

	M'			M
			F	L
1	A C T A G A \$		\$ A C T A G A	
2	C T A G A \$ A		A A \$ A C T A G	G
3	T A G A \$ A C		A A C T A G A \$	A
4	A G A \$ A C T	→ sort →	A A G A \$ A C T	T
5	G A \$ A C T A		C C T A G A \$ A	A
6	A \$ A C T A G		G G A \$ A C R A	A
-	\$ A C T A G A		T T A G A \$ A C	C

Figure 1. BWT formed transformation. M' is the matrix of cyclic rotations before sorting. M is the matrix after sorting. We have included the extra symbol \$ for consistency in the suffixes.

Idx	T	L	F	V	Hr	Hrs	T	(Hr)	F	T	(Hrs)	F
1	A	G	A	5	2	6	A		A	A		A
2	C	A	A	1	4	1	C		A	C		A
3	T	T	A	6	6	4	T		A	T		A
4	A	A	C	2	3	2	A		C	A		C
5	G	A	G	3	5	3	G		G	G		G
6	A	C	T	4	1	5	A		T	A		T

Figure 2: Illustration of the auxiliary arrays for a sequence T=ACTAGA [Adjeroh2002zmpb]

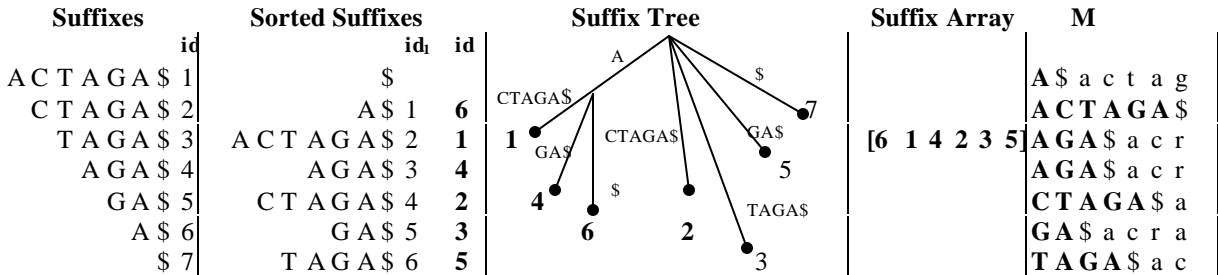


Figure 3. Suffix tree, suffix arrays, and matrix of sorted rotations [Taken from Adjeroh2002zmpb]. The numbers on the leaf nodes in the suffix tree correspond to those on the suffixes (id), which indicate the starting position of the suffix in the sequence. The numbers on the sorted suffixes (id_i) indicate the sorted index of the suffixes. The corresponding position in the sequence (id) is also shown. The labels on each link correspond to substrings in the sequence.

Suffixes		LCP		Sorted Suffixes		SCP
ACTAGACTAS\$ 1	id	1 2 3 4 5 6 7 8 9		\$	id	1 2 3 4 5 6 7 8 9
CTAGACTAS\$ 2		1 - 0 0 1 0 4 0 0 1		A\$ 1		1 - 1 1 1 0
TAGACTAS\$ 3		2 - 0 0 0 0 3 0 0		A C T A \$ 2		2 - 4 1 0
AGACTAS\$ 4		3 - 0 0 0 0 2 0		A C T A G A C T A \$ 3		3 - 1 0
GACTAS\$ 5		4 - 0 1 0 0 1		A G A C T A \$ 4		4 - 0
ACTAS\$ 6		5 - 0 0 0 0		C T A \$ 5		5 - 3 0
CTAS\$ 7		6 - 0 0 1		C T A G A C T A \$ 6		6 - 0
TAS\$ 8		7 - 0 0		G A C T A \$ 7		7 - 0
A\$ 9		8 - 0		T A \$ 8		8 - 2
\$		9 -		T A G A C T A \$ 9		9 -

Figure 4. Nature of the SCP (Original sequence T = ACTAGACTAS\$). The numbers on the top row, leftmost column of the LCP are the starting positions of the corresponding suffixes in S. For the SCP, this is the position of the suffix in the lexically sorted order of all the suffixes.

T₁= ACCGTT; T₂=ACGTC; T₃=CACGTCT

Three Alignments

-ACCGTT--	-ACCGT-T	-ACCGTT--
-A-CGT-C-	-AC-GTC-	-A-CG-TC-
CA-CGT-CT	CAC-GTCT	CA-CG-TCT

	#	&	#	&	\$	\$	#	&	\$	&	#	&	\$	\$	&	#	&	\$
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
\$	1	-	2	2	0													
#	2		-	2	0	C1												
&	3			-	0													
#	4				-	1	1	1	1	1	1	0						
&	5					-	1	1	1	1	1	0						
\$	6								-	1	1	1	0	C2				
#	7									-	4	3	1	0				
&	8										-	3	1	0				
\$	9											-	1	0				
&	10												-	0				
#	11													-	3	2	0	
&	12														-	2	0	C3
\$	13															-	0	
\$	14																-	1
&	15																	-
#	16																	
&	17																	
\$	18																	

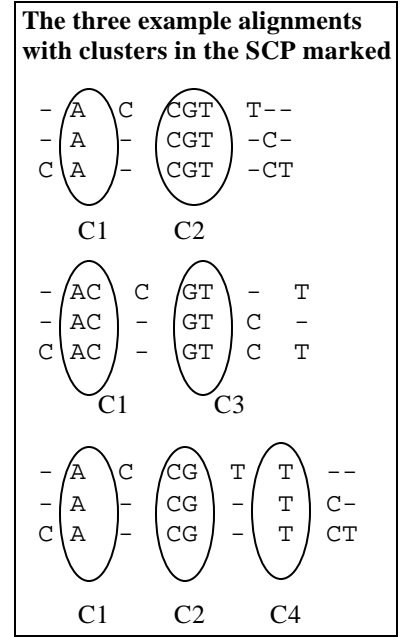
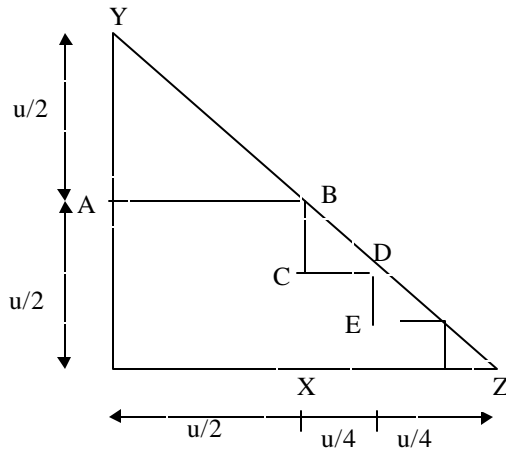


Figure 6. The SCP for combined multiple sequences. The \$, #, & are extra characters needed for to enforce the non-prefix condition in the suffixes. Spaces have been inserted between symbols to show the relationship between the alignment and the different clusters in the SCP. C1, C2, C3, C4 represent the 4 clusters in the SCP. SCP(11,12) is within the sphere of influence of SCP(7,8). Hence, they are likely to have resulted from some common substring.

Figure 5. Binary partitions of the SCP Table and p-paths. The path: A → B → C → D → E → ... → Z defines an p-path.



Suffixes	Sorted Suffixes	Sort id
	\$	
aca aca aca aca aca aca\$	a\$	18
ca aca aca aca aca aca\$	a aca aca aca aca aca\$	2
a aca aca aca aca aca\$	a aca aca aca aca aca\$	3
aca aca aca aca aca\$	a aca aca aca\$	4
ca aca aca aca\$	a aca aca\$	5
a aca aca aca\$	a aca\$	6
aca aca aca\$	aca\$	7
ca aca aca\$	aca aca\$	8
a aca aca\$	aca aca\$	9
aca aca aca\$	aca aca aca\$	10
ca aca aca\$	aca aca aca aca\$	11
a aca aca\$	aca aca aca aca aca\$	12
aca aca\$	ca\$	13
ca aca\$	ca aca\$	14
a aca\$	ca aca aca\$	15
aca\$	ca aca aca aca\$	16
ca\$	ca aca aca aca aca\$	17
a\$	ca aca aca aca aca aca\$	18
\$		

Figure 7a Suffixes list and sorted suffixes for T_1 .

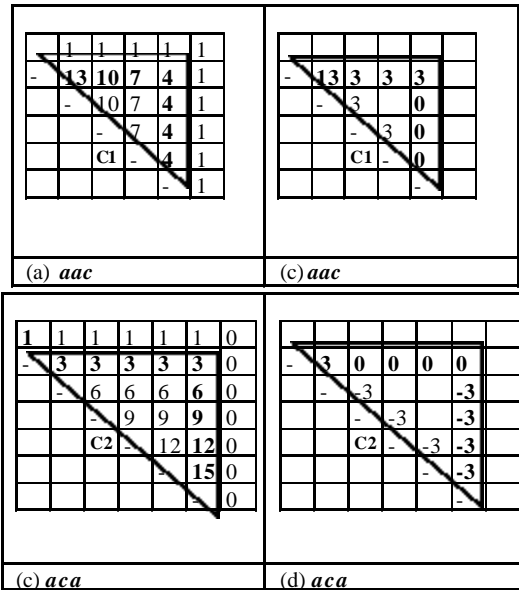


Figure 7c. Left Column: p -periodic neighborhoods for *aac* (a) and *aca* (c) Right Column: corresponding difference triangles for *aac* (b) and *aca* (d):

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	-	0	1	15	0	1	12	0	1	9	0	1	6	0	1	3	0	1
2		-	0	0	14	0	0	11	0	0	8	0	0	5	0	0	2	0
3			-	1	0	13	1	0	10	1	0	7	1	0	4	1	0	1
4				-	0	1	12	0	1	9	0	1	6	0	1	3	0	1
5					-	0	0	11	0	0	8	0	0	5	0	0	2	0
6						-	1	0	10	1	0	7	1	0	4	1	0	1
7							-	0	1	9	0	1	6	0	1	3	0	1
8								-	0	0	8	0	0	5	0	0	2	0
9									-	1	0	7	1	0	4	1	0	1
10										-	0	1	6	0	1	3	0	1
11											-	0	0	5	0	0	2	0
12												-	1	0	4	1	0	1
13													-	0	1	3	0	1
14														-	0	0	2	0
15															-	1	0	1
16																-	0	1
17																	-	0
18																		-

LCP

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	-	1	1	1	1	1	1	1	1	1	1	1	1	1	0			
2		-	13	10	7	4	1	1	1	1	1	1	1	0				
3			-	10	7	4	1	1	1	1	1	1	1	0				
4				-	7	4	1	1	1	1	1	1	1	0				
5					-	4	1	1	1	1	1	1	1	0				
6						-	1	1	1	1	1	1	1	0				
7							-	3	3	3	3	3	0					
8								-	6	6	6	6	0					
9									-	9	9	9	0					
10										-	12	12	0					
11											-	15	0					
12												-	0					
13													-	2	2	2	2	
14														-	5	5	5	5
15															-	8	8	8
16																-	11	11
17																	-	14
18																		-

SCP

Figure 7b Geometry of the SCP: The LCP, SCP and p -periodic neighborhoods for a sequence with tandem arrays The LCP (left) and SCP (right) for string $T_1=abaabaabaabaabaaba$. C1, C2 and C3 represent regions with tandem repeats, corresponding primitive strings *aab*, *aba*, *baa*, respectively.

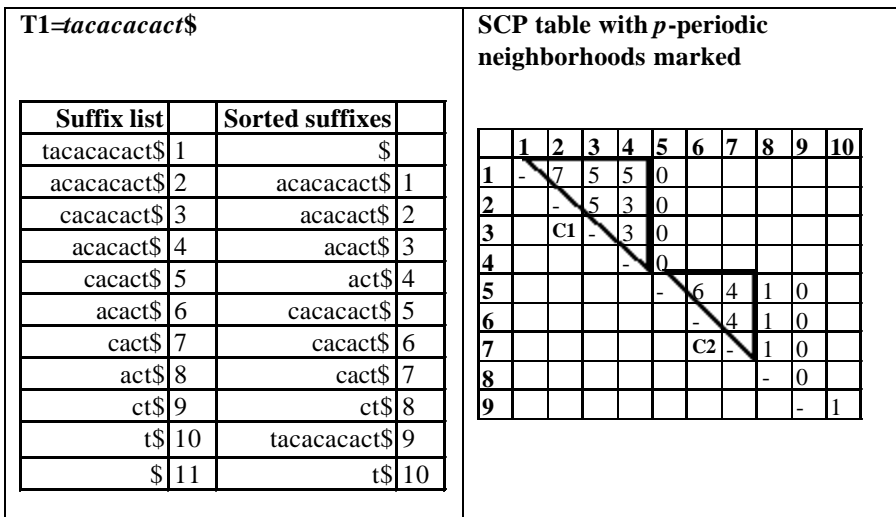


Figure 8a. Suffixes and SCP table for tandem arrays with single occurrence of the basic primitive strings (*ac* and *ca*).

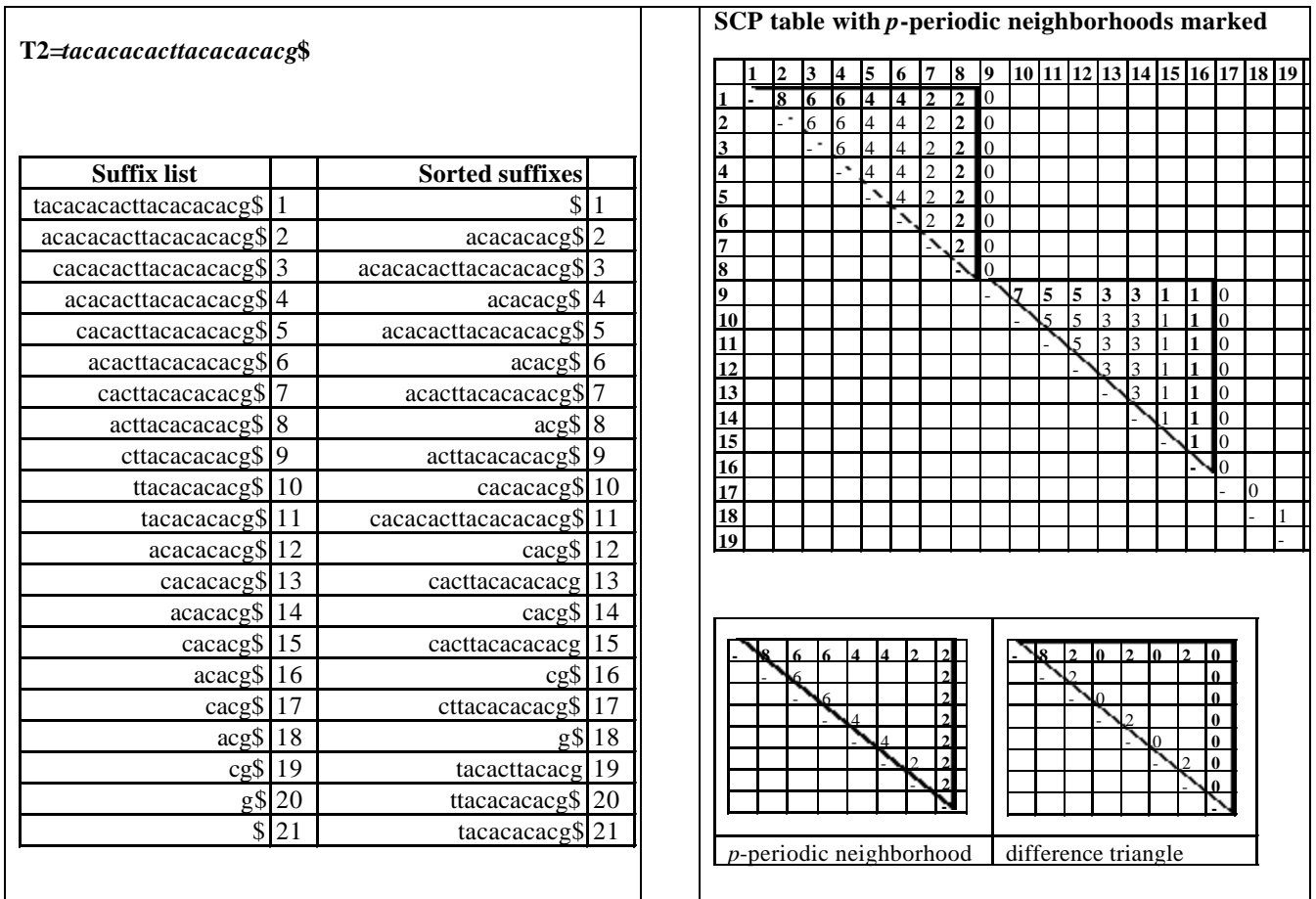


Figure 8: The SCP table and p -periodic neighborhoods for multiple distinct supermaximal tandem arrays with the same basic primitives as in Figure 8b.

Table 1: SCP statistics and computation time for some DNA and Protein files.

File name	size, u (KB)	maxSCP, K_{max}	K_{max}/u	ave nSCP	self- similarity	Time (seconds)		
						AlgorithmC	AlgorithmB	Algorithm A
humEpsBarr	172.28	32442	1.88E-01	3.07E+03	8.71E-06	3.54E+02	2.48E+02	8.26E+02
HUMGHCSA	66.495	408	6.14E-03	2.47E+01	1.55E-05	3.70E+01	3.10E+01	1.00E+02
HUMHDABCD	58.864	107	1.82E-03	2.90E+01	1.75E-05	2.90E+01	2.40E+01	7.70E+01
mitoBP	16.398	15	9.15E-04	6.55E+00	6.81E-05	2.00E+00	2.00E+00	7.00E+00
mitoGallus	16.775	29	1.73E-03	6.64E+00	6.73E-05	3.00E+00	2.00E+00	6.00E+00
mitoLB	17.211	198	1.15E-02	1.07E+01	7.66E-05	3.00E+00	2.00E+00	6.00E+00
mitoMIPACGA	94.192	72	7.64E-04	8.35E+00	1.32E-05	1.02E+02	7.20E+01	2.04E+02
mitoMPOMTCG	186.60	191	1.02E-03	8.91E+00	5.59E-06	2.79E+02	2.79E+02	9.16E+02
VACCG	191.73	1313	6.85E-03	1.97E+01	6.04E-06	4.69E+02	3.28E+02	9.70E+02
YSCCHRIII	316.61	1682	5.31E-03	1.47E+01	3.40E-06	1.57E+03	1.14E+03	2.81E+03
hi(protein)	509.59	343105	6.73E-01	1.16E+05	1.04E-05	2.42E+03	-	-
mj(protein)	448.77	220685	4.92E-01	5.43E+04	3.46E-06	1.30E+03	-	-