

Table 1: Average runtime, Pearson’s correlation coefficient and Fowlkes-Mallows index for $k = 4$ and $k = 8$ and the neighbor-joining algorithm. The ‘reference’ method calculates distances from the original sequences. We show only assembly algorithms that gave the highest and the lowest correlation. From d -type measures, the one with the highest correlation is selected.

Dataset	method	finished	assem. ms	distances ms	corr.	B_4	B_8	FM/upgmak4	FM/upgmak8
Influenza	reference	112/112	0	2,602	1	1	1	1	1
	$\max(R_A , R_B)$	112/112	0	335	.801	.66	.32	0.67	0.32
	$\text{Dist}_{\text{MESSG}}(R_A, R_B)$	107/112	0	899,270	.983	1	1	1	1
	$\text{Dist}_{\text{MESSGq}}(R_A, R_B)$	112/112	0	50,808	.966	1	.97	1	0.98
	$\text{Dist}_{\text{CC}}(C_A, C_B)$ ABySS	87/112	21,628	17,469	.951	.98	.8	0.98	0.87
	$\text{Dist}_{\text{CC}}(C_A, C_B)$ Edena	72/112	285	18,483	.96	1	.85	1	0.88
	$\text{Dist}_{\text{CC}}(C_A, C_B)$ SSAKE	68/112	2,079	17,735	.944	.97	.84	0.97	0.87
	$\text{Dist}_{\text{CC}}(C_A, C_B)$ Velvet	110/112	385	23,567	.958	.99	.91	0.99	0.91
	$\text{Dist}_{\text{CC}}(C_A, C_B)$ SPAdes	43/112	13,529	22,661	.973	.99	.93	0.99	0.95
	$\text{Dist}_{\text{CC}}(C_A, C_B)$ opt ₁₀	112/112	−1	818	.57	.69	.4	0.75	0.39
	$\text{Dist}_{\text{CC}}(C_A, C_B)$ opt ₂₀	112/112	−1	2,927	.814	.88	.5	0.92	0.52
	$\text{Dist}_{\text{CC}}(C_A, C_B)$ opt ₃₀	112/112	−1	5,167	.89	.97	.6	0.99	0.62
	$\text{Dist}_{\text{CC}}(C_A, C_B)$ opt ₄₀	112/112	−1	7,516	.918	.99	.71	1	0.75
	$\text{Dist}_{\text{CC}}(C_A, C_B)$ opt ₅₀	112/112	−1	11,514	.943	1	.84	1	0.88
	$\text{Dist}_{\text{CC}}(C_A, C_B)$ opt ₆₀	112/112	−1	16,060	.964	1	.93	1	0.94
	$\text{Dist}_{\text{CC}}(C_A, C_B)$ opt ₇₀	112/112	−1	18,691	.972	1	.96	1	0.97
	$\text{Dist}_{\text{CC}}(C_A, C_B)$ opt ₈₀	112/112	−1	22,920	.981	1	.97	1	0.98
	$\text{Dist}_{\text{CC}}(C_A, C_B)$ opt ₉₀	112/112	−1	25,813	.984	1	.96	1	0.98
	$\text{Dist}(RC_A, RC_B)$ ABySS	112/112	20,900	508,511	.979	1	1	1	1
	$\text{Dist}(RC_A, RC_B)$ Edena	112/112	233	430,385	.98	1	1	1	1
	$\text{Dist}(RC_A, RC_B)$ SSAKE	112/112	1,655	552,860	.98	1	1	1	1
	$\text{Dist}(RC_A, RC_B)$ Velvet	111/112	378	749,033	.971	1	.99	1	0.99
	$\text{Dist}(RC_A, RC_B)$ SPAdes	112/112	12,380	625,883	.983	1	1	1	1
	$\text{Dist}(RC_A, RC_B)$ opt ₁₀	108/112	−1	961,024	.658	.81	.76	0.84	0.85
	$\text{Dist}(RC_A, RC_B)$ opt ₂₀	108/112	−1	953,892	.92	.97	.95	0.97	0.97
	$\text{Dist}(RC_A, RC_B)$ opt ₃₀	108/112	−1	922,462	.972	1	.99	1	1
	$\text{Dist}(RC_A, RC_B)$ opt ₄₀	109/112	−1	916,372	.971	1	.99	1	1
	$\text{Dist}(RC_A, RC_B)$ opt ₅₀	111/112	−1	907,207	.973	1	.99	1	0.99
	$\text{Dist}(RC_A, RC_B)$ opt ₆₀	112/112	−1	886,136	.975	1	.99	1	0.99
	$\text{Dist}(RC_A, RC_B)$ opt ₇₀	112/112	−1	821,487	.978	1	.99	1	0.99
	$\text{Dist}(RC_A, RC_B)$ opt ₈₀	112/112	−1	693,091	.982	1	.99	1	1
	$\text{Dist}(RC_A, RC_B)$ opt ₉₀	112/112	−1	534,738	.984	1	.99	1	0.99
	Mash	112/112	0	101	.679	.44	.61	0.48	0.58
	d_2	112/112	0	335	.708	.43	.99	0.3	0.67
	d_2^*	112/112	0	389	.837	.4	.9	0.38	0.71
	d_2^{q*}	112/112	0	328	.631	.32	.27	0.37	0.11
	D_2	112/112	0	374	.443	.32	$1.29 \cdot 10^{-3}$	0.68	0.32
	D_2^*	112/112	0	318	−.102	.44	$2.2 \cdot 10^{-3}$	0.5	$3.92 \cdot 10^{-3}$
	D_2^{q*}	112/112	0	312	0	.32	.27	0.37	0.12
	d_2^q	112/112	0	281	.631	.32	.27	0.37	0.11
	D_2^q	112/112	0	327	0	.32	.27	0.37	0.12
	longest contig ABySS	87/112	21,628	1,182	.67	.62	.43	0.63	0.43
	longest contig Edena	72/112	285	1,055	.675	.62	.46	0.64	0.47
	longest contig SSAKE	68/112	2,079	785	.664	.61	.36	0.59	0.35
	longest contig Velvet	110/112	385	38	.569	.46	.23	0.45	0.23
	longest contig SPAdes	43/112	13,529	1,465	.751	.71	.56	0.74	0.56
	longest contig opt ₁₀	112/112	−1	8	.503	.46	.12	0.45	0.11
	longest contig opt ₂₀	112/112	−1	11	.502	.44	.13	0.45	0.13
	longest contig opt ₃₀	112/112	−1	28	.512	.45	.12	0.47	0.13
	longest contig opt ₄₀	112/112	−1	23	.513	.45	.13	0.47	0.13
	longest contig opt ₅₀	112/112	−1	226	.542	.47	.18	0.49	0.18
	longest contig opt ₆₀	112/112	−1	283	.544	.49	.19	0.5	0.19
	longest contig opt ₇₀	112/112	−1	548	.586	.53	.27	0.53	0.26
	longest contig opt ₈₀	112/112	−1	702	.609	.57	.34	0.6	0.35
	longest contig opt ₉₀	112/112	−1	977	.666	.63	.43	0.65	0.44
Hepatitis	reference	9/9	0	1,748,984	1	1	1	1	1
	$\max(R_A , R_B)$	9/9	0	29,340	.181	.72	.83	0.55	0.37
	$\text{Dist}_{\text{MESSG}}(R_A, R_B)$	9/9	0	42,332,682	.965	1	.9	0.99	0.95
	$\text{Dist}_{\text{MESSGq}}(R_A, R_B)$	9/9	0	1,118,585	.897	1	.94	0.91	0.94
	$\text{Dist}_{\text{CC}}(C_A, C_B)$ ABySS	9/9	35,145	48,256,963	.949	1	.91	0.95	0.94
	$\text{Dist}_{\text{CC}}(C_A, C_B)$ Edena	9/9	7,038	44,548,818	.892	1	.84	0.95	0.91
	$\text{Dist}_{\text{CC}}(C_A, C_B)$ SSAKE	5/9	69,156	55,880,178	.901	1	.95	0.98	0.95
	$\text{Dist}_{\text{CC}}(C_A, C_B)$ Velvet	4/9	11,090	59,898,794	.98	1	.99	1	0.97
	$\text{Dist}_{\text{CC}}(C_A, C_B)$ SPAdes	2/9	76,514	31,517,537	.869	1	.89	0.87	0.93
	$\text{Dist}(RC_A, RC_B)$ ABySS	0/9	NaN	NaN	NaN	NaN	NaN	NaN	NaN
	$\text{Dist}(RC_A, RC_B)$ Edena	2/9	2,326	255,000,727	.995	1	.91	1	0.95
	$\text{Dist}(RC_A, RC_B)$ SSAKE	0/9	NaN	NaN	NaN	NaN	NaN	NaN	NaN
	$\text{Dist}(RC_A, RC_B)$ Velvet	0/9	NaN	NaN	NaN	NaN	NaN	NaN	NaN
	$\text{Dist}(RC_A, RC_B)$ SPAdes	0/9	NaN	NaN	NaN	NaN	NaN	NaN	NaN
	Mash	9/9	0	2,350	.967	1	.92	0.96	0.97
	d_2	7/9	0	27,145	.973	1	.86	0.98	0.96
	d_2^*	7/9	0	28,189	.972	1	.9	0.89	0.97
	d_2^{q*}	9/9	0	29,296	.972	1	.9	0.89	0.97
	D_2	8/9	0	30,458	−.181	.93	.43	0.66	0.54
	D_2^*	7/9	0	27,718	−.783	.93	.41	0.66	0.54
	D_2^{q*}	9/9	0	27,151	−.782	.92	.4	0.66	0.54
	d_2^q	9/9	0	27,885	.973	1	.87	0.98	0.96
	D_2^q	7/9	0	31,481	−.187	.93	.41	0.66	0.54
	longest contig ABySS	9/9	35,145	2,493,455	.53	.95	.69	0.65	0.69
	longest contig Edena	9/9	7,038	1,581,613	.515	.92	.76	0.78	0.8
	longest contig SSAKE	5/9	69,156	764,321	.334	.86	.62	0.71	0.66
	longest contig Velvet	4/9	11,090	515	.296	.92	.47	0.64	0.58
	longest contig SPAdes	2/9	76,514	3,242,365	.395	.87	.78	0.82	0.77

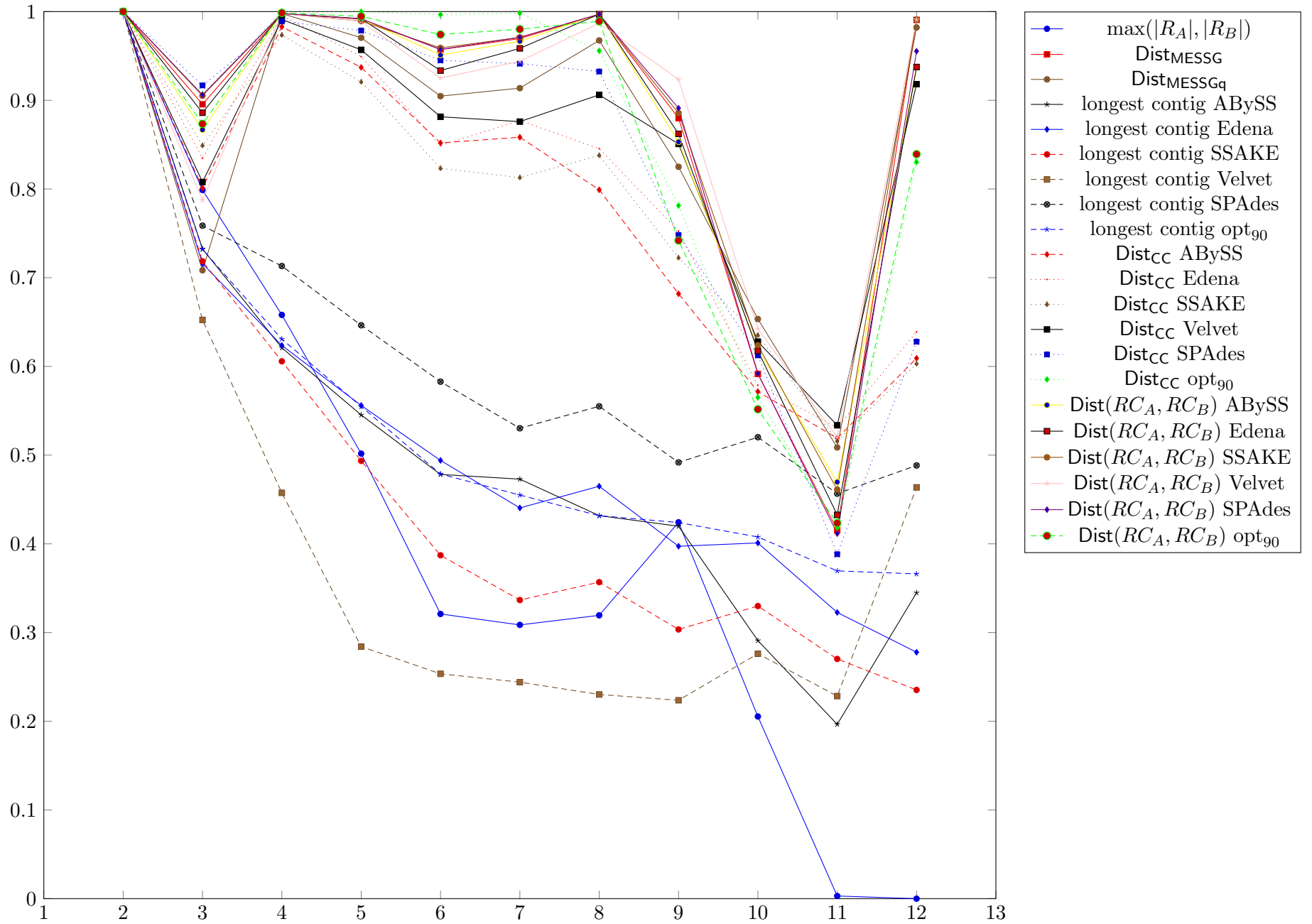


Figure 1: Plot of Fowlkes-Mallows index B_k versus k on influenza dataset. The index compares trees generated by the neighbor-joining algorithm. The tree is compared with the tree generated from the original sequences. If all values are equal to 1, the structures of the trees are the same.

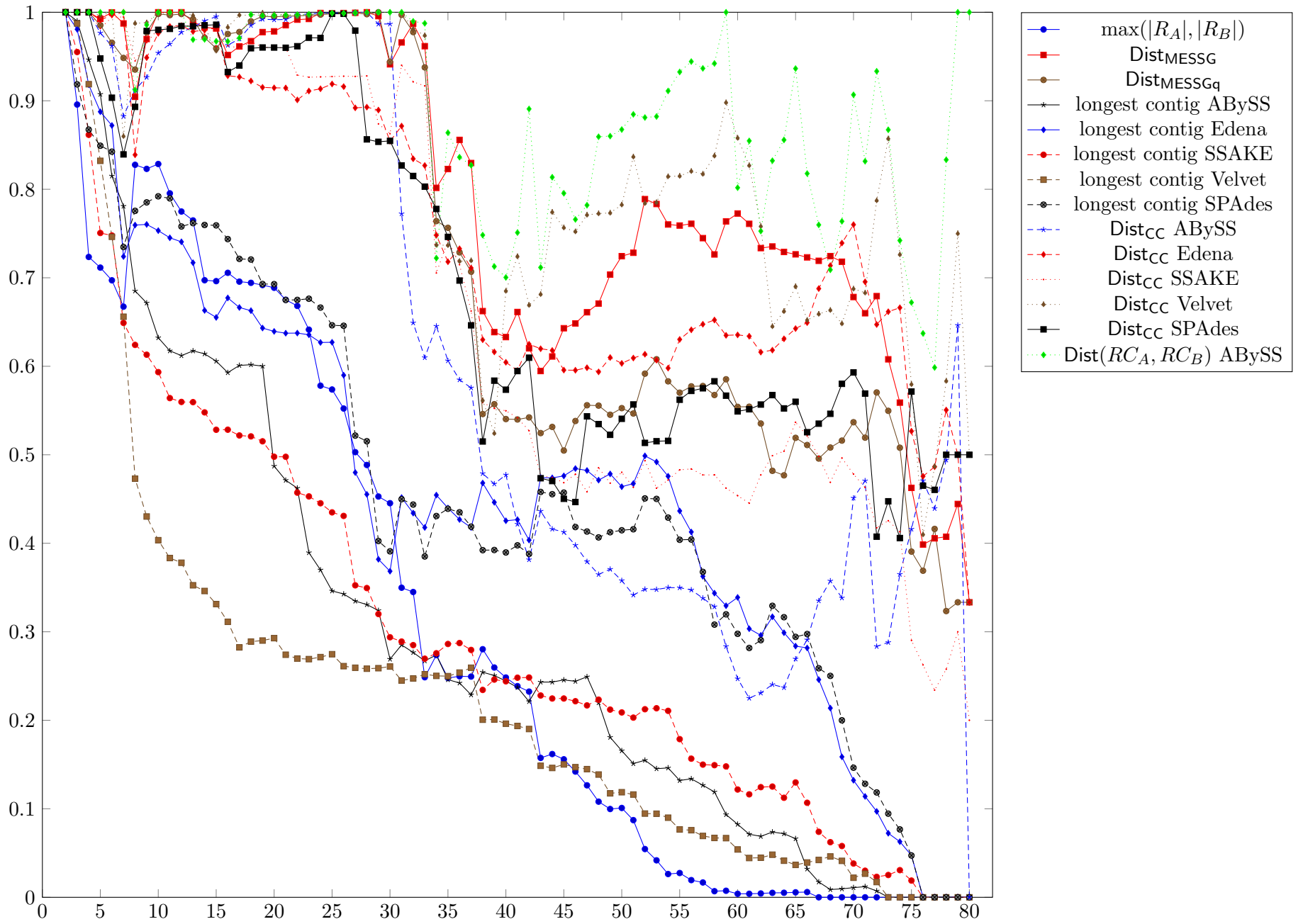


Figure 2: Plot of Fowlkes-Mallows index B_k versus k on hepatitis dataset. The index compares trees generated by the neighbor-joining algorithm.

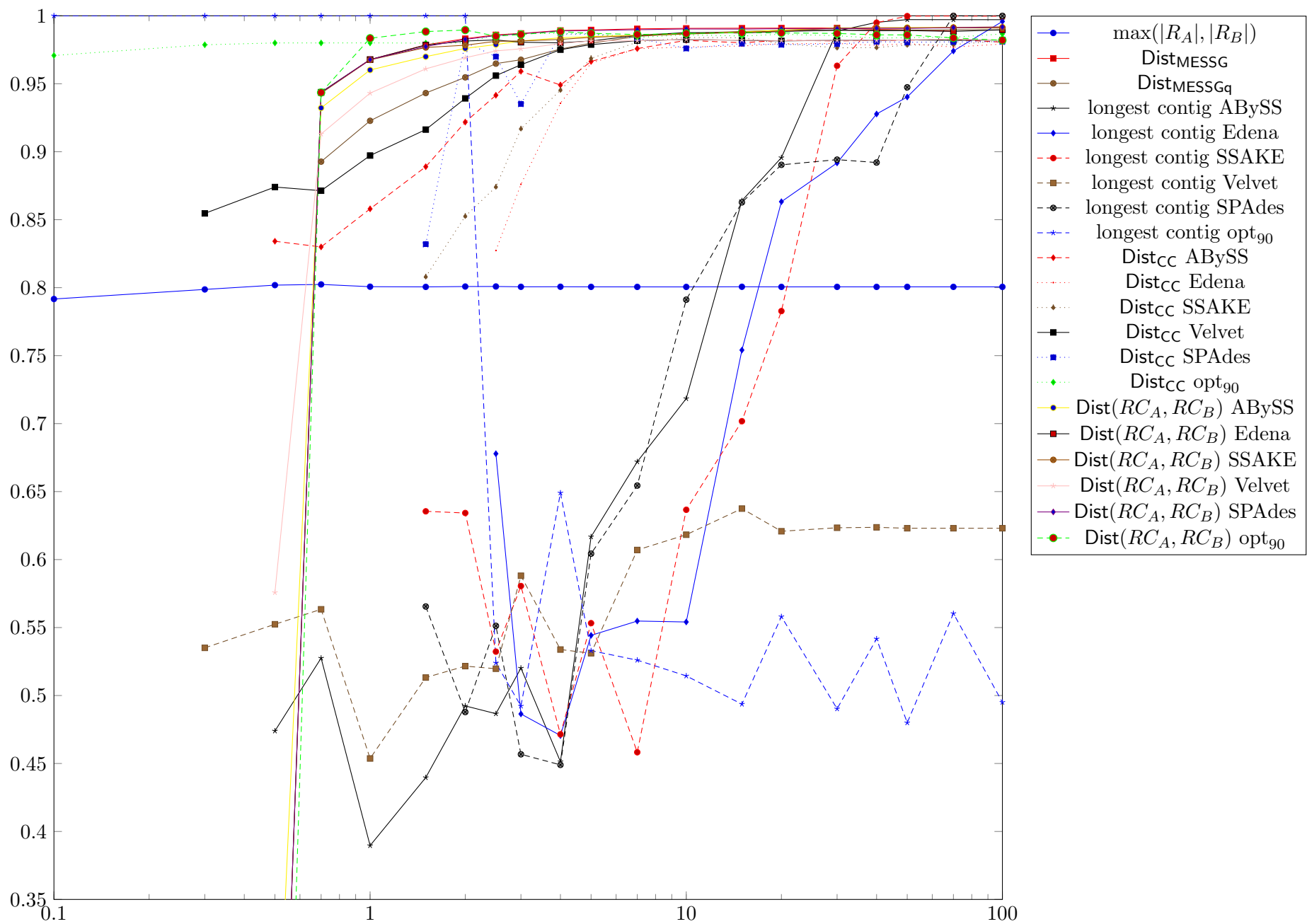


Figure 3: Plot of average Pearson's correlation coefficient for several choices of coverage values.

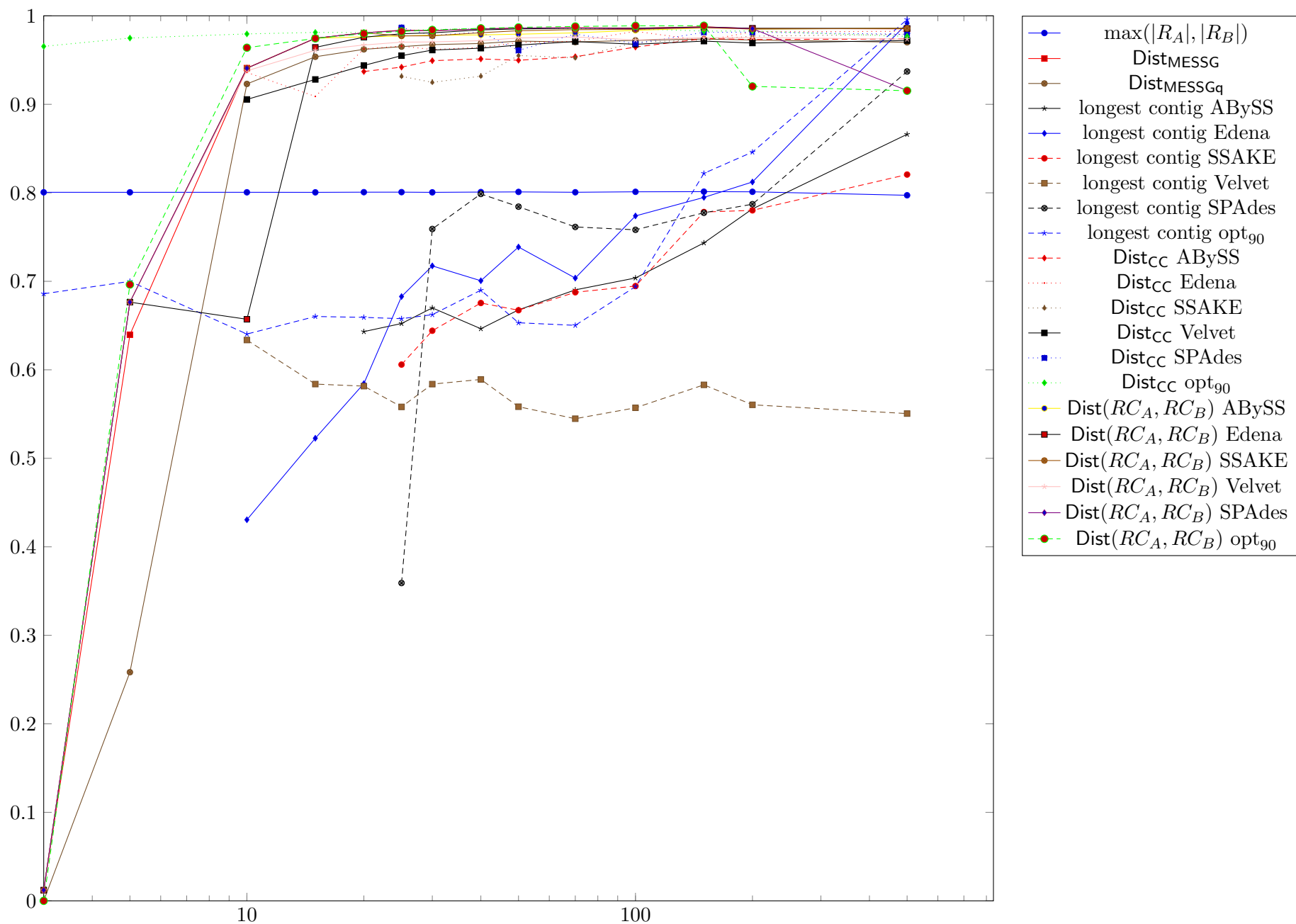


Figure 4: Plot of average Pearson's correlation coefficient for several choices of read length.

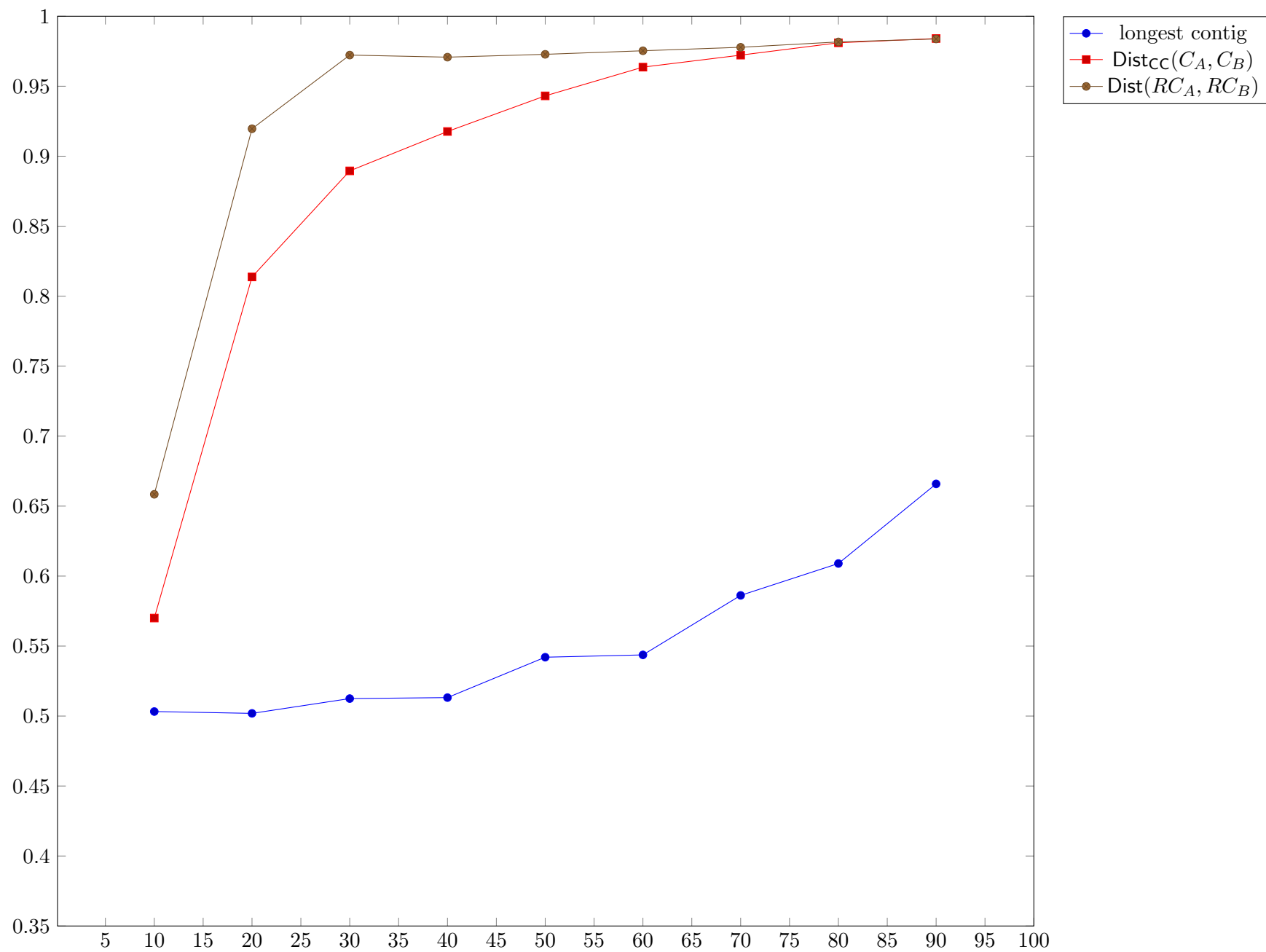


Figure 5: Plot of average Pearson's correlation coefficient on θ parameter for simulated assembly opt_θ on influenza dataset.