Homework 2: Final Exam Preparation

Due: Final Exam day.

Problem 1: Edit Distance

For each pair of words from the list below, determine the edit distance. (Recall, the edit distance is the smallest total number of insertions, deletions and replacements needed to turn one word into another.) Show the appropriate alignment between each pair of words (i.e., the alignment that produces the computed edit distance.) Show all work (i.e., construct the edit distance table).

1. Nucleotide Alphabet.

```
ATTGCTA
GCTATG
ATGCCTA
CATGCA
```

2. Amino Acid Alphabet

```
WALLS
WAIL
SAILS
STALLS
```

Global Alignment via Dynamic programming

For each pair of strings below, use Nudelman-Wensch algorithm to produce the best global alignment. In each case, produce the alignment table, draw the best global alignment, and supply the cost of the alignment. The substitution matrix(ces) to use and the indel penalty are specified for each case below.

1. Nucleotide alphabet. Substitution matrix: $\sigma(X,X) = 5$, $\sigma(X,Y) = -4$, if $X \neq Y$. Indel penalty: $\delta_1 = -3$; $\delta_2 = -5$.

```
CATTAGGTA
ATCATAGATA
```

2. Nucleotide alphabet. Substitution matrix: $\sigma(X, X) = 5$, $\sigma(X, Y) = -4$, if $X \neq Y$. Indel penalty: $\delta_1 = -3$; $\delta_2 = -5$.

```
ATGAATGCGCTAG
ATGATCCTGA
```

3. Amino Acid alphabet. Substitution matrices: PAM250 and BLOSUM62. Indel Penalties: $\delta_1 = -2$; $\delta_2 = -4$; $\delta_3 = -6$.

```
FEARTHEACID
EARTHHEADREST
```

4. Amino Acid alphabet. Substitution matrices: PAM250 and BLOSUM62. Indel Penalties: $\delta_1 = -2$; $\delta_2 = -4$; $\delta_3 = -6$.

```
SPEARED
PEARDEAD
```

5. Amino Acid alphabet. Substitution matrices: PAM250 and BLOSUM62. Indel Penalties: $\delta_1 = -2$; $\delta_2 = -4$; $\delta_3 = -6$.

```
LIKEWIND
LINKWILD
```

Problem 3: Local Alignment: dynamic programming

Using the Smith-Waterman algorithm, find the best local alignment for each of the pairs of strings (and combinations of substitution matrices and indel penalties) from **Problem 2**. Draw the score matrices, the best local alignments and report the alignment scores.

Additionally, find the best local alignments for the following pairs of strings.

1. Nucleotide alphabet. Substitution matrix: $\sigma(X, X) = 5$, $\sigma(X, Y) = -4$, if $X \neq Y$. Indel penalty: $\delta_1 = -3$; $\delta_2 = -5$.

```
ATTAGCCTGTGAT
GGCCTGTCCC
```

2. Nucleotide alphabet. Substitution matrix: $\sigma(X, X) = 5$, $\sigma(X, Y) = -4$, if $X \neq Y$. Indel penalty: $\delta_1 = -3$; $\delta_2 = -5$.

```
TAGGACATTCAACTTT
AAACATCAACGG
```

3. Nucleotide alphabet. Substitution matrix: $\sigma(X, X) = 5$, $\sigma(X, Y) = -4$, if $X \neq Y$. Indel penalty: $\delta_1 = -3$; $\delta_2 = -5$.

```
TAATAGTACTATGCATGTA
CCAATACCCGCAAATGTC
```

4. Amino Acid alphabet. Substitution matrices: PAM250 and BLOSUM62. Indel Penalties: $\delta_1 = -2$; $\delta_2 = -4$; $\delta_3 = -6$.

```
WILLIESDILLIES
SILLSDILES
```

Problem 4: Similarity matrix computation

For the following lists of strings, compute the distance matrix based on the global alignment scores between each pair of the strings. The substitution matrices and the indel penalties are specified for each case.

1. Nucleotide alphabet. Substitution matrix: $\sigma(X, X) = 5$, $\sigma(X, Y) = -4$, if $X \neq Y$. Indel penalty: $\delta_1 = -4$.

```
S1: CAGTCATG
S2: CATATG
S3: ATCATCATG6
S4: TACTATATG
S5: TATTATAG
S6: CATATGGG
```

2. Amino Acid alphabet. Substitution matrices: PAM250 and BLOSUM62. Indel Penalties: $\delta_2 = -4$.

```
S1: WINGS
S2: SINGS
S3: SIGN
S4: ALIGN
S5: HILL
S6: HILLS
S7: SHILL
```

Problem 5: Hierarchical Clustering

Use the agglomerative hierarchical clustering algorithm to construct the dendrogram for each of the distance matrices computed in **Problem 4**. For each matrix use *single link*, *complete link* and *average link* distance measures for recomputing intercluster distances. For each matrix, compare the three dendrograms obtained and describe any differences found (if there are any differences.

In addition, perform the same tasks for the following distance matrix (note, in this case, larger means less similar):

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
S1	0	3	2	9	12	7	10	5	5	8
S2		0	1	7	10	6	12	6	4	6
S3			0	10	7	8	9	8	4	7
S4				0	2	5	3	12	19	12
S5					0	2	4	15	16	8
S6						0	3	11	12	6
S7							0	6	6	3
S8								0	2	4
S9									0	5

Problem 6: Local Alginment. FASTA

For the pairs of strings below (first string the database string, second - the query string), construct the dot plots for matching the substrings of length 3 (i.e., perform the first step of the FASTA algorithm). Show which substrings of length 3 are found in which string (and where), and draw the dot plot.

1. Nucleotide alphabet.

```
Query:  CATGCATGC
DB:     ATCATGCAGGCTCCATGCAAT
```

2. Amino Acid alphabet.

```
Query:  SKILLS
DB:     IKILLILLSANDSAIL
```

Problem 7: Local Alginment: FASTA

For each ungpped local alignment shown below, produce the score using BLOSUM62 substitution matrix.

1.
```
DARNITISAHARDCASE
   |||   |||
 STNITARTHAR
```

2.
```
RINGSOFFIREANDWATER
    ||| |||
    SOFTIREDIDLE
```

3.
```
STRINGOFBEADTHEFTS
   |||| || ||
  GONGOFHEARTHS
```

Problem 7: Local Alginment: FASTA

For each ungpped local alignment shown below, produce the score using BLOSUM62 substitution matrix and check if any local alignments with a gap of length 1 are better. Use gap penalty $\delta = -3$.

1.
```
HITREALLYHARDANDSTRIKE
     ||||
   MYALLYISARDENT
```

2.
```
CLEARSKIESSPELLSSHINE
  |||||| |
 FEARSKISSESBYES
```

3.
```
MYTEAMISWINNINGTHISGAME
  ||||
 STEAMSWINGING
```

Problem 8: Local Alignment: BLAST

For each pair of strings below, find all matches of length 5 and extend them using BLOSUM62 substitution matrix and no gaps.

1.
```
   DB:  LIKEAREALFRIEND
Query:  NEWAREALSPRING
```

2.
```
   DB:  GRINDANDGRINDANDYETAGRIND
Query:  MYRINDGRINS
```