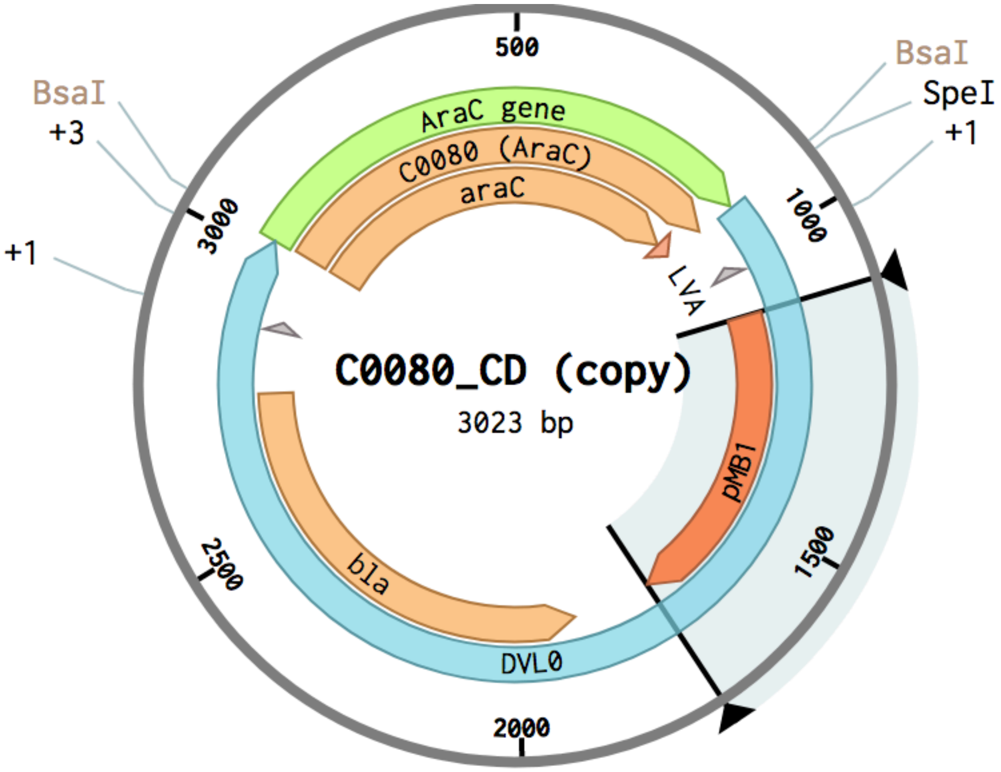
Needleman Wunsch Calculations / Dynamic Programming Application

Comparison between heuristic calculation (with the NCBI Needleman-Wunsch Global Alignment GUI) vs. theoretical calculation

Benchling Diagram for C0080_CD:



Segments:

C0080_CD: [from 842 bp to 900 bp]:

```
| = denotes a sub-segment break (easier to use the NW algorithm
heuristic / improve accuracy of the NW score):

TGTGAAGAA |    AAGTGAATGA |   ATGAATGATG |   TAGCCGTCAA |
     GTTGTGTCAG |   CTGCAAACGA |   GCAAAACTA-
```

C0040_CD: [from 583 bp to 641 bp]:

```
| = denotes a sub-segment break (easier to use the NW algorithm
heuristic / improve accuracy of the NW score):
```

TGCGGATTAG | AAAAACAACT | TAAATGTGA | ATGGGTCCGC |
     TGCAAACGAC | GAAAACTA—

Main formulas:

Individual sub-sequence NW scores:

$$NW_{metric} = \begin{cases} match += 1 \\ mismatch += -1 \\ gap += -2 \end{cases}$$

$$NW_{score} = \left(\left|\frac{match}{mismatch}\right| * \frac{1}{total\ base\ pairs}\right) - (2 * \frac{gap}{total\ base\ pairs})$$

$$Total\ NW_{score} = \sum_{i=1}^{sequences} NW_{scores} = (\left|\frac{\sum match}{\sum mismatch}\right| - \frac{1}{total\ base\ pairs} - (2 * \prod_{i=1}^{gaps} \frac{gap}{total\ base\ pairs})$$ (notice the subtraction change!)

Theoretical Calculation:

Color denoting:
Pink / red: mismatch (-1)
Green: match (+1)
Blue / cyan: gap / indel (-2)

Subsequence 1 Analysis:

C0080_CD	+1	+1	-1	+1	-1	+1	-1	-1	+1	-1
	T	G	T	G	A	A	G	A	A	A
C0040_CD	+1	+1	-1	+1	-1	+1	-1	-1	+1	-1
	T	G	C	G	G	A	T	T	A	G

Matches: 5
Mismatches: 5
Gaps: 0

$$NW_{metric} = \begin{cases} match += 1 \\ mismatch += -1 \\ gap += -2 \end{cases}$$

$$NW_{score} = \left(\left|\frac{match}{mismatch}\right| * \frac{1}{total\ base\ pairs}\right) - (2 * \frac{gap}{total\ base\ pairs})$$

$$NW_{subseq1} = \left(\left|\frac{5}{-5}\right| * \frac{1}{10}\right) - (2 * \frac{0}{10})$$

= (1/10) – (2*0) = **10%**

Subsequence 2 Analysis:

C0080_CD	+1	+1	-1	-1	-1	-1	+1	-1	-1	-1
	A	A	G	T	G	A	A	T	G	A
C0040_CD	+1	+1	-1	-1	-1	-1	+1	-1	-1	-1
	A	A	A	A	A	C	A	A	C	T

Matches: 3

Mismatches: 7

Gaps: 0

$$NW_{metric} = \begin{cases} match += 1 \\ mismatch += -1 \\ gap += -2 \end{cases}$$

$$NW_{score} = \left(\left|\frac{match}{mismatch}\right| * \frac{1}{total\ base\ pairs}\right) - (2 * \frac{gap}{total\ base\ pairs})$$

$$NW_{subseq1} = \left(\left|\frac{3}{-7}\right| * \frac{1}{10}\right) - (2 * \frac{0}{10})$$

= (3/7 * 1/10) – (2*0) = **4.29%**

Subsequence 3 Analysis:

C0080_CD	-1	-1	-1	+1	+1	+1	+1	-1	-1	-1
	A	T	G	A	A	T	G	A	T	G
C0040_CD	-1	-1	-1	+1	+1	+1	+1	-1	-1	-1
	T	A	A	A	A	T	G	T	G	A

Matches: 4

Mismatches: 6

Gaps: 0

$$NW_{metric} = \begin{cases} match += 1 \\ mismatch += -1 \\ gap += -2 \end{cases}$$

$$NW_{score} = \left(\left|\frac{match}{mismatch}\right| * \frac{1}{total\ base\ pairs}\right) - (2 * \frac{gap}{total\ base\ pairs})$$

$\text{NW}_{\text{subseq1}} = \left(\left|\frac{4}{-6}\right| * \frac{1}{10}\right) - (2 * \frac{0}{10})$

= (4/6 * 1/10) – (2*0) = **6.66%**

Subsequence 4 Analysis:

C0080_CD	-1	-1	+1	-1	-1	-1	-1	+1	-1	-1
	T	A	G	C	C	G	T	C	A	A
C0040_CD	-1	-1	+1	-1	-1	-1	-1	+1	-1	-1
	A	T	G	G	G	T	C	C	G	C

Matches: 2
Mismatches: 8
Gaps: 0

$$NW_{metric} = \begin{cases} match{+}= 1 \\ mismatch{+}= -1 \\ gap{+}= -2 \end{cases}$$

$$NW_{score} = \left(\left|\frac{match}{mismatch}\right| * \frac{1}{total\ base\ pairs}\right) - (2 * \frac{gap}{total\ base\ pairs})$$

$\text{NW}_{\text{subseq1}} = \left(\left|\frac{2}{-8}\right| * \frac{1}{10}\right) - (2 * \frac{0}{10})$

= (2/8 * 1/10) – (2*0) = **2.50%**

Subsequence 5 Analysis:

C0080_CD	-1	-1	-1	-1	+1	+1	-1	-1	-1	-1
	C	T	G	C	A	A	A	C	G	A
C0040_CD	-1	-1	-1	-1	+1	+1	-1	-1	-1	-1
	T	G	C	A	A	A	C	G	A	C

Matches: 2
Mismatches: 8
Gaps: 0

$$NW_{metric} = \begin{cases} match{+}= 1 \\ mismatch{+}= -1 \\ gap{+}= -2 \end{cases}$$

$$NW_{score} = \left(\left|\frac{match}{mismatch}\right| * \frac{1}{total\ base\ pairs}\right) - (2 * \frac{gap}{total\ base\ pairs})$$

$\mathrm{NW}_{\mathrm{subseq1}} = \left(\left|\frac{2}{-8}\right| * \frac{1}{10}\right) - (2 * \frac{0}{10})$

= (1/4 * 1/10) – (2*0) = **2.50%**

Subsequence 6 Analysis:

C0080_CD	-1	-1	+1	+1	+1	-1	-1	-1	-2	-2
	C	G	A	A	A	A	C	T	A	-
C0040_CD	-1	-1	+1	+1	+1	-1	-1	-1	-2	-2
	G	A	A	A	A	C	T	A	-	-

Matches: 3
Mismatches: 5
Gaps: 2

$$NW_{metric} = \begin{cases} match += 1 \\ mismatch += -1 \\ gap += -2 \end{cases}$$

$$NW_{score} = \left(\left|\frac{match}{mismatch}\right| * \frac{1}{total\ base\ pairs}\right) - (2 * \frac{gap}{total\ base\ pairs})$$

$\mathrm{NW}_{\mathrm{subseq1}} = \left(\left|\frac{3}{-5}\right| * \frac{1}{10}\right) - (2 * \frac{1}{60})$

= (0.666 * 1/10) – (0.0333) = **2.67 %**

Total Sequence Analysis:

Sequence	Matches	Mismatches	Gaps	Base Pair Range [start,end]
1	5	5	0	[0,10]
2	3	7	0	[11,20]
3	4	6	0	[21,30]
4	2	8	0	[31,40]
5	2	8	0	[41,50]
6	3	5	2	[51,60]
Σ	19	39	2	[0,60]

$$Total\ NW_{score} = \sum_{i=1}^{sequences} NW_{scores} = (\left|\frac{\sum match}{\sum mismatch}\right| - \frac{1}{total\ base\ pairs} - (2 * \prod_{i=1}^{gaps} \frac{gap}{total\ base\ pairs)}$$

$$= \sum_{i=1}^{6} NW_{scores} = (\left|\frac{19}{39}\right| - \frac{1}{58} - (2 * \prod_{i=1}^{gaps} * \frac{gaps}{58)})$$

$$= \sum_{i=1}^{6} NW_{scores} = (\left|\frac{19}{39}\right| - \frac{1}{58} - (2 * \prod_{i=1}^{2} * \frac{2*(2*1)}{58)})$$

= (0.487 - 0.01724) - 0.03448

= 0.46976 - 0.03448

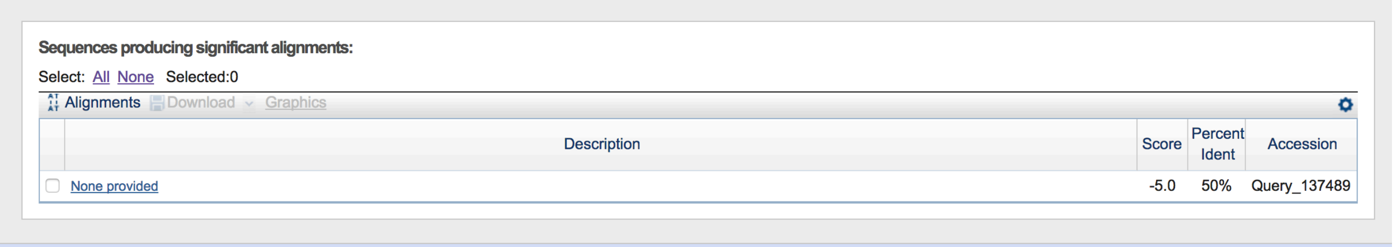
= 0.43528 = **43.53%**

Heuristic Calculation:

Website: (with screenshots of the results):

Sequence 1:

Total NW Score: -5.0



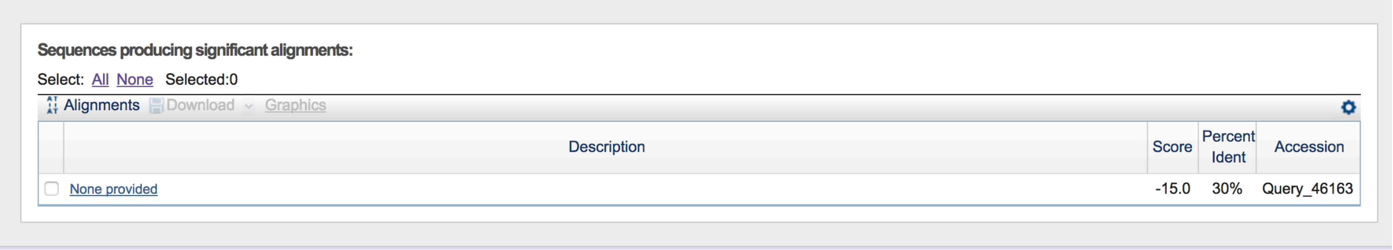
Sequences producing significant alignments:

Select: All None Selected:0

	Description	Score	Percent Ident	Accession
☐	None provided	-5.0	50%	Query_137489

Sequence 2:

Total NW Score: -15.0



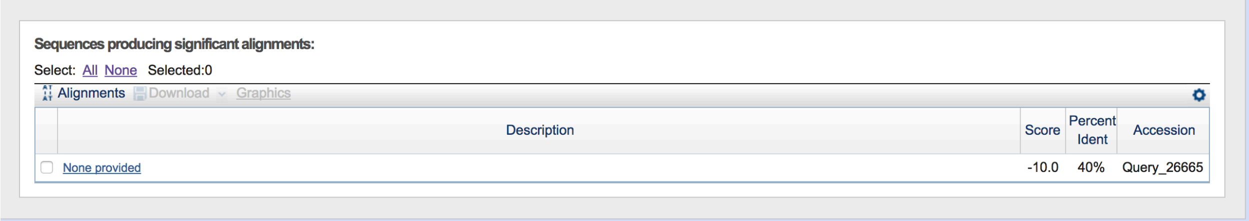
Sequences producing significant alignments:

Select: All None Selected:0

	Description	Score	Percent Ident	Accession
☐	None provided	-15.0	30%	Query_46163

Sequence 3:

Total NW Score: -10.0

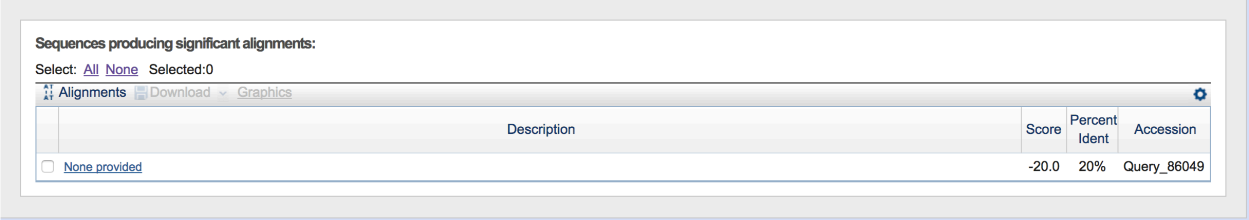


Sequences producing significant alignments:

	Description	Score	Percent Ident	Accession
□	None provided	-10.0	40%	Query_26665

Sequence 4:

Total NW Score: -20.0

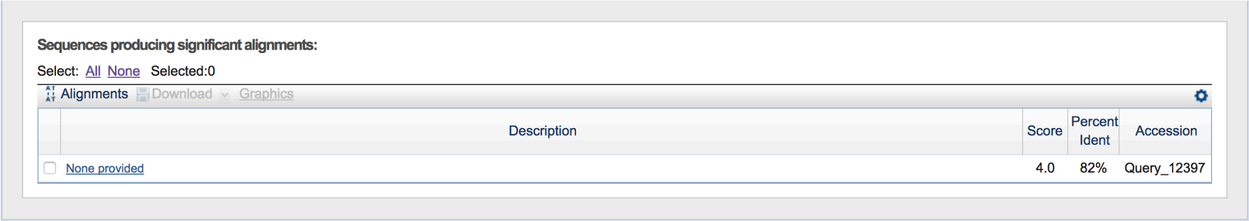


Sequences producing significant alignments:

	Description	Score	Percent Ident	Accession
□	None provided	-20.0	20%	Query_86049

Sequence 5:

Total NW Score: + 4.0

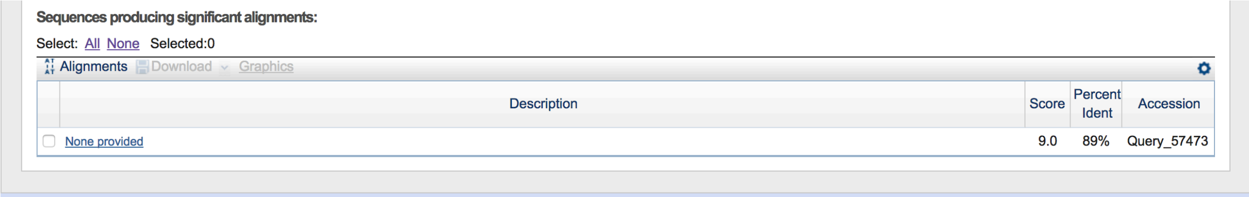


Sequences producing significant alignments:

	Description	Score	Percent Ident	Accession
□	None provided	4.0	82%	Query_12397

Sequence 6:

Total NW Score: + 9.0



Sequences producing significant alignments:

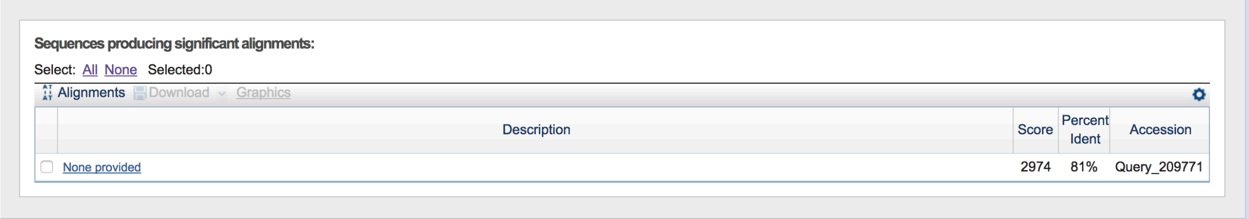
	Description	Score	Percent Ident	Accession
□	None provided	9.0	89%	Query_57473

Total C0080_CD and C0040_CD Global Alignment:

Σ NW Score (all of the 58 base pairs from C0080_CD and C0040_CD together):

+ 40.5 (half of 81% given the overlapping of subsequences) [2776/3424]

Total Sequence Part 1:

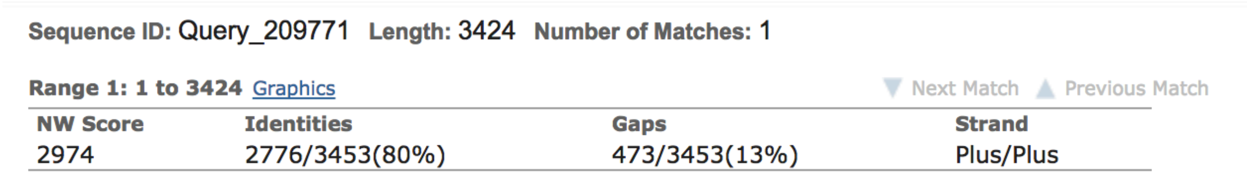


Sequences producing significant alignments:

Select: All None Selected:0

	Description	Score	Percent Ident	Accession
□	None provided	2974	81%	Query_209771

Total Sequence Part 2 (with gap / indel calculations):



Sequence ID: Query_209771 Length: 3424 Number of Matches: 1

Range 1: 1 to 3424 Graphics

NW Score	Identities	Gaps	Strand
2974	2776/3453(80%)	473/3453(13%)	Plus/Plus

Dynamic programming approach (with the first segment only):

C0080_CD: `TGTGAAGAAA`

C0040_CD: `TGCGGATTAG`

Summary of Theoretical vs. Heuristic Approaches:

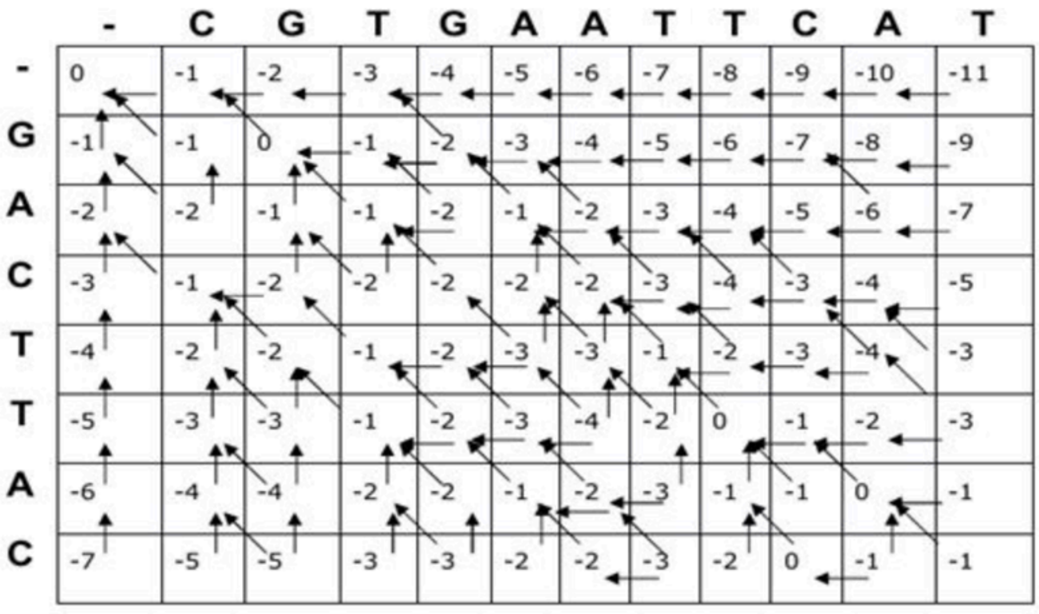
Color Coding:

Theoretical: Red

Heuristic: Blue

Theoretical		Heuristic		
Sequence	**Value**	**Sequence**	**Value**	**Absolute Value**
1	**10%**	1	-5.0	5.0
2	**4.29%**	2	-15.0	15.0
3	**6.66%**	3	-10.0	10.0
4	**2.50%**	4	-20.0	20.0
5	**2.50%**	5		
6	**2.67%**	6		
Σ	**43.53%**	Σ		

Example (from the website: http://vlab.amrita.edu/?sub=3&brch=274&sim=1431&cnt=1):



	-	C	G	T	G	A	A	T	T	C	A	T
-	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11
G	-1	-1	0	-1	-2	-3	-4	-5	-6	-7	-8	-9
A	-2	-2	-1	-1	-2	-1	-2	-3	-4	-5	-6	-7
C	-3	-1	-2	-2	-2	-2	-2	-3	-4	-3	-4	-5
T	-4	-2	-2	-1	-2	-3	-3	-1	-2	-3	-4	-3
T	-5	-3	-3	-1	-2	-3	-4	-2	0	-1	-2	-3
A	-6	-4	-4	-2	-2	-1	-2	-3	-1	-1	0	-1
C	-7	-5	-5	-3	-3	-2	-2	-3	-2	0	-1	-1

Figure 3: Matrix filling with back pointers

Our Dynamic Programming Matrix (F-Matrix)
Note: Screenshot using an Automatic Excel Sheet Program

Site: https://rmtheis.wordpress.com/2010/02/16/microsoft-excel-implementation-of-the-needleman-wunsch-sequence-alignment-algorithm/

Recursive Formula:

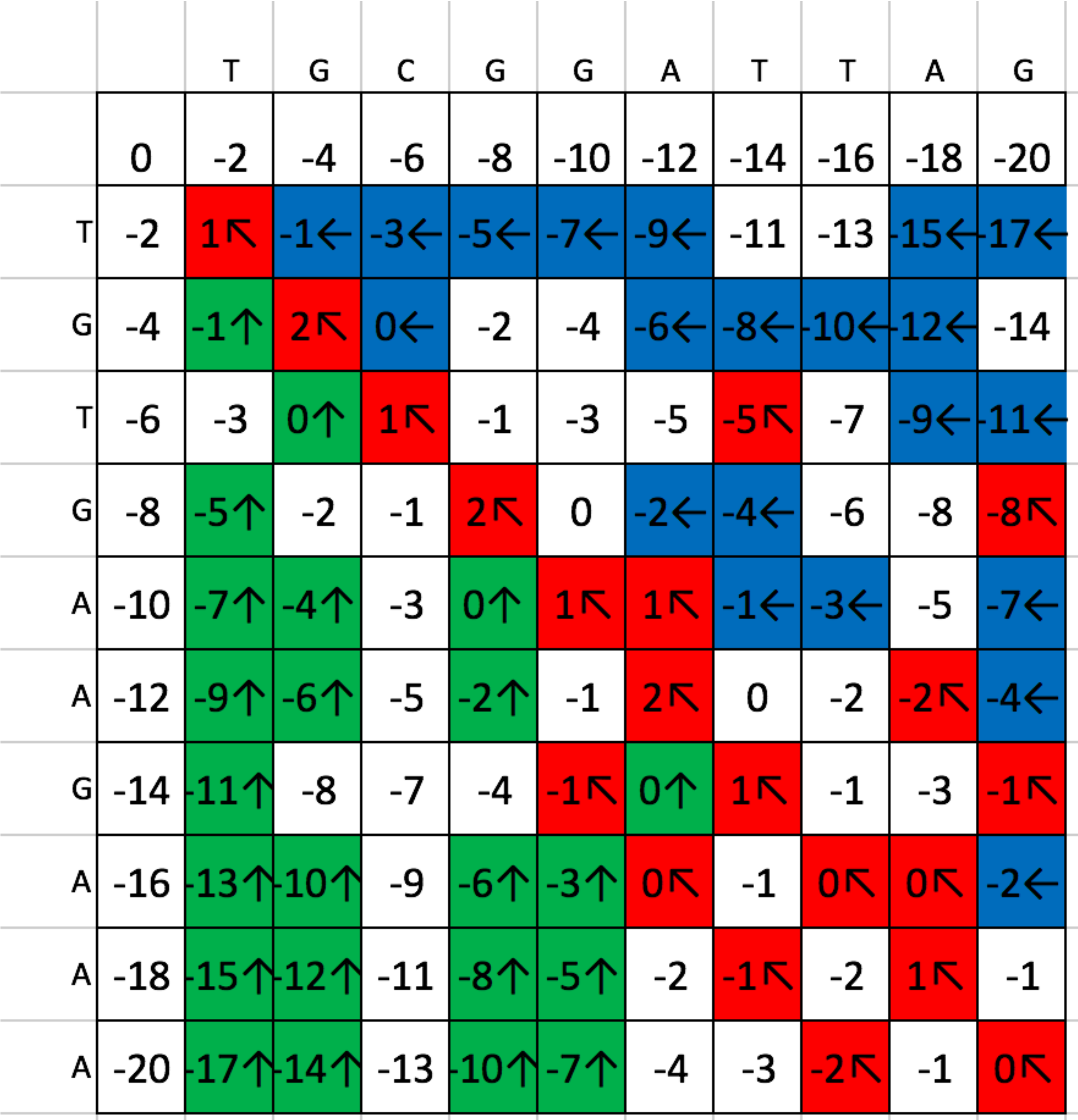
$$M_{i,j} = Max \begin{cases} M_{i-1,j-1} + S_{i,j} \\ M_{i,j-1} + W \\ M_{i-1,j} + W \end{cases}$$

Penalty Metrics:

$$NW_{metric} = \begin{cases} match += 1 \\ mismatch += -1 \\ gap += -2 \end{cases}$$

C0040_CD

C0080 CD



		T	G	C	G	G	A	T	T	A	G
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18	-20
T	-2	1↖	-1←	-3←	-5←	-7←	-9←	-11	-13	-15←	-17←
G	-4	-1↑	2↖	0←	-2	-4	-6←	-8←	-10←	-12←	-14
T	-6	-3	0↑	1↖	-1	-3	-5	-5↖	-7	-9←	-11←
G	-8	-5↑	-2	-1	2↖	0	-2←	-4←	-6	-8	-8↖
A	-10	-7↑	-4↑	-3	0↑	1↖	1↖	-1←	-3←	-5	-7←
A	-12	-9↑	-6↑	-5	-2↑	-1	2↖	0	-2	-2↖	-4←
G	-14	-11↑	-8	-7	-4	-1↖	0↑	1↖	-1	-3	-1↖
A	-16	-13↑	-10↑	-9	-6↑	-3↑	0↖	-1	0↖	0↖	-2←
A	-18	-15↑	-12↑	-11	-8↑	-5↑	-2	-1↖	-2	1↖	-1
A	-20	-17↑	-14↑	-13	-10↑	-7↑	-4	-3	-2↖	-1	0↖

As evidenced by the Dynamic Programming results shown on the above table, the best alignment for the above (shown in white) are the optimal alignments of:

Sequences = {T, GGG, TGGTG, TGGAT, GCGTA, CA, CTT, AT, CATG, CATA}
Residuals = {Calculated Later}

With the following residual values for the NW score: Calculated later