Simulation of the 2JLP Gene Assembly Process in Ciliates*

Md Sowgat Ibne Mahmud and Ian McQuillan*

University of Saskatchewan, Saskatoon, Saskatchewan, Canada,
mdm179@mail.usask.ca, mcquillan@cs.usask.ca **

Abstract. The gene assembly process in ciliates consists of a massive amount of DNA excision from the micronucleus and sometimes the rearrangement of the rest of the DNA sequences. Several models exist that describe certain parts of this process. In this research, a simulation is created and tested with real data to test the feasibility of the 2JLP model. Several parameters are introduced in the model that are used to test ambiguities or edge cases of the biological model. Parameters are systematically varied within the simulation to try to find their optimal values. Interestingly, a negative correlation is found between the degree to which the simulation successfully descrambles genes, and a parameter that is used to filter out scnRNAs that are similar to IES specific sequences from the macronucleus. This provides *in silico* evidence that if a scnRNA consists of both a portion of MDS and IES, then from the perspective of maximizing the accuracy of the descrambling, it is desirable to filter out this scnRNA. The simulator successfully performs the gene assembly process whether the inputs are scrambled or unscrambled DNA sequences. On average, before the proof checking stage that is in the model, the descrambling intermediate genes are 91.1% similar to the descrambled genes. After the proof checking stage, the intermediate genes are 99.4% similar. We hope that this work and further simulations can serve as a foundation for future computational and mathematical study of descrambling, and to help inform and refine the biological model.

Keywords: biological simulation, template guided recombination, scan RNAs, scrambled genes, gene assembly, ciliates, natural computing.

1 Introduction

Ciliates are a group of unicellular protozoa characterized by the presence of hair-like organelles called cilia. Worldwide, 4,500 different species of ciliates are known [1]. Two distinct types of nuclei are present in each cell, called the *micronucleus* and the *macronucleus* [14]. The macronucleus produces all the RNA

* Published in Proceedings of UCNC 2015. The final authenticated version is available online at http://dx.doi.org/10.1007/978-3-319-21819-9_17.

** This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

and proteins needed for day-to-day operations, and the micronucleus remains silent functionally, except after conjugation when certain micronucleus specific genes get expressed [3]. During the period of conjugation, ciliates destroy their macronuclei and exchange haploid micronuclei. Each then constructs a fully functional macronucleus from the micronuclear genome by doing a massive quantity of DNA excision and rearrangement [13, 6, 8].

Micronuclear genes have two classes of DNA sequences— non-coding DNA segments that get excised in the conversion, known as *IESs* (internal eliminated sequences) and segments that are retained, known as *MDSs* (macronuclear destined DNA sequences). A functional macronucleus can be constructed by deleting IESs and merging MDSs from the micronucleus. Different ciliates perform the gene assembly process in different ways. In the case of two genera of ciliates *Tetrahymena* and *Euplotes*, the MDSs of the micronucleus are interrupted by IESs but the MDSs occur in the same order as in the macronucleus. But in the case of stichotrichs (containing genera *Stylonychia* and *Oxytricha*), the MDSs are not only interrupted by IESs, but the MDSs can also occur in a scrambled order. Fig. 1 shows a diagram of a scrambled micronuclear gene, and the descrambled variant.

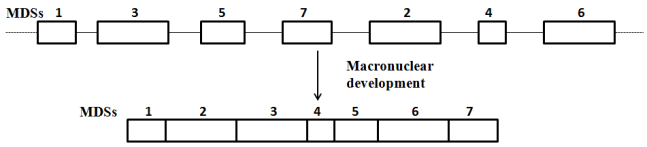

Fig. 1. During macronuclear development, IESs (the lines between the boxes) are excised from the micronucleus and the MDSs are joined in the correct order to yield a macronuclear gene.

In stichotrichous ciliates, IESs are flanked by repeat sequences called *pointers* [18]. These pointers are less than 20 bp in size with one copy of the pointer at the $3'$-end of one MDS and the other copy at the $5'$-end of the next MDS (next MDS according to the correct ordering in the macronucleus) [15, 17]. IESs are excised between two adjacent MDSs along with one copy of the pointer.

There are a variety of biological models and hypotheses that have been created to model the gene assembly process in ciliates, such as the intramolecular model [16], the intermolecular model [7], the *scnRNA model* [9], the *template guided model* [17], and the *2JLP model* [5] (to be described in Section 2). And from a number of those, formal models have been created in an attempt to capture the biological models. All the existing models appear to capture at least part of the gene assembly process, even though some have experimentally verified limitations in scope [10]. More recently, the 2JLP model [5] was created, and involves a combination of the scnRNA model for excising IESs from the micronucleus, and the template-guided model for removing the remaining IESs, for rearranging MDSs, and for a proofreading process.

This paper briefly describes existing formal and biological models as well as known limitations in Section 2. Then, a simulation is presented together with the algorithms to capture and analyze the 2JLP model. The implemented algorithms are provided and discussed in Section 3. Later on, the outcome of the simulation is discussed based on its use with real micronuclear and macronuclear genes. In the simulator, some important parameters are considered such as the minimum value needed for sufficient similarity between scnRNAs (small RNAs) and MDSs, and the minimum value to needed for sufficient similarity between filtered scnRNAs and the new micronucleus in order to identify subwords for deletion. These parameters are used to deal with ambiguities in the biological model and to determine optimal values according to the simulation. Indeed, the primary motivation of the work lies in its potential towards:

- unifying existing models and new aspects into well-defined algorithms while capturing the biological 2JLP model, thereby establishing which aspects of existing models are compatible with each other,
- building simulations to test the feasibility of the model and its consistency with real micronuclear and macronuclear genes,
- and testing and resolving ambiguities of existing models through systematic variation of parameters.

Furthermore, as far as the authors are aware, the use of computer simulations to test a gene assembly model is novel, and may contribute techniques towards new biological models such as the more recent piRNA model [1] for *Oxytricha*, which has some similar aspects.

2 Existing Models

A variety of biological and formal models have emerged that attempt to explain different parts of the gene assembly process. In this section, we briefly describe some of the models, as they relate to the 2JLP model.

Landweber and Kari proposed a model for gene assembly known as the *intermolecular model* [7]. It consists of one unary intramolecular and two binary intermolecular operations of DNA recombination on pointers. Another model for gene assembly was introduced by Prescott et al. [16] and Ehrenfeucht et al. [4] called the *intramolecular model*. It consists of three unary molecular operations based on pointers. One of the major limitations in scope of these models is that they do not discuss the process of pointer identification, as pointers are too short to uniquely identify their other copy.

A model for the gene assembly process was proposed by Mochizuki et al. in 2002 based on small RNAs, called the *scan RNA* (scnRNA) model [9]. They proposed that during the early conjugation period, a RNAi-related pathway starts with a bi-directional transcription of the micronucleus. From that, it generates small RNAs of size $28 - 29$ bp also known as *scnRNAs*. These localize to the parental macronucleus where all scnRNAs that are similar to some segment of the parental macronucleus degrades. The rest of the scnRNAs that fail to degrade

are therefore likely similar to IES-specific sequences. Then these IES-specific scnRNAs travel to the developing macronucleus where they eliminate subsequences that are similar. A limitation of this model is that it does not address MDS reordering. Moreover, the model does not easily explain IES removal for cases where IESs are smaller than scnRNAs.

In a key experiment on the ciliate *Paramecium tetraurelia* (that does not have scrambling, but does have IESs), an *IES* was injected into a macronucleus before mating (so that a portion of the macronuclear gene "looked like" the micronuclear version, with two MDSs separated by an IES) [15]. Then, the ciliate was allowed to traverse into the sexual cycle, after which it was found that this particular *IES* was present in the structure of the new macronucleus. As a result of this experiment, it was thought that some sequence-specific information must be transferred from the parental macronucleus to the new macronucleus. Hence, a biological model of gene assembly was introduced by Prescott et al. in 2003 and is known as the *template guided model* [17]. In this model a molecule (later determined to be RNA [12]) that has been generated from the parental macronucleus is used as a template to guide both IES removal and MDS reordering in the developing macronucleus. A limitation of this model can be seen by examining the notion of *cryptic pointers*, which are direct repeats of length 1–8 that are in proximity to real pointers. In fact, despite not being the real pointers, ciliates frequently use cryptic pointers for splicing. It was observed in an experiment [10] that IESs are deleted randomly and sometimes imprecisely (when IESs are removed based on cryptic pointers) at the middle-late stage of macronuclear development. These become corrected at a later stage.

Despite the limitations of these models, there is indeed evidence that the *scnRNA* model does filter out IESs from the new micronucleus. There is also other evidence that some parts of the template model must also be true, with template molecules being present, and influencing the resulting macronucleus. Based on this, a biological model of gene assembly was proposed by Jönsson et al. [5]. It is known as the *2JLP* model, and it unifies portions of the previous models, which all occur within a temporal procedure (Fig. 2) summarized as follows:

Definition 1 (2JLP Model). *This model can be defined by the following steps:*

1. *During the early period after conjugation, each ciliate generates scnRNAs. The genome of the micronucleus is transcribed bi-directionally and the resulting transcripts generate double-stranded RNA molecules which are eventually processed into scnRNAs.*
2. *These scnRNAs travel to the parental macronucleus and any scnRNAs similar to DNA sequences in the parental macronucleus are degraded.*
3. *In the late conjugation stages, the remaining portion of the scnRNAs (that are similar to IESs) are transferred to the developing new macronucleus, where they target and identify IESs to be eliminated by base pairing between repeats (either real or cryptic pointers).*

4. *At the same time, the template guided model generates template RNAs from the parental macronucleus to guide the alignment of MDSs and their pointer sequences, and produces the new macronucleus.*
5. *In the case of scrambled genes, the template RNAs perform unscrambling of MDSs according to their order in the macronuclear chromosomes. Homologous recombination between the aligned pointers splice out IESs. For IES excision, if cryptic pointers are used instead of real pointers, a proofreading mechanism guided by the template ensures the missing sequences are filled in and the extra sequences are removed to create full-length chromosomes.*

More recently the procedure has been discovered to be different between *Tetrahymena* and *Oxytricha*, where *Tetrahymena* uses scnRNAs from the old micronucleus as in the 2JLP model, whereas *Oxytricha* uses 27bp small RNAs (called piRNAs) from the old macronucleus to mark MDS regions in the developing macronucleus. A simulation involving the latter model is left as future work.

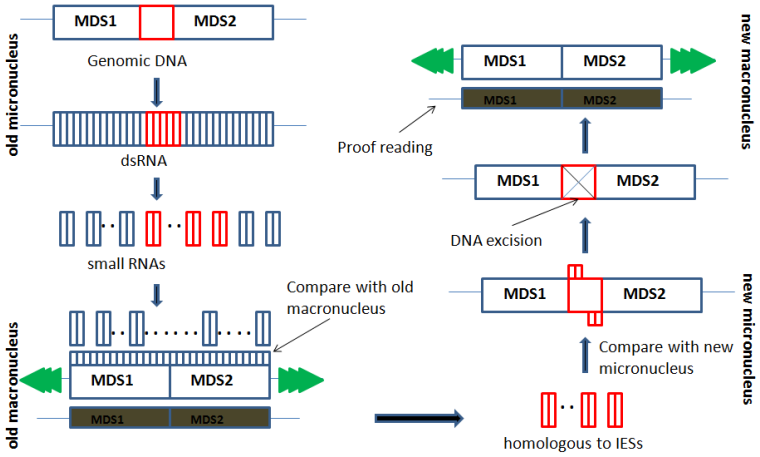

Fig. 2. The 2JLP model combines aspects from the scanRNA model and the template-guided model to explain the gene assembly in ciliates: the whole micronuclear genome is transcribed early in macronuclear development into long double-stranded transcripts, which are processed into small RNAs (scnRNAs). These invade the old macronucleus. There, scnRNAs similar to macronuclear sequences (dark blue) degrade. The rest of the scnRNAs (red) are sent to the new micronucleus for marking and excision of IESs by recruiting chromatin-modifying proteins to the micronuclear-specific sequences. Imprecisely processed sequences will be corrected by a proofreading mechanism that is guided by template RNAs (gray). These template RNAs originate from the old macronucleus. In scrambled genes, the template RNAs guide alignment of micronuclear MDSs in the correct order of the template, creating a new macronucleus.

3 Simulation

This section describes the implementation for the simulation of the 2JLP model. The purpose of developing the simulation is to test the model's feasibility, and determine additional important aspects regarding the gene assembly process by analysing the results with real data. These findings can be helpful for refining the 2JLP model. For example, in the algorithm, certain values are parameterized that were left ambiguous or not described in the biological model of Definition 1. Then, it can be tested which values for the parameters give optimal results.

Fig. 3 shows the flow diagram of the pipeline used to simulate the 2JLP model (each part explained in Section 3.1). Global sequence alignment and semi-global alignment are used within, which are the standard Needleman-Wunsch algorithm to compute the optimal global sequence alignment and its semi-global variant [11]. For scoring alignments, a match score of 1, mismatch score of -1, and gap penalty of -2 are used.

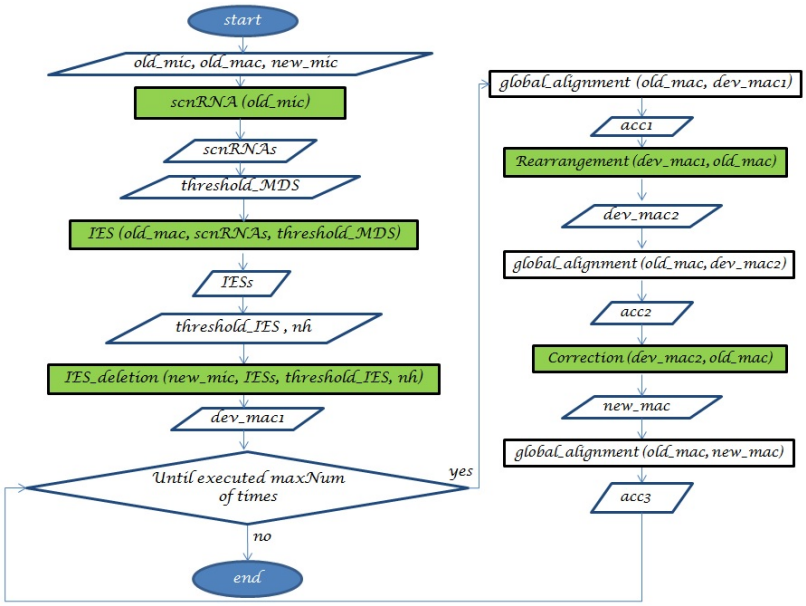

Fig. 3. Flow diagram of the simulator. Major functions of the simulator are represented by green shaded rectangles. The parameters are explained in Table 1. Each part of the pipeline will be explained in Section 3.1.

From Fig. 3, it is shown that the simulator has five major functions (green shaded rectangles). These are *scnRNA*, *IES*, *IES_deletion*, *rearrangement*, and *correction*. Among these, the scnRNA function closely simulates the construction process of scnRNAs (first step of Definition 1). The IES function simulates the mechanism of finding putative IESs (second step of Definition 1). The

IES_deletion function simulates the mechanism of deleting IESs from the new micronucleus (third step of Definition 1). For simulating the construction process of the new macronucleus (forth and fifth steps of Definition 1), both the *rearrangement* and the *correction* (simulating the proof-checking step) functions are used.

3.1 Important Parameters, Algorithms, and Methods of Evaluation

Three important parameters, *threshold_MDS*, *threshold_IES*, and *nh* are considered. They can take on a range of possible values, and all integer values within the range are simulated for each. Their meaning and ranges are described below.

It is possible to determine "how much" of the gene has been descrambled after each stage of the simulation by computing the similarity of the developing gene to the macronuclear gene. If the percent similarity (computed from a global sequence alignment) between the string computed thus far shows similarity to the fully descrambled variant, then it has been largely descrambled. Hence, three percentages are calculated throughout the simulation: *acc1* is computed after deleting IESs from the micronucleus (*dev_mac1*), *acc2* is computed after rearrangment based on templates (*dev_mac2*), and *acc3* is computed after proof checking (*new_mac*).

The implemented algorithm takes three input strings: *old_mic*, *old_mac*, and *new_mic*. Of these, *old_mic* and *old_mac* are a single matching micronuclear gene and macronuclear gene, respectively.

The *scnRNA* function generates all possible scnRNAs through a "sliding window" technique. In the *scnRNA* function, *old_mic* is divided into all subwords of length 28, as is done in step 1 of Definition 1. The output, an array called *scnRNAs* is taken along with *old_mac* as inputs to the *IES* function to generate an array called *IESs*. This function compares each element from *scnRNAs* with the parental macronucleus, and if there is a match that is "similar enough", it gets filtered out as it will largely be MDS specific. However, "similar enough" is ambiguous, and therefore a semi-global alignment between the scnRNA of length 28 and the macronuclear gene is used with the parameter *threshold_MDS*. This is the parameter representing the minimum score needed to classify as "similar". Ultimately, the simulation is tested for all values of this parameter between 1 and 28, as there is no indication in the model as to what degree of similarity is needed.

Table 1. List of the important parameters

Parameter name	Range	Purpose of parameter
threshold_MDS	1 to 28	The minimum score of the semi-global alignment needed for sufficient similarity between scnRNAs and *old_mac*.
threshold_IES	1 to 28	The minimum score of the semi-global alignment needed for sufficient similarity between filtered scnRNAs and *new_mic*.
nh	1 to 20	The size of the neighbourhood indicates the area around where filtered scnRNAs match the developing macronucleus.

Then, *new_mic*, *IESs*, and *threshold_IES* are inputs to the *IES_deletion* function to generate *dev_mac1* (the gene obtained after IES removal in step 3 of Definition 1). This function has two parts. In the first part, it compares all strings of *IESs* with *new_mic* and performs a "marking of matched subwords" (simulated by keeping track of the start and end positions in *new_mic* of where it matches any string in *IESs*). In the second part, it removes each subword from *new_mic* if it has a repeated segment of length between 2 and 20 "close to" the ends of the marked portion. The range of between 2 and 20 is chosen as these are the allowable lengths of pointers. An important aspect of this simulation is that these repeated strings do not need to be the real pointers. Indeed, Möllenbeck et al. [10] show that often cryptic pointers are used for splicing in proximity to the MDS-IES junctions (see Definition 1 and the preceding discussion). Also, it is possible that the repeated sequence is a part of an IES (which would result in a portion of the IES remaining after deletion), but it is also possible that the repeated sequence could be part of an MDS as well (which would result in part of that MDS being missing). That is why in the algorithm, a parameter named "neighbourhood (*nh*)" is taken to address the range of possible distances of cryptic or real pointers to the marked portion. The model dictates that this is close to the MDS-IES junctions and thus the parameter *nh* represents the largest distance allowed, which is simulated for all values up to 20. As the selection of repeats used for splicing is not always the same, we select repeats randomly within the neighbourhood. However, the final descrambled gene will depend on the random values chosen. Therefore, this step is simulated four times (represented by *maxNum* in Fig. 3) for each pointer, and for each value of *nh* to select different repeats from *new_mic*, eventually generating many different values of *dev_mac1*. The value *acc1* measuring the percent similarity is stored at this stage, for each value of *dev_mac1*.

Then, the *rearrangement* function is used to generate *dev_mac2* by taking *dev_mac1*, and *old_mac* as inputs. The main purpose of this function is to rearrange MDSs from *dev_mac1* based on the old macronucleus (*old_mac*). A precise method to predict the order in which MDSs descramble is not known, and therefore, our simulation of this stage is a simplification of the actual biological procedure. Indeed, our method randomly picks a locus from the template (which is the same as the parental macronuclear gene), finds a similar segment in *dev_mac1* (on either the sense or antisense strand), extends in both directions, and repeats until all segments of the template are matched, and then the matched segments are rearranged, creating *dev_mac2*. At this stage, the second percent identity *acc2* is calculated to quantify the degree to which the gene has been descrambled.

At this point, the *correction* function is applied. In this function, the final macronucleus (*new_mac*) is generated by comparing *dev_mac2* and the template, simulated with a sequence alignment. Based on the alignment, extra characters are removed from *dev_mac2* (from gaps along the template) and missing characters are inserted (from gaps in *dev_mac2*) into *dev_mac2*. Then, the final percent identity *acc3* is calculated from the resulting descrambled gene *new_mac*.

4 Results and Analysis

In the simulation, for each set of fixed parameters for *threshold_MDS*, *threshold_IES*, and neighbourhood, results are calculated at three different stages to measure the change from the new micronucleus to the new macronucleus. These three different stages are after the *IES_deletion* function, after the *rearrangement* function, and finally after the *correction* function. The term *accuracy* is defined to represent the degree to which descrambling has occurred at the various stages.

Input data was collected from the IES MDS Database [2]. From there, 13 real micronucleus and macronucleus matching gene pairs of the ciliate *Oxytricha trifallax* are used in the simulation. Although this is a limited number of pairs of genes, the micronuclear data contains 40,844 base pairs and the macronuclear data contains 32,770 base pairs, and also the simulation is run many times randomly choosing different repeats within each neighbourhood, and by trying all combinations of parameters.

Among these 13 input pairs, pair number 7 (the Actin I gene) has a smaller micronuclear sequence (989 bp) than its macronuclear sequence (1553 bp) due to incomplete data. This pair will indeed appear differently in the results. There is a very recent paper [3] on the sequencing and analysis of the micronuclear genome of the ciliate *Oxytricha trifallax*. However, as it is still in draft status, has not been used as further verification of the simulation.

For each input pair, the 15,680 different parameter combination are tested, each generating an average value for *acc1*, *acc2*, and *acc3* across all micronucleus and macronucleus gene pairs. The combination of three parameters that gives the maximum average *acc2* score is considered to be the optimal parameters. The reason the *acc2* accuracy value (after the *rearrangement* function) is used to define and to determine the optimal parameter values is because using the accuracy after the *IES_deletion* function (*acc1*) always gives low accuracies in the case of scrambled genes, as rearrangement has not yet occurred, and taking the accuracies after the *correction* function (*acc3*) often can fix otherwise bad alignments as the templates are used in this stage. Ideally, one would expect that for scrambled genes, *acc2* (used to determine optimal parameters) be "quite high" to account for cryptic pointers occurring in proximity of MDS-IES junctions, but not perfect as cryptic pointers do indeed occur (recall Definition 1 and preceding discussion on cryptic pointers, as sometimes IESs are eliminating around repeats nearby to the real pointers instead of the pointers themselves). Further, *acc3* should be almost perfect to account for proof checking from templates. Thus, using the accuracy after the *rearrangement* function seems to be the best way to calculate the optimal parameters and success of the simulation.

The maximum accuracy values using *acc2*, occurs when the parameters of *threshold_MDS* is 5, *threshold_IES* is 9, and *nh* is 15 (these are the values of the parameters for which the average *acc2* output is maximized across all data). For these optimal parameters, the simulation is run multiple times (at step 3 of Definition 1 was simulated *maxNum* times for each pointer selection) to calculate the average and standard deviation of the accuracies. Table 2 shows, on average over all gene pairs, the *acc1* value is 60.5%, *acc2* value is 88.5%, and *acc3* value is

99.5%. Removing gene pair number 7 (due to having incomplete data) increases the average *acc2* value to 91.1%. Indeed, this number is "quite high" but not perfect as was desirable.

Table 2. For each gene pair (indexed by the first column), the average (Avg) and standard deviation (STD) of *acc1*, *acc2*, and *acc3* are shown for the optimal parameters. The final two rows summarize the average over all 13 genes, and over the 12 genes without gene 7 that has incomplete data, respectively.

pair_no	Avg acc1	STD acc1	Avg acc2	STD acc2	Avg acc3	STD acc3
1	43	1.4	92.9	1.9	99.9	0.1
2	87.2	1.6	94.4	0.4	99.8	0.2
3	53.5	1	84.6	0.4	98.8	0.2
4	61.2	0.8	93.5	0.7	99.6	0.4
5	61.9	0.9	89.4	0.7	99.7	0.3
6	70.1	1	93.3	0.6	99.9	0.1
7	57.6	1.6	58.2	1.8	99.9	0.1
8	43.1	1.5	90.3	1.2	98.9	0.5
9	69.9	0.9	92.6	0.4	99.9	0.1
10	62.4	1	93.3	0.6	99.8	0.3
11	73.1	0.8	89.8	0.3	99.8	0.2
12	46.5	0.5	94.4	0.3	99.8	0.3
13	56.8	1.9	84.4	1.9	97.3	1.5
average	60.5	0.41	88.5	0.61	99.5	0.39
average-7	60.7	0.41	91.1	0.57	99.4	0.4

From the 2JLP model, it can be seen that the macronucleus is generated from the micronucleus in a successive manner. Table 2 shows that *acc2* is greater than *acc1* and *acc3* is greater than *acc2* for all input pairs, and the values of acc2 are quite high, but not perfect, which is exactly what we expect given the nature of cryptic pointers. And indeed, the values of acc3 are almost perfect as proof checking can add in missing, or remove excessive information.

Of interest, in Fig. 4, a scatter plot is shown that shows the relationship between *threshold_MDS* and average alignment scores between the descrambled and parental macronucleus. Here, an alignment score is calculated by dividing it by the size of *old_mac* and multiplying it by 100. Average alignment scores are calculated by the scores after the *rearrangement* function for all input pairs. In the same way, the average alignment scores are calculated for each value of *threshold_MDS* (from 1 to 28) where *threshold_IES* and neighbourhood (*nh*) are fixed with their optimal values. The scatter plot is generated by plotting *threshold_MDS* on the x-axis and average alignment scores on the y-axis. The trend-line equation is $y = -1.3576x + 76.105$ and the square of the correlation coefficient (R^2) value is 0.7435. The slope value is negative which indicates that there is a negative correlation present in between these two variables. If the value of *threshold_MDS* is increased it eventually degrades the value of the alignment

score. As the R^2 value is 0.7435, this indicates approximately 74% of the variation in accuracy can be explained by *threshold_MDS*.

Fig. 4 shows that a lower value of *threshold_MDS* is good from the perspective of maximizing the alignment score for the simulation. These lower scores for *threshold_MDS* occur when shorter pieces of scnRNAs are matched to the old macronucleus at the time of filtering, similar to IES specific sequences from the set of scnRNAs. Thus, if a scnRNA contains part of an MDS and part of an IES, from the perspective of maximizing the alignment score of the simulation, it is desirable to filter out this scnRNA at this stage. This is because if it does not get filtered out then the simulation may discard the matching portion of that scnRNA from the new micronucleus. This may result in an erroneous deletion of an MDS from the micronucleus.

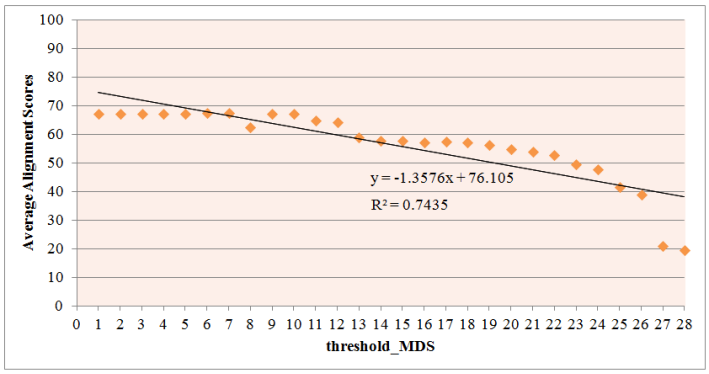

Fig. 4. Relationship between *threshold_MDS* and average alignment scores (*acc2*, after the *rearrangement* function, for all 13 input pairs). Here, *threshold_IES* and *nh* are fixed with the optimal values of *threshold_IES* as 9 and *nh* as 15.

5 Future Directions

Currently, the simulator has been tested only for thirteen pairs of real genes. After the assembly of the micronuclear genome emerges from draft status, a more extensive analysis will be possible. As mentioned is Section 3, a new simulation to capture piRNAs is desirable for *Oxytricha*. Furthermore, a simulation using scnRNAs as done in this paper using *Tetrahymena* data would help to validate the hypothesized model.

References

[1] Bracht, J.R., Fang, W., Goldman, A.D., Dolzhenko, E., Stein, E.M., Landweber, L.F.: Genomes on the edge: Programmed genome instability

in ciliates. Cell 152(3), 406–416 (2013)
[2] Cavalcanti, A.: Ciliate nuclear dimorphism pages. http://oxytricha.princeton.edu/dimorphism/ (2004)
[3] Chen, X., Bracht, J.R., Goldman, A.D., Dolzhenko, E., Clay, D., Swart, E., Perlman, D., Doak, T., Stuart, A., Amemiya, C., Sebra, R., Landweber, L.: The architecture of a scrambled genome reveals massive levels of genomic rearrangement during development. Cell 158(5), 1187–1198 (2014)
[4] Ehrenfeucht, A., Prescott, D.M., Rozenberg, G.: Computational aspects of gene (un)scrambling in ciliates. In: Evolution as Computation, pp. 216–256. Natural Computing Series, Springer Berlin Heidelberg (2002)
[5] Jönsson, F., Postberg, J., Lipps, H.J.: The unusual way to make a genetically active nucleus. DNA Cell Biol. 28(2), 71–8 (2009)
[6] Keil, J.M., Liu, J., McQuillan, I.: Algorithmic properties of ciliate sequence alignment. Theoretical Computer Science 411(6), 919–925 (2010)
[7] Landweber, L.F., Kari, L.: The evolution of cellular computing: nature's solution to a computational problem. Biosystems 52, 3–13 (1999)
[8] Landweber, L.F., Kuo, T.C., Curtis, E.A.: Evolution and assembly of an extremely scrambled gene. Proc Natl Acad Sci USA 97(7), 3298–3303 (2000)
[9] Mochizuki, K., Fine, N.A., Fujisawa, T., Gorovsky, M.A.: Analysis of a piwi-related gene implicates small RNAs in genome rearrangement in Tetrahymena. Cell 110(6), 689–699 (2002)
[10] Möllenbeck, M., Zhou, Y., Cavalcanti, A.R.O., Jönsson, F., Higgins, B.P., Chang, W.J., Juranek, S., Doak, T.G., Rozenberg, G., Lipps, H.J., Landweber, L.F.: The pathway to detangle a scrambled gene. PLoS ONE 3(6), e2330 (2008)
[11] Needleman, S.B., Wunsch, C.D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology 48(3), 443–453 (1970)
[12] Nowacki, M., Vijayan, V., Zhou, Y., Schotanus, K., Doak, T.G., Landweber, L.F.: RNA-mediated epigenetic programming of a genome-rearrangement pathway. Nature 451, 153–158 (2008)
[13] Prescott, D.M.: The unusual organization and processing of genomic DNA in hypotrichous ciliates. Trends in Genetics 8(12), 439–445 (1992)
[14] Prescott, D.M.: The DNA of ciliated protozoa. Microbiol. Rev. 58(2), 233–267 (1994)
[15] Prescott, D.M.: Genome gymnastics: unique modes of DNA evolution and processing in ciliates. Nature reviews Genetics 1(3), 191–198 (2000)
[16] Prescott, D.M., Ehrenfeucht, A., Rozenberg, G.: Molecular operations for DNA processing in hypotrichous ciliates. European Journal of Protistology 37(3), 241–260 (2001)
[17] Prescott, D.M., Ehrenfeucht, A., Rozenberg, G.: Template-guided recombination for IES elimination and unscrambling of genes in stichotrichous ciliates. Journal of Theoretical Biology 222(3), 323–330 (2003)
[18] Verlan, S., Alhazov, A., Petre, I.: A sequence-based analysis of the pointer distribution of stichotrichous ciliates. Biosystems 101(2), 109–116 (2010)