Biol 47800/59500 Homework 4

Work individually on this assignment. Feel free to write a program to perform the alignments if you like (it took me about 1.5 hr to do so).

1. BLOSUM scoring

   a. Explain the meaning of the suffixes 100, 80, 62, etc in the BLOSUM series (i.e., BLOSUM100, BLOSUM80, BLOSUM62).

The number indicates the percent identity between sequences that will be clustered together

   b. Calculate the BLOSUM100 and BLOSUM80 scoring matrix for the block of sequences shown below. Indicate which sequences you cluster together in each case.

```
ACGCTAGCTG
TCCCCTGATG
AAGCAAGCTC
TCCCCGGCTG
ACGCATGATG
```

BLOSUM 100: No sequences are 100% identical so there is no clustering
5*4/2 = 10 transitions per position = 100 total transitions

Foreground counts

	1	2	3	4	5	6	7	8	9	10		Total count
	ATATA	CCACC	GCGCG	CCCCC	TCACA	ATAGT	GGGGG	CACCA	TTTTT	GGCGG		
AA	3				1	1		1				6
AC		4			4			6				14
AG						2						2
AT	6				2	4						12
CC		6	1	10	1			3				21
CG			6							4		10
CT					2							2
GG			3				10			6		19
GT						2						2
TT	1					1			10			12
all	10	10	10	10	10	10	10	10	10	10		100

Background counts

A	3	1			2	2		2				10	0.200
C		4	2	5	2			3		1		17	0.340
G			3			1	5			4		13	0.260
T	2				1	2			5			10	0.200

BLOSUM100 Log-odds

	f = n /c	b= random background	odds ratio r=f/b	log odds (base 10)	log odds (bits)
AA	0.060	0.040	1.500	0.176	0.585
AC	0.140	0.136	1.029	0.013	0.042
AG	0.020	0.104	0.192	-0.716	-2.379
AT	0.120	0.080	1.500	0.176	0.585
CC	0.210	0.116	1.817	0.259	0.861
CG	0.100	0.177	0.566	-0.247	-0.822
CT	0.020	0.136	0.147	-0.833	-2.766
GG	0.190	0.068	2.811	0.449	1.491
GT	0.020	0.104	0.192	-0.716	-2.379
TT	0.120	0.040	3.000	0.477	1.585

BLOSUM80

% identity between sequences

	1	2	3	4	5
1	100	50	70	70	70
2		100	30	80	70
3			100	50	60
4				100	50
5					100

Sequences 1 and 2, and 2 and 4 are 80% identical and will be clustered

Effective number of sequences is 3; 3*2/2 = 3 transitions per position = 30 total transitions

use the same background frequencies as above

	AA(TT)A	CA(CC)C	GG(CC)G	CC(CC)C	TA(CC)A	AA(TG)T	GG(GG)G	CC(AC)A	TT(TT)T	GC(GG)G	total
AA	3				1	1		0.5			5.5
AC		3			2			3.5			8.5
AG						1					1
AT	3				2	3					8
CC		3		6				2			11
CG			3							3	6
CT					1						1
GG			3				6			3	12
GT						0.5					0.5
TT						0.5			6		6.5
all	6	6	6	6	6	6	6	6	6	6	60

Log-odds

f = n /c	b= random background	odds ratio r=f/b	log odds (base 10)	log odds (bits)
0.040	2.292	0.360	1.196	0.040
0.136	1.042	0.018	0.059	0.136
0.104	0.160	-0.795	-2.642	0.104
0.080	1.667	0.222	0.737	0.080
0.116	1.586	0.200	0.665	0.116
0.177	0.566	-0.247	-0.822	0.177
0.136	0.123	-0.912	-3.029	0.136
0.068	2.959	0.471	1.565	0.068
0.104	0.080	-1.096	-3.642	0.104

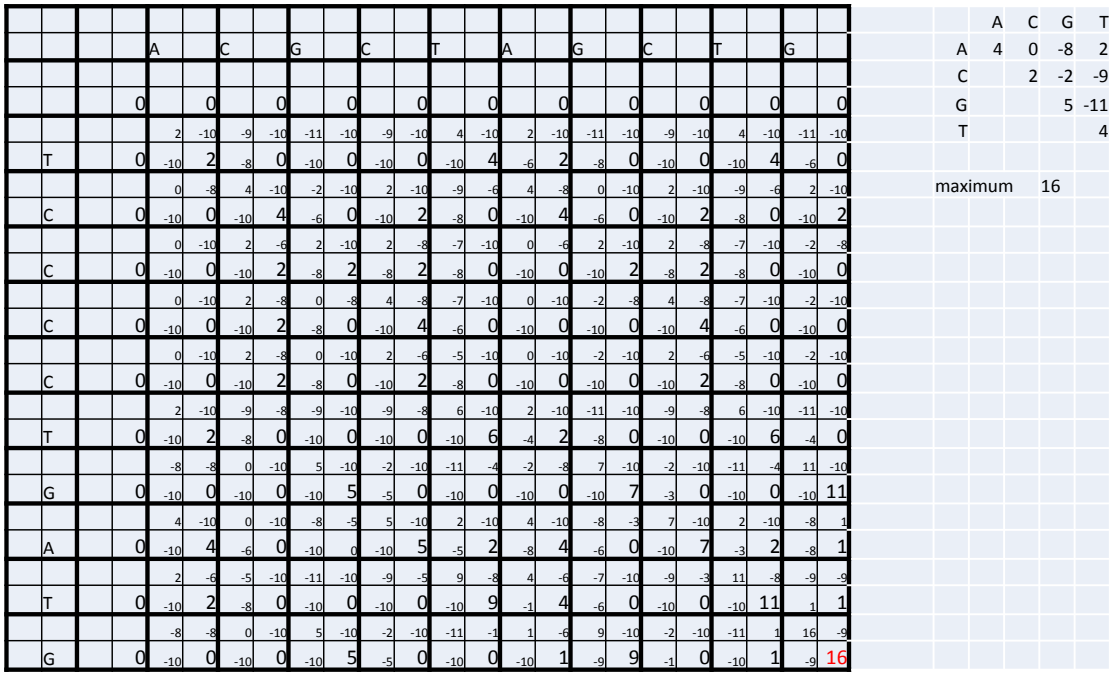
0.117	0.040	2.917	0.465	1.544

2. Using the first two sequences, above, show the local dynamic programming matrix and alignment using your BLOSUM80 matrix from question 3. Round the values in the scoring table to 1 decimal point and multiply by 10 to make this easier.

gap open = 0

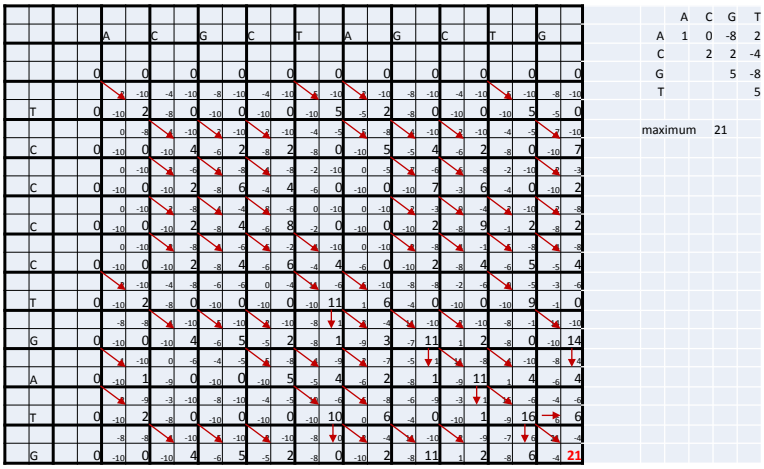
gap extend = -1

score matrix



	A	C	G	T
A	4	0	-8	2
C		2	-2	-9
G			5	-11
T				4
maximum		16		

path matrix



	A	C	G	T
A	1	0	-8	2
C		2	2	-4
G			5	-8
T				5
maximum		21		

Alignment

AGCTG

- : - : :

TGATG