Genome 559:

Introduction to Statistical and Computational Genomics

Professors Jim Thomas and Elhanan Borenstein

TA Seungsoo Kim

Logistics

- Syllabus and web site:

  http://faculty.washington.edu/jht/GS559_2017/

- Should I take this class?
- Grading – see web page, problem sets count most.

Homework format

Hand in homework on paper at beginning of class.

Some parts may be hand-drawn if you like.

Class time structure

Split into three parts:

1) bioinformatics topics
2) Python (programming) topics
3) in class Python exercises

Bioinformatics

- Sequence alignment
- Genome assembly
- Sequence trees
- Molecular evolution
- Gene prediction
- Expression analysis
- Network analysis
- Machine learning
- Large dataset management
- Mass spec peptide identification
- Genotype-phenotype association
- Many others…

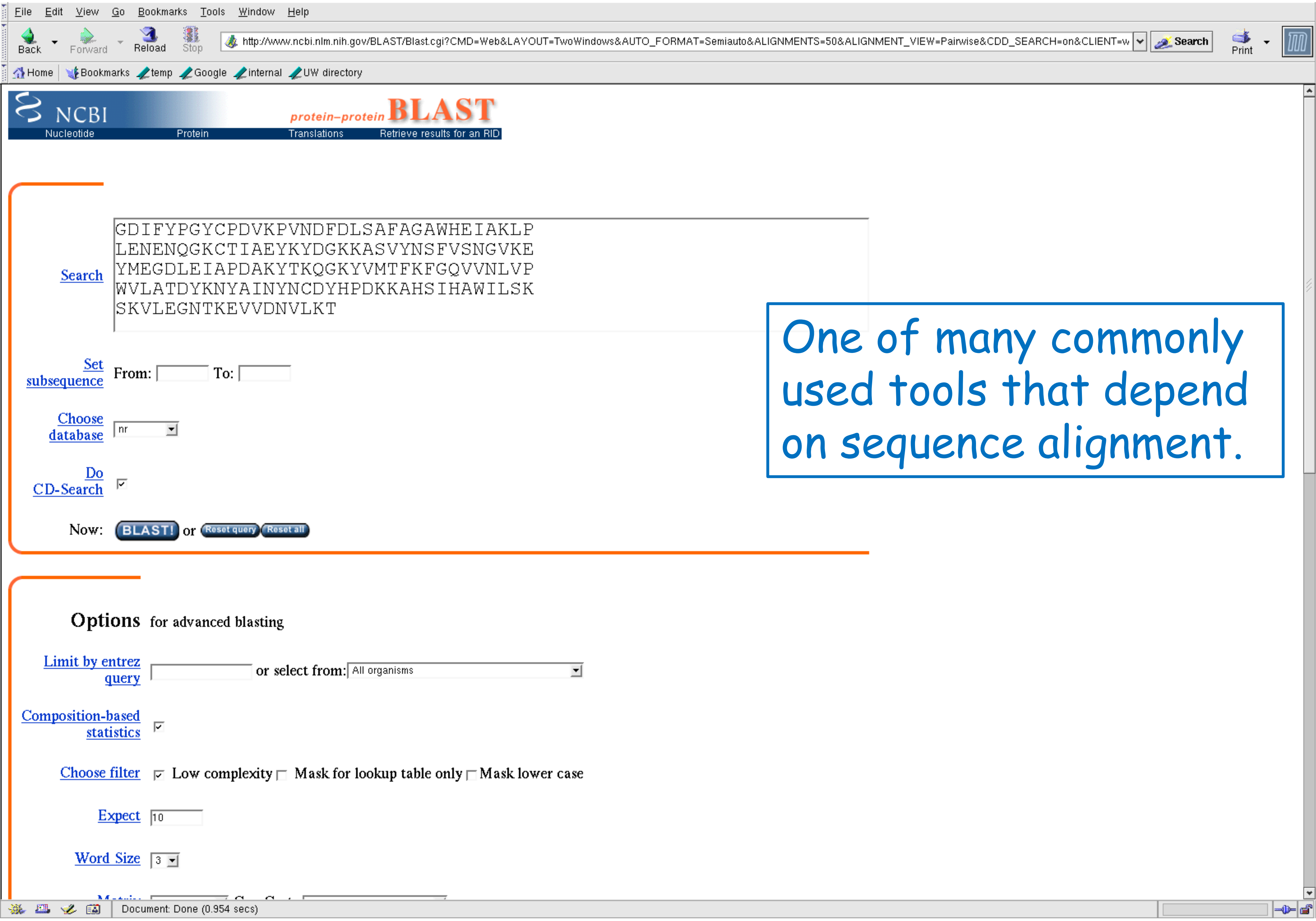
Sequence comparison: Introduction and motivation

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 related sequences in a database.
  - Locate functional elements (motifs or domains).
  - Infer protein or RNA structure, if the structure of a related sequence is known.
  - Analyze sequence evolution.



http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi?CMD=Web&LAYOUT=TwoWindows&AUTO_FORMAT=Semiauto&ALIGNMENTS=50&ALIGNMENT_VIEW=Pairwise&CDD_SEARCH=on&CLIENT=w

Search

```
GDIFYPGYCPDVKPVNDFDLSAFAGAWHEIAKLP
LENENQGKCTIAEYKYDGKKASVYNSFVSNGVKE
YMEGDLEIAPDAKYTKQGKYVMTFKFGQVVNLVP
WVLATDYKNYAINYNCDYHPDKKAHSIHAWILSK
SKVLEGNTKEVVDNVLKT
```

Set subsequence From: To:

Choose database nr

Do CD-Search ☑

Now: BLAST! or Reset query Reset all

Options for advanced blasting

Limit by entrez query or select from: All organisms

Composition-based statistics ☑

Choose filter ☑ Low complexity □ Mask for lookup table only □ Mask lower case

Expect 10

Word Size 3

Document: Done (0.954 secs)

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

Sequence comparison overview

- Problem: Find the "best" alignment between two sequences.
- To solve this problem, we need:
  - a method for scoring alignment quality
  - 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 called dynamic programming.

A simple alignment problem.

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

Scoring alignments

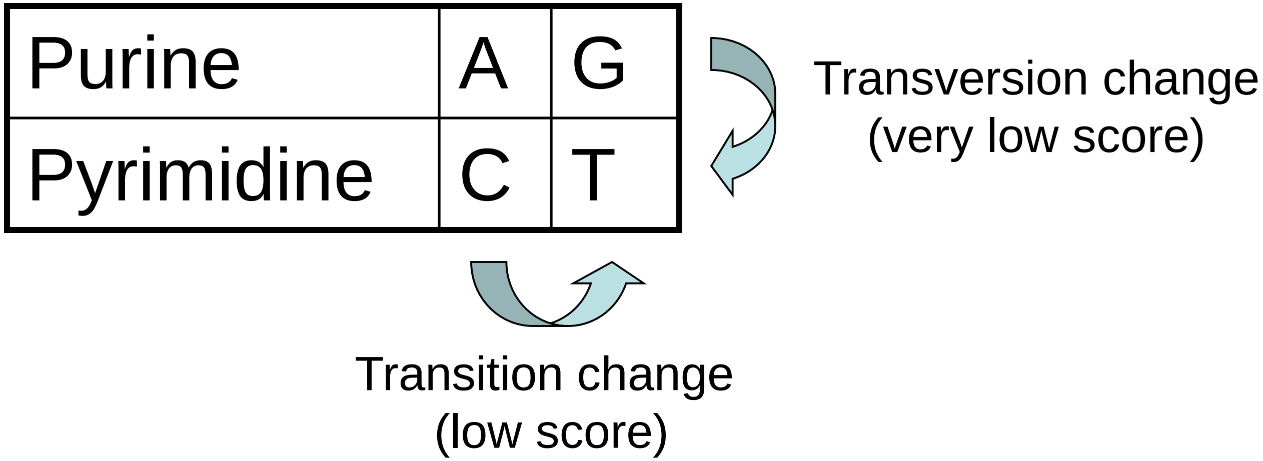
```
GAATC       GAAT-C      -GAAT-C
CATAC       C-ATAC      C-A-TAC

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



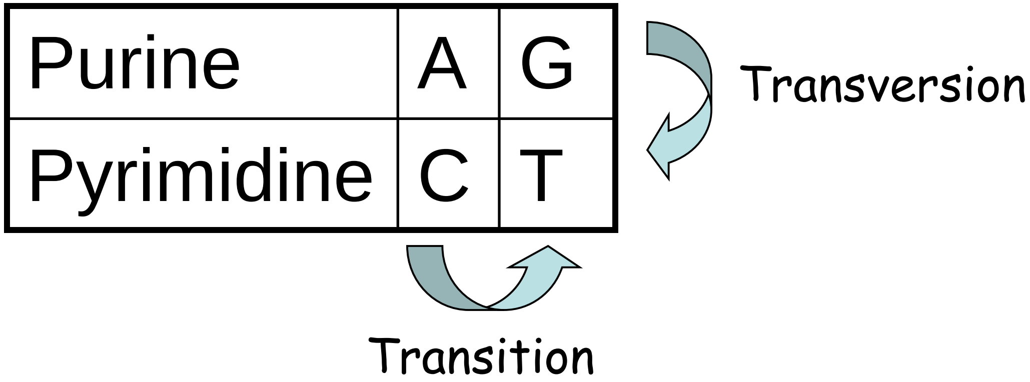
Purine	A	G
Pyrimidine	C	T

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

GAATC

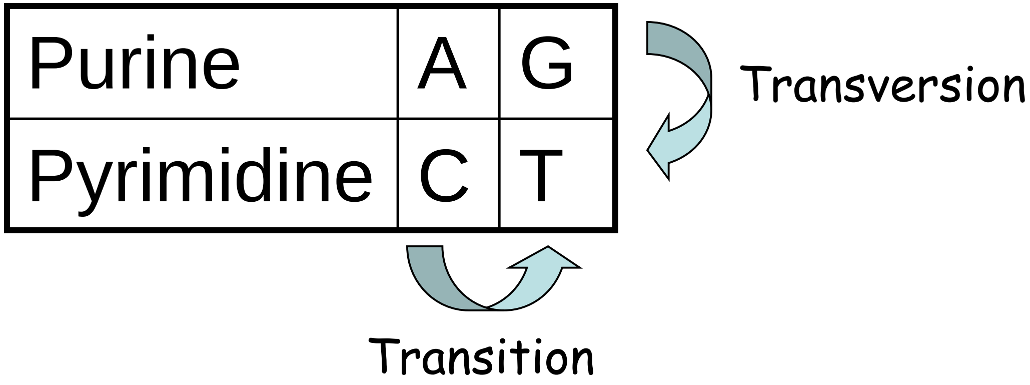
CATAC

-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

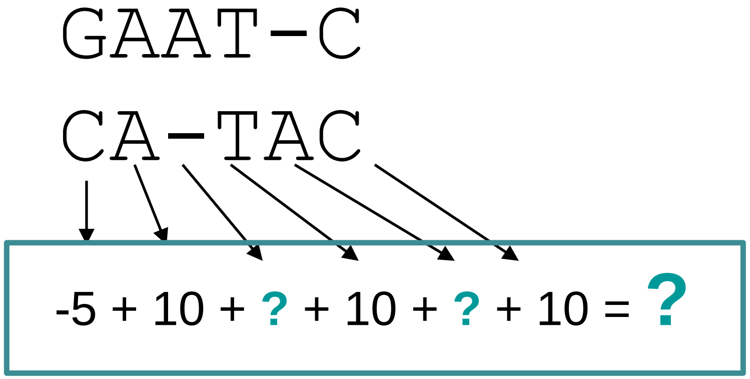
Scoring gaps



Purine	A	G
Pyrimidine	C	T

Transversion

Transition



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

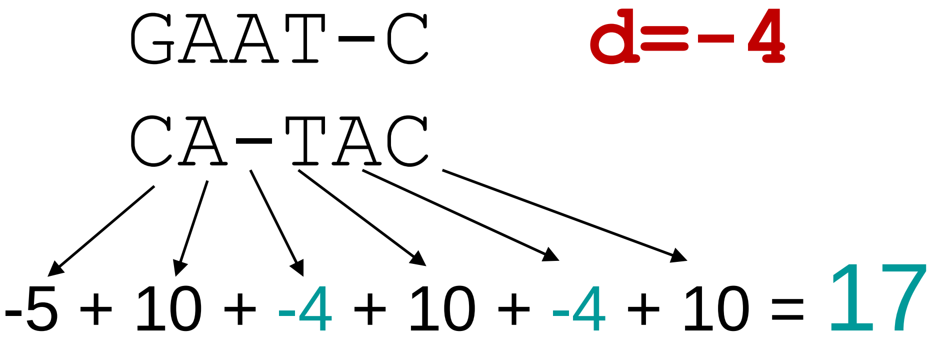
-5 + 10 + ? + 10 + ? + 10 = ?

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:

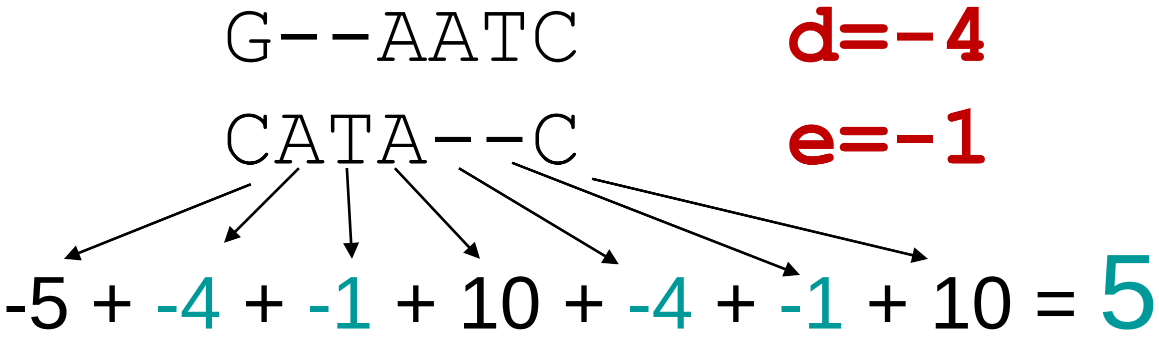


```
GAAT-C
CA-TAC
```

d=-4

-5 + 10 + -4 + 10 + -4 + 10 = 17

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



```
G--AATC
CATA--C
```

d=-4

e=-1

-5 + -4 + -1 + 10 + -4 + -1 + 10 = 5

Why not just allow gaps for free?

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