SDRAP: An Annotation Pipeline for Highly Scrambled Genomes

Jasper Braun[1], Rafik Neme[2], Nataša Jonoska[1], Laura Landweber[3]

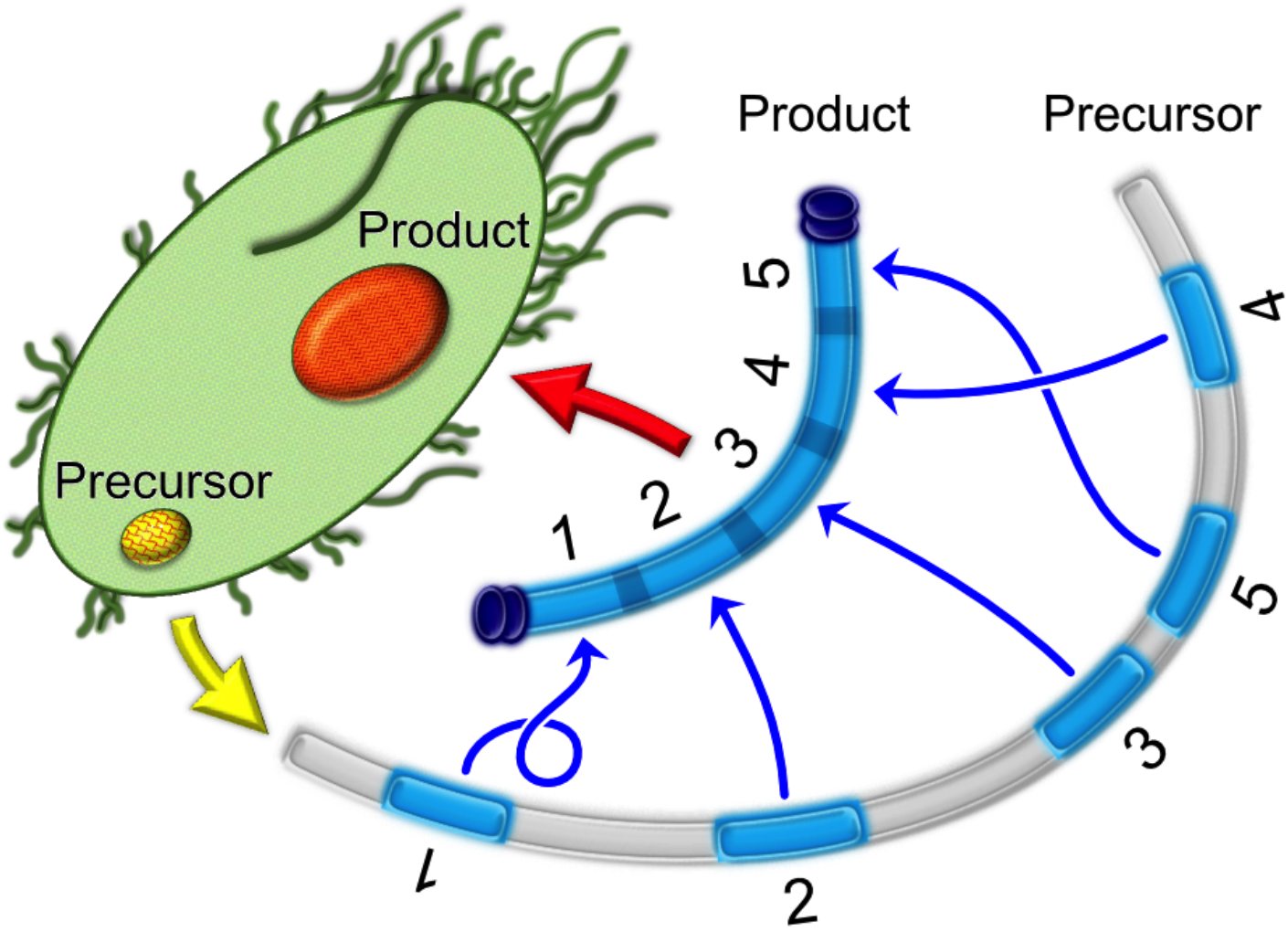
[1] Department of Mathematics & Statistics, University of South Florida
[2] Department of Chemistry and Biology, Universidad del Norte
[3] Department of Biochemistry and Molecular Biophysics, and Biological Sciences, Columbia University

jasperbraun@mail.usf.edu

October 6, 2019

Supported in part by NSF DMS-1800443

DNA Rearrangements

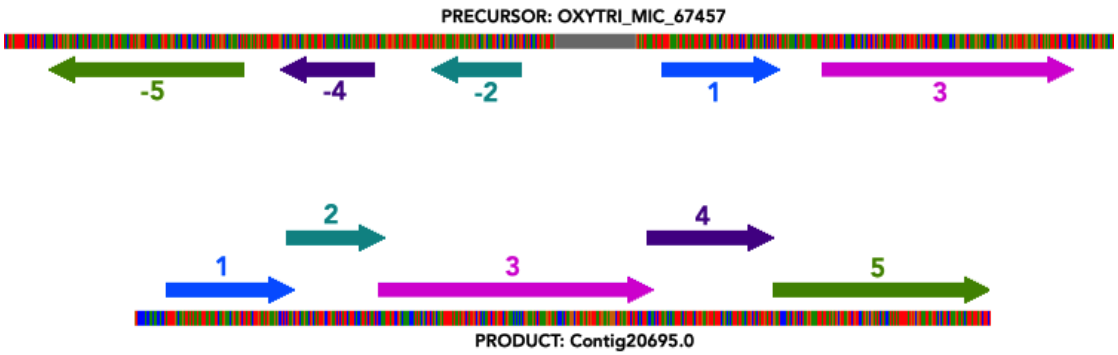


Existing Software

Basic Local Alignment Search Tool (BLAST)[1]:

- finds similar regions
- widely used and efficient

```
Query   403    TGATAATGATAAAGAAGATGATGATGTTGATGAAAA   438
               | ||| ||||||||||||||| |||||   |||||||
Sbjct   6715   TTATATTGATAAAGAAGATGTTGATGAAAATGAAAA   6680
```

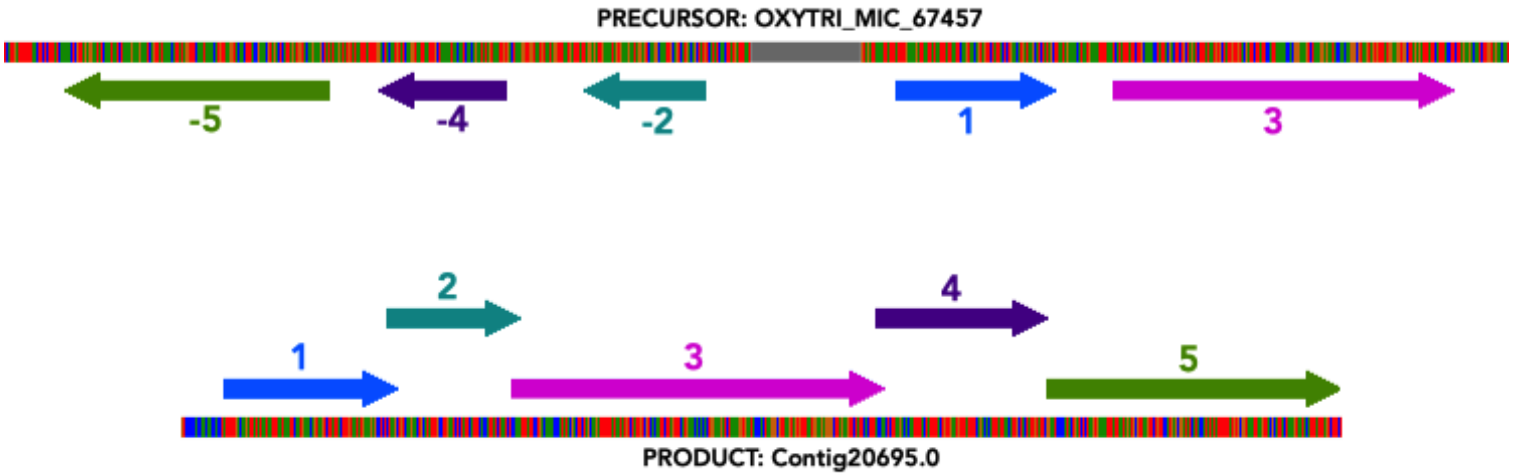


MDS/IES DNA Annotation Software (MIDAS)[4]:

- uses BLAST to find similar regions
- annotates rearranging segments
- identifies scrambled rearrangement maps

Why is MIDAS not enough?

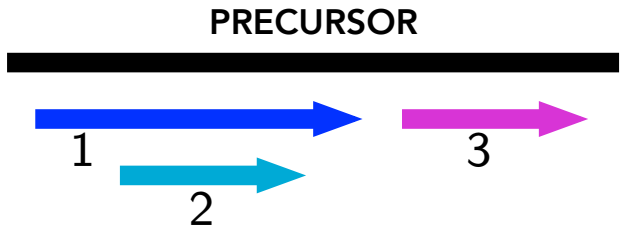
MIDAS compares the **order and orientation** of rearranging segments on the precursor to the order and orientation of the **corresponding** segments on the product. [2]



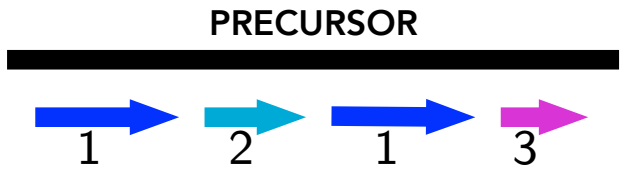
Special Cases

Method used by MIDAS to determine scrambling becomes unclear when applied to certain *"special cases"*

order of segments is difficult to define

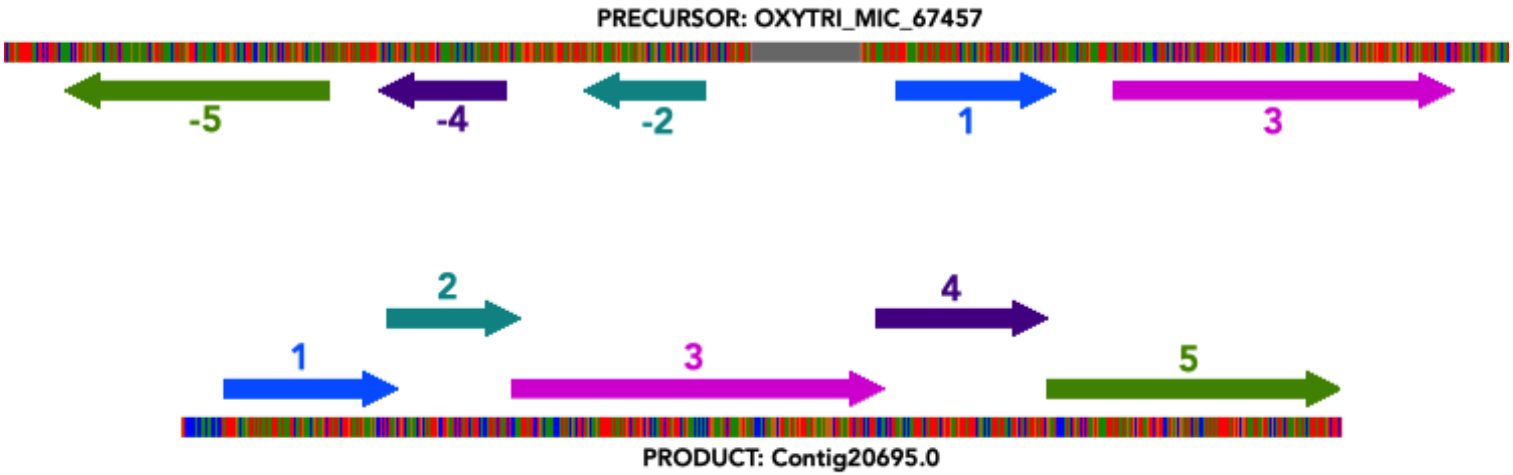


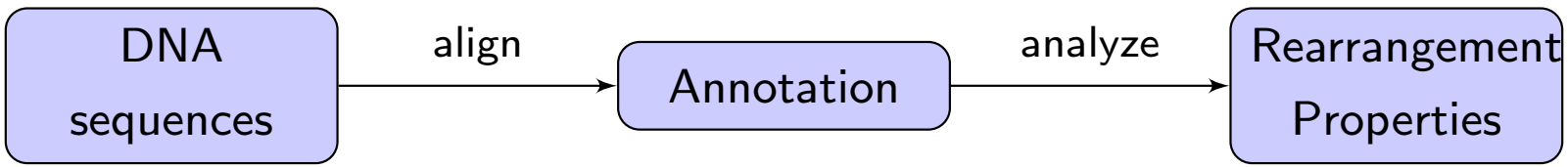
one-to-one map from precursor to product may not exist



SDRAP

- **INPUT:** Precursor and Product nucleotide sequences
- **OUTPUT:** Annotation of rearranging segments and analysis of scrambling for all precursor and product sequences

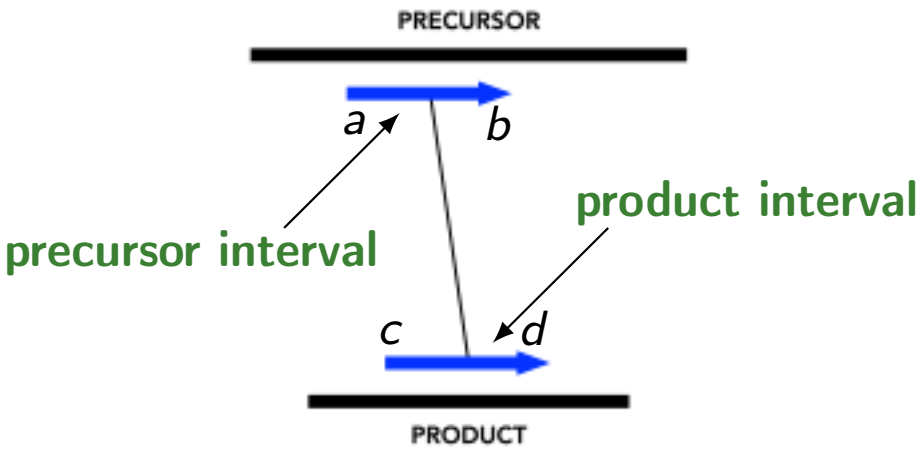




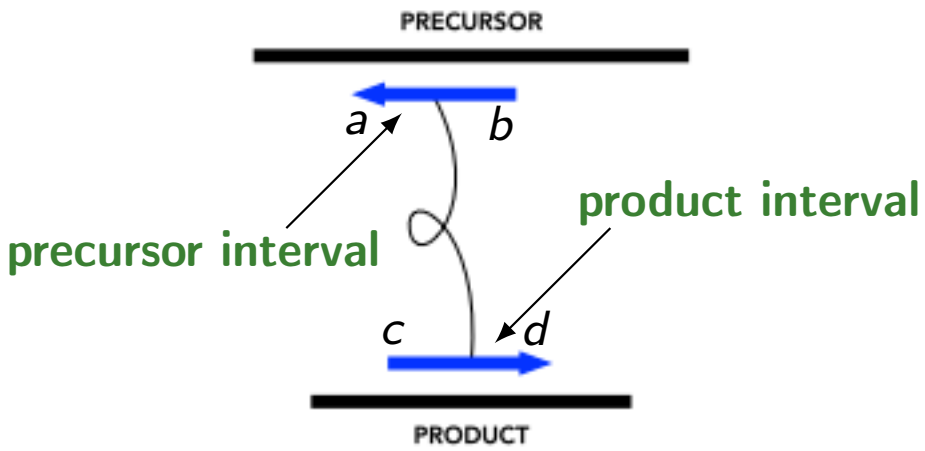
Definition

A **match** consists of:

- an integer interval $[a, b]$ called the **precursor interval**
- an integer interval $[c, d]$ called the **product interval**
- an **orientation** $+$ or $-$



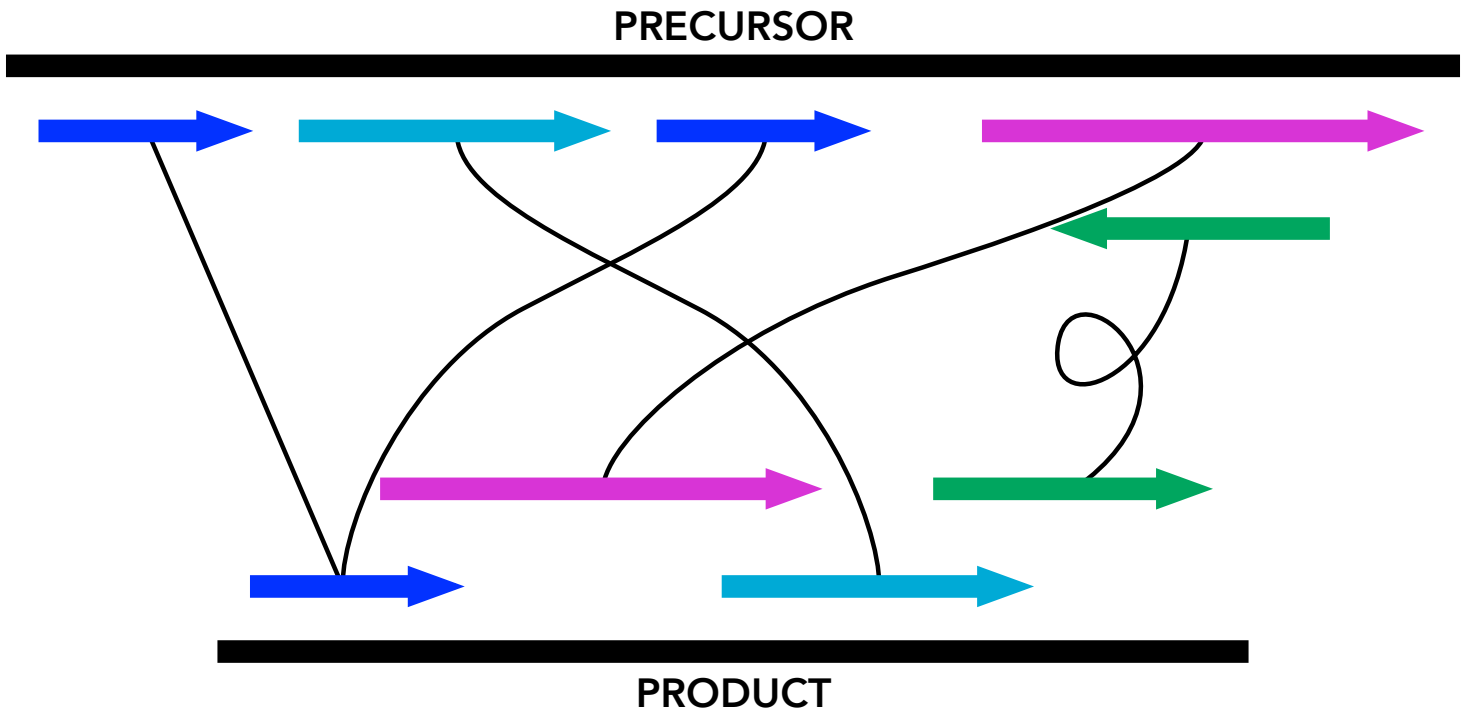
orientation: $+$



orientation: $-$

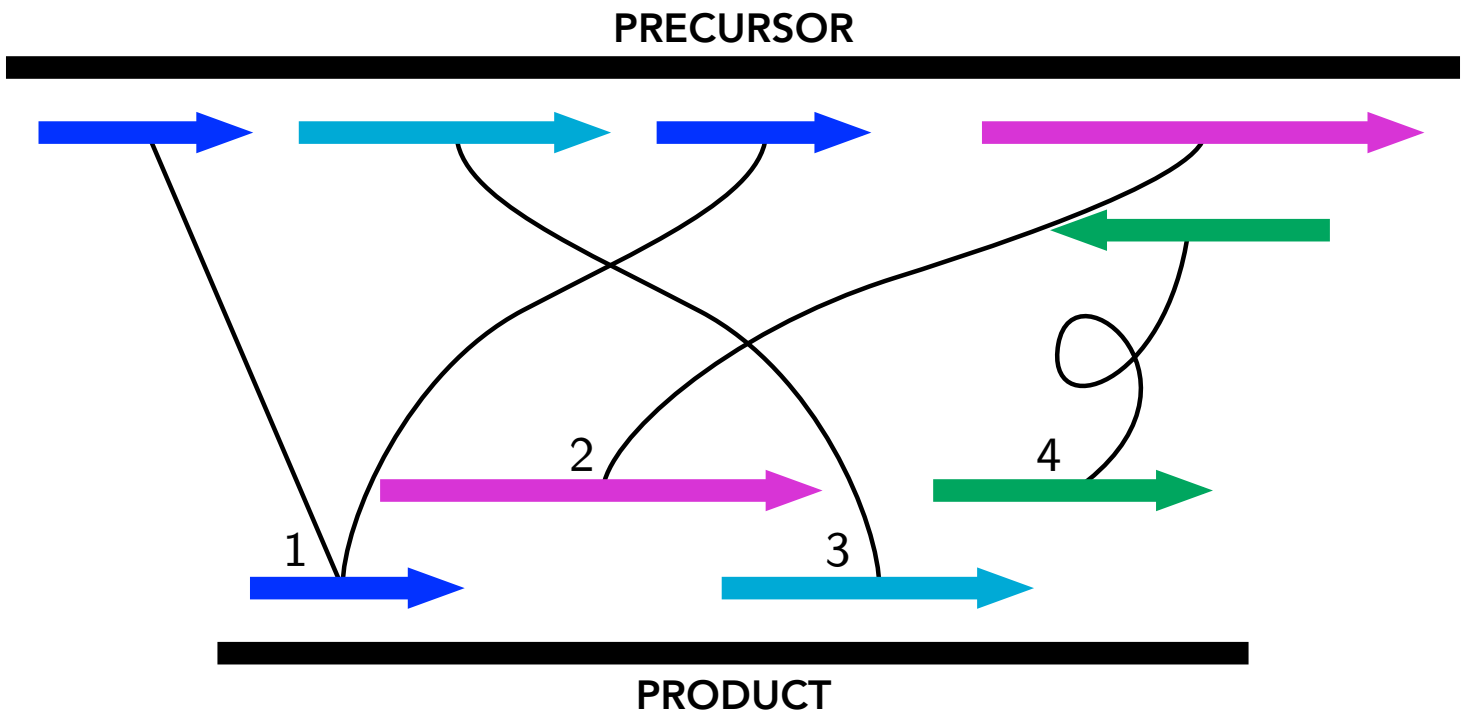
Definition

An **arrangement** is a set of matches where the product interval of none of the matches in the set properly contains the product interval of another.



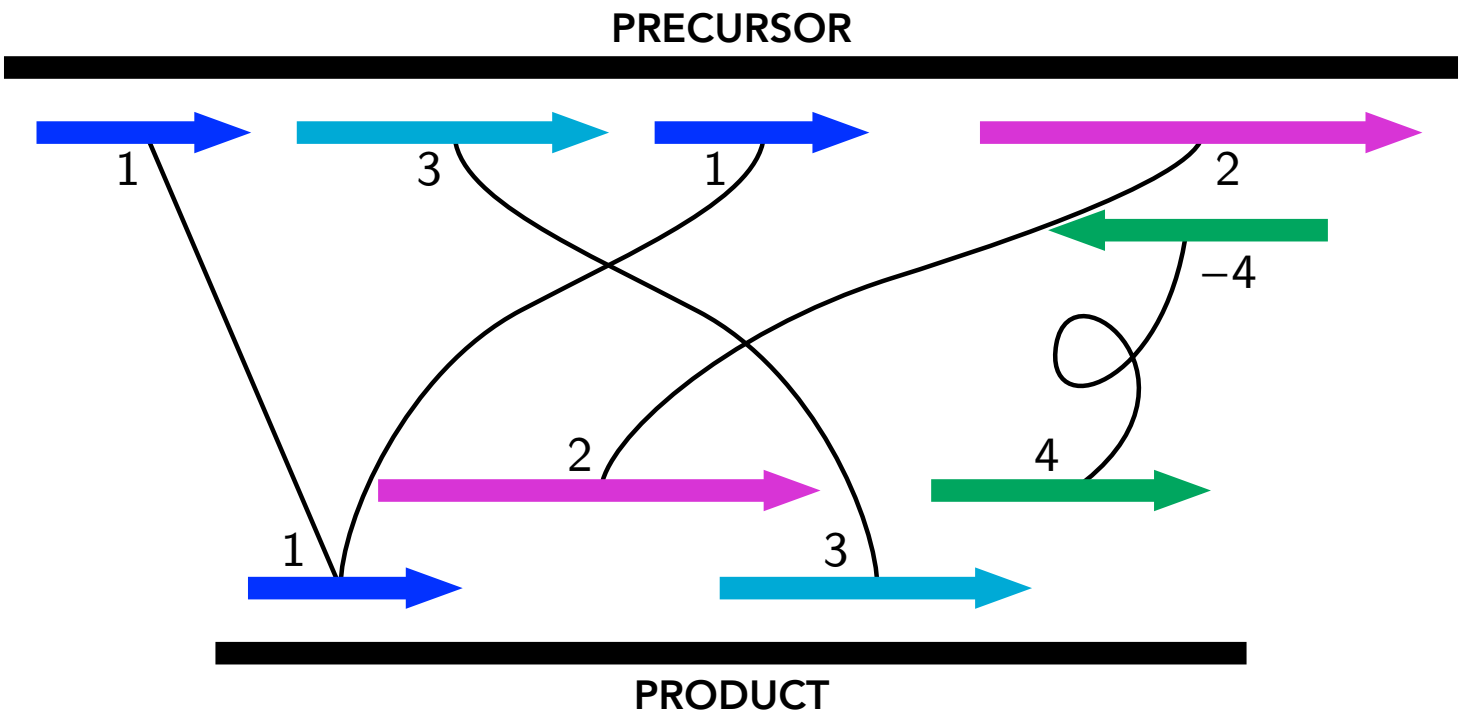
Definition

An **arrangement** is a set of matches where the product interval of none of the matches in the set properly contains the product interval of another.



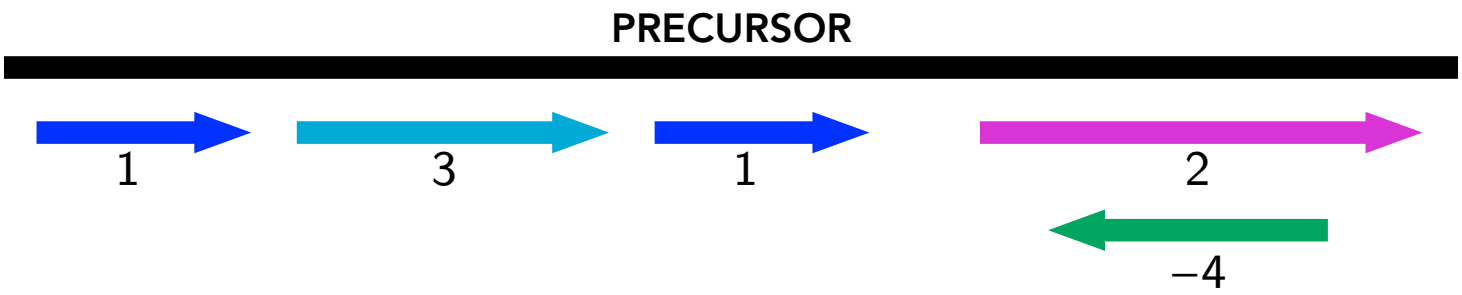
Definition

An **arrangement** is a set of matches where the product interval of none of the matches in the set properly contains the product interval of another.



Definition

An **arrangement** is a set of matches where the product interval of none of the matches in the set properly contains the product interval of another.

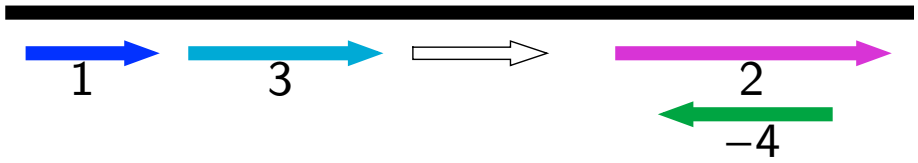


Handling "Special Cases"

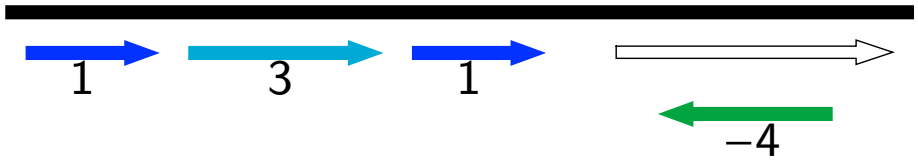
A **well-behaved subarrangement** of an arrangement is a subset of its matches with the following properties:

- no two matches share the same product interval
- no two precursor intervals overlap
- the subset is maximal with the above two properties.

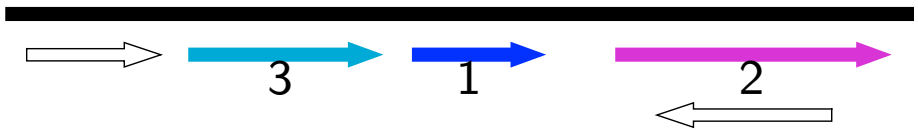
no two matches share the same product interval



no two precursor intervals overlap



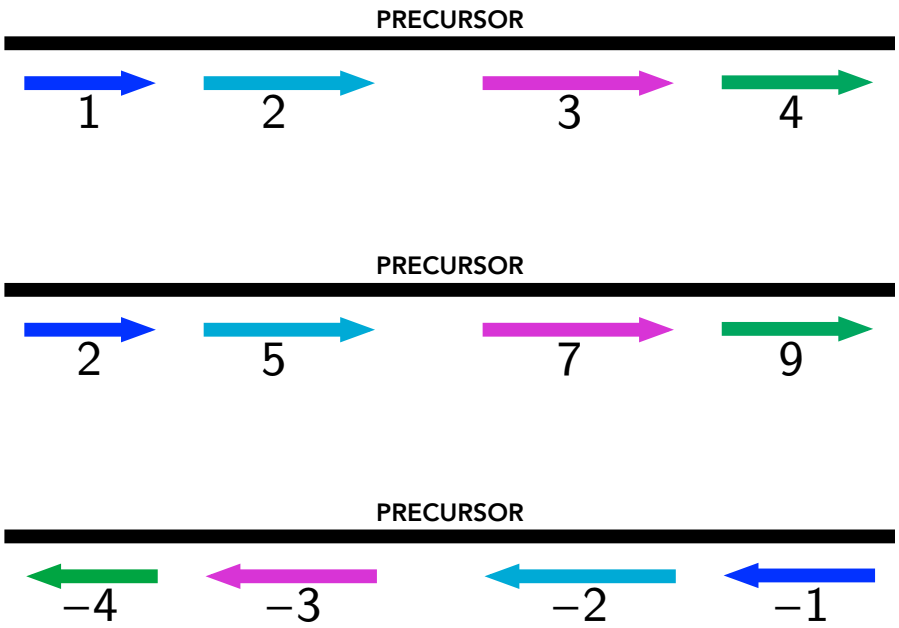
well-behaved



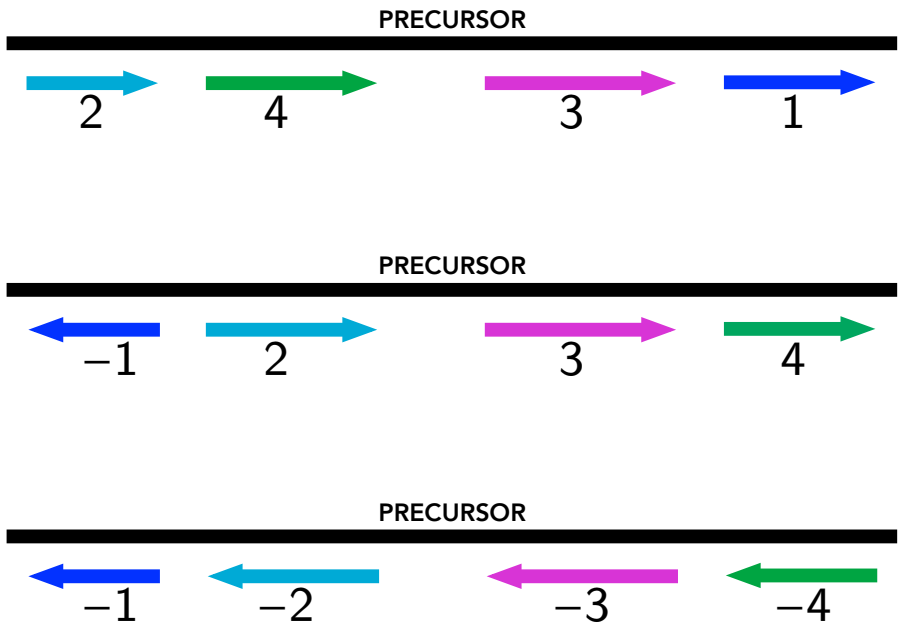
What is Nonscrambled?

A well-behaved subarrangement of an arrangement is **ordered** if the precursor intervals occur in the same order and orientation as the corresponding product intervals.

ordered:



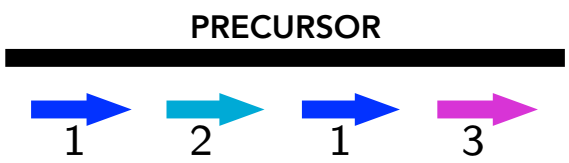
not **ordered**:



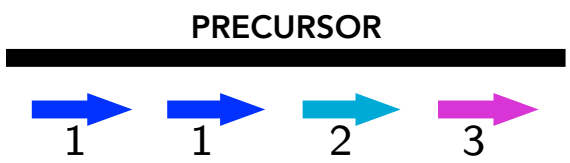
Definition

An arrangement is:

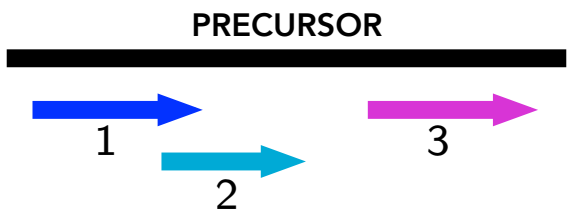
- **weakly ordered** if at least one well-behaved subarrangement is ordered
- **strongly ordered** if all well-behaved subarrangements are ordered



weakly ordered



strongly ordered

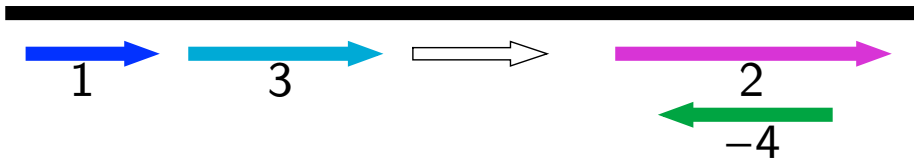


strongly ordered

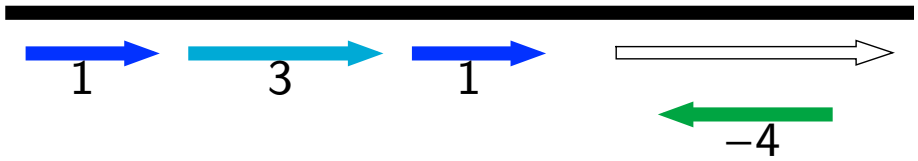
Definition

A subset of an arrangement is **complete** if every product interval from the arrangement belongs to at least one of the matches in the subset

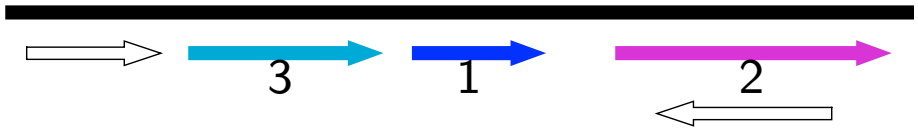
complete



not **complete**



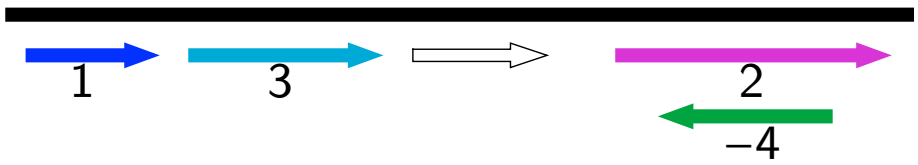
not **complete**



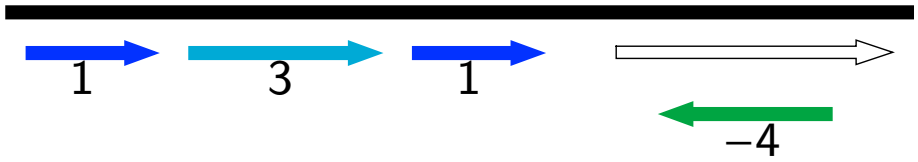
Definition

A subset of an arrangement is **consecutive** if whenever two product intervals belong to matches in the subset, every product interval in-between does, too

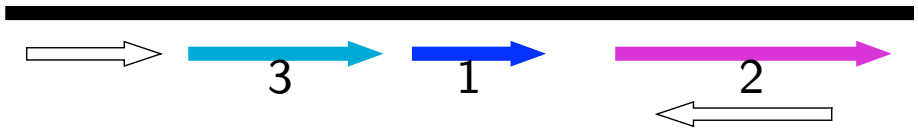
consecutive and **complete**



neither **consecutive**, nor **complete**



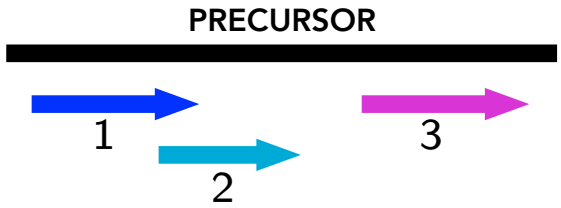
consecutive, but not **complete**



Scrambling

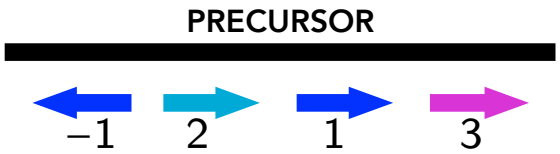
Let $S \subseteq \{\text{ordered, consecutive, complete}\}$. An arrangement is:

- **weakly S-nonscrambled** if at least one well-behaved subarrangement has every property in S
- **strongly S-nonscrambled** if all well-behaved subarragenemts have every property in S



$S = \{\text{ordered, consecutive}\}$

weakly, but not **strongly** **S-nonscrambled**



$S = \{\text{ordered, complete}\}$

neither **weakly**, nor **strongly** **S-nonscrambled**

SDRAP

Scrambled DNA Rearrangements Annotation Pipeline (SDRAP) was implemented as a web application using javascript, PHP, and MySQL (github.com/JasperBraun/SDRAP)

SDRAP was tested on the precursor and product genomes of the ciliate *Oxytricha trifallax* sequenced in [3], and [5], respectively

Data consists of 25,720 precursor sequences and 22,450 product sequences

arrangements where at least 50% of product sequence is covered by matches	multiple matches share same product interval	precursor intervals overlap	multiple matches share same product interval and precursor intervals overlap
97,846	17,108	16,345	8,710

	{**ordered**}	{**ordered, consecutive**}	{**ordered, complete**}
weakly *S*-**nonscrambled**	75,465	73,379	66,683
strongly *S*-**nonscrambled**	67,048	64,599	62,666

Future Work

- Use SDRAP to analyze scrambling patterns in organisms other than *O. trifallax*
- Compare scrambling patterns between different organisms

References

[1] Stephen F Altschul et al. "Basic local alignment search tool". In: *Journal of Molecular Biology* 215.3 (Oct. 1990), pp. 403–410. DOI: 10.1016/S0022-2836(05)80360-2.

[2] Jonathan Burns et al. "<mds_ies_db>: a database of ciliate genome rearrangements". In: *Nucleic Acids Research* 44.D1 (Jan. 2016), pp. D703–D709. DOI: 10.1093/nar/gkv1190.

[3] Xiao Chen et al. "The architecture of a scrambled genome reveals massive levels of genomic rearrangement during development". In: *Cell* 158.5 (Aug. 2014), pp. 1187–1198. DOI: 10.1016/j.cell.2014.07.034.

[4] USF Math-Bio Research Lab. *MDS/IES DNA Annotation Software*. 2015. URL: http://http://knot.math.usf.edu/midas/index.html (visited on 09/28/2019).

[5] Estienne C Swart et al. "The *Oxytricha trifallax* Macronuclear Genome: A Complex Eukaryotic Genome with 16,000 Tiny Chromosomes". In: *PLOS Biology* 11.1 (Jan. 2013), pp. 1–29. DOI: 10.1371/journal.pbio.1001473.

Thanks