Complex Rearrangements in the Highly Scrambled Genome of *O. trifallax*

Lukas Nabergall, Nataša Jonoska, Masahico Saito
University of South Florida

DNA Recombination

Genome rearrangement processes are observed in many species, on both evolutionary and developmental scale. *Oxytricha trifallax*, a species of ciliate, undergoes massive genome rearrangements during the development of a somatic macronucleus (MAC) from a germline micronucleus (MIC) and is used as a model organism to study DNA rearrangements [1] (see Fig. 1).

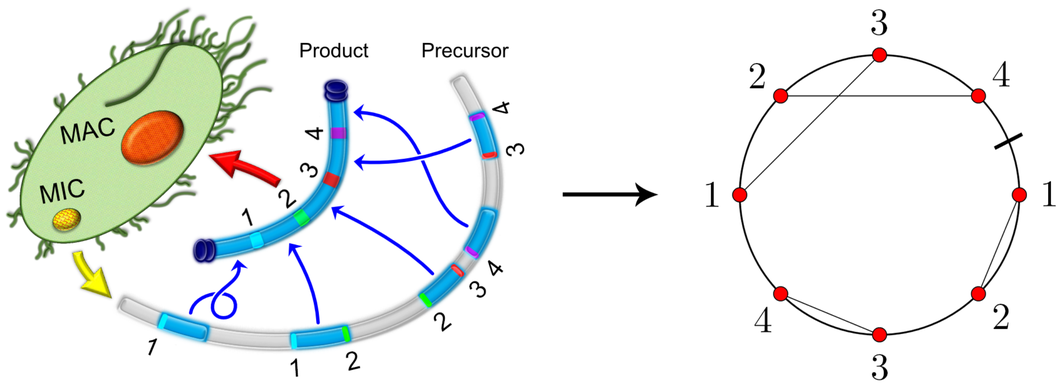

Figure 1: Example of DNA recombination in *O. trifallax* and the corresponding chord diagram schematically representing a micronuclear locus, with pointer list representation 11223434 and repeat words 11, 22, and 3434.

During the copying of DNA from the MIC to the MAC, fragments, known as **macronuclear destined sequences** (MDSs), are often rearranged and inverted. Short sequences of DNA, called **pointers**, located on either side of the MDSs determine the rearrangement by overlapping with their unique copy in the MAC (see Fig. 1).

Mathematical Model

Sequentially labeling the resulting pointers in the MAC determines a labeling of the pointers in the MIC. Symbolically, we represent these pointer lists by **double occurrence words** (DOWs), words where every letter appears twice. Two commonly occurring subsequences in the scrambled genome of *O. trifallax* are the *repeat words* and *return words* (see Figures 1 and 2):

• The **repeat word pattern** is a word of the form

$$a_1 a_2 \cdots a_n \cdots a_1 a_2 \cdots a_n \quad \text{(e.g. } 12341234\text{)},$$

while the **return word pattern** is a word of the form

$$a_1 a_2 \cdots a_n \cdots a_n a_{n-1} \cdots a_1 \quad \text{(e.g. } 123456654321\text{)}.$$

Double occurrence words can be visually represented via **chord diagrams**, where each letter corresponds to the end points of a chord on the outer circle (see Fig. 2).

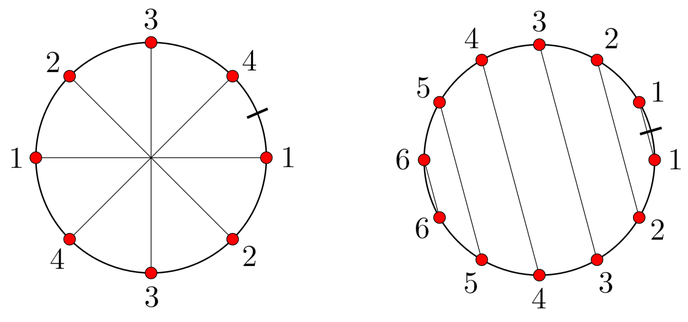

Figure 2: The chord diagrams associated with the repeat word 12341234 (left) and the return word 123456654321 (right).

Repeat and return words can appear nested within one another in the scrambled genome. For example, the word 121342566534 has the return word 5665 nested within the repeat word 3434.

Measuring Rearrangement Complexity

To measure the prevelance of repeat and return words as rearrangement patterns in *O. trifallax*, we define *reductions* on words [2]:

• A **reduction** of a word w is a sequence of words $(u_0, u_1, \ldots, u_n)$ in which $u_0 = w$ and u_{k+1} is obtained from u_k by removing all repeat and return subwords.

We also apply the concept of a reduction to search for other patterns. First, we must define the notion of a pattern.

• A **pattern** is a sequence of words whose symbols are variables. For example, the **tangled cord** T is defined recursively by setting $T_1 = 11$, $T_2 = 1212$, $T_3 = 121323$, and so on (see Fig. 3). The repeat word and return word are also patterns.

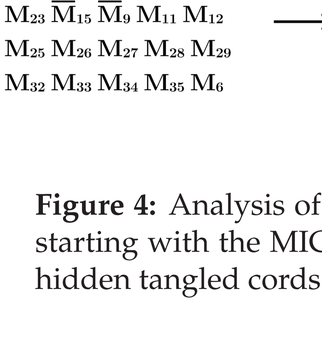

Figure 3: The chord diagram associated with the tangled cord 121324354656.

To measure how nested a set of patterns appears in a given word, we define so-called *pattern indices*:

• Given a set of patterns

$$P = \{p_1, p_2, \ldots, p_k\},$$

define the **pattern index** of a word w by $I_P(w) := \min\{n \mid (w_0, w_1, \ldots, w_n = \epsilon) \text{ is a reduction of } w\}$, where a reduction now involves removing a single instance of a pattern from P per step and ϵ is the empty word.

• For example, the **pattern recurrence index** (PRI) of a word w is the pattern index of w with $P = \{\text{repeat word}, \text{return word}\}$. Similarly, the **tangled index** (TI) of a word w is defined by setting $P = \{\text{tangled cord}, \text{letter}\}$.

Results

Out of 2021 scrambled sequences studied in *O. trifallax*, 1948 reduced to the empty word, implying that they are compositions of nested repeat and returns words [3][4]. More notably, 22 DOWs were identified which retained at least 4 letters after iterative removal of all repeats and returns (see Fig. 4).

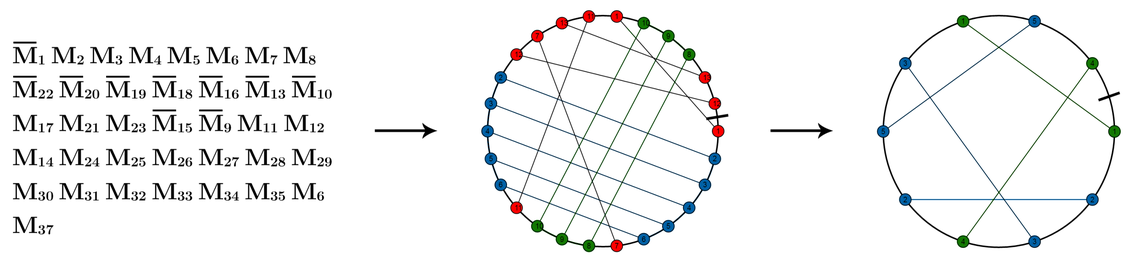

Figure 4: Analysis of *O. trifallax* contig OXYTRI_MAC_9447 [1], starting with the MIC arrangement. Two repeat words and two hidden tangled cords are highlighted in green and blue.

These 22 highly scrambled genes contain hidden tangled cord patterns:

	Average		
	PRI	PRI + T	TI
Highly scrambled cases	3.91	3.59	3.91
Random sample	3.50	3.29	4.36

Table 1: Compared with an identically distributed random sample of 22 DOWs, the 22 highly scrambled cases exhibit significantly lower averages on indices that include the tangled cord pattern. PRI + T is the PRI with tangled cord removals allowed.

References and Acknowledgements

[1] Burns J, Kukushkin D, Lindblad K, Chen X, Jonoska N, Landweber LF, <mds_ies_db>: a database of ciliate genome rearrangements. *Nucleic Acids Res*. **44** (2016) (Database issue) doi: gkv1190.

[2] Arredondo R, Reductions On Double Occurrence Words. Proceedings of the 44th Southeastern International Conference on Combinatorics, Graph Theory and Computing. *Congr. Numer.* **218** (2013) 43-56.

[3] Chen X, et al. The Architecture of a Scrambled Genome Reveals Massive Levels of Genomic Rearrangement during Development. *Cell*, **158**:5 (2014) 1187-1198.

[4] Burns J, et al. Reoccurring patterns of scrambled genes in the encrypted genome of the ciliate *Oxytricha trifallax*. *Journal of Theoretical Biology*, in revision.

Research supported by: NIH R01GM109459.

Special thanks to the other members of the USF Math-Bio Lab.