Scoring Alignments

Genome 373

Genomic Informatics

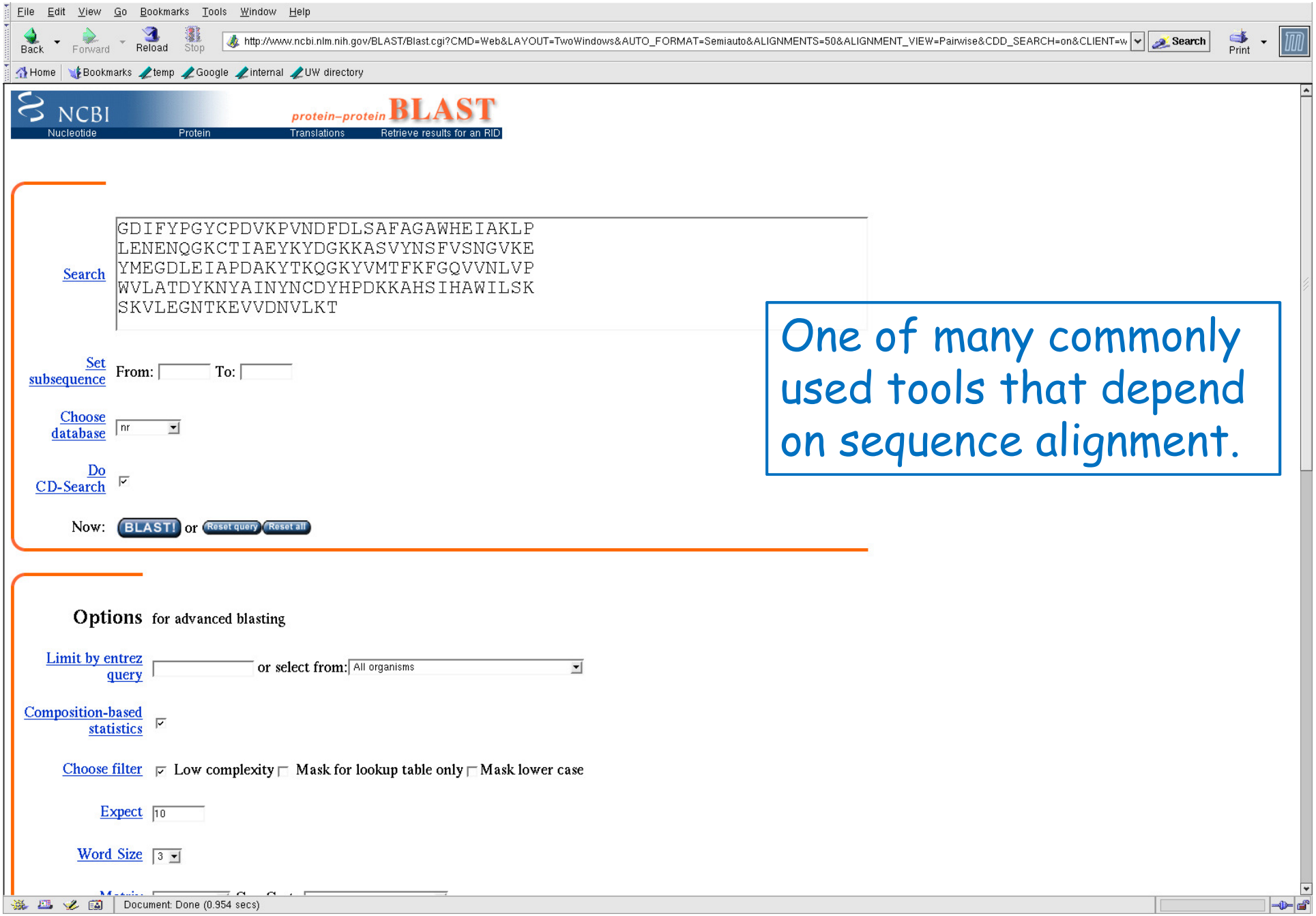
Elhanan Borenstein

Informatic Challenges: Examples

- Sequence comparison:
  - Find the best alignment of two sequences
  - Find the best match (alignment) of a given sequence in a large dataset of sequences
  - Find the best alignment of multiple sequences
- Motif and gene finding
- Relationship between sequences
  - Phylogeny
- Clustering and classification
- Many many many more …

Informatic Challenges: Examples

- **Sequence comparison:**
  - **Find the best alignment of two sequences**
  - **Find the best match (alignment) of a given sequence in a large dataset of sequences**
  - Find the best alignment of multiple sequences
- Motif and gene finding
- Relationship between sequences
  - Phylogeny
- Clustering and classification
- Many many many more …



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

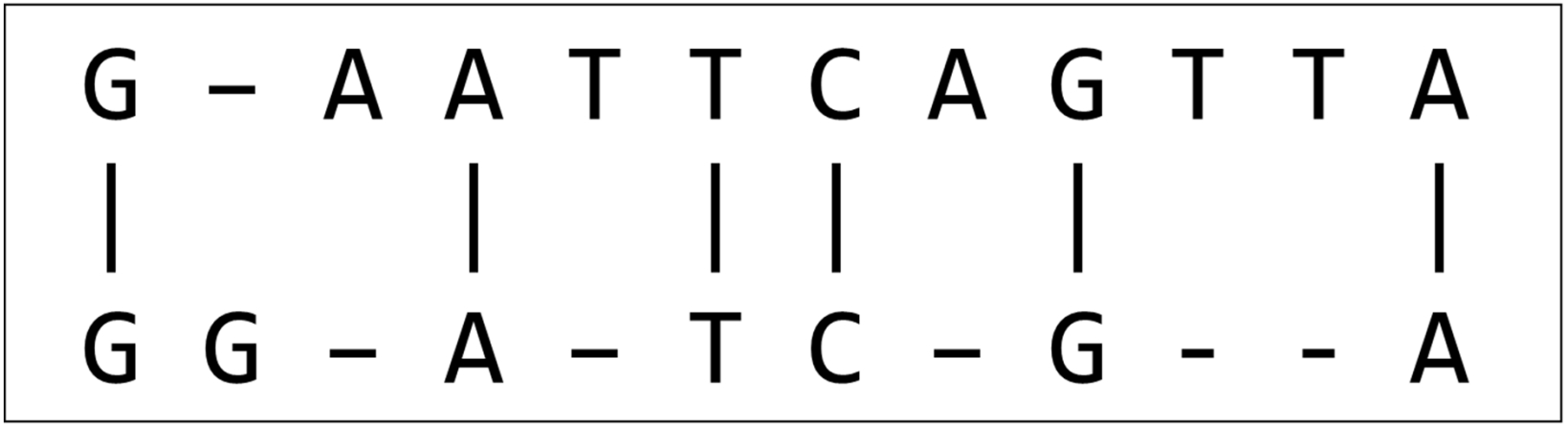
Motivation

- **Why compare/align two protein or DNA sequences?**

Motivation

- **Why compare/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

Sequence Alignment



Mission:

Find the best alignment between two sequences.

This is an optimization problem!

What do we need to solve this problem?

Mission:

Find the best alignment between two sequences.

A "search" algorithm for finding the alignment with the best score

- Dynamic programming

A method for scoring alignments

- Substitution matrix
- Gap penalties

Scoring Alignments

- Find the best alignment of GAATC and CATAC.

```
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.

Scoring Principles

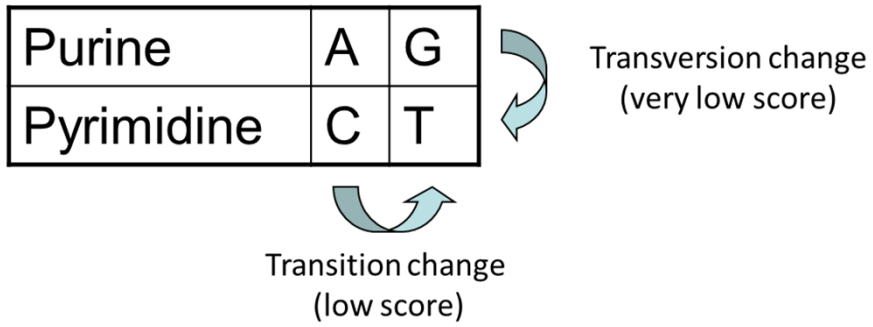
GAATC
CATAC

- Score each locus independently.
- The alignment score will be the sum of the scores in all loci.
- Perfect Matches will get a positive (good) score.
- What about mismatches?

Scoring Principles

GAATC
CATAC

- Score each locus independently.
- The alignment score will be the sum of the scores in all loci.
- Perfect Matches will get a positive (good) score.
- What about mismatches?

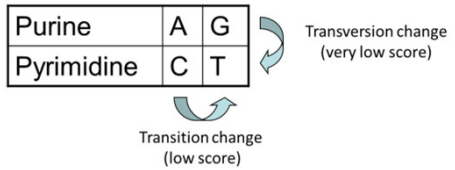


(transitions are typically about 2x as frequent as transversions in real sequences)

Scoring Aligned Bases

- 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



GAATC

CATAC

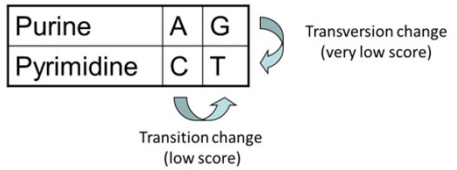
-5 + 10 + -5 + -5 + 10 = 5

What about gaps?

What About Gaps?

- 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



GAAT−C

CA−TAC

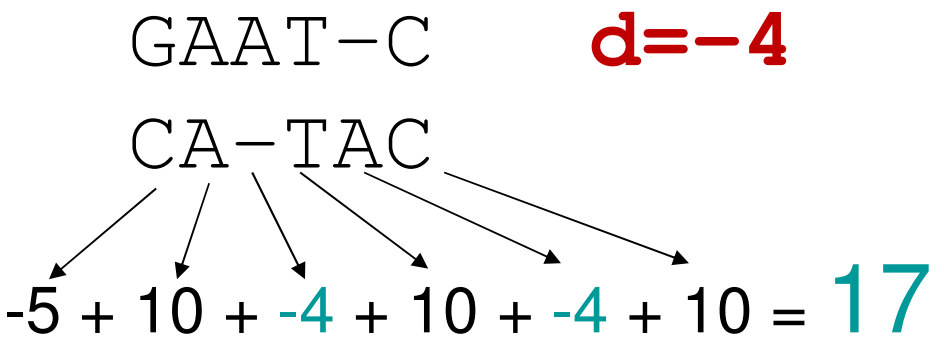
-5 + 10 + ? + 10 + ? + 10 = ?

What do gaps mean?

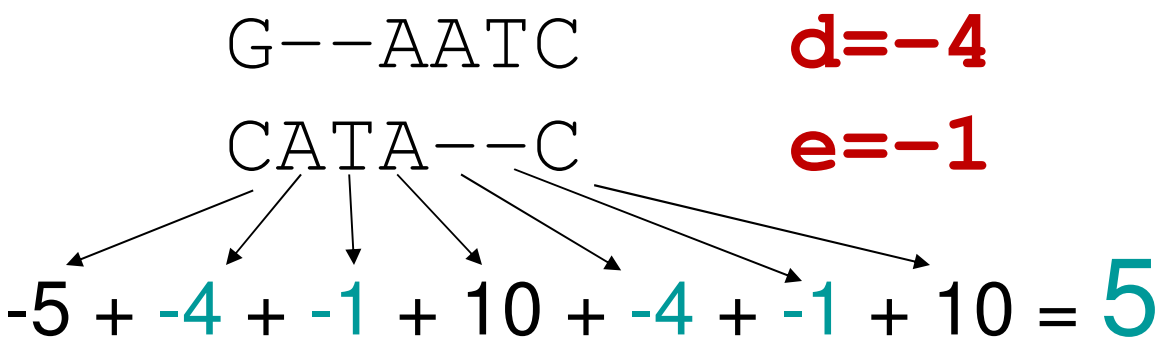
What if gaps have no penalty?

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**:



Same Method Applies to AA

BLOSUM62 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	-4
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1	-4
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1	-4
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1	-1	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3	-2	-4
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3	-1	-4
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-2	-1	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	0	-1	-4
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	-3	-1	-4
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	-3	-1	-4
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	0	1	-1	-4
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	-1	-1	-4
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	-3	-1	-4
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-2	-1	-2	-4
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	0	0	0	-4
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1	0	-4
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4	-3	-2	-4
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-3	-2	-1	-4
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	-3	-2	-1	-4
B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4	1	-1	-4
Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	-4
*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

regular 20 amino acids

ambiguity codes and stop

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

```
YMEGDLEIAPDAK
 +   D  E++PD
VL--DKELSPDGT
```