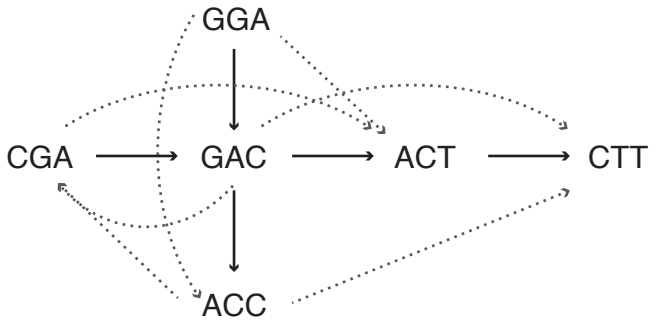
a

K = {CGA, GAC, ACT, CTT, GGA, ACC}

b



k-1 long overlaps

k-2 long overlaps

de Bruijn graph:
(vertex-centric, uni-directional)

overlap graph:

...

c

Name	Strings	Masked superstring	
1. isolated *k*-mers	CGA, GAC, ACT, CTT, GGA, ACC	CGAGACACTCTTGGAACC 10010010010010010 0	ℓ=18 r=6, o=6
2. unitigs	CGA, GAC, GGA, ACC, ACTT	CGAGACGGAACCACTT 1001001001001100	ℓ=16 r=5, o=6
3. simplitigs/SPSS	CGACTT, GGA, ACC	CGACTTGGAACC 111100100100	ℓ=12 r=3, o=6
4. matchtigs	CGACTT, GGACC	CGACTTGGACC 11110011100	ℓ=11 r=2, o=7
5. shortest superstring	GGACCGACTT	GGACCGACTT 1110111100 ① or 1110101100 or 1010111100 ②	ℓ=10 r=2, o=7 r=3, o=6 r=3, o=6

d

enc0:	*"CGACTT\nGG A\nACC"*	*simplitigs/matchtigs delimited by EOL (state-of-the-art)*
enc1:	"GGACCGACTT" "1110111100"	superstring + mask strings
enc2:	"GGAcCGACtt"	mask-cased superstring
enc3:	"GGACCGACTT" [3,1,4]	superstring + mask RLE
enc4:	"GGACCGACTT" [3]	superstring + mask sparse enc. (positions of non-final zeros)

e

① $\mathbb{K}$ = {CGA, 2×GAC, ACT, CTT, GGA, ACC}
$\mathbb{X}$ = {CCG}

② $\mathbb{K}$ = {CGA, 1×GAC, ACT, CTT, GGA, ACC}
$\mathbb{X}$ = {CCG, 1×GAC}