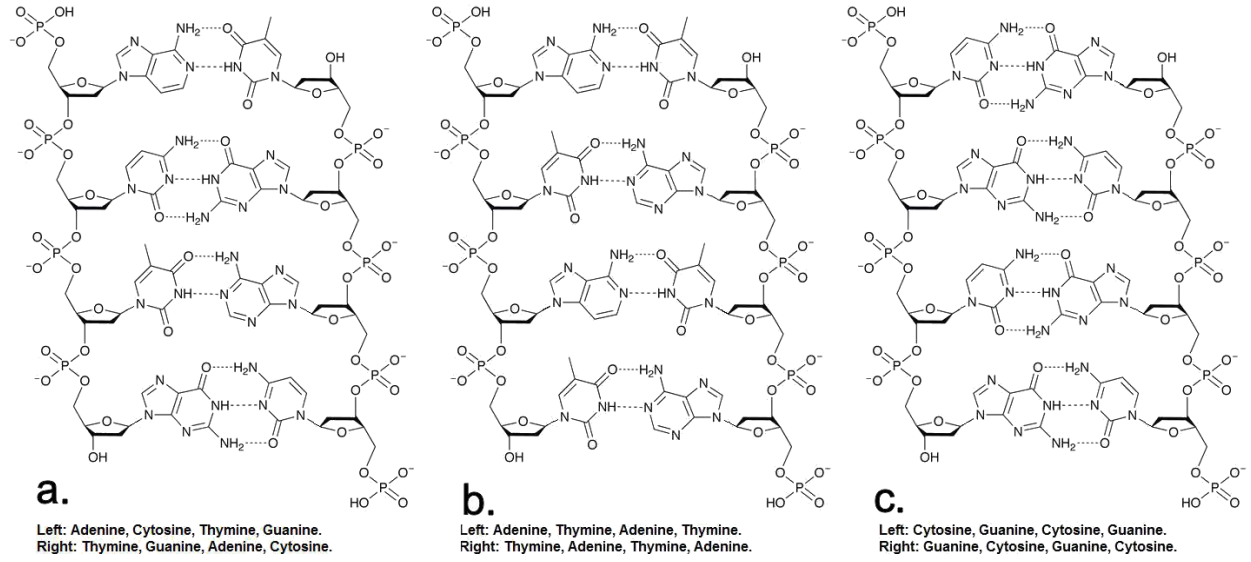


Figure 4: a.) DNA bonds with evenly distributed A-T and C-G base pairs. b.) DNA bonds with more densely distributed A-T base pairs than C-G base pairs. c.) DNA bonds with more densely distributed C-G base pairs than A-T base pairs.

DNA is made up of many base pairs, so each base pair having an extra hydrogen bond significantly strengthens the DNA helix. The concentration and spread of AT base pairs versus GC base pairs plays an important role in bending the DNA to align pointers. We use previously calculated wedge angles and direction angles to calculate the effect of AT base pairs versus CG

base pairs [see Figure 5]. The wedge angle is defined to be the deflection angle of the double helix with respect to each base pair. We find that two AT base pairs lower the amount of needed base pairs by approximately one in order to make a cycle. Similarly, two CG base pairs raise the amount of needed base pairs by approximately one in order to make a cycle. This has to do with the AA wedge angle being the largest as well as the direction angle of the CG-containing sequences being opposite of the DNA bending [14].

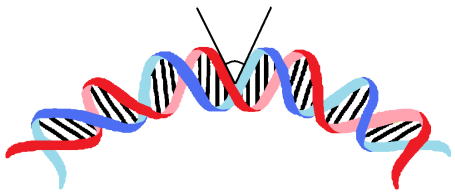


Figure 5: A wedge angle of a bent DNA sequence.

Experiments have shown that the flexibility of DNA greatly varies with different sequences. Certain sequence patterns cause the DNA to bend more than others. We also take this into account in formulating our algorithm. The main sequence pattern responsible for intrinsic DNA curvature is a run of adenines [14]. The base pairs of this particular sequence has a wedge angle of approximately 8.7°, which implies that it bends more readily than most other sequences. Other base pairs which are known to have favorable wedge angles, in addition to AA/TT are AG/CT, GA/TC, CG/CG, and GC/GC [15]. All possible wedge angles have been calculated, but only certain ones have been observed in nature [16]. Another sequence that has shown substantial intrinsic curvature is GGGCCC [17]. We will take into account the presence of these sequences in the DNA cycles we are attempting to form. The presence of any of the following sequences will increase the likelihood of a duplex DNA is less than 100 base pairs to form a cycle.

Again, we use previously calculated wedge angles and direction angles to calculate the effect of sequences with intrinsic bendability on the overall bendability of the DNA. We find that A tracks lower the amount of needed base pairs by approximately one in order to make a cycle. Once again, this has to do with the AA wedge angle being the largest. It is also seen that having repeated A tracks about every 10 base pairs achieves the highest intrinsic bendability [14].

2.3 Enzymes

If by the anaylisis of the DNA sequence a duplex DNA cannot bend due to it's length and nucleotide arrangements, but the molecule is observed to bend, enzymes are likely involved. DNA flexibility is dependent on DNA-protein interactions because many DNA-binding proteins bend DNA. Common proteins known to effect the flexibility of the DNA double helix include CRP and TBP. The cAMP receptor protein (CRP), also known as CAP (catabolite activator protein) is found to have a major role in bending DNA. Both biochemical data and the X-ray-derived structure of the CRP-DNA complex suggests that the interaction of CRP with DNA induces bending of the DNA helix. The X-ray-derived structure suggests that the bend is about 90° in the DNA. However, the bend is not

smooth, and is mostly attributed to two $40 - 45^\circ$ kinks at the pyrimidine - purine steps. Another factor affecting the bending of DNA is changes in DNA sequence outside the CRP binding site. This changes the binding affinity 10-fold and alters the bending angle by up to 30°. These effects are attributed to the influence of certain DNA sequences which increase the ease at which the DNA can bend. As noted, a short sequence of repeated adenines is found to produce bending in the DNA sequence. Crystal structures have also shown that the complex of the TATA-box binding protein (TBP) with DNA involves a bend angle of about 100° [18, 19]. We do not consider the possible presence and influence of these two enzymes in our algorithm as we have no experimental data that confirms expression of these proteins during the recombination process. We also assume that ligase is present during the process due to the influence it has on the cyclization of short DNA sequences. Note that in our previous discussion, we did not assume that ligase or any other enzymes were present or had any effect on the bending of our DNA sequences.

We propose to use previously calculated wedge angles and direction angles to calculate the effect of the TATA-box binding protein on the bendability of DNA [14]. Since the protein is known to bind to the sequence TATAAA, we consider at the angles associated with these base pairs. We find that this sequence lowers the amount of needed base pairs by approximately 10%, which is an astonishing 10 base pairs on a 100 base sequence, in order to make a cycle.

3 Mathematical Model

A *double occurrence word* is a word containing letters from a finite alphabet such that every symbol appears exactly twice. A *semi-double occurrence word* is a double occurrence word in which up to two symbols may appear only once. The term *directed graph* will be used to describe a four-tuple $G = (V, E, \iota, \tau)$, where V is the set of vertices, E is the set of edges, ι is the initial vertex and τ is the final vertex of a given edge. Where there is no confusion, we denote each edge $e(\iota(e), \tau(e))$, where $\iota(e)$ denotes the initial vertex of edge e and $\tau(e)$ denotes the terminal vertex of edge e. The edges are enumerated $E = \{e_1, ..., e_k\}$. A *four-valent rigid vertex* is a vertex of degree four for which a cyclic order of its incident edges is specified. An *assembly graph* is a finite directed graph consisting of four-valent rigid vertices, except for the two endpoints, each of which has degree one. A *semi-assembly graph* is an assembly graph in which up to two vertices may have degree two. A *path* is a sequence of vertices and edges such that for every edge in the sequence, the final vertex of an edge is the initial vertex of the succeeding edge except for the last edge. The initial vertex of the first edge is the first vertex of the path and the terminal vertex of the last edge is the last vertex of the path. A *cycle* refers to a path with no repetition of vertices and edges such that the initial vertex and the terminal vertex are the same [20] [See orange outline in Figure 6]. A *cluster* is a path with allowed repetition of vertices and edges such that the initial vertex and the terminal vertex are the same [See purple outline in Figure 6].

Assembly graphs are used to model possible pathways of unscrambling the micronuclear sequences. The vertices of the graph represent the pointers, or recombination regions, of the DNA sequence. All but two pointers, the first and last in the macronuclear sequence, are present twice in the micronuclear sequence and all pointers are present only once in the macronuclear sequence. The edges of the assembly graphs represent the MDSs and IESs out of the four edges incident to each vertex, two correspond to the two consecutive MDSs which share that pointer, and the other

two correspond to the IESs which are connected to that pointer in the micronuclear sequence. The vertices of the assembly graphs can be *smoothed* in a specific manner to produce the macronuclear DNA rearrangements as the final products. Different orders of smoothing vertices of the assembly graph correspond to the different pathways of unscrambling the micronuclear sequences. The vertices of the assembly graph can be smoothed in one of two ways: *parallel* or *non-parallel* [see Figure 7].

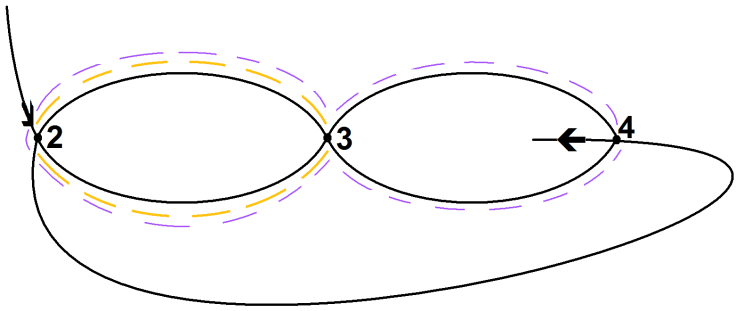


Figure 6: An assembly graph corresponding to the assembly word 234324. The arrows on the graph correspond to the direction of the DNA sequence. One of six cycles in this graph is outlined in orange. A cluster is outlined in purple.

Once all of the vertices are smoothed and the IESs are excised, we are left with a graph (line) representing the unscrambled macronuclear gene, referred to as the *unscrambled graph*. The assembly graph is called a *resulting graph* after some, but not necessarily all, of the vertices have been smoothed (it is the result of the original assembly graph after the smoothings occur).

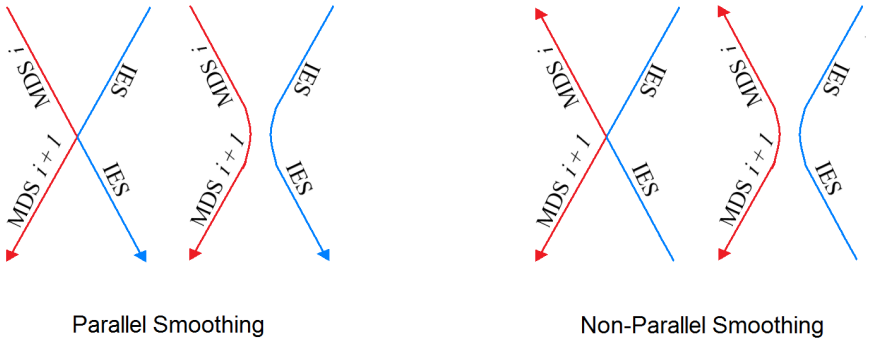


Figure 7: A parallel and a non-parallel smoothing of a vertex of an assembly graph [21].

4 Procedures

In this section, we describe the detailed procedures for implementing each step in the Algorithm, including how to construct the assembly graph and how to find the cycles in the assembly graph.

From the DNA Sequence to the Semi-Double Occurrence Word σ_g:

For a given gene g with micronuclear sequence consisting of k MDSs, M_{i_1}, IES_1, M_{i_2}, ..., M_{i_k}, where $i_1, i_2, ..., i_k$ are integers from 1 to k, form a word $s_g = a_1 a_2 ... a_k$, where $a_j = \begin{cases} i_j & \text{if } MDS_{i_j} \text{ is not inverted} \\ -i_j & \text{if } MDS_{i_j} \text{ is inverted} \end{cases}$. Use this to form the semi-double occurrence word $\sigma_g = b_1 b_2 ... b_{2k}$, where $b_{2j-1} b_{2j} = \begin{cases} a_j(a_j + 1) & \text{if } a_j > 0 \\ (-a_j + 1) - a_j & \text{if } a_j < 0 \end{cases}$. Note that all but two b_j's, 1 and $k+1$, appear twice.

From the Semi-Double Occurrence Word σ_g to the Semi-Assembly Graph γ_g:

Form a graph γ_g with vertex set $V(\gamma_g) = \{b_j : b_j \text{ is a symbol in } \sigma_g : 1 \leq j \leq k+1\}$ and edge set $E = \{e_j(b_j, b_{j+1}) : a \leq j \leq 2k-1\}$. This process creates the semi-assembly graph because all vertices which correspond to symbols that appear twice in σ_g are four-valent vertices but the two vertices which correspond to symbols 1 and $k+1$ that appear only once in σ_g, are only possible two-valent vertices. Before we "remove" the two valent vertices, we use them to label the edges.

From the Semi-Assembly Graph γ_g to the labeled Semi-Assembly Graph Γ_g:

The graph Γ_g is edge labeled with the MDS and IES segments that appear in gene g. For each edge e_j, if j is odd and if $b_j < b_{j+1}$, label the edge e_j $MDSb_j$. If j is odd and $b_j > b_{j+1}$, label the edge e_j $\overline{\text{MDS}b_{j+1}}$. If j is even, label the edge e_j $IES\frac{j}{2}$.

From the Labeled Semi-Assembly Graph Γ_g to the Assembly Graph $\mathcal{G}_g$:

For edges $e_i(a_i, b_i)$ and $e_j(b_i, c_i)$ with labels X and $Y, respectively$, incident at a vertex b_i, create an edge $\tilde{e}_{ij}(a_i, c_i)$ and label $\tilde{e}_{ij}(a_i, c_i)$ with XY. We call this process *gluing edges*, where $\tilde{e}_{ij}(a_i, c_i)$ is the *glued edge*.

If the labels on the edges are both MDS or IES, we say that the edges *share labels*.

In the semi-assembly graph, up to two vertices are not 4-valent, we glue these edges to obtain an assembly graph where all vertices have degree 4. We eliminate the presence of two-valent vertices because recombination does not appear at these vertices. If two edges $e_i(x, 1)$ and $e_j(1, y)$ exist, glue them and create edge $\tilde{e}_{ij}(x, y)$. If two edges $e_i(x, k+1)$ and $e_j(k+1, y)$ exist, glue them and create the glued edge $\tilde{e}_{ij}(x, y)$. The graph obtained in this way is $\mathcal{G}_g$.

Removing Conventional IESs of $\mathcal{G}_g$:

We suspect that conventional IESs are removed via a different process. *Conventional IESs* are represented by edges of the form $e_j(b, b)$ containing only one IES label. Remove all such edges. Place the verticies incident to removed edges in set $P_{0_\mathbf{j}}$ for all $\mathbf{j}$, where $\mathbf{j}$ will index the $\mathbf{j}$th pathway.

Finding Cycles of $\mathcal{G}_g$:

We search for cycles in $\mathcal{G}_g$ to determine whether the corresponding DNA sequence can cyclize and align pointers. We find all the cycles in the graph by using a slightly modified version of an already existing algorithm called the Depth First Search Algorithm [22]. We output the results of the algorithm by creating a set $\mathcal{C}_g = \{C_i : C_i \text{ is a set of edges in } \Gamma_g\}$, where $C_i = \{e_{i_1}, .., e_{i_s}\}$ is a set of edges describing each cycle because there is a sequence $e_{i_1}, ..., e_{i_s}$ that forms a cycle.

Lengths of Cycles:

Define the length of each cycle C_i to be $|C_i| = \sum_{e_j \in C_i} w(e_j)$, where $w(e_j)$ is the number of nucleotides in the label of edge e_j. Create a set $S = \{C_i : |C_i| \geq 100\}$. Define $N = \mathcal{C}_g \setminus S$.

Define function $L : \mathbb{N} \longrightarrow [0, 1]$ and set $L(C_i) = L_{C_i} = \frac{|C_i|}{100}$ for all $C_i \in N$ as the length parameter.

Sequences of Cycles:

Define $|C_i|_A$ to be the number of Adenines (As) in cycle C_i, $|C_i|_G$ to be the number of Guanines (Gs) in cycle C_i, $|C_i|_T$ to be the number of Thymines (Ts) in cycle C_i, and $|C_i|_C$ to be the number of Cytosines (Cs) in cycle C_i. For each cycle $C_i \in N$, define $p_{C_i} = |C_i|_A + |C_i|_T$ and $q_{C_i} = |C_i|_C + |C_i|_G$

Define function $M : \mathbb{N} \longrightarrow \mathbb{R}_+$ with $M(C_i) = M_{C_i} = L_{C_i} + 0.007p_i - 0.007q_i$ and compute M_{C_i} for each cycle $C_i \in N$. If $M_{C_i} \geq 1$, move cycle C_i from N to S.

Construct the Resulting Graphs $\mathcal{G}_{g_{i\mathbf{j}}}$:

Let $D = \{e_i : e_i \in \bigcup_{C_i \in N} C_i\}$, $F = E \setminus D$, and $\mathcal{F} = \{\iota(e_i), \tau(e_i) : e_i \in F\}$. The set $\mathcal{F}$ contains vertices that lie in cycles that can form. Now we define a *gluing set* $G' = \{(e_i(a_i, b_i), e_j(b_i, c_i)) : b_i \in \mathcal{F}$ and e_i and e_j share labels$\}$. Glue the pairs of edges in the gluing set G'. Define the set of vertices $P'_i = \{b_i : (e_i(a_i, b_i), e_j(b_i, c_i)) \in G'_{\mathbf{j}}\}$. Note that P'_i corresponds to the vertices which represent pointers that can be aligned during the ith step of the pathway. Define edge set E'_i to be all non-glued edges in E and all newly glues edges.

Now we begin developing the pathways. We find all the clusters K_i in $\mathcal{G}_{g_{i\mathbf{j}}}$ which contain cycles $C_i \in N$ using a slightly modified version of an already existing algorithm called the Depth First Search Algorithm and define $\mathcal{K}_g$ to be the set of all clusters [22]. Let $H = \{e_i : e_i \in \bigcup_{K_i \in \mathcal{K}_g} K_i\}$ and $\mathcal{H} = \{\iota(e_i), \tau(e_i) : e_i \in H\}$. Now we define an *attaching set* $G'' = \{(e_i(a_i, b_i), e_j(b_i, c_i)) : b_i \in \mathcal{H}$ and e_i and e_j share labels$\}$. Define a gluing set $G_{\mathbf{j}}$ to be the maximal subsets of G'' such that each edge appears in at most one attaching pair in G'', where $\mathbf{j}$ corresponds to the $\mathbf{j}$th pathway. For a fixed $i \in \mathbb{N}$, assign $\mathbf{j} \in \mathbb{N}$ to be the minimum natural number not already assigned to i. Glue the pairs of edges in the gluing set $G_{\mathbf{j}}$. Define the sets of vertices $P''_{i_{\mathbf{j}}} = \{b_i : (e_i(a_i, b_i), e_j(b_i, c_i)) \in G_{\mathbf{j}}\}$. Note that $P''_{i_{\mathbf{j}}}$ corresponds to different sets of vertices which represent pointers that can be aligned during the ith step of each pathway $\mathbf{j}$. Define edge set $E''_{i_{\mathbf{j}}}$ to be all non-glued edges in E'_i and all newly glues edges for each pathway $\mathbf{j}$.

For each $P''_{i_{\mathbf{j}}}$, define $P_{i_{\mathbf{j}}} = P'_i \cup P''_{i_{\mathbf{j}}}$ and $\tilde{E}_{i_{\mathbf{j}}} = E'_i \cup E'''_{i_{\mathbf{j}}}$. The set $P_{i_{\mathbf{j}}}$ corresponds to the ith element in the $\mathbf{j}$th pathway. The edges $\tilde{E}_{i_{\mathbf{j}}}$ and vertices $V_{i_{\mathbf{j}}} = V \setminus \cup P_{i_{\mathbf{j}}}$ describe the resulting graphs $\mathcal{G}_{g_{i\mathbf{j}}}$. Each resulting graph, $\mathcal{G}_{g_{i\mathbf{j}}}$, will be run through the algorithm again starting with finding the cycles of the graphs.

5 Algorithm

Input: Micronuclear DNA sequence with MDSs and IESs annotated.

1. Construct the semi-double occurrence word σ_g.
2. Construct the semi-assembly graph γ_g.
3. Construct the labeled semi-assembly graph Γ_g.

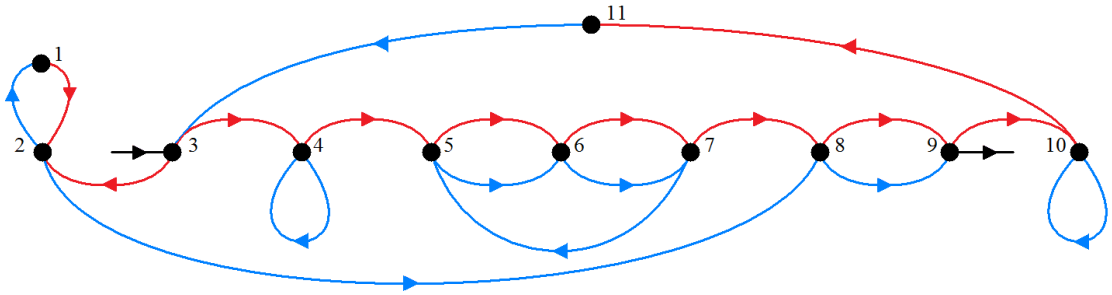
4. Construct the assembly graph $\mathcal{G}_g$ and define the resulting graph $\mathcal{G}_{g_{i\mathbf{j}}}$ to be the empty graph for all $\mathbf{j}$. Set $i = 1$ (1st run through the algorithm) and assign a $\mathbf{j}$ to i.
5. Remove the conventional IESs of $\mathcal{G}_g$ and create sets $P_{0_\mathbf{j}}$ for all $\mathbf{j}$.
6. If $\mathcal{G}_g \neq \mathcal{G}'_g$ find the cycles $\mathcal{C}_g$ of $\mathcal{G}_g$. Otherwise $\mathcal{G}_g = \mathcal{G}'_g$, terminate.
7. Compute L_{C_i} for each $C_i \in \mathcal{C}_g$.
8. Compute M_{C_i} for each $C_i \in \mathcal{N}$.
9. Construct the resulting graphs $\mathcal{G}_{g_{i\mathbf{j}}}$, add $P_{i_\mathbf{j}}$ to $\mathcal{S}_{g_\mathbf{j}}$, and for each $\mathcal{G}_{g_{i\mathbf{j}}}$, increase i by one.
10. Go back to step 6.

Output: Sequences of sets of vertices $\mathcal{S}_{g_\mathbf{j}} = \{\{P_{1_\mathbf{j}}\}\{P_{2_\mathbf{j}}\}...\{P_{s_\mathbf{j}}\}\}$.

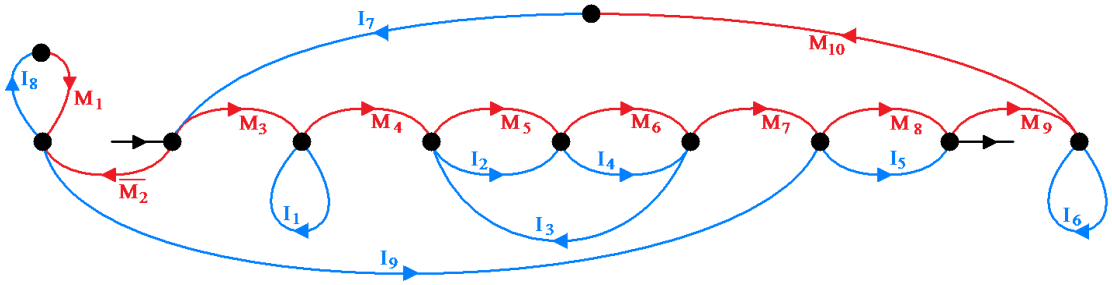
Example: Sterkiella Histriomuscorum Actin I gene

Input: $M_3\ I_1\ M_4\ I_2\ M_6\ I_3\ M_5\ I_4\ M_7\ I_5\ M_9\ I_6\ M_{10}\ I_7\ \overline{M_2}\ I_8\ M_1\ I_9\ M_8$

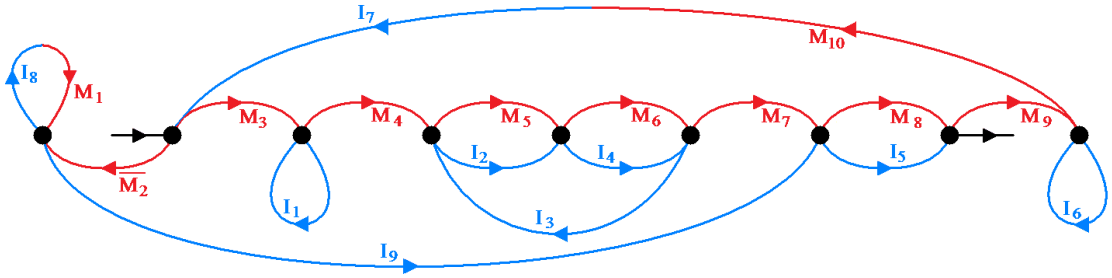
1. σ_g: 3, 4, 4, 5, 6, 7, 5, 6, 7, 8, 9, 10, 10, 11, 3, 2, 1, 2, 8, 9
2. γ_g:



3. Γ_g:

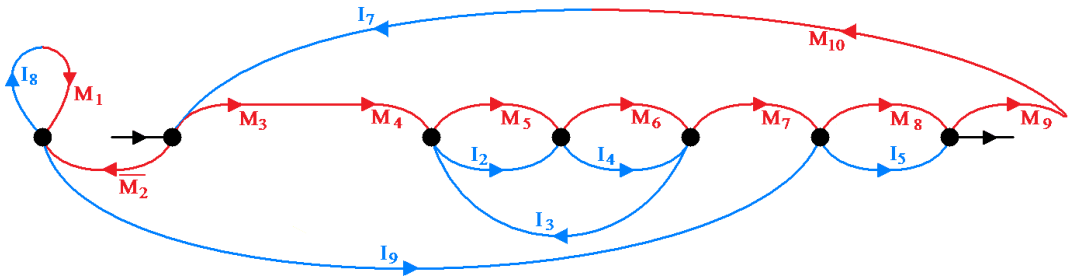


4. $\mathcal{G}_g$:



5. Remove the conventional IESs:

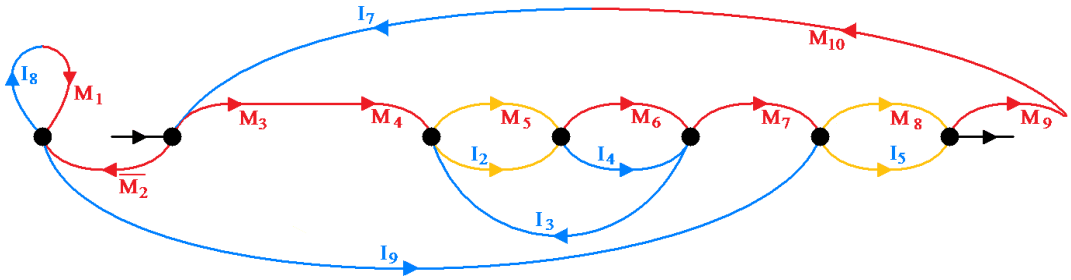
   $P_{0_j} = \{4, 7\}$



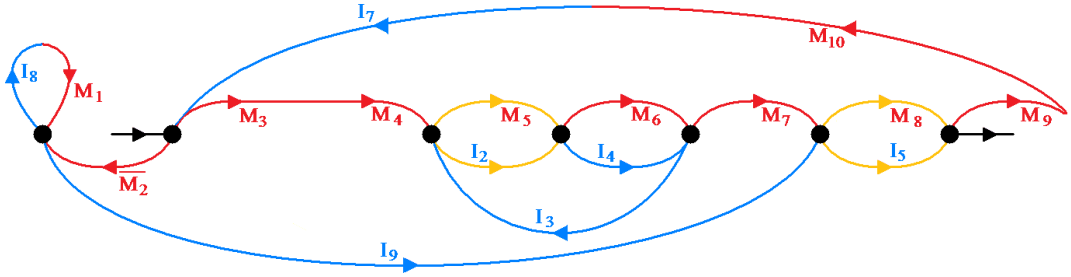
6. Cycles $\mathcal{C}_g$ of $\mathcal{G}_g$:

```
cycle
-----------------
Edge (1, 2), Copies: 1
Edge (2, 3), Copies: 1
Edge (3, 4), Copies: 2
Edge (4, 5), Copies: 2
Edge (5, 6), Copies: 1
Edge (6, 7), Copies: 2
Edge (7, 8), Copies: 1
Edge (1, 8), Copies: 1
```

7. Apply length parameter to $\mathcal{C}_g$:

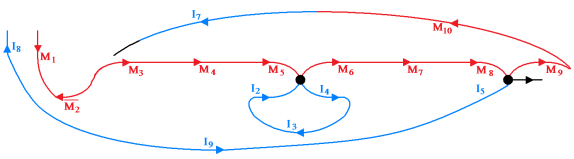


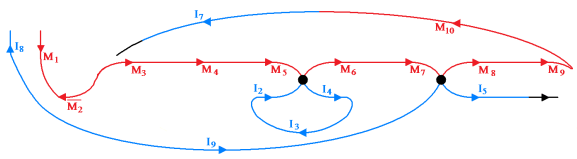
8. Apply sequence parameter to $\mathcal{C}_g$:

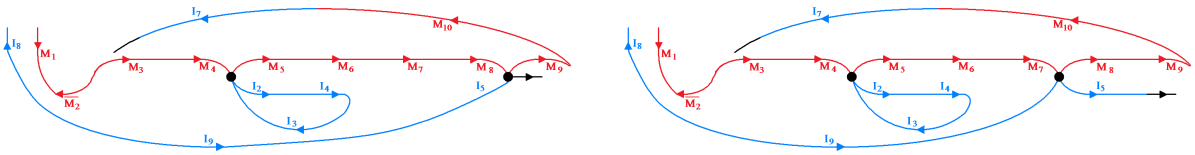


9. Resulting graphs $\mathcal{G}'_{g_{ij}}$

   $P_{1_1} = \{2, 3, 5, 7, 8\}$, $P_{1_2} = \{2, 3, 5, 7, 9\}$, $P_{1_3} = \{2, 3, 6, 7, 8\}$, $P_{1_4} = \{2, 3, 6, 7, 9\}$

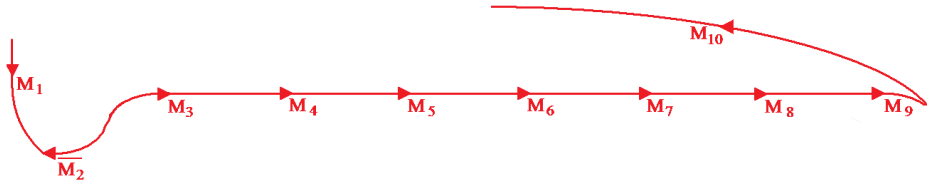






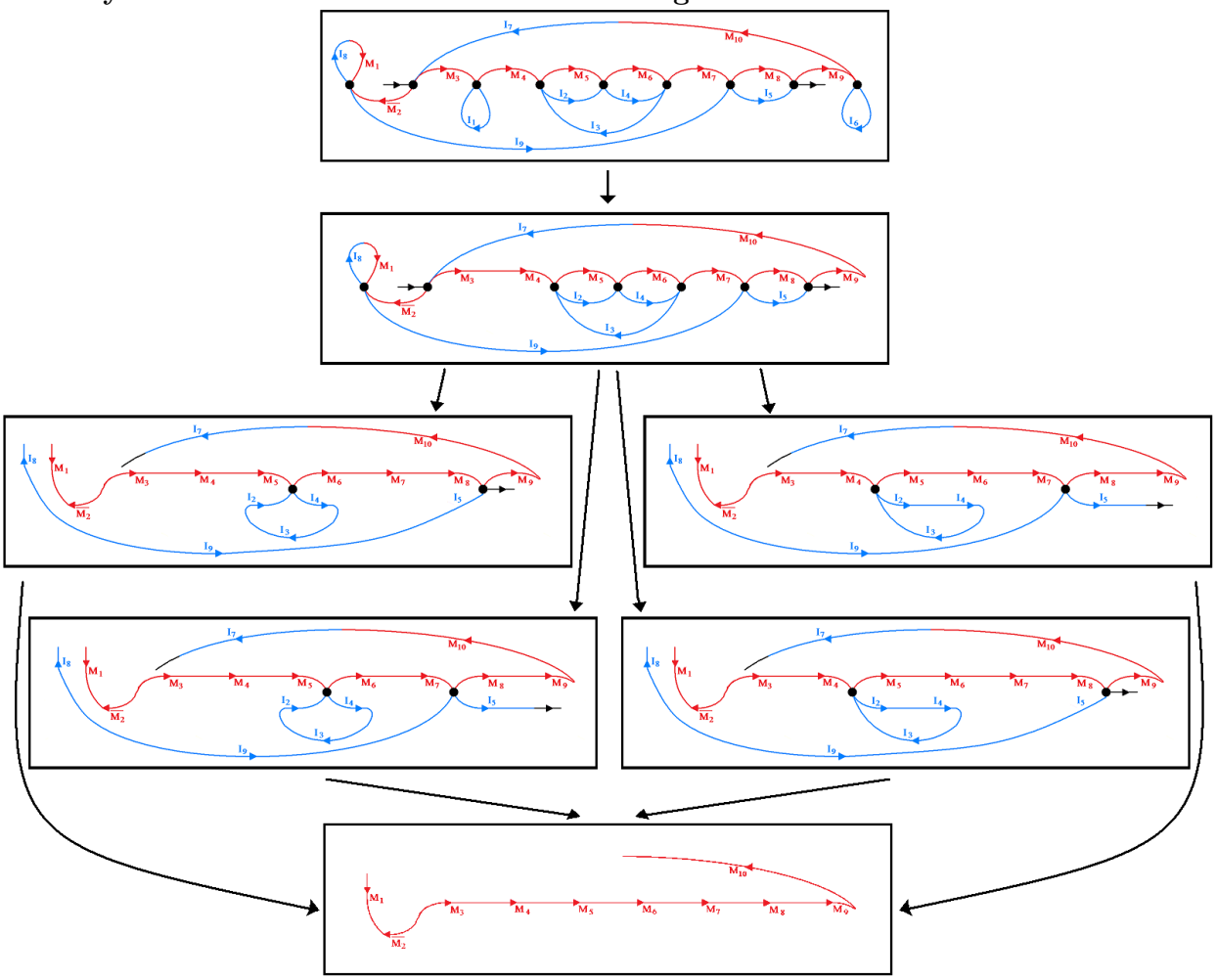
10. Go back to step 5. Repeat through step 9.

Output:
$\mathcal{S}_{g_1} = \{\{4,10\},\{2,3,5,7,8\},\{6,9\}\}$, $\mathcal{S}_{g_2} = \{\{4,10\},\{2,3,5,7,9\},\{6,8\}\}$,
$\mathcal{S}_{g_3} = \{\{4,10\},\{2,3,6,7,8\},\{5,9\}\}$, $\mathcal{S}_{g_4} = \{\{4,10\},\{2,3,6,7,9\},\{5,8\}\}$



Note: All of the data used in the algorithm was obtained from Laura Landwebber's Lab [23].

Pathways: Sterkiella Histriomuscorum Actin I gene



6 Concluding Remarks

The steps of the algorithm will assign a value to each portion of the DNA sequence. It will then remove all of the disjoint portions with the acceptable values. Once all of these portions corresponding to the acceptable values are removed, the new sequence produced will go through the algorithm again. This process repeats until the last IES is excised and the unscrambled sequence is produced or until the algorithm terminates. The order in which the excision of the IESs and the alignment of the pointers occur will be recorded and correspond to the most likely pathway or pathways of the unscrambling of the gene.

If the algorithm terminates without producing the unscrambled graph, additional enzymes are assumed to be present and influence the alignment of pointers. Other common enzymes involved in the manipulation of DNA include *helicase*, *nuclease*, and *topoisomerase*. Helicase refers to a family of enzymes responsible for unwinding the DNA double helix. This enzyme is known to have a role in transcription but may also likely have a role in loosening the double helix to require less energy to bend. Nuclease is a family of enzymes that cuts strands of DNA into smaller pieces, often associated with ligase. Topoisomerase is an enzyme family that performs the functions of nuclease and ligase at the same time. It cuts the DNA and rejoins it back together. These enzymes are also known to have a role in DNA transcription buy may also assist in making DNA easier to bend. Other enzymes that may have a role in chemically modifying DNA and possibly making it easier to bend are *kinase* and *phosphotase* [4]. These enzymes, as well as countless others, are responsible for attaching and detaching certain functional groups or inorganic molecules, such as phosphates, to the DNA. These molecules can have a chemical reaction which may make the DNA easier to bend. We consider these enzymes the last resort for the DNA sequences to bend, We propose that with their involvement, any pointers can align, therefore we do not consider them in the algorithm. However, if the algorithm predicts that bending of a particular portion of the sequence is not likely to occur, these enzymes are assumed to do the job of aligning the pointers, and these steps are considered to occur last in the algorithm, and consequently, in the pathway.

7 References

[1] Prescott, DM. Genome gymnastics: unique modes of DNA evolution and processing in ciliates. Nat Rev Genet. (2000) 191-198.

[2] Mollenbeck M, Zhou Y, Cavalcanti ARO, Jonsson F, Higgins BP, et al. (2008) The Pathway to Detangle a Scrambled Gene. PLoS ONE 3(6): e2330. doi:10.1371/journal.pone.0002330.

[3] R.L. Baldwin, J. Langowski, D. Shore, Formation of small circular DNA molecules via an in vitro site-specifc recombination system, Proc. Natl. Acad. Sci. (USA) 78 (1981) 4833-4837.

[4] Rawn, J. David (1989). Biochemistry, International Edition

[5] Nelson, David L., Michael M. Cox, and Albert L. Lehninger. Lehninger: Principles of Biochemistry. New York: W. H. Freeman and, 2005. Print.

[6] Sutcliffe, J. G. "Supplemental Content." National Center for Biotechnology Information. U.S. National Library of Medicine, 1978. Web. 22 July 2012. <http://www.ncbi.nlm.nih.gov/nuccore/208958>.

[7] K. Abremski, R. Hoess, A. Wierzbicki, Formation of small circular DNA molecules via an in vitro site-specific recombination system, Gene 40 (1985) 325-329.

[8] Vafabakhsh, Reza, and Taekjip Ha. "Extreme Bendability of DNA Less than 100 Base Pairs Long Revealed by Single Cyclization." Science (2012): 1097-1101. Print.
[9] Vologodskii, Alexander, Du Quan, and Maxim Frank-Kamenetskii. "Bending of short DNA helices." Artificial DNA: PNA and XNA (2013): 1-3. Print.
[10] Legon, A. C.; Millen, D. J. (1987). "Angular geometries and other properties of hydrogen-bonded dimers: a simple electrostatic interpretation of the success of the electron-pair model". Chemical Society Reviews 16: 467.
[11] Larson, J. W.; McMahon, T. B. (1984). "Gas-phase bihalide and pseudobihalide ions. An ion cyclotron resonance determination of hydrogen bond energies in XHY- species (X, Y = F, Cl, Br, CN)". Inorganic Chemistry 23 (14): 20292033.
[12] Emsley, J. (1980). "Very Strong Hydrogen Bonds". Chemical Society Reviews 9 (1): 91124.
[13] Verlan, Sergey, Artiom Alhazov, and Ion Petre. "A Sequence-based Analysis of the Pointer Distribution of Stichotrichous Ciliates." Biosystems 101.2 (2010): 109-16. Print.
[14] Bolshoy A, et al.. "Curved DNA without A-A: Experimental Estimation of all 16 DNA Wedge Angles." Proc. Natl. Acad. Sci. USA (1991): 2312-2316. Print.
[15] Ulanovski, Levy, et al. "Curved DNA: Design, Synthesis, and Circularization." Proc. Natl. Acad. Sci. USA (1986): 862-866. Print.
[16] Kabsch W., C. Sander, and E.N. Trifonov. "The Ten Helical Twist Angles of B-DNA." Nucleic Acids Research (1981): 1097-1104. Print.
[17] Geggier, Stephanie, and Alexander Vologodoskii. "Sequence Dependence of DNA Bending Rigidity." PNAS (2010): 14521-14526. Print.
[18] Bates, Andrew D., and Anthony Maxwell. DNA Topology. Oxford: Oxford UP, 2005. Print.
[19] Goodsell, David S., and Richard E. Dickerson. "Bending and Curvature Calculations in B-DNA." Nucleic Acids Research(1994): 5497-5503. Print.
[20] Adams, Colin Conrad., and Robert David. Franzosa. Introduction to Topology: Pure and Applied. Upper Saddle River (NJ): Pearson Prentice Hall, 2008. Print.
[21] A. Angeleska, N. Jonoska. M. Saito, DNA Rearrangement through assembly graphs. Discrete and Applied Math, 157 (2009) pp. 3020-3037.
[21] Tarjan, R. "Depth-First Search and Linear Graph Algorithms." SIAM Journal on Computing. (1972) 1:2, 146-160. Print.
[23] A. Cavalcanti, T.H. Clarke, L. Landweber, MDS IES DB: a database of macronuclear and micronuclear genes in spirotrichous ciliates, Nucleic Acids Research 33 (2005) 396-398.