Accurate Detection of Tandem Repeats from Error-Prone Sequences with EquiRep

Zhezheng Song[1,*] Tasfia Zahin[1,*] Xiang Li[1] Mingfu Shao[1,2]

[1]Department of Computer Science and Engineering, School of Electrical Engineering and Computer Science, The Pennsylvania State University

[2]Huck Institutes of the Life Sciences, The Pennsylvania State University

RECOMB Seoul | 서울 | 2025

PennState College of Engineering | ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

PennState Huck Institutes of the Life Sciences

Tandem Repeats

- Sequence of nucleotides that appears as multiple contiguous, near-identical copies arranged consecutively.

Tandem Repeats

- Sequence of nucleotides that appears as multiple contiguous, near-identical copies arranged consecutively.

- Length of these repeat units vary from a few base pairs in STRs to a hundred base pairs in VNTRs and satellite DNAs.

Tandem Repeats

- Sequence of nucleotides that appears as multiple contiguous, near-identical copies arranged consecutively.
- Length of these repeat units vary from a few base pairs in STRs to a hundred base pairs in VNTRs and satellite DNAs.
- Exact: ACGTACGTACGTACGTACGT

Tandem Repeats

- Sequence of nucleotides that appears as multiple contiguous, near-identical copies arranged consecutively.
- Length of these repeat units vary from a few base pairs in STRs to a hundred base pairs in VNTRs and satellite DNAs.
- Exact: ACGTACGTACGTACGTACGT
- Approximate: ACTTACTACGTCCGTACGGT

Significance

Significance

Tandem repeats make up about **8-10% of the human genome**.

Significance

Tandem repeats make up about **8-10% of the human genome**.

Closely linked to several **neurological and developmental disorders** like Huntington's disease, Friedreich's Ataxia, fragile X syndrome.

Significance

Tandem repeats make up about **8-10% of the human genome**.

Closely linked to several **neurological and developmental disorders** like Huntington's disease, Friedreich's Ataxia, fragile X syndrome.

Satellite repeats are found to be abundant in centromeric regions of many organisms and are essential for studying genome stability and evolutionary dynamics.

Significance

Tandem repeats make up about **8-10% of the human genome**.

Closely linked to several **neurological and developmental disorders** like Huntington's disease, Friedreich's Ataxia, fragile X syndrome.

Satellite repeats are found to be abundant in centromeric regions of many organisms and are essential for studying genome stability and evolutionary dynamics.

Tandem **repeats generated by certain technologies** such as those used for circular molecules can be useful for full-length circular RNA assembly.

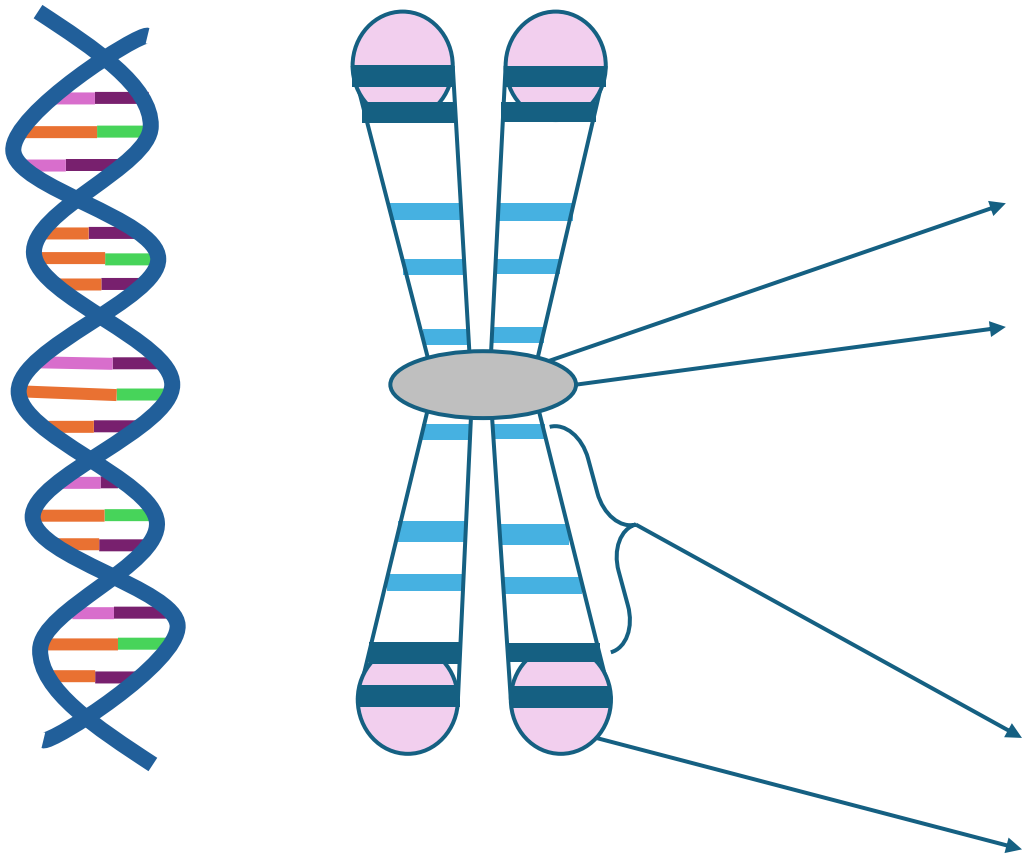
Applications

Applications

Study of Satellite DNAs

Applications

Study of Satellite DNAs



Satellite Family	Size of Repeat Unit (bp)	Location in Human Chromosome
α	170	Centromeres of all chromosomes
β	68	Centromeres of chromosomes 1, 9 13, 14, 15, 21, 22, and Y
Satellite 1	25-48	Centromeres and other regions in heterochromatin of most chromosomes
Satellite 2	5	Most chromosomes
Satellite 3	5	Most chromosomes
Microsatellites	1-10	Widely distributed throughout the chromosome
Minisatellites	10-60	Telomeres of many chromosomes

https://en.wikipedia.org/wiki/Satellite_DNA

Applications

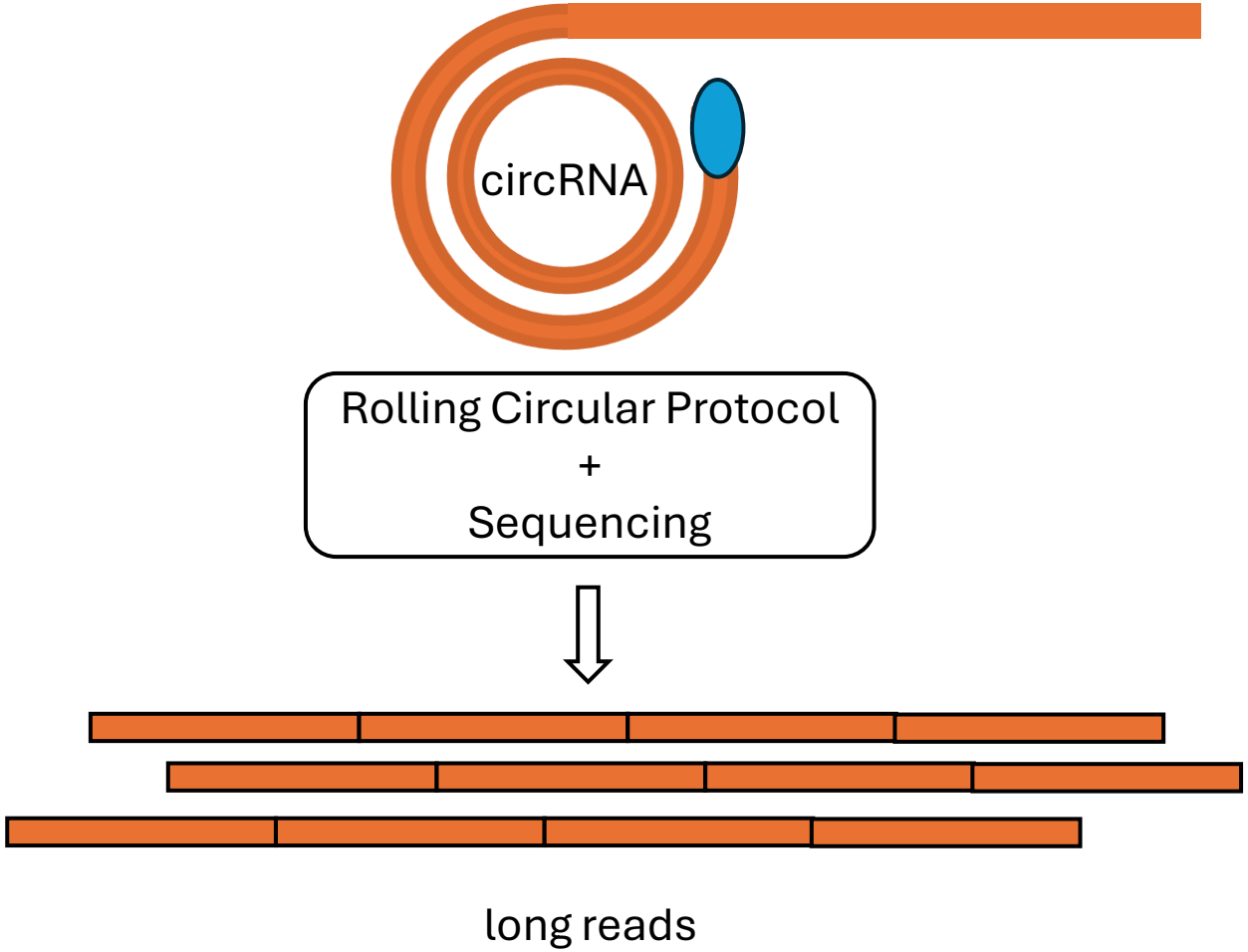
Assemble Circular RNAs from Rolling Circular Reads

Applications

Assemble Circular RNAs from Rolling Circular Reads

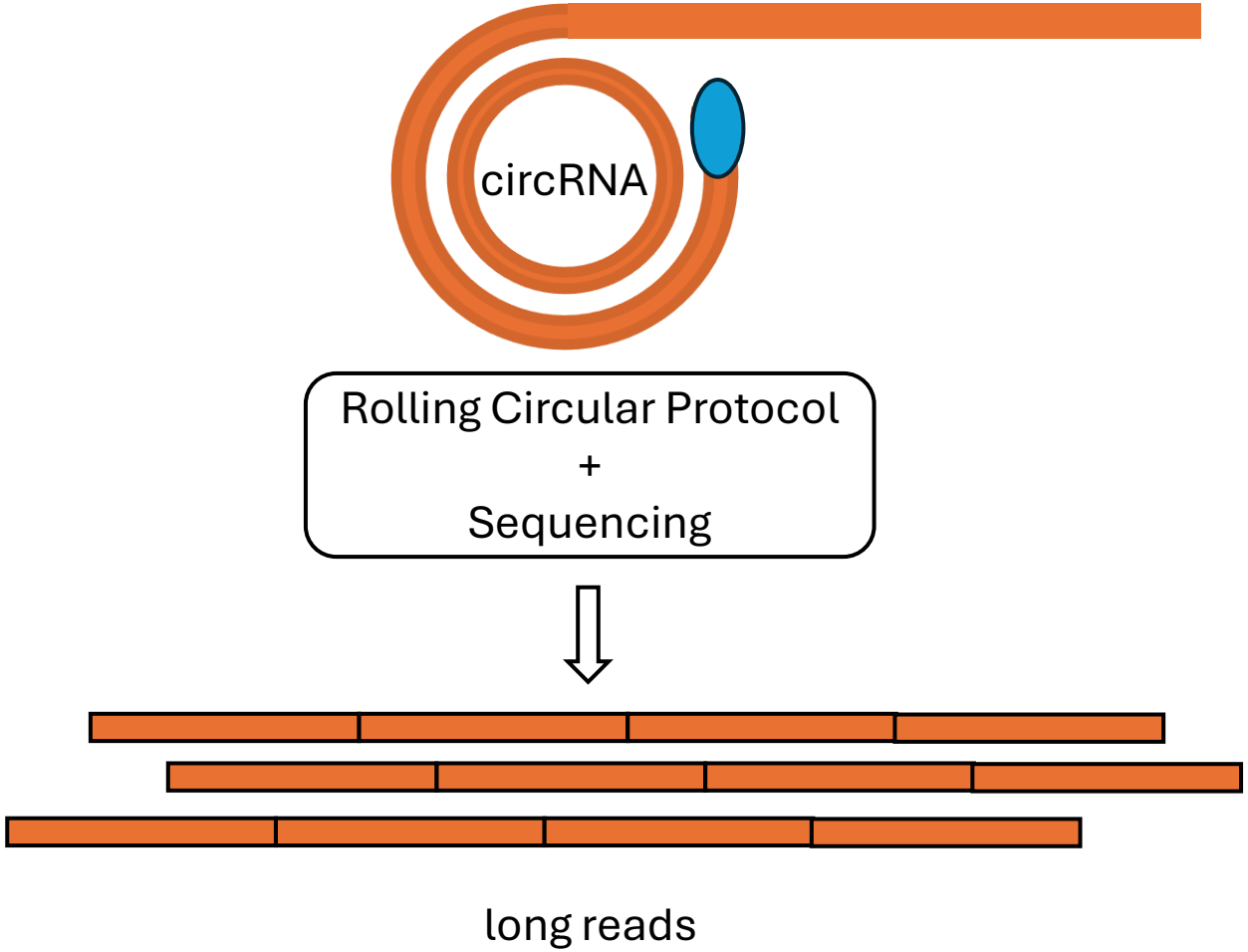
Applications

Assemble Circular RNAs from Rolling Circular Reads



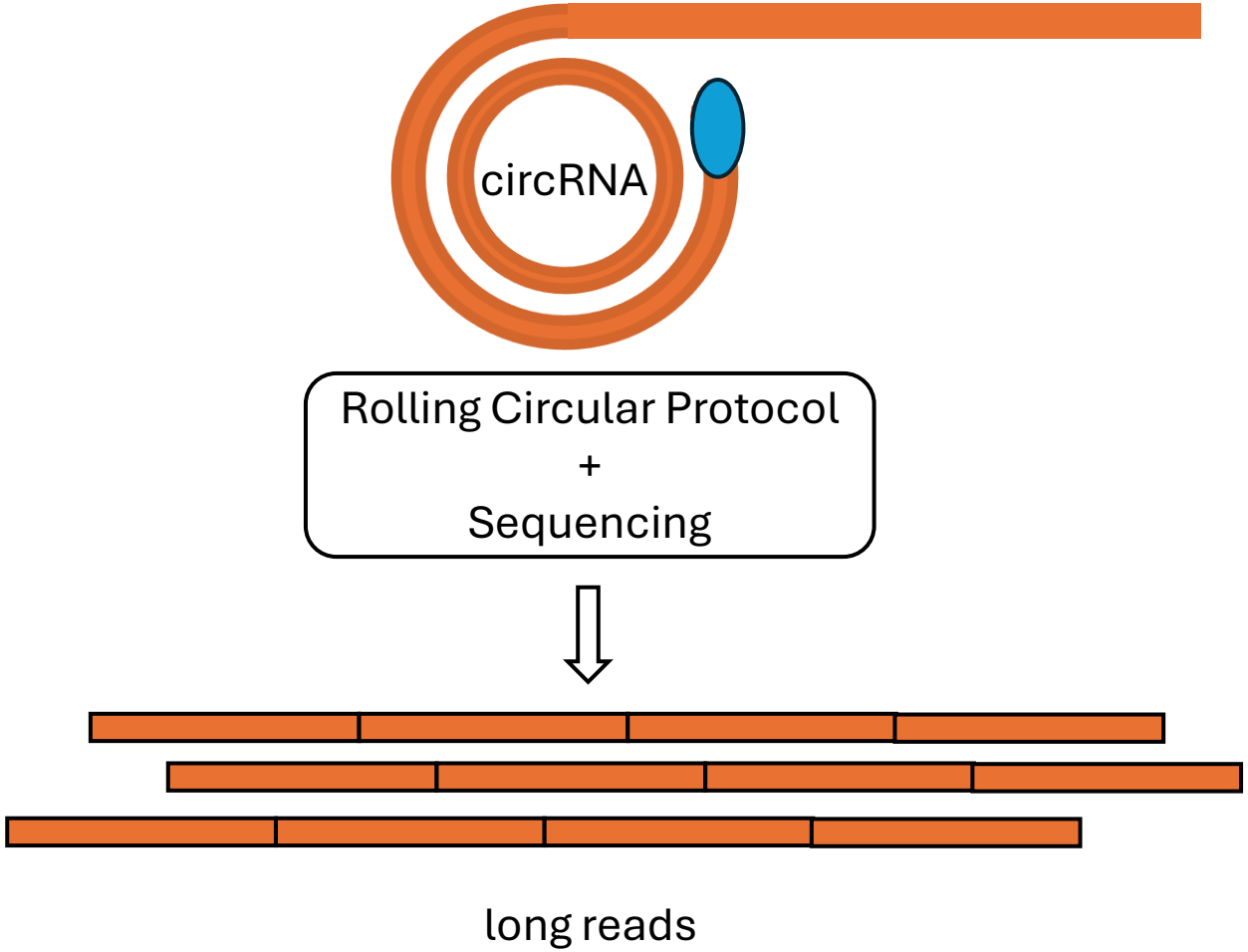
Applications

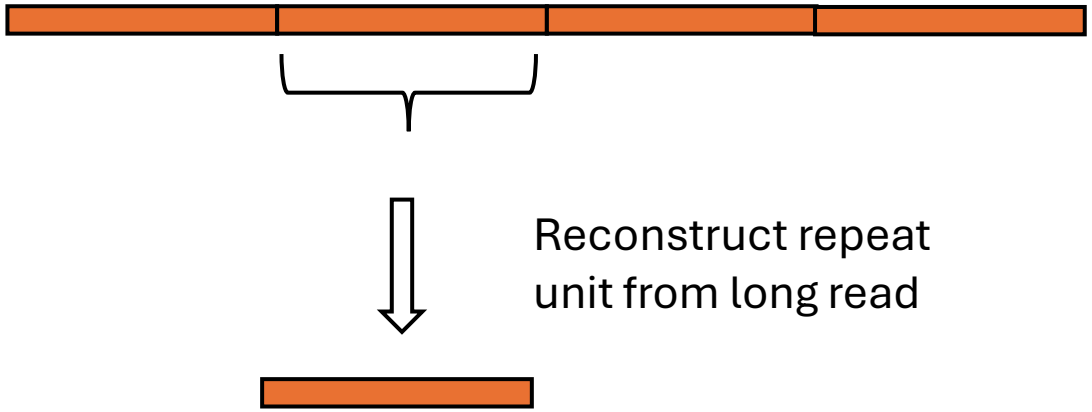
Assemble Circular RNAs from Rolling Circular Reads



Applications

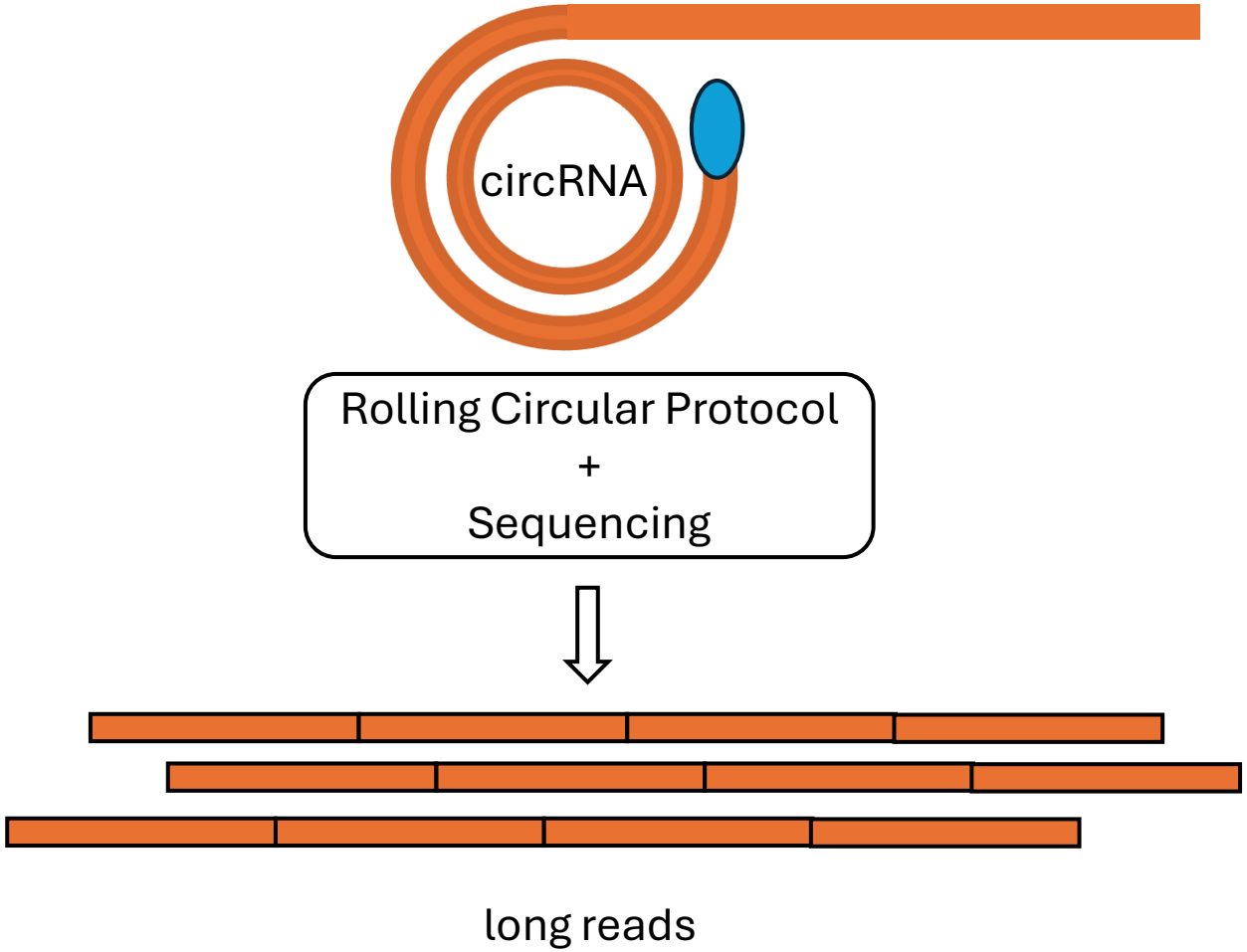
Assemble Circular RNAs from Rolling Circular Reads

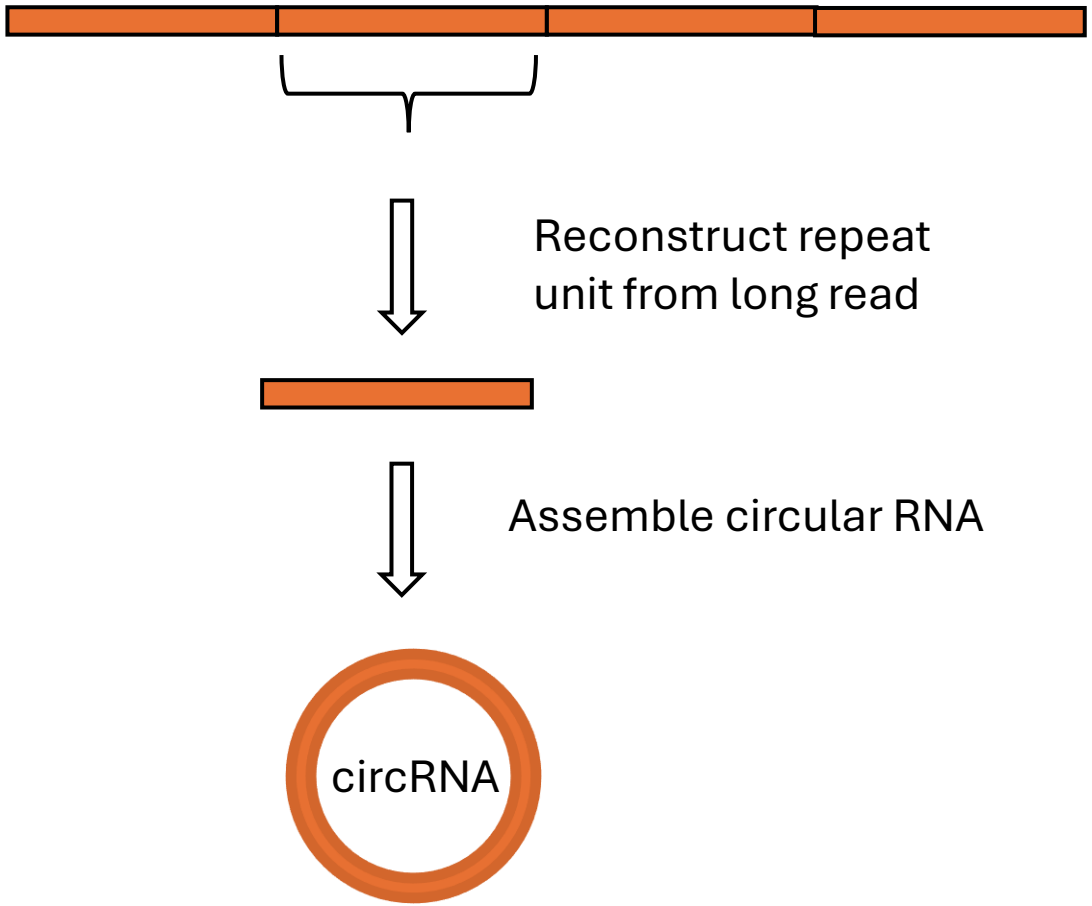




Applications

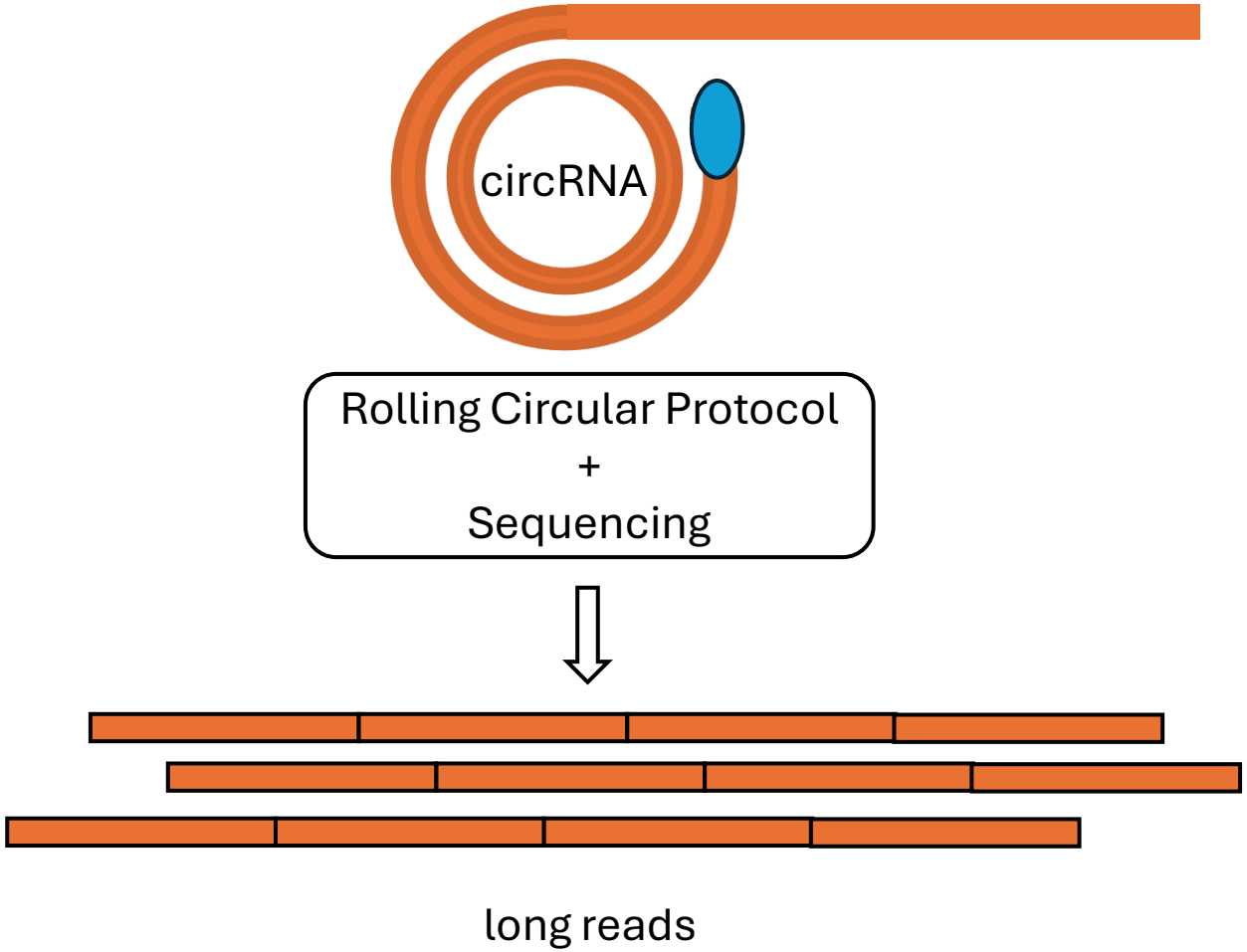
Assemble Circular RNAs from Rolling Circular Reads

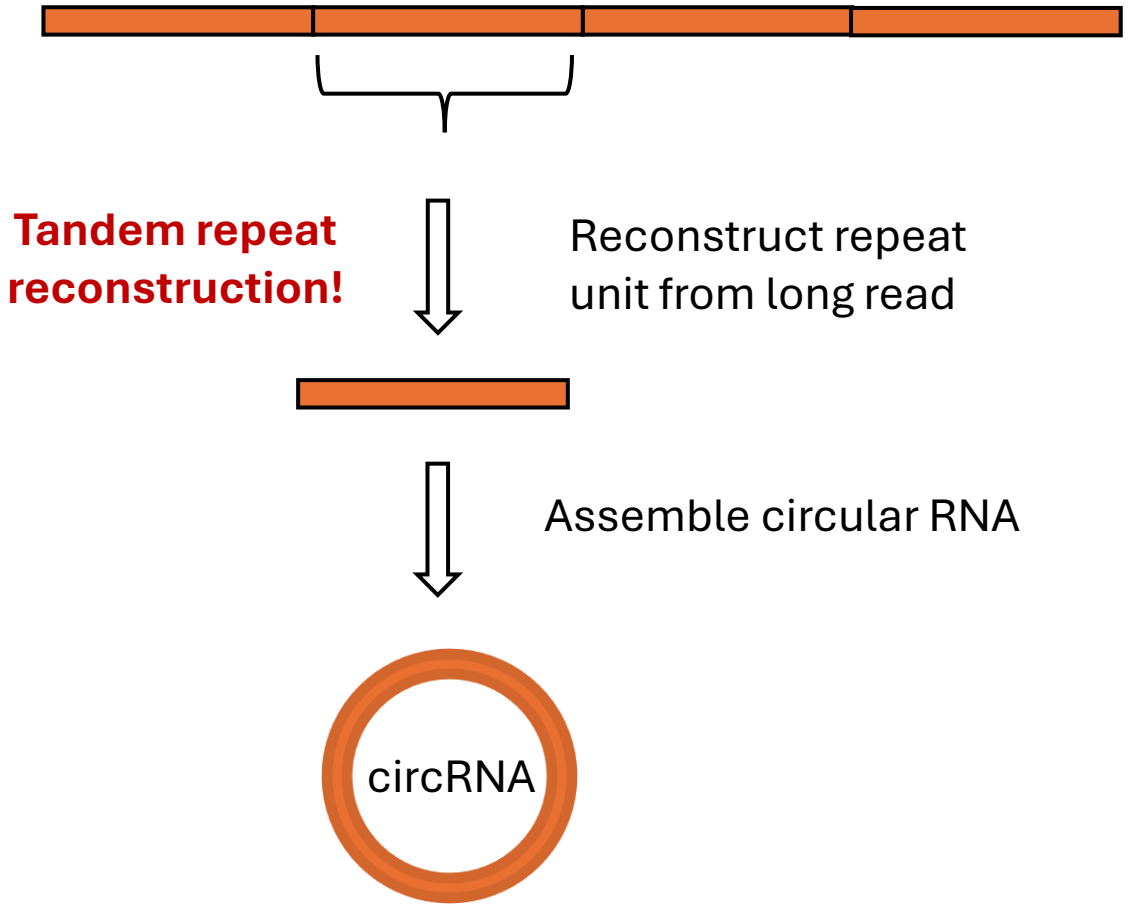




Applications

Assemble Circular RNAs from Rolling Circular Reads





Tandem Repeat Reconstruction

Tandem Repeat Reconstruction

Input: an error-prone (long) sequence/read.

ACTTACTACGTCCGTACGGT

Tandem Repeat Reconstruction

Input: an error-prone (long) sequence/read.

ACTTACTACGTCCGTACGGT

Output: tandem repeat unit (if any).

ACGT

Existing Methods

Existing Methods

- Most tools are designed for reconstruction of short units from relatively low error data.
  - mreps, dot2dot.
  - They often do not perform well with higher repeat lengths and/or lower frequencies.

Existing Methods

- Most tools are designed for reconstruction of short units from relatively low error data.
  - mreps, dot2dot.
  - They often do not perform well with higher repeat lengths and/or lower frequencies.

- Tools capable of managing high error rates are rare, existing ones struggle to achieve satisfactory accuracy in challenging settings.
  - mTR struggles with repeats of low copy numbers.
  - TideHunter compromises accuracy when dealing with repeats of small length.

Existing Methods

- Most tools are designed for reconstruction of short units from relatively low error data.
  - mreps, dot2dot.
  - They often do not perform well with higher repeat lengths and/or lower frequencies.
- Tools capable of managing high error rates are rare, existing ones struggle to achieve satisfactory accuracy in challenging settings.
  - mTR struggles with repeats of low copy numbers.
  - TideHunter compromises accuracy when dealing with repeats of small length.
- Other tools focus on quantification rather than reconstruction.
  - DeepRepeat, ExpansionHunter.

Existing Methods

- Most tools are designed for reconstruction of short units from relatively low error data.
  - mreps, dot2dot.
  - They often do not perform well with higher repeat lengths and/or lower frequencies.
- Tools capable of managing high error rates are rare, existing ones struggle to achieve satisfactory accuracy in challenging settings.
  - mTR struggles with repeats of low copy numbers.
  - TideHunter compromises accuracy when dealing with repeats of small length.
- Other tools focus on quantification rather than reconstruction.
  - DeepRepeat, ExpansionHunter.

EquiRep: a new tool for reconstructing tandem repeat units from error-prone sequences.
- Robust against errors.
- Effective in detecting repeats of large length and low frequency.

Definitions

Definitions

- **Equivalent positions**: i ~ j, if R[i] and R[j] originate from the same letter in the true unit U.

Erroneous Sequence (R)

A C T T A C T A C G T C C G T A C G G T

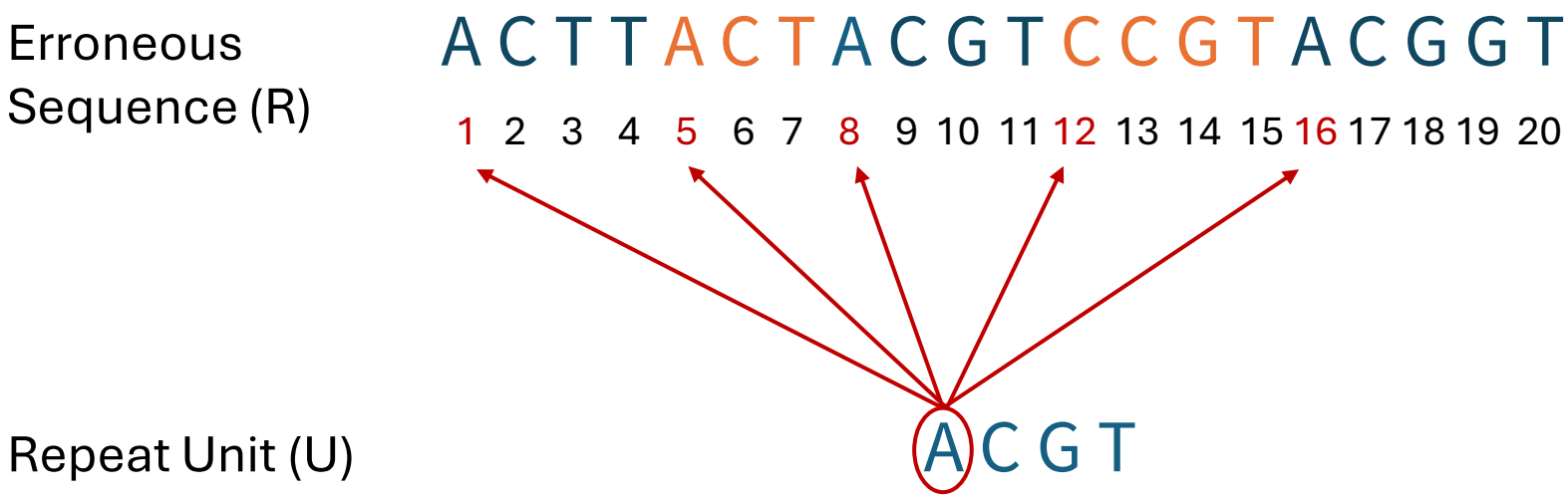
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

Repeat Unit (U)

A C G T

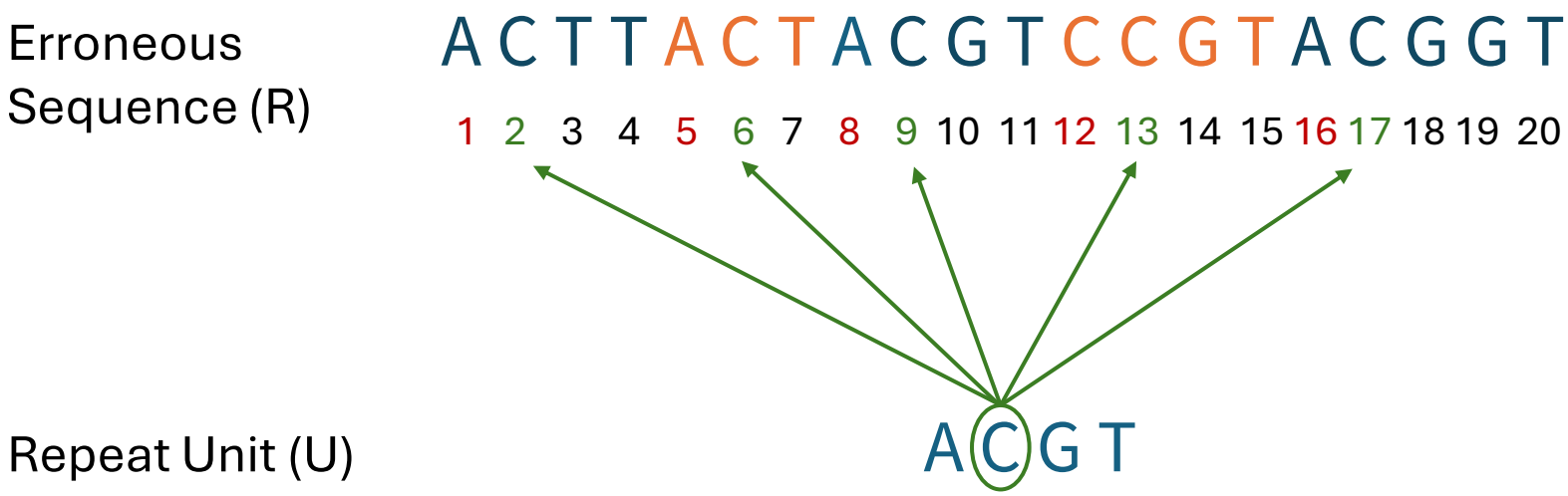
Definitions

- **Equivalent positions**: i ~ j, if R[i] and R[j] originate from the same letter in the true unit U.



Definitions

- **Equivalent positions**: i ~ j, if R[i] and R[j] originate from the same letter in the true unit U.



Definitions

- **Diagonal-free self-alignment**: Local alignment between sequence R and itself, with constraint that the same position cannot be aligned to itself.

Erroneous Sequence (R)

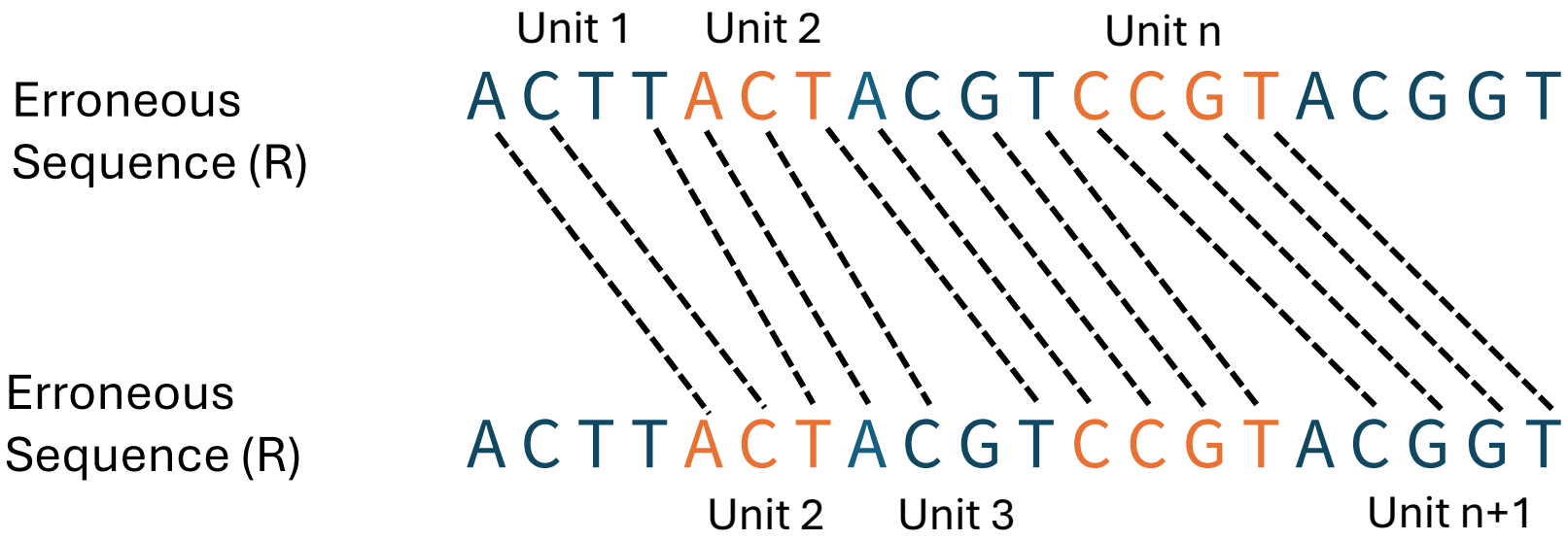
A C T T A C T A C G T C C G T A C G G T

Erroneous Sequence (R)

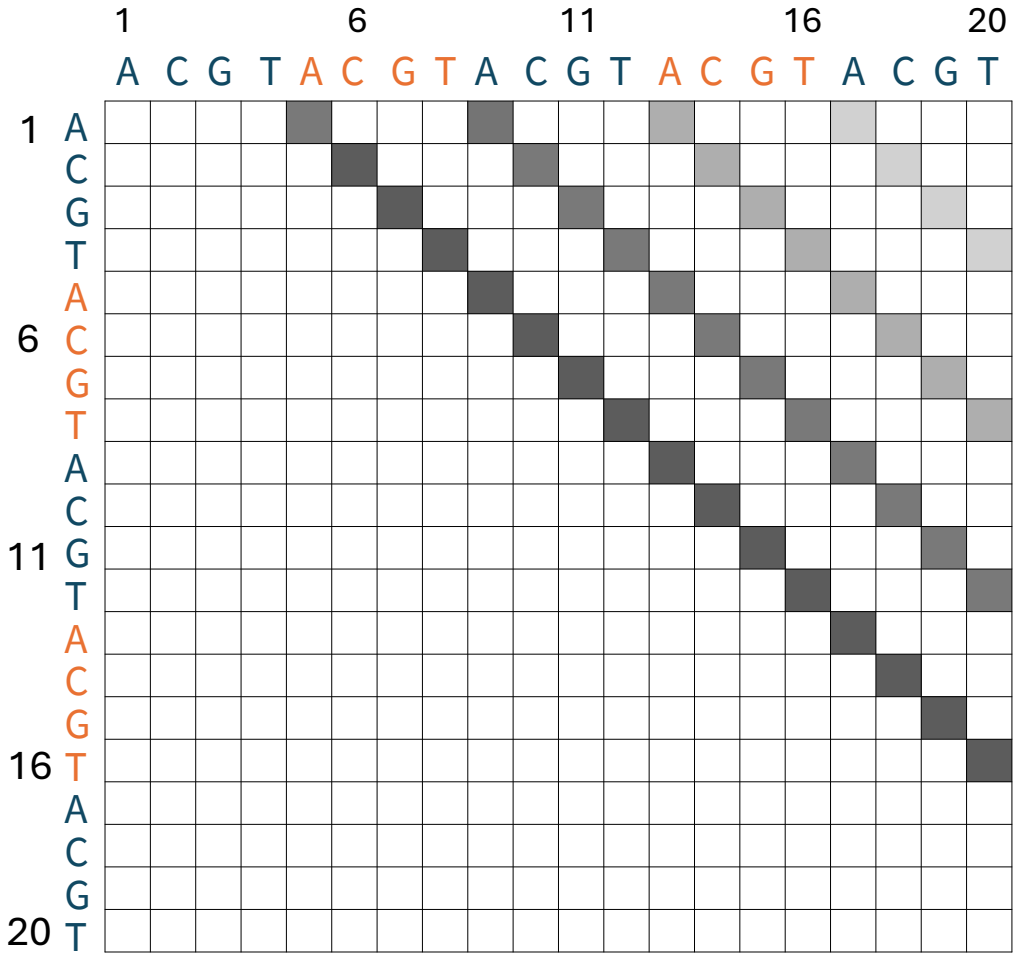
A C T T A C T A C G T C C G T A C G G T

Definitions

- **Diagonal-free self-alignment**: Local alignment between sequence R and itself, with constraint that the same position cannot be aligned to itself.

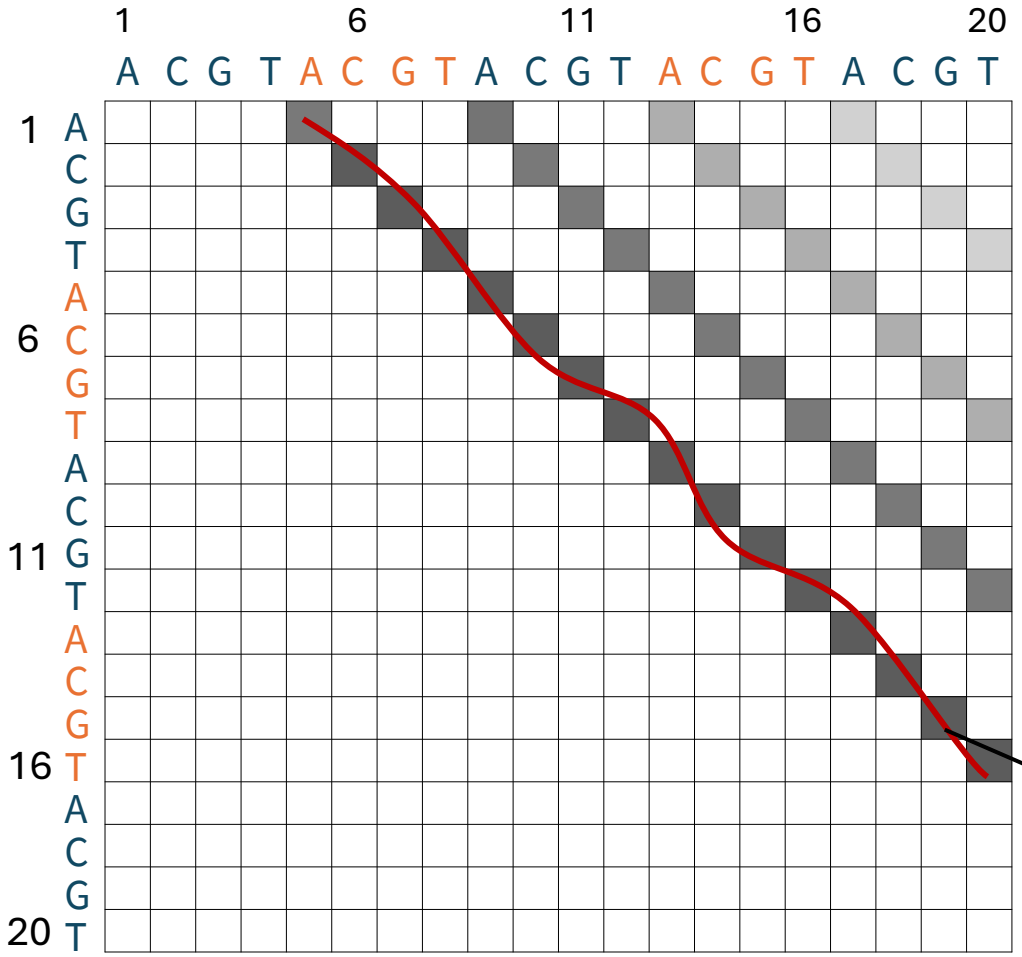


Diagonal-free Self-alignment (Ideal)

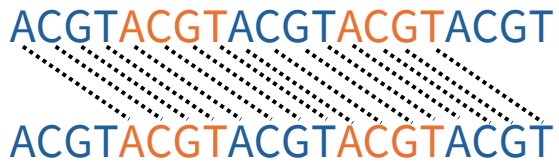


Ideal DP Table

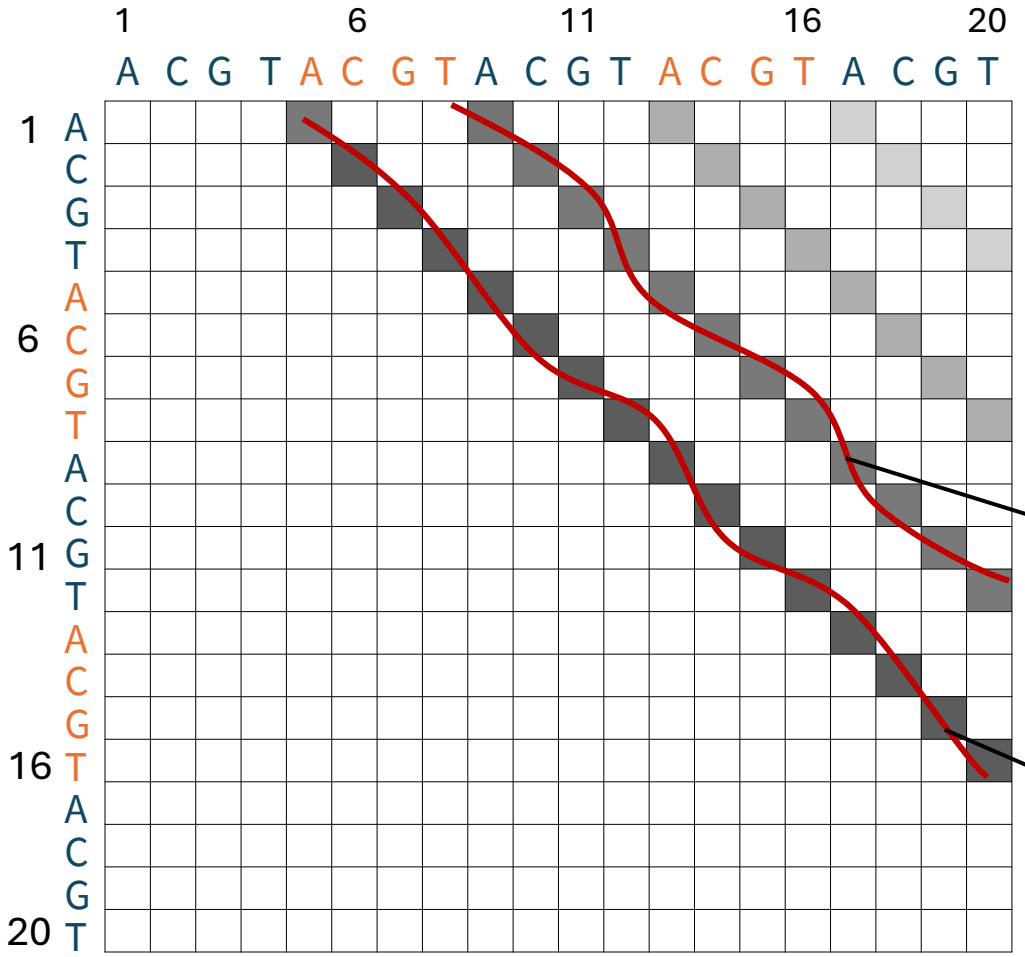
Diagonal-free Self-alignment (Ideal)



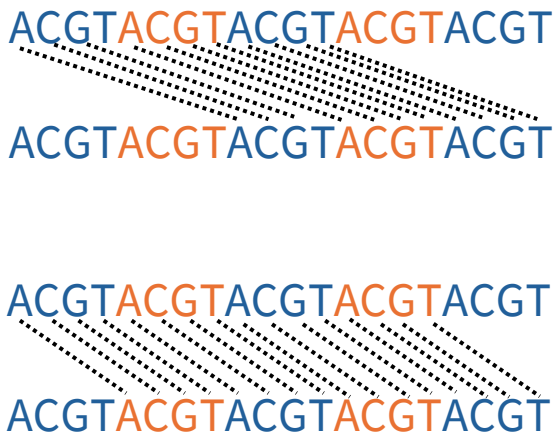
Ideal DP Table



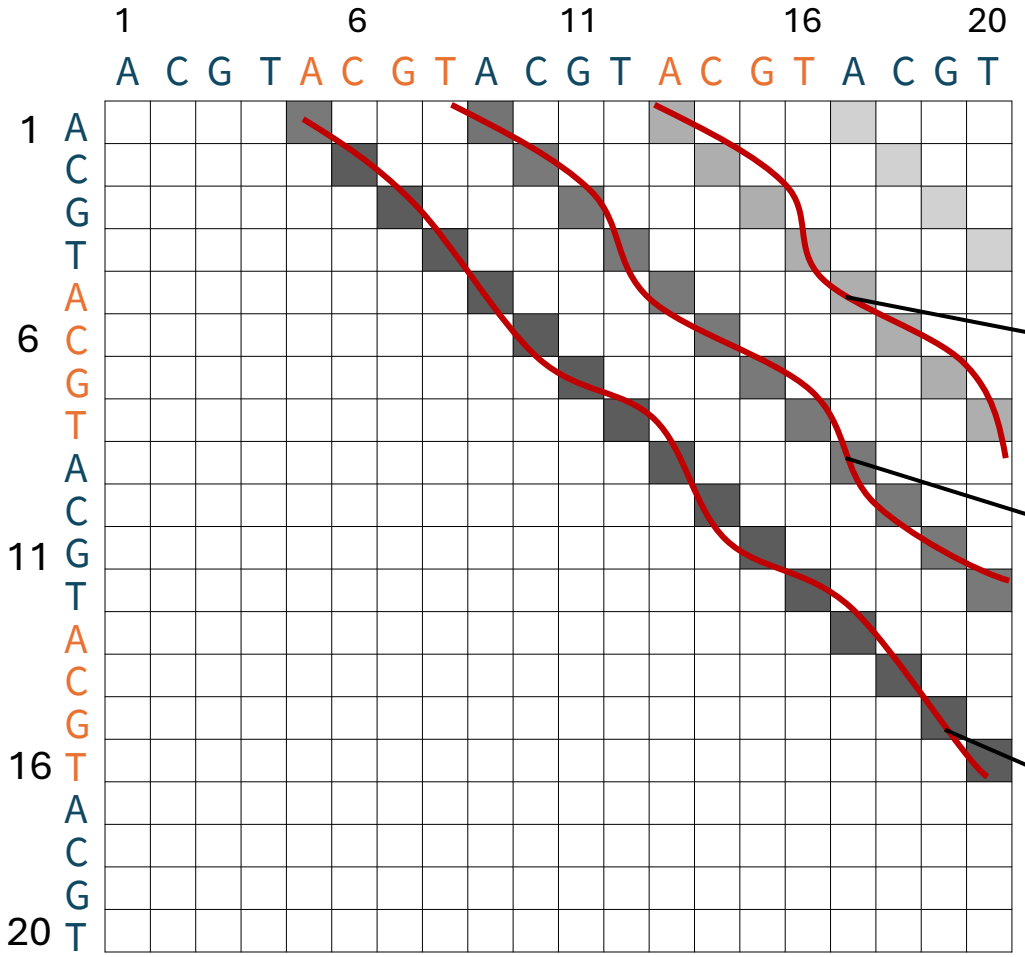
Diagonal-free Self-alignment (Ideal)



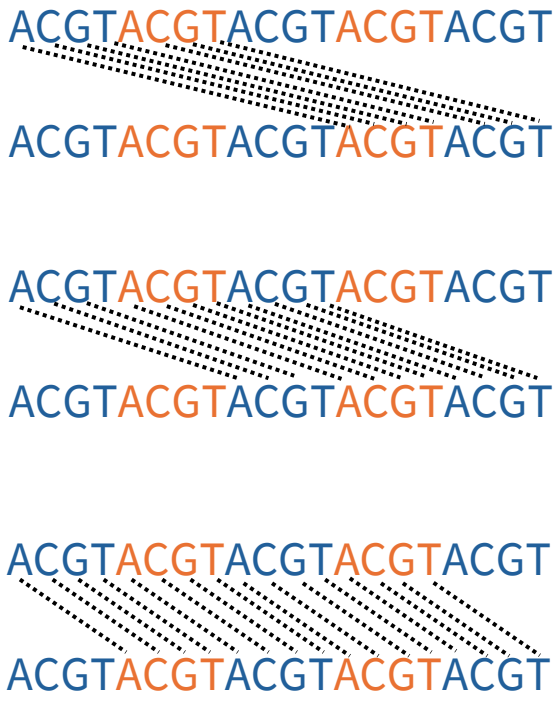
Ideal DP Table



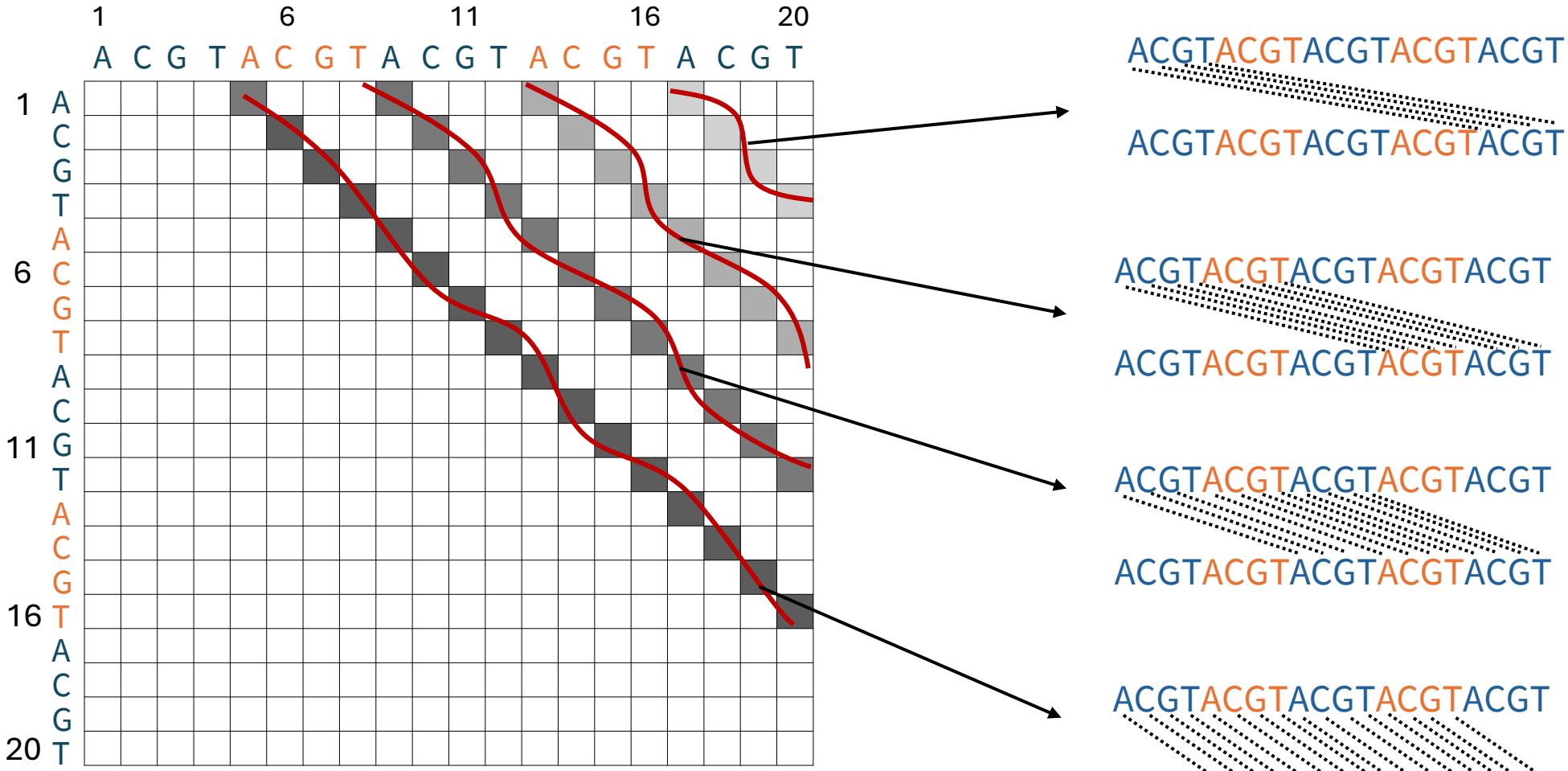
Diagonal-free Self-alignment (Ideal)



Ideal DP Table

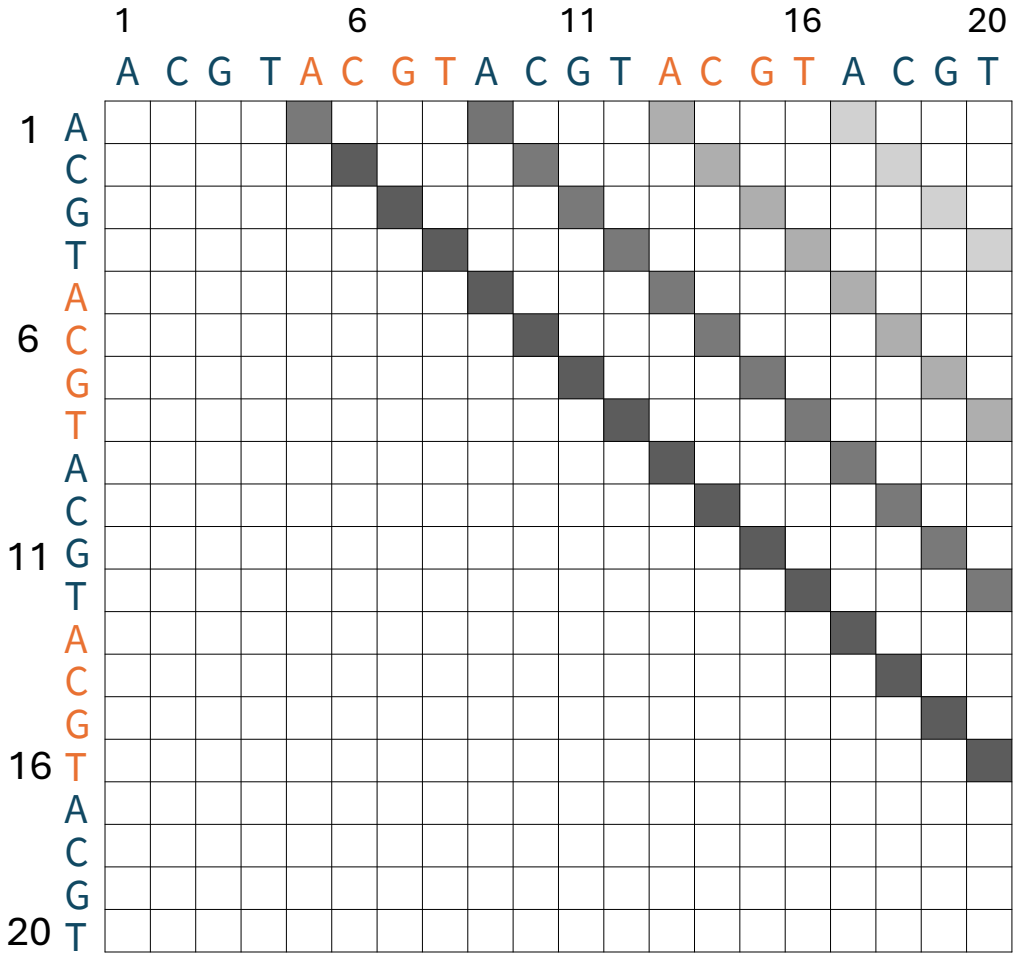


Diagonal-free Self-alignment (Ideal)



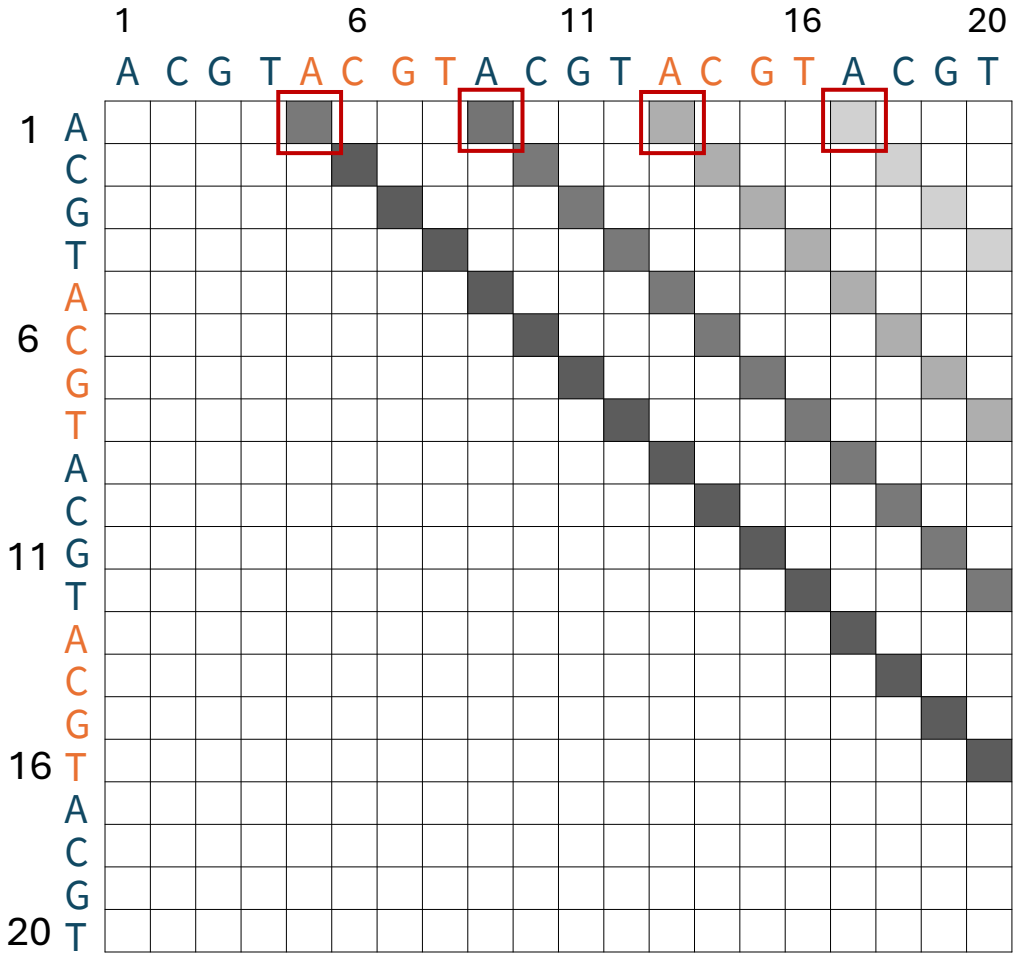
Ideal DP Table

Diagonal-free Self-alignment (Ideal)



Ideal DP Table

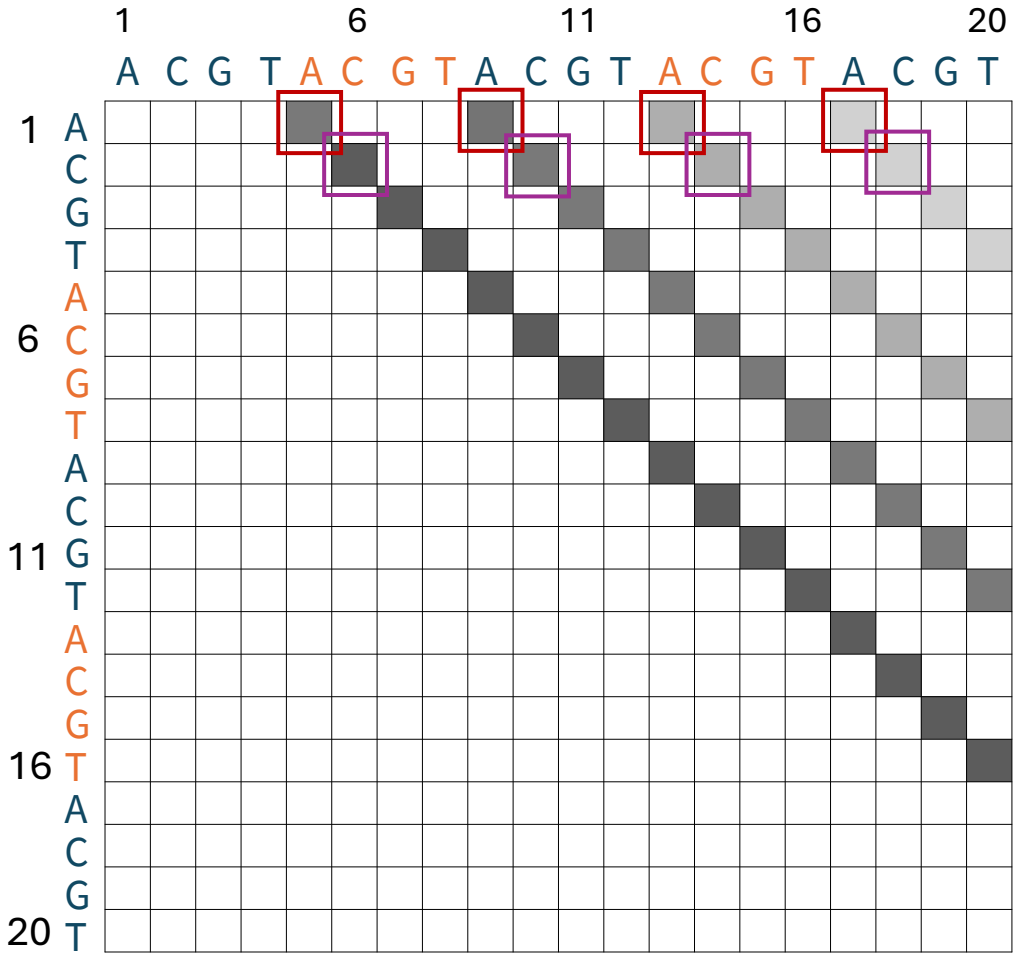
Diagonal-free Self-alignment (Ideal)



Ideal DP Table

Class 1: {1,5,9,13,17}

Diagonal-free Self-alignment (Ideal)

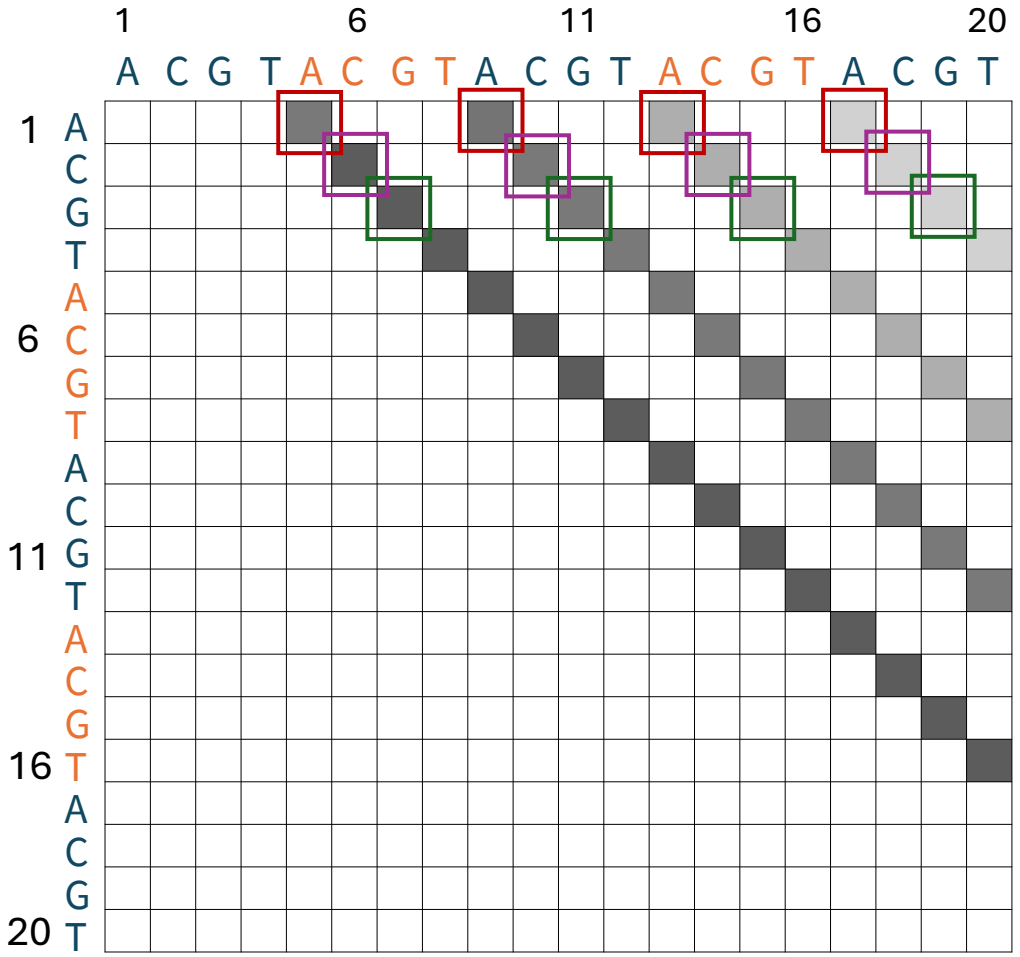


Ideal DP Table

Class 1: {1,5,9,13,17}

Class 2: {2,6,10,14,18}

Diagonal-free Self-alignment (Ideal)



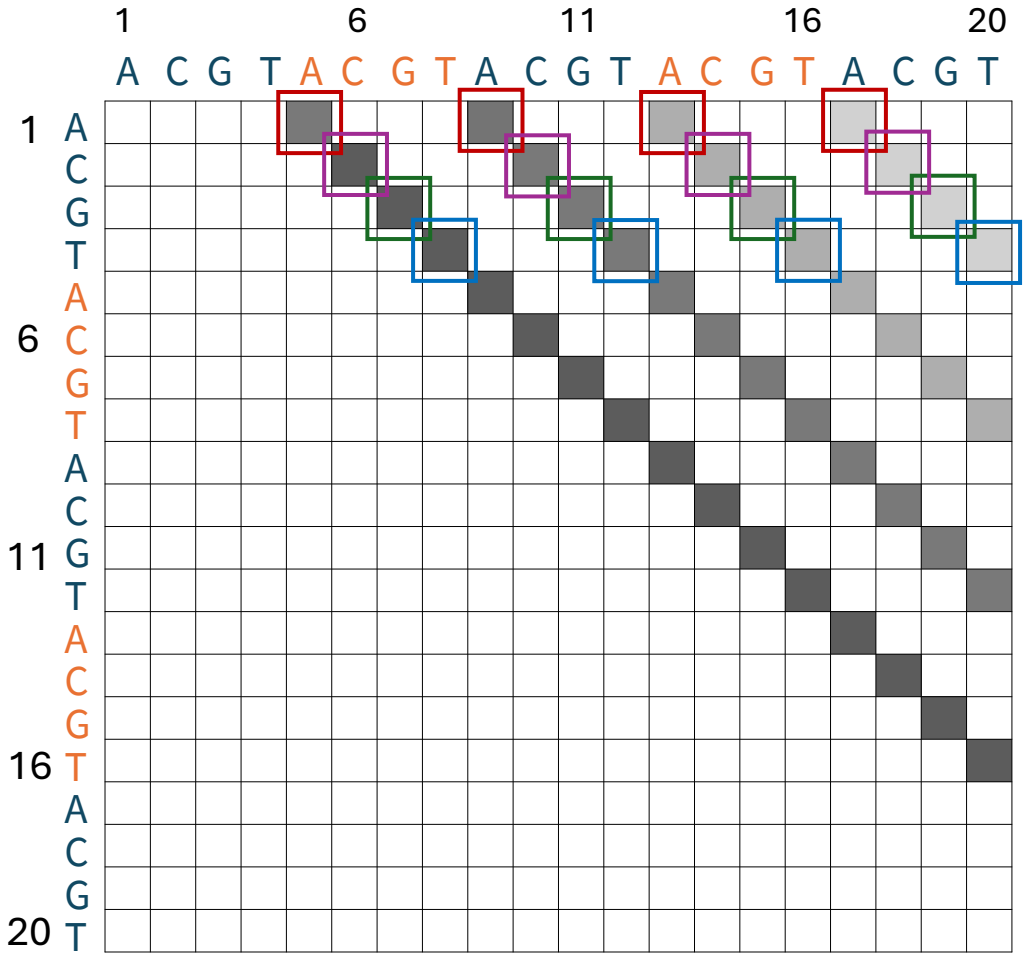
Ideal DP Table

Class 1: {1,5,9,13,17}

Class 2: {2,6,10,14,18}

Class 3: {3,7,11,15,19}

Diagonal-free Self-alignment (Ideal)



Ideal DP Table

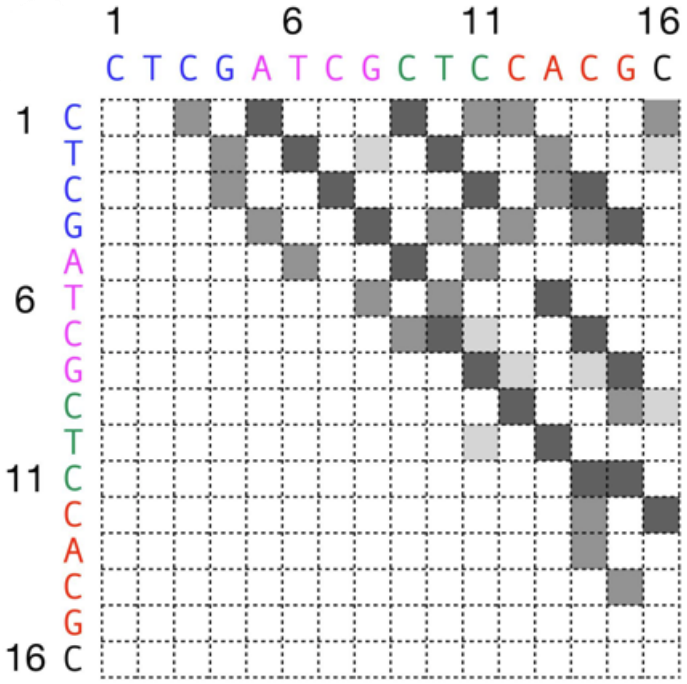
Class 1: {1,5,9,13,17}

Class 2: {2,6,10,14,18}

Class 3: {3,7,11,15,19}

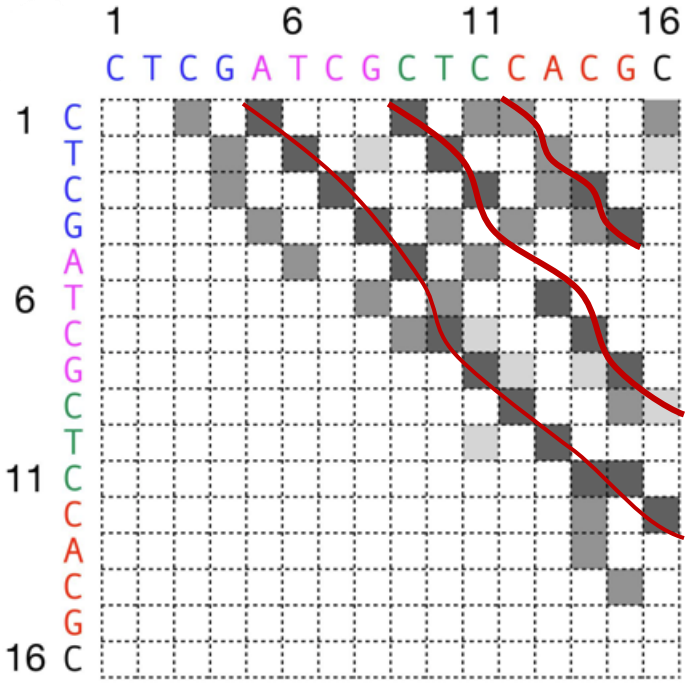
Class 4: {4,8,12,16,20}

Diagonal-free Self-alignment (Actual)



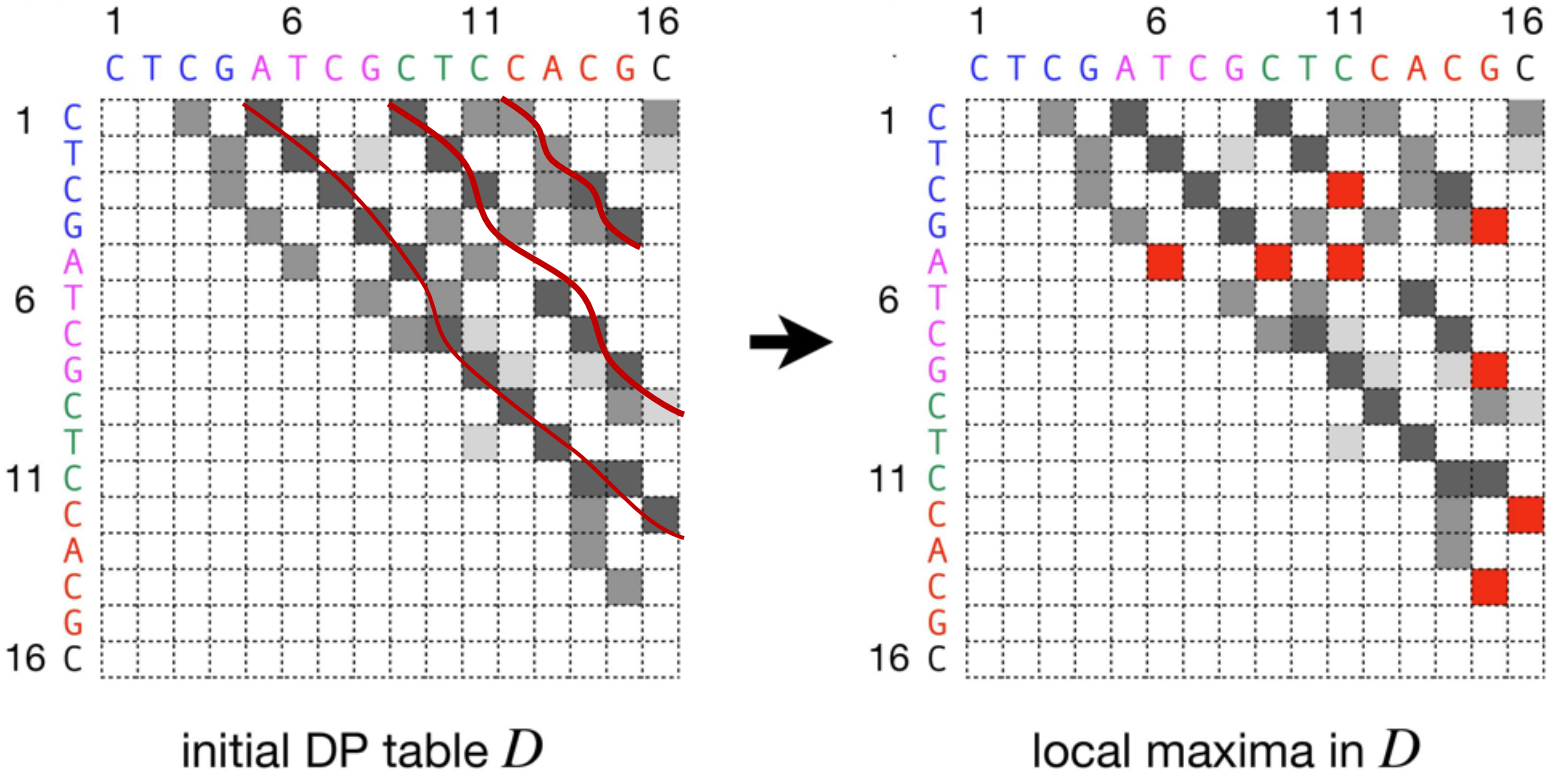
initial DP table D

Diagonal-free Self-alignment (Actual)



initial DP table D

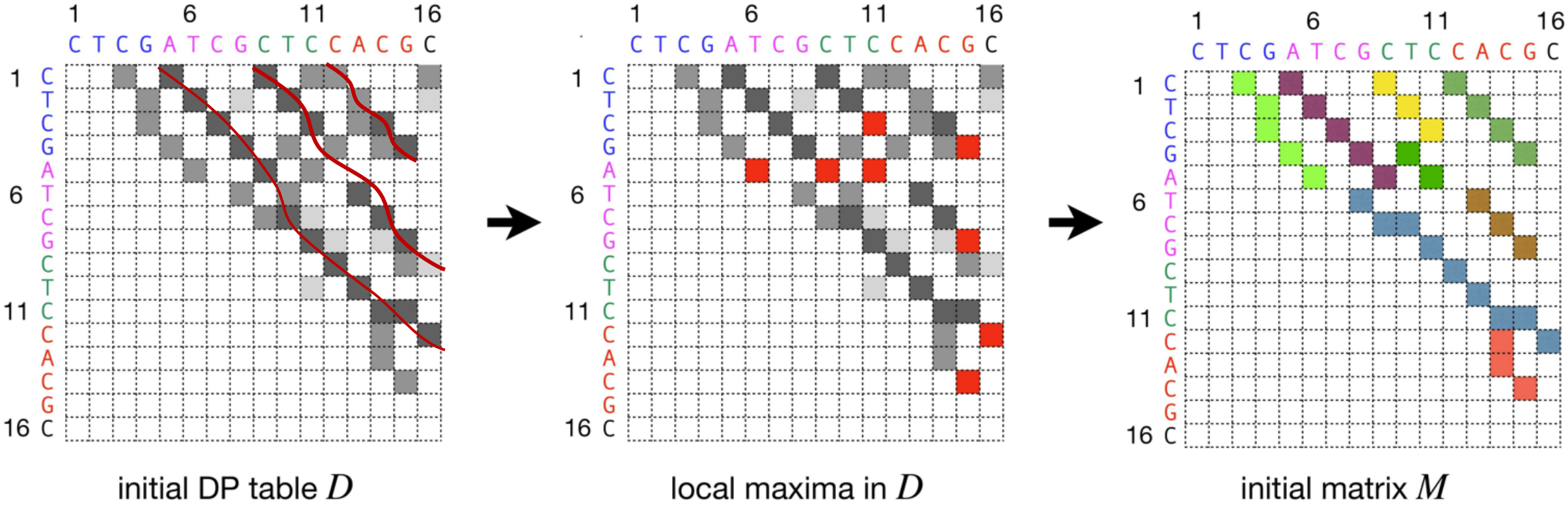
Constructing Initial Matrix M



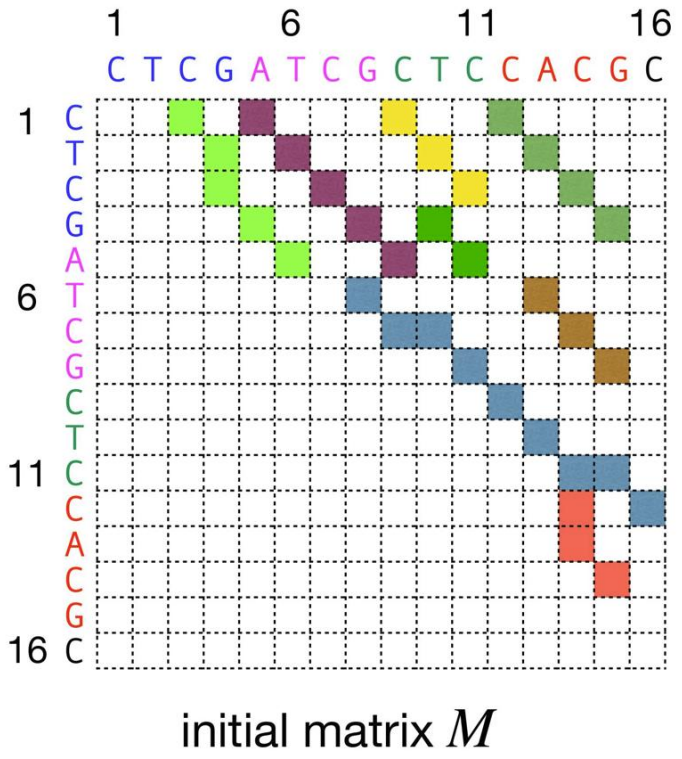
initial DP table D

local maxima in D

Constructing Initial Matrix M

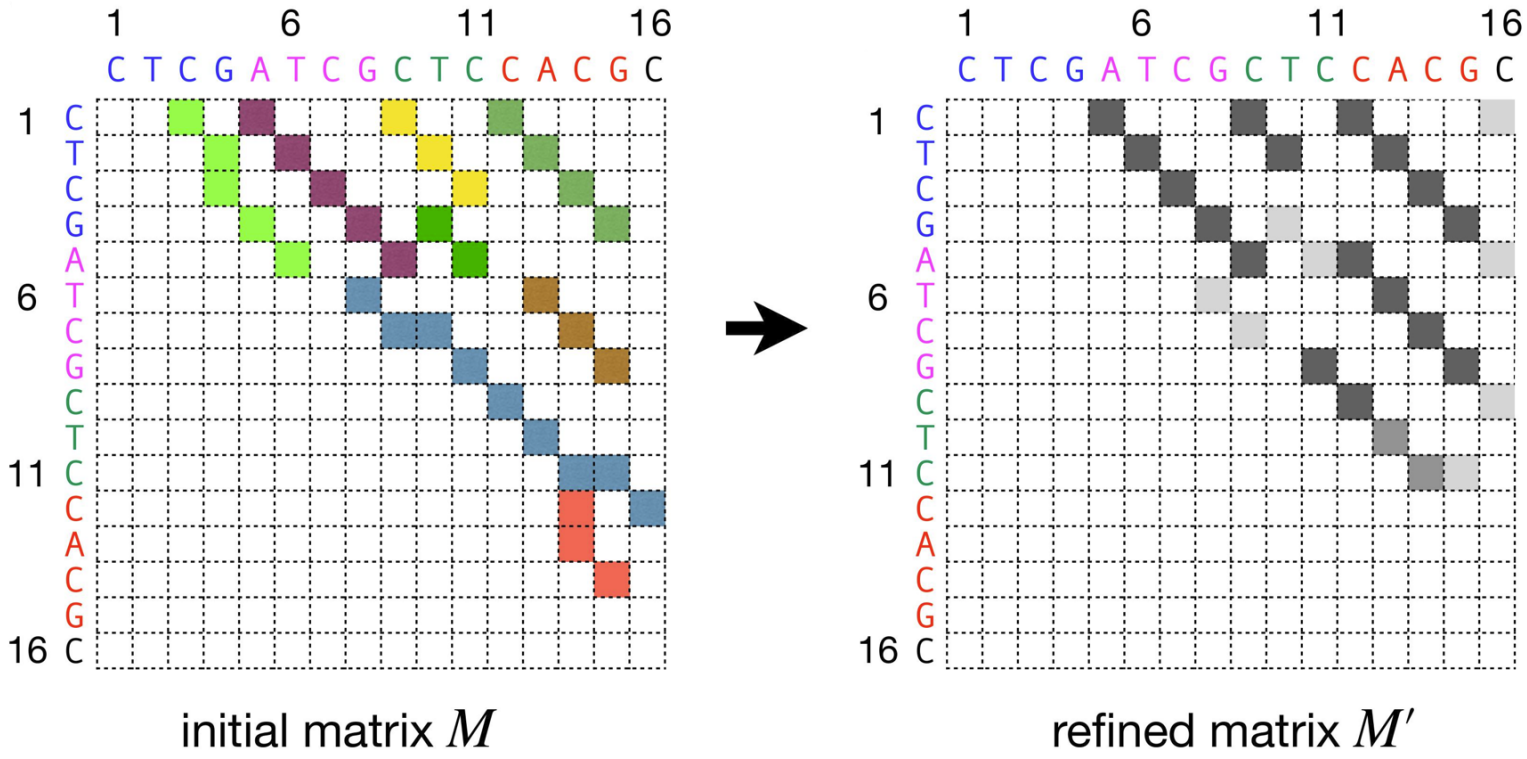


Constructing Refined Matrix M’



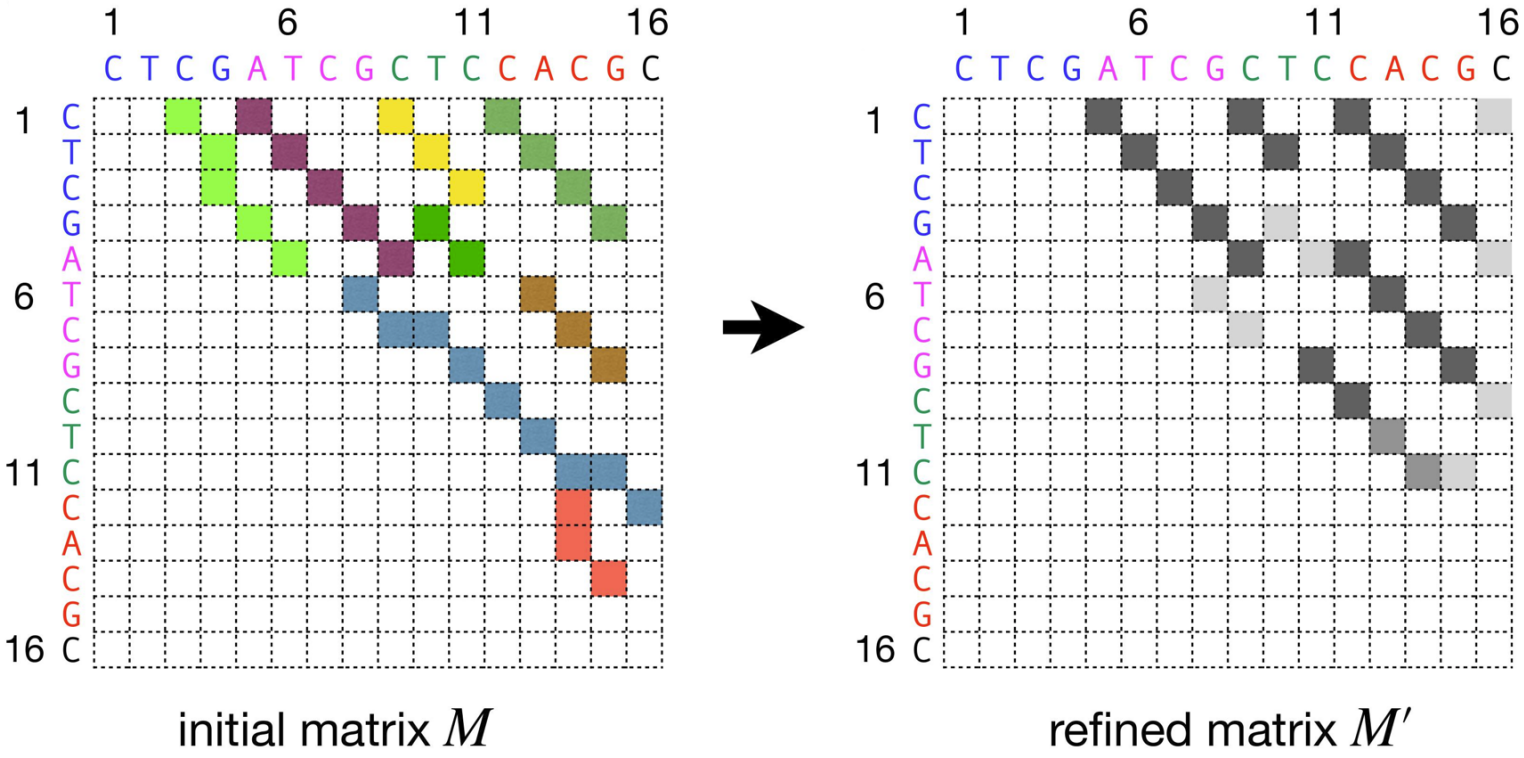
initial matrix M

Constructing Refined Matrix M’



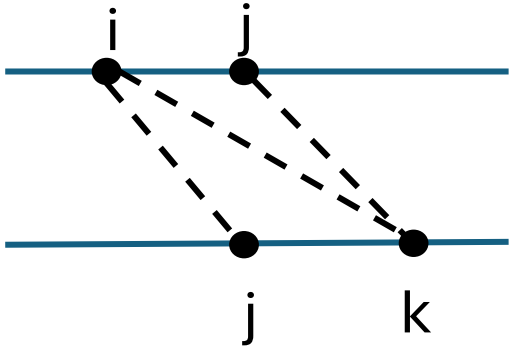
initial matrix M refined matrix M'

Constructing Refined Matrix M’



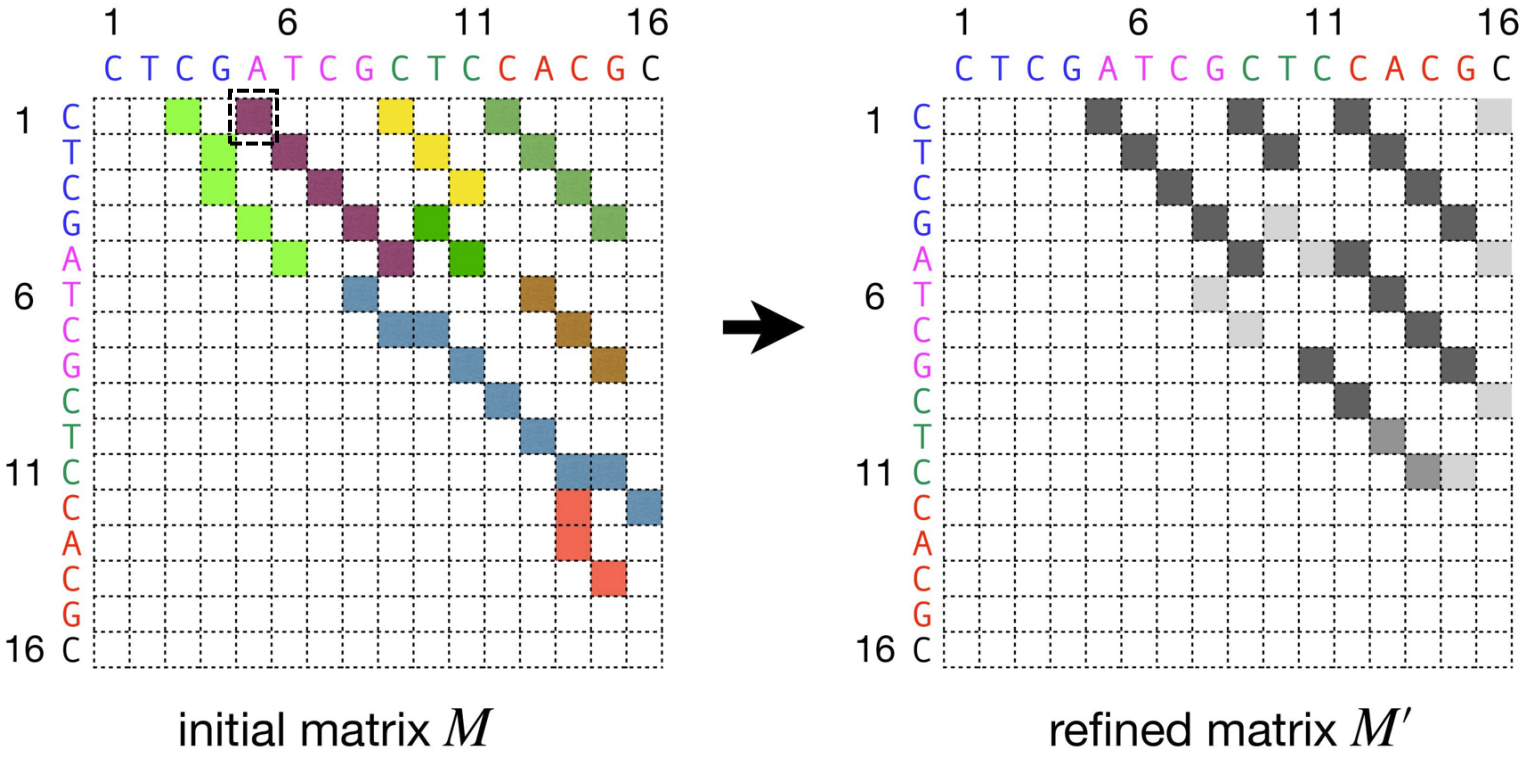
initial matrix M

refined matrix M'



$$\delta \leftarrow \min\{M[i,j], M[i,k], M[j,k]\}$$
$$M'[i,j] \leftarrow M'[i,j] + \delta$$
$$M'[j,k] \leftarrow M'[j,k] + \delta$$
$$M'[i,k] \leftarrow M'[i,k] + \delta$$

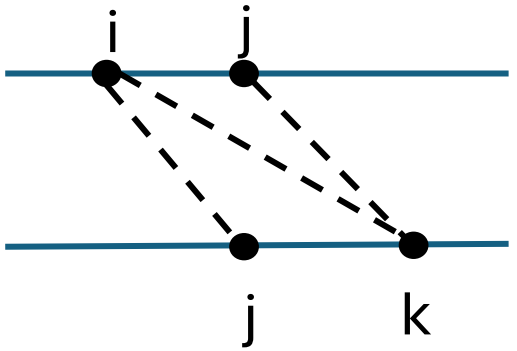
Constructing Refined Matrix M’



initial matrix M

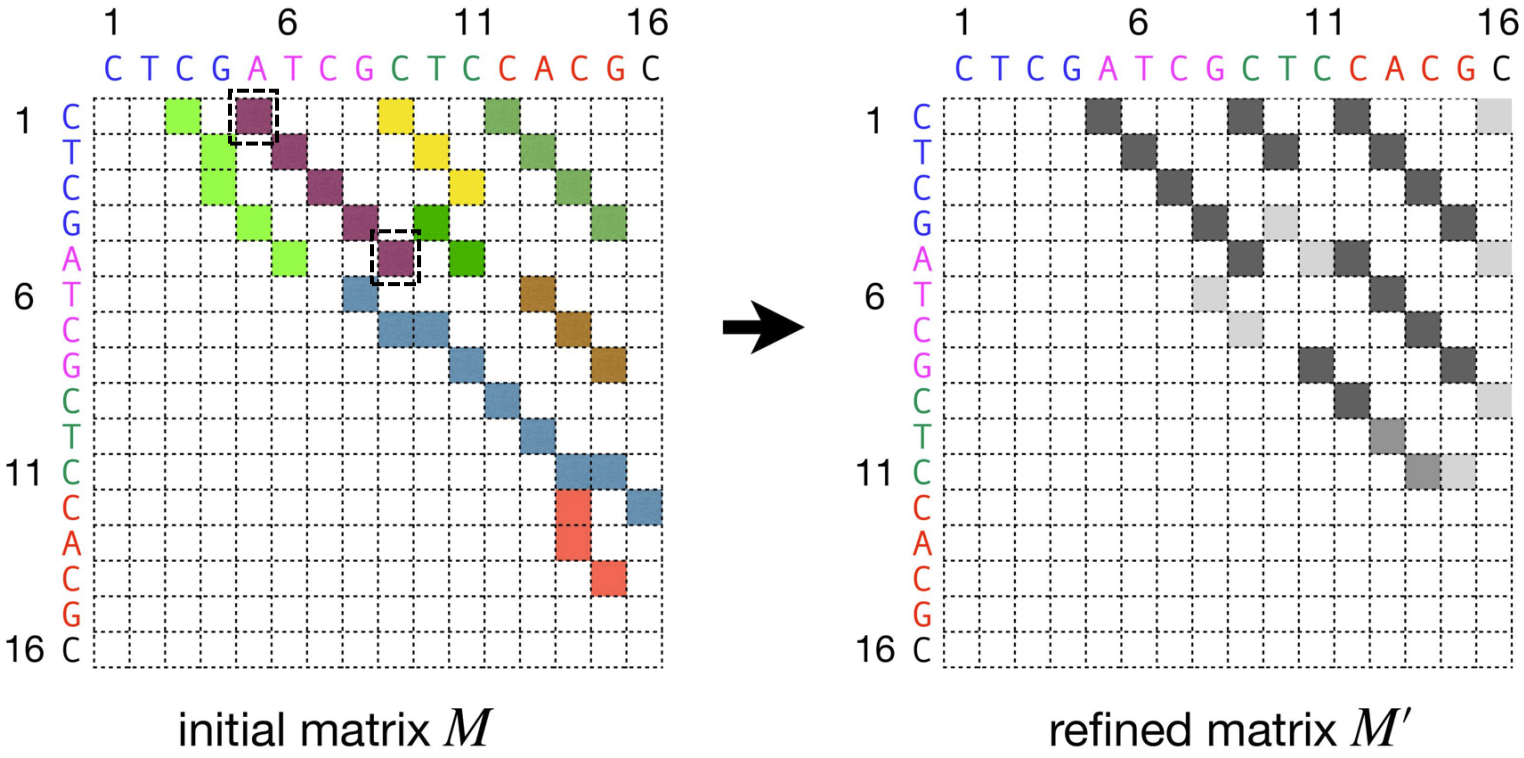
refined matrix M'

$M\,[1, 5] > 0$



$$\delta \leftarrow \min\{M\,[i, j], M\,[i, k], M\,[j, k]\}$$
$$M'\,[i, j] \leftarrow M'\,[i, j] + \delta$$
$$M'\,[j, k] \leftarrow M'\,[j, k] + \delta$$
$$M'\,[i, k] \leftarrow M'\,[i, k] + \delta$$

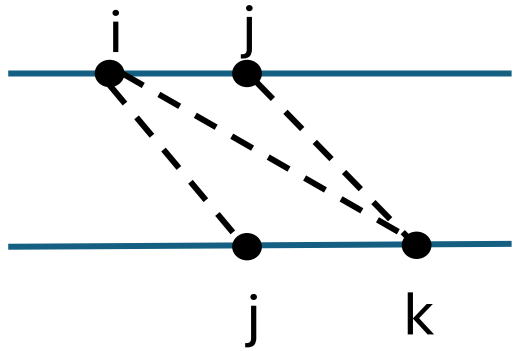
Constructing Refined Matrix M’



initial matrix M

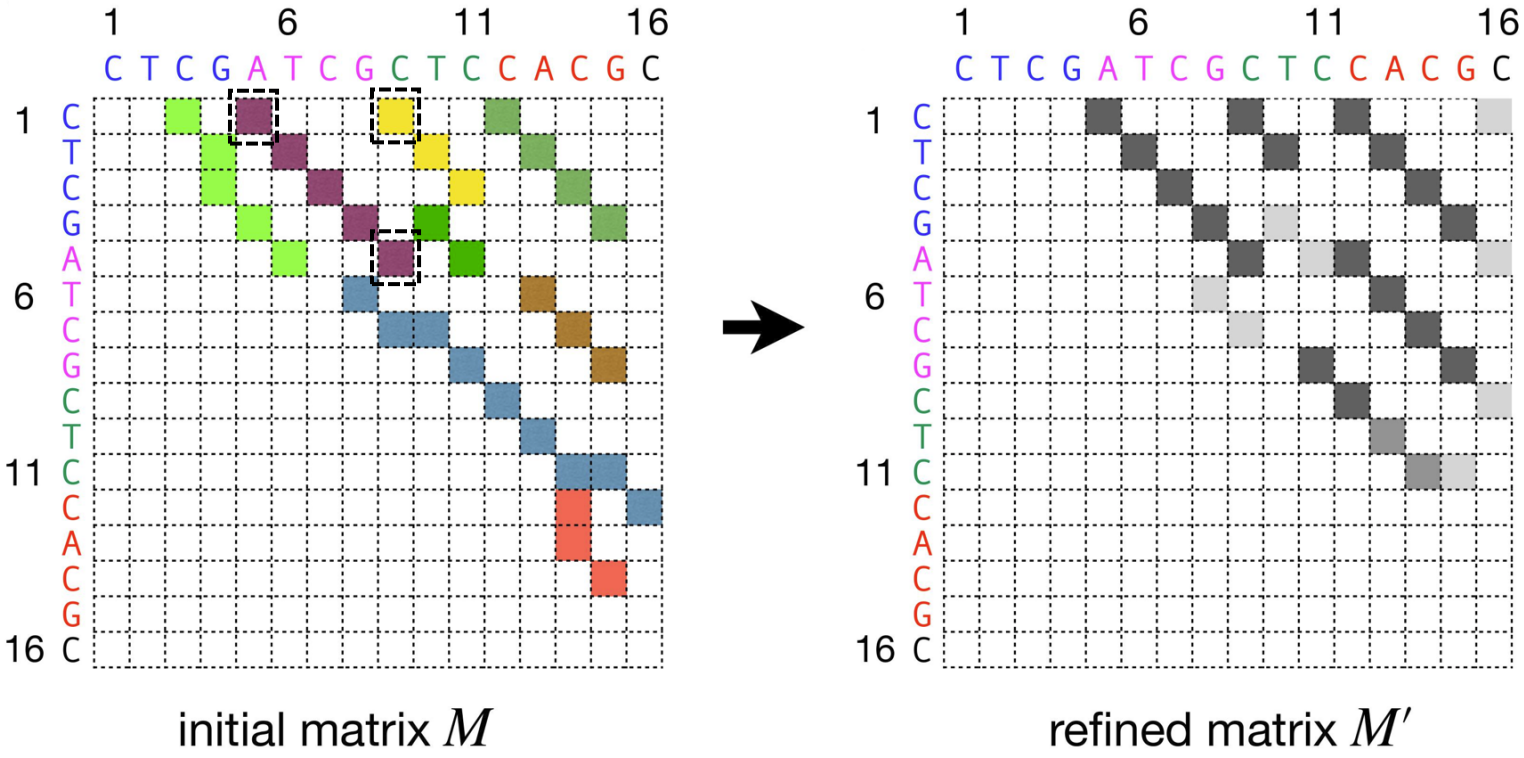
refined matrix M'

$$M\,[1,5] > 0$$
$$M\,[5,9] > 0$$



$$\delta \leftarrow \min\{M\,[i,j], M\,[i,k], M\,[j,k]\}$$
$$M'\,[i,j] \leftarrow M'\,[i,j] + \delta$$
$$M'\,[j,k] \leftarrow M'\,[j,k] + \delta$$
$$M'\,[i,k] \leftarrow M'\,[i,k] + \delta$$

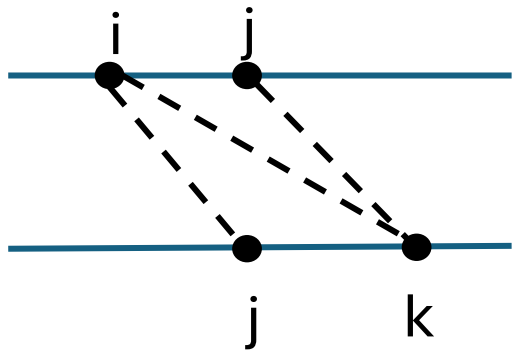
Constructing Refined Matrix M’



initial matrix M

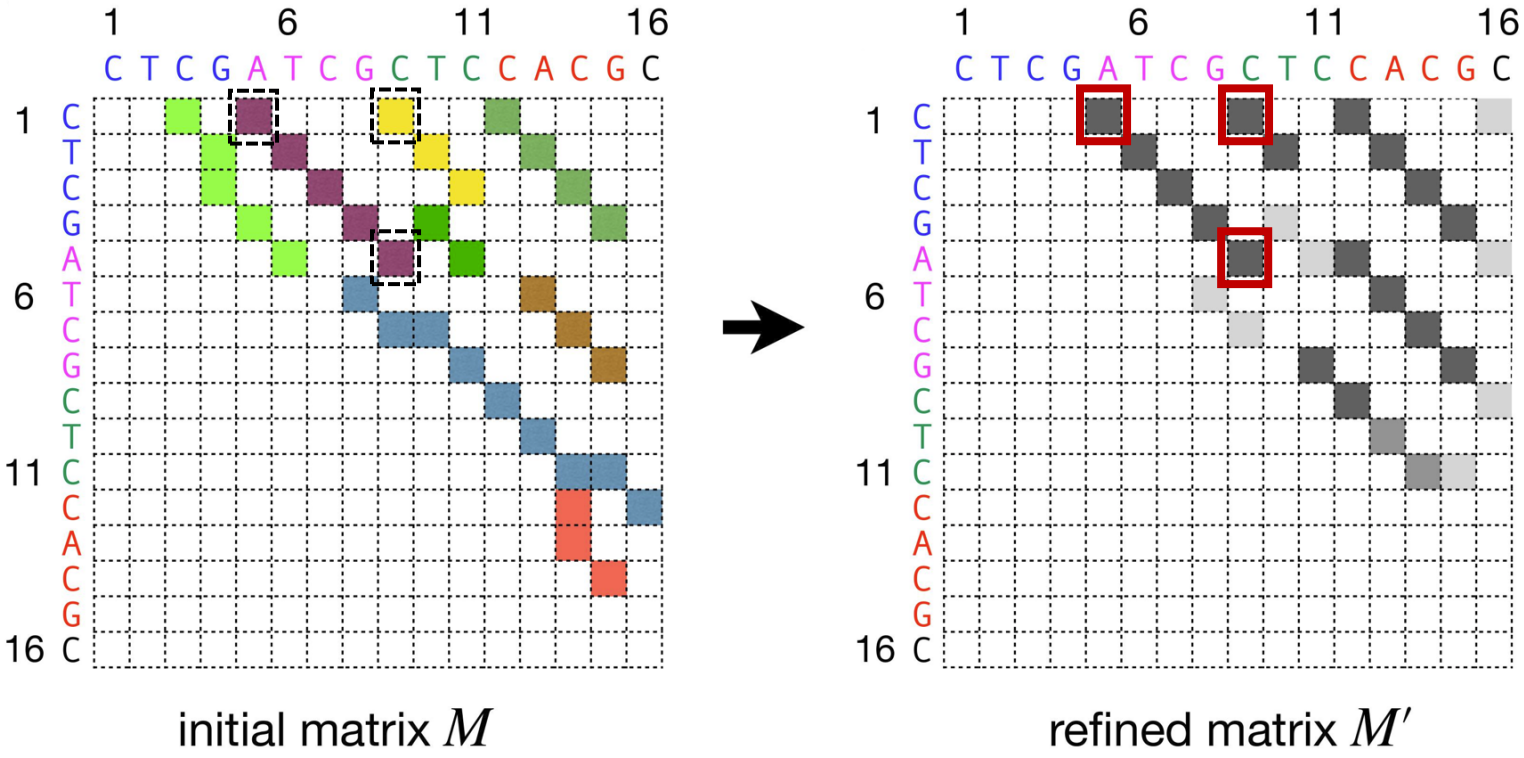
refined matrix M'

$$M\,[1,5] > 0$$
$$M\,[5,9] > 0$$
$$M\,[1,9] > 0$$



$$\delta \leftarrow \min\{M\,[i,j], M\,[i,k], M\,[j,k]\}$$
$$M'\,[i,j] \leftarrow M'\,[i,j] + \delta$$
$$M'\,[j,k] \leftarrow M'\,[j,k] + \delta$$
$$M'\,[i,k] \leftarrow M'\,[i,k] + \delta$$

Constructing Refined Matrix M’

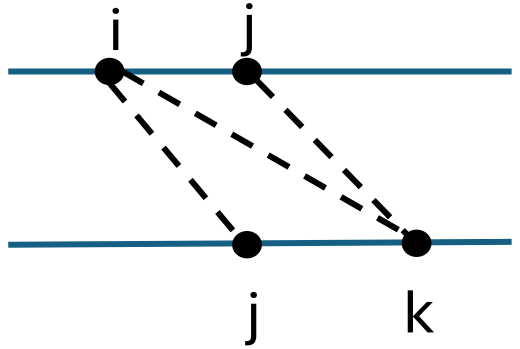


initial matrix M

refined matrix M'

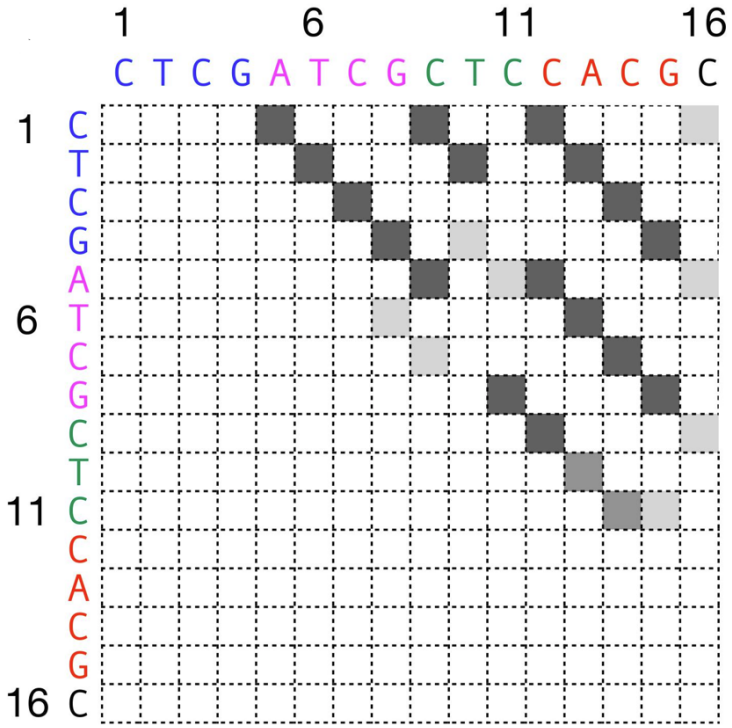
$$M\,[1,5] > 0$$
$$M\,[5,9] > 0$$
$$M\,[1,9] > 0$$

$$M'\,[1,5] \leftarrow M'\,[1,5] + \delta$$
$$M'\,[5,9] \leftarrow M'\,[5,9] + \delta$$
$$M'\,[1,9] \leftarrow M'\,[1,9] + \delta$$



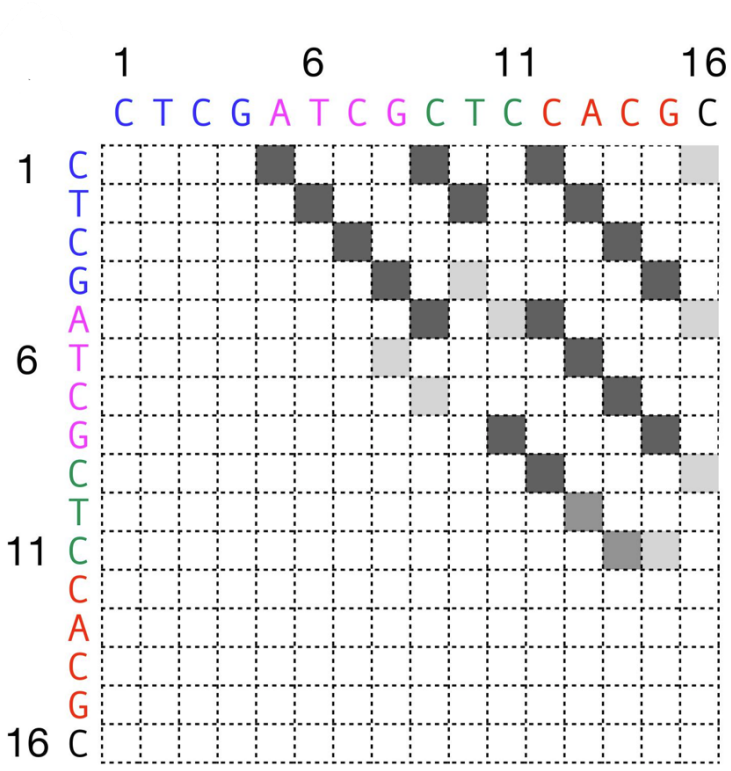
$$\delta \leftarrow \min\{M\,[i,j], M\,[i,k], M\,[j,k]\}$$
$$M'\,[i,j] \leftarrow M'\,[i,j] + \delta$$
$$M'\,[j,k] \leftarrow M'\,[j,k] + \delta$$
$$M'\,[i,k] \leftarrow M'\,[i,k] + \delta$$

Constructing Equivalent Positions

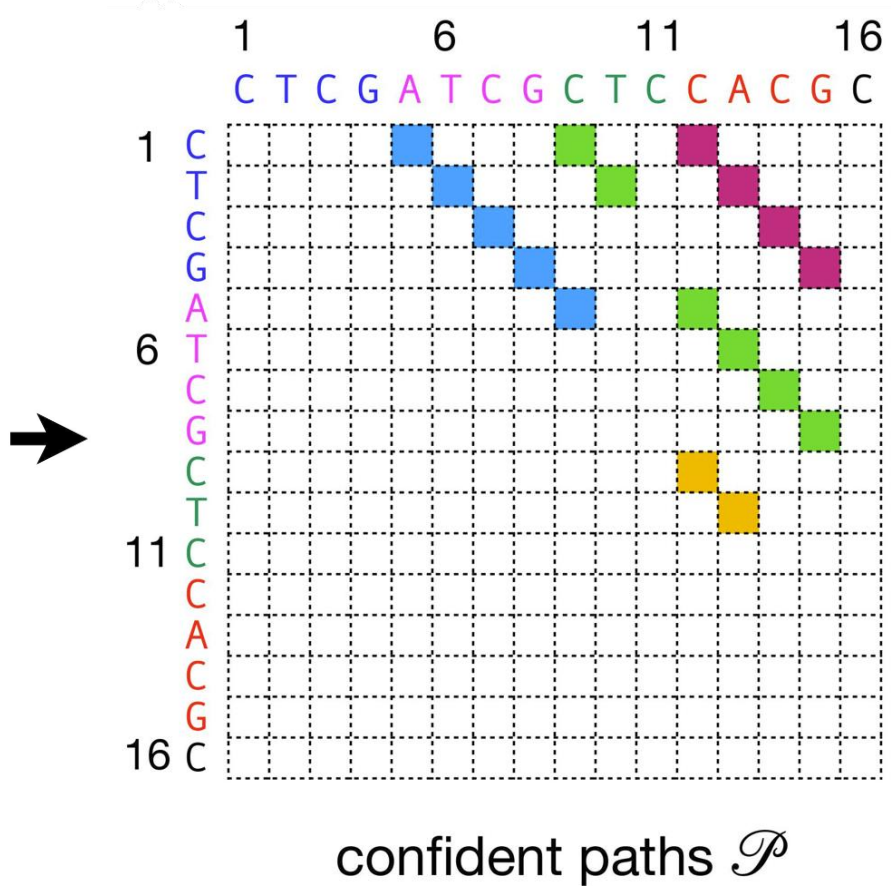


refined matrix M'

Constructing Equivalent Positions

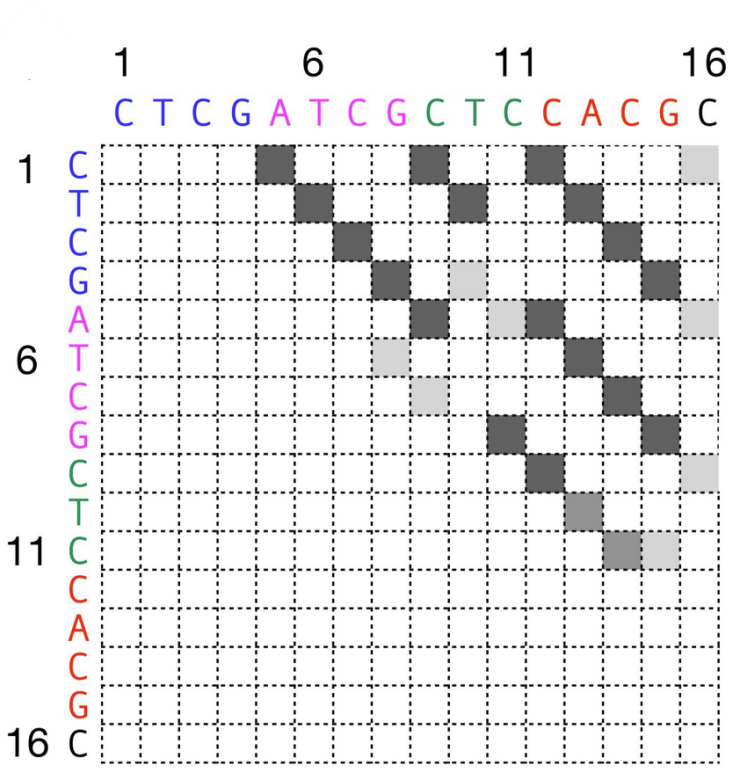


refined matrix M'

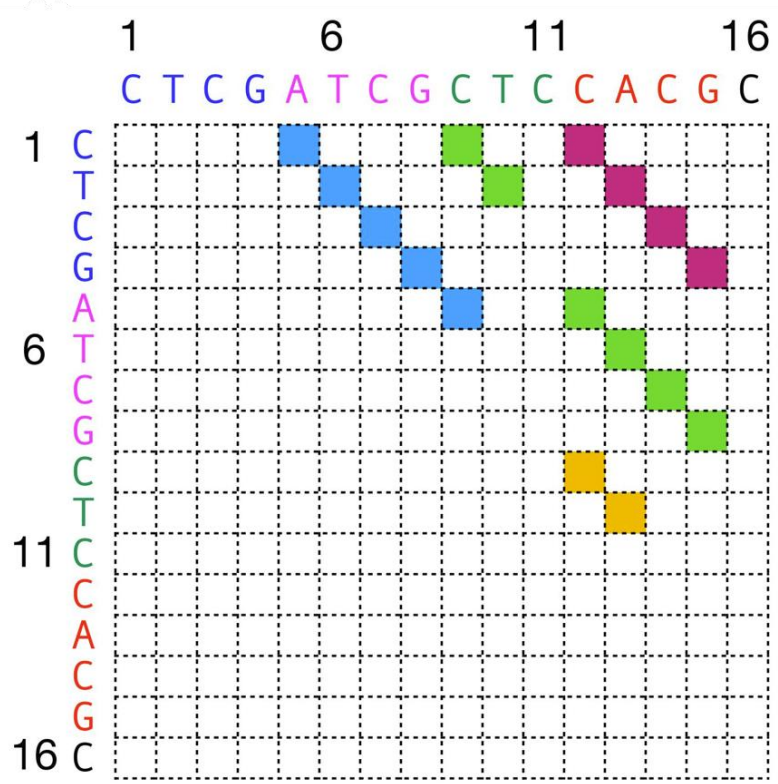


confident paths $\mathscr{P}$

Constructing Equivalent Positions



refined matrix M'



confident paths $\mathscr{P}$

1,5,9,12 (C)

2,6,10,13 (T)

3,7,14 (C)

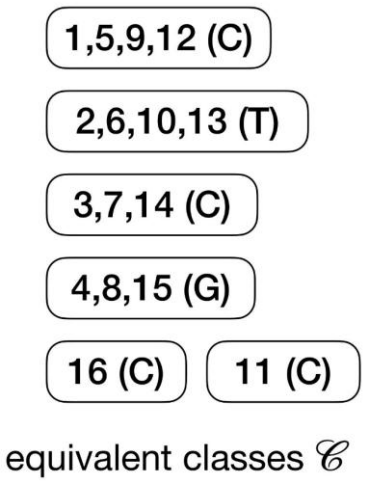
4,8,15 (G)

11 (C)

16 (C)

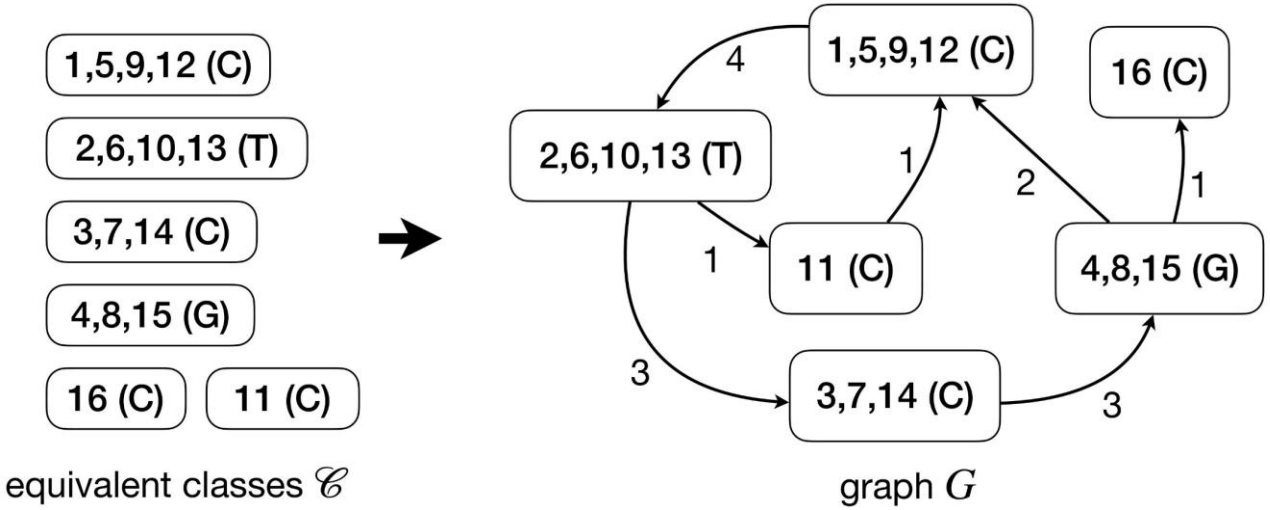
equivalent classes $\mathscr{C}$

Building Weighted Graph

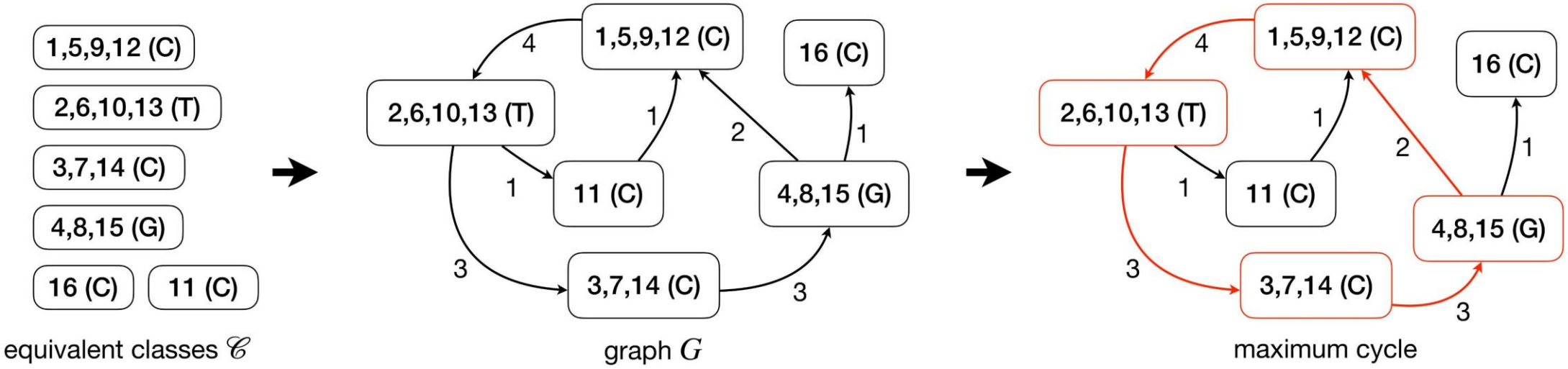


equivalent classes $\mathscr{C}$

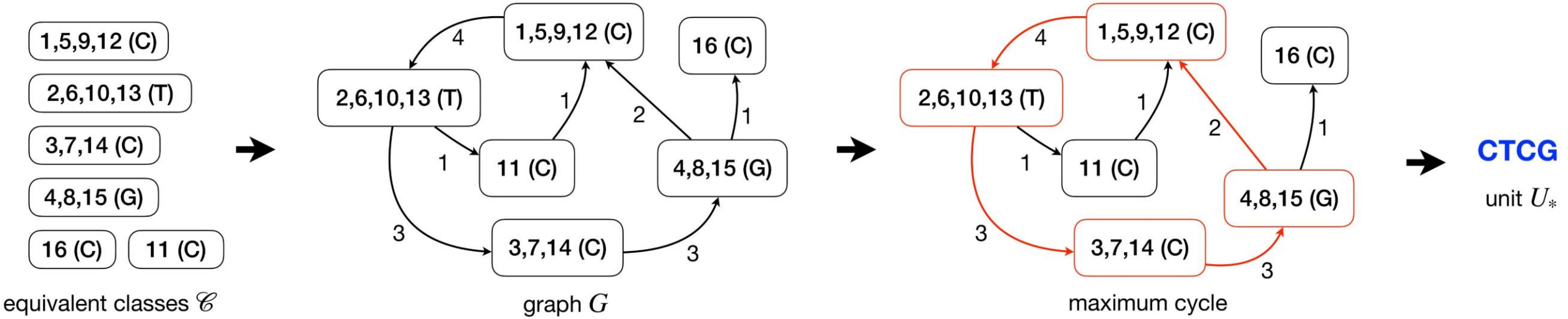
Building Weighted Graph



Building Weighted Graph



Building Weighted Graph



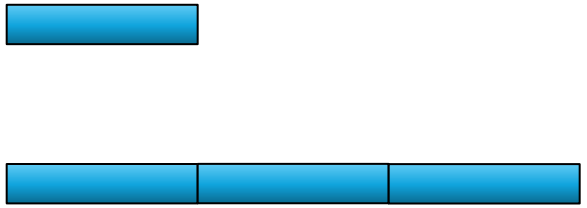
Experimental Setup: Simulations

Experimental Setup: Simulations

- Generate a random unit U of length L.

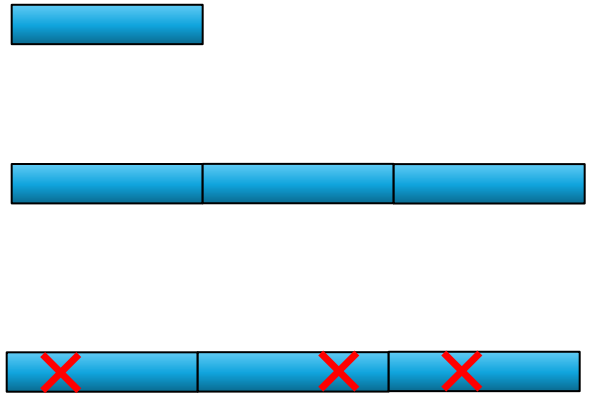
Experimental Setup: Simulations

- Generate a random unit U of length L.
- Concatenate multiple copies U with frequency F (number of copies).



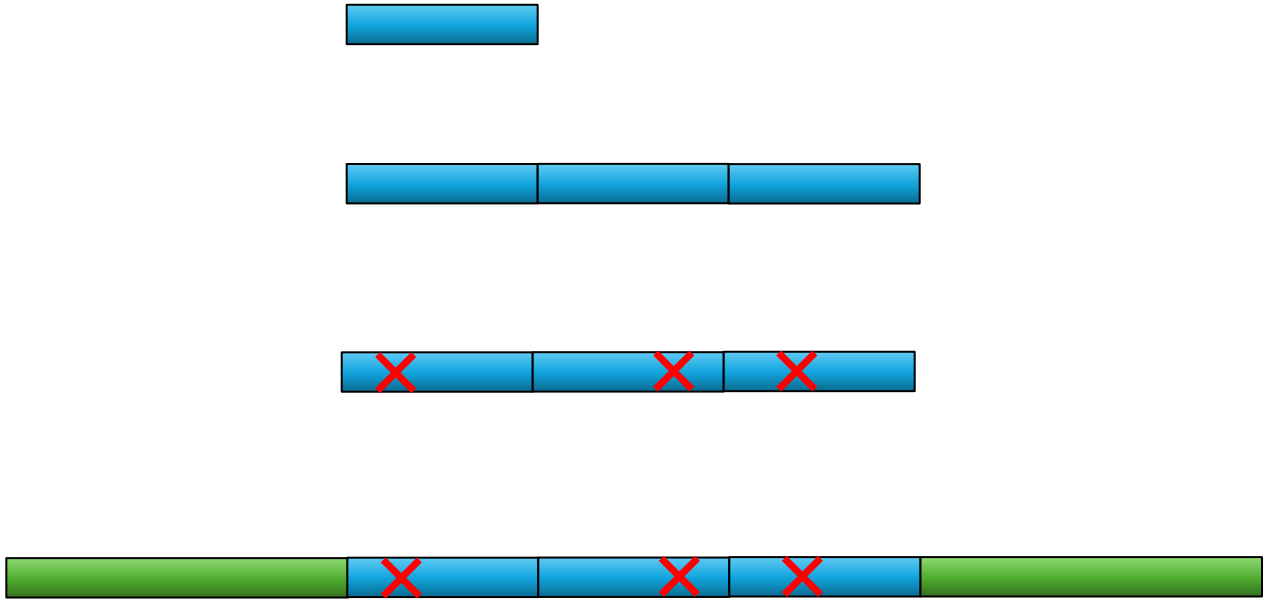
Experimental Setup: Simulations

- Generate a random unit U of length L.
- Concatenate multiple copies U with frequency F (number of copies).
- Introduce random errors: insertions, deletions, and substitutions at equal probabilities at rate R.



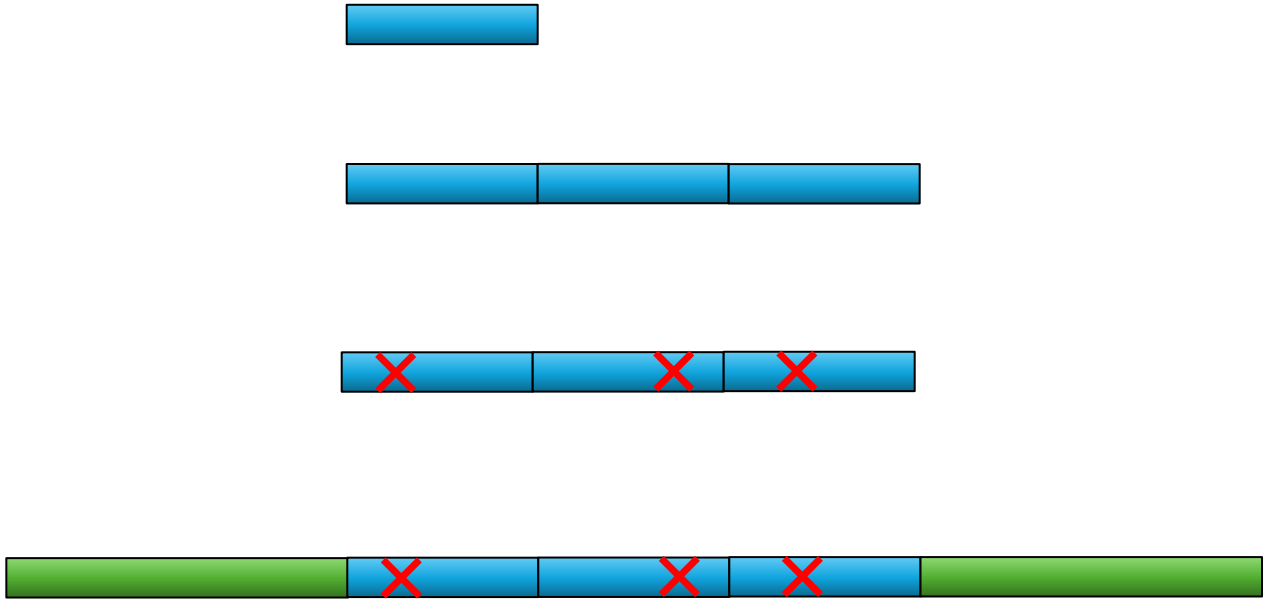
Experimental Setup: Simulations

- Generate a random unit U of length L.
- Concatenate multiple copies U with frequency F (number of copies).
- Introduce random errors: insertions, deletions, and substitutions at equal probabilities at rate R.
- Insert random strings at both sides of the concatenated string.



Experimental Setup: Simulations

- Generate a random unit U of length L.
- Concatenate multiple copies U with frequency F (number of copies).
- Introduce random errors: insertions, deletions, and substitutions at equal probabilities at rate R.
- Insert random strings at both sides of the concatenated string.
- Generate data for different combinations of L, F, R.



Evaluation

Evaluation

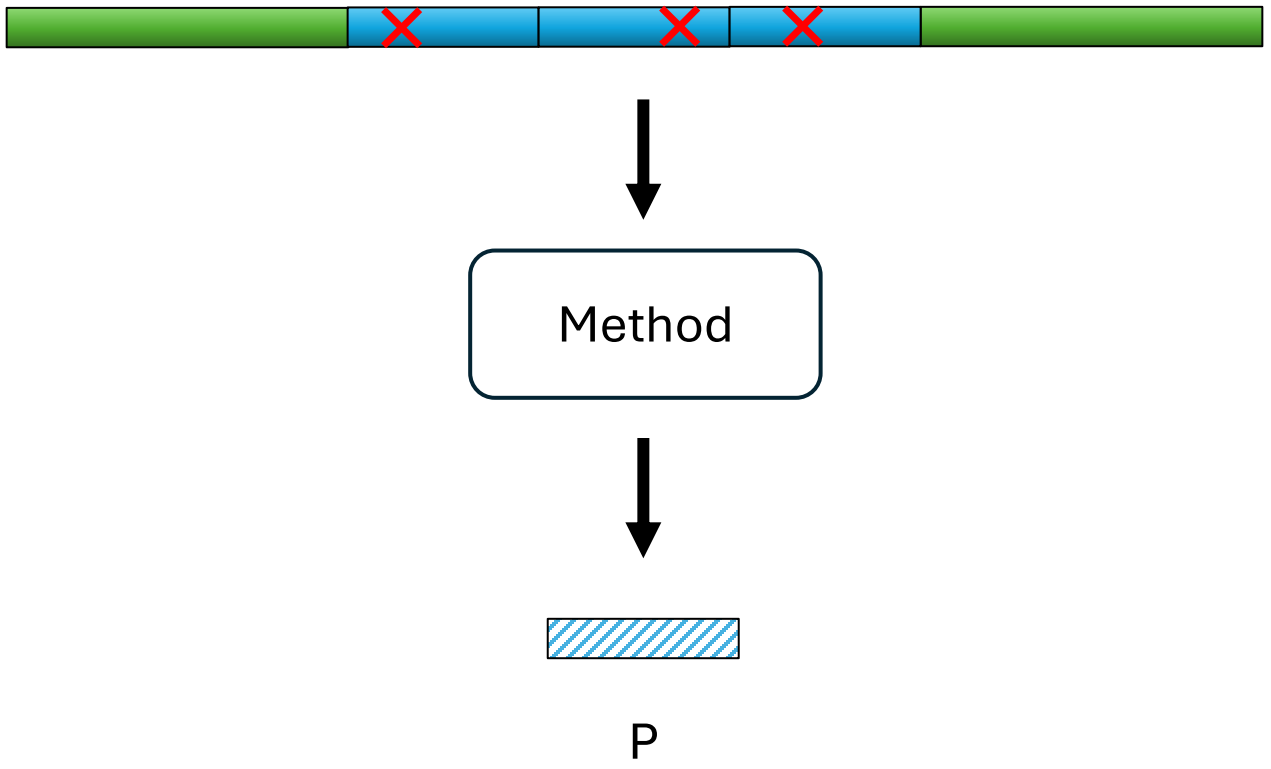
- **Normalized rotation-aware edit distance:**

Evaluation

- **Normalized rotation-aware edit distance:**

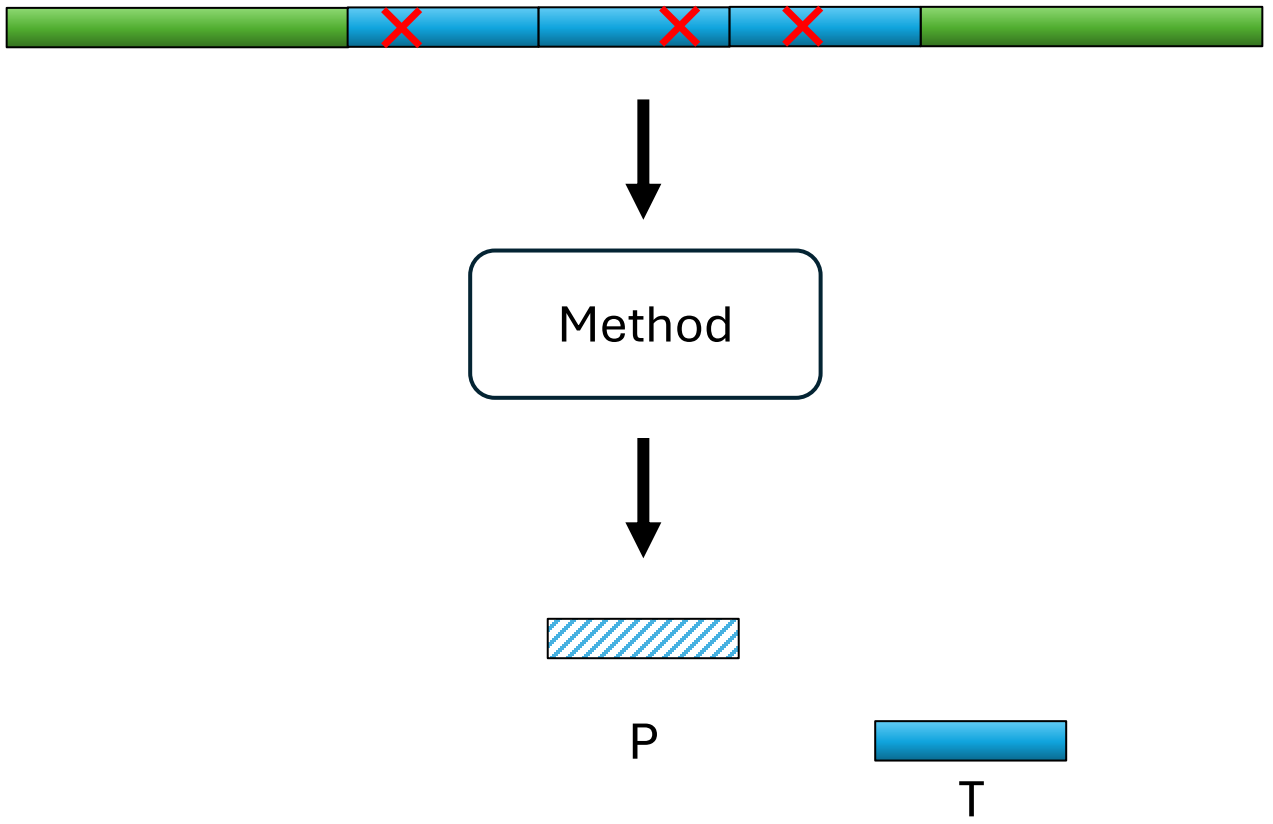
Evaluation

- **Normalized rotation-aware edit distance:**
  - Let P be the predicted unit.



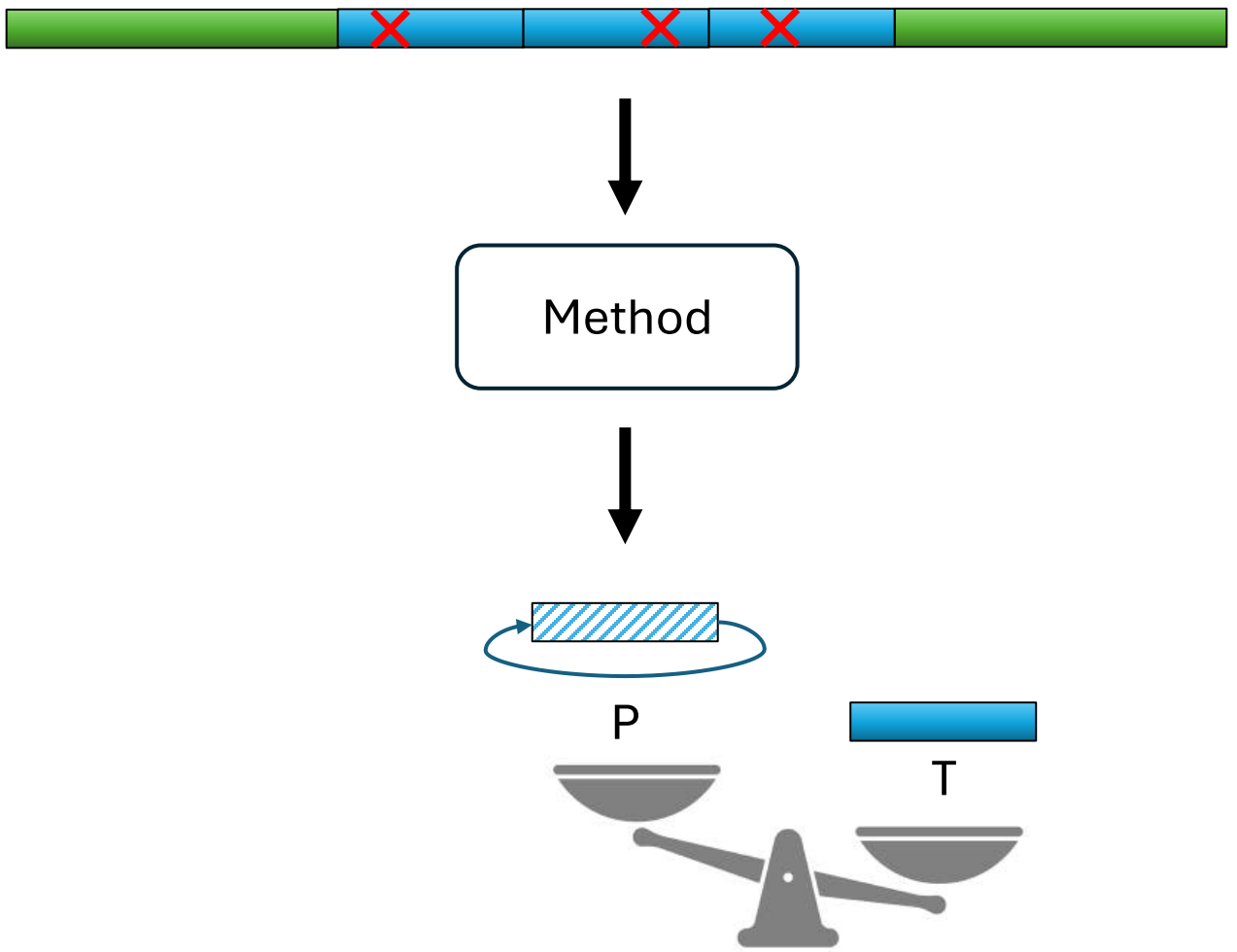
Evaluation

- **Normalized rotation-aware edit distance:**
  - Let P be the predicted unit.
  - Let T be a ground truth repeat unit.



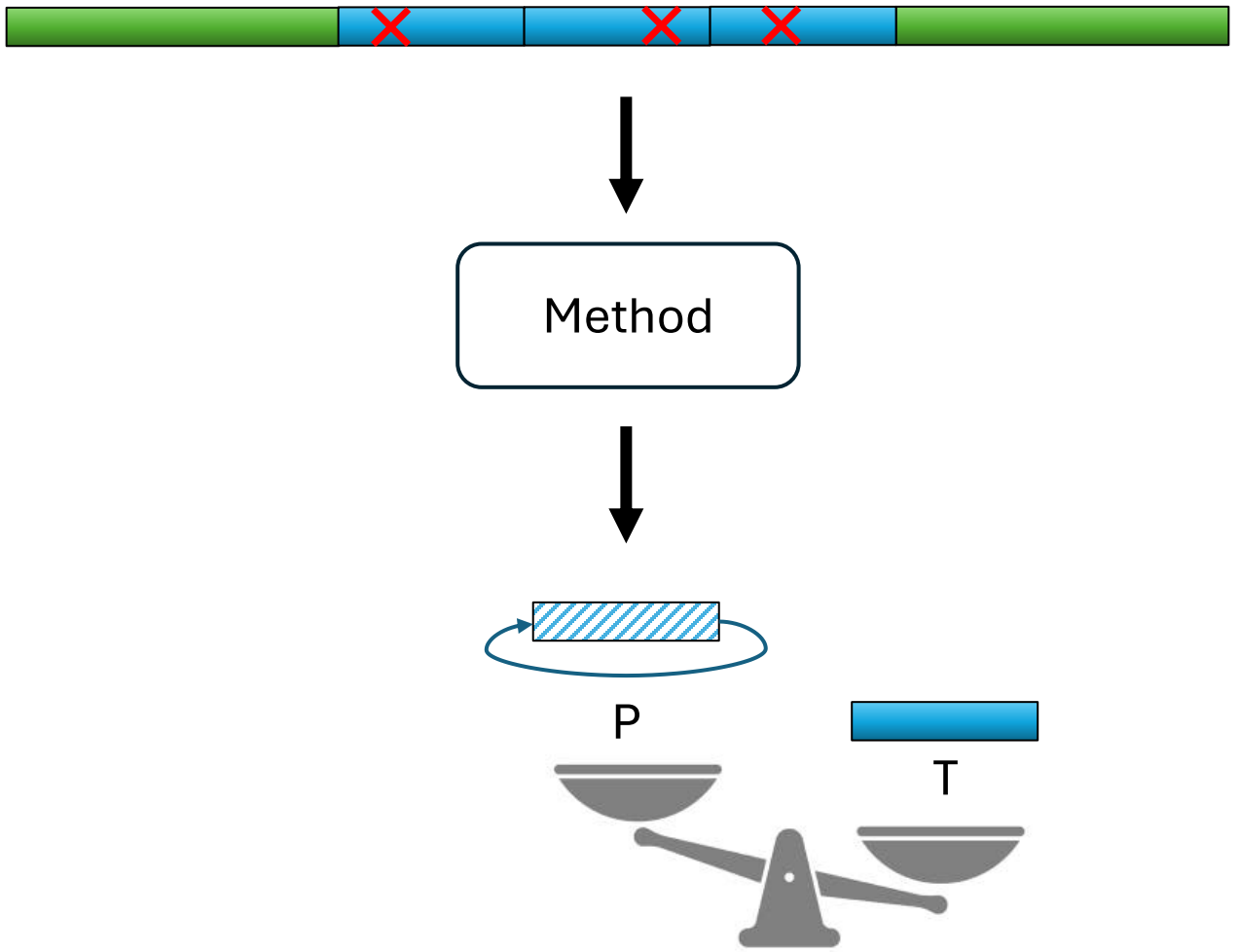
Evaluation

- **Normalized rotation-aware edit distance:**
  - Let P be the predicted unit.
  - Let T be a ground truth repeat unit.
  - We calculate the edit distance between T and all possible rotations of P, and take the minimum value.

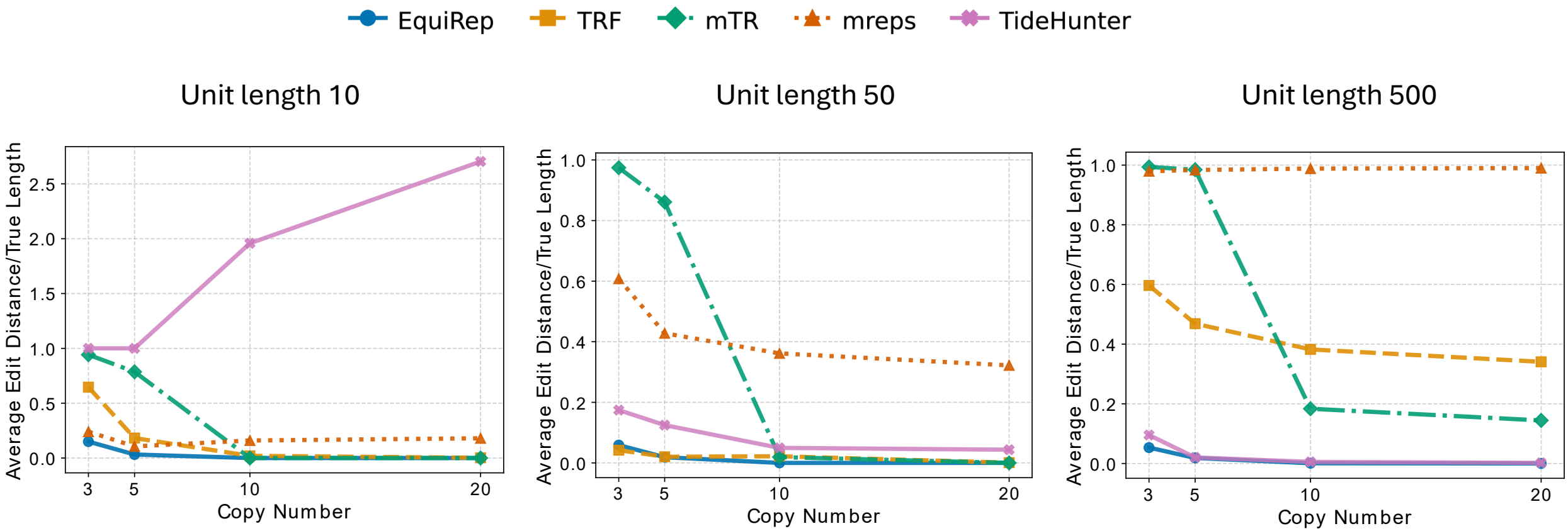


Evaluation

- **Normalized rotation-aware edit distance:**
  - Let P be the predicted unit.
  - Let T be a ground truth repeat unit.
  - We calculate the edit distance between T and all possible rotations of P, and take the minimum value.
  - We divide the minimum distance by the true unit length for normalization.



Results on Simulations



Average normalized rotation-aware edit distance at 10% error rate

Experimental Setup: C. elegans Centromere

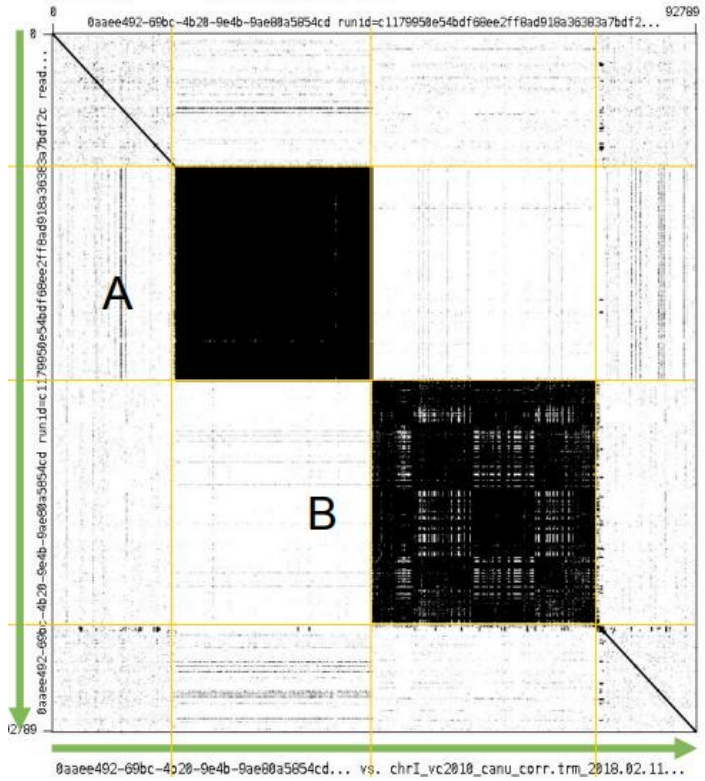
- We adopt a dataset reported in *Yoshimura et al*. that studied the assembly of C. elegans genome using Nanopore long-reads data.

Yoshimura, Jun, et al. "Recompleting the Caenorhabditis elegans genome." *Genome research* 29.6 (2019): 1009-1022.

Experimental Setup: C. elegans Centromere

- We adopt a dataset reported in *Yoshimura et al.* that studied the assembly of C. elegans genome using Nanopore long-reads data.

- We collect the raw long reads that are aligned to centromere and extract rough repeat regions using dotplots.

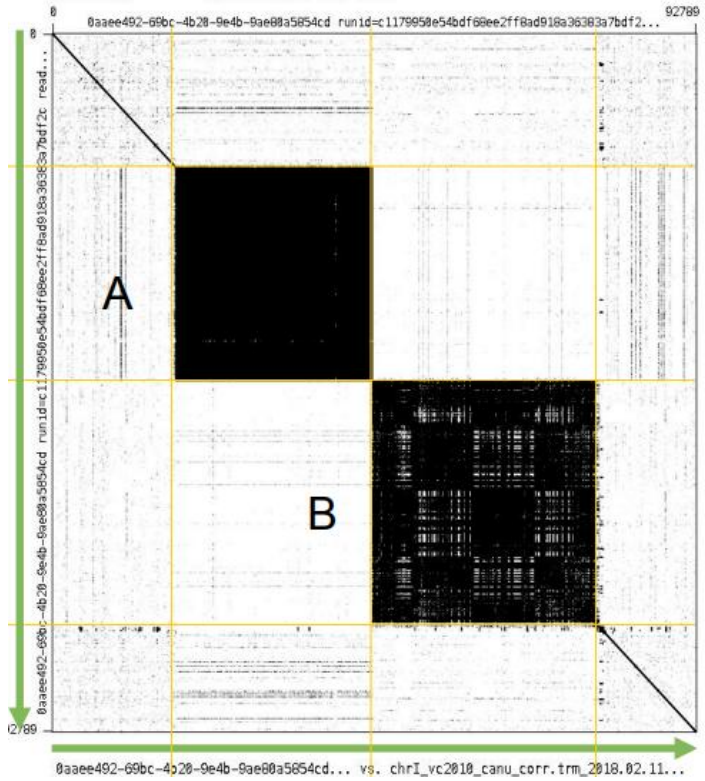


Dotplot from *Yoshimura, Jun, et al.*

Yoshimura, Jun, et al. "Recompleting the Caenorhabditis elegans genome." *Genome research* 29.6 (2019): 1009-1022.

Experimental Setup: C. elegans Centromere

- We adopt a dataset reported in *Yoshimura et al.* that studied the assembly of C. elegans genome using Nanopore long-reads data.
- We collect the raw long reads that are aligned to centromere and extract rough repeat regions using dotplots.
- The ground-truth sequence of the unit is available, which are obtained by curating from PacBio HIFI datasets.

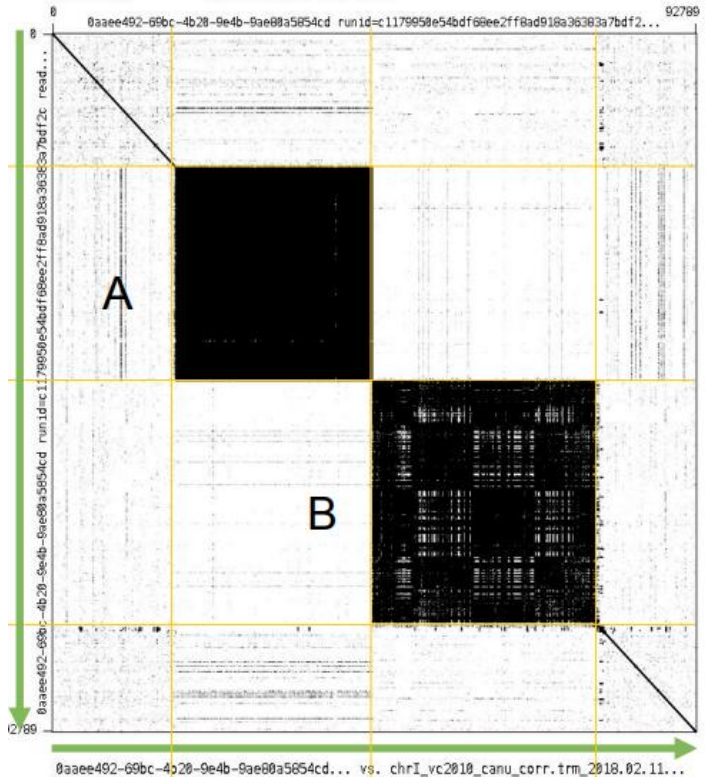


Dotplot from *Yoshimura, Jun, et al.*

Yoshimura, Jun, et al. "Recompleting the Caenorhabditis elegans genome." *Genome research* 29.6 (2019): 1009-1022.

Experimental Setup: C. elegans Centromere

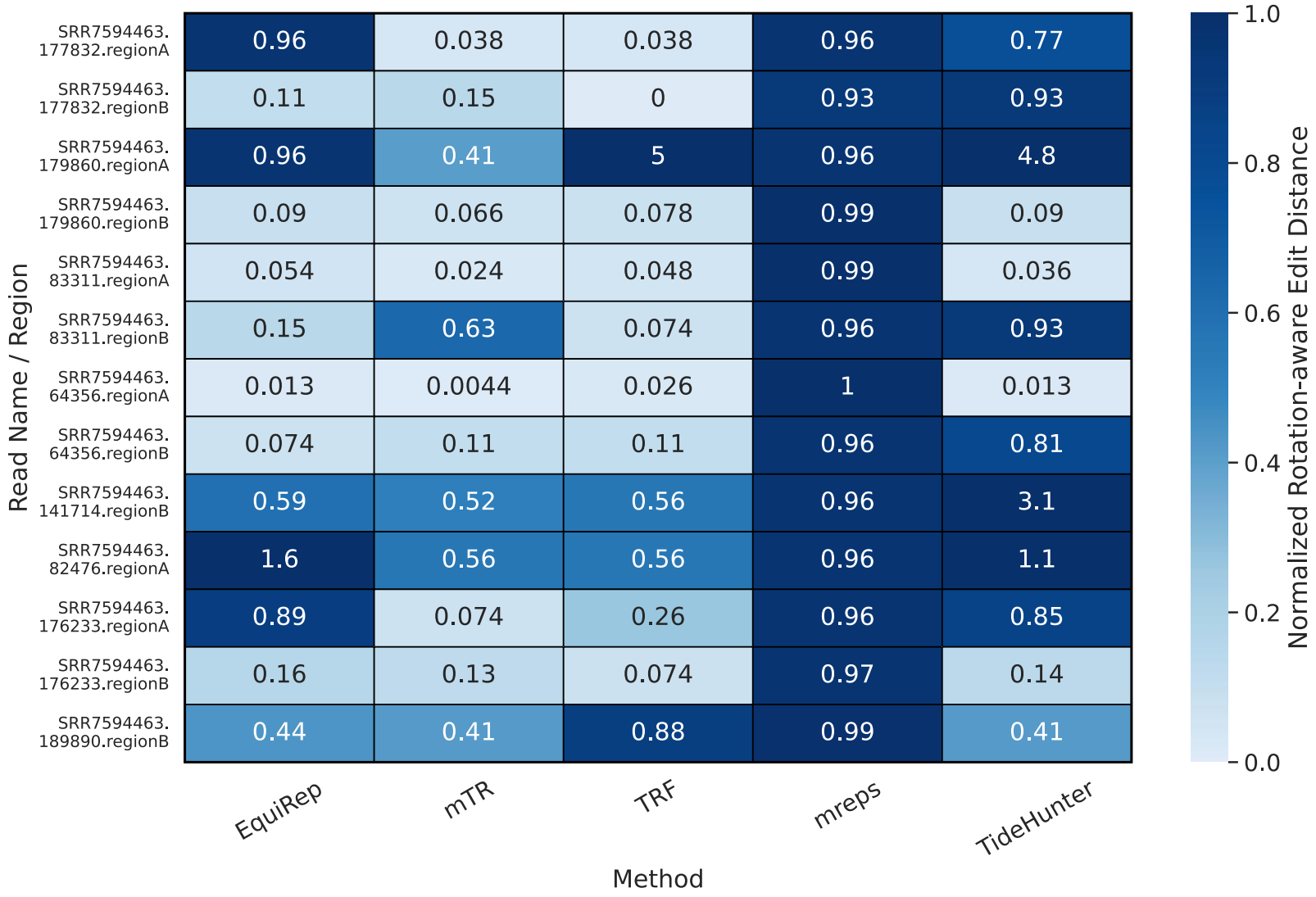
- We adopt a dataset reported in *Yoshimura et al.* that studied the assembly of C. elegans genome using Nanopore long-reads data.
- We collect the raw long reads that are aligned to centromere and extract rough repeat regions using dotplots.
- The ground-truth sequence of the unit is available, which are obtained by curating from PacBio HIFI datasets.
- We calculate the normalized rotation-aware edit distance between the predicted unit and the ground truth unit.



Dotplot from *Yoshimura, Jun, et al.*

Yoshimura, Jun, et al. "Recompleting the Caenorhabditis elegans genome." *Genome research* 29.6 (2019): 1009-1022.

Results on C. elegans Centromere Data



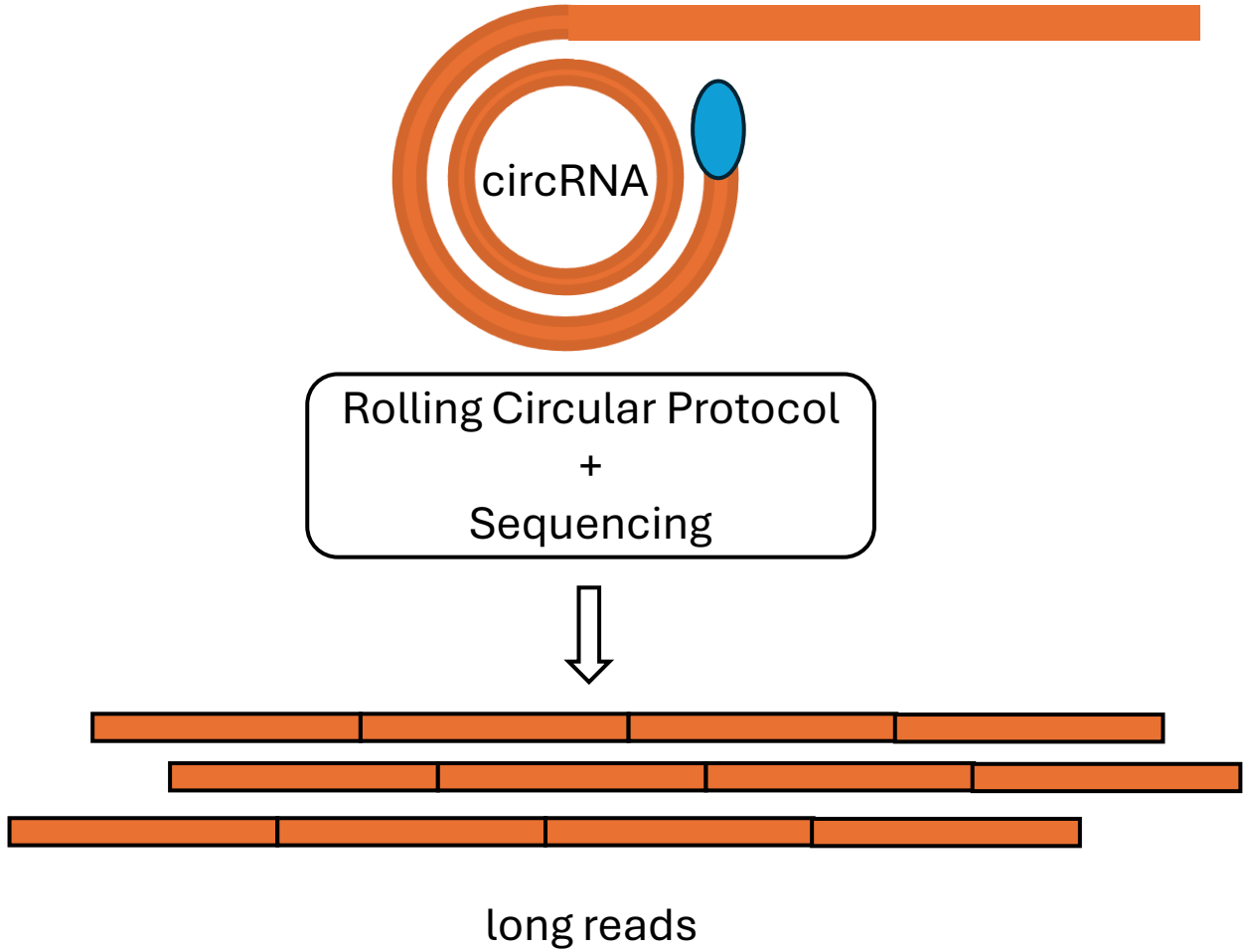
Experimental Setup: RCA Human Tissue

- This set of real data is a Rolling Circle Amplification based Nanopore sequencing protocol from *Xin et al*.

Xin, Ruijiao, et al. "isoCirc catalogs full-length circular RNA isoforms in human transcriptomes." *Nature communications* 12.1 (2021): 266.

Experimental Setup: RCA Human Tissue

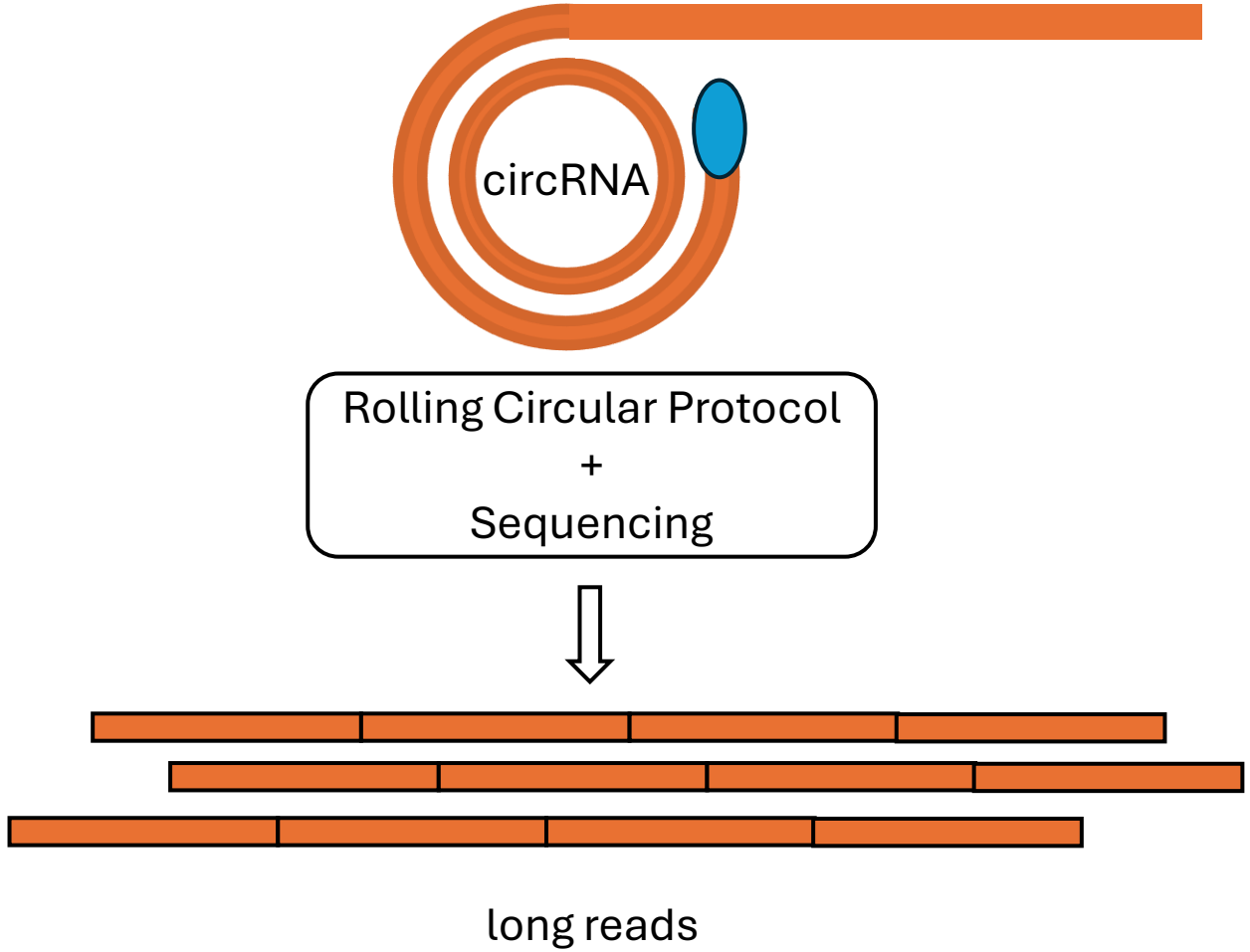
- This set of real data is a Rolling Circle Amplification based Nanopore sequencing protocol from *Xin et al*.



Xin, Ruijiao, et al. "isoCirc catalogs full-length circular RNA isoforms in human transcriptomes." *Nature communications* 12.1 (2021): 266.

Experimental Setup: RCA Human Tissue

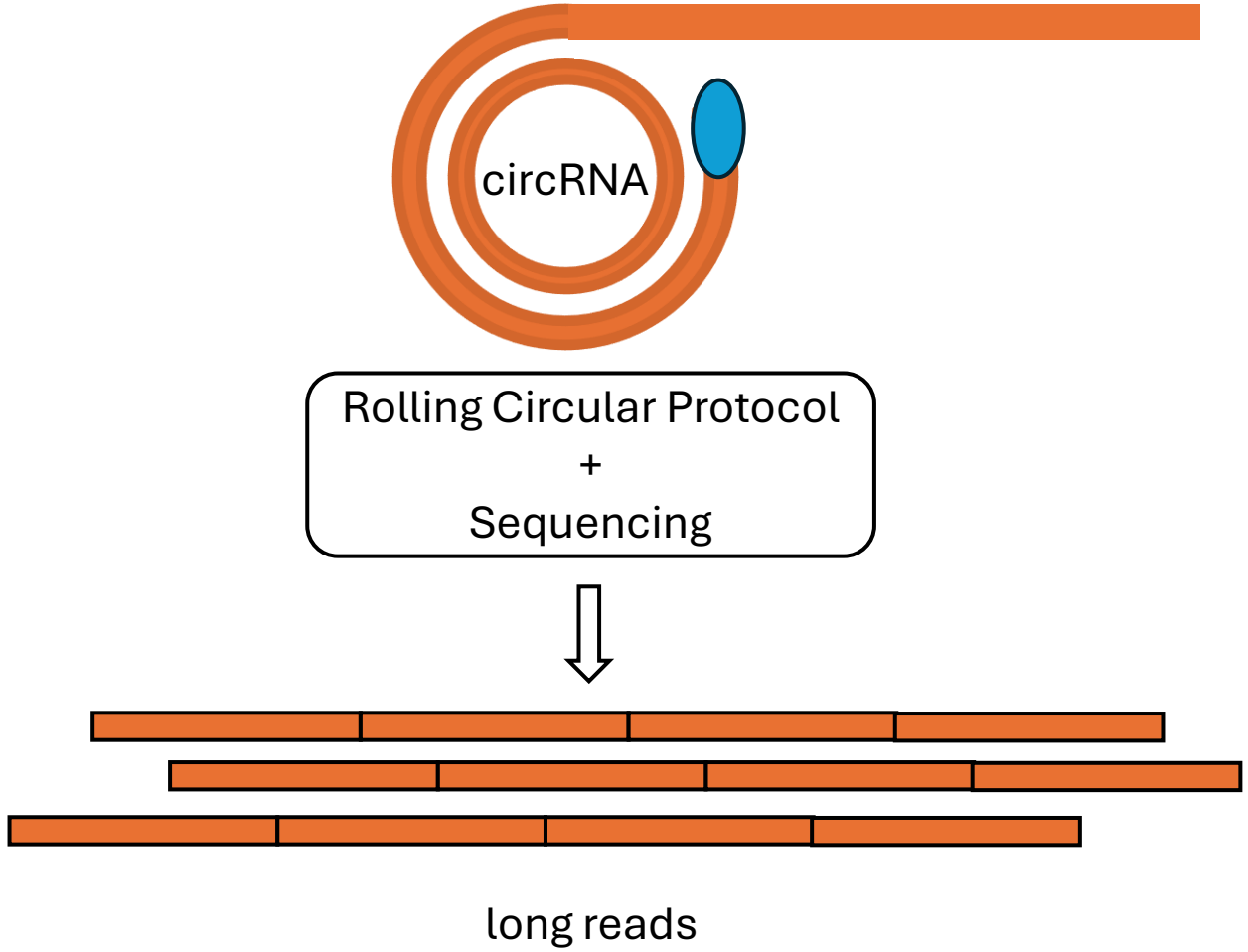
- This set of real data is a Rolling Circle Amplification based Nanopore sequencing protocol from *Xin et al*.
- This dataset has been used to detect a catalogue of full-length circular RNAs from 12 human tissues.



Xin, Ruijiao, et al. "isoCirc catalogs full-length circular RNA isoforms in human transcriptomes." *Nature communications* 12.1 (2021): 266.

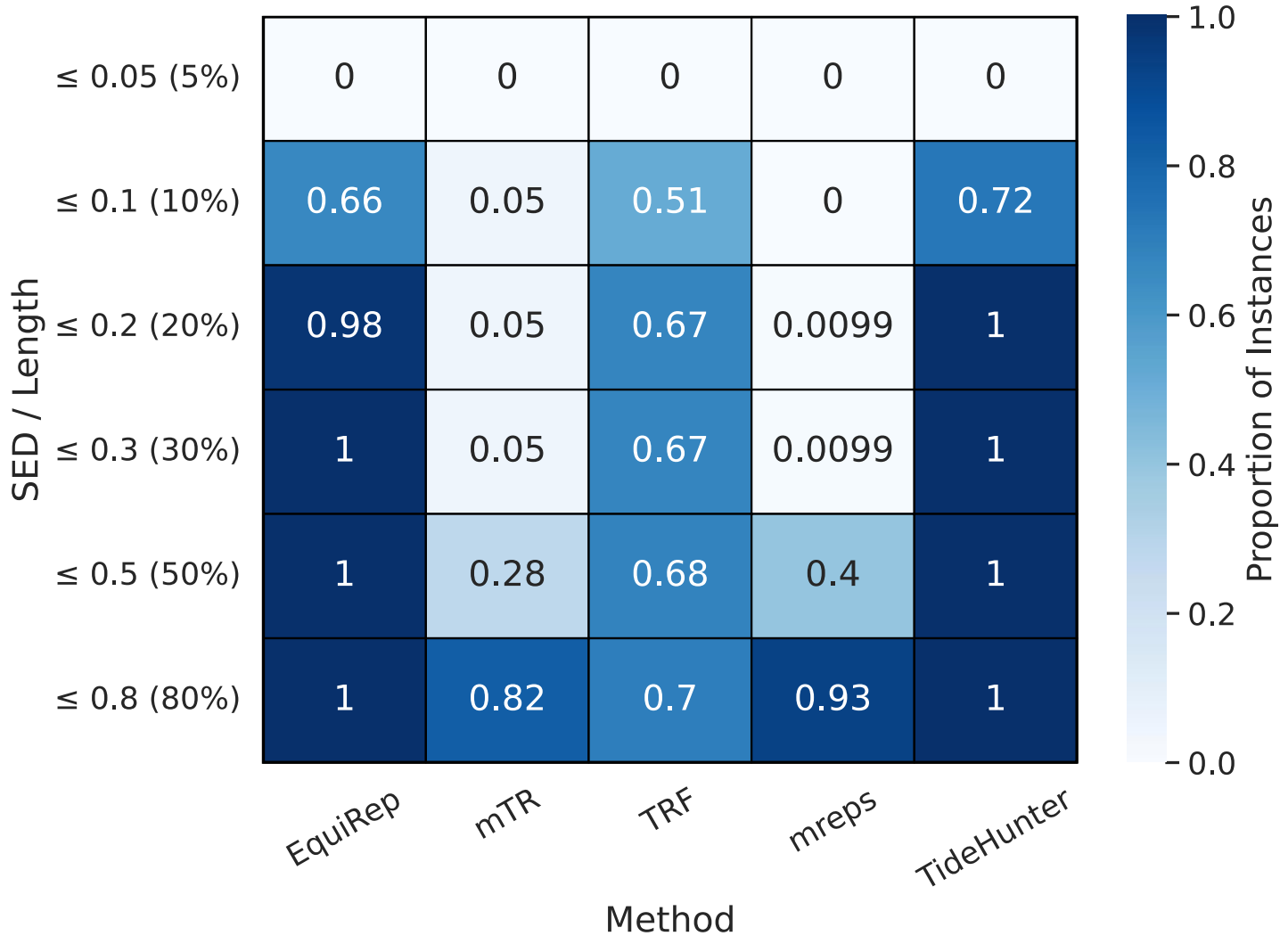
Experimental Setup: RCA Human Tissue

- This set of real data is a Rolling Circle Amplification based Nanopore sequencing protocol from *Xin et al*.
- This dataset has been used to detect a catalogue of full-length circular RNAs from 12 human tissues.
- We collect a subset of 101 sequences of the Nanopore long reads from the human prostate tissue for analysis.



Xin, Ruijiao, et al. "isoCirc catalogs full-length circular RNA isoforms in human transcriptomes." *Nature communications* 12.1 (2021): 266.

Results on RCA Data



Summary

Summary

- We present EquiRep, a new tool for reconstructing the tandem repeat unit from error-prone sequences.

Summary

- We present EquiRep, a new tool for reconstructing the tandem repeat unit from error-prone sequences.

- EquiRep identifies equivalent positions within a sequence by combining self-local alignment with an innovative refinement step that reliably reduces noise.

Summary

- We present EquiRep, a new tool for reconstructing the tandem repeat unit from error-prone sequences.
- EquiRep identifies equivalent positions within a sequence by combining self-local alignment with an innovative refinement step that reliably reduces noise.
- We present results that show EquiRep's robustness against errors and effectiveness in reconstructing repeats of large length and low frequency.

Summary

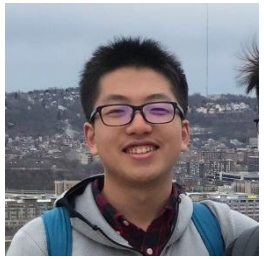
- We present EquiRep, a new tool for reconstructing the tandem repeat unit from error-prone sequences.
- EquiRep identifies equivalent positions within a sequence by combining self-local alignment with an innovative refinement step that reliably reduces noise.
- We present results that show EquiRep's robustness against errors and effectiveness in reconstructing repeats of large length and low frequency.
- Given the scarcity of tools that can reliably reconstruct long, error-prone repeat units, we expect EquiRep to be widely used.

Summary

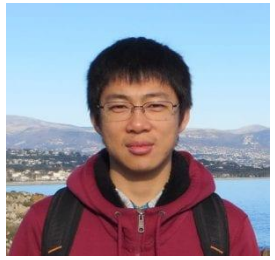
- We present EquiRep, a new tool for reconstructing the tandem repeat unit from error-prone sequences.
- EquiRep identifies equivalent positions within a sequence by combining self-local alignment with an innovative refinement step that reliably reduces noise.
- We present results that show EquiRep's robustness against errors and effectiveness in reconstructing repeats of large length and low frequency.
- Given the scarcity of tools that can reliably reconstruct long, error-prone repeat units, we expect EquiRep to be widely used.
- Tool availability: https://github.com/Shao-Group/EquiRep.

Acknowledgements

- Co-authors


Zhezheng Song


Xiang Li


Mingfu Shao

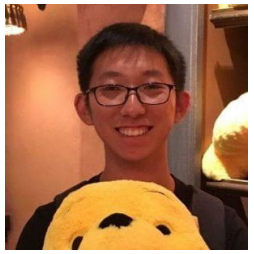
- Funding support

- Members of Shao Group


Ke Chen


Xin Yuan


Qian Shi


Carl Zang


Aaryan Mahesh


Ajmain Yasar


Irtesam Mahmud


Abhishek Talesara