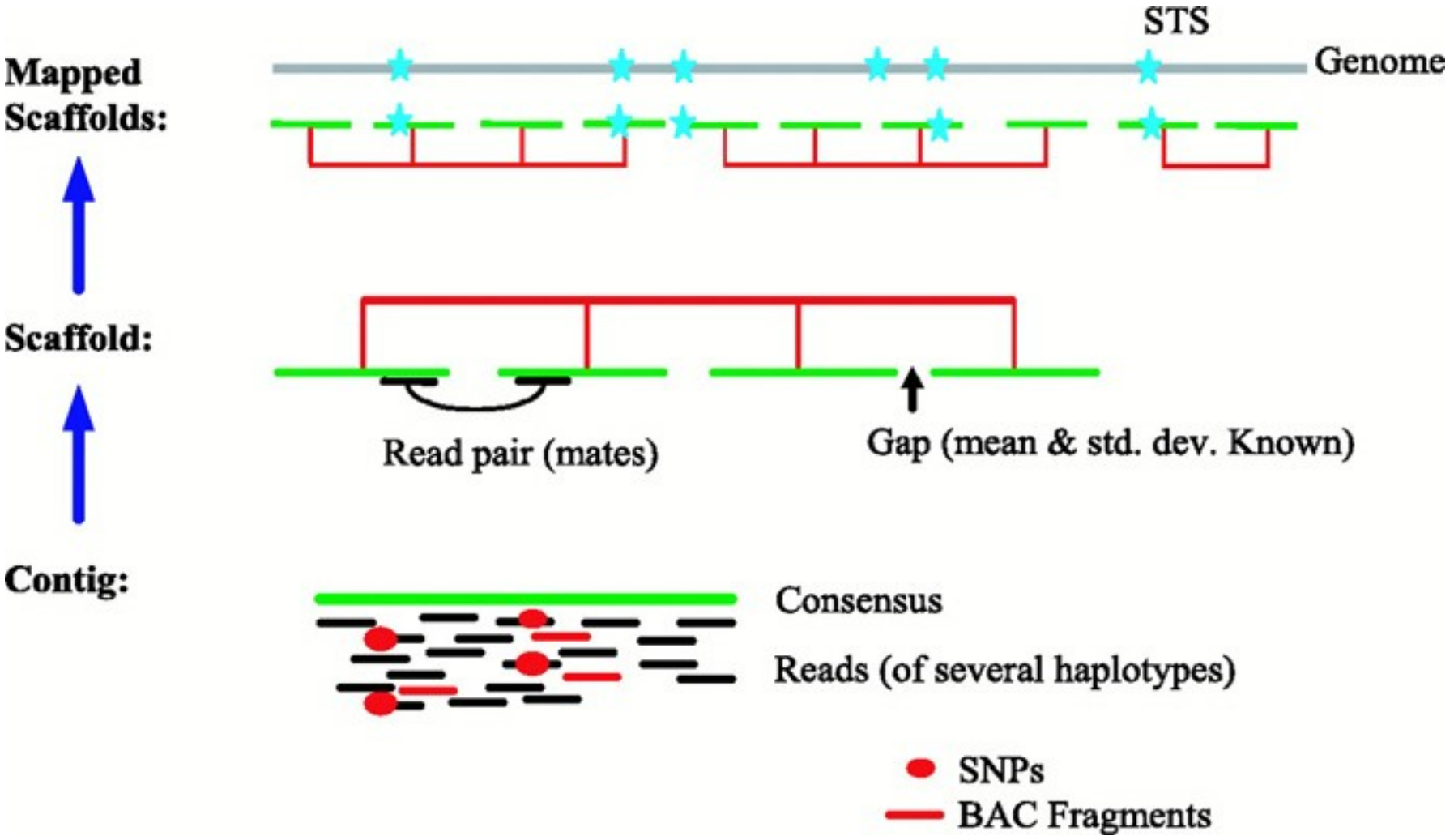
Announcement for Homework

GigAssembler

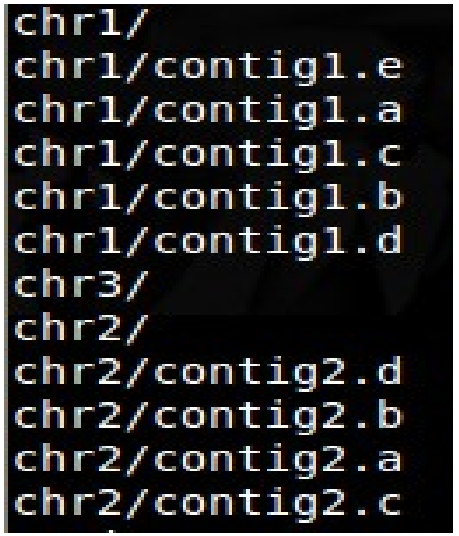
Genome Assembly: A big picture



http://www.nature.com/scitable/content/anatomy-of-whole-genome-assembly-20429

GigAssembler – Preprocessing

1. Decontaminating & Repeat Masking.
2. Aligning of mRNAs, ESTs, BAC ends & paired reads against initial sequence contigs.
   - psLayout → BLAT
3. Creating an input directory (folder) structure.



http://www.triazzle.com; The image from http://www.dangilbert.com/port_fun.html

Reference: Jones NC, Pevzner PA, Introduction to Bioinformatics Algorithms, MIT press

RepBase + RepeatMasker

```
taejoon@fourierseq:~/RepBase/RepBase15.05.fasta$ ls -a
.           dcotrep.ref  mamsub.ref   rodsub.ref
..          diarep.ref   mcotrep.ref  simple.ref
angrep.ref  drorep.ref   mousub.ref   spurep.ref
appendix    fngrep.ref   nemrep.ref   synrep.ref
athrep.ref  fugrep.ref   oryrep.ref   tmplanrep.ref
bctrep.ref  grasrep.ref  plnrep.ref   tmpnemrep.ref
cbrrep.ref  humrep.ref   prirep.ref   tmpxenrep.ref
celrep.ref  humsub.ref   prisub.ref   version
chlrep.ref  invrep.ref   pseudo.ref   vrtrep.ref
cinrep.ref  invsub.ref   ratsub.ref   zebrep.ref
cinunc.ref  mamrep.ref   rodrep.ref
```

```
>MER51D ERV1  Homo sapiens
tgaggcaggagaaaatagcagagggaattggaagttggataaagggagaatgagtaaaagcangagagca
gaagcaaggtaaagaggcgggtgagcaagaagcaagataagaagcagaagttgagcagccaaaacaaaag
taagatnanaaagaagtgagtaaggagcccacatggctggctagatccagaccaaaccagtaaggggcag
ctcctcagagatgggcatgtacattagagagaaaaagtatccttaaaatgaccccgtatgataatcagct
cattaaagctcatgcatatggactgcatatcatgcatgtacttaaaattatgggatggaggtgacgcgca
agawgtcacaagcacacaggggccatagkattaagtaactaagcaacccacctatcaatcaaaaggcaga
tgctggctagagattaggcagccttgggaagagaagaaaaaaaaaacacataaaaagacccaaagtacac
caaactgacgctgatctcatttcgcagaggtcagcccactctcccctctctgagagtgtaatactgtgct
taataaacttttgctgctttgctatctgtgtgtgtcttgtccaattctttgtttgggacaccaagagcct
ggaactgcacrgcaccakctggtaaca
>MIRb SINE2/tRNA  Mammalia
cagaggggcagcgtggtgcagtggaaagagcacgggctttggagtcaggcagacctgggttcgaatcctg
gctctgccacttactagctgtgtgaccttgggcaagtcacttaacctctctgagcctcagtttcctcatc
tgtaaaatggggataataatacctacctcgcagggttgttgtgaggattaaatgagataatgcatgtaaa
gcgcttagcacagtgcctggcacacagtaagcgctcaataaatggtagctctattatt
>LTR45  ERV1  Homo sapiens
tgtaaccgcgggaccagcccaaactgggcctactctgttgataacaaaatgtcaagttaccttgtaggta
taacagagcccaaaactgcaagtcatgtagcccgggcatgtgcaatagaaaaagctttgacctctaacaa
cacccagaaccaatgattcctcccctcggaaccaagaagaccgggacatgaccggaacctgaatgccgga
actctttcagaagcaaaggggtccgttggcccggaagatctggggctaaaatctgcctcaacatacctta
ccgtaaatggtcaaatttgaagccctccaatcagaccctgccaagccaacattcctaaatccttttccctt
gccctctgatcccttaaaacttgccccagaccccaaatcggggagacagatttgagcccacctcctgtct
ccttgctggccggttttgcaataaagcctttcttttctcaaaagctggtgccatagttattggcttctgt
gtgcatcaggcagcaagcccatttgctcgataaca
>MER80B hAT Homo sapiens
cagggcttcttaaccagaggtccatggatgggcttcaggaggtctgtgaaccctctgaaattatatacaa
aaatgttgtgtatatgtgcatatatgtatttttctggggagagggttcatagctttcatcagattctcaa
aggggtctatgatctmaaaaaggttaagaagccctg
```

GigAssembler: Build merged sequence contigs (“rafts”)

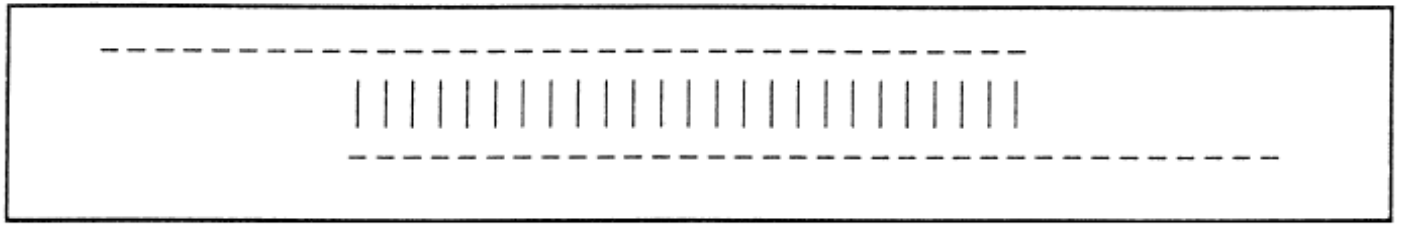
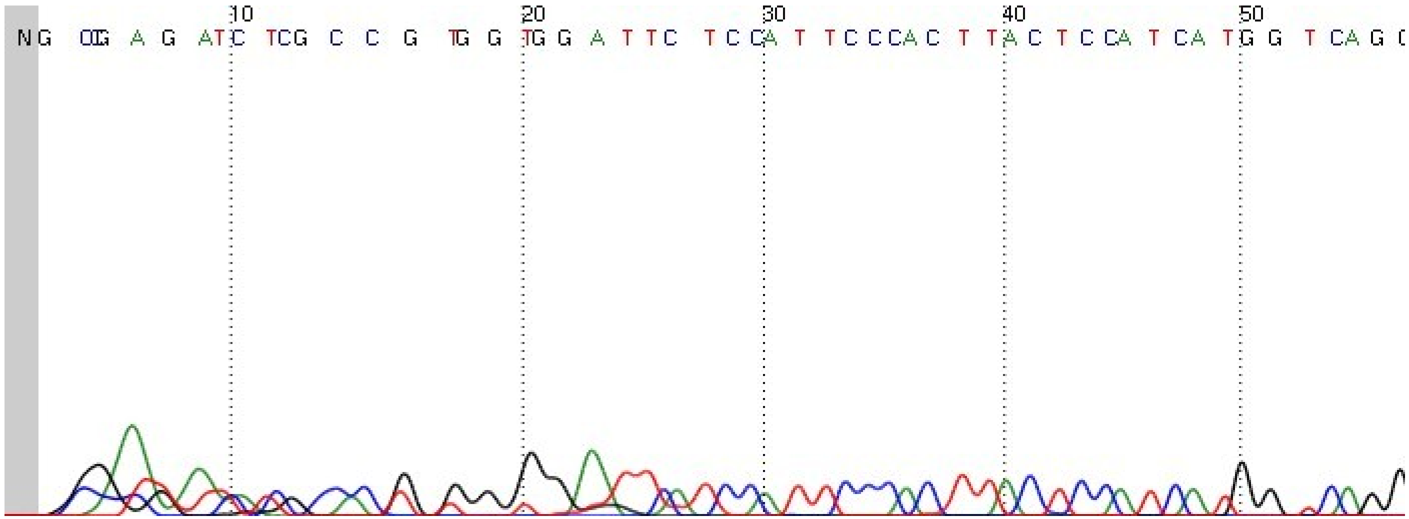


Figure 1 Two sequences overlapping end to end. The sequences are represented as dashes. The aligning regions are joined by vertical bars. End-to-end overlap is an extremely strong indication that two sequences should be joined into a contig.

Sequencing quality (Phred Score)



>gnl|ti|2299297598 name:*fww01a01.x1* NCBI Accession: AC243936 Mate pair:*2299297599*

Quality score: not available | >0 - <20 | >=20 - <40 | >=40 - <60 | >=60 - <80 | >=80 - <100

0	3	6	6	6	6	6	8	8	8	8	8	8	8	8	8	6	6	8	6	6	7	13	19	10	10	10	22	13	22	22	35	35	36	30
35	32	35	36	35	36	35	35	28	28	28	30	29	28	28	28	25	23	19	21	3	0	3	13	14	10	10	15	15	15	13	9	6	6	7
7	6	6	6	6	6	6	11	11	12	16	11	10	12	22	23	30	24	23	24	31	37	44	53	53	57	38	38	35	35	35	35	31	31	37
37	37	27	21	15	8	6	6	6	6	6	14	12	16	18	28	25	31	34	43	37	36	36	50	50	50	53	53	53	57	57	57	57	57	57
57	57	57	57	57	57	59	59	57	59	59	59	68	50	50	50	50	50	53	57	57	57	57	57	68	68	68	68	68	68	68	68	68	68	68
68	68	68	68	68	68	50	51	51	51	51	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68
68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68
68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68
68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68
68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	62	62
62	62	62	62	62	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68
68	68	68	68	68	62	68	68	68	68	68	68	68	68	62	62	68	68	68	68	68	68	68	68	68	62	62	62	62	62	68	68	68	68	68
68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68
68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68
68	68	68	68	68	68	62	62	62	68	68	68	62	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	68	62	62	62	68	68	68
59	59	59	57	57	68	68	68	68	68	68	68	68	68	68	68	68	59	59	59	68	68	68	68	68	59	59	59	59	57	59	59	59	59	68

Sequencing quality (Phred Score)

$$Q = -10 \log_{10} P$$

Base-calling Error Probability

or

$$P = 10^{\frac{-Q}{10}}$$

Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90 %
20	1 in 100	99 %
30	1 in 1000	99.9 %
40	1 in 10000	99.99 %
50	1 in 100000	99.999 %

http://en.wikipedia.org/wiki/Phred_quality_score

GigAssembler: Build merged sequence contigs (“rafts”)

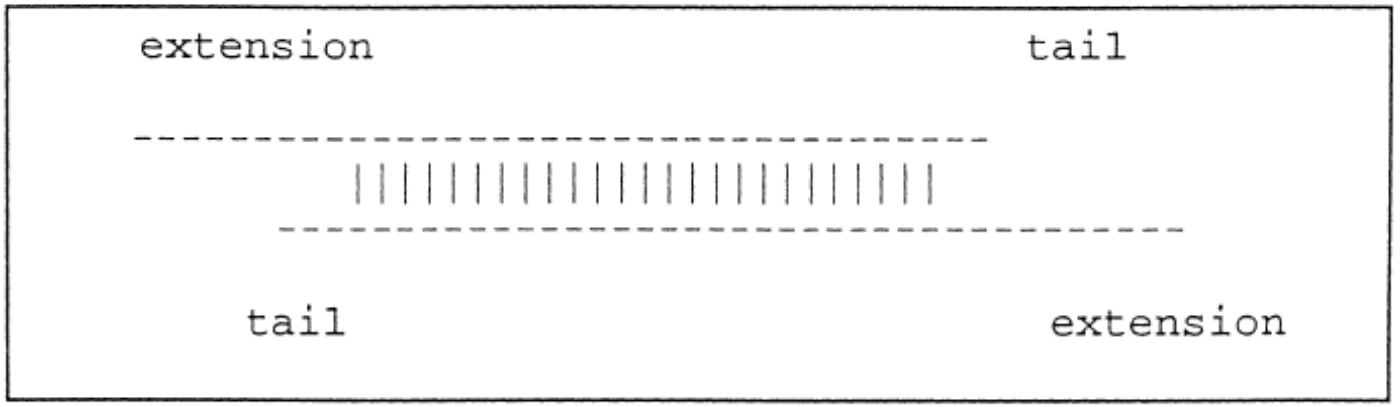


Figure 2 Two sequences with tails. The nonaligning regions on either side can be classified into ‘extensions’ and ‘tails.’ Short tails are fairly common even when two sequences should be joined into a contig because of poor quality sequence near the ends and occasional chimeric reads. Long tails, however, are generally a sign that the alignment is merely due to the sequences sharing a repeating element.

GigAssembler: Build merged sequence contigs (“rafts”)

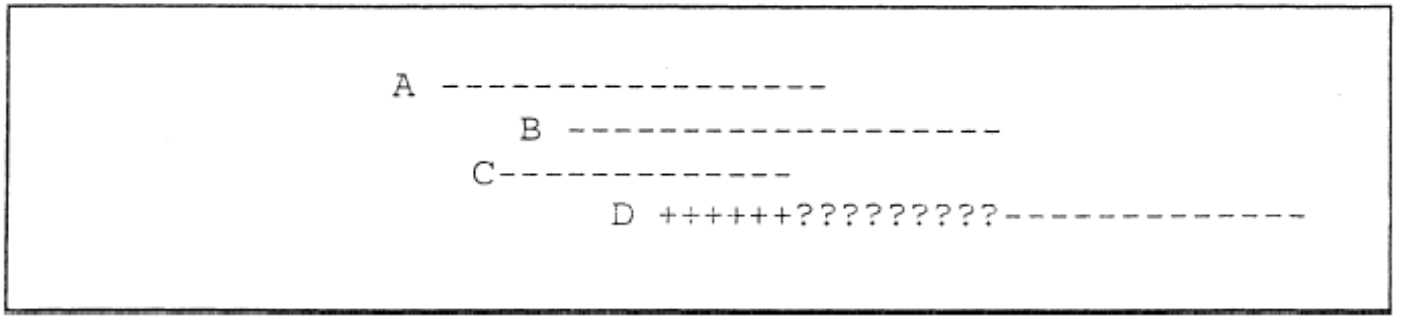


Figure 3 Merging into a raft. A contig (‘raft’) of three sequences: A, B, and C has already been constructed by `GigAssembler`. The program now examines an alignment between sequence C and a new sequence, D, to see whether D should also be added to the raft. The parts of D marked with +s are compatible with the raft because of the C/D alignment. The program must also check that the parts of D marked with ?s are compatable with the raft by examining other alignments.

GigAssembler: Build sequenced clone contigs (“barges”)

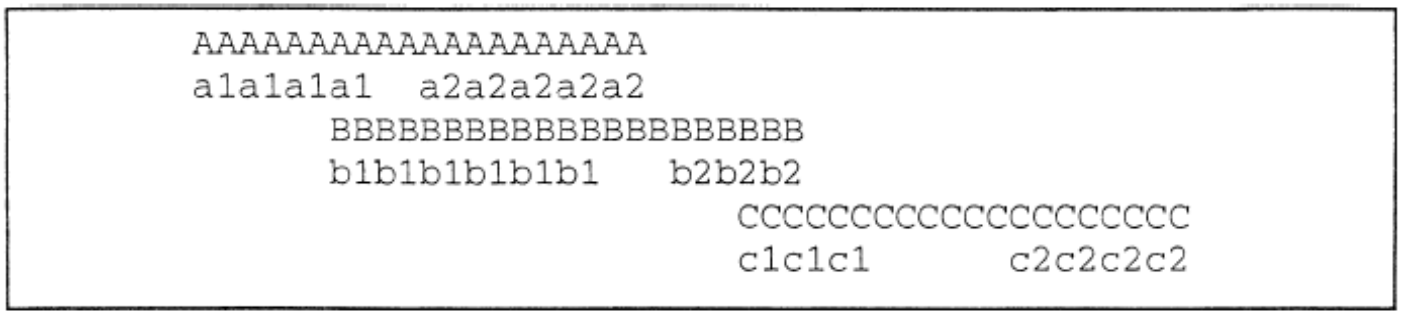


Figure 4 Three overlapping draft clones: A, B, and C. Each clone has two initial sequence contigs. Note that initial sequence contigs a1, b1, and a2 overlap as do b2 and c1.

GigAssembler: Build a “raft-ordering” graph

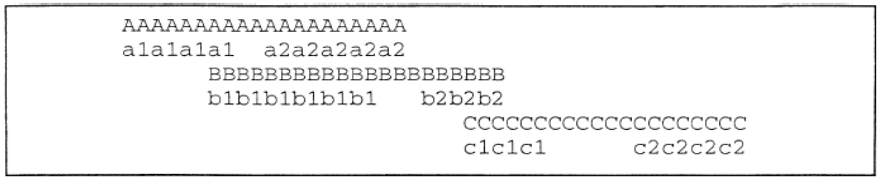


Figure 4 Three overlapping draft clones: A, B, and C. Each clone has two initial sequence contigs. Note that initial sequence contigs a1, b1, and a2 overlap as do b2 and c1.

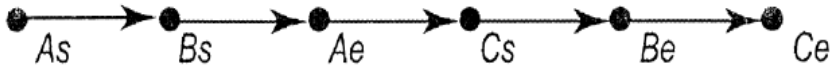


Figure 5 Ordering graph of clone starts and ends. This represents the same clones as in Fig. 4. (As) The start of clone A; (Ae) the end of clone A. Similarly Bs, Be, Cs, and Ce represent the starts and ends of clones B and C.

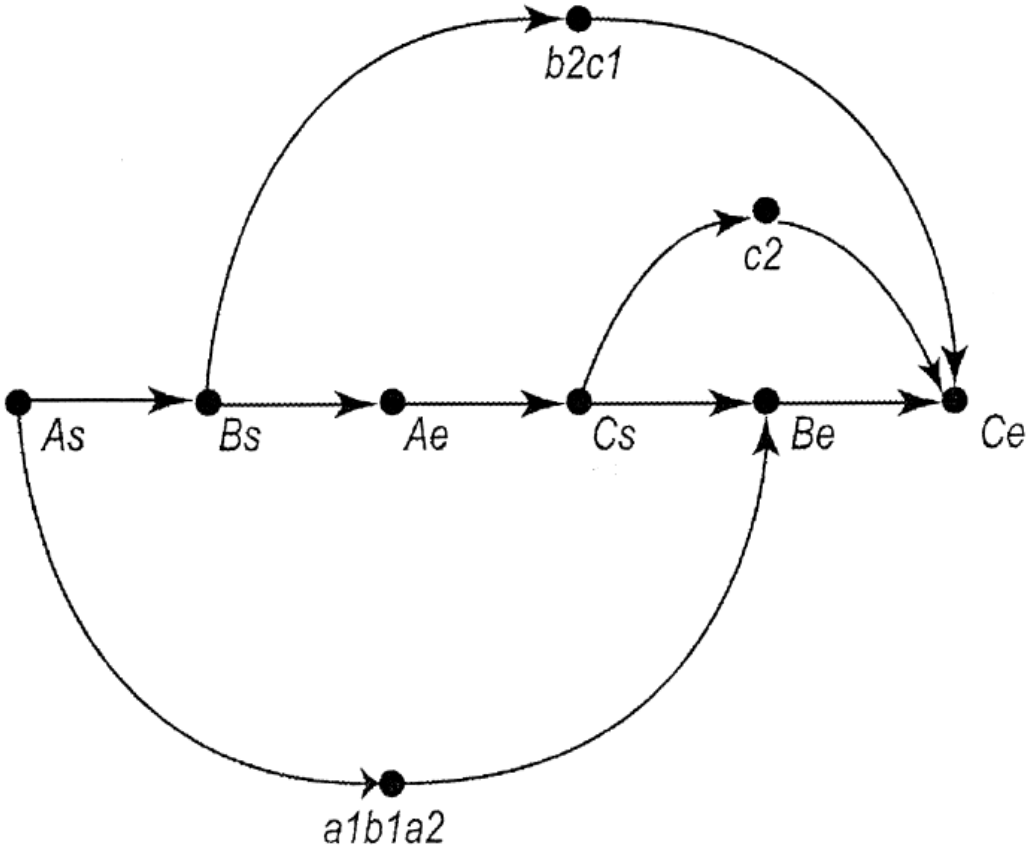


Figure 6 Ordering graph after adding in rafts. The initial sequence contigs shown in Fig. 4 are merged into rafts where they overlap. This forms three rafts: a1b1a2, b2c1, and c2. These rafts are constrained to lie between the relevant clone ends by the addition of additional ordering edges to the graph shown in Fig. 5.

GigAssembler: Build a “raft-ordering” graph

- Add information from mRNAs, ESTs, paired plasmid reads, BAC end pairs: building a “bridge”
  - Different weight to different data type: (mRNA ~ highest)
  - Conflicts with the graph as constructed so far are rejected.
- Build a sequence path through each raft.
- Fill the gap with N.
  - 100: between rafts
  - 50,000: between bridged barges

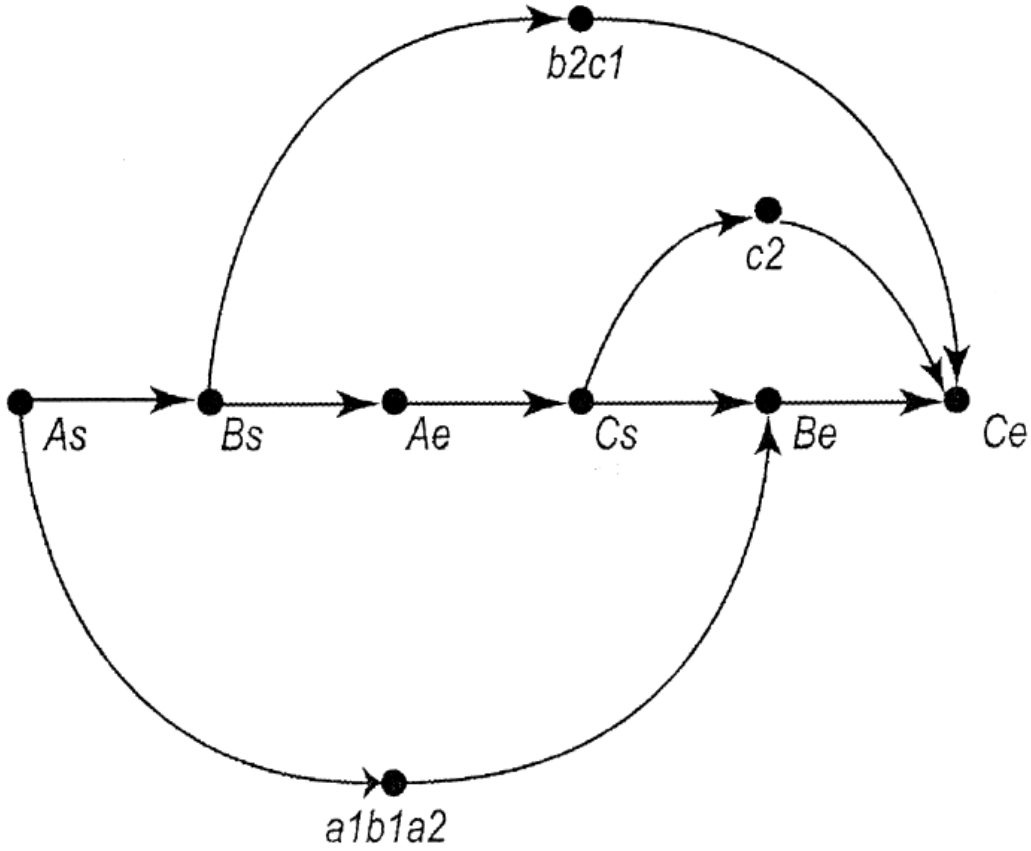
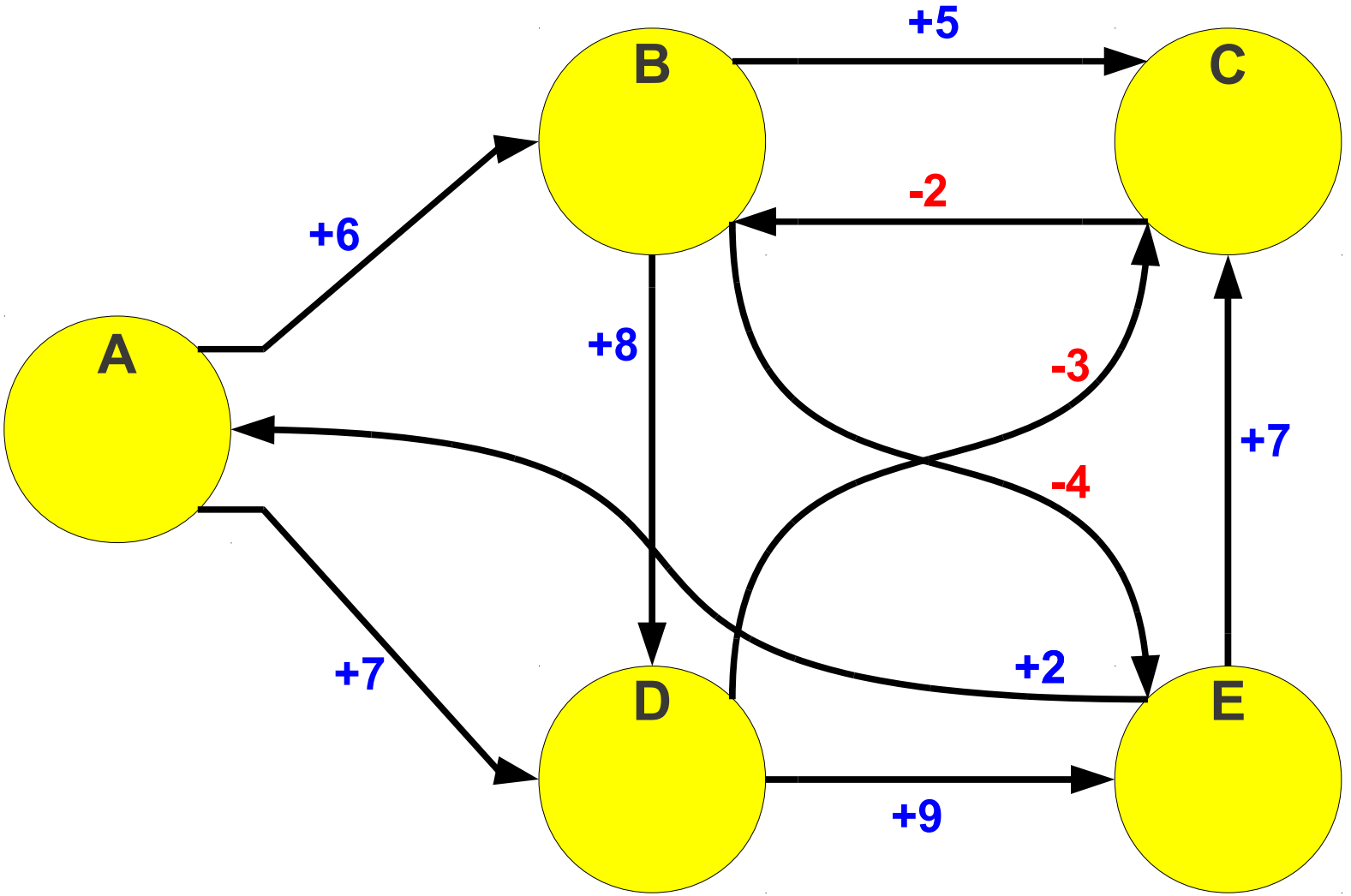


Figure 6 Ordering graph after adding in rafts. The initial sequence contigs shown in Fig. 4 are merged into rafts where they overlap. This forms three rafts: a1b1a2, b2c1, and c2. These rafts are constrained to lie between the relevant clone ends by the addition of additional ordering edges to the graph shown in Fig. 5.

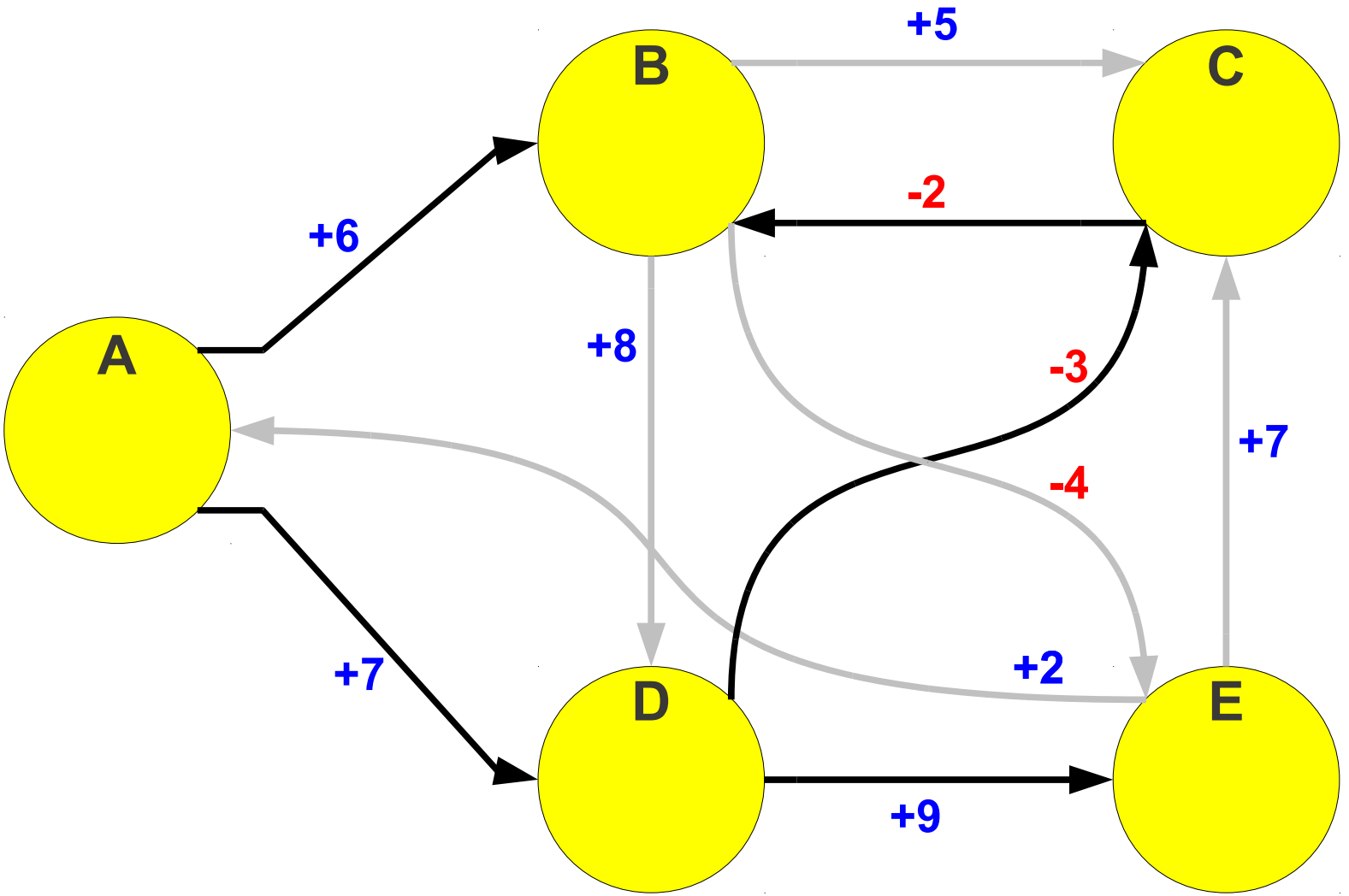
Bellman-Ford algorithm

http://compprog.wordpress.com/2007/11/29/one-source-shortest-path-the-bellman-ford-algorithm/

Find the shortest path to all nodes.

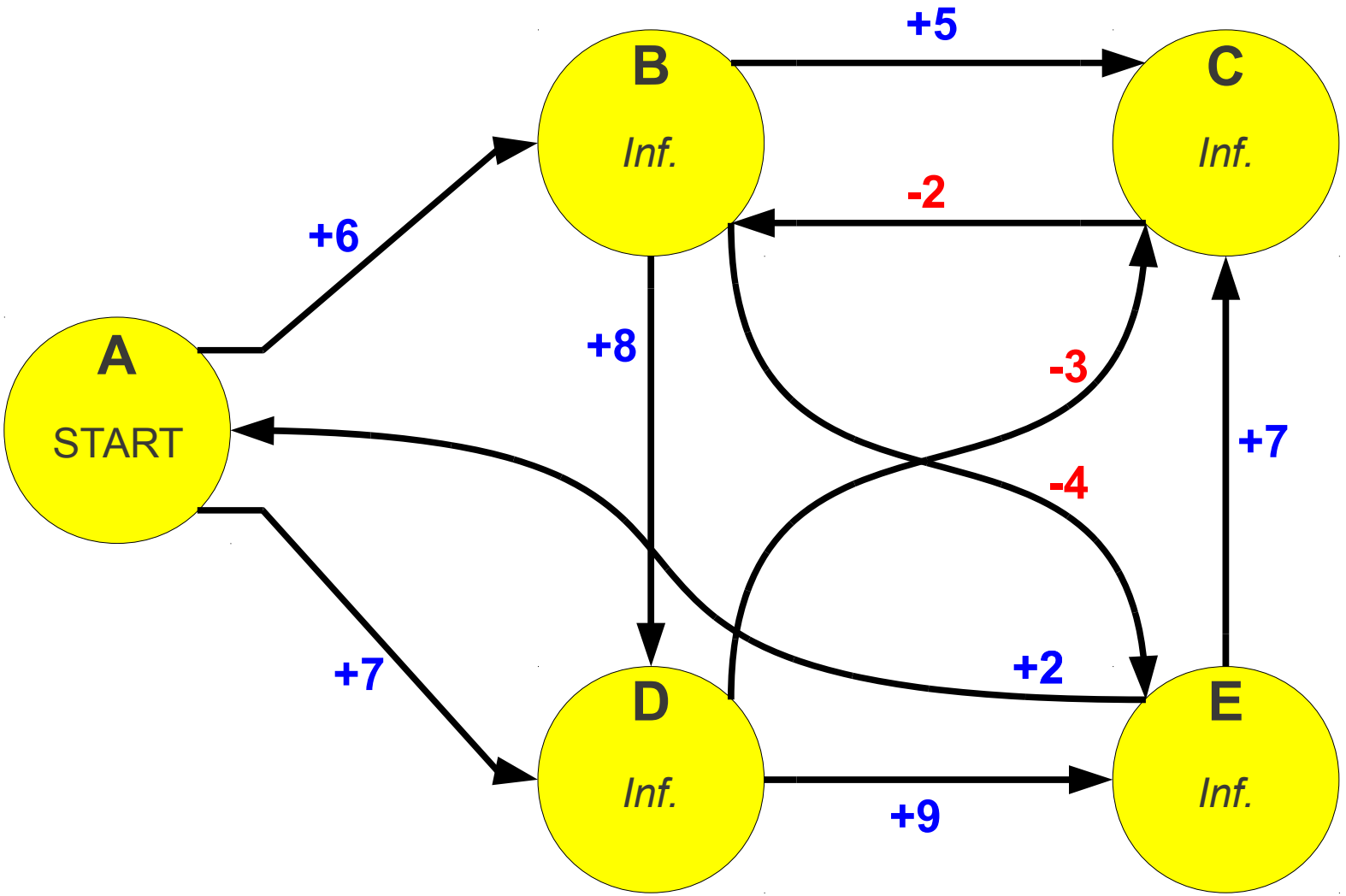


Find the shortest path to all nodes.



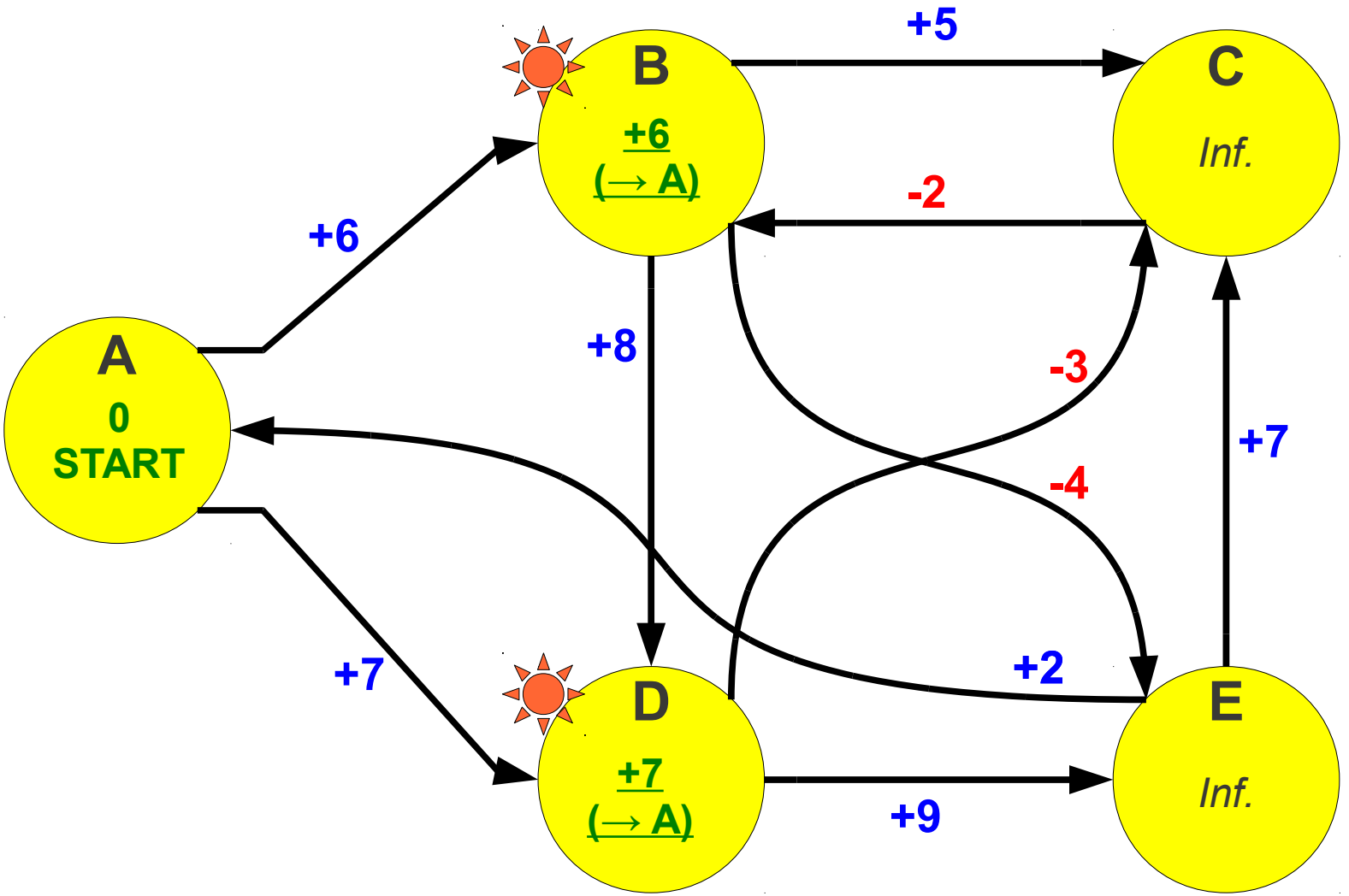
Find the shortest path to all nodes.

Take every edge and try to relax it; (N – 1 times where N is the number of nodes)



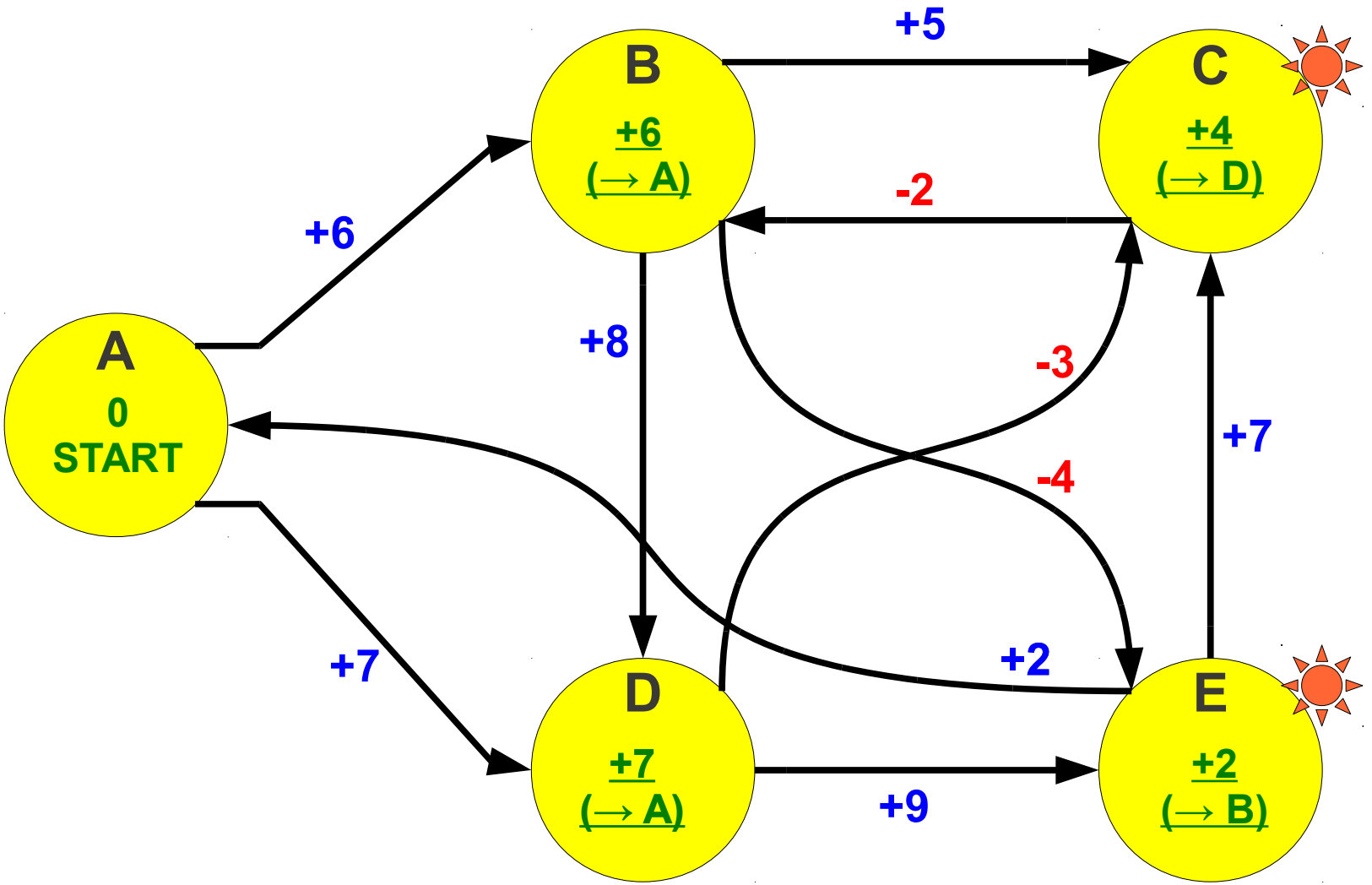
Find the shortest path to all nodes.

Take every edge and try to relax it; (N – 1 times where N is the number of nodes)



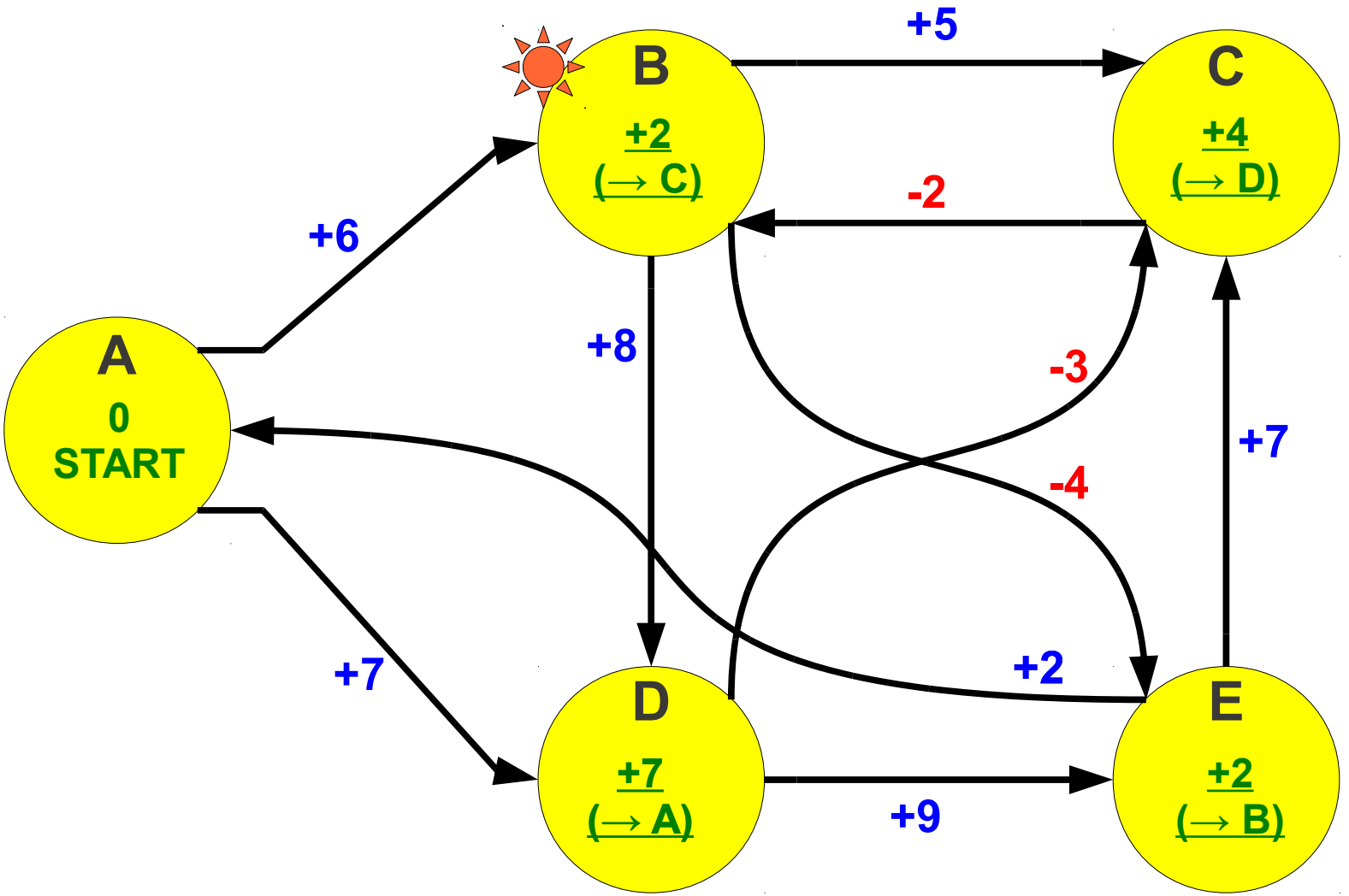
Find the shortest path to all nodes.

Take every edge and try to relax it; (N – 1 times where N is the number of nodes)



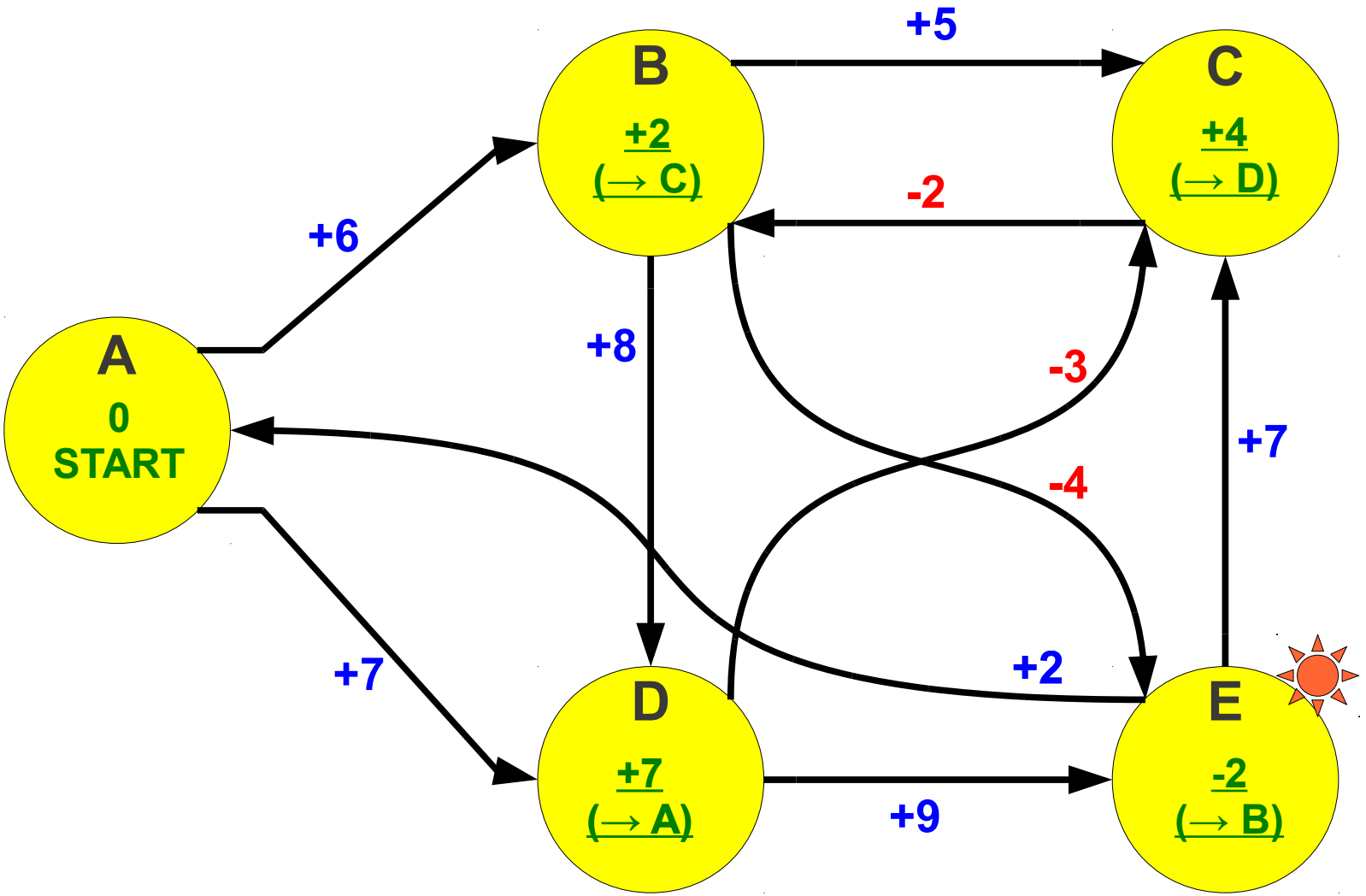
Find the shortest path to all nodes.

Take every edge and try to relax it; (N – 1 times where N is the number of nodes)

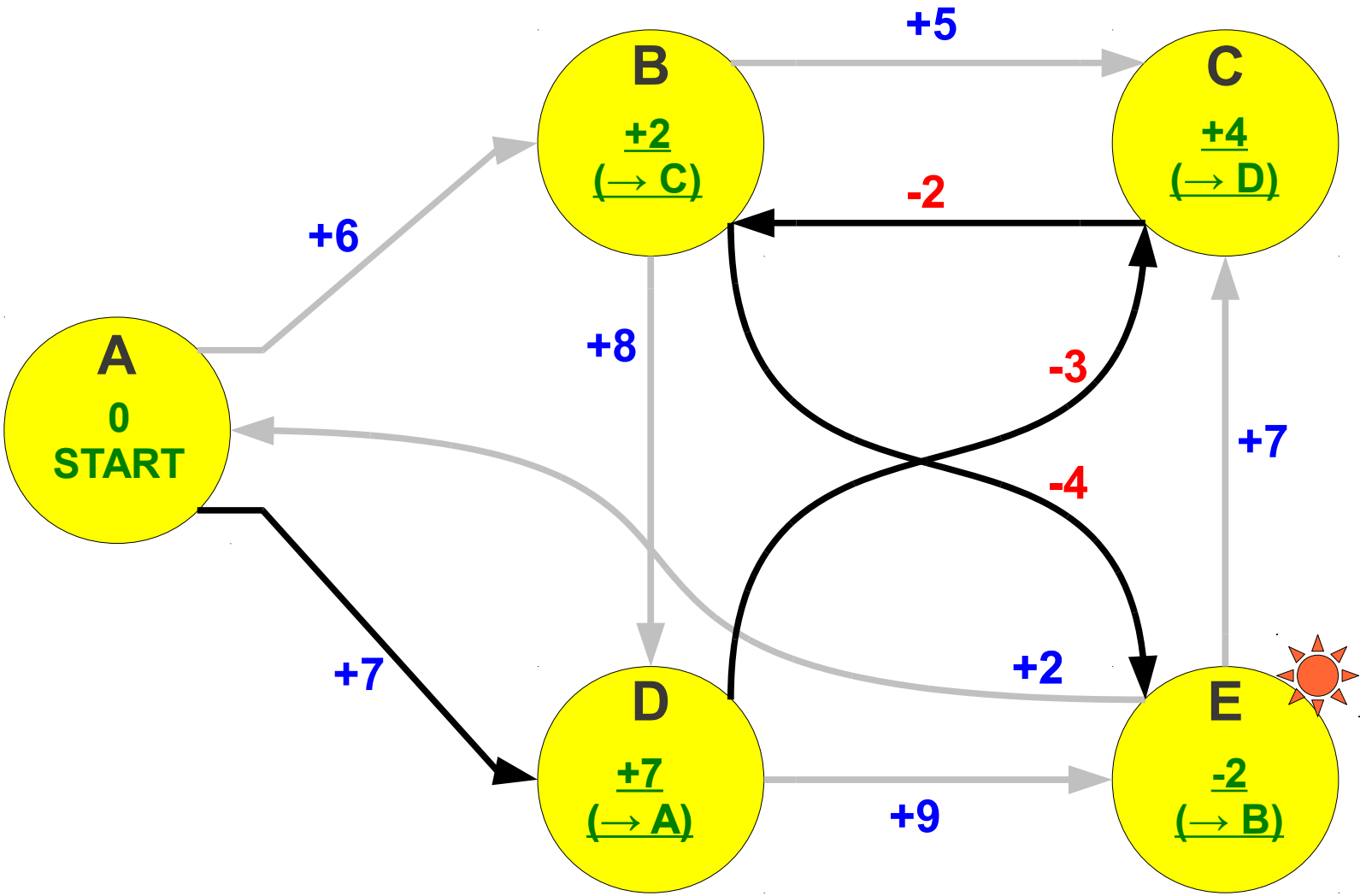


Find the shortest path to all nodes.

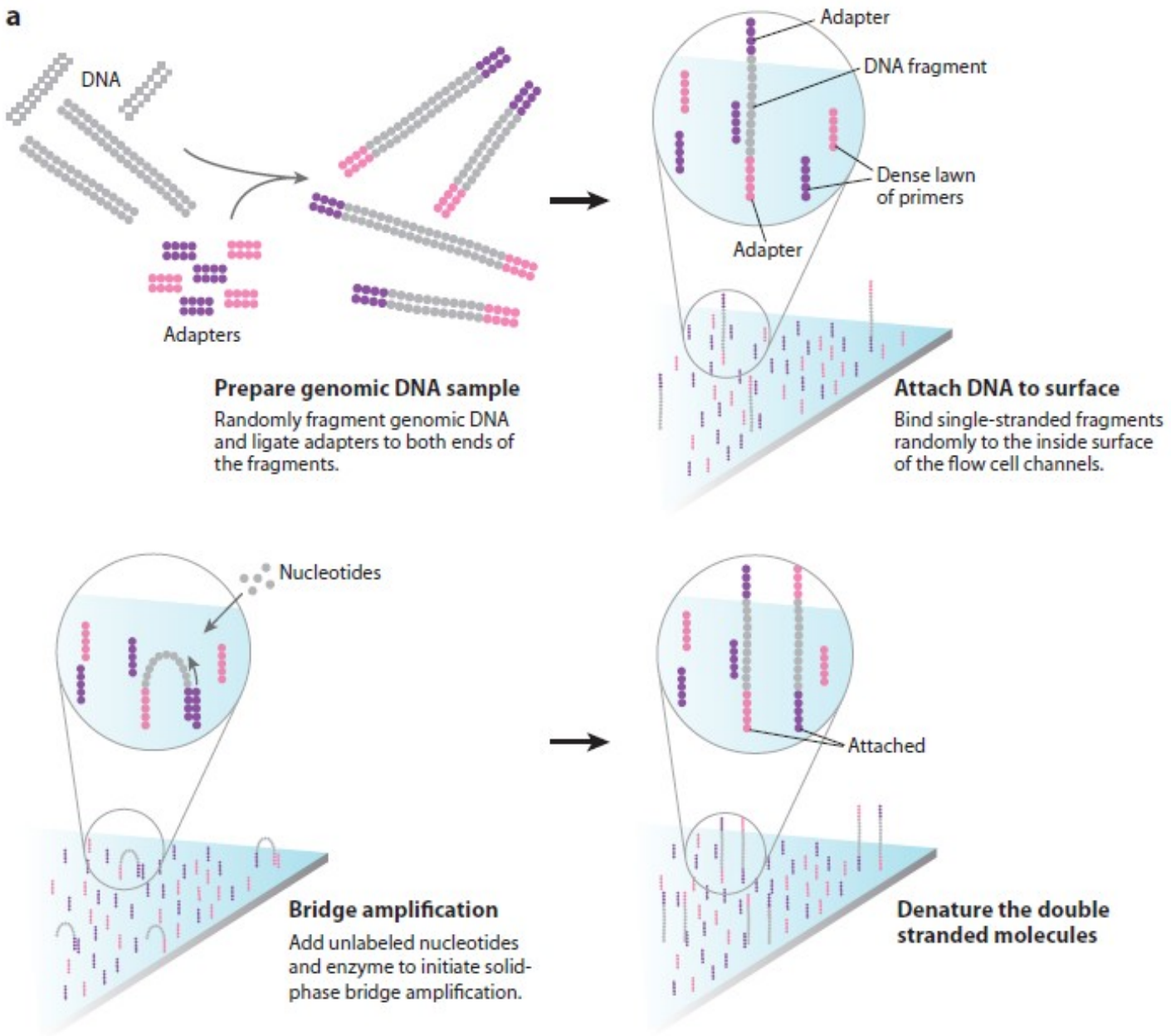
Take every edge and try to relax it; (N – 1 times where N is the number of nodes)



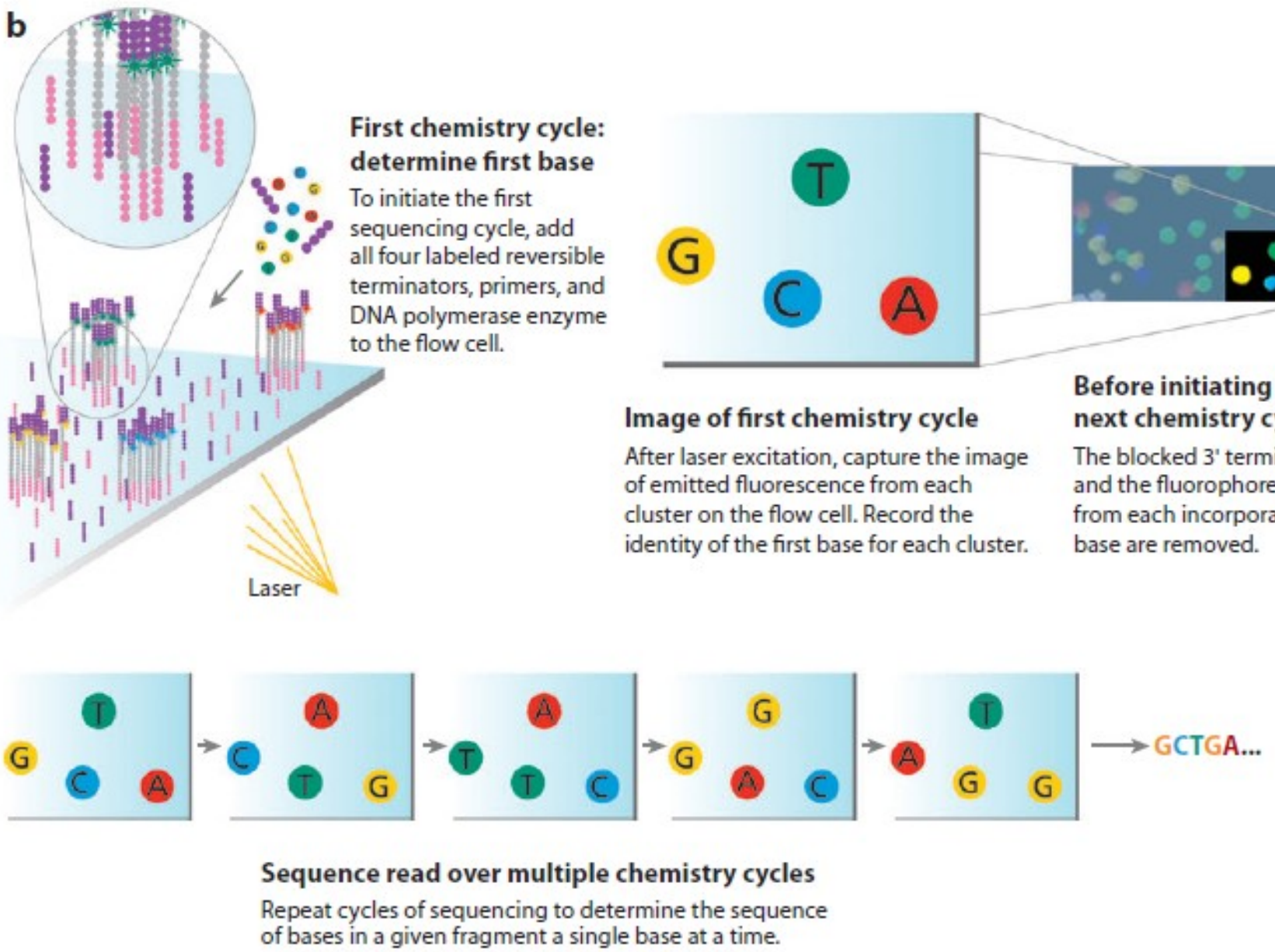
Answer: A-D-C-B-E



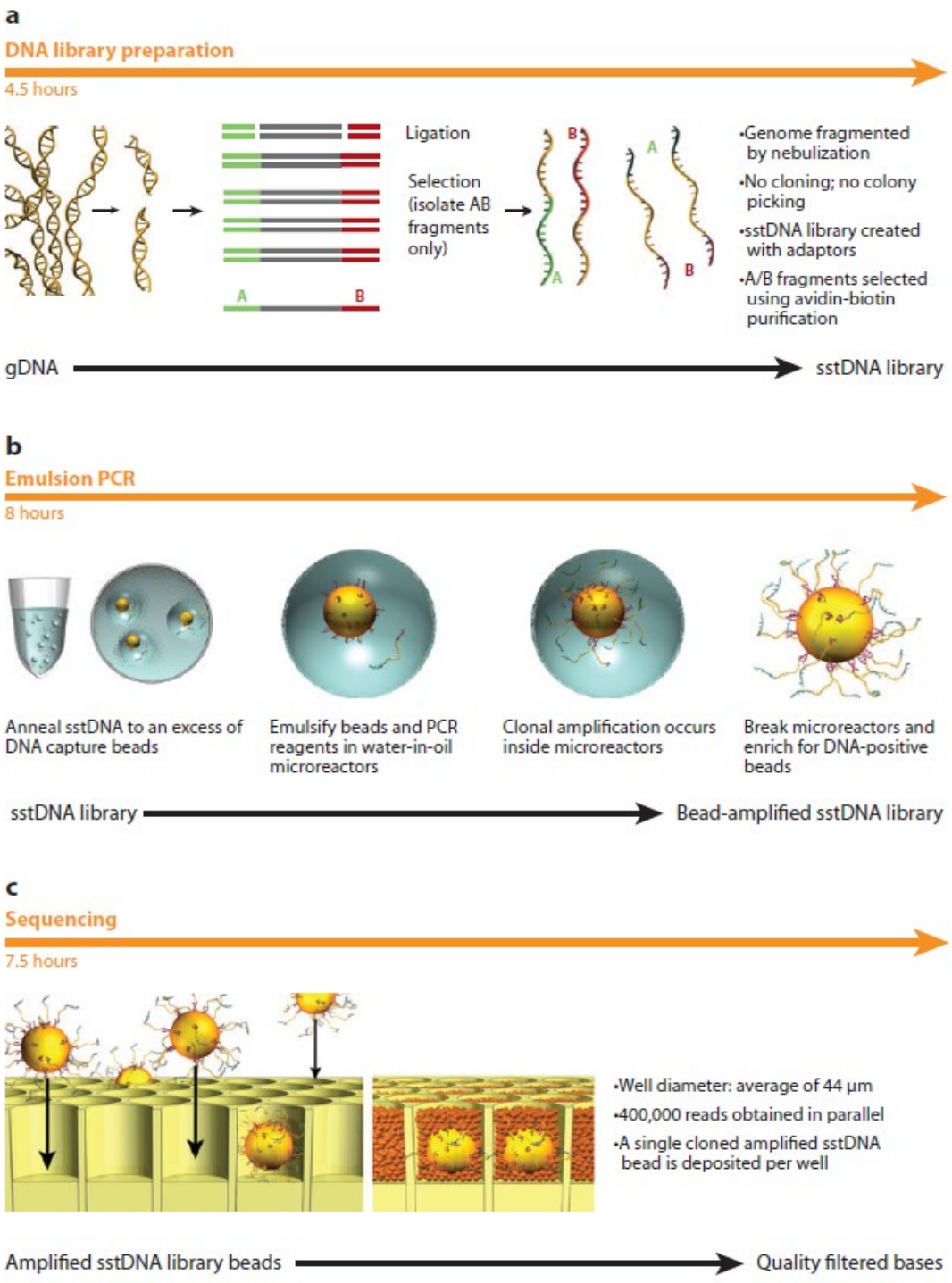
Next-generation sequencing



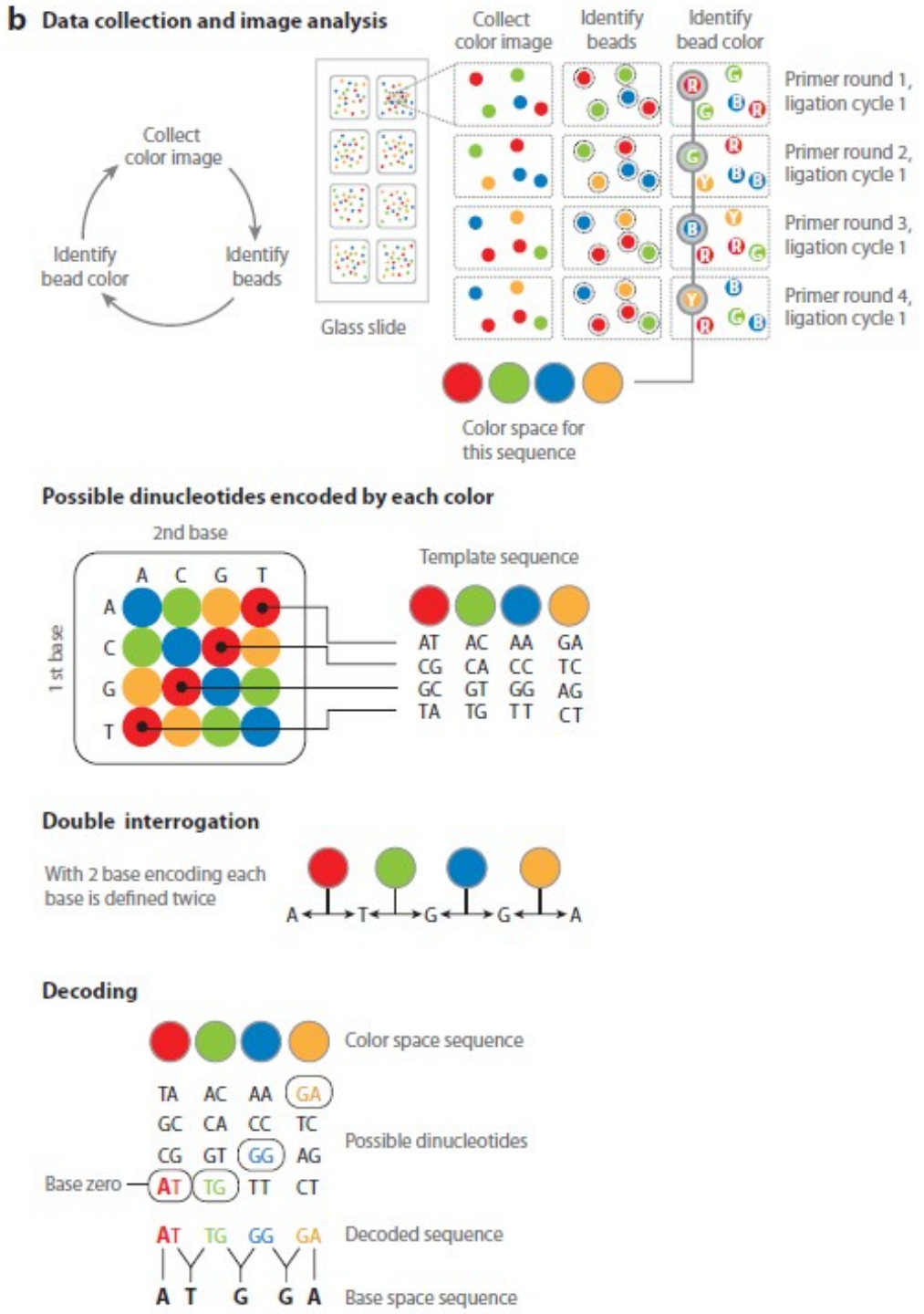
Mardis ER, *Annu. Rev. Genomics Hum. Genet.*, 2008



Mardis ER, *Annu. Rev. Genomics Hum. Genet.*, 2008



Mardis ER, *Annu. Rev. Genomics Hum. Genet.*, 2008



Mardis ER, *Annu. Rev. Genomics Hum. Genet.*, 2008

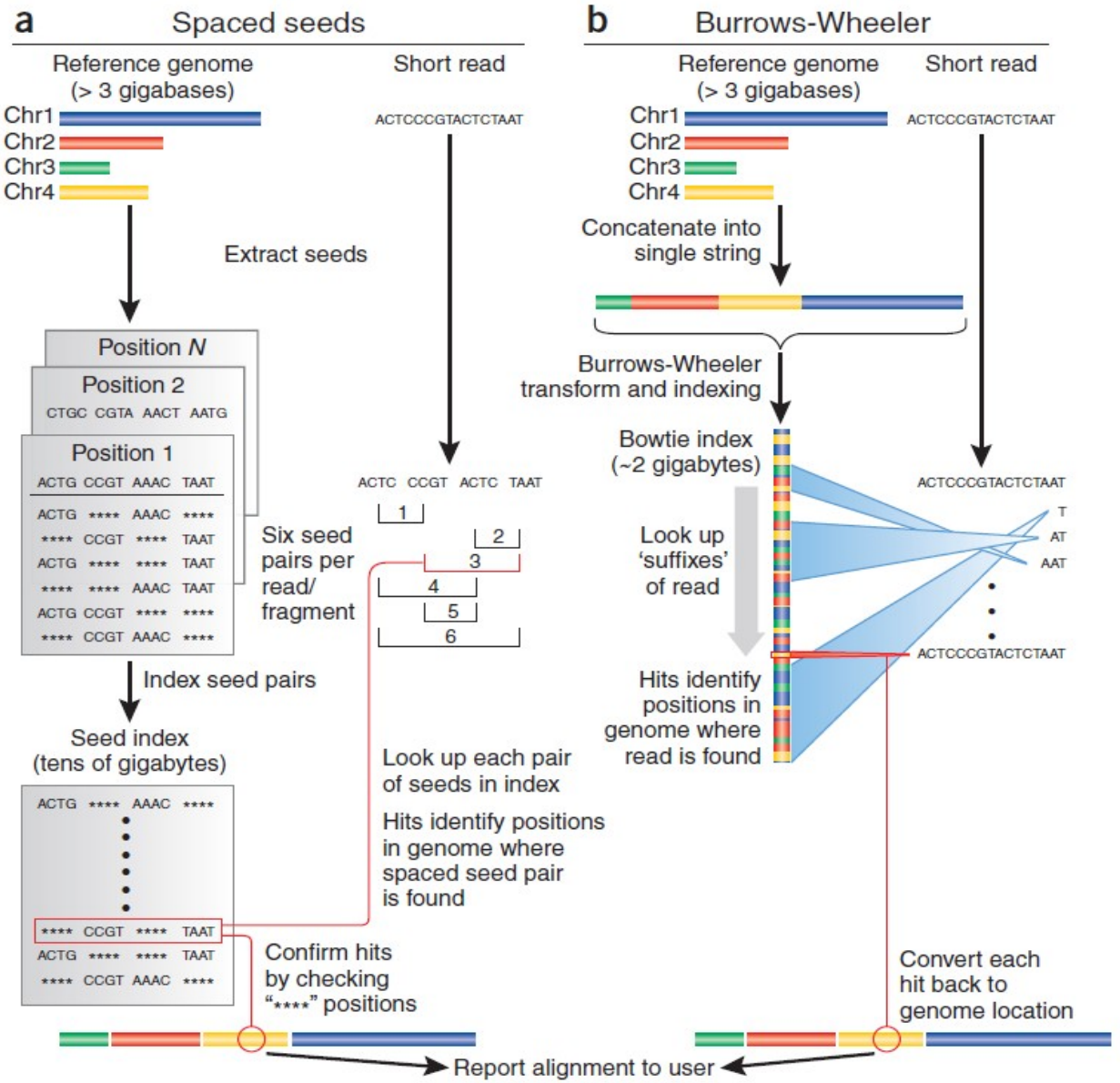
Mapping program

Table 1 A selection of short-read analysis software

Program	Website	Open source?	Handles ABI color space?	Maximum read length
Bowtie	http://bowtie.cbcb.umd.edu	Yes	No	None
BWA	http://maq.sourceforge.net/bwa-man.shtml	Yes	Yes	None
Maq	http://maq.sourceforge.net	Yes	Yes	127
Mosaik	http://bioinformatics.bc.edu/marthlab/Mosaik	No	Yes	None
Novoalign	http://www.novocraft.com	No	No	None
SOAP2	http://soap.genomics.org.cn	No	No	60
ZOOM	http://www.bioinfor.com	No	Yes	240

Trapnell C, Salzberg SL, *Nat. Biotech.*, 2009

Two strategies in mapping



Trapnell C, Salzberg SL, *Nat. Biotech.*, 2009

Real data: environment samples

```
-rw-r--r-- 1 taejoon marcotte  15M 2011-03-05 17:08 V3BC21.F3.csfasta
-rw-r--r-- 1 taejoon marcotte  32M 2011-03-08 11:54 V3BC21.F3_QV.qual
-rw-r--r-- 1 taejoon marcotte  43M 2011-03-05 17:09 V3BC22.F3.csfasta
-rw-r--r-- 1 taejoon marcotte  92M 2011-03-08 11:55 V3BC22.F3_QV.qual
-rw-r--r-- 1 taejoon marcotte  68M 2011-03-05 17:09 V3BC23.F3.csfasta
-rw-r--r-- 1 taejoon marcotte 151M 2011-03-08 11:56 V3BC23.F3_QV.qual
-rw-r--r-- 1 taejoon marcotte  38M 2011-03-05 17:09 V3BC24.F3.csfasta
-rw-r--r-- 1 taejoon marcotte  84M 2011-03-08 11:56 V3BC24.F3_QV.qual
-rw-r--r-- 1 taejoon marcotte  38M 2011-03-05 17:09 V3BC25.F3.csfasta
-rw-r--r-- 1 taejoon marcotte  85M 2011-03-08 11:56 V3BC25.F3_QV.qual
```

```
-rw-r--r-- 1 taejoon marcotte  12M 2011-03-05 17:10 V3BC21.F5.csfasta
-rw-r--r-- 1 taejoon marcotte 5.0M 2011-03-08 12:01 V3BC21.F5_QV.qual
-rw-r--r-- 1 taejoon marcotte  33M 2011-03-05 17:11 V3BC22.F5.csfasta
-rw-r--r-- 1 taejoon marcotte  64M 2011-03-08 12:01 V3BC22.F5_QV.qual
-rw-r--r-- 1 taejoon marcotte  53M 2011-03-05 17:11 V3BC23.F5.csfasta
-rw-r--r-- 1 taejoon marcotte 103M 2011-03-08 12:00 V3BC23.F5_QV.qual
-rw-r--r-- 1 taejoon marcotte  30M 2011-03-05 17:11 V3BC24.F5.csfasta
-rw-r--r-- 1 taejoon marcotte  57M 2011-03-08 12:00 V3BC24.F5_QV.qual
-rw-r--r-- 1 taejoon marcotte  30M 2011-03-05 17:12 V3BC25.F5.csfasta
-rw-r--r-- 1 taejoon marcotte  59M 2011-03-08 12:00 V3BC25.F5_QV.qual
```

Real data: environment samples

```
taejoon@cygnus:~/project/UTpond/F3$ head V3BC25.F3_QV.qual
>853_52_1477_F3
16 7 10 10 8 4 4 4 4 4 4 7 5 8 7 5 4 5 4 10 5 11 4 4 6 9 4 8 5 14 6 4 11 11 15 6 5 13 6 4 6 5 5 8 11
 6 6 4 7 16
>853_65_616_F3
4 4 10 27 27 4 4 13 10 4 5 29 7 6 13 7 5 17 6 13 6 8 6 19 5 4 6 6 10 21 13 11 27 10 12 6 24 9 4 6 9
4 12 25 4 8 8 6 11 24
>853_80_1163_F3
30 29 27 31 31 32 33 32 31 9 17 7 27 33 20 29 7 12 8 22 33 4 9 25 26 5 4 25 19 23 8 4 26 10 33 15 7
23 28 16 25 16 11 16 26 4 11 11 26 6
>853_82_1751_F3
14 33 5 24 14 25 28 12 12 23 31 19 10 27 20 27 22 8 26 22 6 28 28 28 8 24 33 23 31 28 27 24 20 19 26
 17 28 16 28 28 27 20 31 32 5 17 32 31 17 30
>853_85_1401_F3
27 32 33 23 25 31 4 26 0 6 8 0 28 8 20 24 0 18 6 11 12 4 26 23 4 4 4 11 12 6 24 4 26 6 6 10 4 27 14
12 22 6 25 23 8 27 12 26 25 14
taejoon@cygnus:~/project/UTpond/F3$ head V3BC25.F3.csfasta
>853_52_1477_F3
T313333132332323221230133331013023232233332330223
>853_65_616_F3
T111312100113332203210331020210121203313212231032 23
>853_80_1163_F3
T012332123031232330123031210223232030033330300300 01
>853_82_1751_F3
T033210332332121122330111011123122133102330323123 33
>853_85_1401_F3
T133023133021313130031320201323332030201023212300 33
```

Real data: environment samples

```
taejoon@cygnus:~/project/UTpond/NCBI.bacteria$ head -n 15 V3BC25.F3.NCBI_bacteria.gmapper_out
#FORMAT: readname contigname strand contigstart contigend readstart readend readlength score editstring
>853_168_733_F3 Acetohalobium_arabaticum|>gi|302390797|ref|NC_014378.1| -       138156  138205  1       50      50      425     13T1G27A6
>853_168_733_F3 Acetohalobium_arabaticum|>gi|302390797|ref|NC_014378.1| -       433985  434034  1       50      50      425     13T1G27A6
>853_168_733_F3 Acetohalobium_arabaticum|>gi|302390797|ref|NC_014378.1| -       796056  796105  1       50      50      425     13T1G27A6
>853_168_733_F3 Acetohalobium_arabaticum|>gi|302390797|ref|NC_014378.1| +       1424800 1424849 1       50      50      425     13T1G27A6
>853_168_733_F3 Acetohalobium_arabaticum|>gi|302390797|ref|NC_014378.1| -       10971   11020   1       50      50      425     13T1G27A6
>860_574_319_F3 Alcanivorax_borkumensis|>gi|110832861|ref|NC_008260.1|  -       531853  531900  3       50      50      455     17T30
>860_574_319_F3 Alcanivorax_borkumensis|>gi|110832861|ref|NC_008260.1|  +       2261669 2261716 3       50      50      455     17T30
>860_574_319_F3 Alcanivorax_borkumensis|>gi|110832861|ref|NC_008260.1|  -       404128  404175  3       50      50      455     17T30
>860_574_319_F3 Allochromatium_vinosum|>gi|288939764|ref|NC_013851.1|   +       2026084 2026130 4       50      50      445     16T30
>860_574_319_F3 Allochromatium_vinosum|>gi|288939764|ref|NC_013851.1|   +       2906744 2906790 4       50      50      445     16T30
>853_866_1426_F3        Azospirillum_sp.|>gi|288956841|ref|NC_013854.1| -       2179881 2179930 1       50      50      400     21G11C2G2G10
>863_1722_361_F3        Azospirillum_sp.|>gi|288956841|ref|NC_013854.1| +       521498  521547  1       50      50      450     11G3A34
>863_1722_361_F3        Acetobacter_pasteurianus|>gi|258541105|ref|NC_013209.1| -       2768567 2768616 1       50      50      450     11G3A34
>863_1722_361_F3        Anoxybacillus_flavithermus|>gi|212637849|ref|NC_011567.1|       +       11016   11065   1       50      50      450     11G3A34
taejoon@cygnus:~/project/UTpond/NCBI.bacteria$
taejoon@cygnus:~/project/UTpond/NCBI.bacteria$
taejoon@cygnus:~/project/UTpond/NCBI.bacteria$ head -n 15 V3BC25.F5.NCBI_bacteria.gmapper_out
#FORMAT: readname contigname strand contigstart contigend readstart readend readlength score editstring
>853_562_985_F5-BC       Amycolatopsis_mediterranei|>gi|300781937|ref|NC_014318.1|       +       5449850 5449883 2       35      35      298     16x10x5x3
>853_562_985_F5-BC       Azospirillum_sp.|>gi|288956841|ref|NC_013854.1| +       2070763 2070797 1       35      35      283     4C12x10x5x3
>853_562_985_F5-BC       Azotobacter_vinelandii|>gi|226942170|ref|NC_012560.1|   +       2958122 2958156 1       35      35      283     4C12x10x4x4
>853_562_985_F5-BC       Azotobacter_vinelandii|>gi|226942170|ref|NC_012560.1|   -       4827430 4827464 1       35      35      283     4C12x10x5x3
>853_562_985_F5-BC       Aromatoleum_aromaticum|>gi|56475432|ref|NC_006513.1|    -       1714032 1714066 1       35      35      283     4C12x10x5x3
>857_160_628_F5-BC       Acidovorax_sp.|>gi|121592436|ref|NC_008782.1|   +       1825323 1825356 2       35      35      315     4C29
>857_160_628_F5-BC       Alkalilimnicola_ehrlichii|>gi|114319166|ref|NC_008340.1|       +       658523  658556  2       35      35      315     4C29
>857_160_628_F5-BC       Acidithiobacillus_ferrooxidans|>gi|198282148|ref|NC_011206.1|   +       610447  610481  1       35      35      308     4x1x1x29
>857_160_628_F5-BC       Acidovorax_ebreus|>gi|222109225|ref|NC_011992.1|        -       2100565 2100598 2       35      35      290     4C25C3
>857_160_628_F5-BC       Acidovorax_avenae|>gi|120608714|ref|NC_008752.1|        +       2931102 2931135 2       35      35      290     4C3C25
>853_1779_1130_F5-BC     Acidiphilium_cryptum|>gi|148259021|ref|NC_009484.1|     -       2893817 2893850 1       34      35      284     5x5x20x2x2
>863_1722_361_F5-BC      Anoxybacillus_flavithermus|>gi|212637849|ref|NC_011567.1|       -       246812  246846  1       35      35      286     19T3x9C2
>863_1722_361_F5-BC      Acidobacterium_capsulatum|>gi|225871699|ref|NC_012483.1|        -       2864186 2864220 1       35      35      286     19T3x8C3
>863_1722_361_F5-BC      Anoxybacillus_flavithermus|>gi|212637849|ref|NC_011567.1|       -       84667   84701   1       35      35      286     19T3x9C2
taejoon@cygnus:~/project/UTpond/NCBI.bacteria$
```