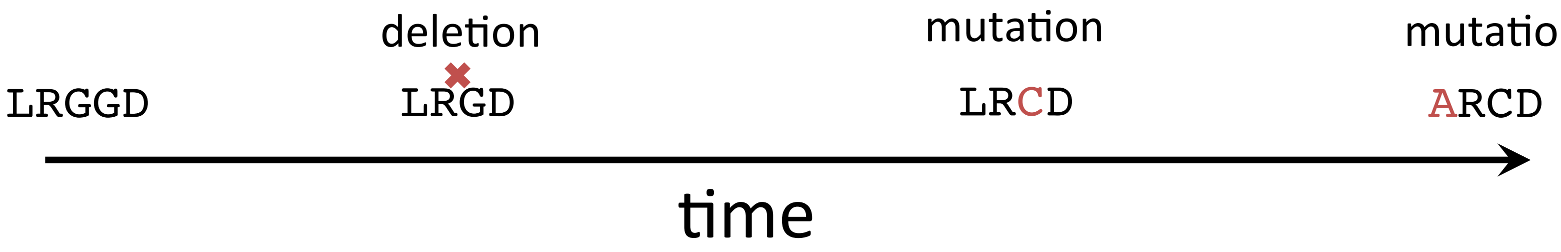
Sequence alignments

Genetic sequences change over time



deletion

mutation

mutation

LRGGD

LRGD

LRCD

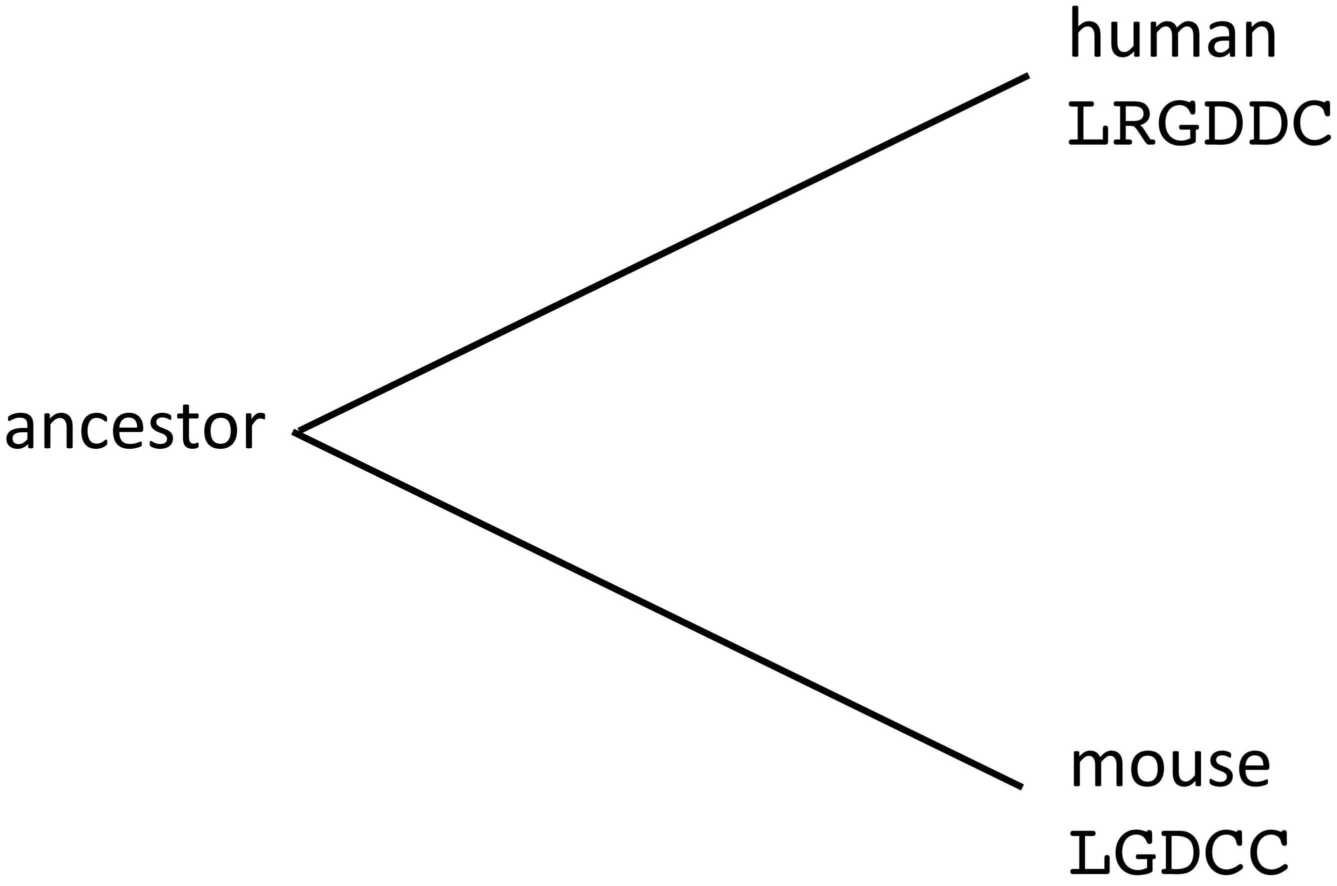
ARCD

time

Relationship between original and final sequence:

```
LRGGD          LRGGD
AR-CD    or    ARC-D
```

In practice: we only know sequences from extant organisms



ancestor

human

LRGDDC

mouse

LGDCC

We need to align these sequences to compare them

human	mouse
LRGDDC	LGDCC

```
LRGDDC     LRGDDC-     LRGDDC
L-GDCC     L-GD-CC     -LGDCC
```

Which alignment is correct?

We need to score the alignment

Example:

- match = +1
- mismatch = -1
- gap = 0

```
LRGDDC      score = 1+0+1+1-1+1
L-GDCC            = 3
```

```
LRGDDC-     score = 1+0+1+1+0+1+0
L-GD-CC           = 4
```

```
LRGDDC      score = 0-1+1+1-1+1
-LGDCC            = 1
```

We need to score the alignment

Example:

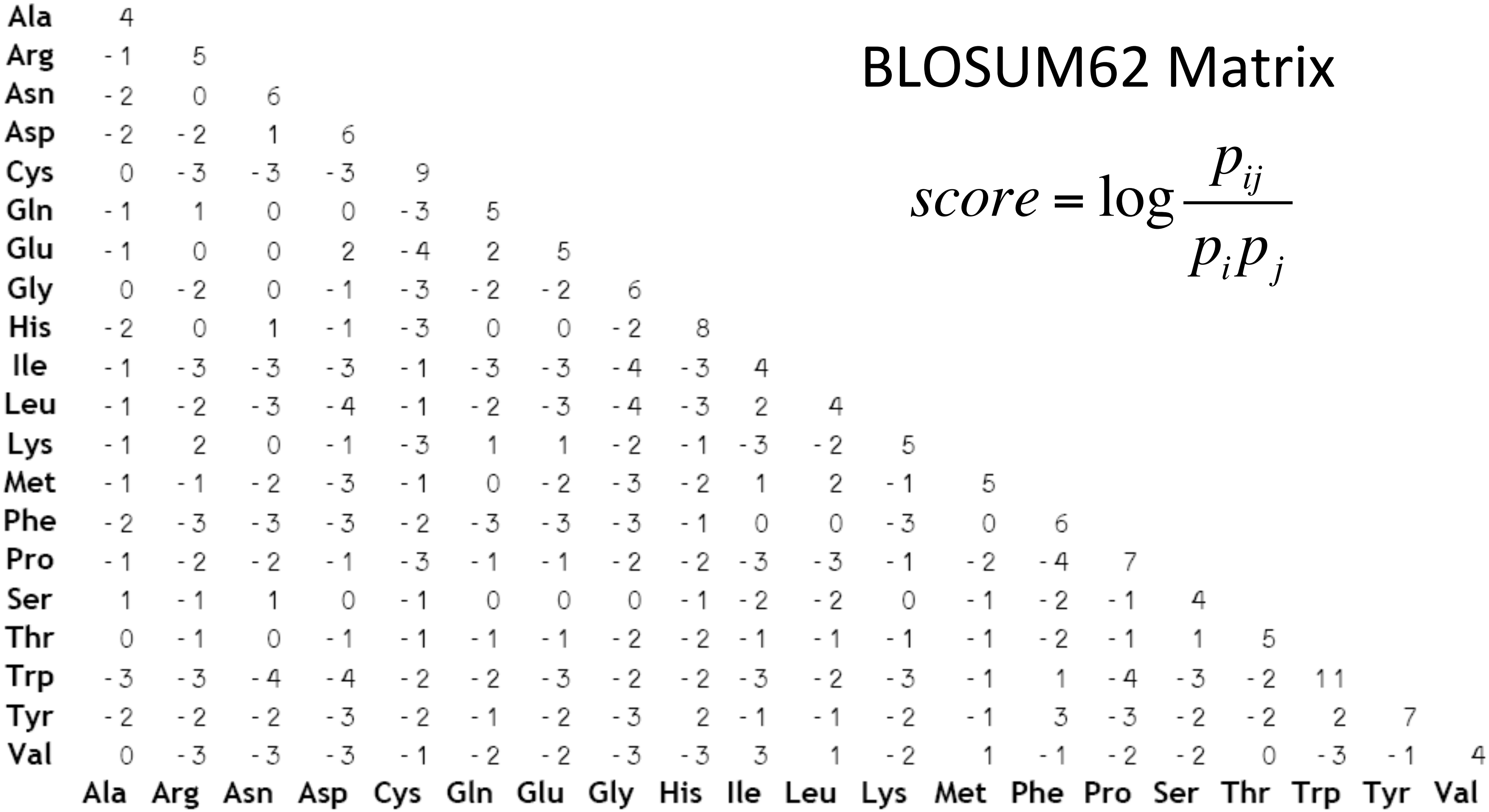
- match = +1
- mismatch = -1
- gap = -2

```
LRGDDC     score = 1-2+1+1-1+1
L-GDCC           = 1
```

```
LRGDDC-    score = 1-2+1+1-2+1-2
L-GD-CC          = -2
```

```
LRGDDC     score = -2-1+1+1-1+1
-LGDCC           = -1
```

We often score by amino-acid similarity



Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4
	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val

BLOSUM62 Matrix

$$score = \log \frac{p_{ij}}{p_i p_j}$$

http://commons.wikimedia.org/wiki/File:BLOSUM62.gif

Gaps in alignments are called “indels”

```
LRGDDC
L-GDCC
```

indel

Can you guess why?

How do we find the best alignment given a scoring system?

Global alignment: Needleman-Wunsch algorithm

Example: align GCAT and GAT

Scoring: match = 1, mismatch = -1, gap = -1

	-	G	C	A	T
-					
G					
A					
T					

How do we find the best alignment given a scoring system?

Global alignment: Needleman-Wunsch algorithm

Example: align `GCAT` and `GAT`

Scoring: match = 1, mismatch = -1, gap = -1

	-	G	C	A	T
-	0				
G					
A					
T					

Alignment:

```
-
-
```

How do we find the best alignment given a scoring system?

Global alignment: Needleman-Wunsch algorithm

Example: align GCAT and GAT

Scoring: match = 1, mismatch = -1, gap = -1

	-	G	C	A	T
-	0	-1			
G					
A					
T					

Alignment:

```
-G
--
```

How do we find the best alignment given a scoring system?

Global alignment: Needleman-Wunsch algorithm

Example: align GCAT and GAT

Scoring: match = 1, mismatch = -1, gap = -1

	–	G	C	A	T
–	0	-1	-2		
G					
A					
T					

Alignment:

```
–GC
–––
```

How do we find the best alignment given a scoring system?

Global alignment: Needleman-Wunsch algorithm

Example: align GCAT and GAT

Scoring: match = 1, mismatch = -1, gap = -1

	–	G	C	A	T
–	0	-1	-2	-3	-4
G					
A					
T					

Alignment:
```
–GCAT
-----
```

How do we find the best alignment given a scoring system?

Global alignment: Needleman-Wunsch algorithm

Example: align `GCAT` and `GAT`

Scoring: match = 1, mismatch = -1, gap = -1

	–	G	C	A	T
–	0	-1	-2	-3	-4
G	-1				
A					
T					

Alignment:

––
–G

How do we find the best alignment given a scoring system?

Global alignment: Needleman-Wunsch algorithm

Example: align GCAT and GAT

Scoring: match = 1, mismatch = -1, gap = -1

	–	G	C	A	T
–	0	-1	-2	-3	-4
G	-1				
A	-2				
T	-3				

Alignment:

```
----
-GAT
```

How do we find the best alignment given a scoring system?

Global alignment: Needleman-Wunsch algorithm

Example: align GCAT and GAT

Scoring: match = 1, mismatch = -1, gap = -1

	–	G	C	A	T
–	0	-1	-2	-3	-4
G	-1	?			
A	-2				
T	-3				

How do we find the best alignment given a scoring system?

Global alignment: Needleman-Wunsch algorithm

Example: align GCAT and GAT

Scoring: match = 1, mismatch = -1, gap = -1

	–	G	C	A	T
–	0	-1	-2	-3	-4
G	-1	↑ -2			
A	-2				
T	-3				

Alignment:

–G–

––G

How do we find the best alignment given a scoring system?

Global alignment: Needleman-Wunsch algorithm

Example: align GCAT and GAT

Scoring: match = 1, mismatch = -1, gap = -1

	–	G	C	A	T
–	0	-1	-2	-3	-4
G	-1	← -2			
A	-2				
T	-3				

Alignment:

––G
–G–

How do we find the best alignment given a scoring system?

Global alignment: Needleman-Wunsch algorithm

Example: align `GCAT` and `GAT`

Scoring: match = 1, mismatch = -1, gap = -1

	–	G	C	A	T
–	0	-1	-2	-3	-4
G	-1	↖ 1			
A	-2				
T	-3				

Alignment:

```
–G
–G
```

How do we find the best alignment given a scoring system?

Global alignment: Needleman-Wunsch algorithm

Example: align GCAT and GAT

Scoring: match = 1, mismatch = -1, gap = -1

	–	G	C	A	T
–	0	-1	-2	-3	-4
G	-1	↖ 1	← 0		
A	-2				
T	-3				

Alignment:

–GC
–G–

How do we find the best alignment given a scoring system?

Global alignment: Needleman-Wunsch algorithm

Example: align GCAT and GAT

Scoring: match = 1, mismatch = -1, gap = -1

	–	G	C	A	T
–	0	-1	-2	-3	-4
G	-1	1	0		
A	-2	0			
T	-3				

Alignment:

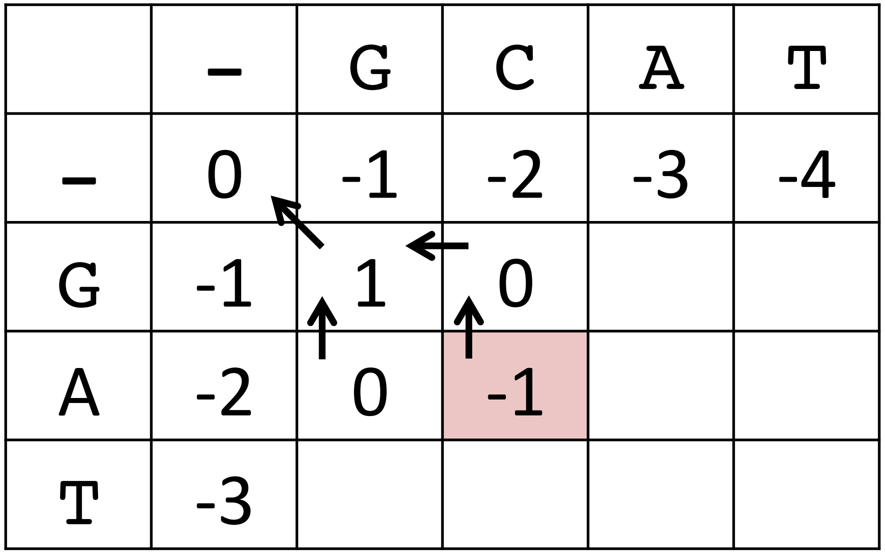
–G–
–GA

How do we find the best alignment given a scoring system?

Global alignment: Needleman-Wunsch algorithm

Example: align GCAT and GAT

Scoring: match = 1, mismatch = -1, gap = -1



	–	G	C	A	T
–	0	-1	-2	-3	-4
G	-1	1	0		
A	-2	0	-1		
T	-3				

Alignment:

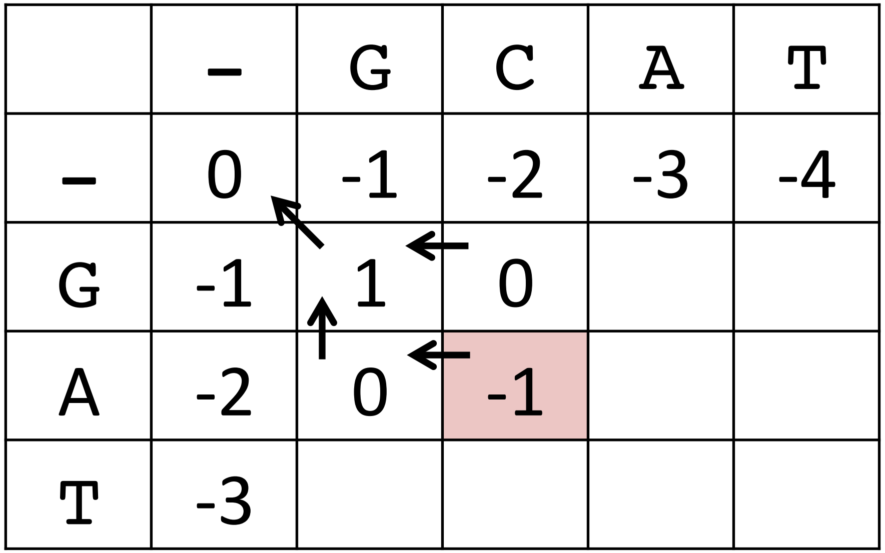
–GC–
–G–A

How do we find the best alignment given a scoring system?

Global alignment: Needleman-Wunsch algorithm

Example: align GCAT and GAT

Scoring: match = 1, mismatch = -1, gap = -1



	–	G	C	A	T
–	0	-1	-2	-3	-4
G	-1	1	0		
A	-2	0	-1		
T	-3				

Alignment:

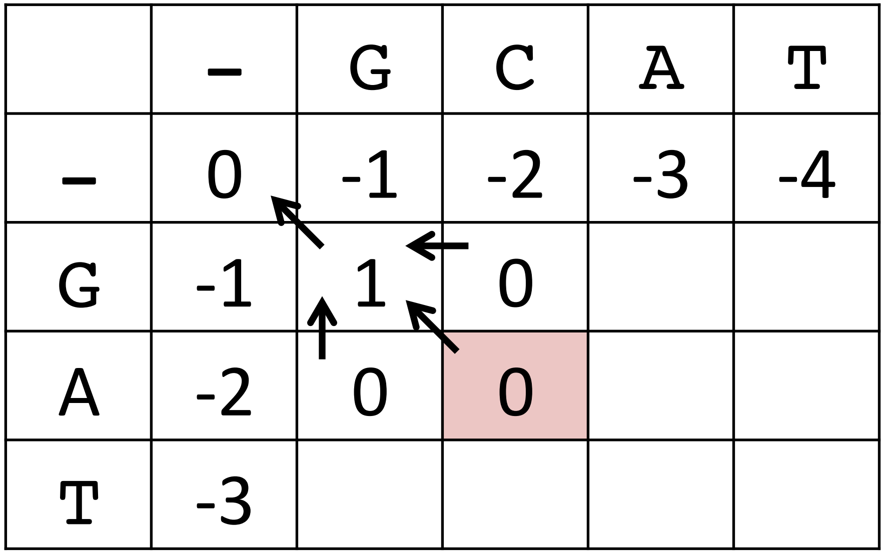
```
–G–C
–GA–
```

How do we find the best alignment given a scoring system?

Global alignment: Needleman-Wunsch algorithm

Example: align GCAT and GAT

Scoring: match = 1, mismatch = -1, gap = -1



	–	G	C	A	T
–	0	-1	-2	-3	-4
G	-1	1	0		
A	-2	0	0		
T	-3				

Alignment:
–GC
–GA

How do we find the best alignment given a scoring system?

Global alignment: Needleman-Wunsch algorithm

Example: align GCAT and GAT

Scoring: match = 1, mismatch = -1, gap = -1

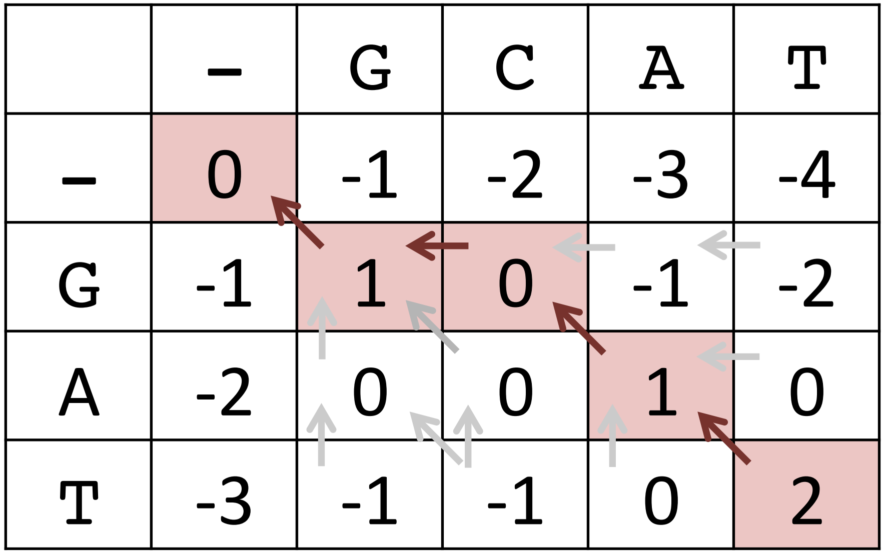
	–	G	C	A	T
–	0	-1	-2	-3	-4
G	-1	↖ 1	← 0	← -1	← -2
A	-2	↑ 0	↖ 0	↖ 1	← 0
T	-3	↑ -1	↖ ↑ -1	↑ 0	↖ 2

How do we find the best alignment given a scoring system?

Global alignment: Needleman-Wunsch algorithm

Example: align GCAT and GAT

Scoring: match = 1, mismatch = -1, gap = -1



	–	G	C	A	T
–	0	-1	-2	-3	-4
G	-1	1	0	-1	-2
A	-2	0	0	1	0
T	-3	-1	-1	0	2

Alignment:

–GCAT
–G–AT

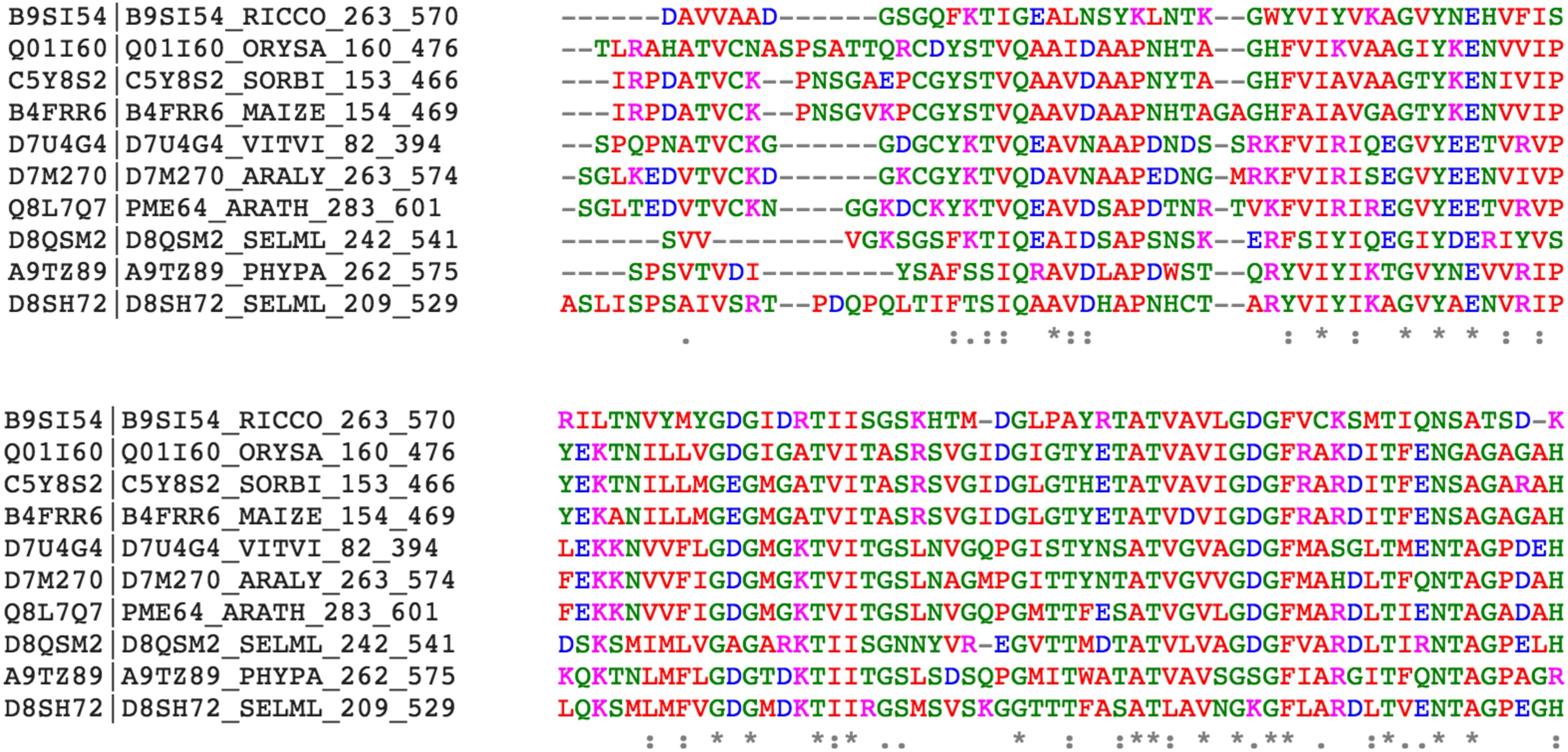
Now try on your own

Align ATGCT and ATTACA

Scoring: match = 1, mismatch = -1, gap = -1

	-	A	T	T	A	C	A
-							
A							
T							
G							
C							
T							

Multiple sequence alignment (MSA)



```
B9SI54|B9SI54_RICCO_263_570     ------DAVVAAD------GSGQFKTIGEALNSYKLNTK--GWYVIYVKAGVYNEHVFIS
Q01I60|Q01I60_ORYSA_160_476     --TLRAHATVCNASPSATTQRCDYSTVQAAIDAAPNHTA--GHFVIKVAAGIYKENVVIP
C5Y8S2|C5Y8S2_SORBI_153_466     ---IRPDATVCK--PNSGAEPCGYSTVQAAVDAAPNYTA--GHFVIAVAAGTYKENIVIP
B4FRR6|B4FRR6_MAIZE_154_469     ---IRPDATVCK--PNSGVKPCGYSTVQAAVDAAPNHTAGAGHFAIAVGAGTYKENVVIP
D7U4G4|D7U4G4_VITVI_82_394      --SPQPNATVCKG------GDGCYKTVQEAVNAAPDNDS-SRKFVIRIQEGVYEETVRVP
D7M270|D7M270_ARALY_263_574     -SGLKEDVTVCKD------GKCGYKTVQDAVNAAPEDNG-MRKFVIRISEGVYEENVIVP
Q8L7Q7|PME64_ARATH_283_601      -SGLTEDVTVCKN----GGKDCKYKTVQEAVDSAPDTNR-TVKFVIRIREGVYEETVRVP
D8QSM2|D8QSM2_SELML_242_541     ------SVV--------VGKSGSFKTIQEAIDSAPSNSK--ERFSIYIQEGIYDERIYVS
A9TZ89|A9TZ89_PHYPA_262_575     ----SPSVTVDI--------YSAFSSIQRAVDLAPDWST--QRYVIYIKTGVYNEVVRIP
D8SH72|D8SH72_SELML_209_529     ASLISPSAIVSRT--PDQPQLTIFTSIQAAVDHAPNHCT--ARYVIYIKAGVYAENVRIP
                                       .               :.::  *::           : * :  * * * : :

B9SI54|B9SI54_RICCO_263_570     RILTNVYMYGDGIDRTIISGSKHTM-DGLPAYRTATVAVLGDGFVCKSMTIQNSATSD-K
Q01I60|Q01I60_ORYSA_160_476     YEKTNILLVGDGIGATVITASRSVGIDGIGTYETATVAVIGDGFRAKDITFENGAGAGAH
C5Y8S2|C5Y8S2_SORBI_153_466     YEKTNILLMGEGMGATVITASRSVGIDGLGTHETATVAVIGDGFRARDITFENSAGARAH
B4FRR6|B4FRR6_MAIZE_154_469     YEKANILLMGEGMGATVITASRSVGIDGLGTYETATVDVIGDGFRARDITFENSAGAGAH
D7U4G4|D7U4G4_VITVI_82_394      LEKKNVVFLGDGMGKTVITGSLNVGQPGISTYNSATVGVAGDGFMASGLTMENTAGPDEH
D7M270|D7M270_ARALY_263_574     FEKKNVVFIGDGMGKTVITGSLNAGMPGITTYNTATVGVVGDGFMAHDLTFQNTAGPDAH
Q8L7Q7|PME64_ARATH_283_601      FEKKNVVFIGDGMGKTVITGSLNVGQPGMTTFESATVGVLGDGFMARDLTIENTAGADAH
D8QSM2|D8QSM2_SELML_242_541     DSKSMIMLVGAGARKTIISGNNYVR-EGVTTMDTATVLVAGDGFVARDLTIRNTAGPELH
A9TZ89|A9TZ89_PHYPA_262_575     KQKTNLMFLGDGTDKTIITGSLSDSQPGMITWATATVAVSGSGFIARGITFQNTAGPAGR
D8SH72|D8SH72_SELML_209_529     LQKSMLMFVGDGMDKTIIRGSMSVSKGGTTTFASATLAVNGKGFLARDLTVENTAGPEGH
                                     : : * *   *:* ..      *  :  :**: * *.** .  :*..* *    :
```

Software to generate MSAs

- MAFFT
  (very good, very fast)
  http://mafft.cbrc.jp/alignment/software/
- Clustal Omega
  (very good, very fast)
  http://www.ebi.ac.uk/Tools/msa/clustalo/
- PRANK
  (extremely good, very slow)
  http://wasabiapp.org/software/prank/