Algorithms for Bioinformatics: Basic Sequence Analysis

Jan Michael C. Yap
Core Facility for Bioinformatics
jcyap@up.edu.ph

PGC
PHILIPPINE GENOME CENTER

Outline

- Sequence Alignment
- Sequence Database Searching: BLAST

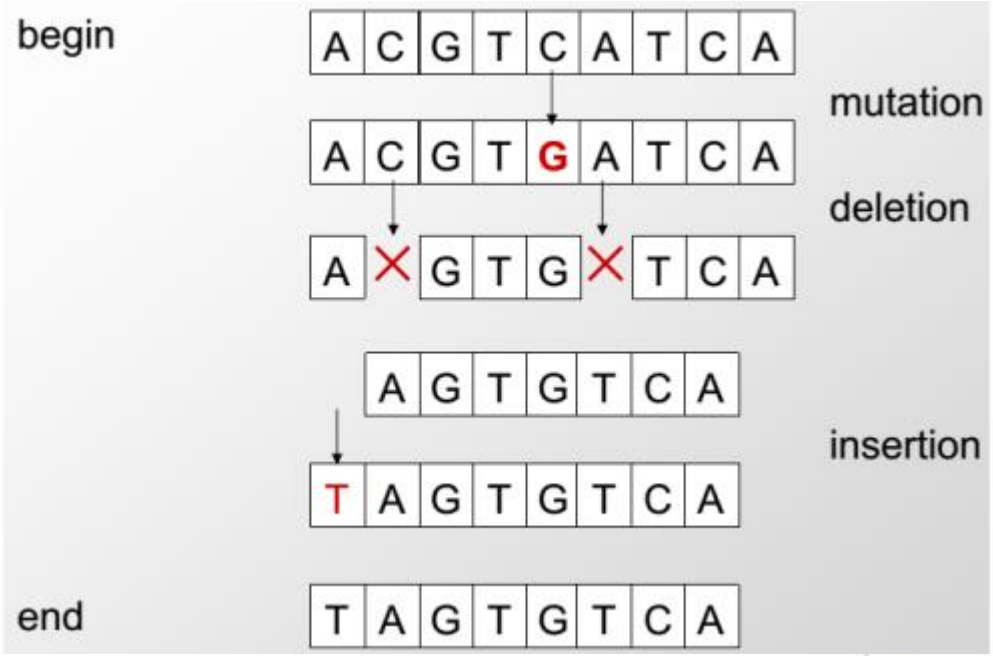
PGC
PHILIPPINE GENOME CENTER

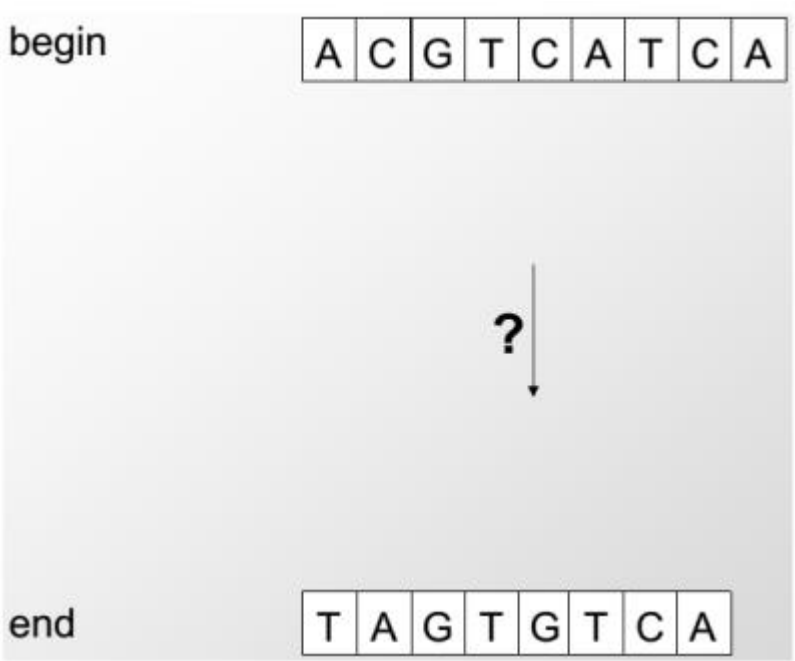
Outline

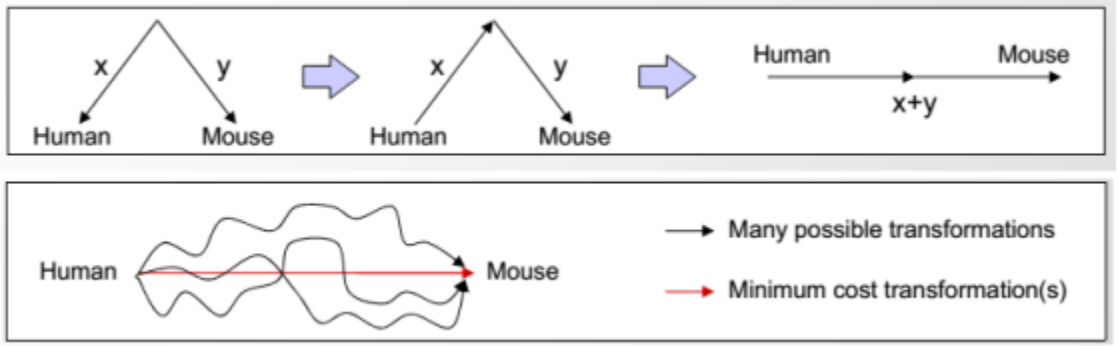
- Sequence Alignment
- Sequence Database Searching: BLAST

PGC
PHILIPPINE GENOME CENTER

Basis for Sequence Similarity



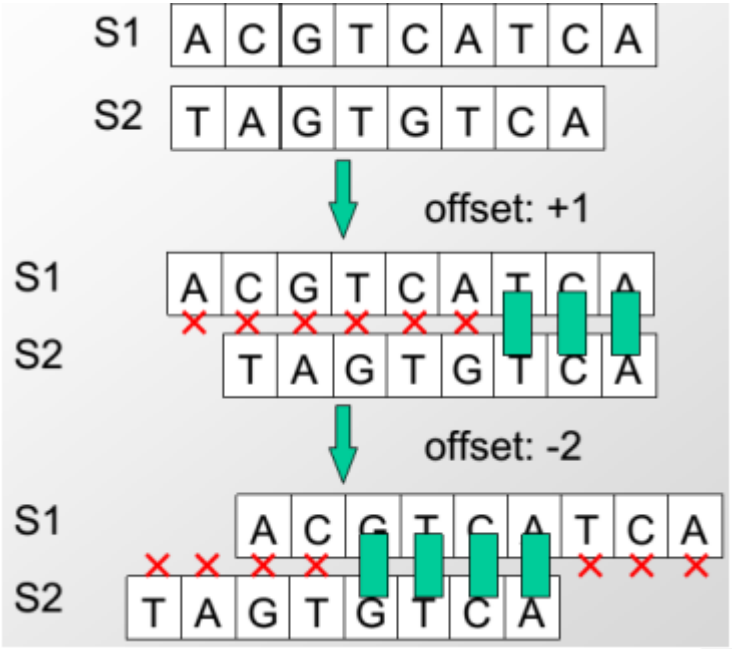


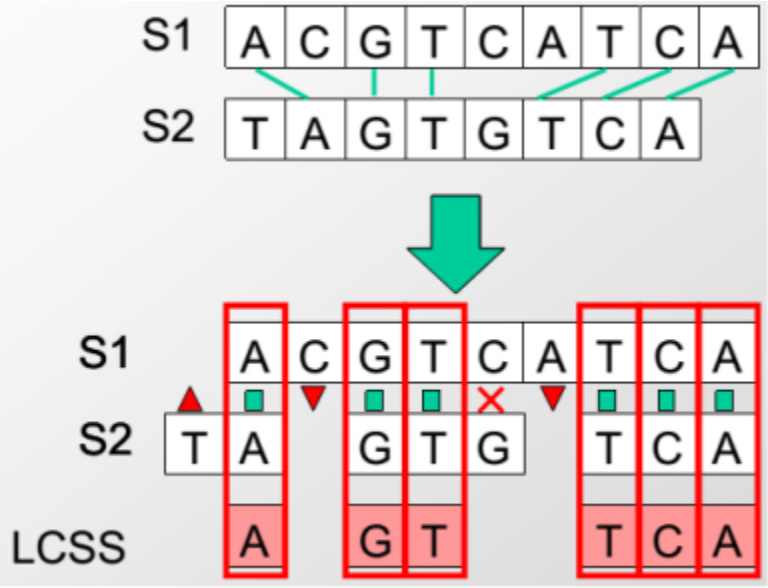


http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-047-computational-biology-genomes-networks-evolution-fall-2008/lecturenotes/MIT6_047f08_lec02_slide02.pdf

PGC
PHILIPPINE GENOME CENTER

Viewing the Sequence Similarity Problem Computationally





Scoring function:
Match(x,x) = +1
Mismatch(A,G)= -½
Mismatch(C,T)= -½
Mismatch(x,y) = -1

	A	G	T	C
A	+1	-½	-1	-1
G	-½	+1	-1	-1
T	-1	-1	+1	-½
C	-1	-1	-½	+1

purine pyrimid.

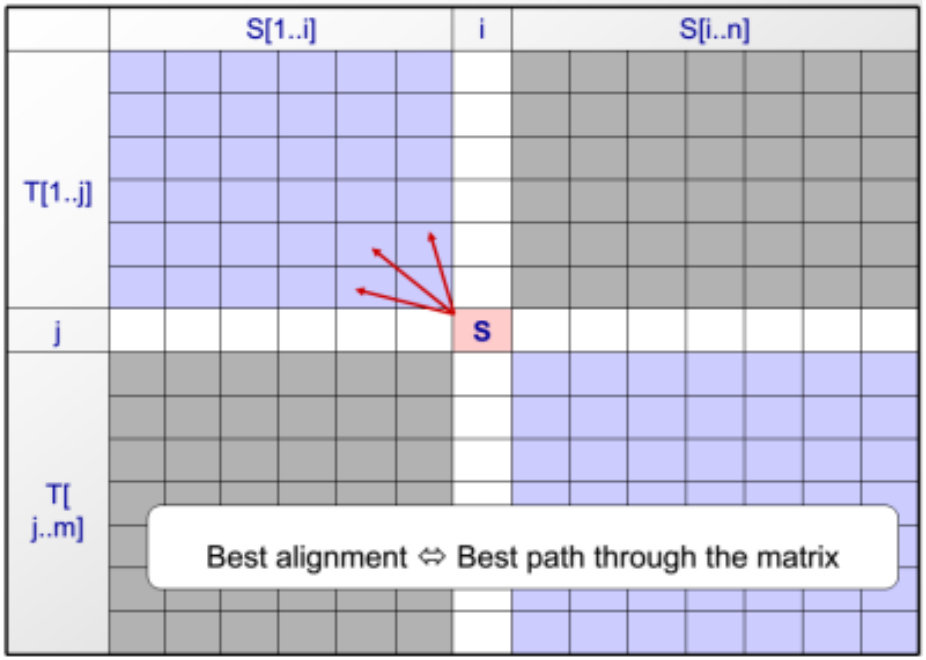
Transitions:
A⇔G, C⇔T common (lower penalty)

Transversions:
All other operations

http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-047-computational-biology-genomes-networks-evolution-fall-2008/lecturenotes/MIT6_047f08_lec02_slide02.pdf

PGC PHILIPPINE GENOME CENTER

Setting Things Up: Data Structure and Scoring



- Local update: Compute next alignment based on previous alignment (table look-up)
- Compute scores for prefixes of increasing length
- Compute the score of a cell from its neighbors

$$F(i,j) = \max\left(F(i-1,j) - gap, F(i,j) + score, F(i,j-1) - gap\right)$$

http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-047-computational-biology-genomes-networks-evolution-fall-2008/lecturenotes/MIT6_047f08_lec02_slide02.pdf

PGC
PHILIPPINE GENOME CENTER

Global Alignment: Needleman-Wunsch Algorithm

		A	T	G
	0			
G				
G				
A				
A				
T				
G				
T				
I/D	-2			
Mismatch	-1			
Match	1			

		A	T	G
	0	-2	-4	-6
G	-2			
G	-4			
A	-6			
A	-8			
T	-10			
G	-12			
T	-14			
I/D	-2			
Mismatch	-1			
Match	1			

Rosenberg MS (2009). Sequence Alignment: Methods, Concepts, and Strategies. University of California Press

PGC
PHILIPPINE GENOME CENTER

Global Alignment: Needleman-Wunsch Algorithm

		A	T	G
	0	-2	-4	-6
G	-2			
G	-4			
A	-6			
A	-8			
T	-10			
G	-12			
T	-14			
I/D	-2			
Mismatch	-1			
Match	1			

		A	T	G
	0	-2	-4	-6
G	-2			
G	-4			
A	-6			
A	-8			
T	-10			
G	-12			
T	-14			

I/D	-2	Direction	Align	Score
Mismatch	-1	D:	A G	0 + -1 = -1
Match	1			
		U:	A- -G	-2 + -2 = -4
		L:	-A G-	-2 + -2 =-4

Rosenberg MS (2009). Sequence Alignment: Methods, Concepts, and Strategies. University of California Press

PGC PHILIPPINE GENOME CENTER

Global Alignment: Needleman-Wunsch Algorithm

		A	T	G
	0	-2	-4	-6
G	-2			
G	-4			
A	-6			
A	-8			
T	-10			
G	-12			
T	-14			

I/D	-2	Direction	Align	Score
Mismatch	-1	D:	A G	0 + -1 = -1
Match	1			
		U:	A- -G	-2 + -2 = -4
		L:	-A G-	-2 + -2 =-4

		A	T	G
	0	-2	-4	-6
G	-2	-1(D)		
G	-4			
A	-6			
A	-8			
T	-10			
G	-12			
T	-14			

I/D	-2	Direction	Align	Score
Mismatch	-1	**D:**	**A** **G**	**0 + -1 = -1**
Match	1			
		U:	A- -G	-2 + -2 = -4
		L:	-A G-	-2 + -2 =-4

Rosenberg MS (2009). Sequence Alignment: Methods, Concepts, and Strategies. University of California Press

PGC PHILIPPINE GENOME CENTER

Global Alignment: Needleman-Wunsch Algorithm

		A	T	G
	0	-2	-4	-6
G	-2			
G	-4			
A	-6			
A	-8			
T	-10			
G	-12			
T	-14			

I/D	-2	Direction	Align	Score
Mismatch	-1	D:	A G	0 + -1 = -1
Match	1			
		U:	A- -G	-2 + -2 = -4
		L:	G- -A	-2 + -2 =-4

		A	T	G
	0	-2	-4	-6
G	-2	-1 (D)		
G	-4			
A	-6			
A	-8			
T	-10			
G	-12			
T	-14			

I/D	-2	Direction	Align	Score
Mismatch	-1	**D:**	**A** **G**	**0 + -1 = -1**
Match	1			
		U:	A- -G	-2 + -2 = -4
		L:	G- -A	-2 + -2 =-4

Rosenberg MS (2009). Sequence Alignment: Methods, Concepts, and Strategies. University of California Press

PGC PHILIPPINE GENOME CENTER

Global Alignment: Needleman-Wunsch Algorithm

		A	T	G
	0	-2	-4	-6
G	-2	-1 (D)		
G	-4			
A	-6			
A	-8			
T	-10			
G	-12			
T	-14			

I/D	-2	Direction	Align	Score
Mismatch	-1	**D:**	**A** / **G**	**0 + -1 = -1**
Match	1			
		U:	A- / -G	-2 + -2 = -4
		L:	G- / -A	-2 + -2 =-4

		A	T	G
	0	-2	-4	-6
G	-2	-1(D)	-3(D/L)	
G	-4			
A	-6			
A	-8			
T	-10			
G	-12			
T	-14			

I/D	-2	Direction	Align	Score
Mismatch	-1	**D:**	**AT** / **-G**	**-2 + -1 = -3**
Match	1			
		U:	AT- / --G	-4 + -2 = -6
		L:	**AT** / **G-**	**-1 + -2 =-3**

Rosenberg MS (2009). Sequence Alignment: Methods, Concepts, and Strategies. University of California Press

PGC PHILIPPINE GENOME CENTER

Global Alignment: Needleman-Wunsch Algorithm

		A	T	G
	0	-2	-4	-6
G	-2	-1(D)	-3(D/L)	
G	-4			
A	-6			
A	-8			
T	-10			
G	-12			
T	-14			

I/D	-2	Direction	Align	Score
Mismatch	-1	**D:**	**AT** / **-G**	**-2 + -1 = -3**
Match	1	U:	AT- / --G	-4 + -2 = -6
		L:	**AT** / **G-**	**-1 + -2 =-3**

		A	T	G
	0	-2	-4	-6
G	-2	-1(D)	-3(D/L)	-3(D)
G	-4			
A	-6			
A	-8			
T	-10			
G	-12			
T	-14			

I/D	-2	Direction	Align	Score
Mismatch	-1	**D:**	**ATG** / **--G**	**-4 + 1 = -3**
Match	1	U:	ATG- / ---G	-6 + -2 = -8
		L:	ATG / -G-	-3 + -2 =-5

Rosenberg MS (2009). Sequence Alignment: Methods, Concepts, and Strategies. University of California Press

PGC PHILIPPINE GENOME CENTER

Global Alignment: Needleman-Wunsch Algorithm

		A	T	G
	0	-2	-4	-6
G	-2	-1(D)	-3(D/L)	-3(D)
G	-4	-3(D/U)		
A	-6			
A	-8			
T	-10			
G	-12			
T	-14			

		A	T	G
	0	-2	-4	-6
G	-2	-1(D)	-3(D/L)	-3(D)
G	-4	-3(D/U)	-2(D)	-2(D)
A	-6	-3(D)	-4(D/U)	-3(D)
A	-8	-5(D/U)	-4(D)	-5(D/U)
T	-10	-7(U)	-4(D)	-5(D)
G	-12	-9(U)	-6(U)	-3(D)
T	-14	-11(U)	-8(U)	-5(U)

I/D	-2	Direction	Align	Score
Mismatch	-1	**D:**	**-A** / **GG**	**-2 + -1 = -3**
Match	1			
		U:	**A-** / **GG**	**-1 + -2 = -3**
		L:	--A / GG-	-4 + -2 =-6

Rosenberg MS (2009). Sequence Alignment: Methods, Concepts, and Strategies. University of California Press

PGC PHILIPPINE GENOME CENTER

Global Alignment: Needleman-Wunsch Algorithm

		A	T	G
	0	-2	-4	-6
G	-2	-1(D)	-3(D/L)	-3(D)
G	-4	-3(D/U)	-2(D)	-2(D)
A	-6	-3(D)	-4(D/U)	-3(D)
A	-8	-5(D/U)	-4(D)	-5(D/U)
T	-10	-7(U)	-4(D)	-5(D)
G	-12	-9(U)	-6(U)	-3(D)
T	-14	-11(U)	-8(U)	-5(U)

		A	T	G
	0	-2	-4	-6
G	**-2**	-1(D)	-3(D/L)	-3(D)
G	**-4**	-3(D/U)	-2(D)	-2(D)
A	**-6**	**-3(D)**	-4(D/U)	-3(D)
A	-8	**-5(D/U)**	-4(D)	-5(D/U)
T	-10	-7(U)	**-4(D)**	-5(D)
G	-12	-9(U)	-6(U)	**-3(D)**
T	-14	-11(U)	-8(U)	**-5(U)**

BLUE:	G	G	A	A	T	G	T
	-	-	-	A	T	G	-
RED:	G	G	A	A	T	G	T
	-	-	A	-	T	G	-

Rosenberg MS (2009). Sequence Alignment: Methods, Concepts, and Strategies. University of California Press

PGC
PHILIPPINE GENOME CENTER

Semi-Global Alignment: Modified Needleman-Wunsch Algorithm

		A	T	G
	0			
G				
G				
A				
A				
T				
G				
T				
I/D	-2			
Mismatch	-1			
Match	1			

		A	T	G
	0	**0**	**0**	**0**
G	**0**			
G	**0**			
A	**0**			
A	**0**			
T	**0**			
G	**0**			
T	**0**			
I/D	-2			
Mismatch	-1			
Match	1			

Rosenberg MS (2009). Sequence Alignment: Methods, Concepts, and Strategies. University of California Press

Semi-Global Alignment: Modified Needleman-Wunsch Algorithm

		A	T	G
	0	**0**	**0**	**0**
G	**0**			
G	**0**			
A	**0**			
A	**0**			
T	**0**			
G	**0**			
T	**0**			
I/D	-2			
Mismatch	-1			
Match	1			

		A	T	G
	0	**0**	**0**	**0**
G	**0**	-1(D)	-1(D)	-1(D)
G	**0**	-1(D)	-2(D)	-2(D)
A	**0**	1(D)	-2(D/L)	-3(D)
A	**0**	**1(D)**	0(D)	-2(L)
T	**0**	-1(D)	**2(D)**	0(L)
G	**0**	-1(D)	0(U)	**3(D)**
T	**0**	-1(D)	0(D)	1(U)
I/D	-2			
Mismatch	-1			
Match	1			

BLUE:	G	G	A	A	T	G	T
	-	-	-	A	T	G	-

Rosenberg MS (2009). Sequence Alignment: Methods, Concepts, and Strategies. University of California Press

PGC
PHILIPPINE GENOME CENTER

Local Alignment: Smith-Waterman Algorithm

		A	T	G
	0	**0**	**0**	**0**
G	**0**			
G	**0**			
A	**0**			
A	**0**			
T	**0**			
G	**0**			
T	**0**			

I/D	-2
Mismatch	-1
Match	1

		A	T	G
	0	0	0	0
G	0			
G	0			
A	0			
A	0			
T	0			
G	0			
T	0			

I/D	-2	Direction	Align	Score
Mismatch	-1	D:	A	0 + -1 = -1
Match	1		G	
		U:	A-	-2 + -2 = -4
			-G	
		L:	-A	-2 + -2 = -4
			G-	
		IF D, U, and L < 0, SET TO 0		

Rosenberg MS (2009). Sequence Alignment: Methods, Concepts, and Strategies. University of California Press

PGC
PHILIPPINE GENOME CENTER

Local Alignment: Smith-Waterman Algorithm

		A	T	G
	0	0	0	0
G	0	0		
G	0			
A	0			
A	0			
T	0			
G	0			
T	0			

I/D	-2	Direction	Align	Score
Mismatch	-1	D:	A G	0 + -1 = -1
Match	1			
		U:	A- -G	-2 + -2 = -4
		L:	-A G-	-2 + -2 = -4
		IF D, U, and L < 0, SET TO 0		

		A	T	G
	0	0	0	0
G	0	0	0	0
G	0	0	0	0
A	0	1		
A	0			
T	0			
G	0			
T	0			

I/D	-2	Direction	Align	Score
Mismatch	-1	**D:**	**--A** **GGA**	**0 + 1 = 1**
Match	1			
		U:	-A- GGA	0 + -2 = -2
		L:	---A GGA-	0 + -2 = -2
		IF D, U, and L < 0, SET TO 0		

Rosenberg MS (2009). Sequence Alignment: Methods, Concepts, and Strategies. University of California Press

PGC
PHILIPPINE GENOME CENTER

Local Alignment: Smith-Waterman Algorithm

		A	T	G
	0	0	0	0
G	0	0	0	0
G	0	0	0	0
A	0	1	0	0
A	0	1		
T	0			
G	0			
T	0			

I/D	-2	Direction	Align	Score
Mismatch	-1	**D:**	**---A** / **GGAA**	**0 + 1 = 1**
Match	1			
		U:	-A-- / GGAA	1 + -2 = -1
		L:	----A / GGAA-	0 + -2 = -2
		IF D, U, and L < 0, SET TO 0		

		A	T	G
	0	0	0	0
G	0	0	0	0
G	0	0	0	0
A	0	1	0	0
A	0	1	0	
T	0			
G	0			
T	0			

I/D	-2	Direction	Align	Score
Mismatch	-1	**D:**	**---AT** / **GGAAA**	**1 + -1 = 0**
Match	1			
		U:	-AT- / GGAA	0 + -2 = -2
		L:	---AT / GGAA-	0 + -2 = -2
		IF D, U, and L < 0, SET TO 0		

Rosenberg MS (2009). Sequence Alignment: Methods, Concepts, and Strategies. University of California Press

PGC PHILIPPINE GENOME CENTER

Local Alignment: Smith-Waterman Algorithm

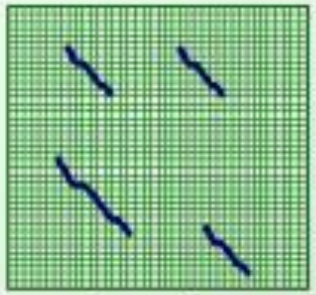
		A	T	G
	0	0	0	0
G	0	0	0	0
G	0	0	0	0
A	0	**1(D)**	0	0
A	0	**1(D)**	0(D)	0
T	0	0	**2(D)**	0
G	0	0	0	**3(D)**
T	0	0	**1(D)**	1(U)

A, ATG, T

Rosenberg MS (2009). Sequence Alignment: Methods, Concepts, and Strategies. University of California Press

PGC PHILIPPINE GENOME CENTER

Pairwise Alignment Algorithms: Summary

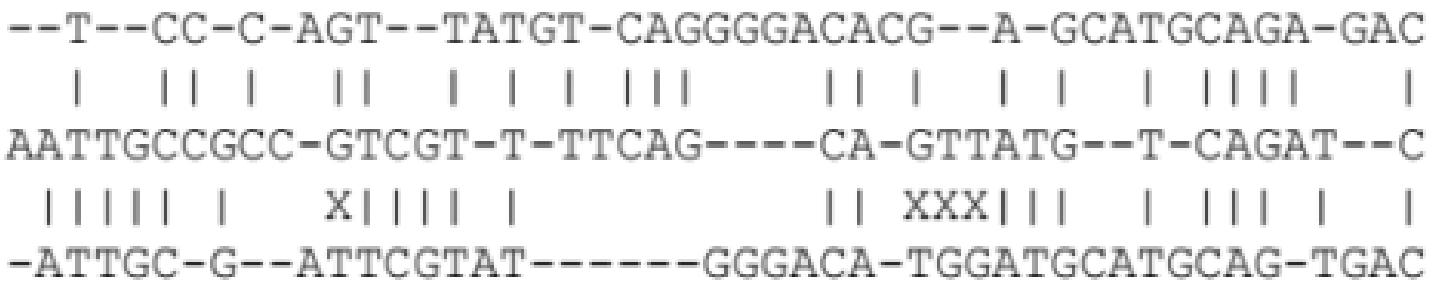


	Global	Local	Semi-global
Initialization	Top left	Top row/left col.	Top row
Iteration:max	$F(i-1, j)-d$ $F(i, j-1)-d$ $F(i-1, j-1)+s(x_i, y_j)$	0 $F(i-1, j)-d$ $F(i, j-1)-d$ $F(i-1, j-1)+s(x_i, y_j)$	$F(i-1, j)-d$ $F(i, j-1)-d$ $F(i-1, j-1)+s(x_i, y_j)$
Termination	Bottom right	Anywhere	Right column

http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-047-computational-biology-genomes-networks-evolution-fall-2008/lecturenotes/MIT6_047f08_lec02_slide02.pdf

PGC
PHILIPPINE GENOME CENTER

Multiple Sequence Alignment: Progressive Alignment

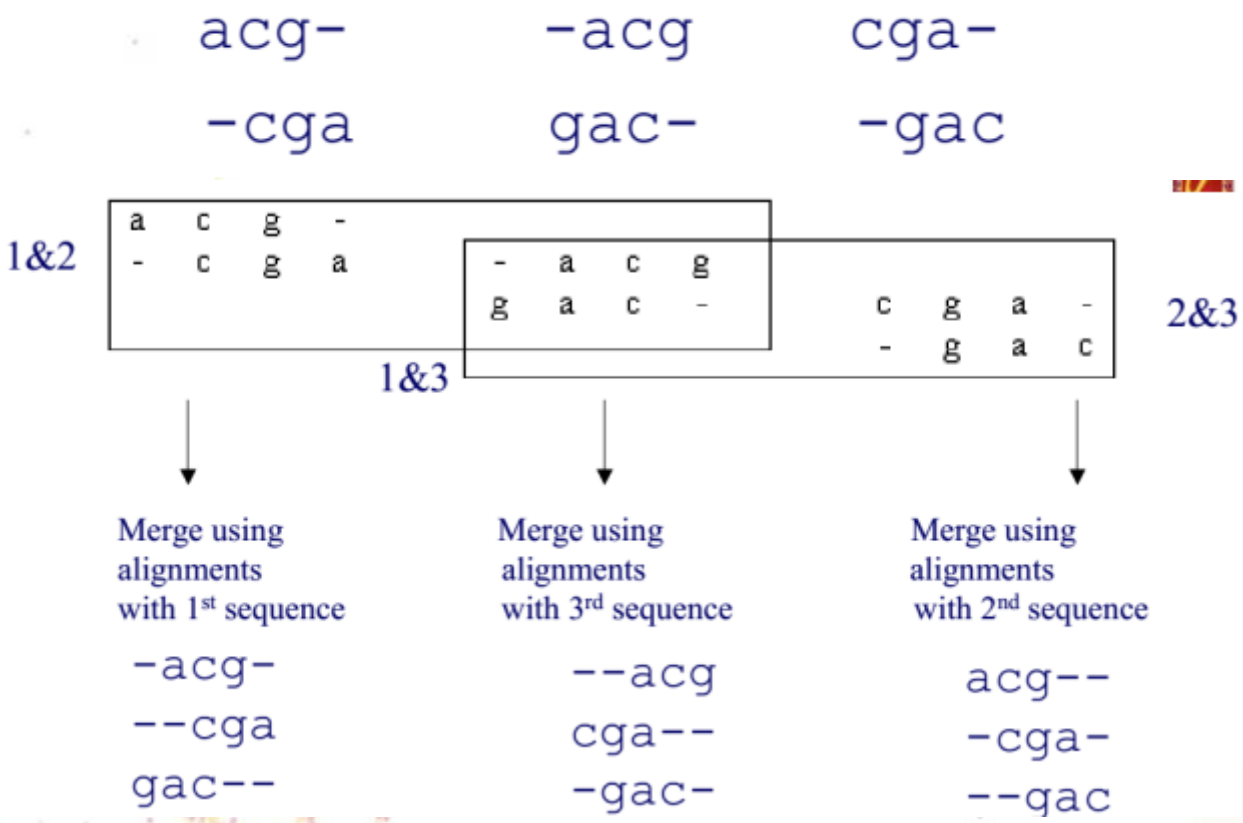


Jones NC and Pevzner PA (2004). Introduction to Bioinformatics Algorithms.

PGC
PHILIPPINE GENOME CENTER

Multiple Sequence Alignment: Progressive Alignment

- Align the following: acg, cga, gac



https://www.cs.princeton.edu/~mona/Lecture/MSA-slides.pdf

PGC PHILIPPINE GENOME CENTER

Outline

- Sequence Alignment
- **Sequence Database Searching: BLAST**

PGC
PHILIPPINE GENOME CENTER

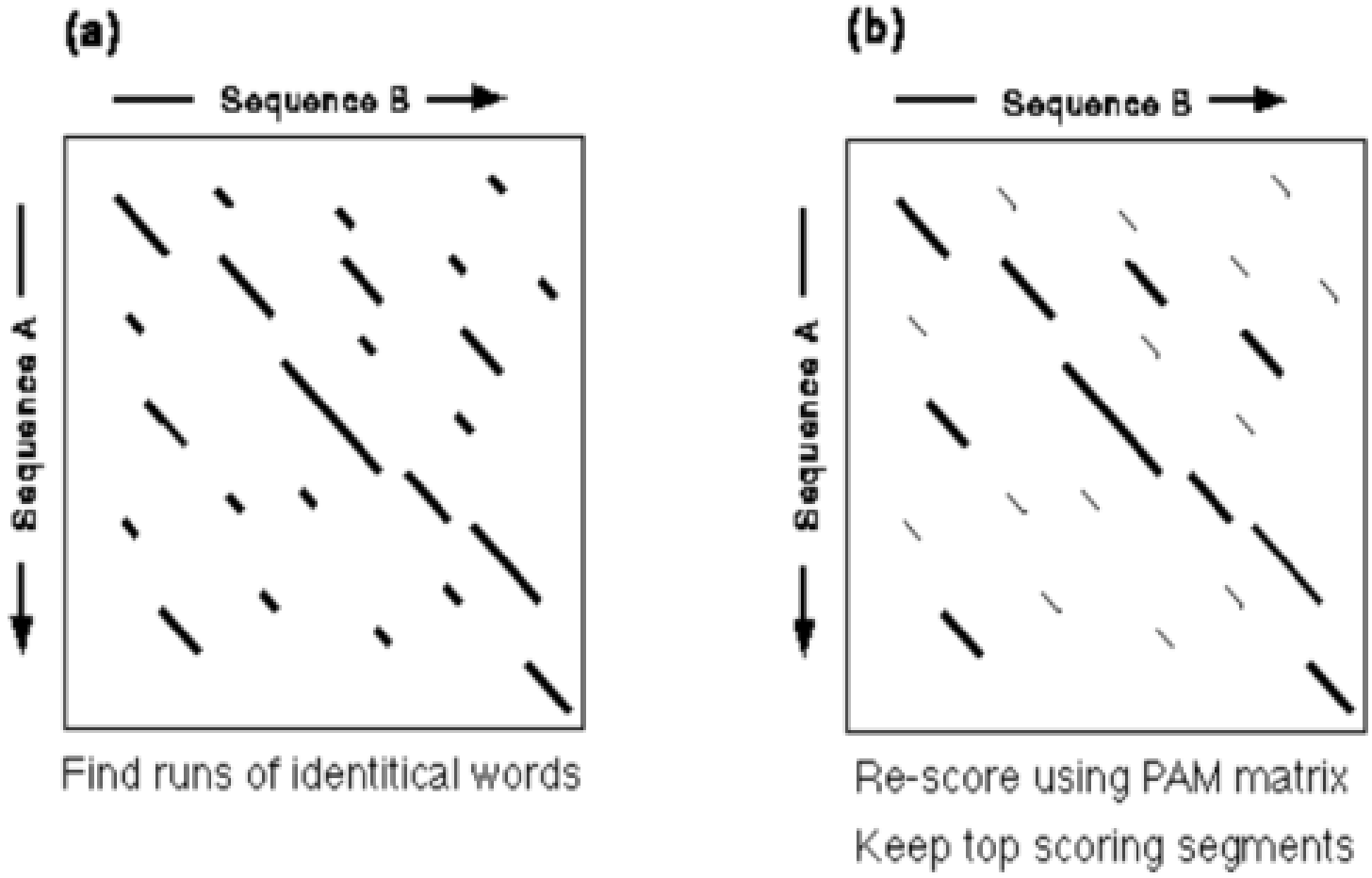
BLAST

- BLAST is a refinement of **FASTA**
- Employs **heuristics** for local alignment
  - **Discard irrelevant sequences** and **perform exact local alignment** on the remaining sequences
- Designed **specifically for database searches**
- BLAST produces several short segments called **High Scoring Segment Pairs (HSPs)**

http://www.fing.edu.uy/inco/grupos/bioinf/bioinfo1/material/Blast_bioinfoI_set09.pdf

PGC
PHILIPPINE GENOME CENTER

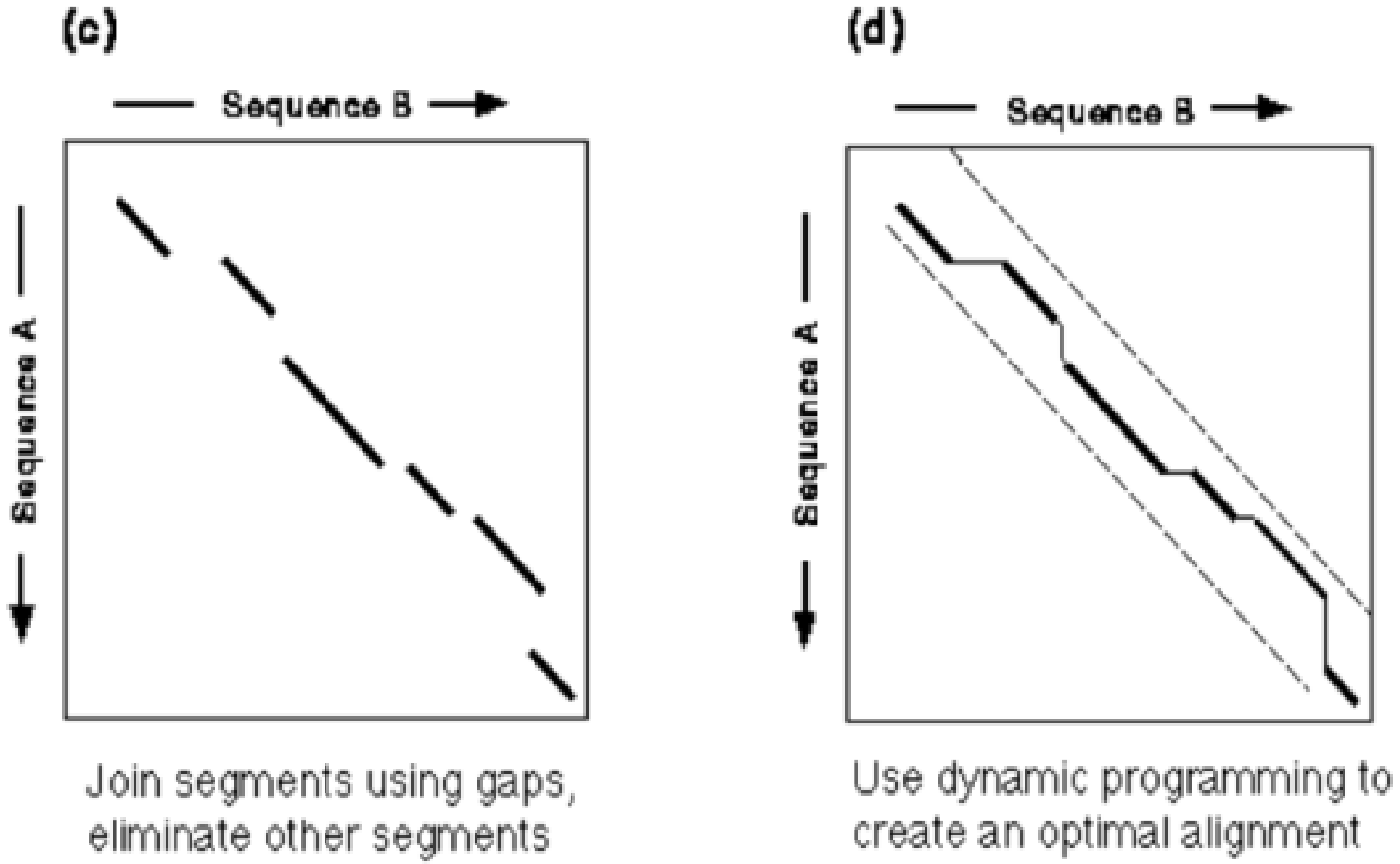
FASTA



http://www.fing.edu.uy/inco/grupos/bioinf/bioinfo1/material/Blast_bioinfoI_set09.pdf

PGC
PHILIPPINE GENOME CENTER

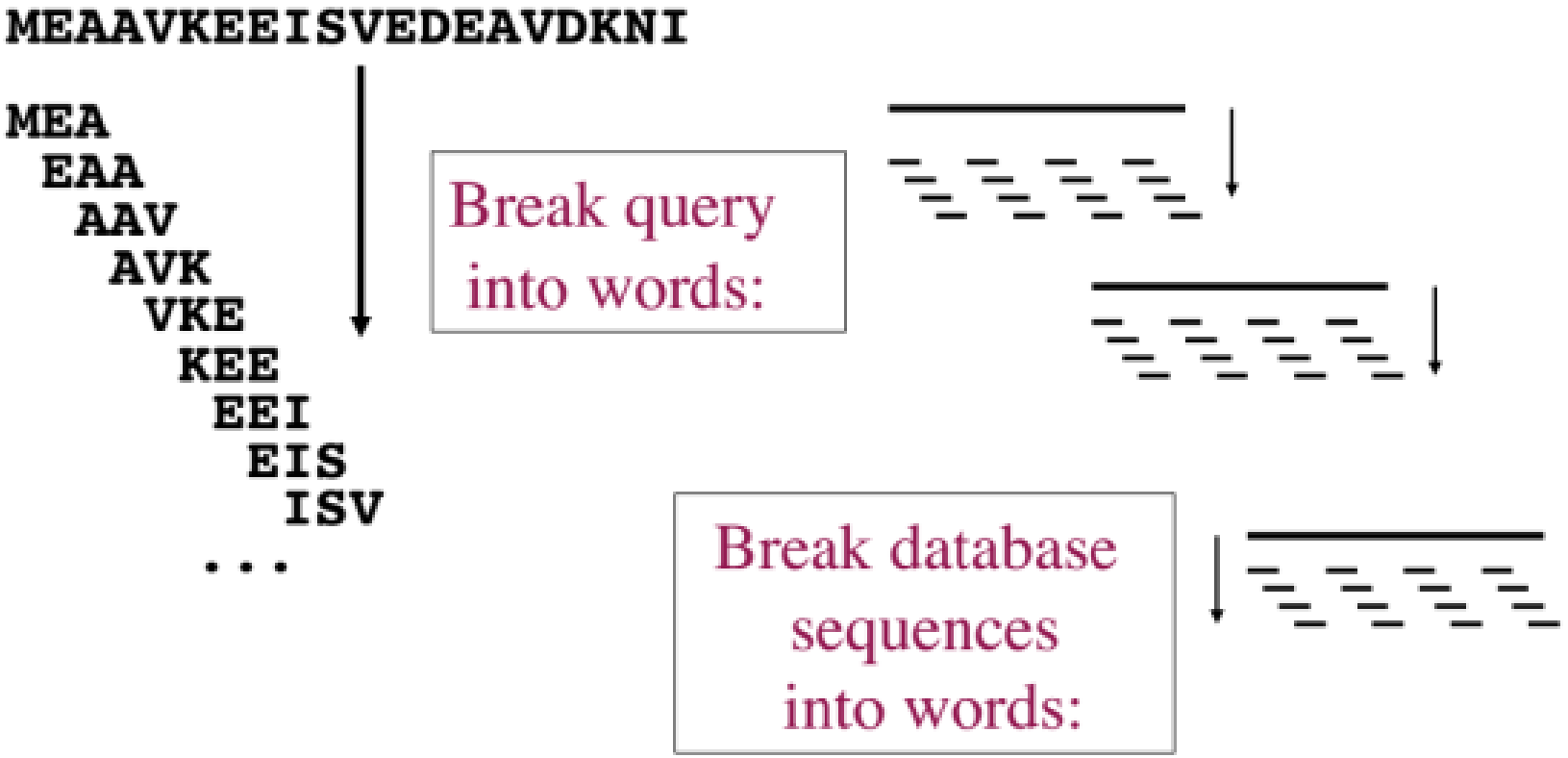
FASTA



http://www.fing.edu.uy/inco/grupos/bioinf/bioinfo1/material/Blast_bioinfoI_set09.pdf

PGC
PHILIPPINE GENOME CENTER

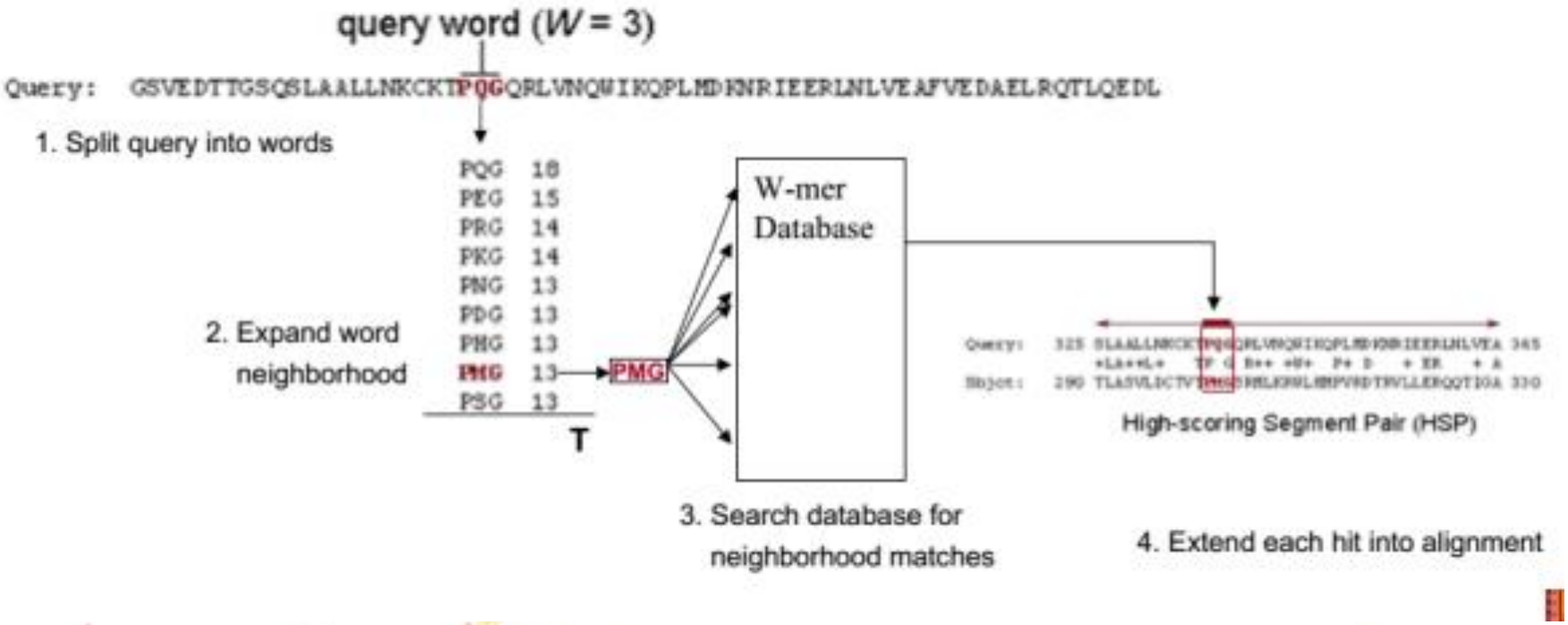
BLAST Preprocessing



http://www.fing.edu.uy/inco/grupos/bioinf/bioinfo1/material/Blast_bioinfoI_set09.pdf

PGC
PHILIPPINE GENOME CENTER

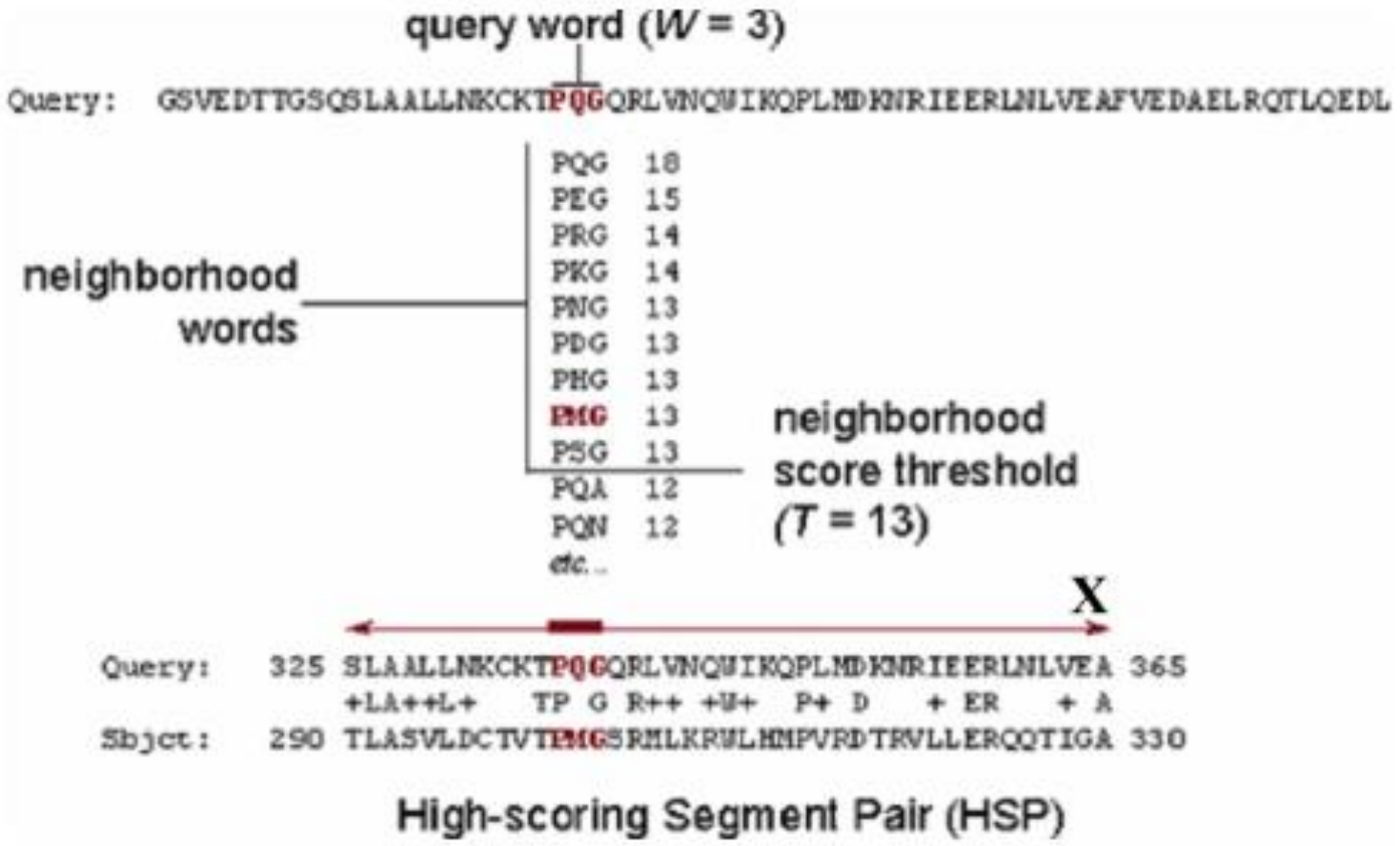
Sequence Database Searching: BLAST Algorithm



http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-047-computational-biology-genomes-networks-evolution-fall-2008/lecturenotes/MIT6_047f08_lec03_slide03.pdf

PGC
PHILIPPINE GENOME CENTER

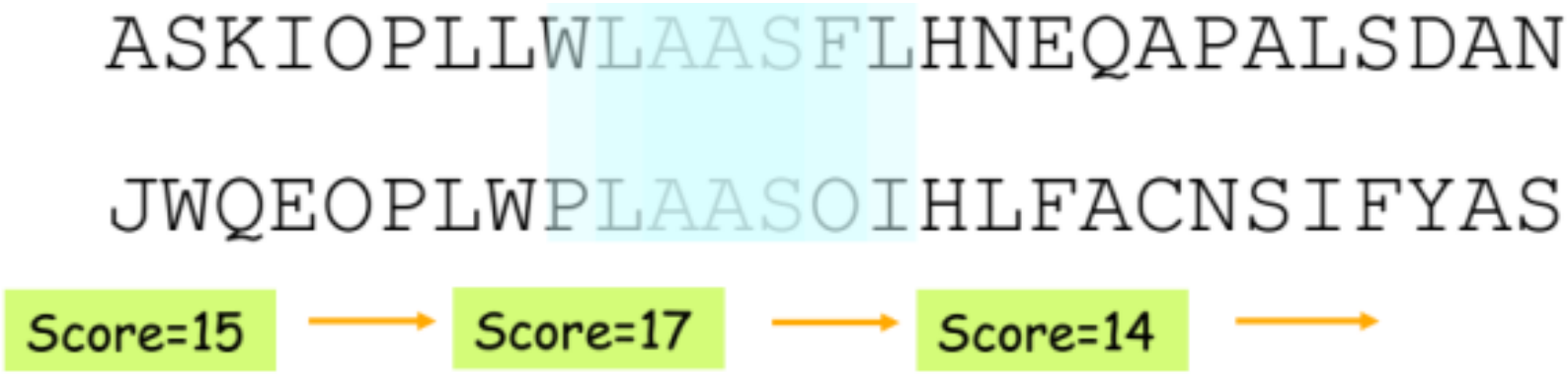
Sequence Database Searching: BLAST Heuristic



http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-047-computational-biology-genomes-networks-evolution-fall-2008/lecturenotes/MIT6_047f08_lec03_slide03.pdf

PGC
PHILIPPINE GENOME CENTER

BLAST Alignment Extension



- Stop extending if the score of the current alignment drops at least X points below the maximum score (obtained so far)
  - X is called the **alignment extension decrement threshold**
- Retain HSPs by discarding segments with score less than the **segment score threshold**

http://www.fing.edu.uy/inco/grupos/bioinf/bioinfo1/material/Blast_bioinfoI_set09.pdf

BLAST Scoring

- E-value[1]

$$E - value = KMNe^{-\lambda S}$$

  - K and λ are "normalization" parameters, as determined by the scoring matrix and gap penalties used
  - M is the length of the query sequence
  - N is the total length of the sequences stored in the database

- Bit score[2]
  - Bit score S'

$$S' = \frac{\lambda S - \ln K}{\ln 2}$$

  - Formula for E-value using S'

$$E - value = MN2^{-S'}$$

1 http://www.fing.edu.uy/inco/grupos/bioinf/bioinfo1/material/Blast_bioinfoI_set09.pdf
2 https://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html#head3

PGC PHILIPPINE GENOME CENTER

THANK YOU VERY MUCH! ☺

PGC
PHILIPPINE GENOME CENTER