Velvet: Algorithms for de novo short read assembly using de Bruijn graphs

Daniel R. Zerbino and Ewan Birney

2008

Problem

- de novo assembly is the problem of figuring out the genome sequence without no prior information
- for example, sequencing a species for the first time
- inputs reads short, can be modelled as randomly broken pieces of the genome
- plus possibly paired end data

Illustration

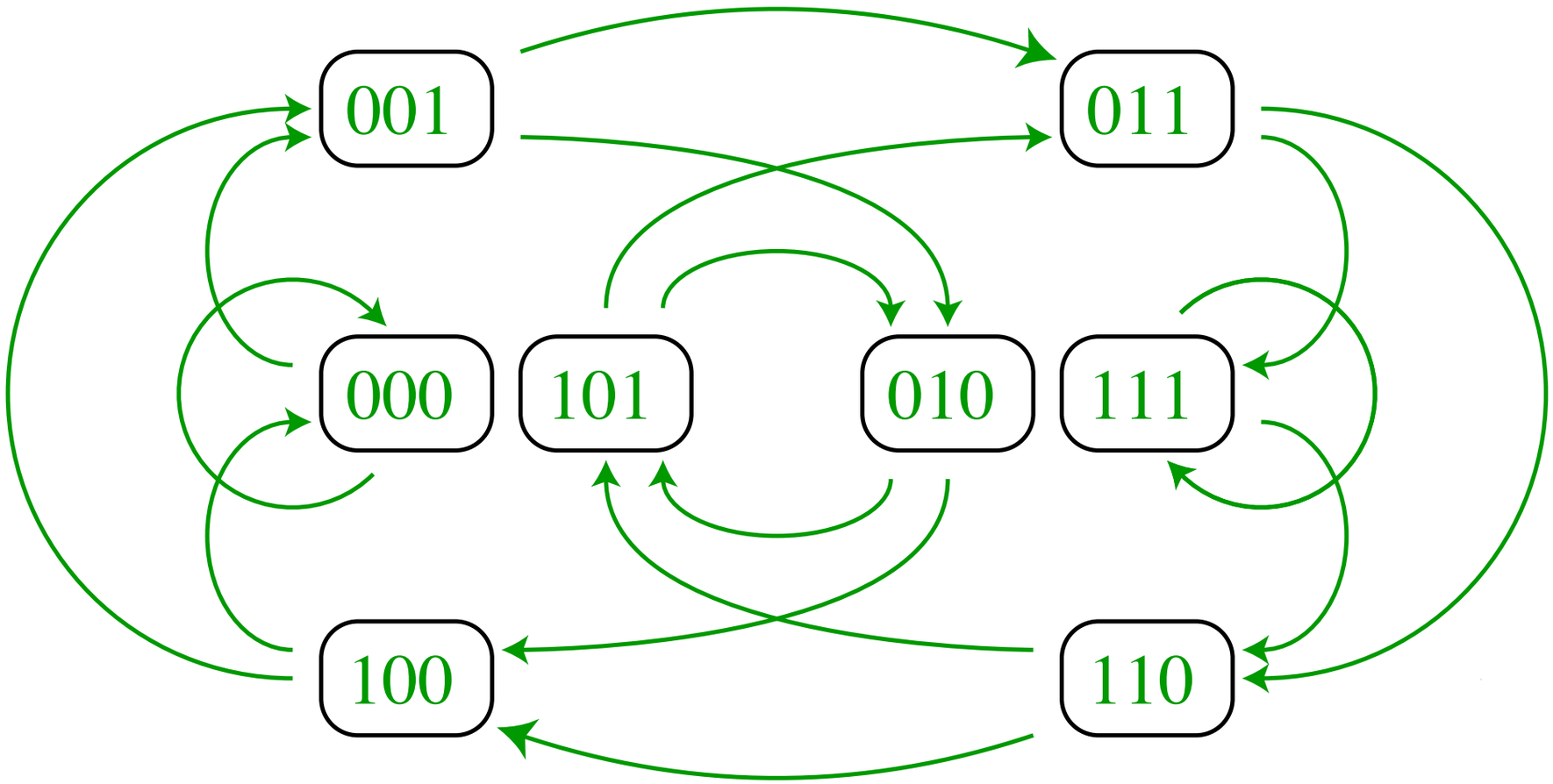
TAGTCGAGGCTTTAGATCCGATGAGGCTTTAGAGACAG

AGTCGAG CTTTAGA CGATGAG CTTTAGA
GTCGAGG TTAGATC ATGAGGC GAGACAG
GAGGCTC ATCCGAT AGGCTTT GAGACAG
AGTCGAG TAGATCC ATGAGGC TAGAGAA
TAGTCGA CTTTAGA CCGATGA TTAGAGA
CGAGGCT AGATCCG TGAGGCT AGAGACA
TAGTCGA GCTTTAG TCCGATG GCTCTAG
TCGACGC GATCCGA GAGGCTT AGAGACA
TAGTCGA TTAGATC GATGAGG TTTAGAG
GTCGAGG TCTAGAT ATGAGGC TAGAGAC
AGGCTTT ATCCGAT AGGCTTT GAGACAG
AGTCGAG TTAGATT ATGAGGC AGAGACA
GGCTTTA TCCGATG TTTAGAG
CGAGGCT TAGATCC TGAGGCT GAGACAG
AGTCGAG TTTAGATC ATGAGGC TTAGAGA
GAGGCTT GATCCGA GAGGCTT GAGACAG

Challenges

- Is this possible with short short reads?
  $4^{25}=10^{15}$ 10^9 bp in the human genome
- repeats in the genome
- mistakes in the sequencing reads
- mistakes in the sequencing biology/ chemistry

de Bruijn graphs

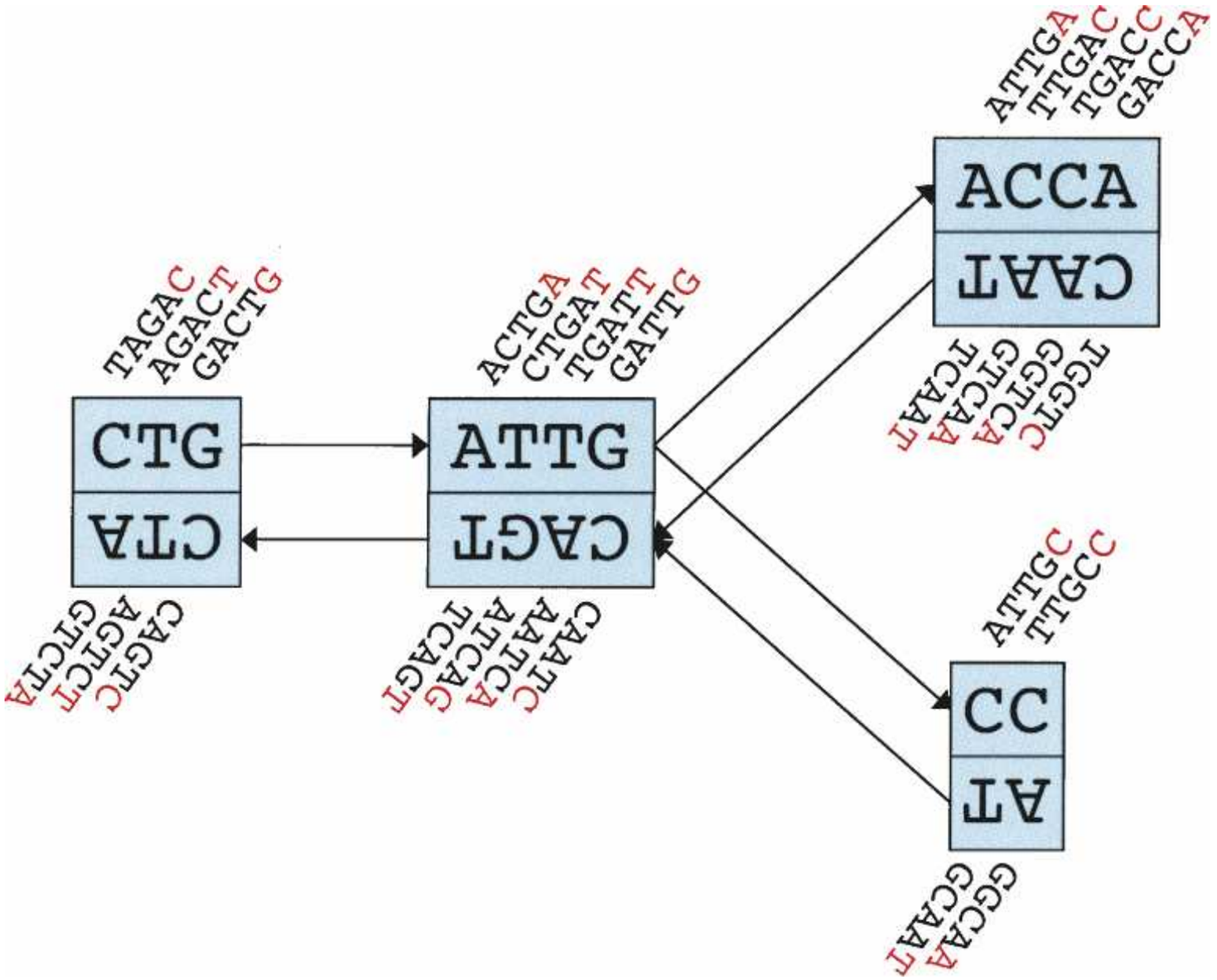


- each sequence is a path through the graph
- the outer path: 1001100

construction

- sequence of each read is parsed into k-mers
- typical k=21 for read length of 25
- series of matches(k-1 long) are aligned together called a block
- the information of each block is the last bp of each k-mer in of the block

alignment


ATTGA
TTGAC
TGACC
GACCA
ACCA
CAAT
TAGAC
AGACT
GACTG
CTG
CTA
CAGTC
AGTCT
GTCTA
ACTGA
CTGAT
TGATT
GATTG
ATTG
CAGT
CAATC
AATCA
TCAGT
TGGTC
GGTCA
GTCAA
TCAAT
ATTGC
TTGCC
CC
AT
GGCAA
GCAAT

links

- a directed link is drawn if there exists a (k-1)long match between two blocks
- if everything is perfect, an underlying sequence follows all links in the de Bruijn graph while tracing through every block
- however, due to the noisy measurement and sequence repeats, many more steps are required

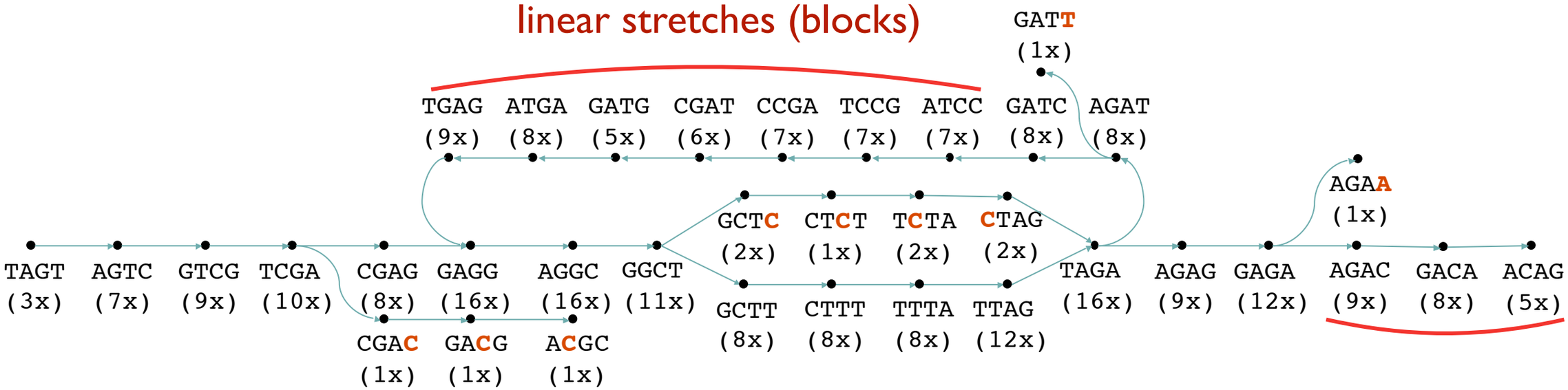
Example

TAGTCGAGGCTTTAGATCCGATGAGGCTTTAGAGACAG

AGTCGAG CTTTAGA CGATGAG CTTTAGA
GTCGAGG TTAGATC ATGAGGC GAGACAG
GAGGCTC ATCCGAT AGGCTTT GAGACAG
AGTCGAG TAGATCC ATGAGGC TAGAGAA
TAGTCGA CTTTAGA CCGATGA TTAGAGA
CGAGGCT AGATCCG TGAGGCT AGAGACA
TAGTCGA GCTTTAG TCCGATG GCTCTAG
TCGACGC GATCCGA GAGGCTT AGAGACA
TAGTCGA TTAGATC GATGAGG TTTAGAG
GTCGAGG TCTAGAT ATGAGGC TAGAGAC
AGGCTTT ATCCGAT AGGCTTT GAGACAG
AGTCGAG TTAGATT ATGAGGC AGAGACA
GGCTTTA TCCGATG TTTAGAG
CGAGGCT TAGATCC TGAGGCT GAGACAG
AGTCGAG TTTAGATC ATGAGGC TTAGAGA
GAGGCTT GATCCGA GAGGCTT GAGACAG

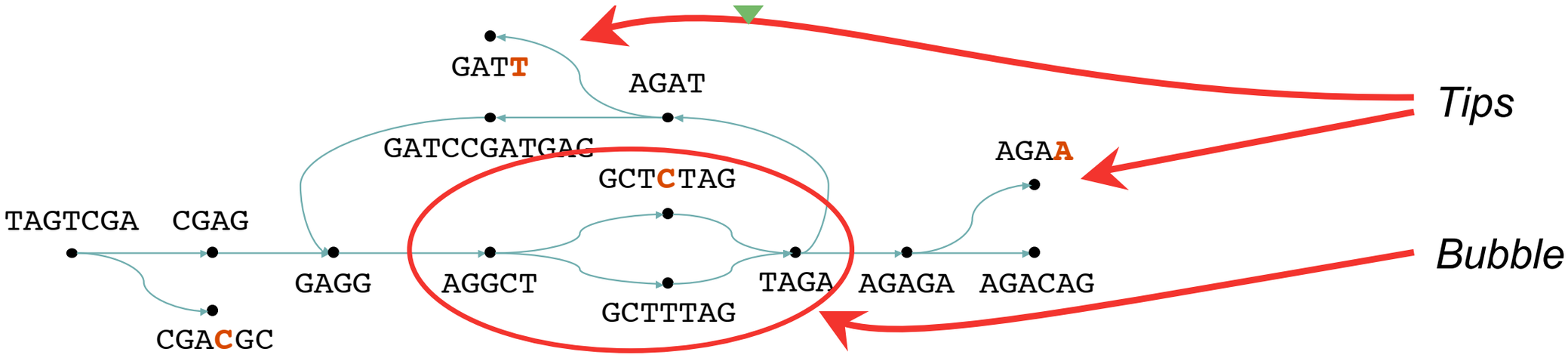
red = sequencing mistakes

Example



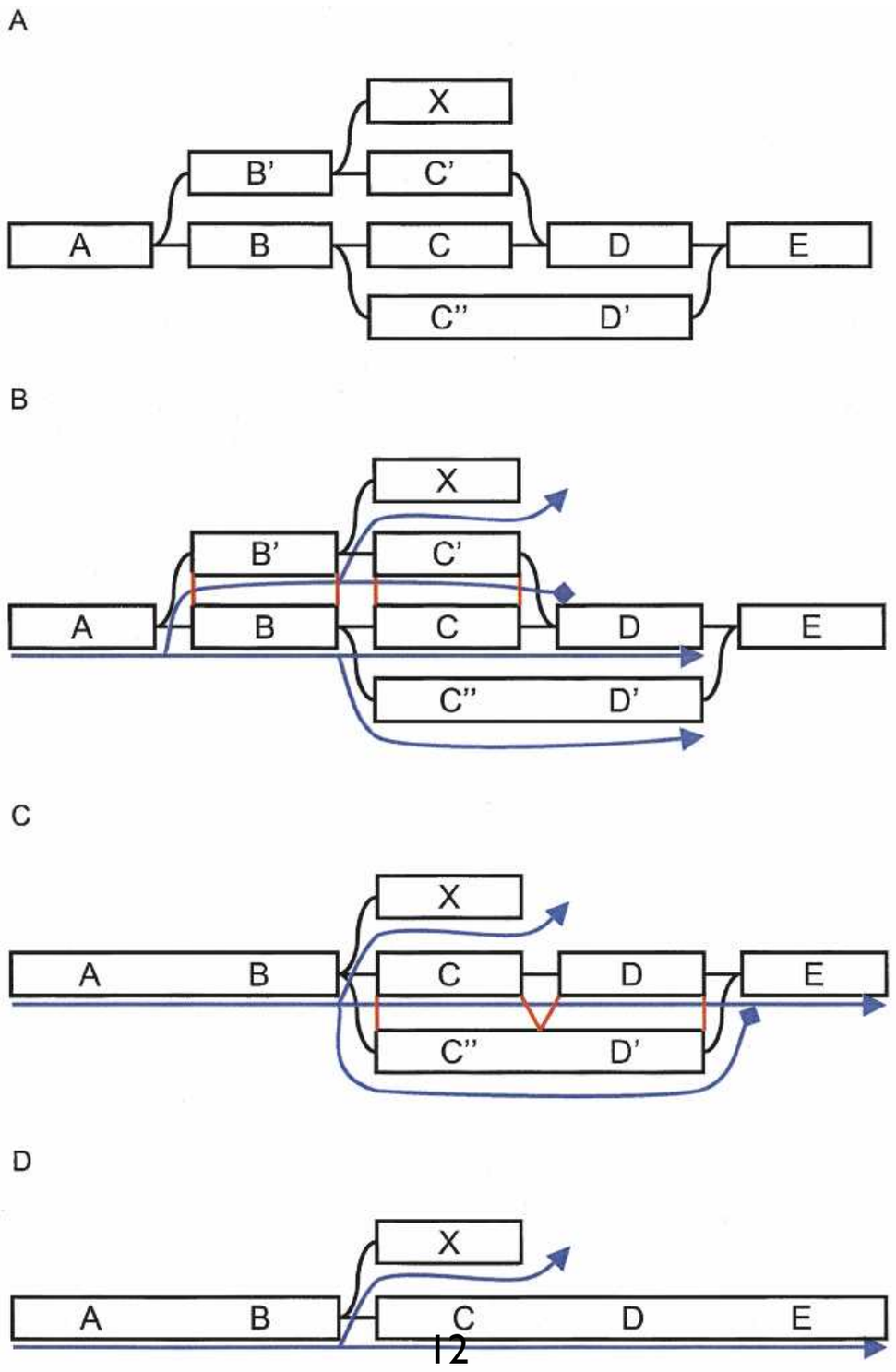
- read data form example is parsed into 4-mers
- matches are found and the de Bruijn graph is constructed
- problems ?

merging blocks

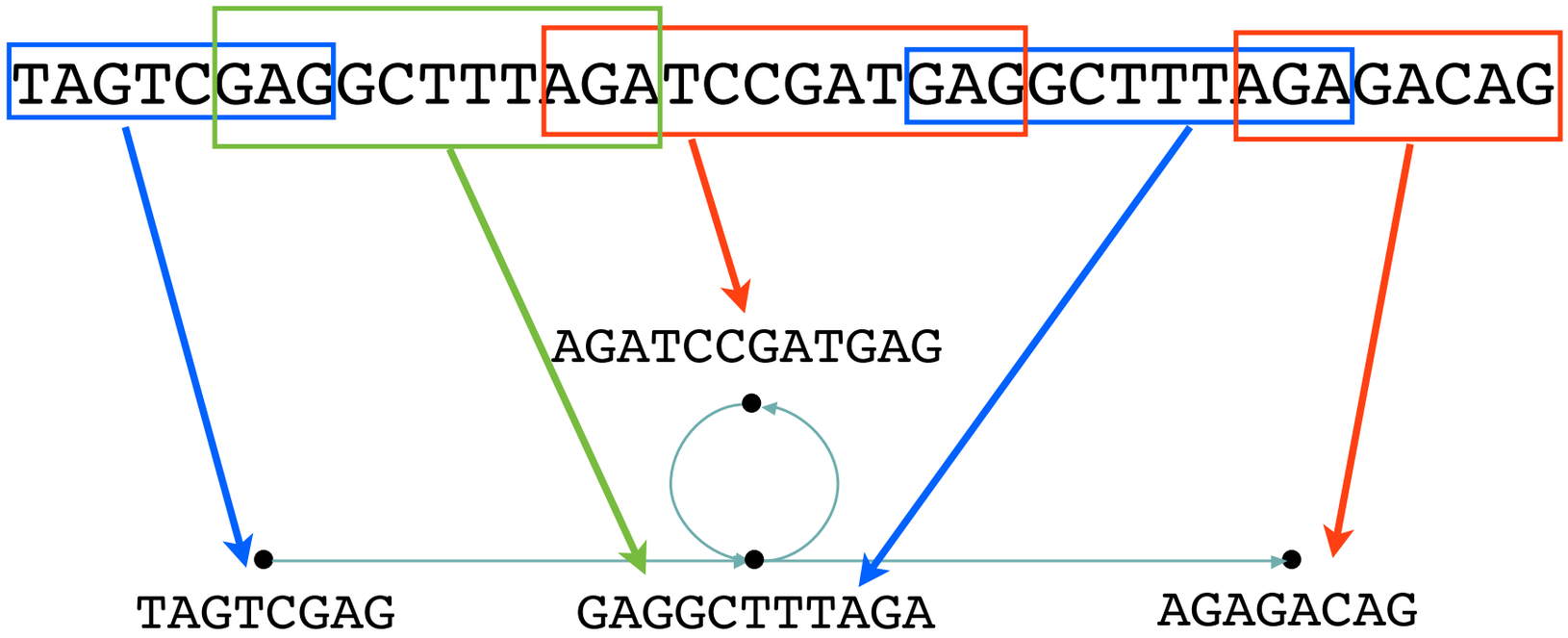


- hanging tips(blocks that do not connect to anything) are likely due to mistakes, especially low-coverage ones
- bubbles(cycles of the in the graph) is likely due to errors

Tour Bus algorithm for bubble removal



result of example

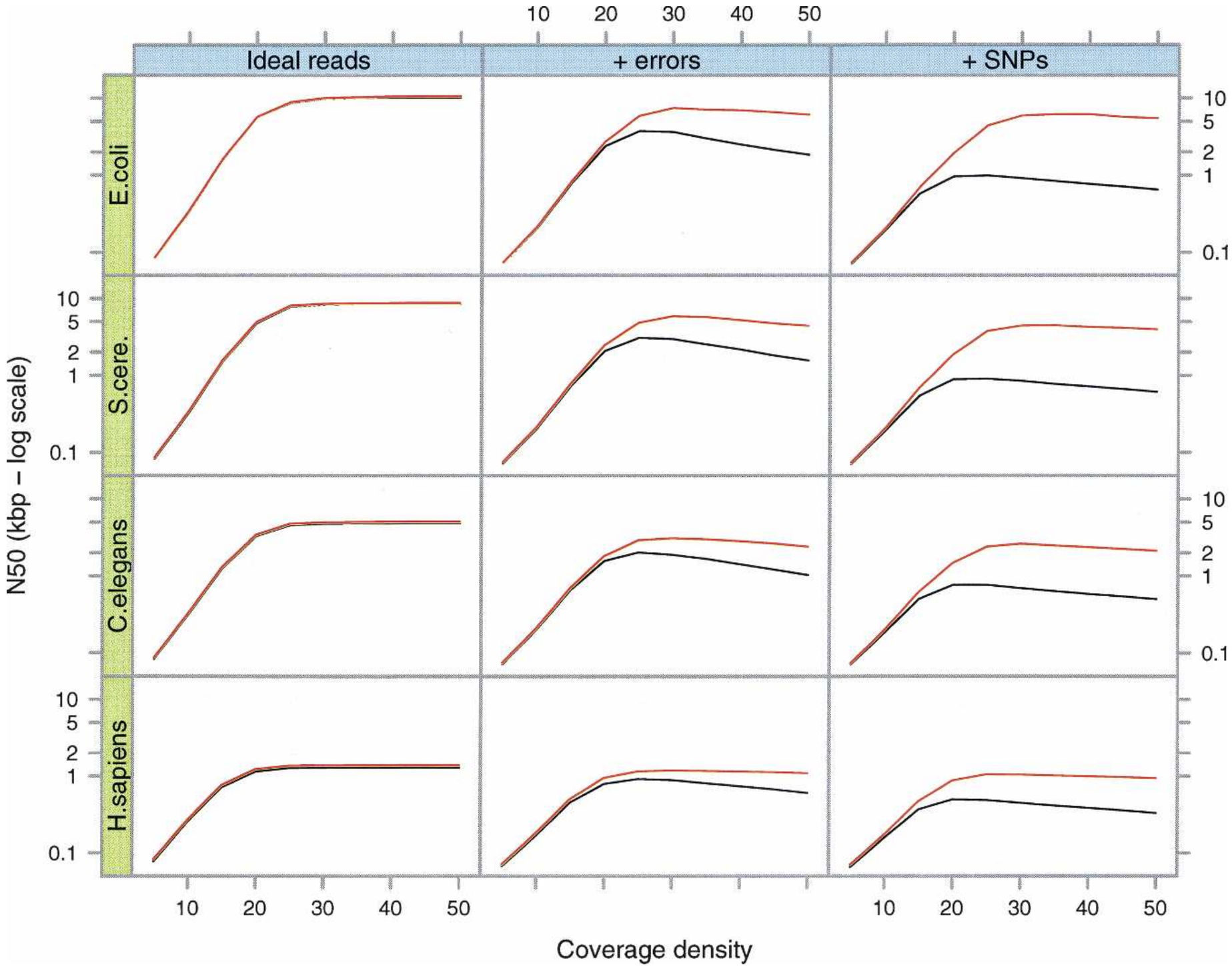


- in this example, sequence length=38 bp, read length=7bp , coverage~10X, error rate~ 3%, with one major repeat = 11bp
- k is chosen to be 5 bp
- Velvet is able to resolve this toy example!

N50

- in problems of practical size, it is unlikely that any large genome can possibly be assembled through read data alone (more experiment needed)
- it is harder for to measure performance
- one measure of how well an assembler performs is the N50 (median length weighted contig length)

simulated results



real reads

- a 173 kbp human BAC was sequenced by Solexa with a coverage of 970X
- read length are 35 bp
- k set to 31
- an virtual ideal sequencer(error free, gap free) that looks at the reference sequence is compared with Velvet

experimental reads

Table 1. Efficiency of the Velvet error-correction pipeline on the BAC data set

Step	No. of nodes	N50 (bp)	Maximum length (bp)	Coverage (percent >50 bp)	Coverage (percent >100 bp)
Initial	1,353,791	5	7	0	0
Simplified	945,377	5	80	4.3	0.2
Tips clipped	4898	714	5037	93.5	78.7
Tour Bus	1147	1784	7038	93.4	90.1
Coverage cutoff	685	1958	7038	92.0	90.0
Ideal	620	2130	9045	93.7	91.9

conclusion

- Velvet is able to a reasonably well job of error removal efficiently with short reads
- complex genome assembly is difficult due to repeats
- de novo genome assembly is not a solved problem

pair end results