3 Extension of global alignment

3.1 Introduction of score matrix

We will expand our simple scoring scheme to score matrices. This expansion allows us to solve general alignment problems with DNA, RNA, and protein sequences.

Extension of a scoring scheme to a score matrix

The matrix below is equivalent with match: 1 and mismatch: 0.

	a	b
a	1	0
b	0	1

Example of a DNA score matrix

The matrix below is equivalent with match: 5 and mismatch: -4.

	A	T	G	C
A	5	-4	-4	-4
T		5	-4	-4
G			5	-4
C				5

Applications of score matrix

Score matrices are more flexible than the simple scoring scheme. For instance, they can be used for the following cases.

- DNA pairs
- RNA pairs
- Similarity of protein sequences by amino acid properties

DNA pairs (Watson-Crick pairs)

A thymine pairs with an adenine, and a cytosine pairs with a guanine.

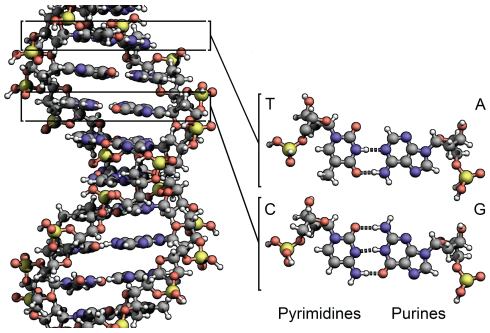

Figure 3.1: Watson-Crick pairs (source: Zephyris, Wikimedia Commons)

Example of score matrix for DNA pairs
The matrix reflects the difference of hydrogen bonds.

	A	T	G	C
A	5	-4	-4	-4
T		5	-4	-4
G			5	-4
C				5

Example of DP for DNA pairs
You can use DP to find a DNA alignment with Watson-Crick pairs. For instance, the DP table below is used to solve the optimal alignment for two DNA sequences: q = AC and d = GT with gap penalty g = 4.

DP table:

q/d		G	T
	0	-4	-8
A	-4	-3	0
C	-8	1	-3

Alignment:

```
q: AC-
d: -GT
```

RNA pairs

A single stand of RNA can form a 3D structure that has a biological function. The secondary structure of RNA is a two-dimensional representation of the structure.

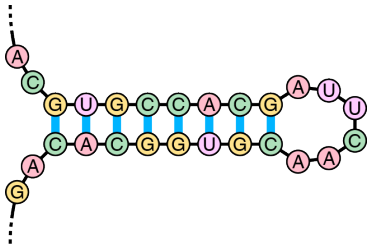

Figure 3.2: RNA stem-loop (source: Sakurambo, Wikimedia Commons)

Wobble pairs
Wobble pairs are not canonical Watson-Crick pairs, but they can still form hydrogen bonds.

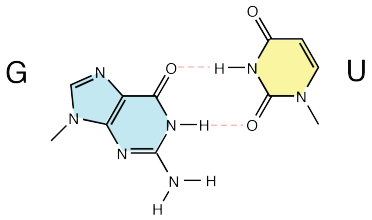

Figure 3.3: GU wobble pairs
(modifed from the original version by Fdardel, Wikimedia Commons)

Example of DP for RNA pairs

You can form the following DP table for two RNA sequences: q = AU and d = UGA with gap penalty g = 9.

DP table:

q/d		U	G	A
	0	-9	-18	-27
A	-9	5	-4	-11
U	-18	-4	7	1

Alignment:

```
q: A-U
d: UGA
```

Similarity of protein sequences

Amino acids can be categorized into several groups by their properties. Proteins alignments often need to take these properties into consideration.

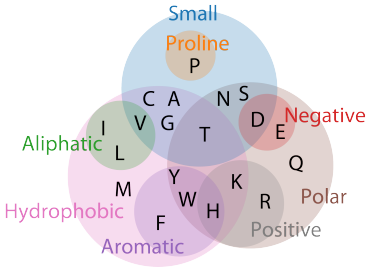

Figure 3.4: Venn diagram of amino acid properties

Example of a protein score matrix

It can be used to compare the similarity between two protein sequences.

	A	**R**	**N**	**D**	**C**	**Q**	**E**	**G**	**H**	**I**	**L**	**K**	**M**	**F**	**P**	**S**	**T**	**W**	**Y**	**V**
A	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
R	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
N	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
D	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
C	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
Q	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3
E	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
G	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
H	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
I	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
L	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
K	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
M	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
F	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
P	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
S	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
T	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
W	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
Y	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
V	7	4	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	72	4	17

Table 3.1: Mutation probability matrix for the evolutionary distance of 250 PAMs (in percentage) (Chapter 22: A model of evolutionary change in proteins, Dayhoff and Schwartz, Atlas of Protein Sequence and Structure, 1978)

Exercise 3.1

1. Use the DNA score matrix below with g = 10 and find the optimal alignment for q = TG and d = TCG.

	A	T	G	C
A	5	-4	-4	-4
T		5	-4	-4
G			5	-4
C				5

2. 250 PAM mutation matrix can not directly be used for global alignments. Explain what kind of matrix you need for calcualting alignment scores.