*Name:*________________________________

ID : _____________________________________

ECS 129: Structural Bioinformatics

March 18, 2024

Notes:

1) The final exam is open book, open notes.
2) The final is divided into 2 parts and graded over 80 points.
3) You can answer directly on these sheets (preferred), or on loose paper.
4) Please write your name at least on the front page!
5) Please, check your work! If possible, show your work when multiple steps are involved.

Part I (15 questions, each 4 points; total 60 points)
(These questions are multiple choices; in each case, find the most **plausible** answer)

1) Sickle cell anemia results from a single amino acid change in the human beta globin, from Glu to Val. The most probable corresponding mutation at the DNA level is:

a) Substitutions of 3 nucleotides
b) Insertion of 2 nucleotides
c) Deletion of 1 nucleotide
d) Substitution of 1 nucleotide
e) Insertion of 3 nucleotides

The codons for Glu are GAA and GAG, while those for Val are GUU, GUC, GUA, and GUG. Mutations of the second nucleotide A of GAA and GAG to a U will lead to GUA and GUG, both coding for Val.

2) In the dynamic programming matrix below, what is the score in the cell identified with an interrogation mark (?). Assume that the score for a perfect match is set to 10, the score of a mismatch is set to 0, and gap penalties are ignored.

	A	Y	F	W	G	G
A	10	0	0	0	0	0
Y	0	20	10	10	10	10
G	0	10	20	20	**30**	

a) 10
b) 20
c) 30
d) 40
e) 0

3) Which of the following statements on the
Needleman and Wunsch algorithm for pair-wise sequence alignment is most likely true?

a) The Needleman-Wunsch algorithm is based on dynamic programming and as such requires additive scores,
b) The Needleman-Wunsch algorithm is based on dynamic programming and as such requires multiplicative scores,
c) The Needleman-Wunsch algorithm generates a single optimal alignment,
d) The Needleman-Wunsch algorithm has a complexity of $O(N^5)$ and as such can only be used on very short sequences,
e) None of the above.

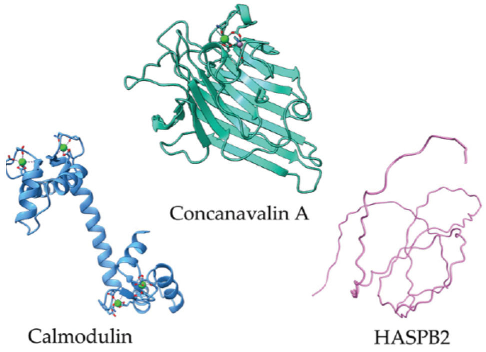
This is a consequence of the key idea of dynamic programming.

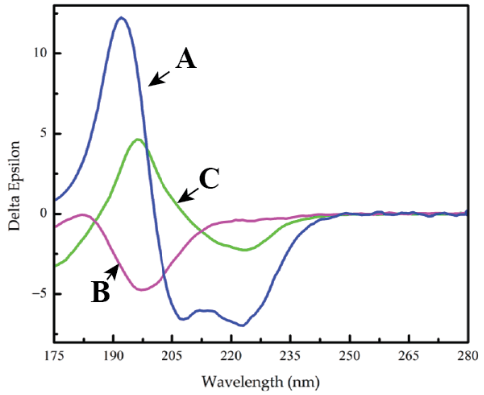
4) You are given a single strand of DNA. You are told that this sequence contains as many purines as pyrimidines, as many guanines as thymine, and that its complementary strand contains 10% more cytosine than thymine. How much cytosine (in percent) does it contain?

a) 10%,
b) 20%,
c) 30%,
d) 40%,
e) Not enough information

We know:
A+T+G+C = 100
A+G=C+T
G = T
cC = cT + 10 (where c means on the complement strand), i.e. G = A + 10
Solving this leads to C = 20, A = 20, G = 30, T = 30.

5) The figure below shows the CD spectra of 3 proteins on the same plot.





Identify A, B, and C:

a) A: HASPB2, B: Calmodulin, C: Concanavalin A
b) A: Concanavalin A, B: Calmodulin, C: HASPB2
c) A: HASPB2, B: Concanavalin A, C: Calmodulin
d) A: Calmodulin, B: HASPB2, C: Concanavalin A
e) A: Calmodulin, B: Concanavalin A, C: HASPB2

6) The cDNA corresponding to a small peptide is ATGTATGATCAATG**C**AGCGGCCTTTA TAG. The corresponding amino acid sequence is Met-Tyr-Asp-Glu-Cys-Ser-Gly-Pro-Leu. A mutation occurs at the DNA level, with the C at position 15 being substituted with T. What effect do you think this mutation might have on the expression of this gene?

a) It introduces a stop codon and the peptide will be shorter
b) The Cys in position 5 of the protein sequence will be replaced with Trp
c) The Start and Stop codons won't be in phase anymore and the gene won't be expressed
d) This is a silent mutation as it will have no impact on the protein sequence
e) None of the above

C 15 belongs to the codon TGC (or, in RNA, UGC) corresponding to Cys. If C 15 becomes T, the codon is changed to TGT, or in RNA, to UGU, which also codes for Cys. This is a “silent” mutation.

7) A protein sequence contains one ASP residue. You want to create a new protein sequence, with this ASP being replaced with a TRP. To do this, you first generate the DNA corresponding to the original protein (with your own choice for the codons you use), then mutate this DNA to get the sequence corresponding to the new protein. What is the minimum number of mutations needed?

a) 1
b) 2
c) 3
d) 0
e) None of the above

The codons for ASP are GAU and GAC, while the codon for TRP is UGG. In both options for ASP, you need to replace all nucleotides… therefore you need at least 3 mutations.

8) We want to find the best alignment(s) between the DNA sequences AGTATCT and AGATGC. The scoring scheme S is defined as follows: S(i,j) = 1 if i = j, and S(i,j) = 0 otherwise. There is a constant gap penalty of -1 (penalty for the first position counts; see table below). The score Sbest and the number N of optimal alignments are (show your final dynamic programming matrix and the best possible alignment (s) for full credit):

	A	T	T	A	T	T	C
A	**1**	-1	-1	0	-1	-1	-1
T	-1	**2**	**1**	0	1	1	0
A	0	0	2	**2**	1	1	1
T	-1	1	2	2	**3**	2	1
T	-1	1	1	2	3	**4**	2
C	-1	0	1	1	2	3	**5**

a) Sbest = 5, N = 2
b) Sbest = 3, N = 1
c) Sbest = 5, N = 1
d) Sbest = 3, N = 2
e) None of the above

The best score is 15 and there are two possible alignments:

```
ATTATTC
AT-ATTC
```

And

```
ATTATTC
A-TATTC
```

*Name:*__________________________________

ID : _______________________________________

9) Which combination of program / substitution matrix will most likely give you the best alignment between two sequences that are very similar?

a) BLAST / Blosum45
b) Dynamic programming / Blosum45
c) BLAST / Blosum90
d) Dynamic programming / Blosum90
e) BLAST / Blosum10

10) The best alignment found between the 2 DNA sequences TATATTC and ATCTC is:

```
TATATTC
-AT-CTC
```

The scoring scheme S is defined as follows: S(i,i) = 2X and S(i,j) = X if i and j are different. There is a constant gap penalty of -1. **Note that a gap at the beginning does not count**. The score of this alignment is 26. What is the value of X?

a) X = 2
b) X = 3
c) X = 4
d) X = 5
e) None of the above

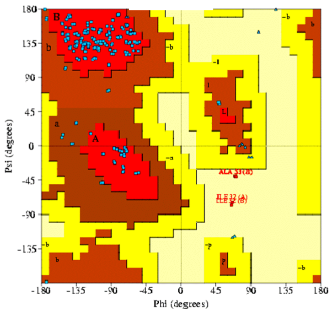
The score of the alignment is 0 (gap at the beginning) + 2X (A-A) + 2X (T-T) -1 (gap) + X (T-C) + 2X (T-T) + 2X (C-C)., i.e. 9X-1. As this is equal to 26, we get X = 3.

11) How many possible alignments, with no internal gaps, can you form when you compare a sequence of length 7 with a sequence of length 12? (Note that an alignment must have at least one letter match between the 2 sequences)

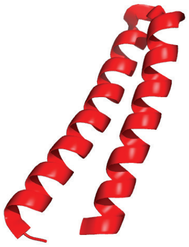
a) 7
b) 12
c) 19
d) 17
e) 18

The number of alignments is n+m-1 = 12+7-1 = 18

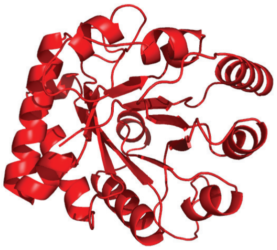
12) The Ramachandran plot of the protein structure 1bww is given below.



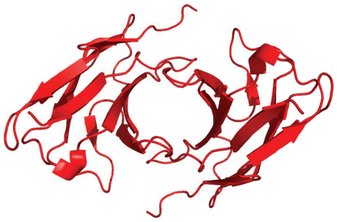
Which of the model of protein structures given below is most likely the corresponding structure



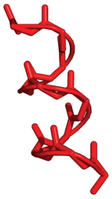
a)



b)



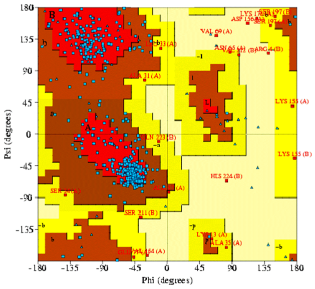
c)



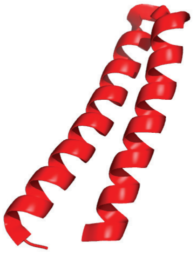
d)

The answer is c.

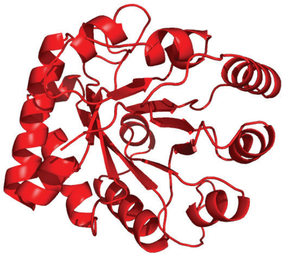
13) The Ramachandran plot of the protein structure 1tim is given below.



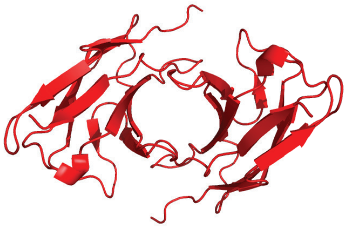
Which of the model of protein structures given below is most likely the corresponding structure:



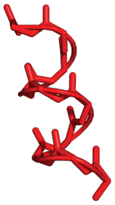
a)



b)



c)



d)

The answer is b.

14) The codon for Tryptophane is UGG. How many different amino acids (not including Tryptophane) could possibly result from substitutions of the first base, the second base, or both (the third base will always be G)?

a) 16
b) 13
c) 14
d) 12
e) Not enough information available

There are 16 codons that end with G. Those codons code for theoretically 16 amino acids, but one of them is the STOP codon, there are 2 repeats (Leu and Arg), and we do not count TRP, so the answer is 12.

15) Only one of those statements is correct when referring to the RMSD between two protein structures:

a) RMSD is computed from the dihedral angles of a structure
b) The larger the RMSD, the closer the two structures are
c) RMSD is equal to the score of the sequence alignment between two protein structures, squared
d) The expected RMSD between two experimental structures of the same protein in the same condition is expected to be below 1 Å
e) The expected RMSD between two experimental structures of the same protein in the same condition is expected to be above 5 Å

Part II (2 problems, total 20 points)

Problem 1 (10 points)

a) (5 points) Perform a global alignment of the two peptides AGPES and GCAET. The scoring scheme S is defined as follows: S(i,j) = 10 if i=j, and S(i,j) = 0 otherwise. There is a constant gap penalty of -1. **Note that a gap at the beginning counts**. After filling out the matrix, circle the traceback path(s) and write the optimal alignment(s). Note that if there are multiple traceback paths, write out all the optimal alignments.

	A	G	P	E	S
G	0	**9**	-1	-1	-1
C	-1	0	**9**	8	8
A	**9**	-1	**8**	9	8
E	-1	9	8	**18**	9
T	-1	8	9	8	**18**

There are 3 possible alignments:

```
AG-PES
-GCAET
```

```
AGP-ES
-GCAET
```

```
--AGPES
GCA--ET
```

All 3 with a score of 18.

b) (5 points) Perform the same global alignment, but now using the BLOSUM62 matrix provided in the appendix to define the score. There is again a constant gap penalty of -1. **Note that a gap at the beginning counts**. After filling the matrix, circle the traceback path(s) and write the optimal alignment(s). Note that if there are multiple traceback paths, write out all the optimal alignments. Explain possible differences with part a).

	A	G	P	E	S
G	0	**5**	-3	-3	-1
C	-1	-3	2	0	3
A	3	-1	**3**	1	2
E	-2	1	3	**8**	2
T	-2	3	5	3	**9**

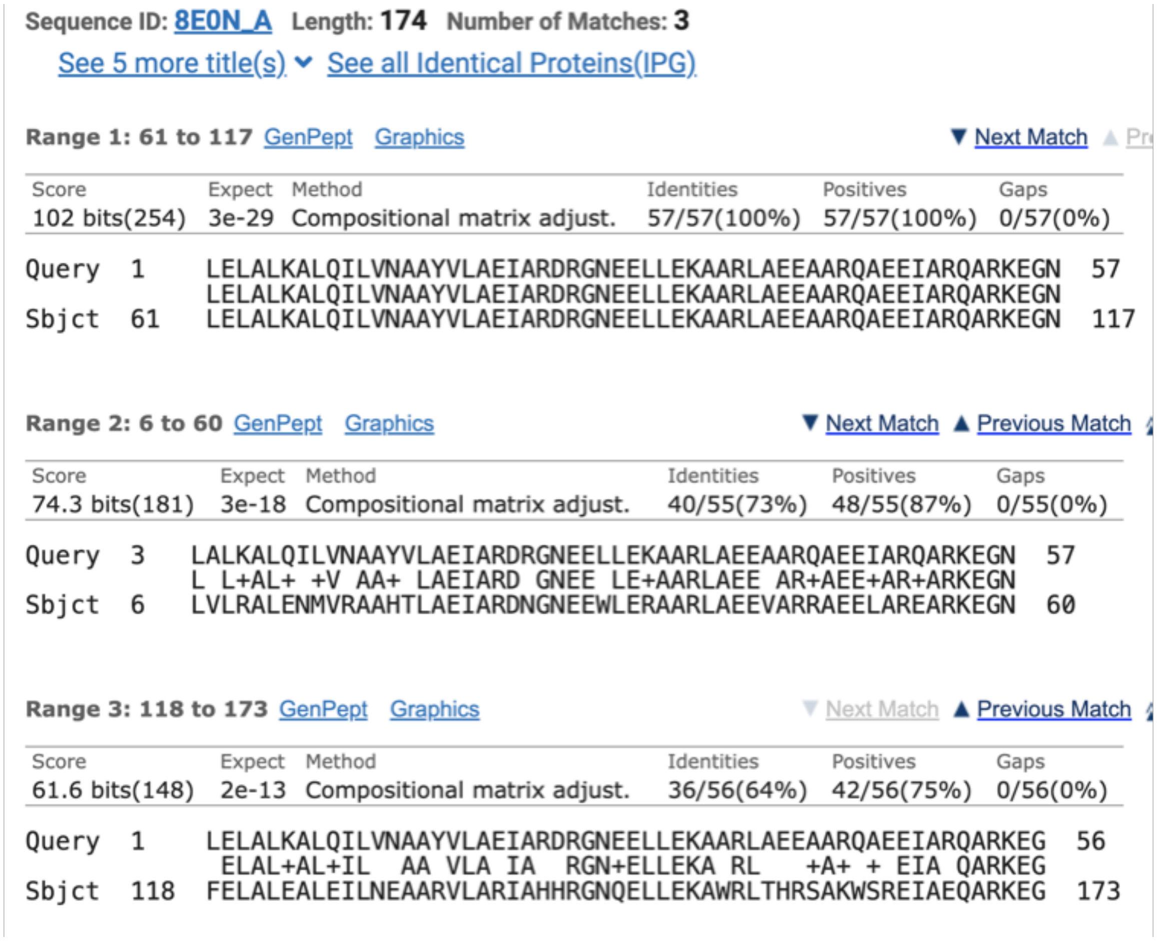
There is a unique optimal alignment of score 9:

```
AG-PES
-GCAET
```

BLOSUM62 is more specific and distinguishes between different mismatches.

Name:________________________________
ID : ______________________________________

Problem 2 (10 points)
A friend working in a biochemistry lab gave you a small sample. They told you that they know it contains a fragment of a protein, F, but they don't know more and would like you to investigate. You first find its sequence using Edman degradation: the fragment contains 57 amino acids. You decide then to check if the structure of this protein has already been studied. You use BLAST for this. Surprisingly, the top hit with BLAST shows three matches within the sequence of 8E0N, a protein that is found in anti-aging serum. Here are the corresponding BLAST results:

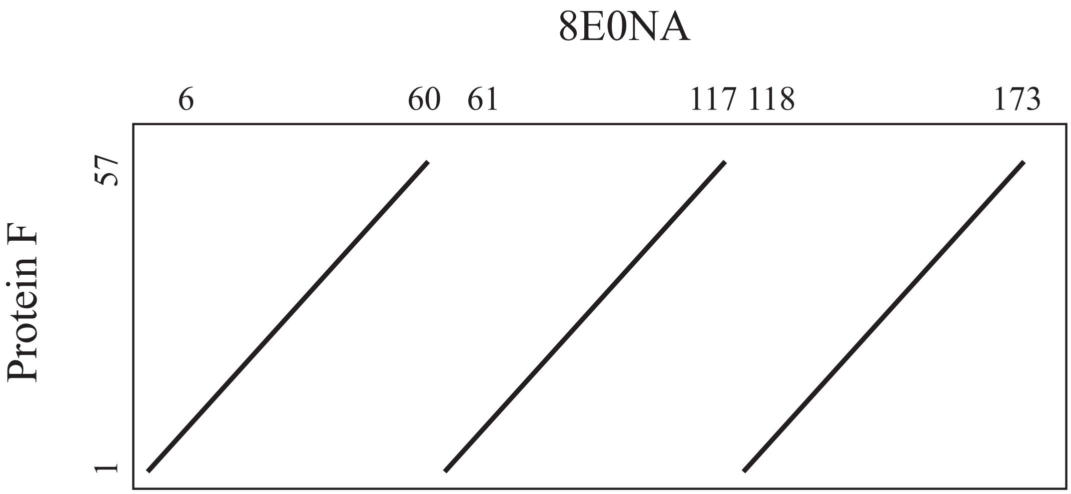


```
Sequence ID: 8E0N_A  Length: 174  Number of Matches: 3
   See 5 more title(s) ⌄  See all Identical Proteins(IPG)

Range 1: 61 to 117  GenPept  Graphics                              ▼ Next Match  ▲ Pr

Score          Expect  Method                         Identities     Positives      Gaps
102 bits(254)  3e-29   Compositional matrix adjust.   57/57(100%)    57/57(100%)    0/57(0%)

Query  1     LELALKALQILVNAAYVLAEIARDRGNEELLEKAARLAEEAARQAEEIARQARKEGN  57
             LELALKALQILVNAAYVLAEIARDRGNEELLEKAARLAEEAARQAEEIARQARKEGN
Sbjct  61    LELALKALQILVNAAYVLAEIARDRGNEELLEKAARLAEEAARQAEEIARQARKEGN  117


Range 2: 6 to 60  GenPept  Graphics                  ▼ Next Match  ▲ Previous Match

Score          Expect  Method                         Identities     Positives      Gaps
74.3 bits(181) 3e-18   Compositional matrix adjust.   40/55(73%)     48/55(87%)     0/55(0%)

Query  3    LALKALQILVNAAYVLAEIARDRGNEELLEKAARLAEEAARQAEEIARQARKEGN  57
            L L+AL+ +V AA+ LAEIARD GNEE LE+AARLAEE AR+AEE+AR+ARKEGN
Sbjct  6    LVLRALENMVRAAHTLAEIARDNGNEEWLERAARLAEEVARRAEELAREARKEGN  60


Range 3: 118 to 173  GenPept  Graphics              ▼ Next Match  ▲ Previous Match

Score          Expect  Method                         Identities     Positives      Gaps
61.6 bits(148) 2e-13   Compositional matrix adjust.   36/56(64%)     42/56(75%)     0/56(0%)

Query  1     LELALKALQILVNAAYVLAEIARDRGNEELLEKAARLAEEAARQAEEIARQARKEG  56
              ELAL+AL+IL  AA VLA IA  RGN+ELLEKA RL   +A+ + EIA QARKEG
Sbjct  118   FELALEALEILNEAARVLARIAHHRGNQELLEKAWRLTHRSAKWSREIAEQARKEG  173
```

a) (4 points) BLAST found three alignments. Are these alignments significant? Justify your answer

The three alignments have E-values of 3x10^{-29}, 3x10^{-18}, and 2x10^{-13} that are all significant.

b) (4 points) Based on these results from BLAST, draw schematically the dotplot between your protein and 8E0NA. Only show the major correspondences between the two sequences



c) (2 points) From these results, can you say anything about the structure of your protein, and the structure of 8E0NA ? Justify your answer.

While we cannot say anything about the specifics of the structure of our protein, 8E0NA seems to be a protein with 3 structural repeats, with each repeat equivalent to the structure of our protein.

*Name:*________________________________

ID : ______________________________________

*Name:*________________________________

ID : ____________________________________

Appendix A: Genetic Code

	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met/Start	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Appendix B: Blosum62 matrix

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9	-1	-1	-3	0	-3	-3	-3	-4	-3	-3	-3	-3	-1	-1	-1	-1	-2	-2	-2
S	-1	4	1	-1	1	0	1	0	0	0	-1	-1	0	-1	-2	-2	-2	-2	-2	-3
T	-1	1	4	1	-1	1	0	1	0	0	0	-1	0	-1	-2	-2	-2	-2	-2	-3
P	-3	-1	1	7	-1	-2	-1	-1	-1	-1	-2	-2	-1	-2	-3	-3	-2	-4	-3	-4
A	0	1	-1	-1	4	0	-1	-2	-1	-1	-2	-1	-1	-1	-1	-1	-2	-2	-2	-3
G	-3	0	1	-2	0	6	-2	-1	-2	-2	-2	-2	-2	-3	-4	-4	0	-3	-3	-2
N	-3	1	0	-2	-2	0	6	1	0	0	-1	0	0	-2	-3	-3	-3	-3	-2	-4
D	-3	0	1	-1	-2	-1	1	6	2	0	-1	-2	-1	-3	-3	-4	-3	-3	-3	-4
E	-4	0	0	-1	-1	-2	0	2	5	2	0	0	1	-2	-3	-3	-3	-3	-2	-3
Q	-3	0	0	-1	-1	-2	0	0	2	5	0	1	1	0	-3	-2	-2	-3	-1	-2
H	-3	-1	0	-2	-2	-2	1	1	0	0	8	0	-1	-2	-3	-3	-2	-1	2	-2
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5	2	-1	-3	-2	-3	-3	-2	-3
K	-3	0	0	-1	-1	-2	0	-1	1	1	-1	2	5	-1	-3	-2	-3	-3	-2	-3
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5	1	2	-2	0	-1	-1
I	-1	-2	-2	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4	2	1	0	-1	-3
L	-1	-2	-2	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4	3	0	-1	-2
V	-1	-2	-2	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4	-1	-1	-3
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6	3	1
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	2
W	-2	-3	-3	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11