An Annotation Algorithm for DNA Rearrangements

Jasper Braun1, Natasa Jonoska1, Laura F. Landweber2, Rafik Neme2

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. 1804433/1764366.

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

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

An arrangement \( A \) can be represented by a \textit{precursor representation}, which consists of its precursor segments labelled with \( I(A) = \{1, 1, 3, 2, -4\} \).

An arrangement \( A \) is \textit{ordered} if \( M \) is a \textit{maximal non repeating and non overlapping arrangement} to be \textit{non scrambled}. Algorithm looks at all such subarrangements of an arrangement:

Figure 4: Subarrangements of the arrangement \( A \) 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.

\[
\text{ordered} \quad \text{ordered} \quad \text{not ordered}
\]

We consider three basic properties of arrangements:

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

\[
\text{ordered} \quad \text{ordered} \quad \text{not ordered}
\]

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

\[
\text{consecutive} \quad \text{consecutive} \quad \text{not consecutive}
\]

User specifies which combination of these properties defines a maximal non repeating and non overlapping arrangement to be \textit{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 \textit{weakly non scrambled}, if the user only requires ordered to define non scrambled, it is \textit{strongly non scrambled}.

Figure 8: Complete (all product segments matched)

\[
\text{complete} \quad \text{complete} \quad \text{not complete}
\]

Computation of Arrangements

Let \( \mathcal{H}_0 \) be the set of 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.

Table 1: Data

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].

References


