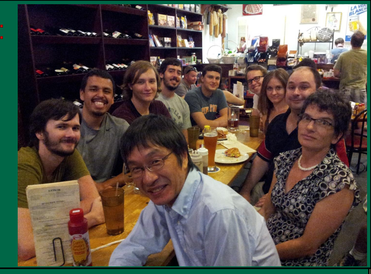
Patterns in scrambled chromosomes of ciliates

Denys Kukushkin, Department of Mathematics and Statistics

Current Lab Members:

- Jonathan Burns
- Daria Karpenko
- Denys Kukushkin
- Nataša Jonoska
- Masahico Saito
- Rick Wallace
- Micah Wine



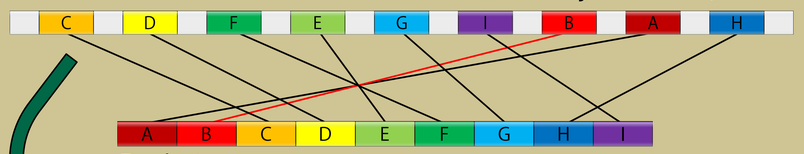
Abstract

Taking ciliates as model organisms, we study homologous DNA rearrangement processes. We use graphs and abstract words to capture patterns in thousands of scrambled genes of a recently sequenced genome. We observe common patterns that can explain complexities of all scrambled genes.

Methodology

- Precursor contains gene segments in a scrambled order. Product has gene segments arranged in the right order with non-coding DNA removed

Precursor — *Oxytricha nova*, Actin I[2]

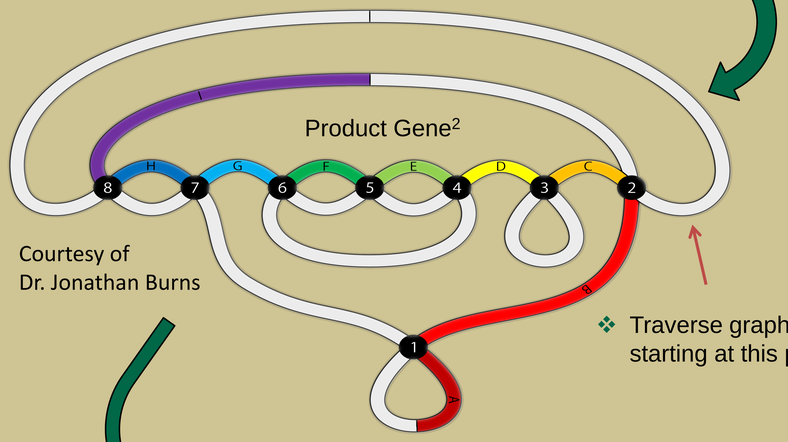


Product

- Mark repetitive sequences at the ends of each gene segment with letters

Precursor gene

- Construct graph by connecting similar letters



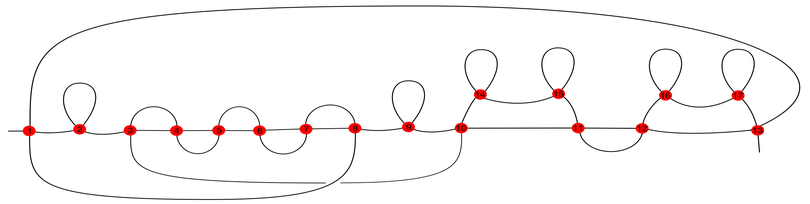
Courtesy of Dr. Jonathan Burns

- Traverse graph starting at this point
- Write down the abstract word

2 3 3 4 5 6 4 5 6 7 8 2 1 1 7 8

Recombination patterns in graphs and abstract words

- Given a precursor gene we build a corresponding graph and abstract word. For example, consider Contig20991.0.0[3] of *Oxytricha trifallax*

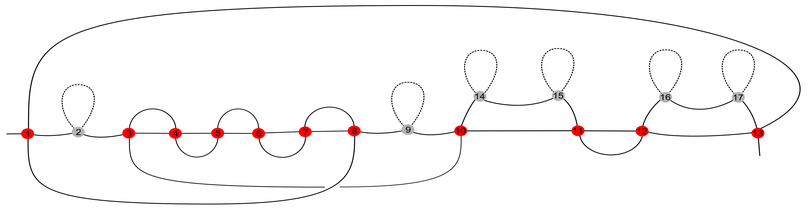


1 2 2 3 4 5 6 7 8 9 9 10 11 12 13 1 8 7 6 5 4 3 10 14 14 15 15 11 12 16 16 17 17 13

- Graph loops correspond to non-coding DNA regions in between consecutive gene segments. We mark these regions

Step 1

- Such DNA regions are removed first during rearrangement process[4]. We remove graph loops

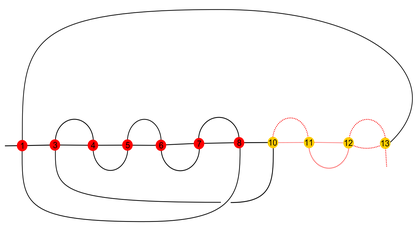


1 2 2 3 4 5 6 7 8 9 9 10 11 12 13 1 8 7 6 5 4 3 10 14 14 15 15 11 12 16 16 17 17 13

- We then look for patterns and continue graph reduction

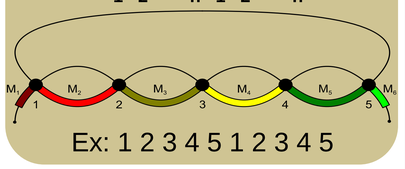
Step 2

- We identify longest sub-repeat word pattern and remove it



Repeat word pattern[5]

$a_1a_2\ldots a_na_1a_2\ldots a_n$

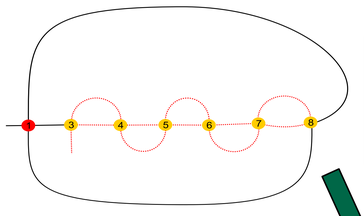


Ex: 1 2 3 4 5 1 2 3 4 5

1 3 4 5 6 7 8 10 11 12 13 1 8 7 6 5 4 3 10 11 12 13

Step 3

- We identify and remove longest sub-return word pattern

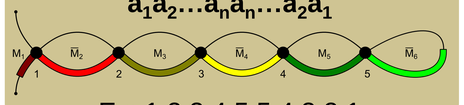


1 3 4 5 6 7 8 1 8 7 6 5 4 3

- We got reduced graph

Return word pattern[5]

$a_1a_2\ldots a_na_n\ldots a_2a_1$



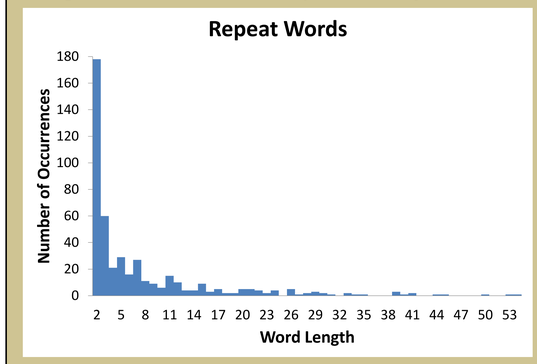
Ex: 1 2 3 4 5 5 4 3 2 1

1 1

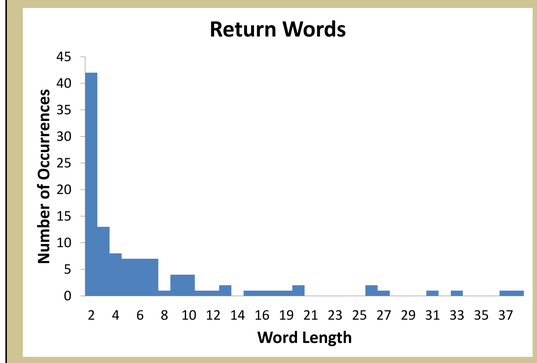
- Note: in general reduced graph may still be complex

Results and Conclusions

- The outlined process is used to estimate the complexity of the scrambled genes in the sequencing data of *Oxytricha trifallax* obtained by Chen, et al.[3] by analyzing recombination graphs and abstract words of 15811 genes
- Step 1 showed that 13084 genes correspond to "all loops" graphs. Hence, these genes do not contain any types of scrambling. As a result, 1893 genes are left after Step 1 to analyze



- Step 2 showed that 464 genes correspond to the repeat word pattern. The histogram on the left depicts the distribution of repeat words compared to the word length
- After Step 2, we are left with 381 genes to analyze



- Step 3 showed that 111 genes correspond to the return word pattern. The histogram on the left depicts the distribution of return words
- After Steps 1,2, and 3 only 215 remained to analyze for further complex recombination patterns

References

1. A. Angeleska, N. Jonoska. M. Saito, DNA Rearrangement through assembly graphs, *Discrete and Applied Math*, 157:14, pp. 3020-3037 (2009)
2. D. M. Prescott, A. F. Greslin, Scrambled actin I gene in the micronucleus of Oxytricha nova, *Developmental Genetics*, 13:1, pp. 66-74 (1992)
3. X. Chen et al., The Architecture of a Scrambled Genome Reveals Massive Levels of Genomic Rearrangement during Development, *Cell* , 158:5, pp. 1187 – 1198 (2014)
4. M. Möllenbeck et al., The Pathway to Detangle a Scrambled Gene, *PLoS ONE*, 3:6 (2008)
5. R. Arredondo, Reductions On Double Occurrence Words, Proceedings of the Forty-fourth Southeastern International Conference on Combinatorics, *Graph Theory and Computing*, Congr. Numer. 218, pp. 43-56 (2013)

Supported by: NSF CCF-1117254
NSF DMS-0900671
NIH R01GM109459-01

Web Resources:

USF Math-Bio Homepage (*http://knot.math.usf.edu*)

The Landweber Lab (http://www.princeton.edu/~lfl/)