

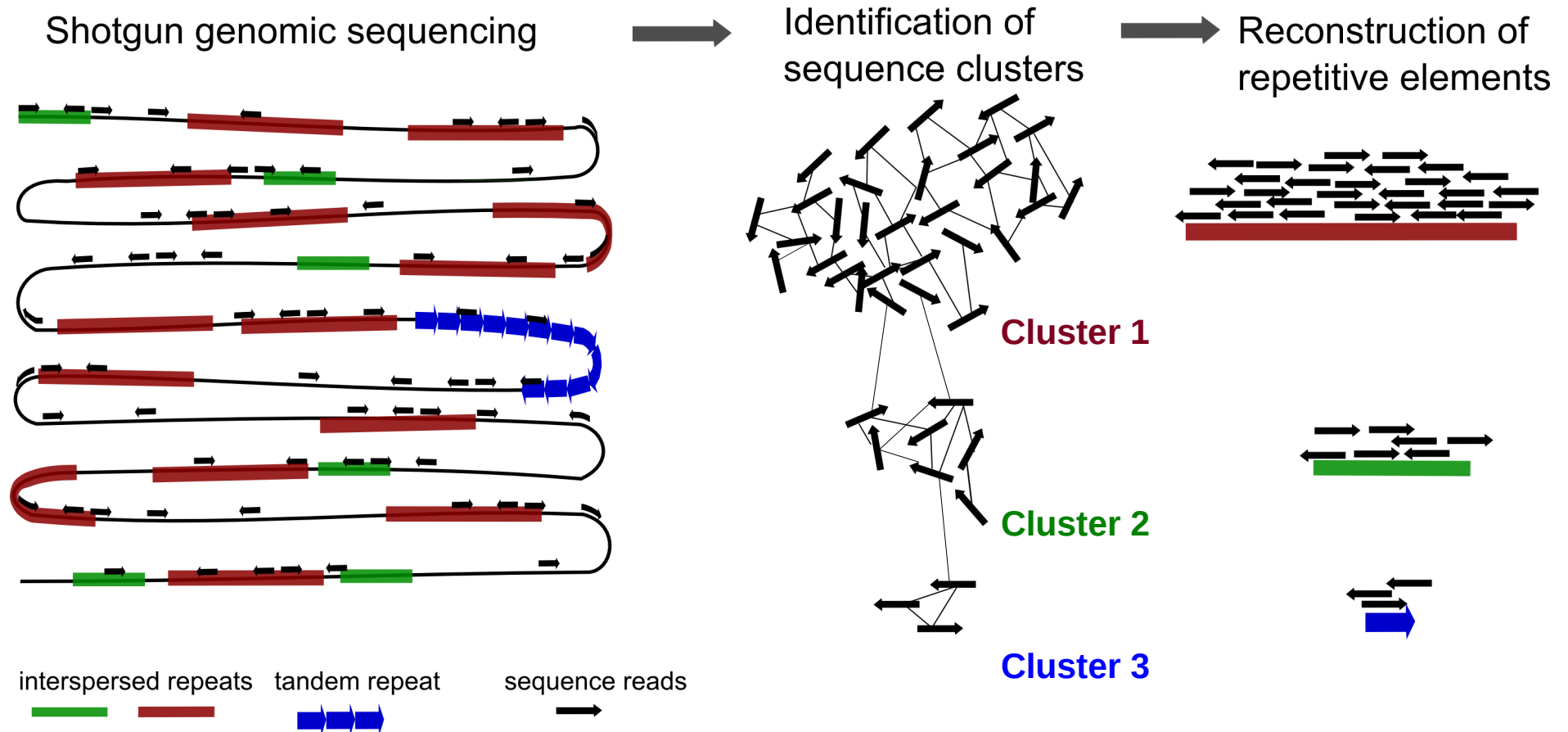
Tandem Repeat Analyzer - TAREAN

Extension of RepeatExplorer clustering

Automated identification of clusters derived from tandem repeats
and reconstruction of monomer consensus sequence

Tandem Repeat Analyzer - TAREAN

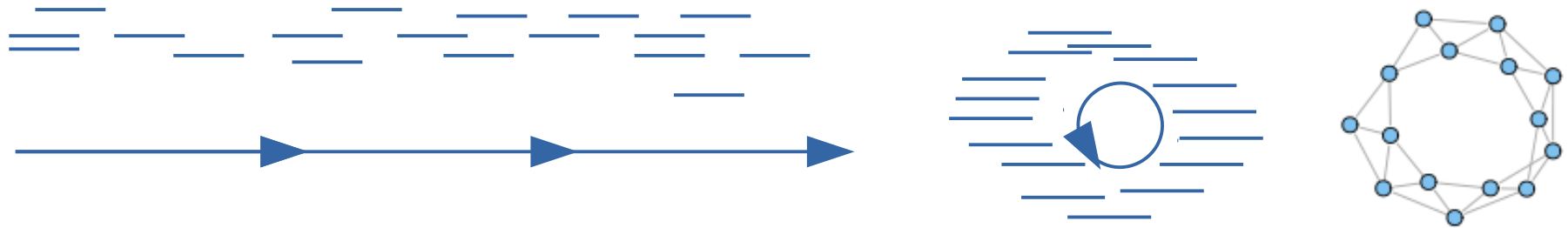
Extension of RepeatExplorer clustering



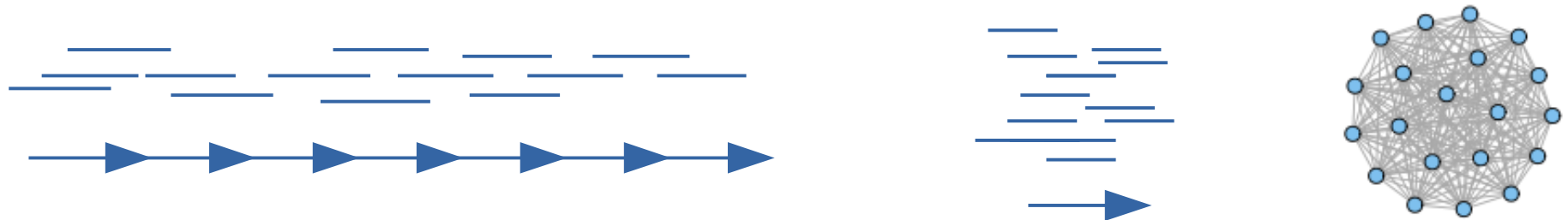
Tandem Repeat Analyzer - TAREAN

Extension of RepeatExplorer clustering

Read length $\ll$ monomer



Read length $\geq$ monomer



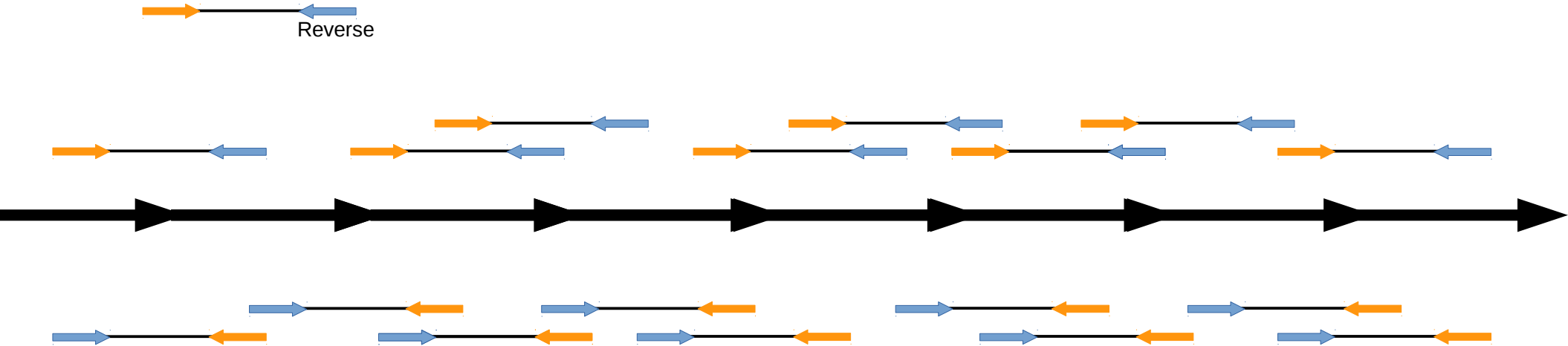
Tandem Repeat Analyzer - TAREAN

Principle

Paired-End Sequencing

Forward

Reverse

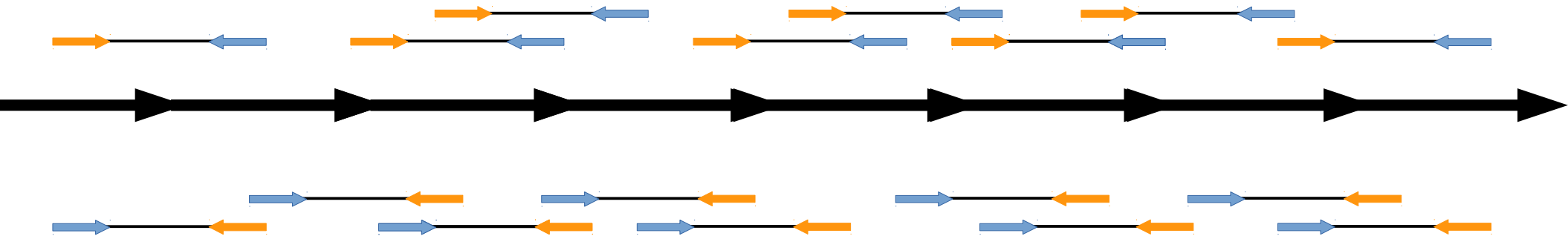


Tandem Repeat Analyzer - TAREAN

Principle

Paired-End Sequencing

Forward
Reverse



All-to-all comparison - blast

query	subject	pid	e.value	strand
66706r	130626f	93.22	3e-20	+
65114r	95482r	95.35	4e-16	-
32080f	22009f	93.65	3e-23	+
85661f	22009f	97.14	9e-14	-
66706r	9071r	95.74	2e-18	+
130626f	9071r	95.83	5e-43	+
66706r	147917r	95.74	2e-18	-

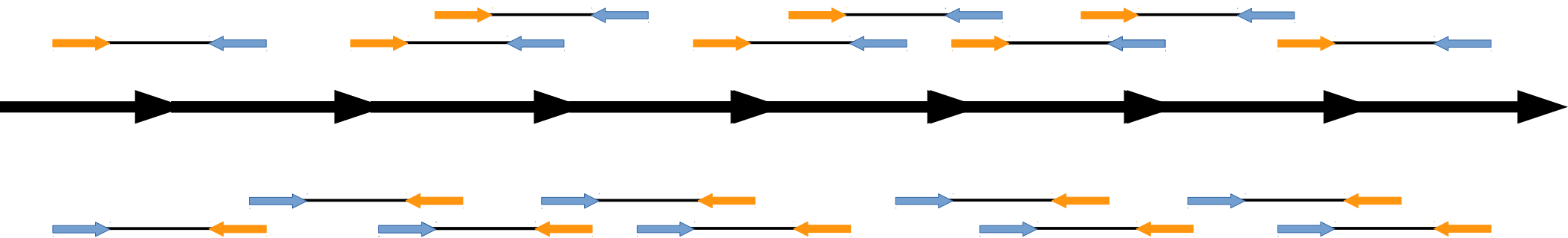
...

Tandem Repeat Analyzer - TAREAN

Principle

Paired-End Sequencing

Forward
Reverse



All-to-all comparison - blast

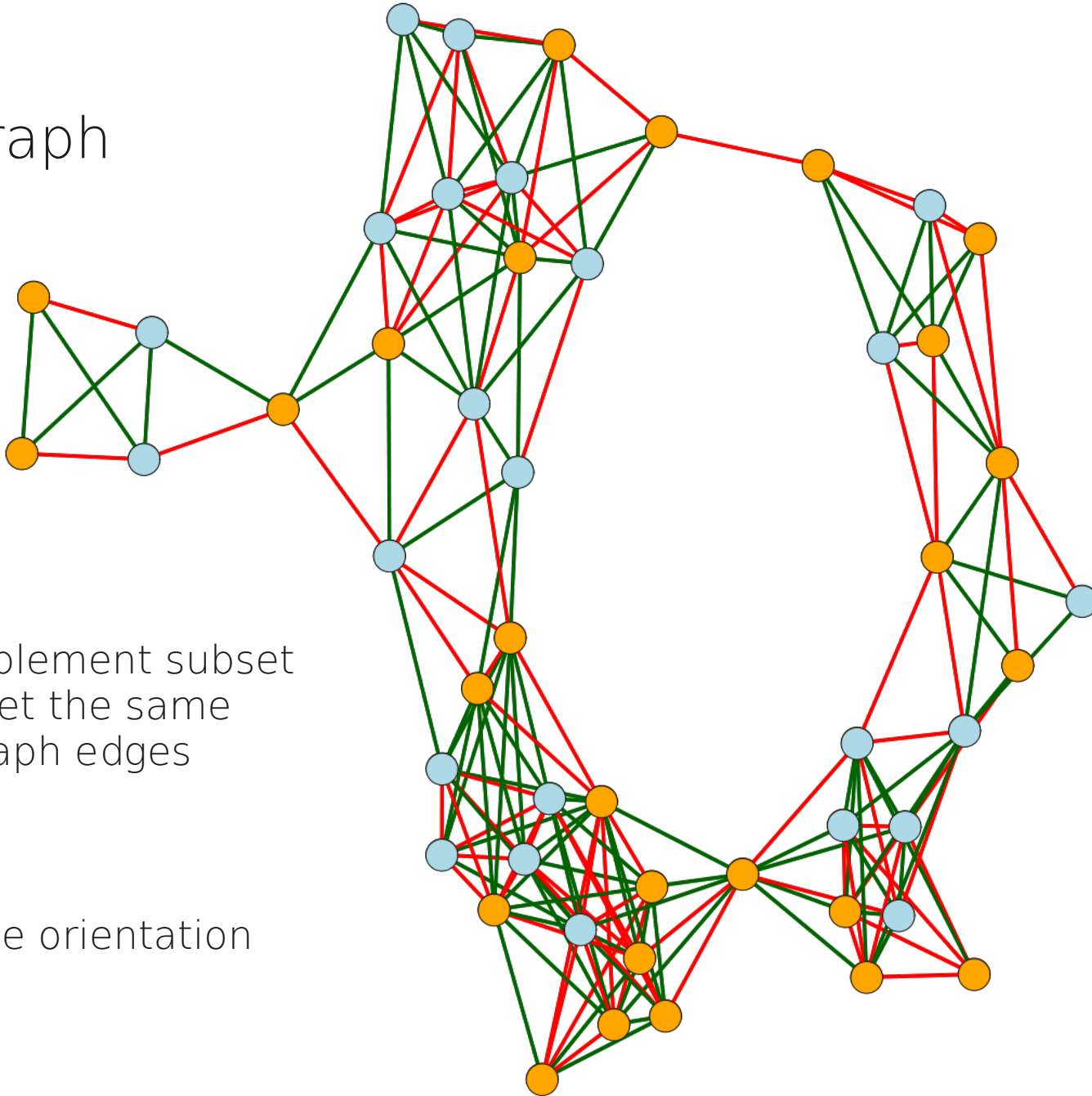
query	subject	pid	e.value	strand
66706f	130626r	93.22	3e-20	+
65114f	95482f	95.35	4e-16	-
32080r	22009r	93.65	3e-23	+
85661r	22009r	97.14	9e-14	-
66706f	9071f	95.74	2e-18	+
130626r	9071f	95.83	5e-43	+
66706f	147917f	95.74	2e-18	-

...

Tandem Repeat Analyzer - TAREAN

Principle

Signed graph



Reverse complement subset
of reads to get the same
sign on all graph edges

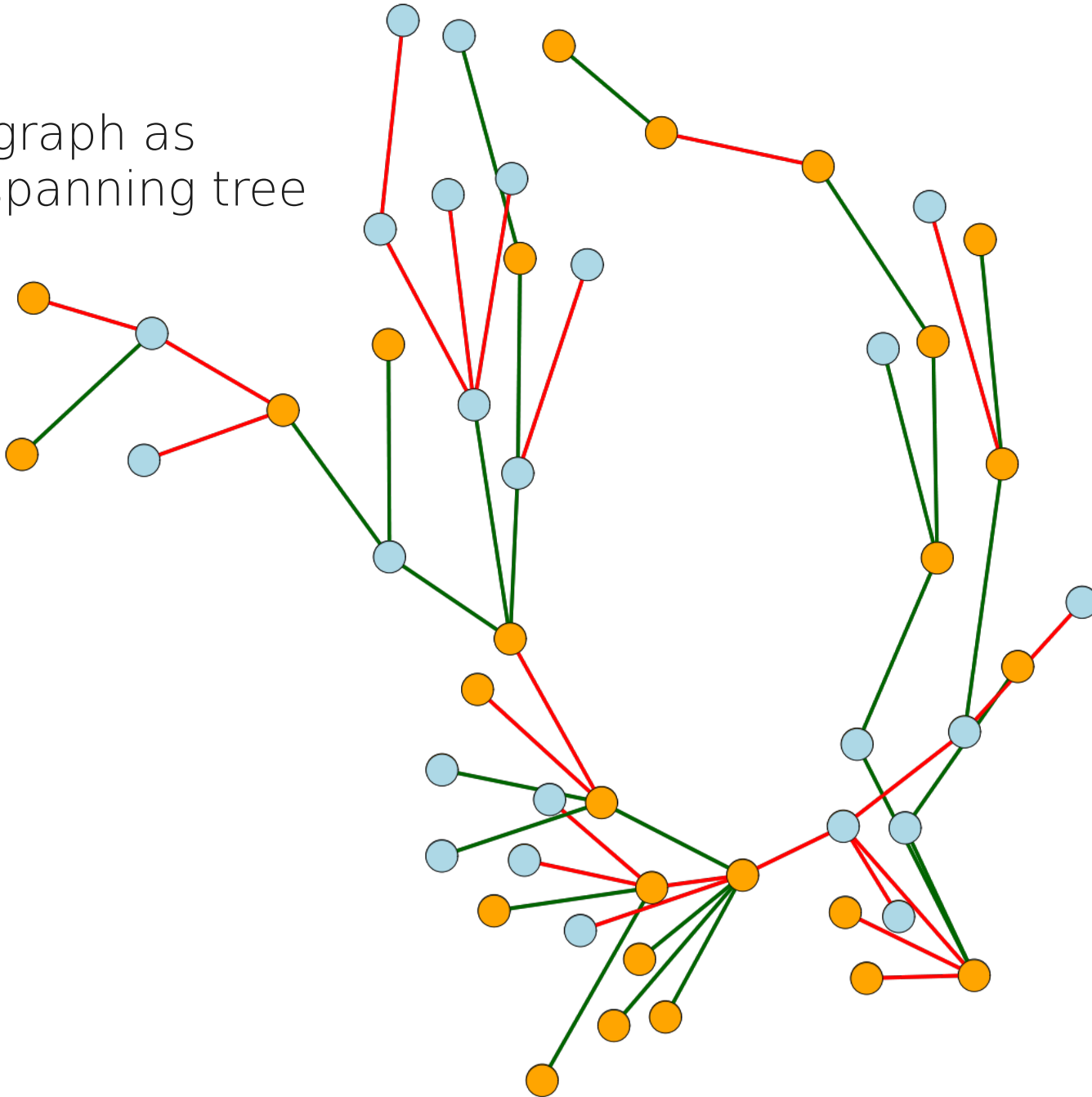


All reads same orientation

Tandem Repeat Analyzer - TAREAN

Principle

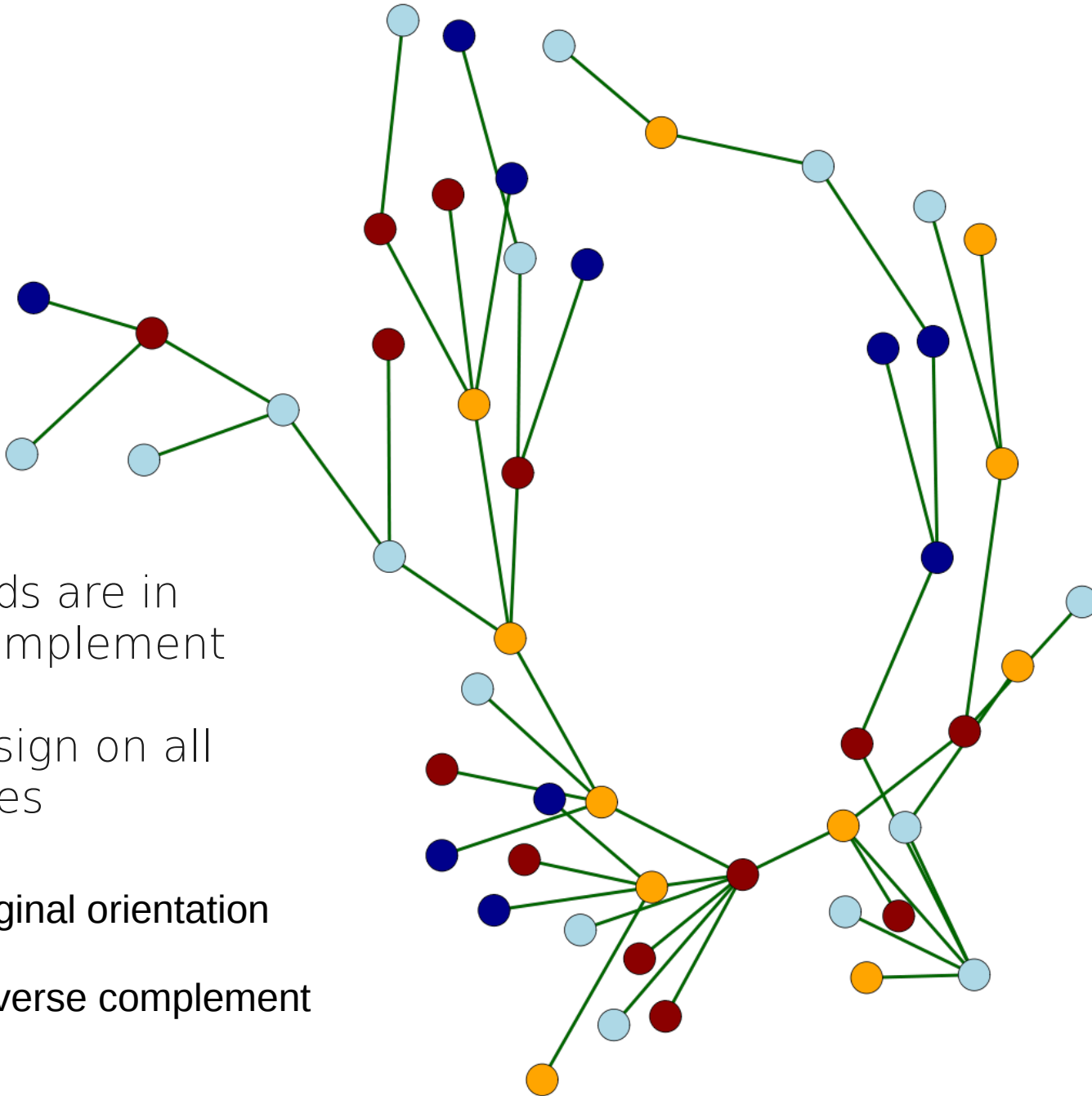
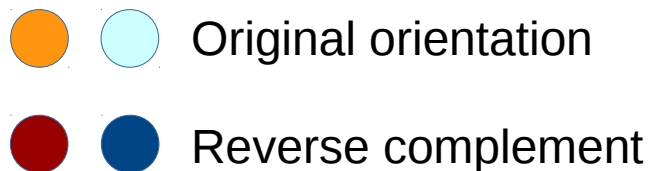
Simplified graph as
minimum spanning tree



Tandem Repeat Analyzer - TAREAN

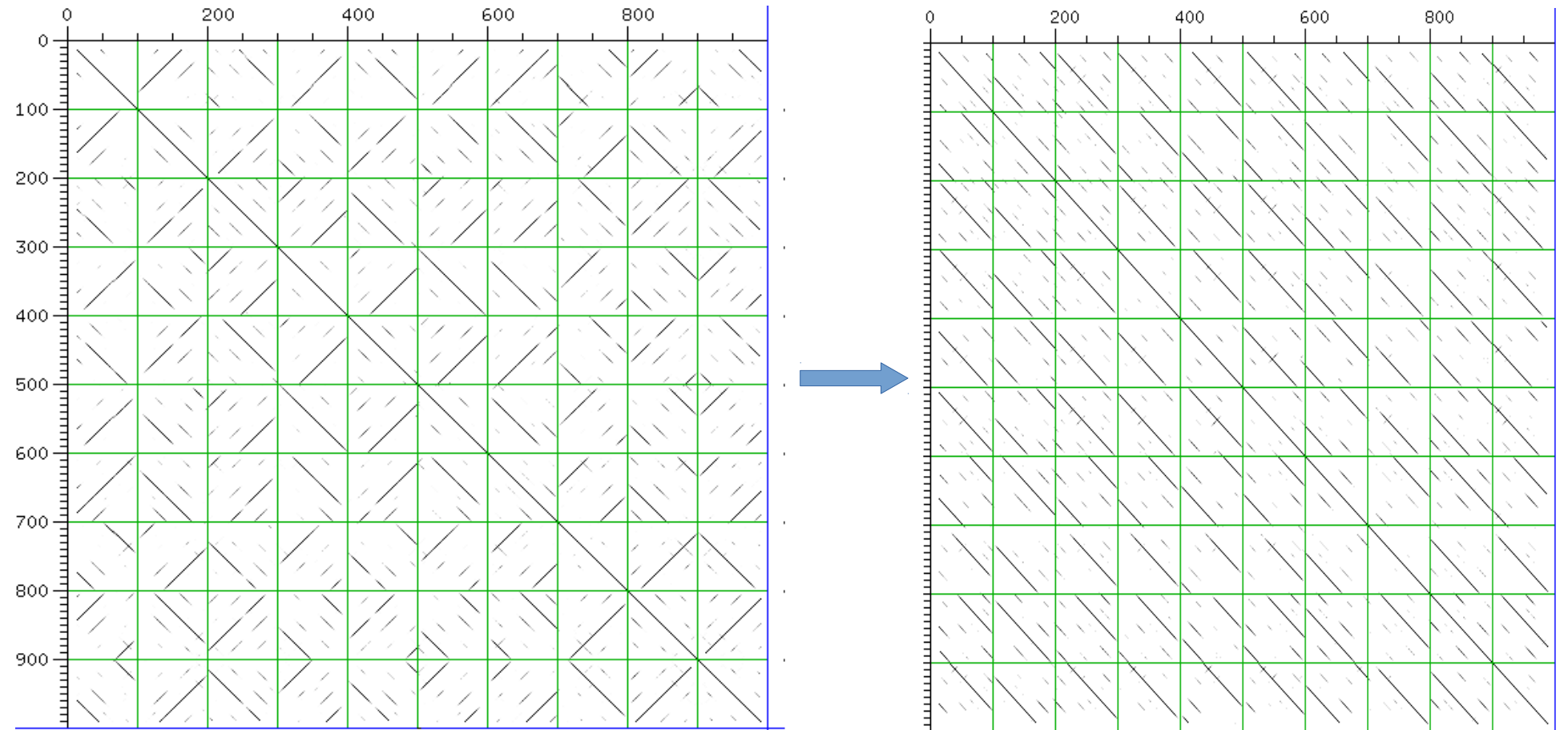
Principle

- ~50% reads are in reverse complement
- the same sign on all graph edges



Tandem Repeat Analyzer - TAREAN

Principle

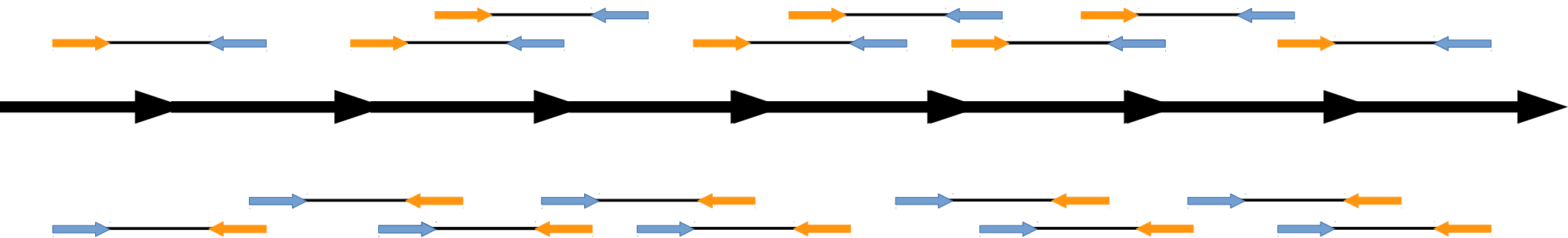


Tandem Repeat Analyzer - TAREAN

Principle

Paired-End Sequencing

Forward
Reverse



All-to-all comparison - blast

query	subject	pid	e.value	strand
66706f	130626r	93.22	3e-20	+
65114f	95482f	95.35	4e-16	-
32080r	22009r	93.65	3e-23	+
85661r	22009r	97.14	9e-14	-
66706f	9071f	95.74	2e-18	+
130626r	9071f	95.83	5e-43	+
66706f	147917f	95.74	2e-18	-

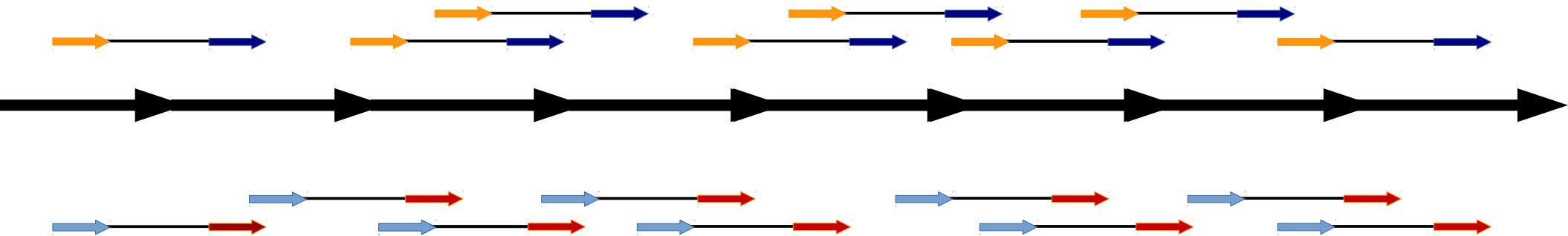
...

Tandem Repeat Analyzer - TAREAN

Principle

Paired-End Sequencing

Forward
Reverse



All-to-all comparison - blast

query	subject	pid	e.value	strand
66706f	130626r	93.22	3e-20	+
65114f	95482f	95.35	4e-16	-
32080r	22009r	93.65	3e-23	+
85661r	22009r	97.14	9e-14	-
66706f	9071f	95.74	2e-18	+
130626r	9071f	95.83	5e-43	+
66706f	147917f	95.74	2e-18	-

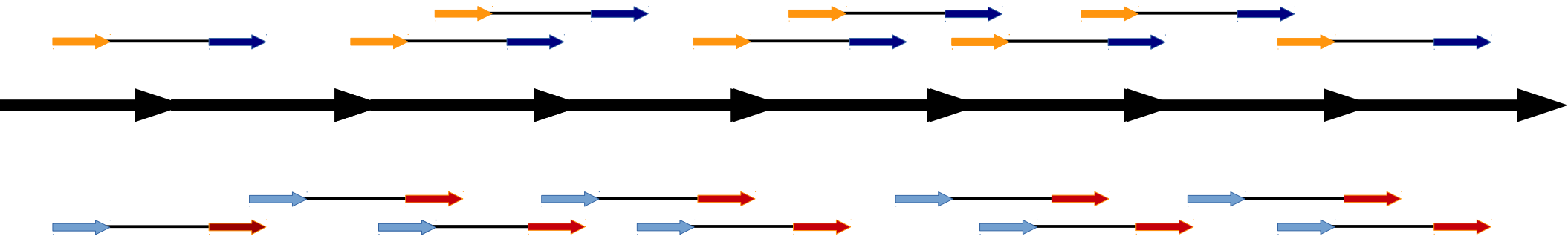
...

Tandem Repeat Analyzer - TAREAN

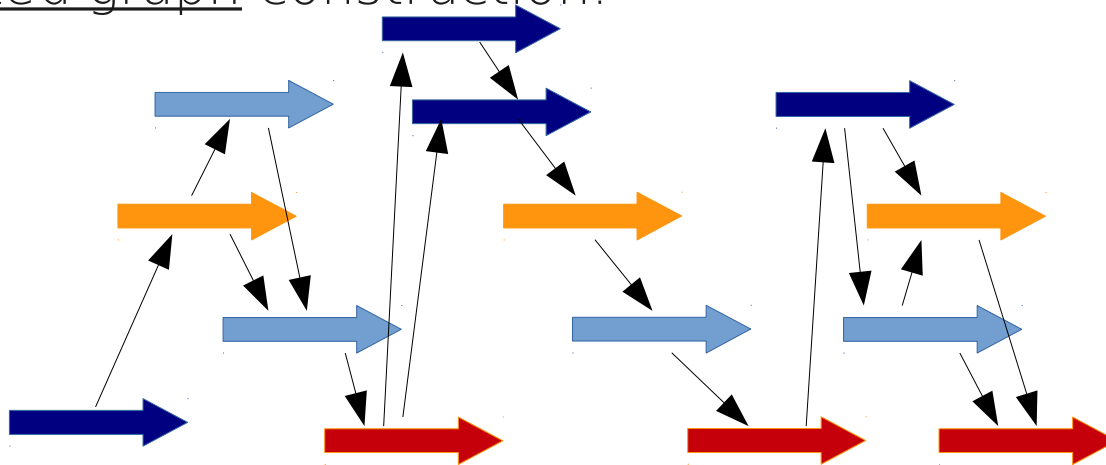
Principle

Paired-End Sequencing

Forward
Reverse



All-to-all comparison – blast – position of alignment and same orientation enable directed graph construction:

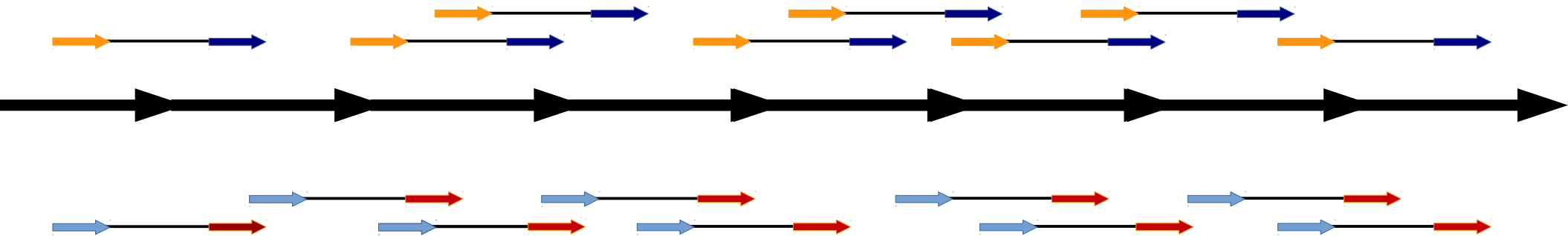


Tandem Repeat Analyzer - TAREAN

Principle

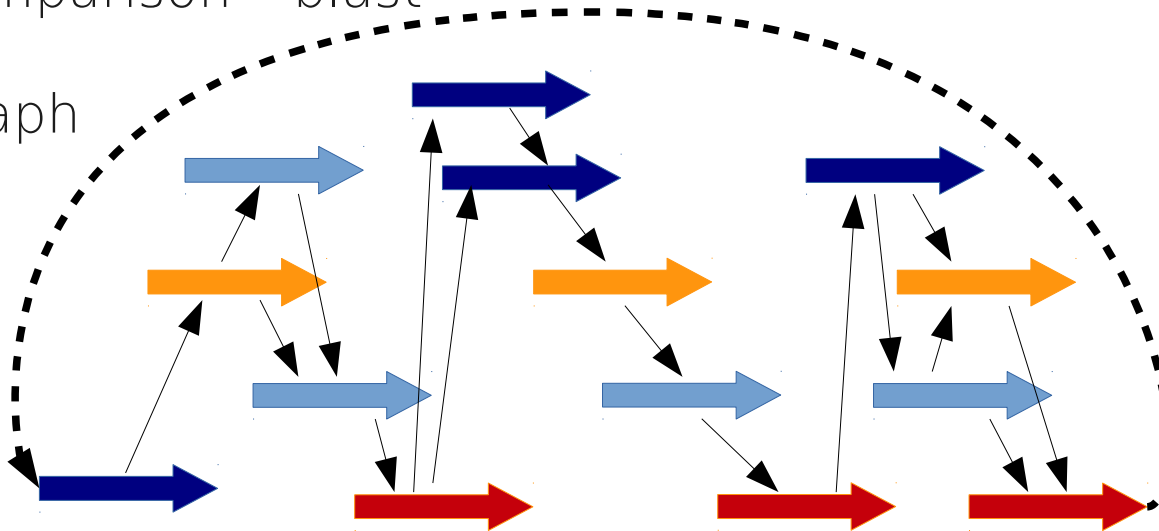
Paired-End Sequencing

Forward
Reverse



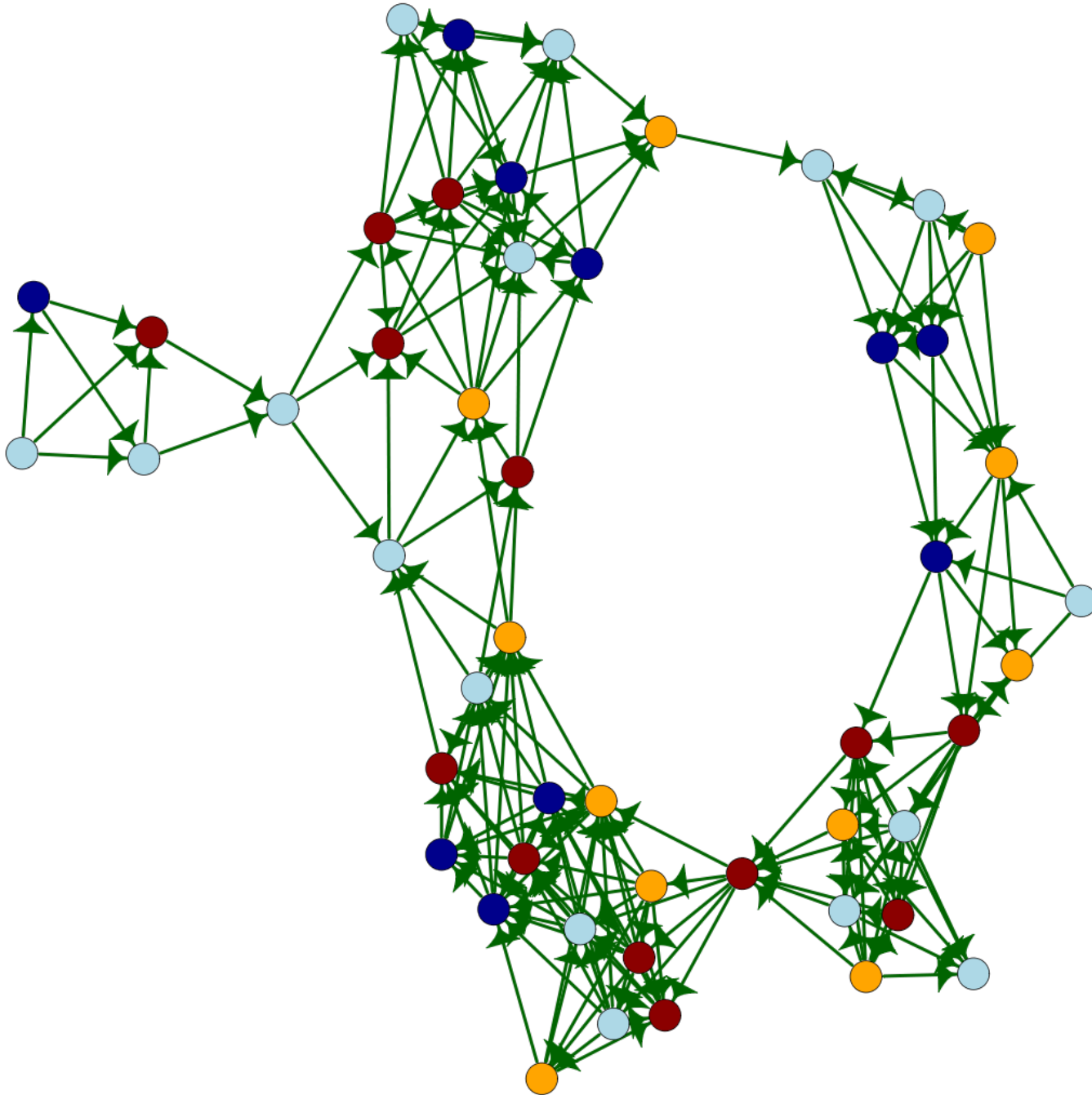
All-to-all comparison - blast

Directed graph



Tandem Repeat Analyzer - TAREAN

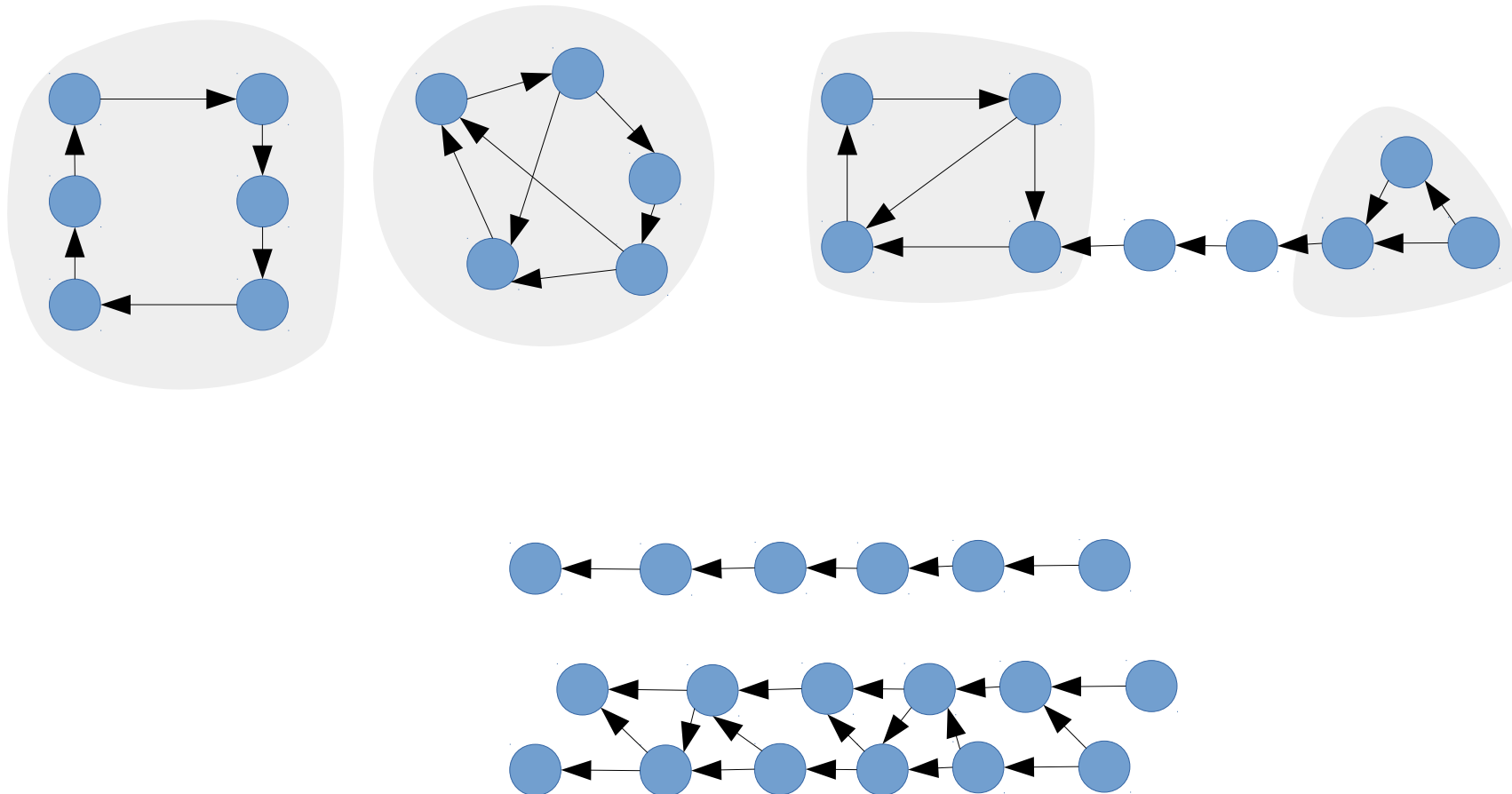
Principle



Tandem Repeat Analyzer - TAREAN

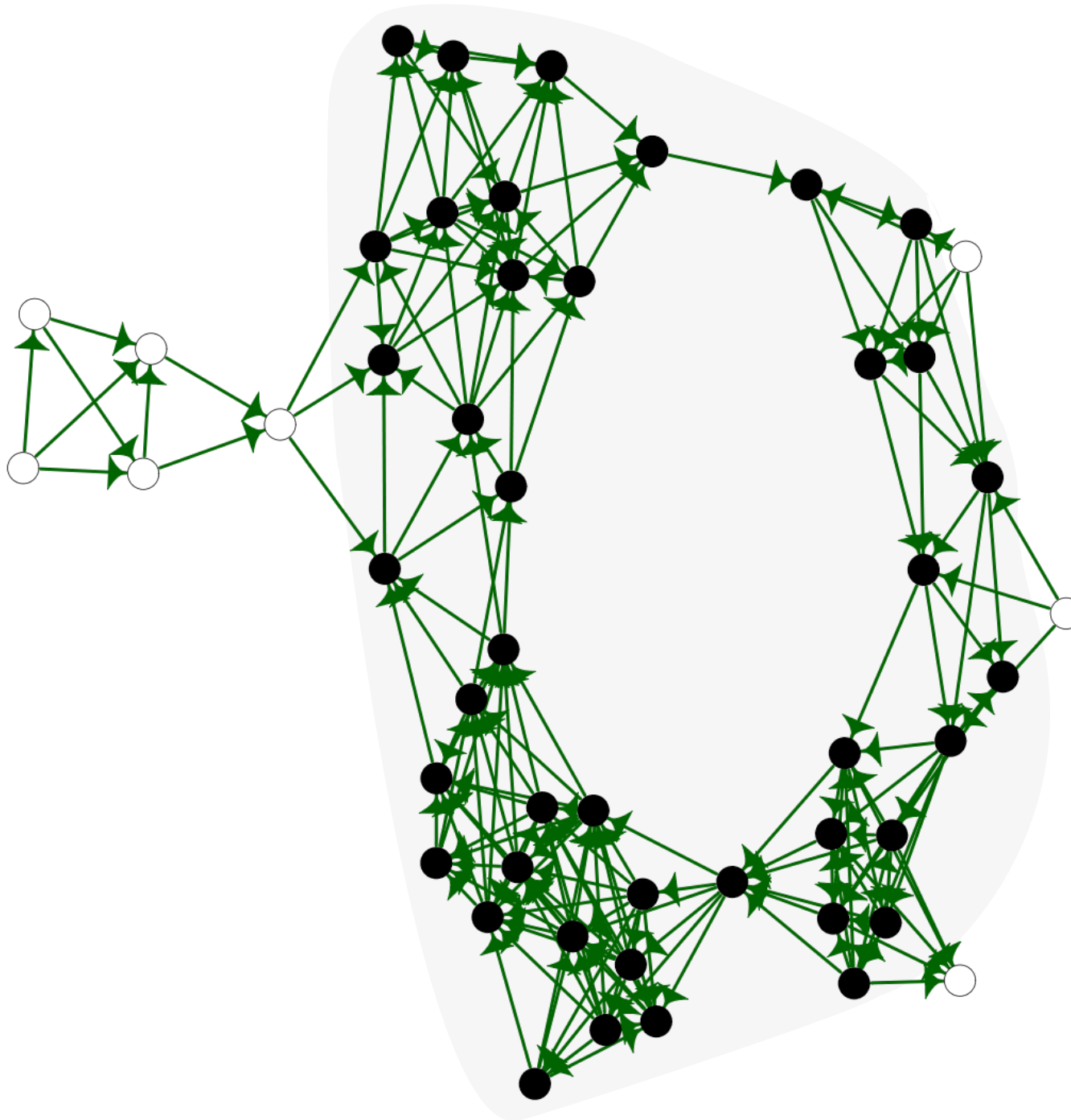
Principle

A directed graph is called **strongly connected** if there is a path in **each direction** between each pair of vertices of the graph.



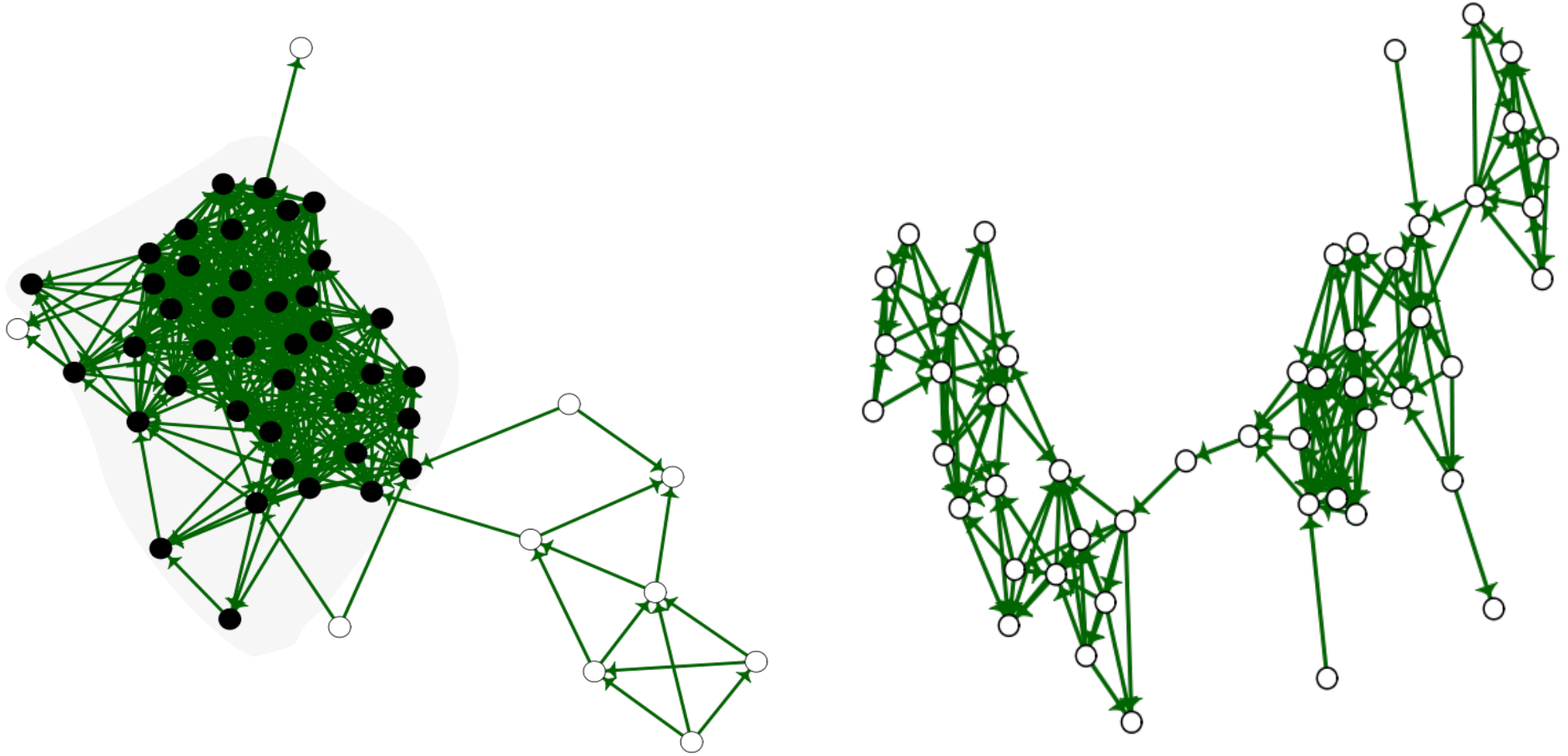
Tandem Repeat Analyzer - TAREAN

Principle



Tandem Repeat Analyzer - TAREAN

Principle



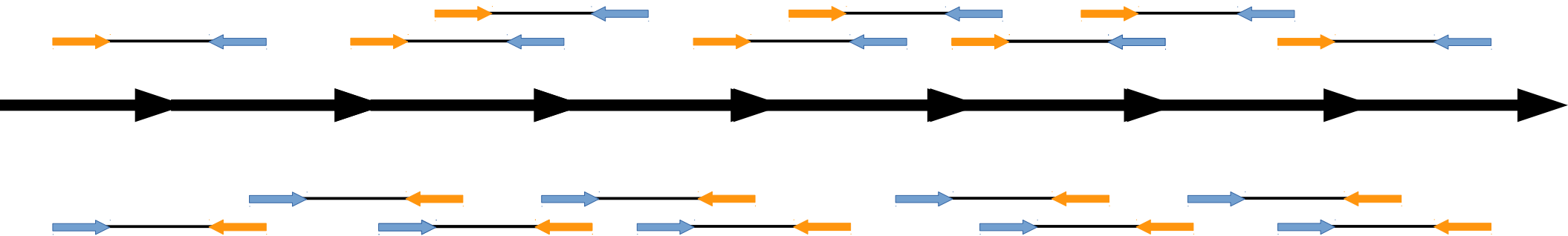
$$\text{Loop index} = \frac{\text{size of the largest strongly connected components}}{\text{Total graph size}}$$

Tandem Repeat Analyzer - TAREAN

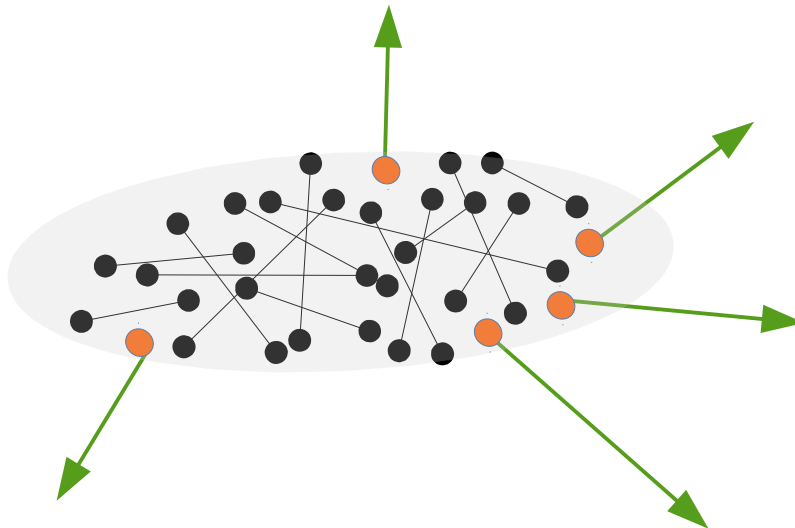
Principle

Paired-End Sequencing

Forward
Reverse



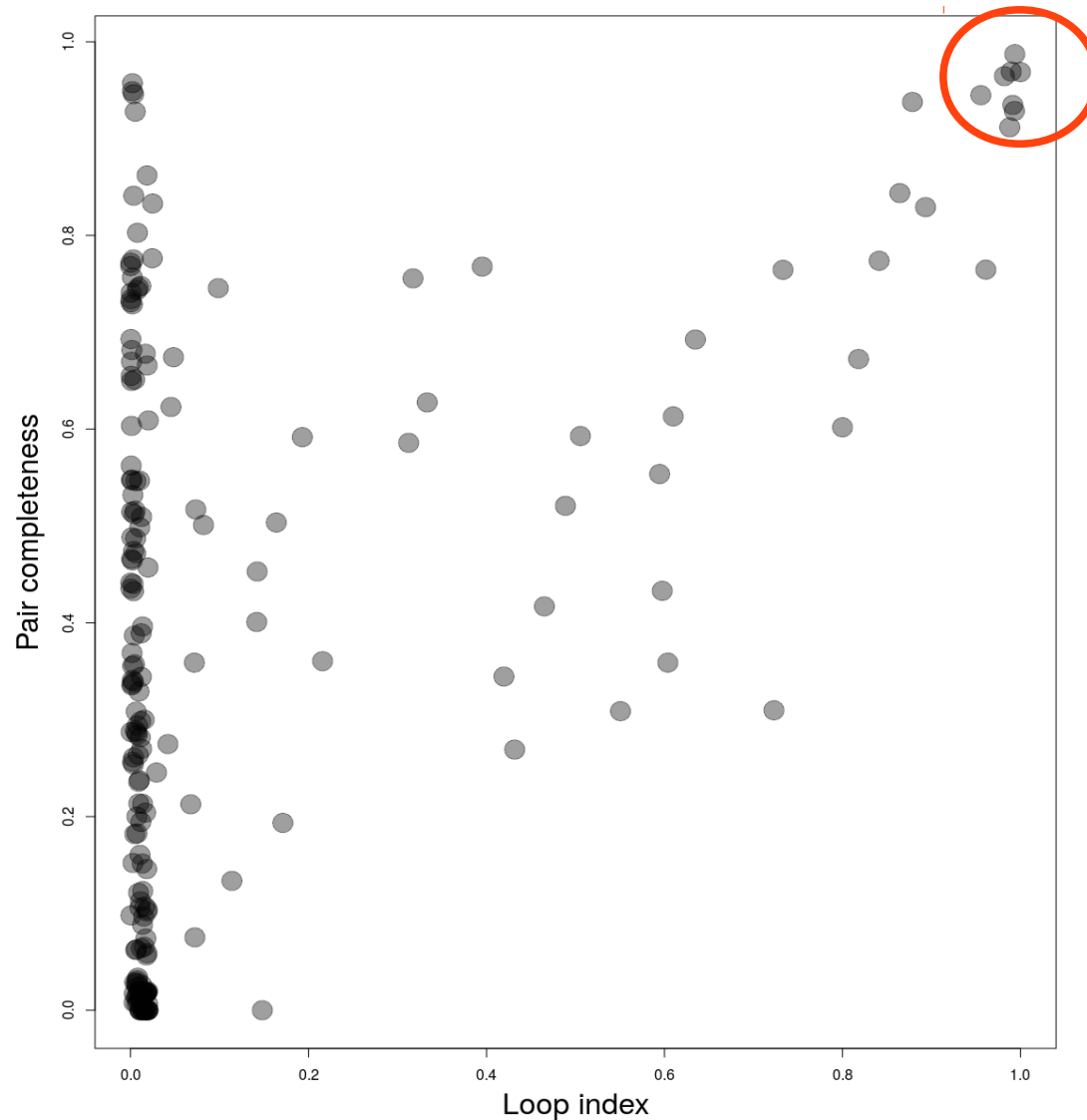
Pair completeness = fraction of complete pairs in cluster



Tandem Repeat Analyzer - TAREAN

Principle

Satellites tend to have pair completeness and loop index close to 1



Tandem Repeat Analyzer - TAREAN

kmer based tandem repeat monomer consensus reconstruction:

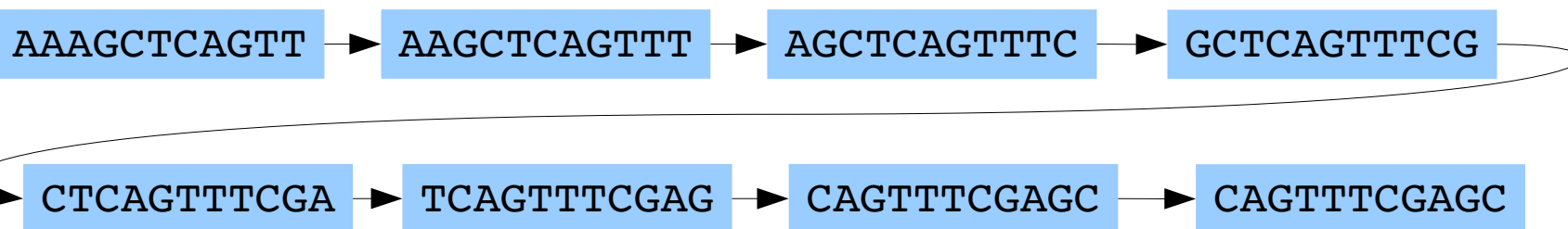
Sequence read:

AAAGCTCAGTTTCGAGCCAGAGACCAGAAAGTGTGGGAGCTTACAGCGCAACTTCAGCAAGAGCGGAG

AAAGCTCAGTT
AAGCTCAGTTT
AGCTCAGTTTC
GCTCAGTTTCG
CTCAGTTTCGA
TCAGTTTCGAG
CAGTTTCGAGC
AGTTTCGAGCC
.....

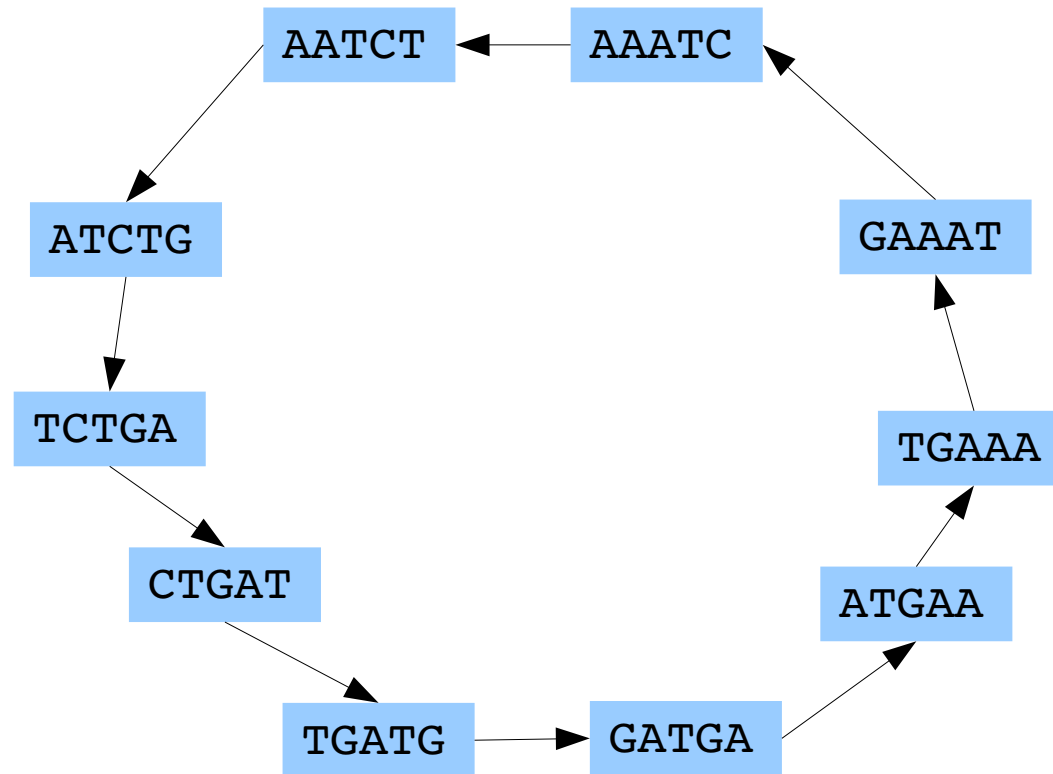
11mers

de Bruijn graph



Tandem Repeat Analyzer - TAREAN

Consensus reconstruction:

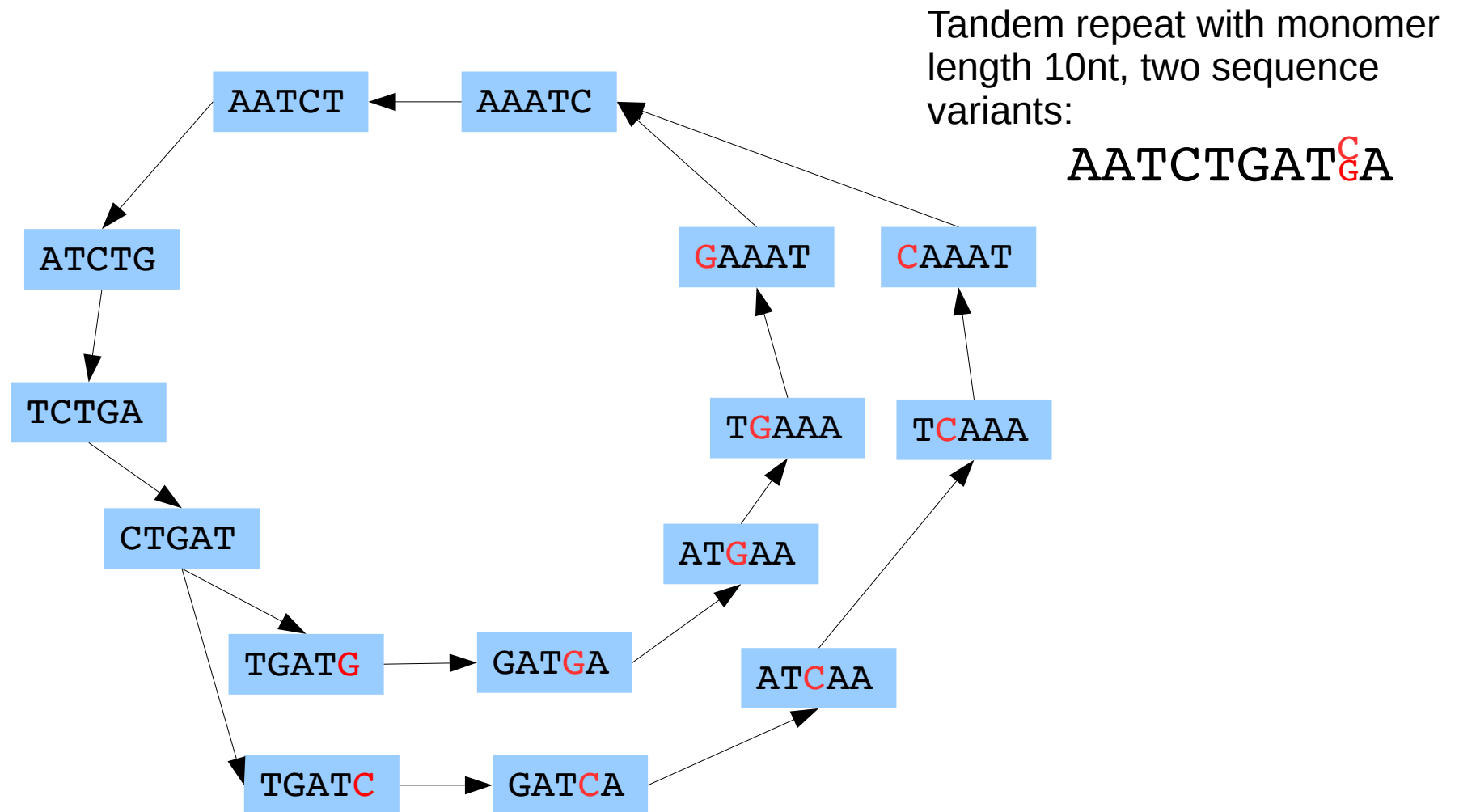


Tandem repeat with monomer
length 10nt:

AATCTGATGA

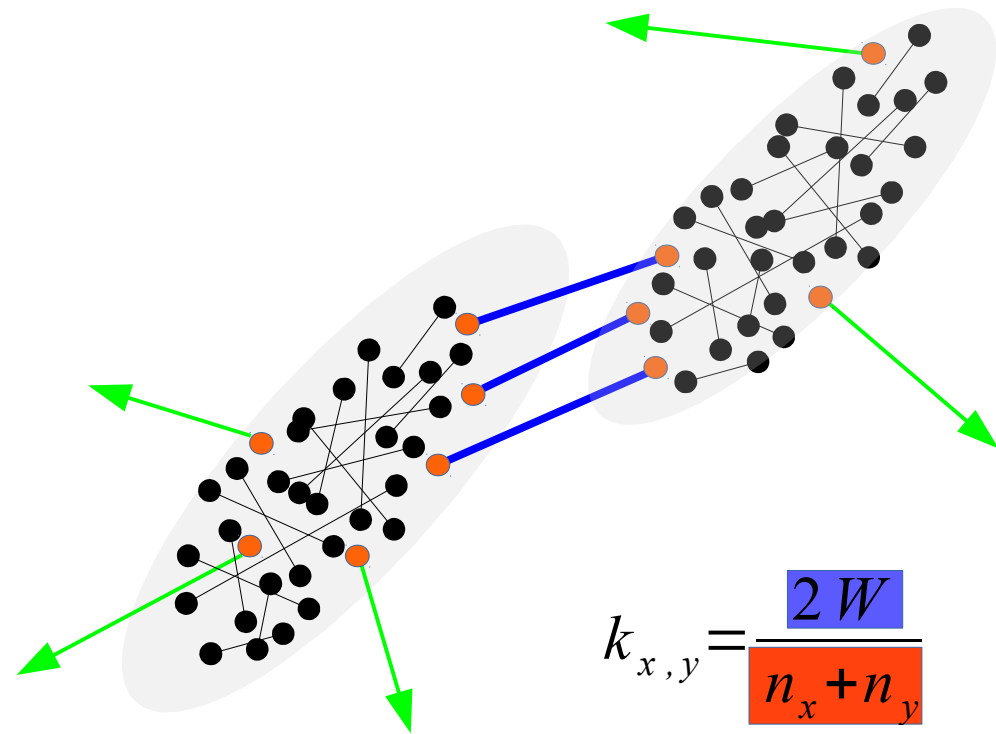
Tandem Repeat Analyzer - TAREAN

Consensus reconstruction:



Tandem Repeat Analyzer - TAREAN

Cluster merging



$$k_{x,y} = \frac{2W}{n_x + n_y}$$

Merging of related clusters

W number of reads pairs shared between clusters x and y
 n_x and n_y is number of reads in cluster x and cluster y with absent read mate within the same cluster respectively

Suitable $k_{x,y}$ cutoff 0.05 – 0.2

full connection: $k_{x,y} = 1$

no connection $k_{x,y} = 0$

Tandem Repeat Analyzer - TAREAN

Results: 500k paired reads, *P. sativum*

Run statistics:

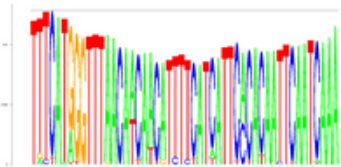
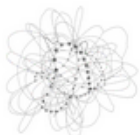
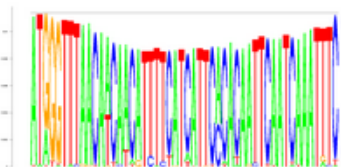
Number of input sequences: 4525544
Number of sequences in clusterings: 500000
Paired sequences: TRUE
Number of putative satellites: 8
Proportion of sequences in analyzed clusters : 66 %

Putative satellites:

Cluster	Genome Proportion[%]	Size real	Satellite probability	Consensus length	Consensus	Kmer analysis	Graph layout	Loop index	Pair completeness	Kmer coverage	V	E	Pbs score	The longest ORF length	Similarity based annotation
14	CL14	1.300	6470	0.9439340866644385	50	TTTCATGGGTTTAAACAACAACATTTTCATCATTACACAATTCAATCAAA	report	0.9913446676970634	0.9348086124401914	0.50	6470	6256255	0	82	0.02% organelle/plastid%
43	CL43	0.200	996	0.9441168259917678	65	CAAAACAAGAACAATAAACAATCAATCTTCTCATTTCTTGATAACAATATACAAAATTTAG	report	0.9879518072289156	0.9117082533589251	0.65	996	406309	0	43	
51	CL51	0.150	768	0.9935709746605441	54	TACGGACGAGTATTGACAAGAAGATGATTAAACGACGAGTGTTGAAAAACAA	report	0.9895833333333334	0.9692307692307692	0.51	768	205652	0	71	
58	CL58	0.120	622	0.9863672152615018	164	ATGAAATGTGGTAAAGTTGAAAGTAGGTACCAACATTTATCCTAAATATGGTATTTGTATACATGTTTTACCCGCATTT ATTAAGAAATGCGGTAAATTAACAAGAAAGTGCAACATTTGTCTTAAAGTATGTTGTTATAATGTTTTTCATGCATT TATT	report	0.9935691318327974	0.9872204472843451	0.58	622	42830	0	56	
84	CL84	0.059	297	0.9439340866644385	960	GTCCAAACCGTTGTGTTTCATCGTCTTCTCATGATGTTGGTACCAACCGCTACAATCACCCTCATCATCGTCATCAT TCTGGCAGCAGTTATAATCATCATCCTCTTCATTATTTCTTGAACTTCGTATCATCAACTTCGCAAAATCAAAACAA ATATAAGTTTCACTAAACTAAAAGTGAAATTTAAATAAGCTACAAAGATCTATTATTACCCCTGACTACTAGATTTTC CATATGGCTATCTATGACATCCATTTCCCTCTTAGTCTTCCAGCTTTTGTCCAATTGGTTATCTTTTCTGTATTA GACGTTGTTGTCTTCCAATTCCTCAATGAGGAGGCTACAAAAAATGATAAACAGAAATTAACAAATACAATGTGA GAAAGATCATAGTAATGTAAGAACTGTAACTTTAGTAATTCAAAAACCTAGAAAAATCAGTAAAAATGTAGTAATA CAATGGCTAAAAATGTAATCATCAACATGAAAACTATCAAGAAAAATGATTAAACGCGTGAGGAATTTGCAAAATCA AGAAAAAGGCAATTAACATAAGATCGTTAAATCATCATTATCTAACGCAAGAAACTATAAAATGCAAGCATGACAAA CTTGCAATTCAGAAAAAGAAGATCATATAATAGCAGTTCTTAAACAGAAACACGAGAAAAATGCAAAAAACATGC ATGATTATGGTTATGTTTATCTGCAACAAGAAAGTTTGTATAGGATGTTGTTATTTGAGCAAGAAAGTTCTGTATAG GATCATCGTTATCTGAGCAAGAAGGTTCTCAGTATGAGCAACTTCTCGGGCTGTTCTGTTGAGGTTCTTTGAGGGA GCTTCATTCAATGTCAGGTTTCTGATGATCATGTTTATCTGAGCAAGATAAGGTTCTGTACGGGAGCATACTATTG	report	0.9932659932659933	0.9285714285714286	0.40	297	5761	0	191	
89	CL89	0.055	277	0.9812880760410964	280	AGAAGTATTTAAAAATCGCAAAAAAGAGGTGCAACACAGGACTTCCAGGAGGTCAACCTCTAGTACTACTCTCG CCCAAGCACGCTTAACGCGGAGTTCTGATGGGATCCGGTGATTAGTGCTGGTATGATCGACCTGATATCTCATGCGC TTTGATTGTATATATTTTATTCTTTCACCTATTAAACGGCAAAACAAATAATTAAACGCTGAAATGCTGTTCTCTCT ACTTTTTCCGACTCTTTTACGCCGATTAAACGAATAAAAA	report	0.9819494584837545	0.9645390070921985	0.74	277	10488	0	61	56.32% 5S_rDNA/5S_rDNA%

Tandem Repeat Analyzer - TAREAN

Results: 500k paired reads, *P. sativum*

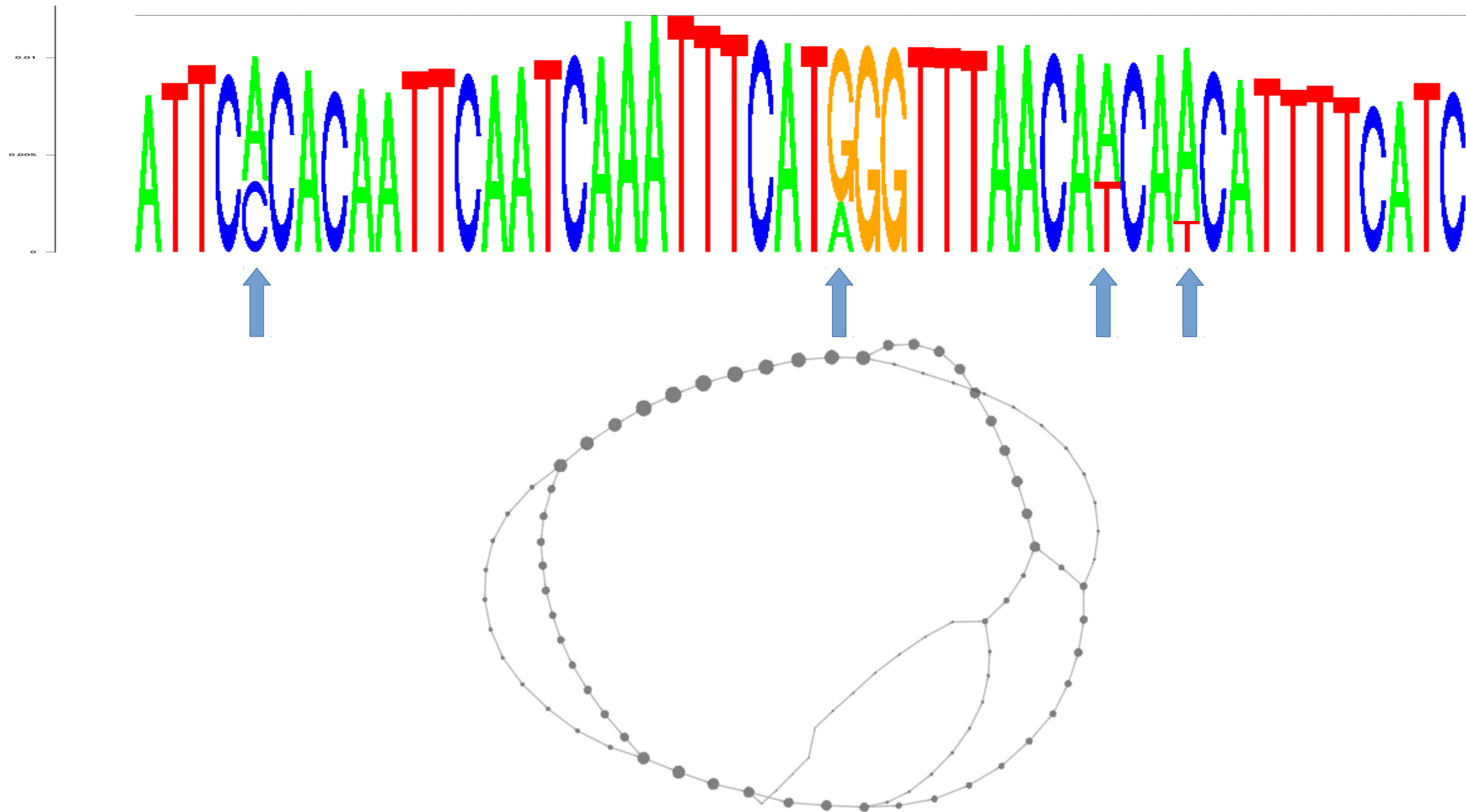
	kmer	variant	total_score	monomer_length	score_bn	consensus	graph_image	logo_image
1	15	2	0.503	50	8.3e-03	TTTCATGGGTTTAACAACAACATTTTCATCATTACACAATTCAATCAAA		
2	11	1	0.486	50	7.5e-03	ATTCACACAATTCAATCAAATTTTCATGGGTTTAACAACAACATTTTCATC		
3	19	2	0.479	50	8.4e-03	ATGGGGTTTAACAACAACATTTTCATCATTACACAATTCAATCAAATTTTC		

PisTRB

Tandem Repeat Analyzer - TAREAN

Results: 500k paired reads, *P. sativum*

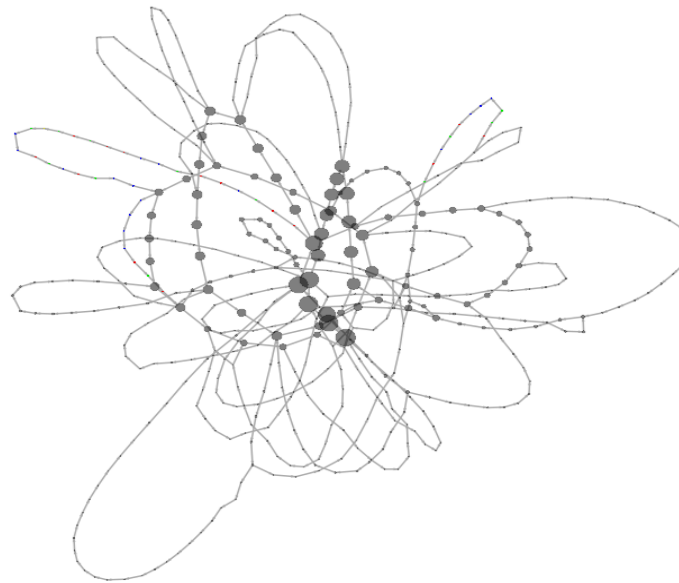
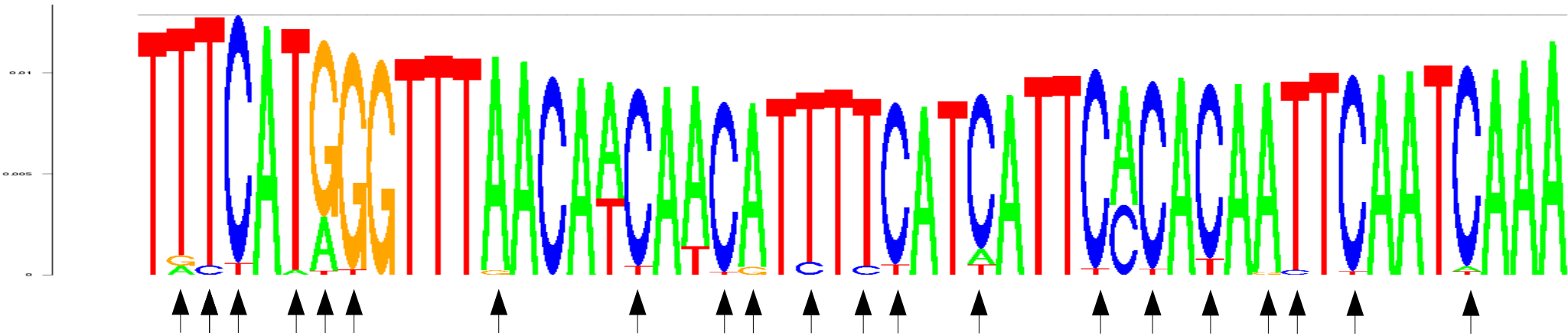
PisTRB – consensus based on 11mers



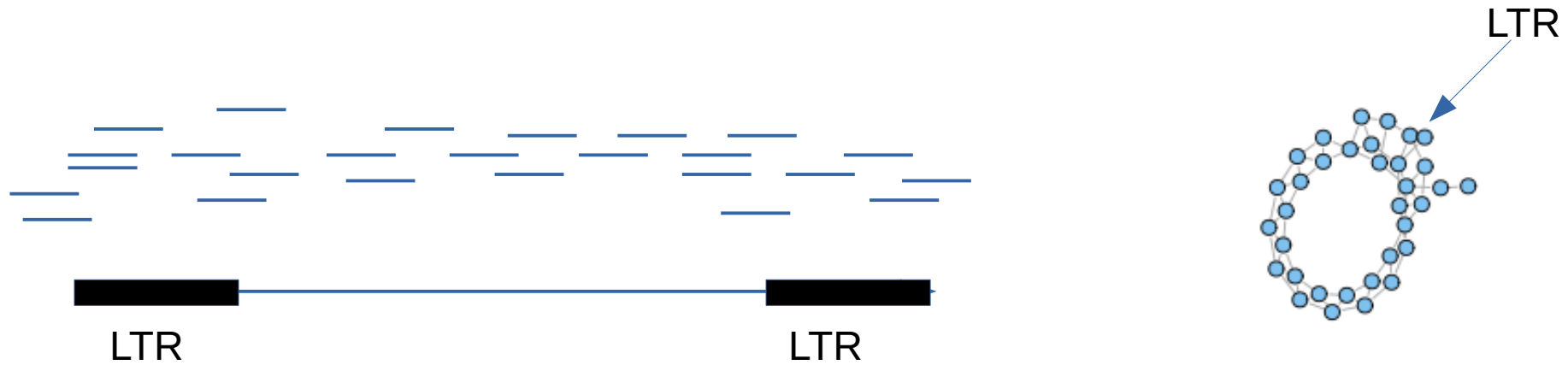
Tandem Repeat Analyzer - TAREAN

Results: 500k paired reads, *P. sativum*

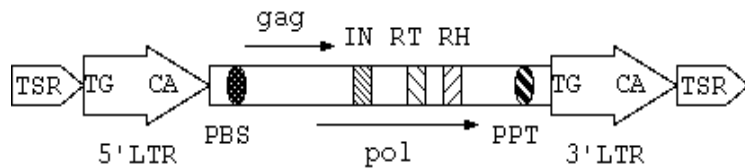
PisTRB – consensus based on 15mers



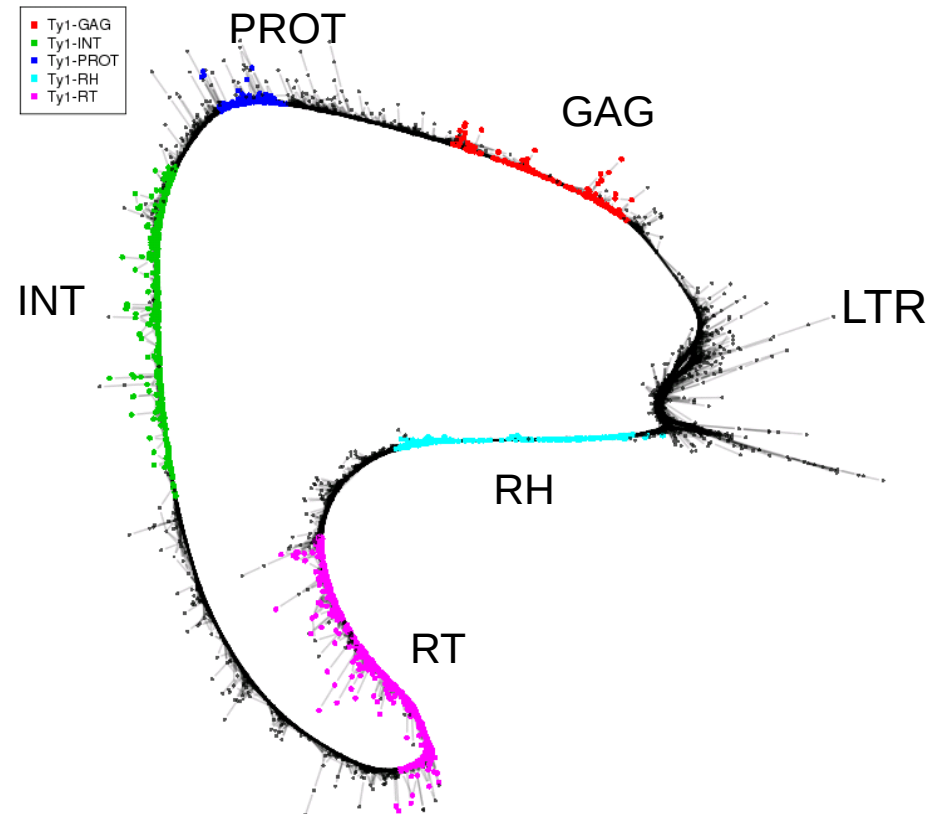
Tandem Repeat Analyzer - TAREAN



LTR-retrotransposons



- Circular layout, loop index $\rightarrow 1$
- frequently high pair completeness
- detection of long ORF
- Identification of PBS



Results: 500k paired reads, *P. fulvum*

Proportion of sequences in analyzed clusters : 71 %

Cluster	Genome Proportion[%]	Size real	Satellite probability	Consensus length	Consensus	Kmer analysis	Graph layout	Loop index	Pair completeness	Kmer coverage	V	E	Pbs score	The longest ORF length	Similarity based annotation
3	CL3	4.400	21863	0.7483552307194397	6264	TTTAGAAAAATCCCATTTTGGTAAAAAGTTATTTTCACGTGAAAGGGAAAAAGGCGAGAAAGGGCAAAAGCCAGAGAACGAAGGTTGAAGGAAGGAGAGCTTGGAGCTCGGAAGCTGCCGATTAACTCAGTGAAGGGGGTTATCATCGTTTATGGGTATTATGGGTTAAACATGTAATGGGTAGTGAAGCCGTTGAATTGACCTAATTCGGATTATGAATGCTGAAAAATTGTTGATGGATGAGTTATGTAAAACTGAAATTGAATCATAGTAGTGTGAGTCATAATTGCTGGAAGTGTAGCTTTTACGGAATTGGAATCGGAGTCCCGAAGTCTCTCAACGGCGGAAAAATGCGGAGAATCTGCATTCTGCTTGTGTAGCCGAGGAACAGCTTCTGTCTTGGCTTACCGGTTAACCGGTTAACCGGGCGTTAACCGGTTAACACTGTTGAAAAATGGAATAATGATATTTTAATGTGTGTACCTAATTTGGTGGTTGGCCTATATCGGTTGATGATATAGTAGGGATTATTTCCCGTGTGTTTGAATGATATAGGTTAGTAGAGTGTGCTAATACTGCTGATTGATTATGTGGCATGATGATATGCTGTTGTGATAAAATATGTGATGATGTGTGATGATGATGATGCTGTGAATGATGATTATGATGATGATGCTGTAATAGACTGTTTATGGCTTAGAGTGTGAGCATGTCTATTCTTGATATTGTTGTTGATGTTGCATTTGCTAGGTGATTAGCTTGACACTGTGCGCCCA	report	0.9568677674610072	0.8537391894183484	0.451451948051948	21863	1787265	11.279959999999999	1491	

[illegible]

Tandem Repeat Analyzer - TAREAN

- Simple repeat are not necessary detected due to dust filtering in all-to-all megablast step (e.g. telomeric)
- Clusters derived from some contaminants with circular DNA (plasmids, bacteriophages, viruses) look like tandem repeats → similarity based annotation
- Cluster fragmentation → merging
- ~500k paired reads sufficient for major satellite detection
- Alternative monomers, 'chimeric tandem repeats'

Tandem Repeat Analyzer - TAREAN

Galaxy implementation:

Tandem Repeat Analyzer (version 1.0.0)

paired NGS reads:

4: VGR_paired.fas

Input must be interlaced paired reads from paired-end sequencing in single fasta file. All pairs must be complete

Sample size:

80000

Threshold for cluster merging:

No merging

No merging

0.2

0.3

0.4

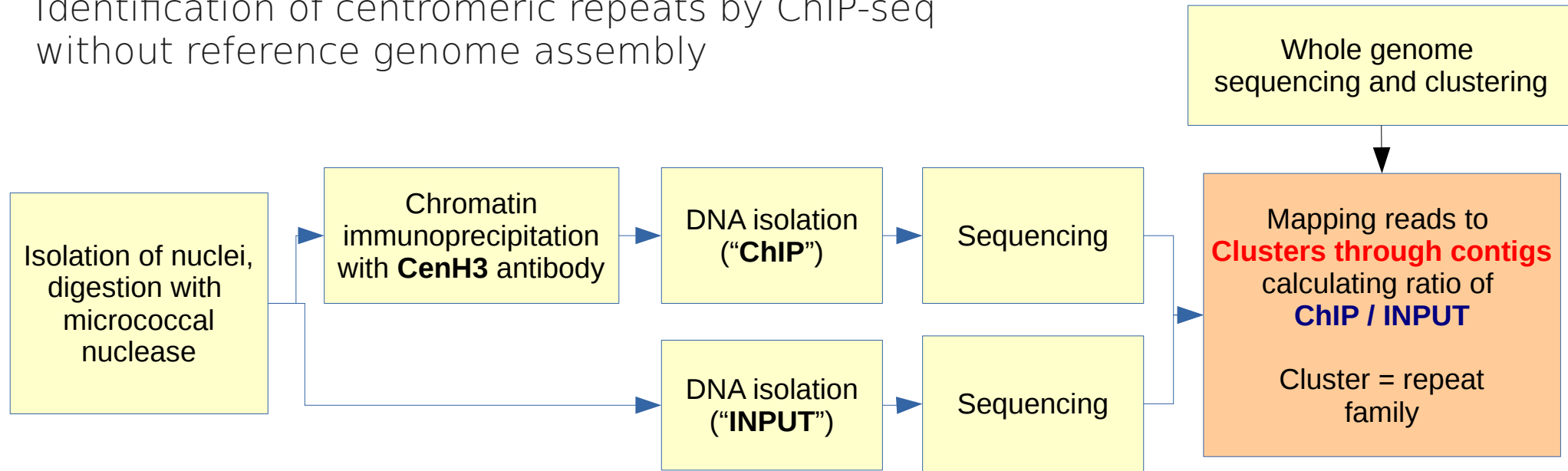
0.5

0.7

Tandem Repeat Analyzer

ChIP-Seq Mapper

Tool based on Neumann et al. 2012, PLoS Genetics
Identification of centromeric repeats by ChIP-seq
without reference genome assembly



Reference = contigs from clustering

Advantage:

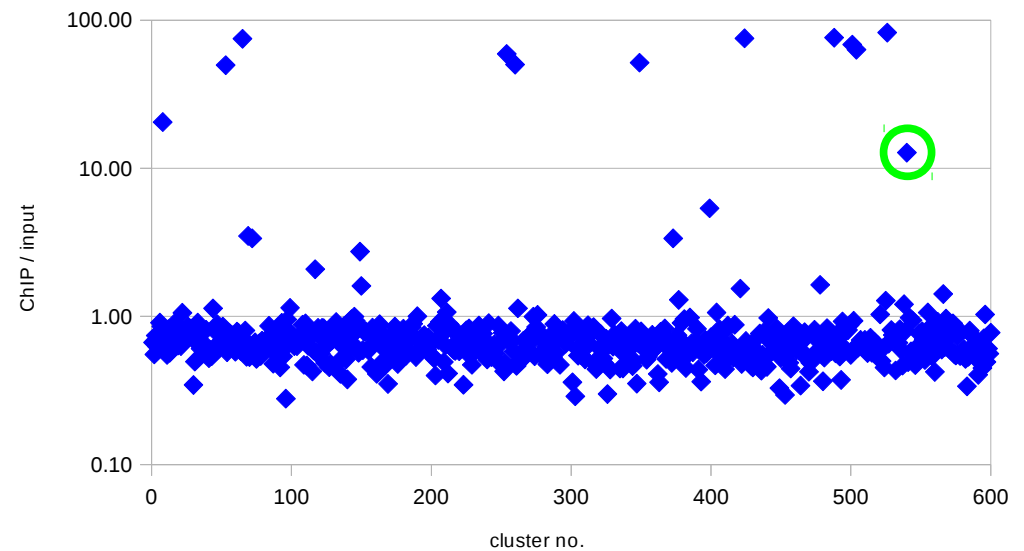
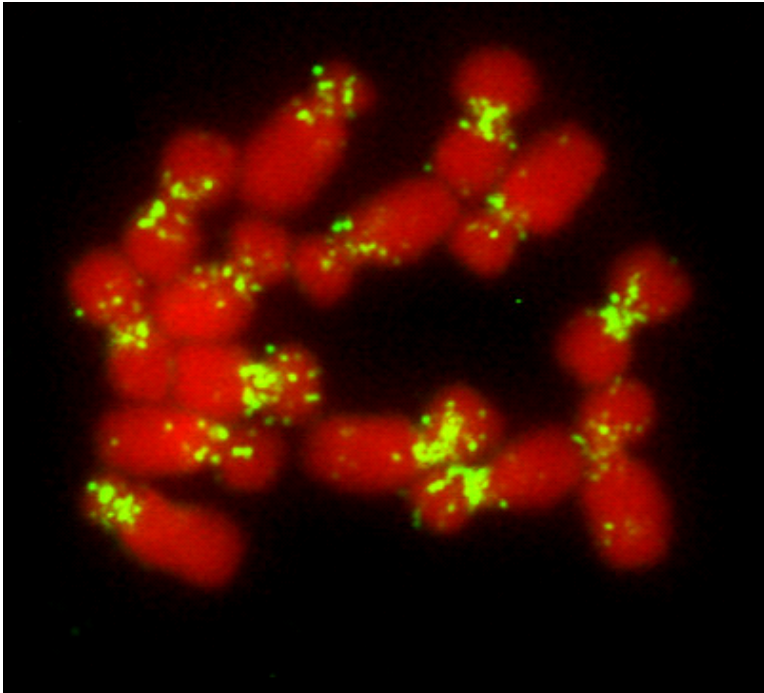
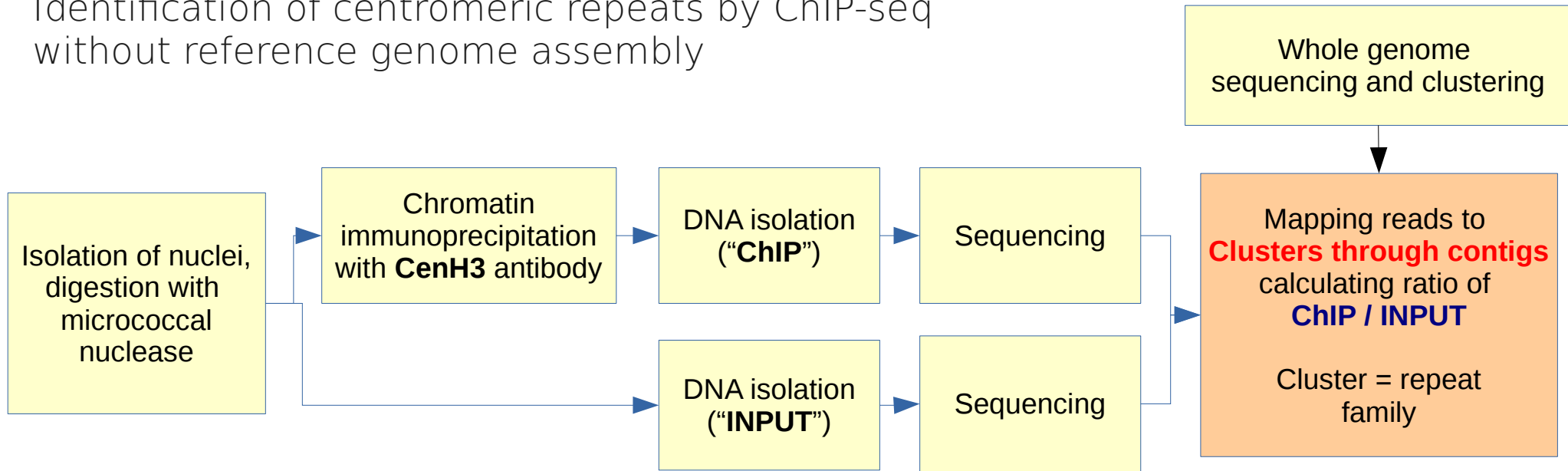
- relatively fast, less computationally demanding
- sufficient to identify all major centromeric repeats
- ChIP enrichment value is calculated for all clusters

Disadvantage:

- misses most single/low copy sequences
- insufficient to identify low-abundant repeats

ChIP-Seq Mapper

Tool based on Neumann et al. 2012, PloS Genetics
Identification of centromeric repeats by ChIP-seq
without reference genome assembly



ChIP-Seq Mapper

Tool based on Neumann et al. 2012, PloS Genetics
Identification of centromeric repeats by ChIP-seq
without reference genome assembly

Galaxy implementation:

ChIP-Seq Mapper (version 0.1.0)

Chip Sequences:

1: Input_100k

NGS data in fasta format

Input Sequences:

2: ChIP_100k

NGS data in fasta format

Reference - Contig Sequences:

3: Contigs from clustering_minRD5

Contigs obtained from RepeatExplorer clustering pipeline in fasta file

Number of clusters to be shown in graph:

200

Execute

Reference – Contigs from clustering

>CL10Contig2870 (546-72.8-39765)

ATGGAAGCTGGAAAAAATCAAGTCATCTGAAAAAGTCAA
CTACAGAGTCAAACCTTTGACTTTTTGGGGTTTTGGTCAA
ATTTGAATTTCTCAAGTCAAGTATCATGGAAATATGT

>CL3Contig28 (156-32.8-76565)

CTTATTTTCAACCCCAATTGGCTTGTGCTTCATCATCAT
ATACATTCATACATATCATTGGTTCTATGGTTATGGGAT
CAAGGTATCTTTCTATTGGTTTTTTCAAGCACTAGAAGA
TGAAGATCTGATTATCAGTGCAAGGAGGATGGTATATCC
TCATGTGTCA

...

1	2	3	4
"Cluster"	"Chip_Hits"	"Input_Hits"	"Ratio Chip/Input"
0	631	3225	0.195658914728682
1	1256	1780	0.70561797752809
2	189	992	0.190524193548387
3	185	897	0.206243032329989
4	158	711	0.222222222222222
5	148	897	0.164994425863991
6	197	866	0.227482678983834
7	171	870	0.196551724137931
8	76	737	0.103120759837178
9	143	722	0.198060041828255

ChIP-Seq Mapper

Tool based on Neumann et al. 2012, PloS Genetics
Identification of centromeric repeats by ChIP-seq
without reference genome assembly

Galaxy implementation:

ChIP-Seq Mapper (version 0.1.0)

ChIP Sequences:

1: Input_100k

NGS data in fasta format

Input Sequences:

2: ChIP_100k

NGS data in fasta format

Reference - Contig Sequences:

3: Contigs from clustering_minRD5

Contigs obtained from RepeatExplorer clustering pipeline in fasta file

Number of clusters to be shown in graph:

200

Execute

ChIP and Input data:

Uniform length
~ 100 nt

About 10^6 reads is sufficient

New tools availability

<https://galaxy-elixir.cerit-sc.cz/>

Development status – alpha version

Use it and report bugs!

