Zürcher Hochschule
für Angewandte Wissenschaften

A study of dynamics of Indels using ProPIP, PRANK and MAFFT

Student: Eldhose Poulose, Master's Thesis 2020
Supervisors: Dr. Manuel Gil, Dr. Maria Anisimova, Dr. Massimo Maiolo
ACGT Department, Institute of Applied Simulation (IAS), ZHAW, Campus Reidbach, 8820 Wädenswil, Switzerland

Abstract

Evolutionary changes happens over time in the genomes of species. Two of the most important mechanisms of evolutionary changes are substitutions and indels (insertions and deletions). Analyses of genomic sequences typically rely on multiple sequence alignments (MSA), which infer indels. The Applied Computational Genomics Team has developed an MSA estimator (ProPIP). It relies on an explicit evolutionary model of indel (as opposed to more traditional aligners) termed Poisson Indel Process (PIP).

In this thesis we analyse and compare the MSAs inferred by ProPIP and two state of art aligners PRANK and MAFFT . We use simulated as well as real data in this study. In particular, we are interested in the suitability of PIP (which is a single character indel model) to infer long indels. We also examine the location of the indel events on the phylogenies implied by the different aligners.

MSA Evaluation Methods

Main Results

(100,8)	True(id)	MAFFT v7.453		PRANK v.170427		ProPIP	
		id	Xid	id	Xid	id	Xid
nIndels	11511	9279	8971	10226	9692	19539	14418
Max-IL	37	45	45	38	38	30	33
Mean	3.116	3.654	3.780	3.306	3.488	2.041	2.767
Median	1	2	2	2	2	1	1
SD	3.850	4.394	4.506	3.307	3.489	2.132	3.321

Table 6.1: The summary statistics of the 'true' Indel length distribution of INDELible data (True(id)) is compared with Indel length and Indel block distribution statistics (See Section 5.2 and 5.3) generated by MAFFT v7.453, PRANK v.17042, and ProPIP. Note: The 'id' represents indel length distribution and 'Xid' represents indel block distribution

In particular, we are interested in the suitability of PIP (which is a single character indel model) to infer long indels. We also examine the location of the indel events on the phylogenies implied by the different aligners.

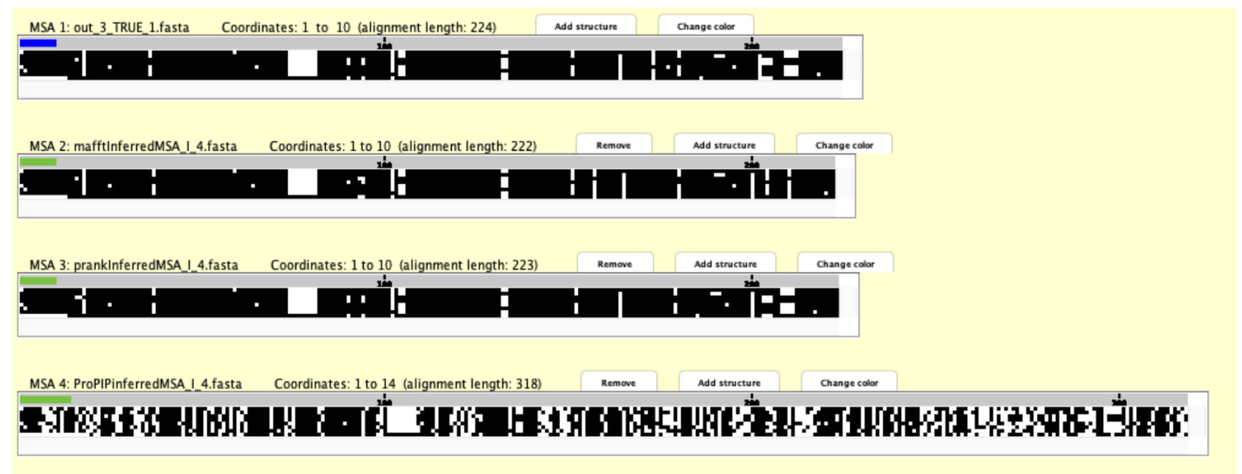
MSA 1: out_3_TRUE_1.fasta Coordinates: 1 to 10 (alignment length: 224)
Add structure
Change color
MSA 2: mafftInferredMSA_I_4.fasta Coordinates: 1 to 10 (alignment length: 222)
Remove
Add structure
Change color
MSA 3: prankInferredMSA_I_4.fasta Coordinates: 1 to 10 (alignment length: 223)
Remove
Add structure
Change color
MSA 4: ProPIPinferredMSA_I_4.fasta Coordinates: 1 to 14 (alignment length: 318)
Remove
Add structure
Change color

Figure 6.1: The Pixel Plot[1] (Section 5.5). A simulated 'true' MSA using INDELible is compared with MSA's generated by MAFFT v7.453 (MSA 2), PRANK v.17042 (MSA 3), and ProPIP (MSA 4). Note: black pixel represents Characters and white pixel represents Indels.

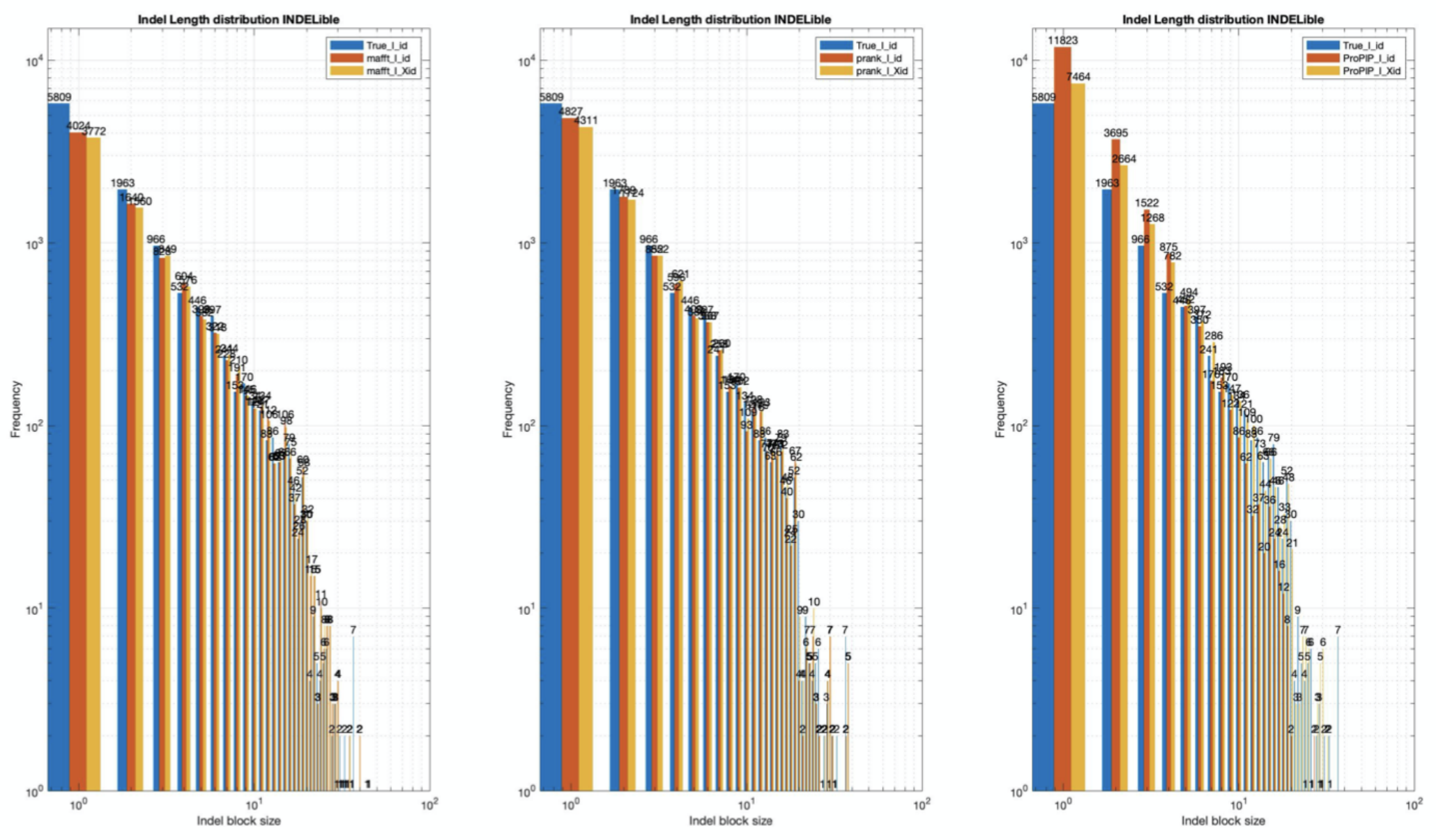

Figure 6.2: The Log-Log Plot. The 'true' Indel length distribution of INDELible data is compared with Indel length and Indel block distribution (See Section 5.2 and 5.3) generated by MAFFT v7.453, PRANK v.17042, and ProPIP. Note: The legend blue represents 'true' Indel length distribution, red represents inferred indel length distribution and yellow represents the inferred indel block distribution.

(100,8)	True (id)	k0.05		k0.10		k0.25		k0.50		k2		k3	
		id	Xid	id	Xid	id	Xid	id	Xid	id	Xid	id	Xid
nIndels	11511	13725	10466	13745	10618	14562	11060	16030	12231	25071	17307	30713	20091
Max-IL	37	30	30	30	30	30	30	30	32	43	105	37	118
Mean	3.116	2.119	2.779	2.161	2.797	2.140	2.817	2.102	2.755	2.090	3.028	2.1663	3.312
Median	1	1	1	1	1	1	1	1	1	1	2	1	2
SD	3.850	2.207	3.316	2.261	3.352	2.226	3.376	2.255	3.294	2.230	3.985	2.347	4.740

Table 6.3: The summary statistics of the 'true' Indel length distribution of INDELible data (True(id)) is compared with Indel length and Indel block distribution statistics (See Section 5.2 and 5.3) generated by ProPIP with k=0.05, ProPIP with k=0.10, ProPIP with k=0.25, ProPIP with k=0.50, ProPIP with k=2, ProPIP with k=3. Note: The 'id' represents indel length distribution and 'Xid' represents indel block distribution.

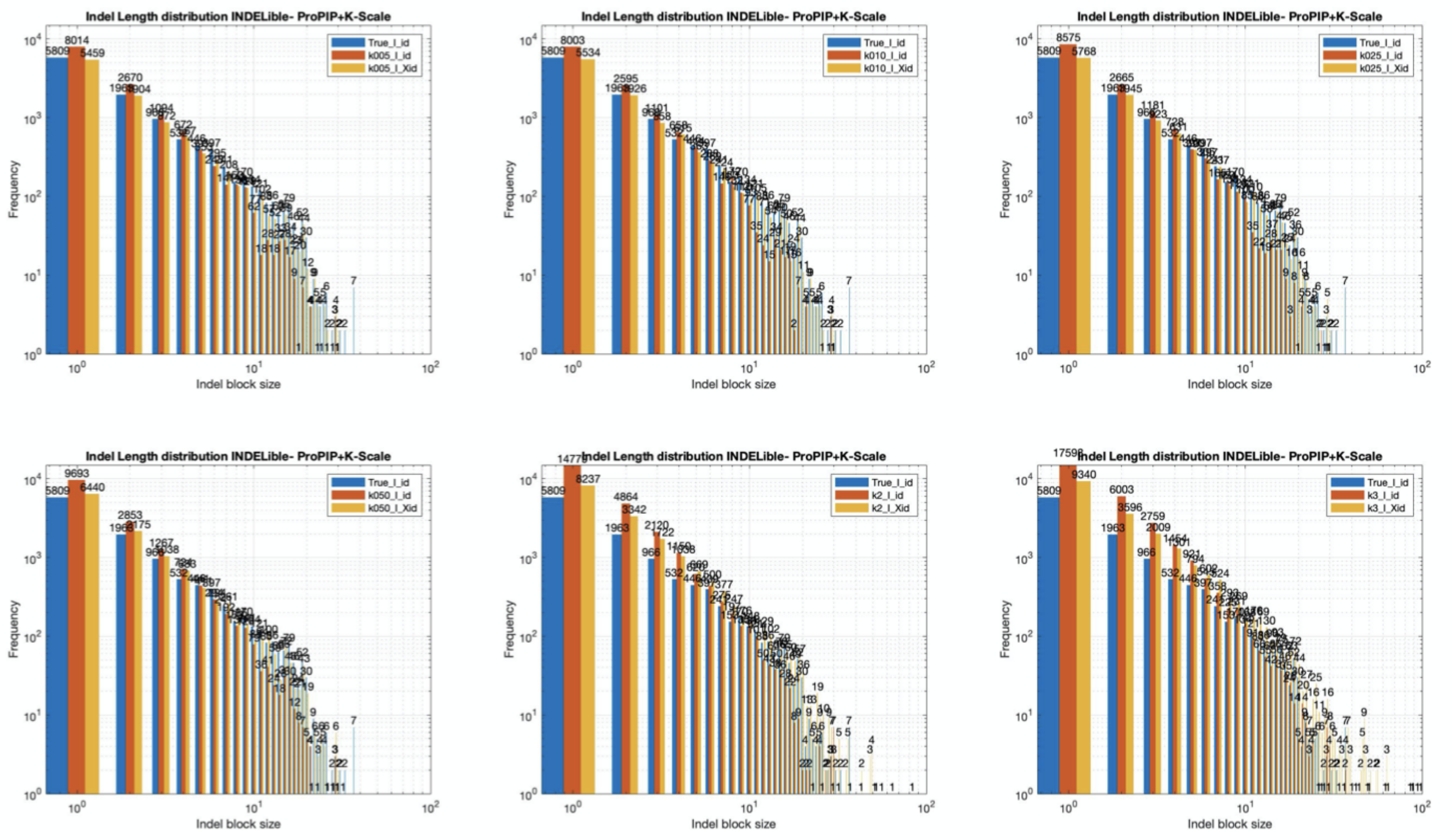

Figure 6.6: The Log-Log Plot. The 'true' Indel length distribution of INDELible data is compared with Indel length and Indel block distribution (See Section 5.2 and 5.3) generated by (from top-left to bottom-right) ProPIP with k=0.05, ProPIP with k=0.10, ProPIP with k=0.25, ProPIP with k=0.50, ProPIP with k=2, and ProPIP with k=3. Note: The legend blue represents 'true' Indel length distribution, red represents inferred indel length distribution and yellow represents the inferred indel block distribution.

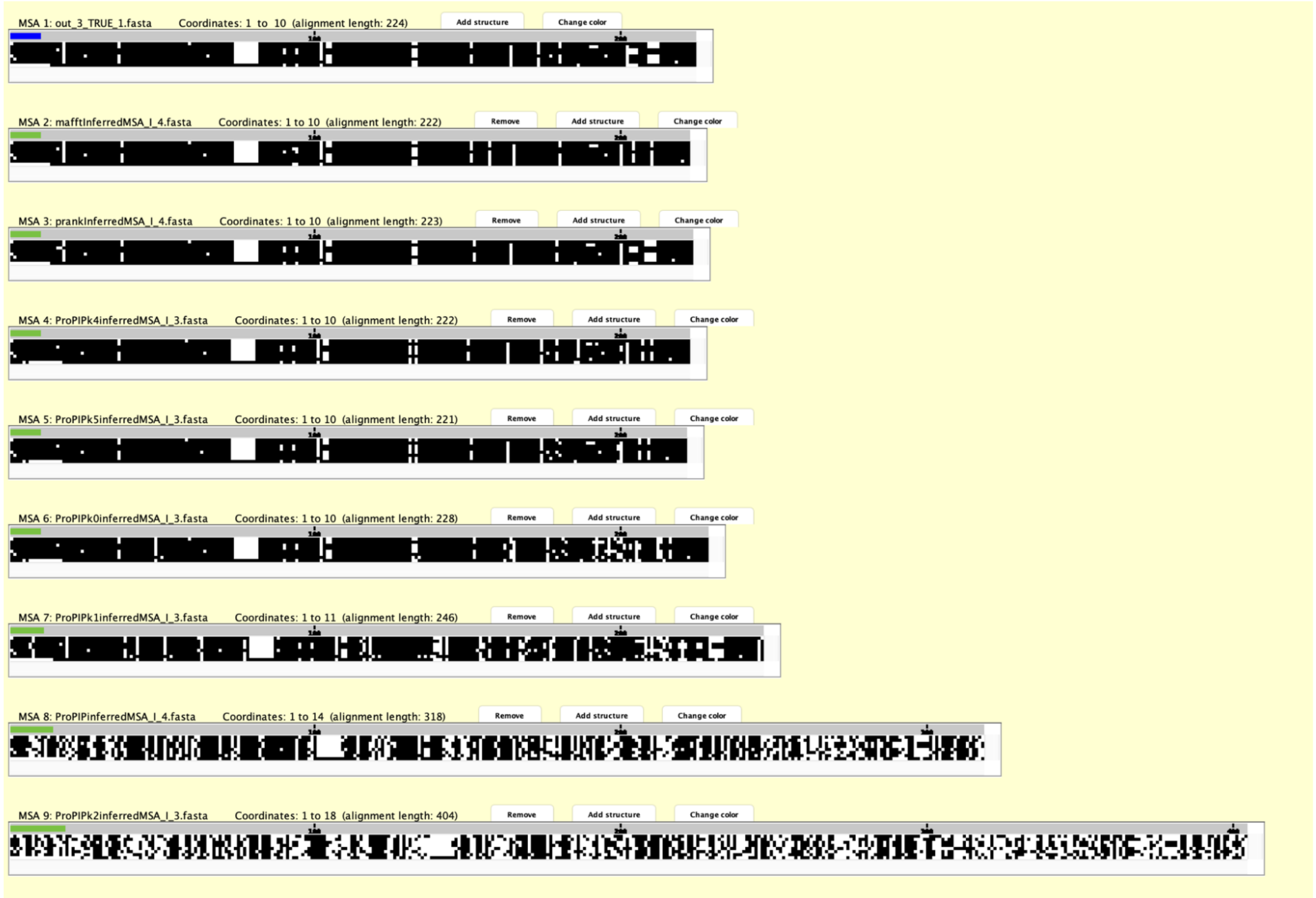
MSA 1: out_3_TRUE_1.fasta Coordinates: 1 to 10 (alignment length: 224)
Add structure
Change color
MSA 2: mafftInferredMSA_I_4.fasta Coordinates: 1 to 10 (alignment length: 222)
Remove
Add structure
Change color
MSA 3: prankInferredMSA_I_4.fasta Coordinates: 1 to 10 (alignment length: 223)
Remove
Add structure
Change color
MSA 4: ProPIPk4inferredMSA_I_3.fasta Coordinates: 1 to 10 (alignment length: 222)
Remove
Add structure
Change color
MSA 5: ProPIPk5inferredMSA_I_3.fasta Coordinates: 1 to 10 (alignment length: 221)
Remove
Add structure
Change color
MSA 6: ProPIPk0inferredMSA_I_3.fasta Coordinates: 1 to 10 (alignment length: 228)
Remove
Add structure
Change color
MSA 7: ProPIPk1inferredMSA_I_3.fasta Coordinates: 1 to 11 (alignment length: 246)
Remove
Add structure
Change color
MSA 8: ProPIPinferredMSA_I_4.fasta Coordinates: 1 to 14 (alignment length: 318)
Remove
Add structure
Change color
MSA 9: ProPIPk2inferredMSA_I_3.fasta Coordinates: 1 to 18 (alignment length: 404)
Remove
Add structure
Change color

Figure 6.5: The Pixel Plot[1] (Section 5.5). A simulated 'true' MSA using INDELible is compared with MSA's generated by MAFFT v7.453 (MSA 2), PRANK v.170427 (MSA 3), ProPIP with k=0.05 (MSA 4), ProPIP with k=0.10 (MSA 5), ProPIP with k=0.25 (MSA 6), ProPIP with k=0.50 (MSA 7), ProPIP with k=1 (MSA 8), and ProPIP with k=2 (MSA 9). Note: black pixel represents Characters and white pixel represents Indels.

Conclusion and Future works

References

1. M.Maiolo, X.Zhang, M.Gil, and M.Anisimova. Progressive multiple sequence alignment with indel evolution.BMC Bioinformatics, 19(1):1–8, 2018.
2. M.Maiolo Dissertation, https://serval.unil.ch/en/notice/serval:BIBD24577D3A885, 2019.
3. L.Gatti M.Maiolo, ProPIP, castor aligner, manual, https://github.com/acg-team/castoraligner, 2017.
4. W.Fletcher and Z.Yang, INDELible:A flexible simulator of biological sequence evolution.Molecular Biology and Evolution, 26(8):1879–1888, 2009
5. C.L.Anderson, C.L.Strope, and E.N.Moriyama. SuiteMSA:Visual tools for multiplesequence alignment comparison and molecular sequence simulation.BMC Bioinformatics,