An Annotation Algorithm for DNA Rearrangements

Jasper Braun[1], Nataša Jonoska[1], Laura F. Landweber[2], Rafik Neme[2]

[1] University of South Florida: Department of Mathematics and Statistics

[2] Columbia University: Department of Biochemistry and Molecular Biophysics, and Biological Sciences

This work has been supported by NIH grant No. R01 GM109459 and NSF grants No. 1800443/1764366.

Introduction

Complex genome rearrangements are observed in many organisms and most notably during the mating process of some species of ciliates, such as *Oxytricha trifallax*. During conjugation, the ***precursor*** germline micronucleus is reorganized to form the ***product*** genome of the somatic macronucleus.

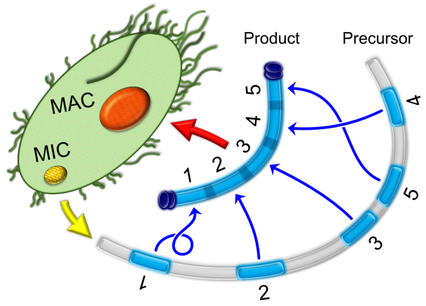


Figure 1: Segments from the micronucleus reorganize through deletion, inversion and reordering to form the macronuclear genome in *O. trifallax*.

Our algorithm aligns precursor and product sequences, and determines whether or not the segments in the precursor are scrambled.

Definitions

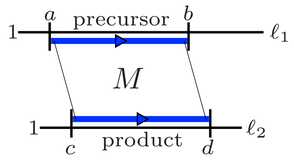


Figure 2: A ***match***

$$M = (\mathrm{Prec}(M), \mathrm{Prod}(M), \sigma(M)) = ([a,b],[c,d],\text{orientation})$$

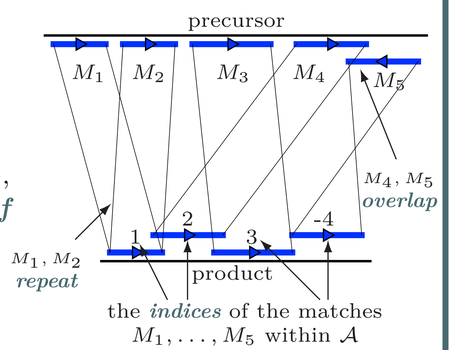


Figure 3: An ***arrangement*** $\mathcal{A} = \{M_1, M_2, M_3, M_4, M_5\}$, the ***multiset of indices of*** $\mathcal{A}$ is $I(\mathcal{A}) = \{1, 1, 3, 2, -4\}$.

An arrangement $\mathcal{A}$ can be represented by a ***precursor representation***, which consists of its precursor segments labelled with $I(\mathcal{A})$.

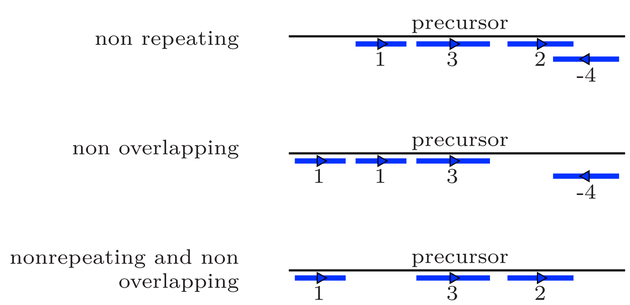
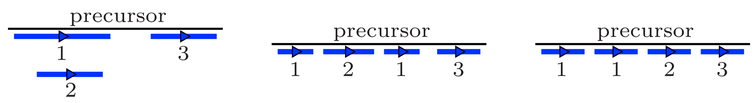


Figure 4: ***Subarrangements*** of the arrangement $\mathcal{A}$ in Figure 3.

Arrangement Properties

We extend the procedure in [1] and [2] to consider the full complexity of the types of arrangements that can appear.

Figure 5: Arrangements for which scrambling is not easily defined.



We consider three basic properties of arrangements:

Figure 6: ***Ordered*** (Indices in the precursor representation are increasing and have same sign)

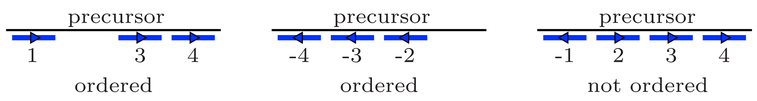


Figure 7: ***Consecutive*** (consecutive portion of product segments matched)

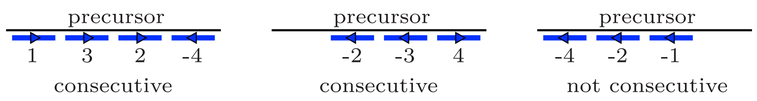
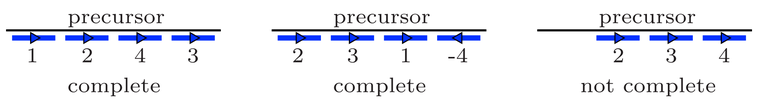


Figure 8: ***Complete*** (all product segments matched)



User specifies which combination of these properties defines a maximal non repeating and non overlapping arrangement to be ***non scrambled***. Algorithm looks at all such subarrangements of an arrangement:

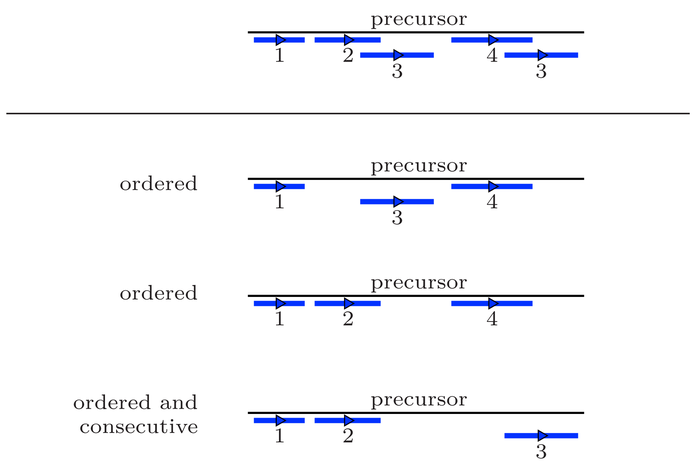


Figure 9: An arrangement and its maximal non repeating and non overlapping subarrangements. If user specifies ordered and consecutive to define non scrambled, this arrangement is ***weakly non scrambled***, if the user only requires ordered to define non scrambled, it is ***strongly non scrambled***

Computation of Arrangements

Let $\mathcal{H}_0$ be the of set high-scoring pairs between a precursor and a product, each having length, bitscore and percent identity above user-defined thresholds. Assume $\mathcal{H}_0$ is sorted by bitscore and percent identity. The procedure is described by the flowchart in Figure 10.

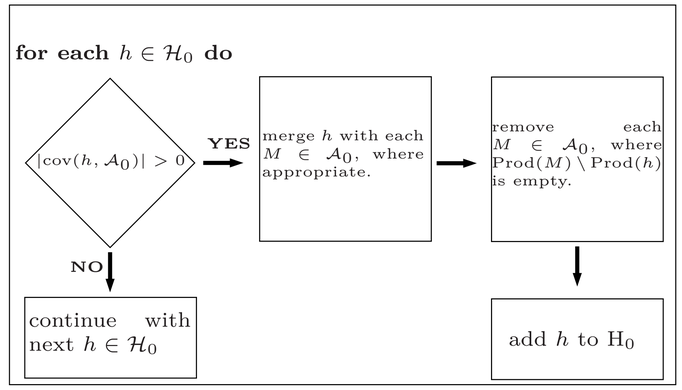


Figure 10: $\mathrm{cov}(h, \mathcal{A}_0)$ refers to the regions $\mathrm{Prod}(h)$ covers, which are not already covered by $\mathcal{A}_0$.

Data

Table 1 was obtained by applying the algorithm to the 25,720 precursor and 22,450 product contigs of *O. trifallax* from [3] and [4].

non scrambled	ordered	ordered and consecutive	ordered and complete
arrangements	89,272	89,272	89,272
non repeating	81,783	81,783	81,783
non overlapping	64,247	64,247	64,247
weakly non scrambled	72,342	64,224	57,154
strongly non scrambled	61,217	58,492	54,738

Table 1: Data

References

[1] Burns et al. <mds_ies_db>: a database of ciliate genome rearrangements. *Nucleic Acids Research*, 44(D1):D703–D709, November 2015. URL: https://doi.org/10.1093/nar/gkv1190.

[2] Burns et al. Recurring patterns among scrambled genes in the encrypted genome of the ciliate *oxytricha trifallax*. *Journal of Theoretical Biology*, 410:171–180, December 2016. URL: https://doi.org/10.1016/j.jtbi.2016.08.038.

[3] Chen et al. The architecture of a scrambled genome reveals massive levels of genomic rearrangement during development. *Cell*, 158:1187–1198, August 2014. URL: https://doi.org/10.1016/j.cell.2014.07.034.

[4] Swart et al. The *oxytricha trifallax* macronuclear genome: A complex eukaryotic genome with 16,000 tiny chromosomes. *PLOS Biology*, 11, January 2013. URL: https://doi.org/10.1371/journal.pbio.1001473.