Genome 559:

Introduction to Statistical and Computational Genomics

Professors Jim Thomas and Elhanan Borenstein

Logistics

- Syllabus and web site:

  http://faculty.washington.edu/jht/GS559_2012/

- Should I take this class?
- Grading
- Send homework by email ATTACHMENT.

Homework format

Attach your answers as a simple text file (NOT Word or HTML etc). I may need to run your programs, so the formatting has to be correct (especially tabs). If you need figures, attach them separately or hand them in on paper in class.

Name your email attached file as follows:

`GS559_MichelleObama_PS1.txt`
`GS559_MichelleObama_PS2.txt`
etc.

Please stick with this format exactly - it makes it a lot easier for my bookkeeping.

If you are unsure whether your Python format is correct in what you send, use copy and paste to save the code in a new file and be sure that the new file runs as a Python program.

Class time structure

Roughly split into thirds:

First, bioinformatic topics
Second, Python topics
Third, in class Python exercises

Sequence comparison: Introduction and motivation

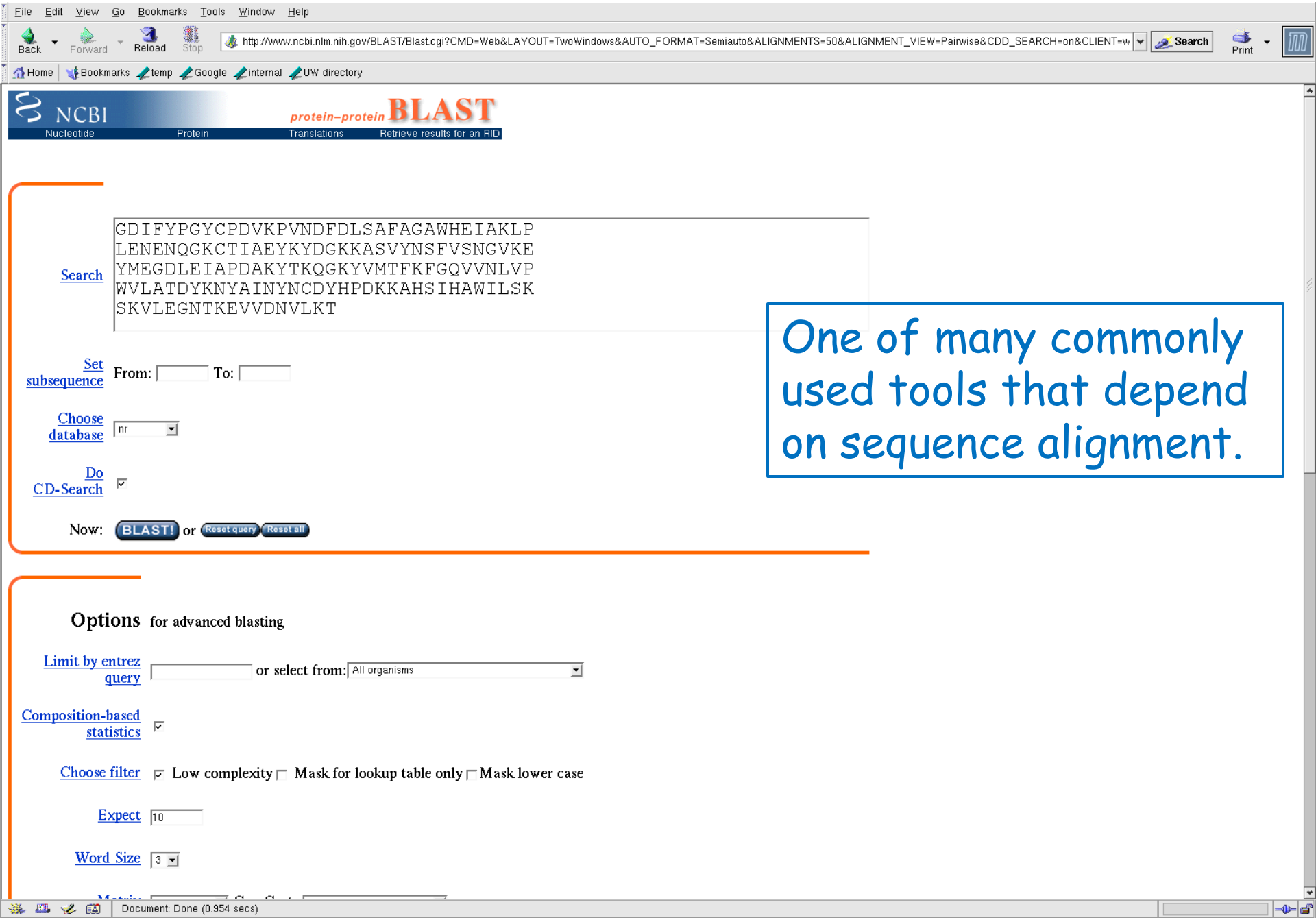
Prof. James H. Thomas

Motivation

- Why align two protein or DNA sequences?

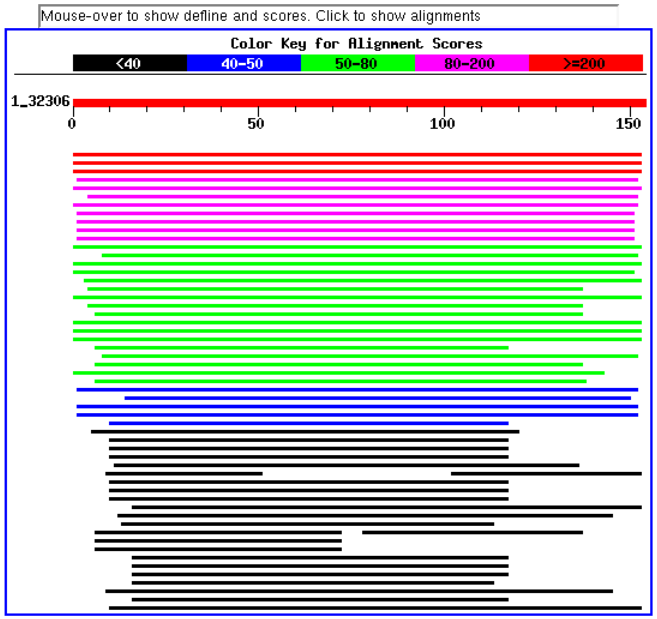
Motivation

- Why align two protein or DNA sequences?
  - Determine whether they are descended from a common ancestor (homologous).
  - Infer a common function.
  - Locate functional elements (motifs or domains).
  - Infer protein or RNA structure, if the structure of one of the sequences is known.
  - Analyze sequence evolution



One of many commonly used tools that depend on sequence alignment.

Distribution of 100 Blast Hits on the Query Sequence

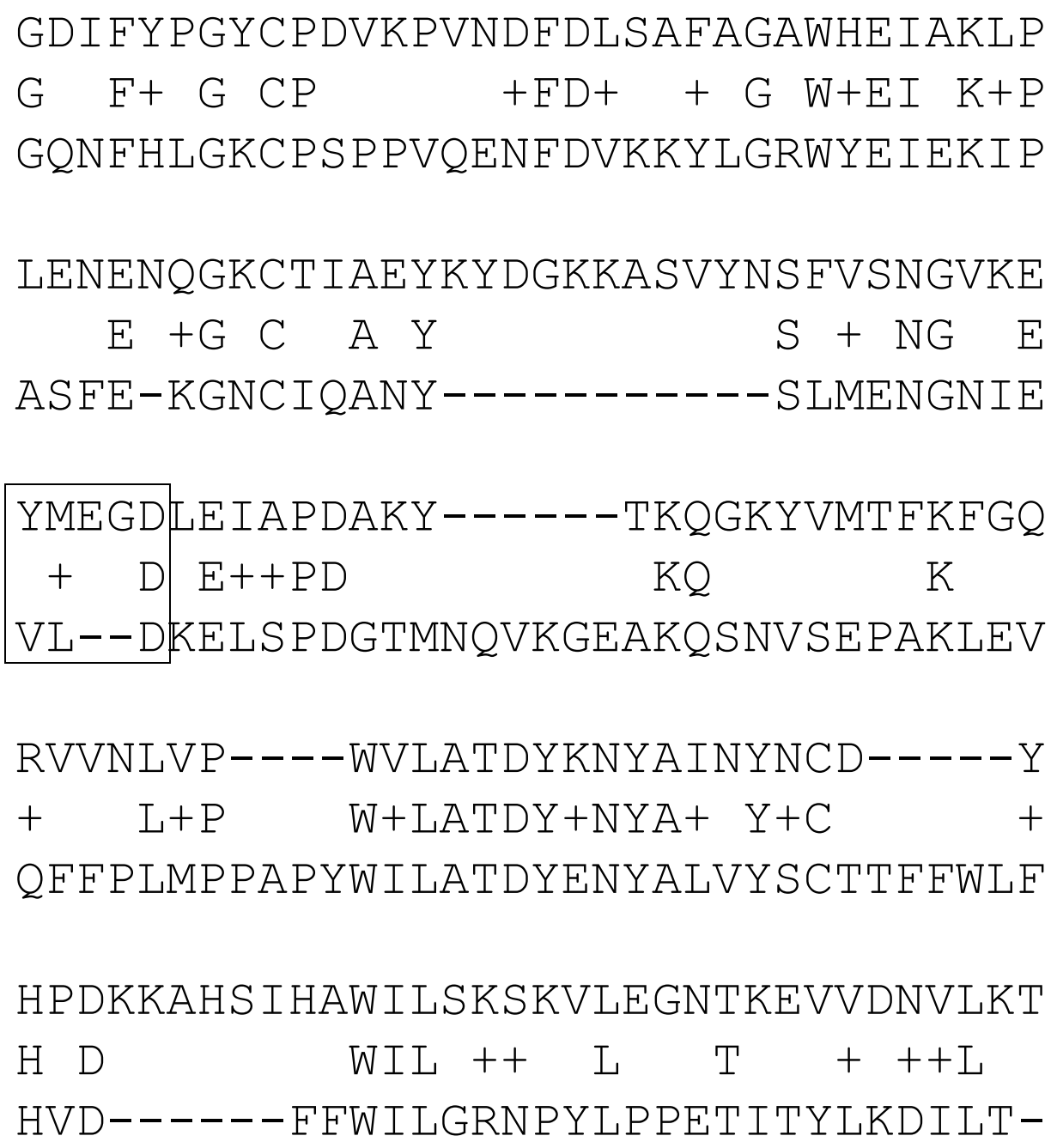


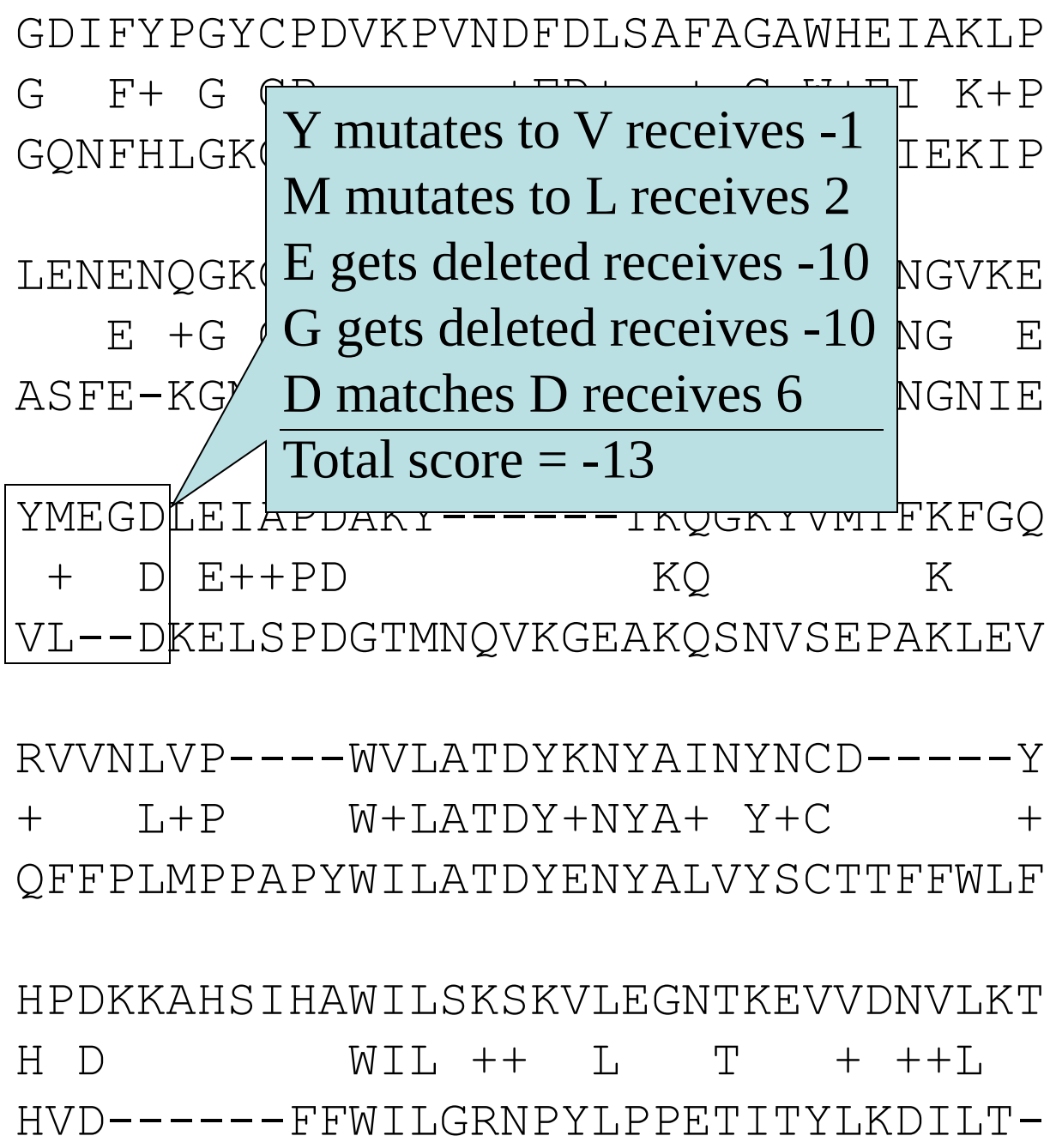
Related Structures

Sequences producing significant alignments:	Score (bits)	E Value	
gi|124151|sp|P00305|ICYA_MANSE Insecticyanin A form (Blue b...	304	4e-82	
gi|124527|sp|Q00630|ICYB_MANSE Insecticyanin B form precurs...	301	2e-81	
gi|102968|pir||S22400 insecticyanin A - tobacco hornworm >g...	287	4e-77	
gi|13928531|dbj|BAB47155.1| Bombyrin [Bombyx mori]	144	7e-34	
gi|18857921|dbj|BAB85482.1| biliverdin binding protein-I [S...	142	2e-33	
gi|1146408|gb|AAA85089.1| gallerin	132	3e-30	
gi|18642498|dbj|BAB84676.1| biliverdin binding protein-II [...	115	3e-25	
gi|34810780|pdb|1N0S|A Chain A, Engineered Lipocalin Flua I...	107	7e-23	S
gi|1705433|sp|P09464|BBP_PIEBR Bilin-binding protein precur...	104	6e-22	
gi|229695|pdb|1BBP|A Chain A, Bilin Binding Protein (BBP) >...	103	7e-22	S
gi|33357253|pdb|1KYQ|A Chain A, Engineered Lipocalin Pica16	97	1e-19	S

Sequence comparison overview

- Problem: Find the "best" alignment between two sequences.
- To solve this problem, we need:
  - a method for scoring alignments
  - an algorithm for finding the alignment with the best score
- The alignment score is calculated using:
  - a substitution matrix
  - gap penalties
- The main algorithm for finding the best alignment is dynamic programming.


GDIFYPGYCPDVKPVNDFDLSAFAGAWHEIAKLP
G F+ G CP +FD+ + G W+EI K+P
GQNFHLGKCPSPPVQENFDVKKYLGRWYEIEKIP
LENENQGKCTIAEYKYDGKKASVYNSFVSNGVKE
E +G C A Y S + NG E
ASFE-KGNCIQANY----------SLMENGNIE
YMEGDLEIAPDAKY------TKQGKYVMTFKFGQ
+ D E++PD KQ K
VL--DKELSPDGTMNQVKGEAKQSNVSEPAKLEV
RVVNLVP----WVLATDYKNYAINYNCD-----Y
+ L+P W+LATDY+NYA+ Y+C +
QFFPLMPPAPYWILATDYENYALVYSCTTFFWLF
HPDKKAHSIHAWILSKSKVLEGNTKEVVDNVLKT
H D WIL ++ L T + ++L
HVD------FFWILGRNPYLPPETITYLKDILT-


GDIFYPGYCPDVKPVNDFDLSAFAGAWHEIAKLP
G F+ G
GQNFHLGK
IEKIP
LENENQGK
NGVKE
E +G
NG E
ASFE-KG
NGNIE
YMEGDLEIAPDAKY------IKQGKYVMIFKFGQ
+ D E++PD KQ K
VL--DKELSPDGTMNQVKGEAKQSNVSEPAKLEV
RVVNLVP----WVLATDYKNYAINYNCD-----Y
+ L+P W+LATDY+NYA+ Y+C +
QFFPLMPPAPYWILATDYENYALVYSCTTFFWLF
HPDKKAHSIHAWILSKSKVLEGNTKEVVDNVLKT
H D WIL ++ L T + ++L
HVD------FFWILGRNPYLPPETITYLKDILT-
Y mutates to V receives -1
M mutates to L receives 2
E gets deleted receives -10
G gets deleted receives -10
D matches D receives 6
Total score = -13

A simple alignment problem.

- Problem: find the best pairwise alignment of GAATC and CATAC.

Scoring alignments

```
GAATC      GAAT-C     -GAAT-C
CATAC      C-ATAC     C-A-TAC

GAATC-     GAAT-C     GA-ATC
CA-TAC     CA-TAC     CATA-C
```

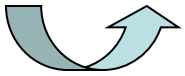
(some of a very large number of possibilities)

- We need a way to measure the quality of a candidate alignment.
- Alignment scores consist of: a substitution matrix (aka score matrix) and a gap penalty.

Scoring aligned bases

Purine	A	G
Pyrimidine	C	T

Transversion change
(very low score)



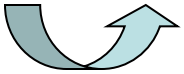
Transition change
(low score)

Transitions are typically about 2x as frequent as transversions in real sequences.

Scoring aligned bases

Purine	A	G
Pyrimidine	C	T

Transversion



Transition

```
GAATC
CATAC
```

-5 + 10 + -5 + -5 + 10 = 5

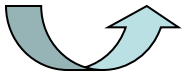
A reasonable substitution matrix:

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Scoring aligned bases

Purine	A	G
Pyrimidine	C	T

Transversion
(expensive)



Transition
(cheap)

```
GAAT-C
CA-TAC
```

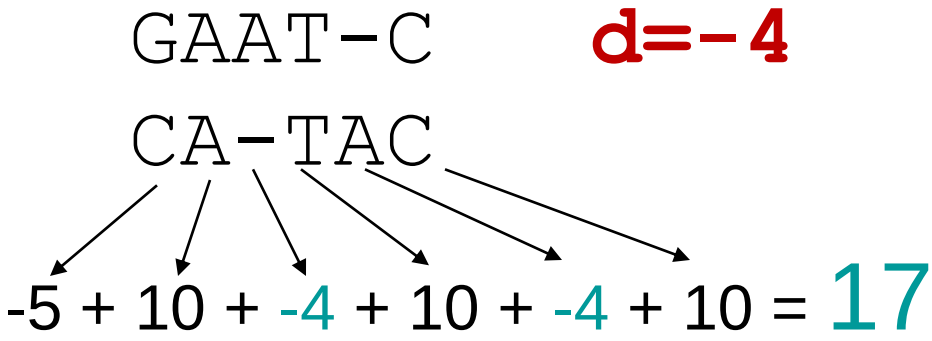
-5 + 10 + ? + 10 + ? + 10 = ?

A reasonable substitution matrix:

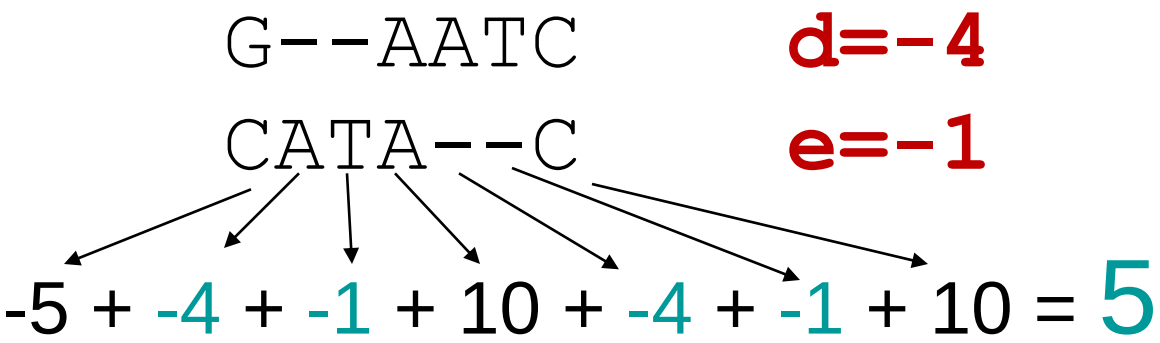
	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Scoring gaps

- Linear gap penalty: every gap receives a score of d:



- Affine gap penalty: opening a gap receives a score of d; extending a gap receives a score of e:



You should be able to ...

- Explain why sequence comparison is useful.
- Define *substitution matrix* and different types of *gap penalties*.
- Compute the score of an alignment, given a substitution matrix and gap penalties.

BLOSUM 62 (amino acid score matrix)

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