Sequence annotation algorithm for identifying genetic building blocks

Jonathan Burns, Denys Kukushkin*, and Nataša Jonoska

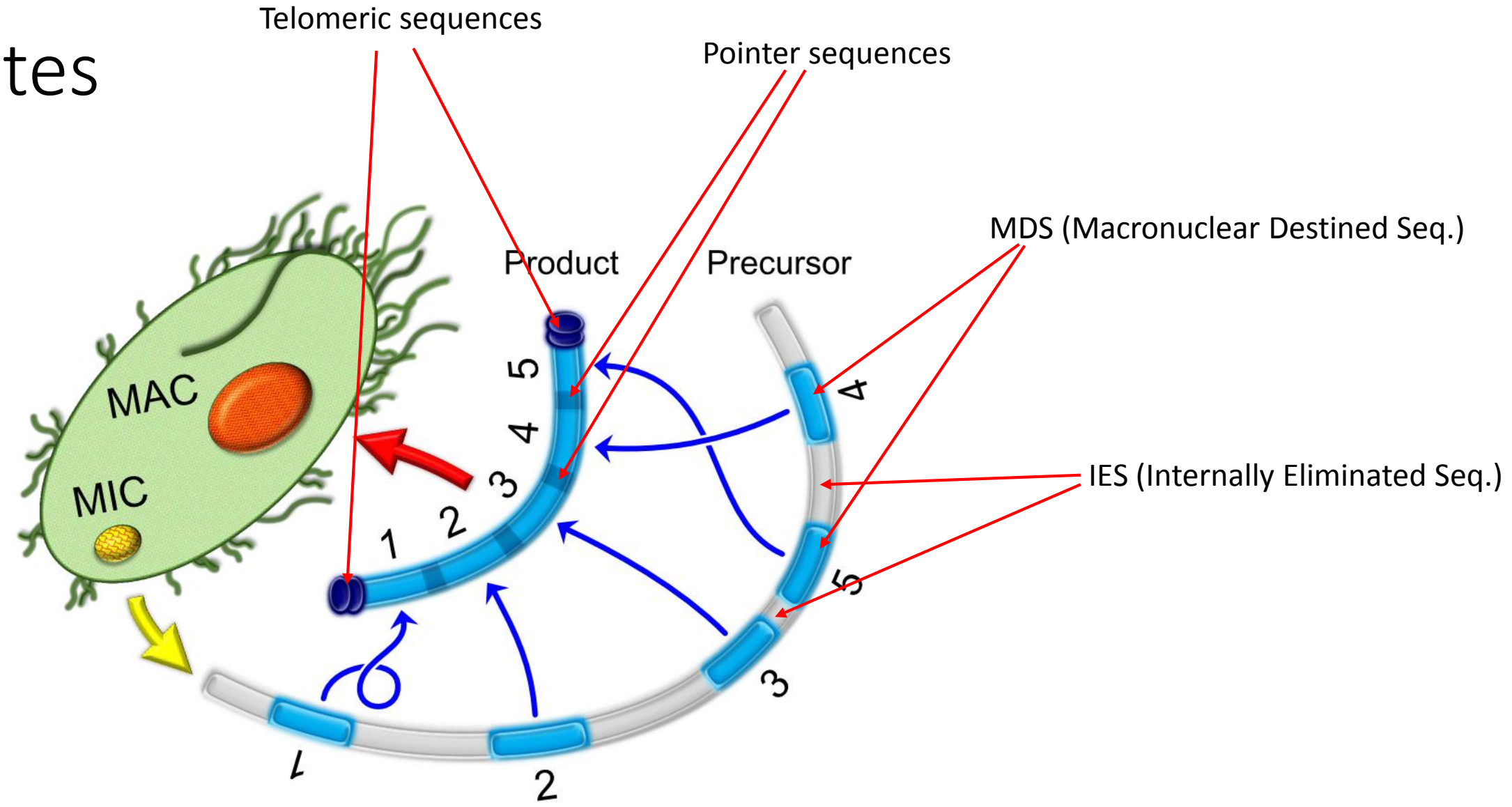
Department of Mathematics and Statistics

University of South Florida

Outline

- Genetic building blocks in ciliates
- Algorithm for identifying genome building blocks
- Algorithm Implementation Results

Ciliates


Telomeric sequences
Pointer sequences
MDS (Macronuclear Destined Seq.)
IES (Internally Eliminated Seq.)
MAC
MIC
Product
Precursor
1
2
3
4
5

Algorithm: Input/Output

- Input
  - ❖MAC nucleotide sequences
  - ❖MIC nucleotide sequences
  - ❖Regular expression for the macronuclear telomeric sequence

- Output
  - ❖MDS annotation of MAC
  - ❖IES annotation of MIC

Algorithm: Main Steps

1. Annotate telomeric sequences and exclude them from the further steps
2. Run BLAST and identify matching sequences between the precursor (MIC) and the product (MAC)
3. Process obtained sequences resolving overlaps
4. Run BLAST again to search for smaller matching sequences between the precursor and the product
5. Process obtained sequences similarly to step 3
6. Output final MDS/IES annotation

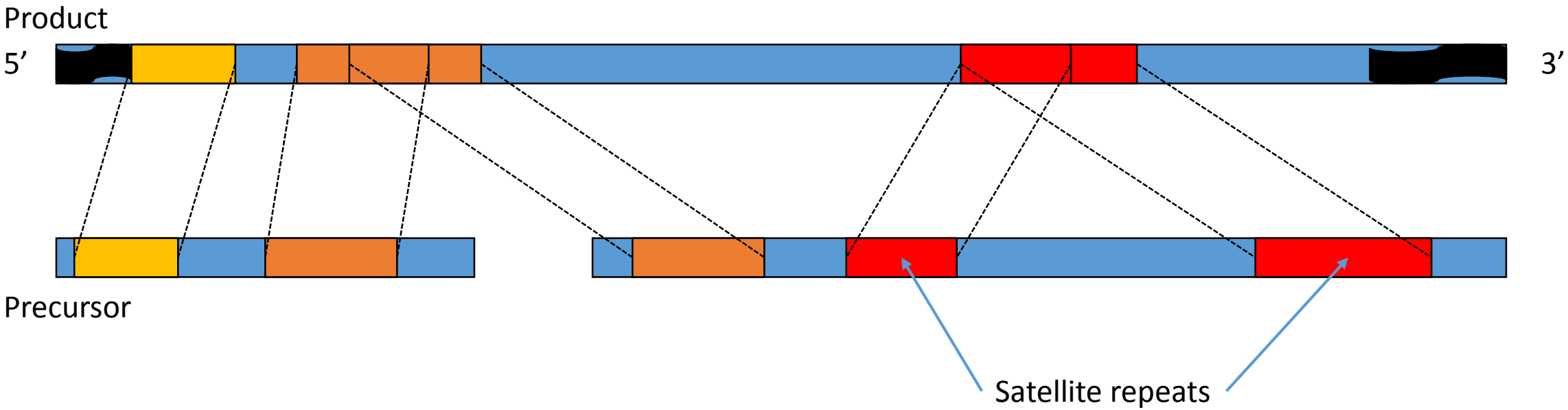
Algorithm: Step 1

- Annotating telomeric sequences

- For example (using regular expression):
  - C{0,4} (AAAACCCC)+ A{0,4} – *Oxytricha trifallax*
  - (TT){0,1} (GGGGTT)+ (GGGG){0,1} – *Tetrahymena thermophila*

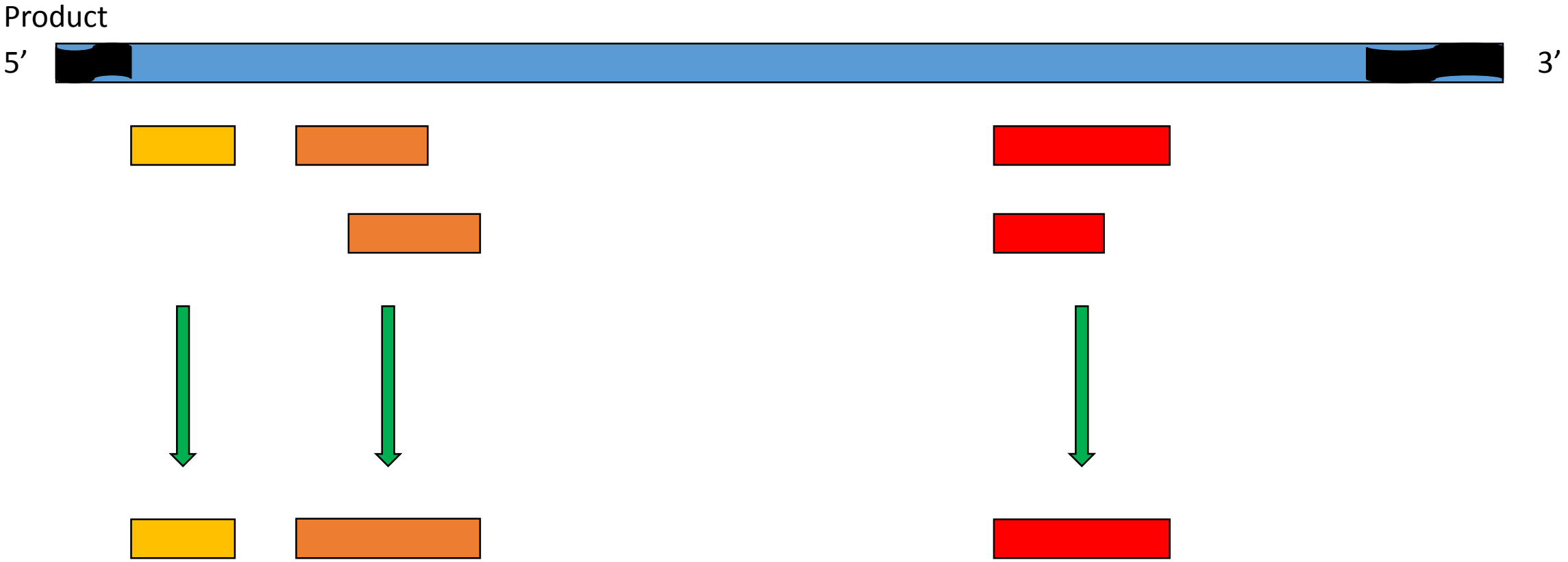
Algorithm: Step 2

- Get high scoring pairs (HSPs) using BLAST
  -task: megablast, -word_size: 28, -dust: no, -ungapped



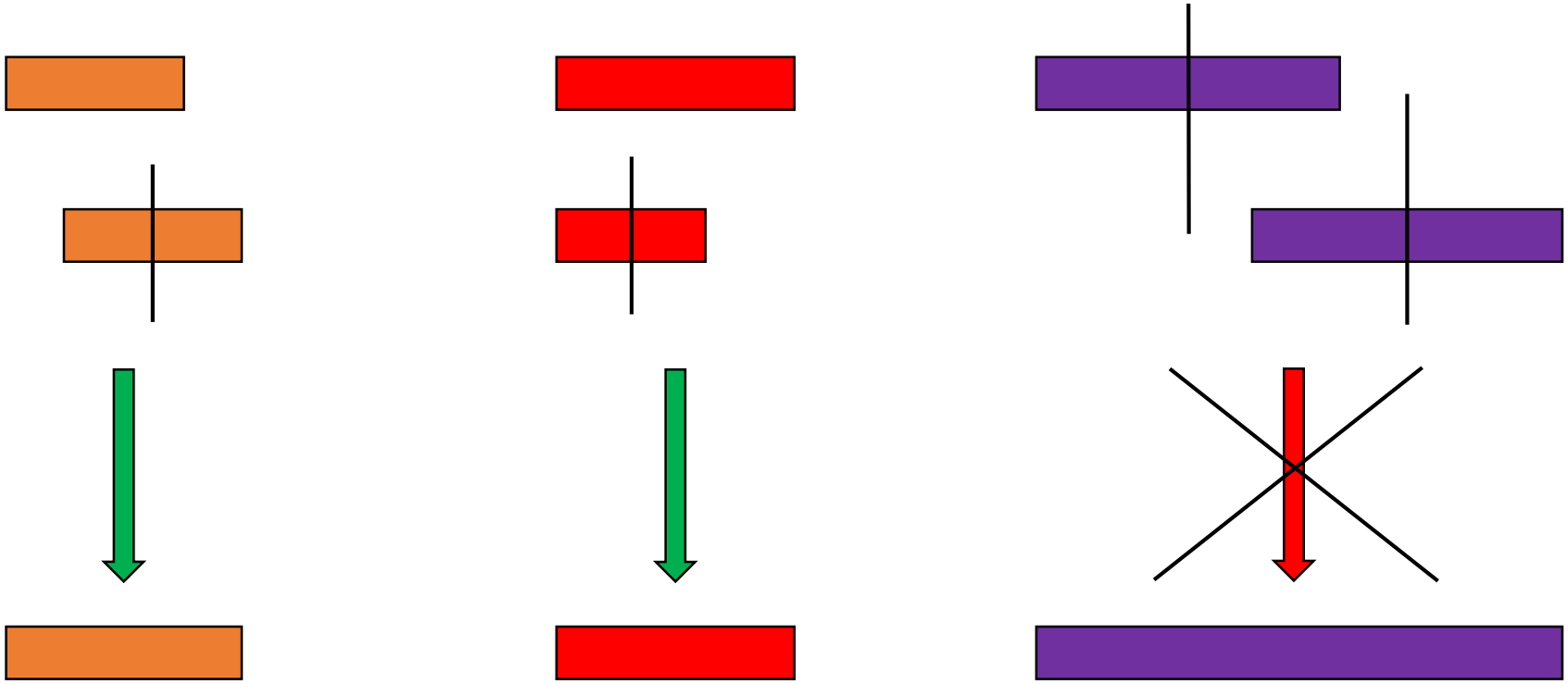
Algorithm: Step 3

- Check for overlaps and consider merging overlapping HSPs



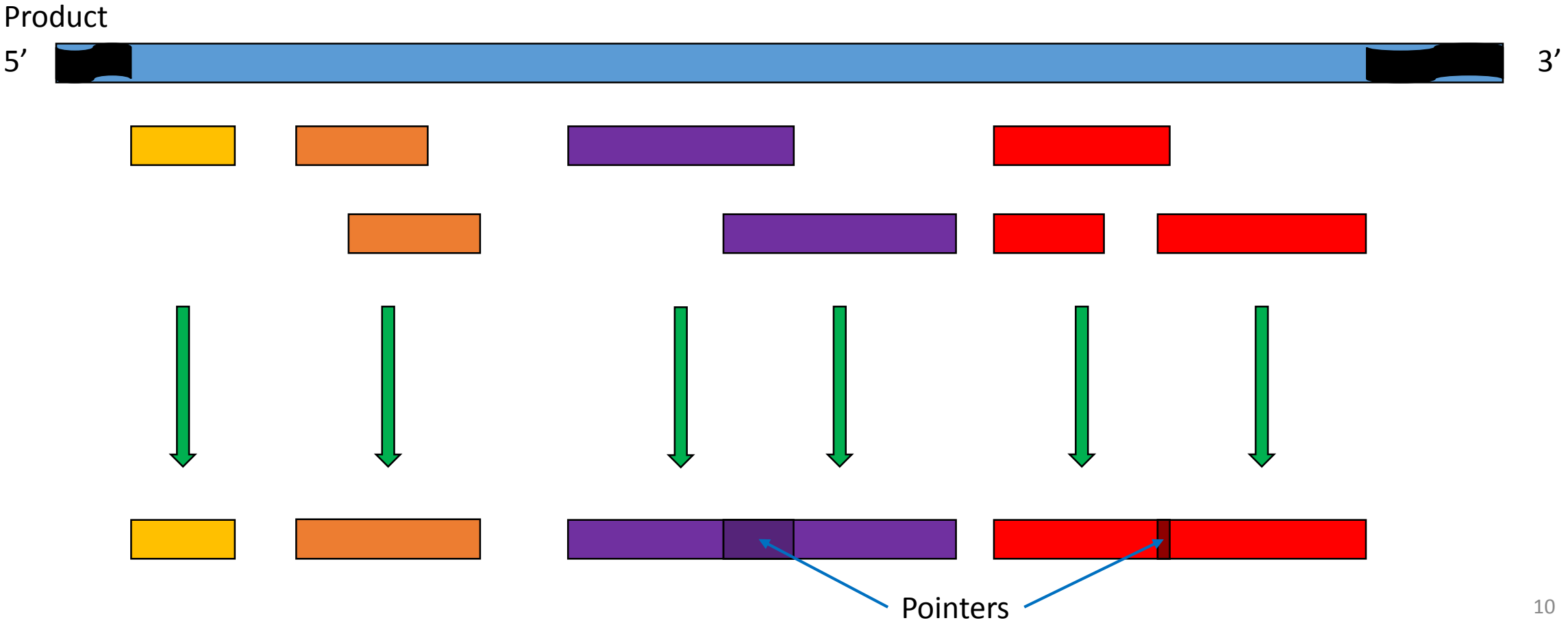
Algorithm: Step 3 – Merging criteria

- If one HSP overlaps with the other at least 50%, then merge them



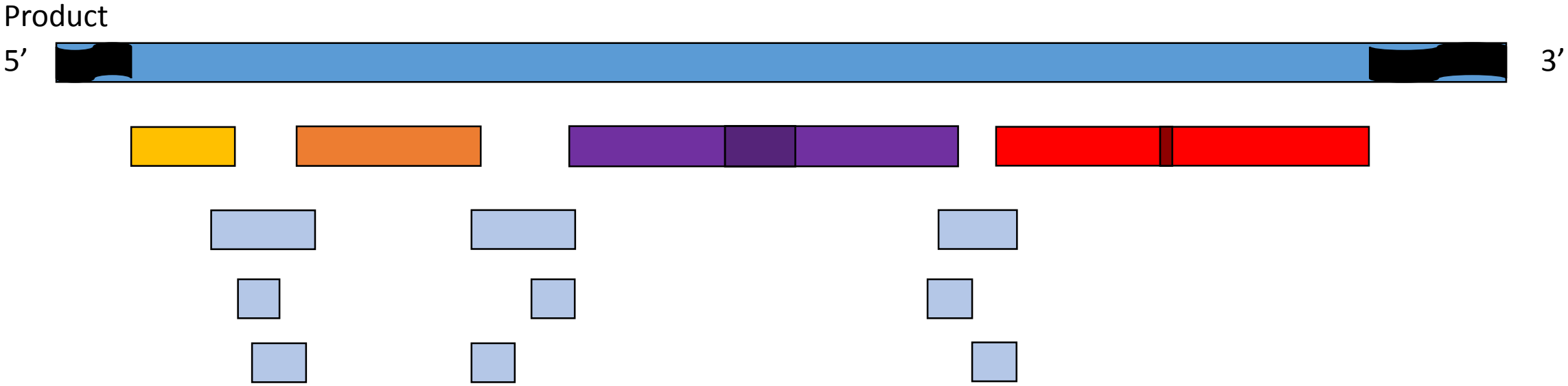
Algorithm: Step 3 - Continued

- The resulting HSPs are considered to be initial MDS annotation



Algorithm: Step 4

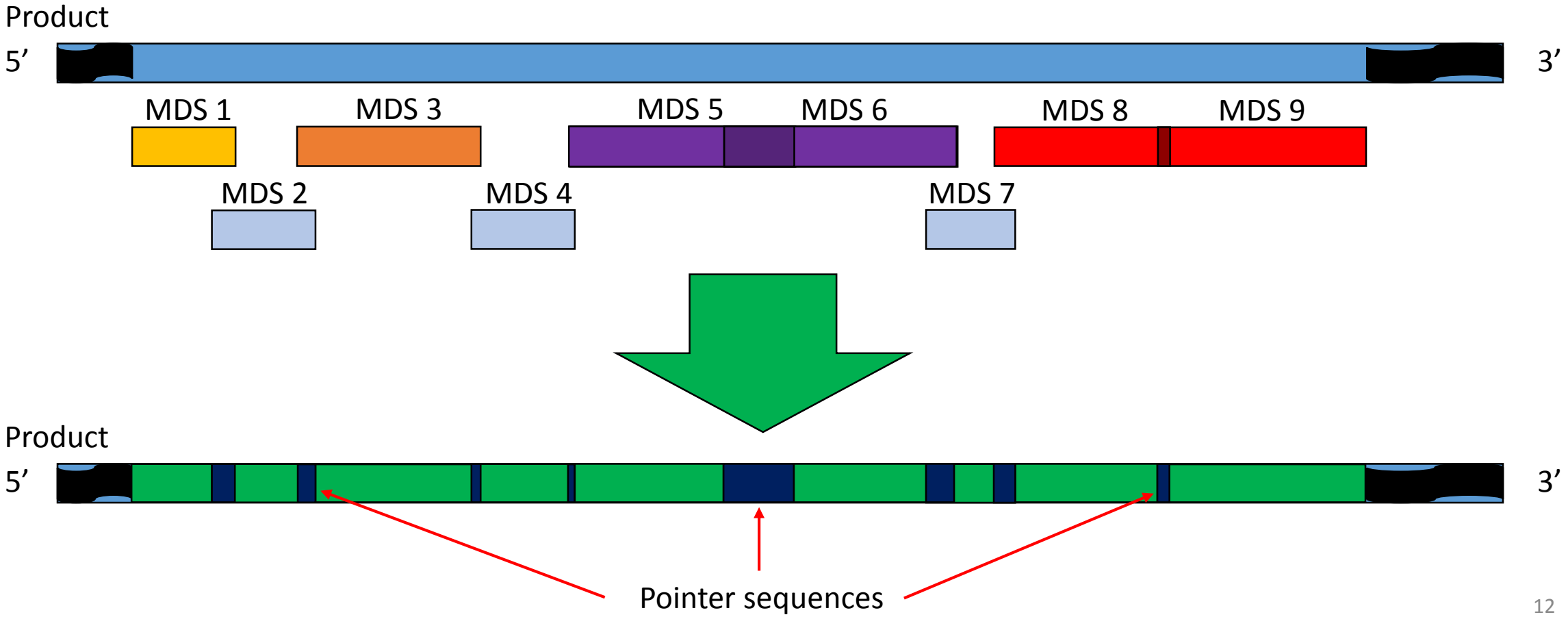
- Get smaller high scoring pairs (HSPs) using BLAST to fill the gaps
  -task: blastn-short, -word_size: 12, -dust: no, -ungapped



- Since HSPs are small, we can get a lot of small “noisy” pieces

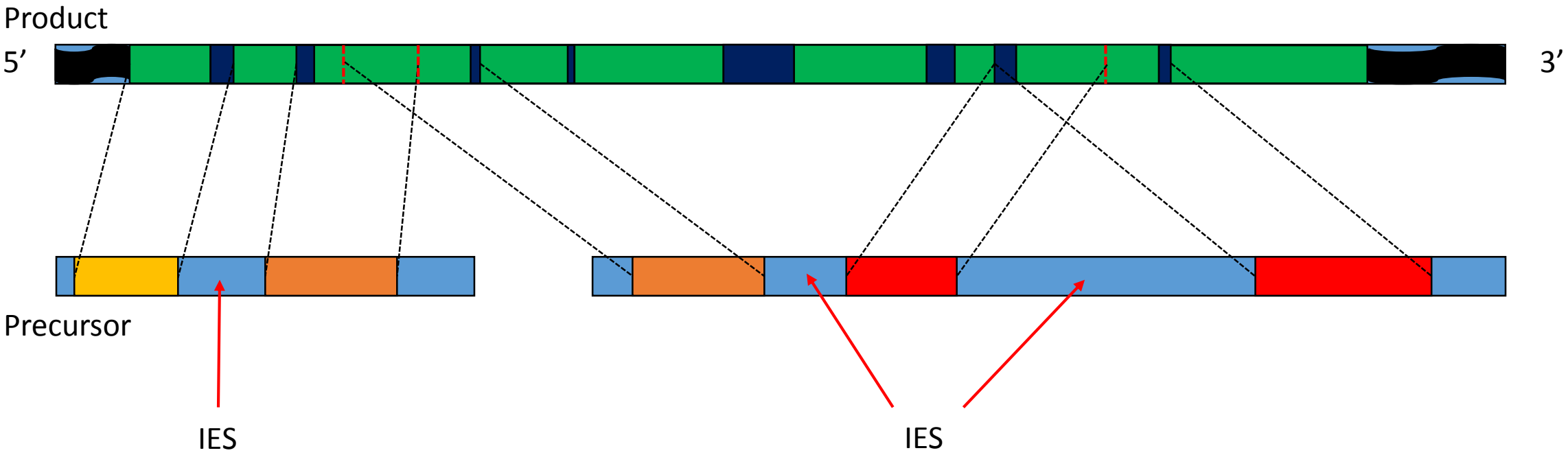
Algorithm: Step 5

- Using the same merging criteria, process HSPs to get final MDS annotation



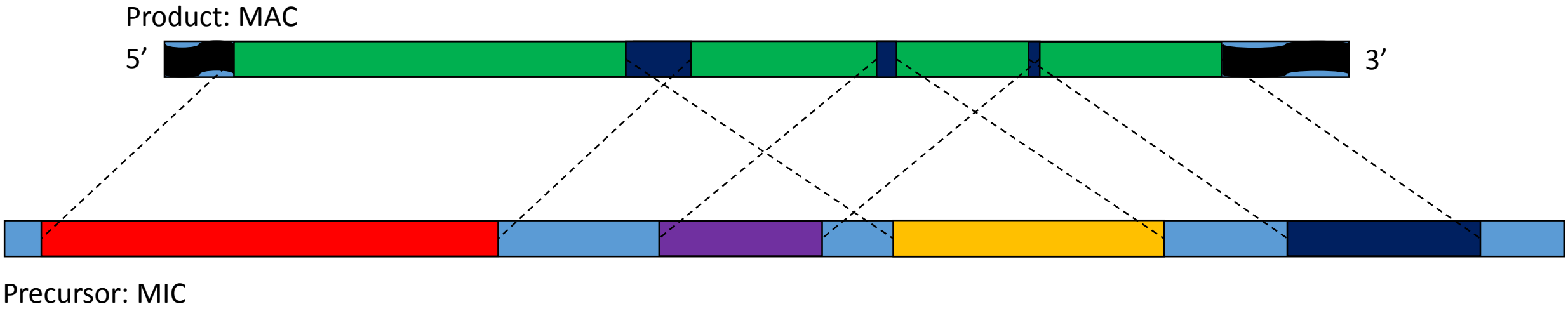
Algorithm: Step 6

- IESs correspond to the nucleotide sequences that were not used in the annotation process and occur in between HSPs in Precursor DNA



Results: *Oxytricha trifallax*

- Processed 22,460 MAC sequences and 25,720 MIC sequences
- Identified 5,464 MAC sequences that are fully covered by at least one MIC sequence



Results: *Oxytricha trifallax*

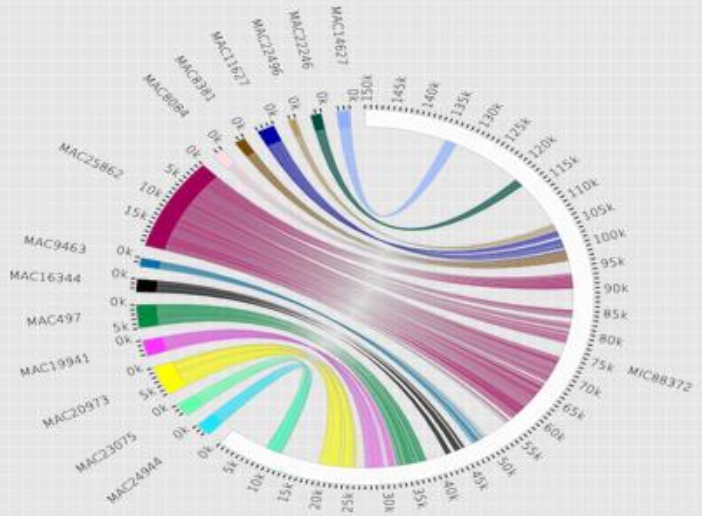
- Out of these 5,464, the program identified 1,538 to be scrambled in the MIC

- Identified 278,656 MDSs (previously reported ≥225,000 in Chen, et al., 2014)
  - More MDSs due to using small HSPs in the annotation process

http://oxytricha.princeton.edu/mds_ies_db/

Welcome!

The **<mds_ies_db>** is a dynamic database featuring ciliate genome rearrangement annotations that include MDS, IES, and pointer annotations for *Oxytricha* and *Tetrahymena*.



What is the <mds_ies_db>?

The <mds_ies_db> is a database that uses customized searches and dynamic charts to elucidate the extensive genome rearrangement process occurring in some species of ciliates.

Ciliates contain two types of nuclei, a somatic macronucleus (MAC) and a germline micronucleus (MIC), with distinct genomes. During conjugation the MAC is disintegrated, and the MIC undergoes genome editing which includes large-scale deletion and segment rearrangement to form a new MAC (see Ciliate Biology for further details).

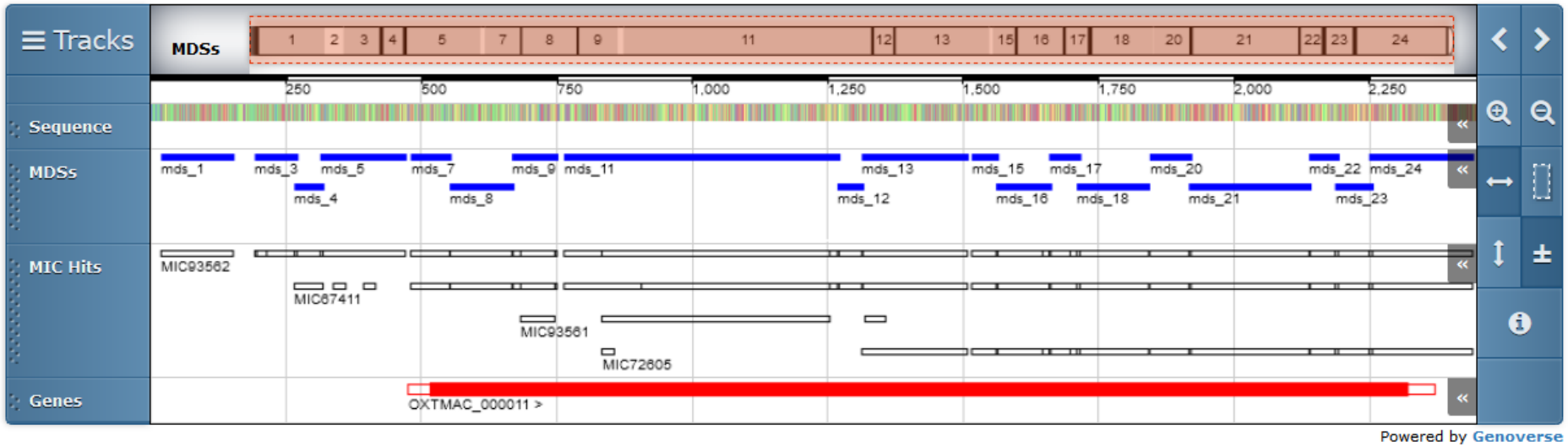
The <mds_ies_db> contains annotations for the MIC segments that are retained in the

What's New?

2015-09-06	Statistics page added
2015-08-11	Home page updated
2015-08-05	Help page added
2015-08-01	About us page added
2015-07-20	*Tetrahymena thermophila* and *Oxytricha trifallax* MDS annotation of MIC uploaded

OXYTRI_MAC_1 (*Oxytricha trifallax JRB310*, Macronuclear Contig)

Help



Powered by Genoverse

Chord Diagram | MDS-IES | Hits Table | Downloads

DNA Information

Sequence Information

DNA Sequence: 2,457 nt
Telomeres: 5′ and 3′ Telomeres

Cross References

OxyDB: GBrowse
GenBank: AMCR01013193

MDS Information

MDS Information

MDS Count: 24
MIC Matches: 4

Gene Information

Future work

- Extend program to also identify genes that go through the rearrangement process
- Improve high scoring pair merging procedure to filter out the noise
- Provide a GUI interface for the implemented program

References

- Chen, Xiao et al. The Architecture of a Scrambled Genome Reveals Massive Levels of Genomic Rearrangement during Development, *Cell* (2014), Volume 158 , Issue 5 , 1187 - 1198
- Burns, Jonathan, Kukushkin, Denys, Jonoska, Nataša <mds_ies_db>: A database of ciliate genome rearrangement, Submitted

This work is supported in part by the NSF grants CCF-1117254 and DMS-0900671 and NIH grant R01GM109459-03

Algorithm Implementation

- Implemented as a Python program: MDS/IES DNA Annotation Software
- Main website: http://knot.math.usf.edu/midas/
- Github repository: https://github.com/j-t-burns/MI-ASS
- Program output was stored in the database: http://oxytricha.princeton.edu/mds_ies_db/