A Bit-Parallel, General Integer-Scoring Sequence Alignment Algorithm

GARY BENSON, YOZEN HERNANDEZ, & JOSHUA LOVING

BIOINFORMATICS PROGRAM
BOSTON UNIVERSITY
JLOVING@BU.EDU

Introduction: Problem Description

Input:

- Sequences X and Y
- Integer weights M; I; G
  match; mismatch; indel or gap that define a similarity or distance scoring function S

Output:

- Calculate the global alignment score for X and Y

Introduction

Global Alignment – Needleman-Wunsch
Alignment Scoring Matrix

		Sequence X							
		-	A	C	T	G	C	A	A
Sequence Y	-	0	-5	-10	-15	-20	-25	-30	-35
	A	-5	2	-3	-8	-13	-18	-23	-28
	G	-10	-3	-1	-6	-6	-11	-16	-21
	T	-15	-8	-6	1	-4	-9	-14	-19
	C	-20	-13	-6	-4	-2	-2	-7	-12
	A	-25	-18	-11	-9	-7	-5	0	-5
	A	-30	-23	-16	-14	-12	-10	-3	2

Match = 2, Mismatch = -3, Indel = -5

Introduction

	-	A	C	T	G	C	A	A
-	0	-5	-10	-15	-20	-25	-30	-35
A	-5	2	-3	-8	-13	-18	-23	-28
G	-10	-3	-1	-6	-6	-11	-16	-21
T	-15	-8	-6	1	-4	-9	-14	-19
C	-20	-13	-6	-4	-2	-2	-7	-12
A	-25	-18	-11	-9	-7	-5	0	-5
A	-30	-23	-16	-14	-12	-10	-3	2

Match = 2, Mismatch = -3, Indel = -5

Integer Scores

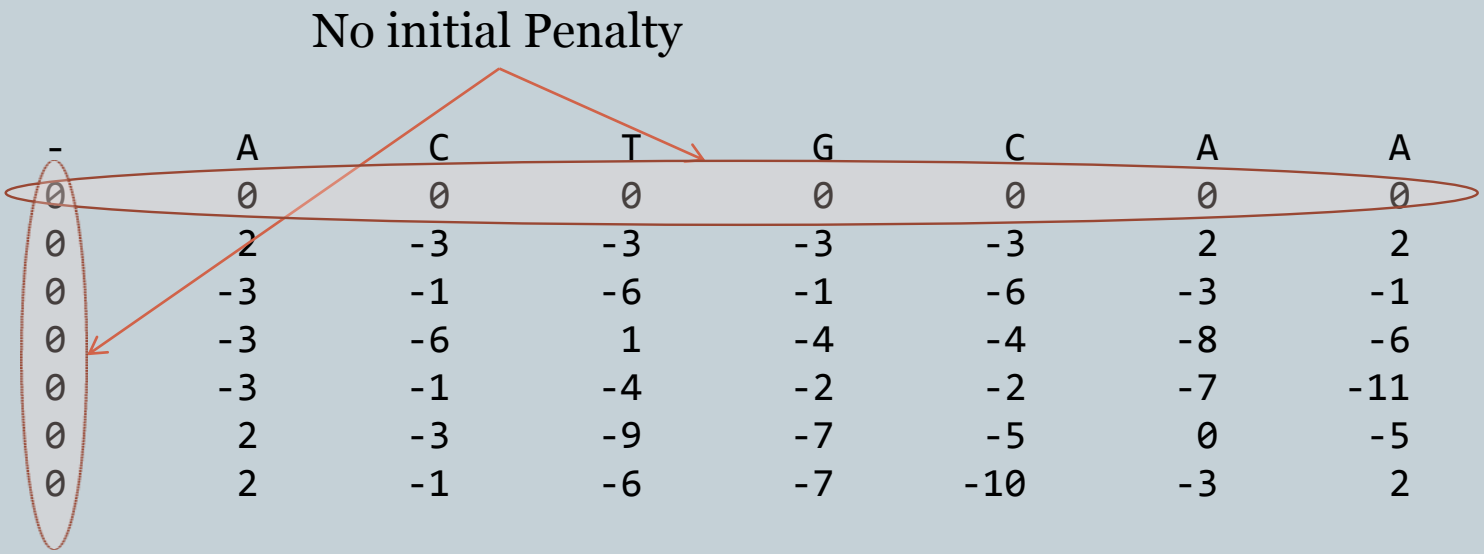
Introduction

Penalty from beginning

	-	A	C	T	G	C	A	A
-	0	-5	-10	-15	-20	-25	-30	-35
A	-5	2	-3	-8	-13	-18	-23	-28
G	-10	-3	-1	-6	-6	-11	-16	-21
T	-15	-8	-6	1	-4	-9	-14	-19
C	-20	-13	-6	-4	-2	-2	-7	-12
A	-25	-18	-11	-9	-7	-5	0	-5
A	-30	-23	-16	-14	-12	-10	-3	2

Match = 2, Mismatch = -3, Indel = -5

Introduction



	-	A	C	T	G	C	A	A
-	0	0	0	0	0	0	0	0
A	0	2	-3	-3	-3	-3	2	2
G	0	-3	-1	-6	-1	-6	-3	-1
T	0	-3	-6	1	-4	-4	-8	-6
C	0	-3	-1	-4	-2	-2	-7	-11
A	0	2	-3	-9	-7	-5	0	-5
A	0	2	-1	-6	-7	-10	-3	2

Match = 2, Mismatch = -3, Indel = -5

Needleman-Wunsch Alignment

	-	A	C	T	G	C	A	A
-	0	-5	-10	-15	-20	-25	-30	-35
A	-5							
G	-10							
T	-15							
C	-20							
A	-25							
A	-30							

Match = 2, Mismatch = -3, Indel = -5

Needleman-Wunsch Alignment

	-	A	C	T	G	C	A	A
-	0	-5	-10	-15	-20	-25	-30	-35
A	-5	2						
G	-10							
T	-15							
C	-20							
A	-25							
A	-30							

Match = 2, Mismatch = -3, Indel = -5

Needleman-Wunsch Alignment

	-	A	C	T	G	C	A	A
-	0	-5	-10	-15	-20	-25	-30	-35
A	-5	2	-3					
G	-10							
T	-15							
C	-20							
A	-25							
A	-30							

Match = 2, Mismatch = -3, Indel = -5

Needleman-Wunsch Alignment

	-	A	C	T	G	C	A	A
-	0	-5	-10	-15	-20	-25	-30	-35
A	-5	2	-3	-8				
G	-10							
T	-15							
C	-20							
A	-25							
A	-30							

Match = 2, Mismatch = -3, Indel = -5

Needleman-Wunsch Alignment

	-	A	C	T	G	C	A	A
-	0	-5	-10	-15	-20	-25	-30	-35
A	-5	2	-3	-8	-13			
G	-10							
T	-15							
C	-20							
A	-25							
A	-30							

Match = 2, Mismatch = -3, Indel = -5

Needleman-Wunsch Alignment

	-	A	C	T	G	C	A	A
-	0	-5	-10	-15	-20	-25	-30	-35
A	-5	2	-3	-8	-13	-18		
G	-10							
T	-15							
C	-20							
A	-25							
A	-30							

Match = 2, Mismatch = -3, Indel = -5

Needleman-Wunsch Alignment

	-	A	C	T	G	C	A	A
-	0	-5	-10	-15	-20	-25	-30	-35
A	-5	2	-3	-8	-13	-18	-23	
G	-10							
T	-15							
C	-20							
A	-25							
A	-30							

Match = 2, Mismatch = -3, Indel = -5

Needleman-Wunsch Alignment

	-	A	C	T	G	C	A	A
-	0	-5	-10	-15	-20	-25	-30	-35
A	-5	2	-3	-8	-13	-18	-23	-28
G	-10							
T	-15							
C	-20							
A	-25							
A	-30							

Match = 2, Mismatch = -3, Indel = -5

Needleman-Wunsch Alignment

	-	A	C	T	G	C	A	A
-	0	-5	-10	-15	-20	-25	-30	-35
A	-5	2	-3	-8	-13	-18	-23	-28
G	-10	-3						
T	-15							
C	-20							
A	-25							
A	-30							

Match = 2, Mismatch = -3, Indel = -5

Needleman-Wunsch Alignment

	-	A	C	T	G	C	A	A
-	0	-5	-10	-15	-20	-25	-30	-35
A	-5	2	-3	-8	-13	-18	-23	-28
G	-10	-3	-1					
T	-15							
C	-20							
A	-25							
A	-30							

Match = 2, Mismatch = -3, Indel = -5

Needleman-Wunsch Alignment

	-	A	C	T	G	C	A	A
-	0	-5	-10	-15	-20	-25	-30	-35
A	-5	2	-3	-8	-13	-18	-23	-28
G	-10	-3	-1	-6				
T	-15							
C	-20							
A	-25							
A	-30							

Match = 2, Mismatch = -3, Indel = -5

Needleman-Wunsch Alignment

	-	A	C	T	G	C	A	A
-	0	-5	-10	-15	-20	-25	-30	-35
A	-5	2	-3	-8	-13	-18	-23	-28
G	-10	-3	-1	-6	-6			
T	-15							
C	-20							
A	-25							
A	-30							

Match = 2, Mismatch = -3, Indel = -5

Needleman-Wunsch Alignment

	-	A	C	T	G	C	A	A
-	0	-5	-10	-15	-20	-25	-30	-35
A	-5	2	-3	-8	-13	-18	-23	-28
G	-10	-3	-1	-6	-6	-11		
T	-15							
C	-20							
A	-25							
A	-30							

Match = 2, Mismatch = -3, Indel = -5

Needleman-Wunsch Alignment

	-	A	C	T	G	C	A	A
-	0	-5	-10	-15	-20	-25	-30	-35
A	-5	2	-3	-8	-13	-18	-23	-28
G	-10	-3	-1	-6	-6	-11	-16	
T	-15							
C	-20							
A	-25							
A	-30							

Match = 2, Mismatch = -3, Indel = -5

Needleman-Wunsch Alignment

	-	A	C	T	G	C	A	A
-	0	-5	-10	-15	-20	-25	-30	-35
A	-5	2	-3	-8	-13	-18	-23	-28
G	-10	-3	-1	-6	-6	-11	-16	-21
T	-15							
C	-20							
A	-25							
A	-30							

Match = 2, Mismatch = -3, Indel = -5

Bit-parallel alignment

	-	A	C	T	G	C	A	A
-	0	-5	-10	-15	-20	-25	-30	-35
A	-5							
G	-10							
T	-15							
C	-20							
A	-25							
A	-30							

Match = 2, Mismatch = -3, Indel = -5

Integer Scores

Bit-parallel alignment

	-	A	C	T	G	C	A	A
-	0	-5	-10	-15	-20	-25	-30	-35
A	-5	2	-3	-8	-13	-18	-23	-28
G	-10							
T	-15							
C	-20							
A	-25							
A	-30							

Match = 2, Mismatch = -3, Indel = -5

Integer Scores

Bit-parallel alignment

	-	A	C	T	G	C	A	A
-	0	-5	-10	-15	-20	-25	-30	-35
A	-5	2	-3	-8	-13	-18	-23	-28
G	-10	-3	-1	-6	-6	-11	-16	-21
T	-15							
C	-20							
A	-25							
A	-30							

Match = 2, Mismatch = -3, Indel = -5

Integer Scores

Bit-parallel alignment

		-	A	C	T	G	C	A	A
	-	0	-5	-10	-15	-20	-25	-30	-35
A		-5	2	-3	-8	-13	-18	-23	-28
G		-10	-3	-1	-6	-6	-11	-16	-21
T		-15	-8	-6	1	-4	-9	-14	-19
C		-20							
A		-25							
A		-30							

Match = 2, Mismatch = -3, Indel = -5

Integer Scores

Bit-parallel alignment

	-	A	C	T	G	C	A	A
-	0	-5	-10	-15	-20	-25	-30	-35
A	-5	2	-3	-8	-13	-18	-23	-28
G	-10	-3	-1	-6	-6	-11	-16	-21
T	-15	-8	-6	1	-4	-9	-14	-19
C	-20	-13	-6	-4	-2	-2	-7	-12
A	-25							
A	-30							

Match = 2, Mismatch = -3, Indel = -5

Integer Scores

Bit-parallel alignment

	-	A	C	T	G	C	A	A
-	0	-5	-10	-15	-20	-25	-30	-35
A	-5	2	-3	-8	-13	-18	-23	-28
G	-10	-3	-1	-6	-6	-11	-16	-21
T	-15	-8	-6	1	-4	-9	-14	-19
C	-20	-13	-6	-4	-2	-2	-7	-12
A	-25	-18	-11	-9	-7	-5	0	-5
A	-30							

Match = 2, Mismatch = -3, Indel = -5

Integer Scores

Bit-parallel alignment

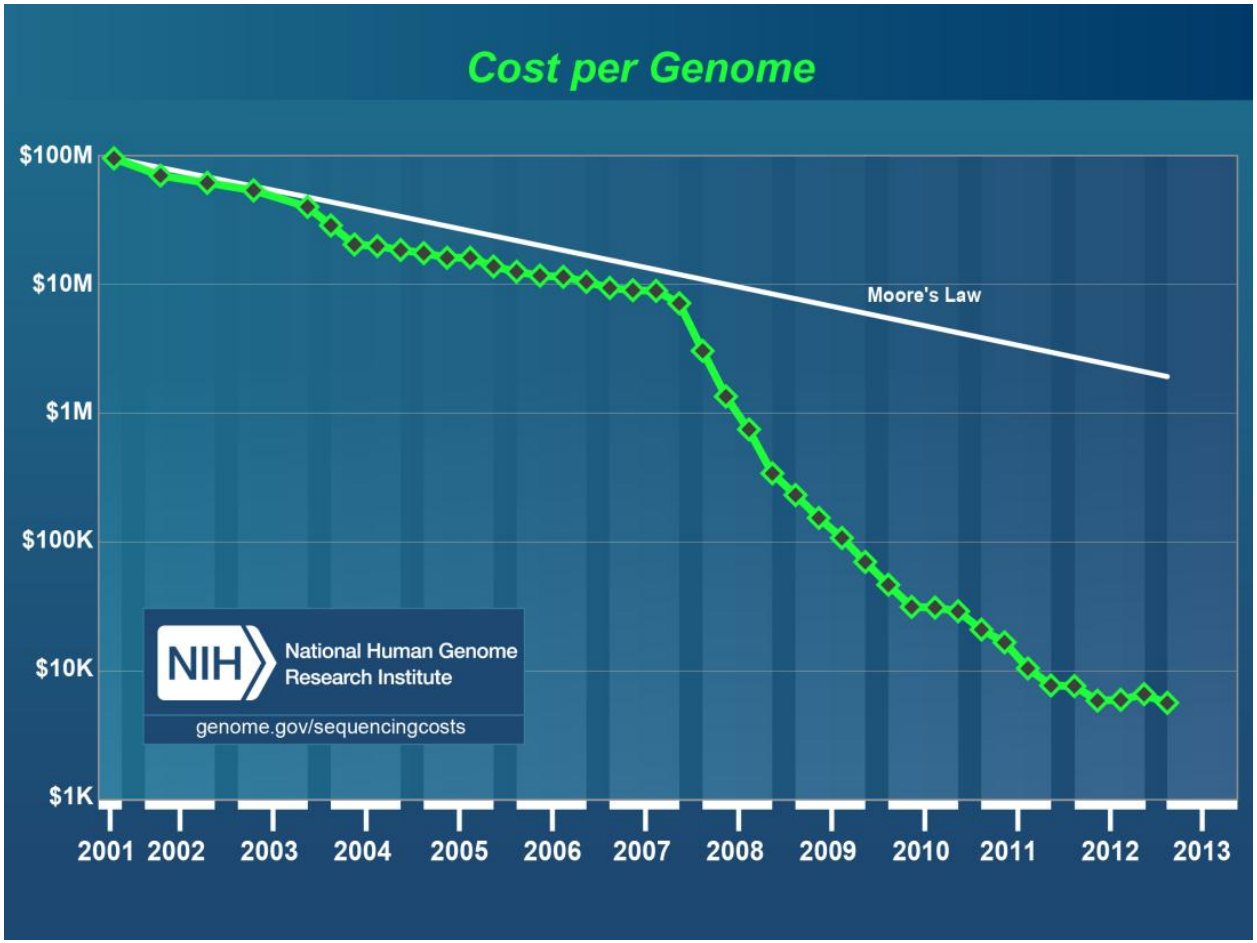
	-	A	C	T	G	C	A	A
-	0	-5	-10	-15	-20	-25	-30	-35
A	-5	2	-3	-8	-13	-18	-23	-28
G	-10	-3	-1	-6	-6	-11	-16	-21
T	-15	-8	-6	1	-4	-9	-14	-19
C	-20	-13	-6	-4	-2	-2	-7	-12
A	-25	-18	-11	-9	-7	-5	0	-5
A	-30	-23	-16	-14	-12	-10	-3	2

Match = 2, Mismatch = -3, Indel = -5

Integer Scores

Motivation

- Cheaper sequencing of DNA means that larger datasets are being generated
- Sequence analysis of such large datasets can be accelerated by faster alignment algorithms



Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcosts. Accessed June 10, 2013.

Earlier Bit-parallel Pattern Matching Algorithms

- Longest Common Subsequence (LCS) (Allison & Dix, 1986; Crochemore et al, 2001; Hyyro, 2004)
- Unit-cost edit-distance (Myers, 99; Hyyro et al, 2005)
- K-differences (*agrep*; Wu-Manber, 92)
- Regular expression search (Navarro, 04)
- Arbitrary weights edit-distance (Bergeron&Hamel, 02)

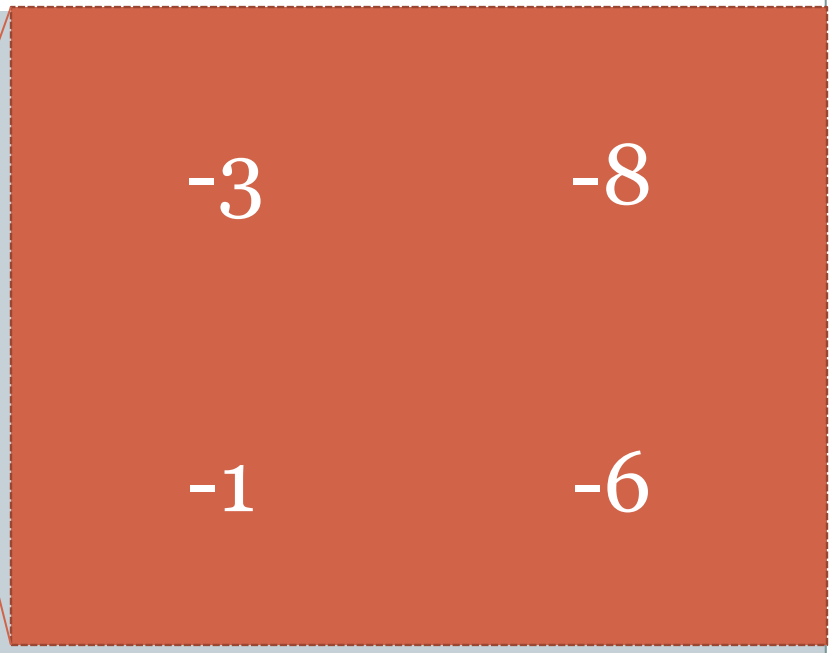
Algorithm Foundation

	-	A	C	T	G	C	A	A
-	0	-5	-10	-15	-20	-25	-30	-35
A	-5	2	-3	-8	-13	-18	-23	-28
G	-10	-3	-1	-6	-6	-11	-16	-21
T	-15	-8	-6	1	-4	-9	-14	-19
C	-20	-13	-6	-4	-2	-2	-7	-12
A	-25	-18	-11	-9	-7	-5	0	-5
A	-30	-23	-16	-14	-12	-10	-3	2

Match = 2, Mismatch = -3, Indel = -5

Algorithm Foundation

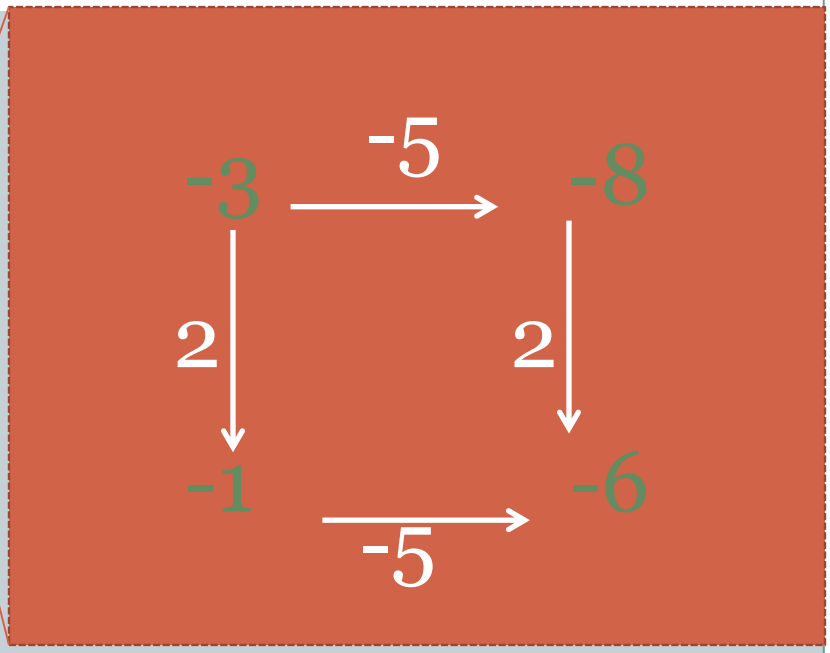
	-	A	C	T
-	0	-5	-10	-15
A	-5	2	-3	-8
G	-10	-3	-1	-6
T	-15	-8	-6	1
C	-20	-13	-6	-4
A	-25	-18	-11	-9
A	-30	-23	-16	-14



Match = 2, Mismatch = -3, Indel = -5

Algorithm Foundation

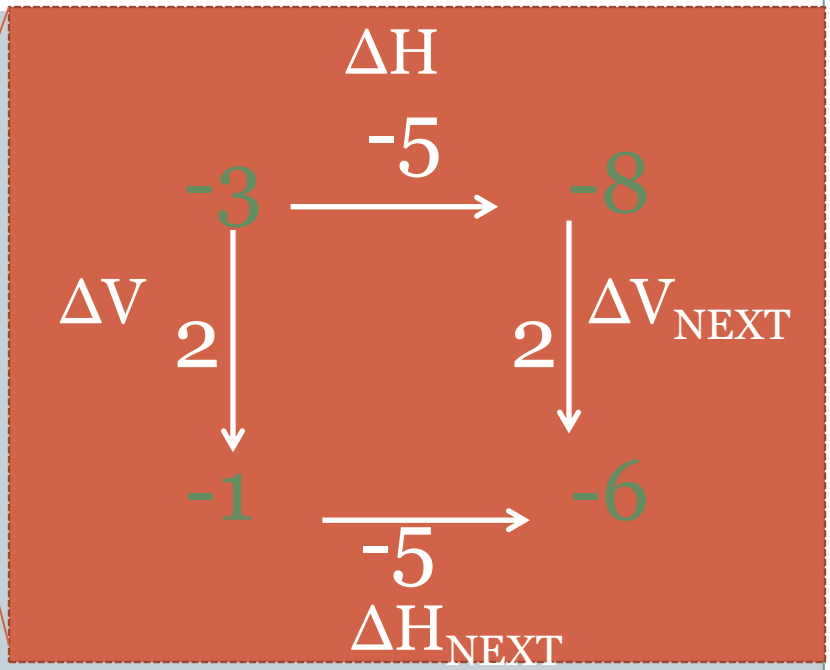
	-	A	C	T
-	0	-5	-10	-15
A	-5	2	-3	-8
G	-10	-3	-1	-6
T	-15	-8	-6	1
C	-20	-13	-6	-4
A	-25	-18	-11	-9
A	-30	-23	-16	-14



Match = 2, Mismatch = -3, Indel = -5

Algorithm Foundation

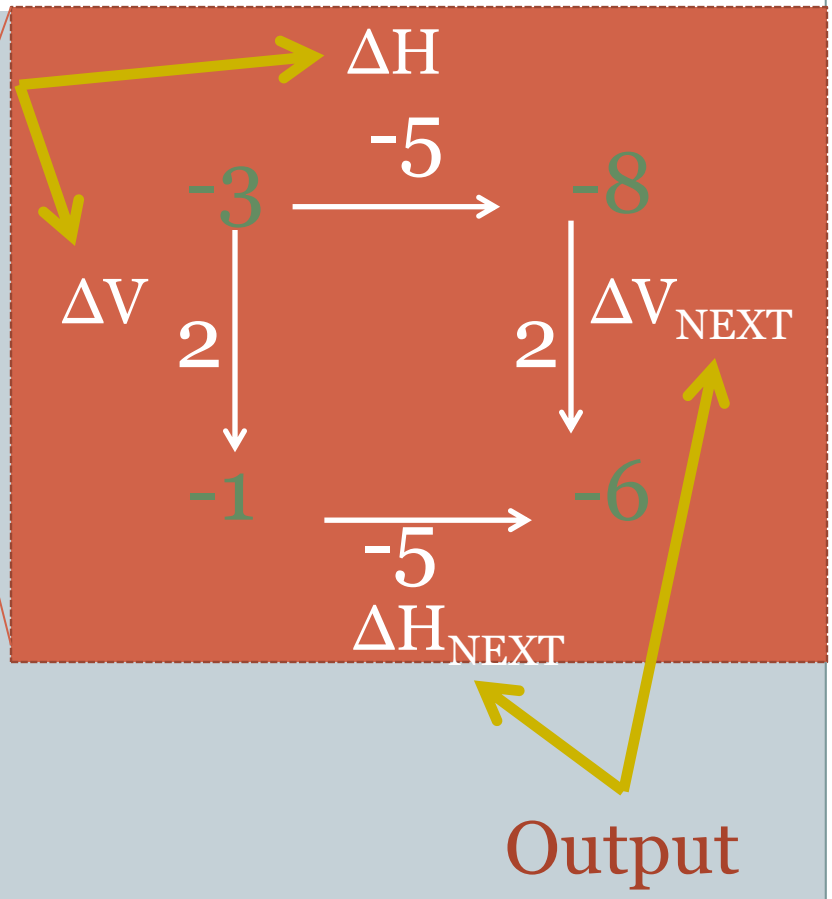
	-	A	C	T
-	0	-5	-10	-15
A	-5	2	-3	-8
G	-10	-3	-1	-6
T	-15	-8	-6	1
C	-20	-13	-6	-4
A	-25	-18	-11	-9
A	-30	-23	-16	-14



Match = 2, Mismatch = -3, Indel = -5

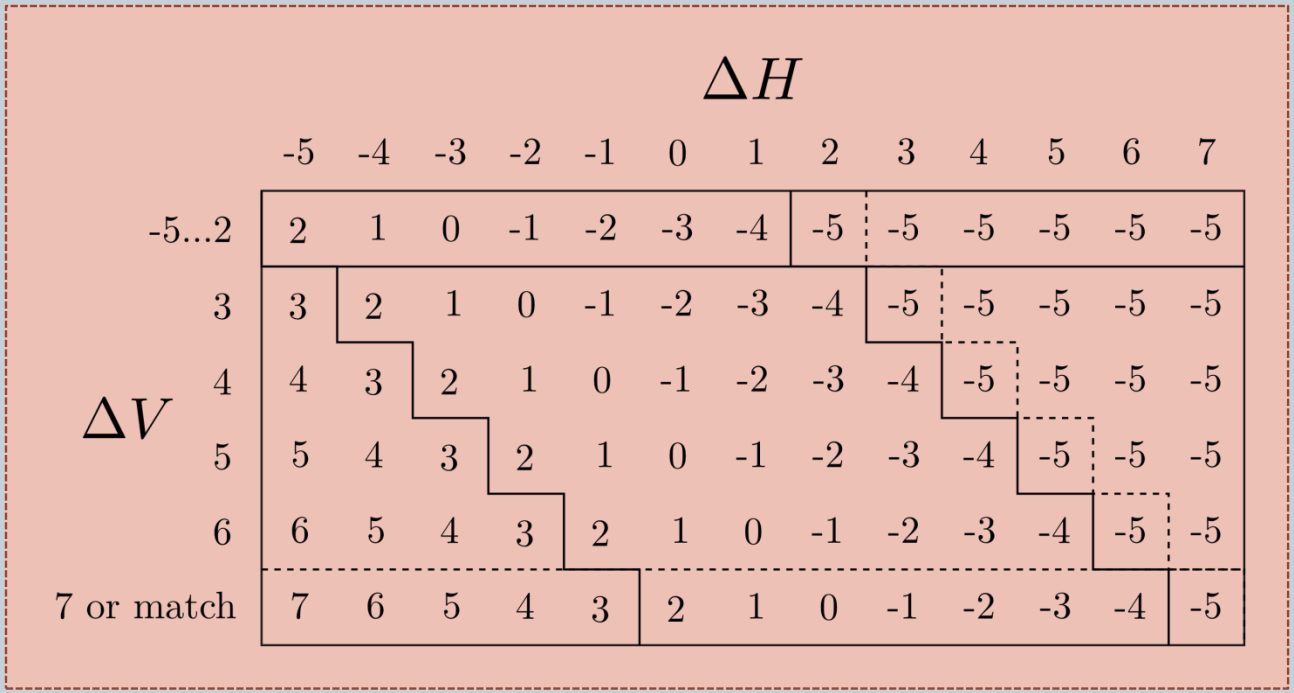
Algorithm Foundation

	-	A	C	T
-	0	-5	-10	-15
A	-5	2	-3	-8
G	-10	-3	-1	-6
T	-15	-8	-6	1
C	-20	-13	-6	-4
A	-25	-18	-11	-9
A	-30	-23	-16	-14



Match = 2, Mismatch = -3, Indel = -5

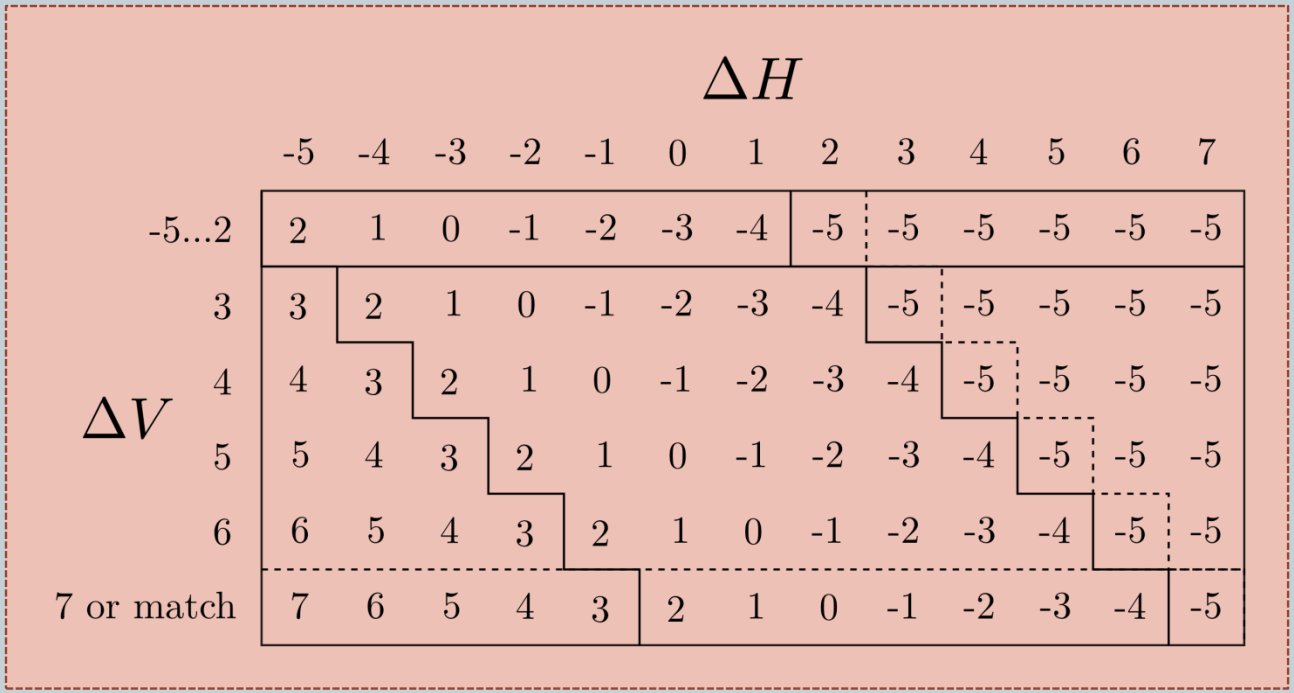
Function Table



ΔV \ ΔH	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7
-5...2	2	1	0	-1	-2	-3	-4	-5	-5	-5	-5	-5	-5
3	3	2	1	0	-1	-2	-3	-4	-5	-5	-5	-5	-5
4	4	3	2	1	0	-1	-2	-3	-4	-5	-5	-5	-5
5	5	4	3	2	1	0	-1	-2	-3	-4	-5	-5	-5
6	6	5	4	3	2	1	0	-1	-2	-3	-4	-5	-5
7 or match	7	6	5	4	3	2	1	0	-1	-2	-3	-4	-5

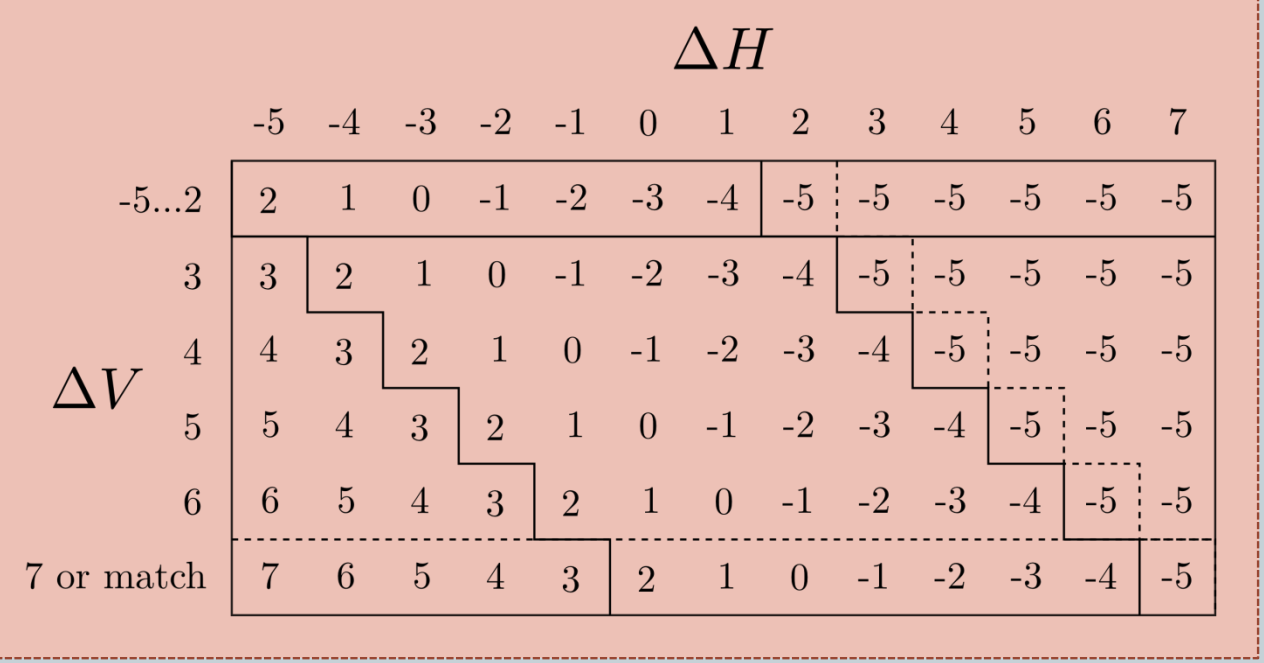
ΔV_{NEXT} output values given ΔV and ΔH input values

What is the range of differences?



ΔV \ ΔH	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7
-5...2	2	1	0	-1	-2	-3	-4	-5	-5	-5	-5	-5	-5
3	3	2	1	0	-1	-2	-3	-4	-5	-5	-5	-5	-5
4	4	3	2	1	0	-1	-2	-3	-4	-5	-5	-5	-5
5	5	4	3	2	1	0	-1	-2	-3	-4	-5	-5	-5
6	6	5	4	3	2	1	0	-1	-2	-3	-4	-5	-5
7 or match	7	6	5	4	3	2	1	0	-1	-2	-3	-4	-5

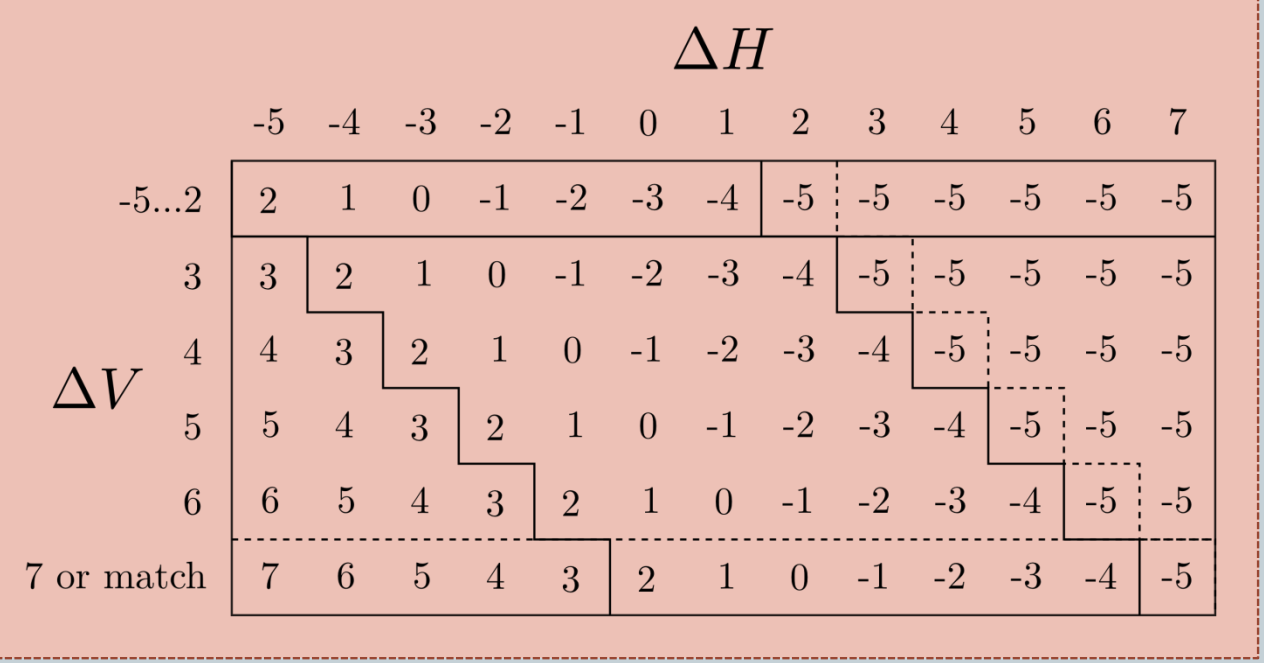
What is the range of differences?



ΔV \ ΔH	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7
-5...2	2	1	0	-1	-2	-3	-4	-5	-5	-5	-5	-5	-5
3	3	2	1	0	-1	-2	-3	-4	-5	-5	-5	-5	-5
4	4	3	2	1	0	-1	-2	-3	-4	-5	-5	-5	-5
5	5	4	3	2	1	0	-1	-2	-3	-4	-5	-5	-5
6	6	5	4	3	2	1	0	-1	-2	-3	-4	-5	-5
7 or match	7	6	5	4	3	2	1	0	-1	-2	-3	-4	-5

Match = 2,
Mismatch = -3,
Indel = -5

What is the range of differences?

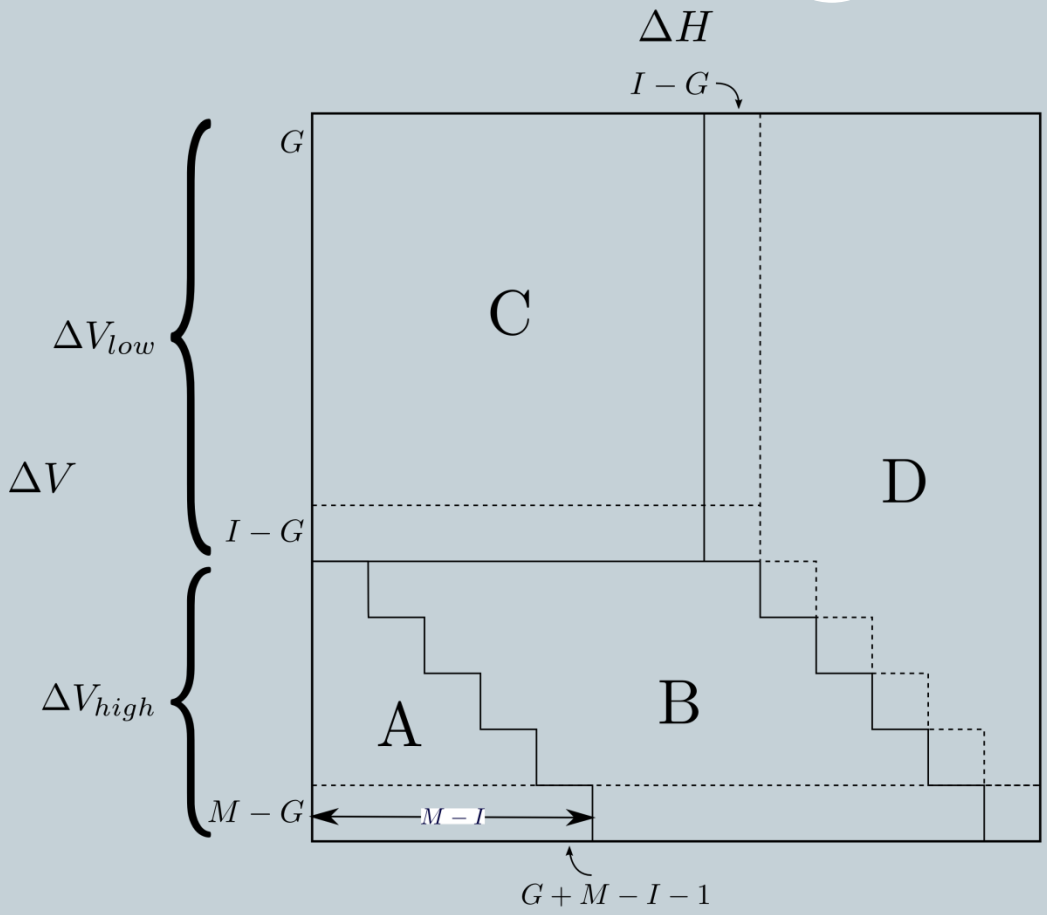


ΔV \ ΔH	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7
-5...2	2	1	0	-1	-2	-3	-4	-5	-5	-5	-5	-5	-5
3	3	2	1	0	-1	-2	-3	-4	-5	-5	-5	-5	-5
4	4	3	2	1	0	-1	-2	-3	-4	-5	-5	-5	-5
5	5	4	3	2	1	0	-1	-2	-3	-4	-5	-5	-5
6	6	5	4	3	2	1	0	-1	-2	-3	-4	-5	-5
7 or match	7	6	5	4	3	2	1	0	-1	-2	-3	-4	-5

Match = 2,
Mismatch = -3,
Indel = -5

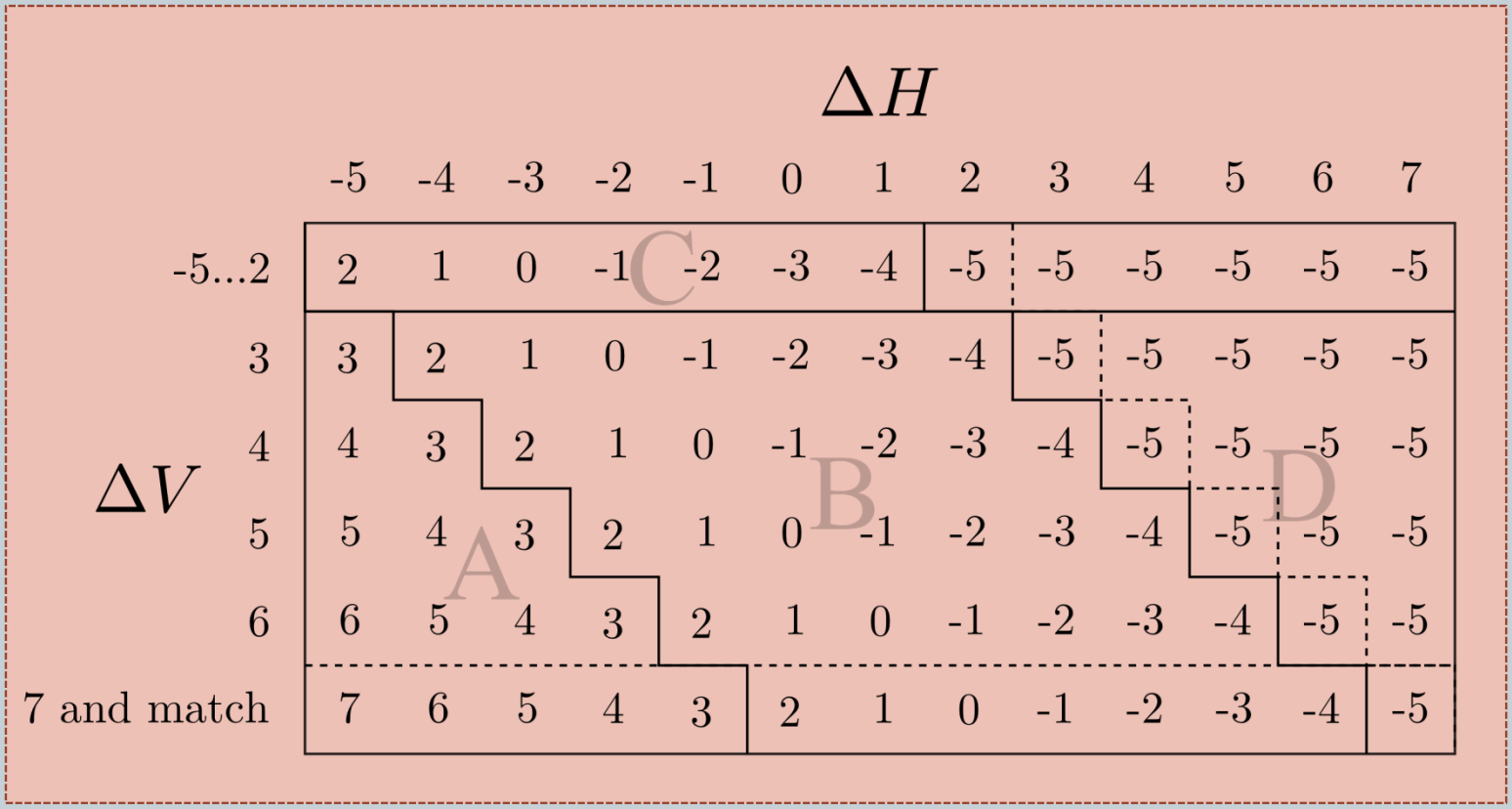
Minimum Value = Indel = -5
Maximum Value = Match – Indel = 2 - (- 5) = 7

Generalized Function Table



M = match score
I = mismatch score
G = indel (gap) penalty

Zones in Example Function Table



Bit-parallel Representation

- Bit-vectors: computer words 64 bits long
- A bit-vector for each possible difference, both horizontally and vertically (ΔV and ΔH)
- A set of Match vectors (MatchA, MatchC, MatchG, MatchT in the DNA case)
- We keep track of match positions because they are a special case in the function table.

Example ΔH Bit-vector Storage

ΔH values

	-	A	C	T	G	C	A	A
C	-20	7 → -13	7 → -6	2 → -4	2 → -2	0 → -2	-5 → -7	-5 → -12

ΔH Bit-Vectors

7	1	1	0	0	0	0	0
6	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0
2	0	0	1	1	0	0	0
1	0	0	0	0	0	0	0
0	0	0	0	0	1	0	0
-1	0	0	0	0	0	0	0
-2	0	0	0	0	0	0	0
-3	0	0	0	0	0	0	0
-4	0	0	0	0	0	0	0
-5	0	0	0	0	0	1	1

Example Match Vectors

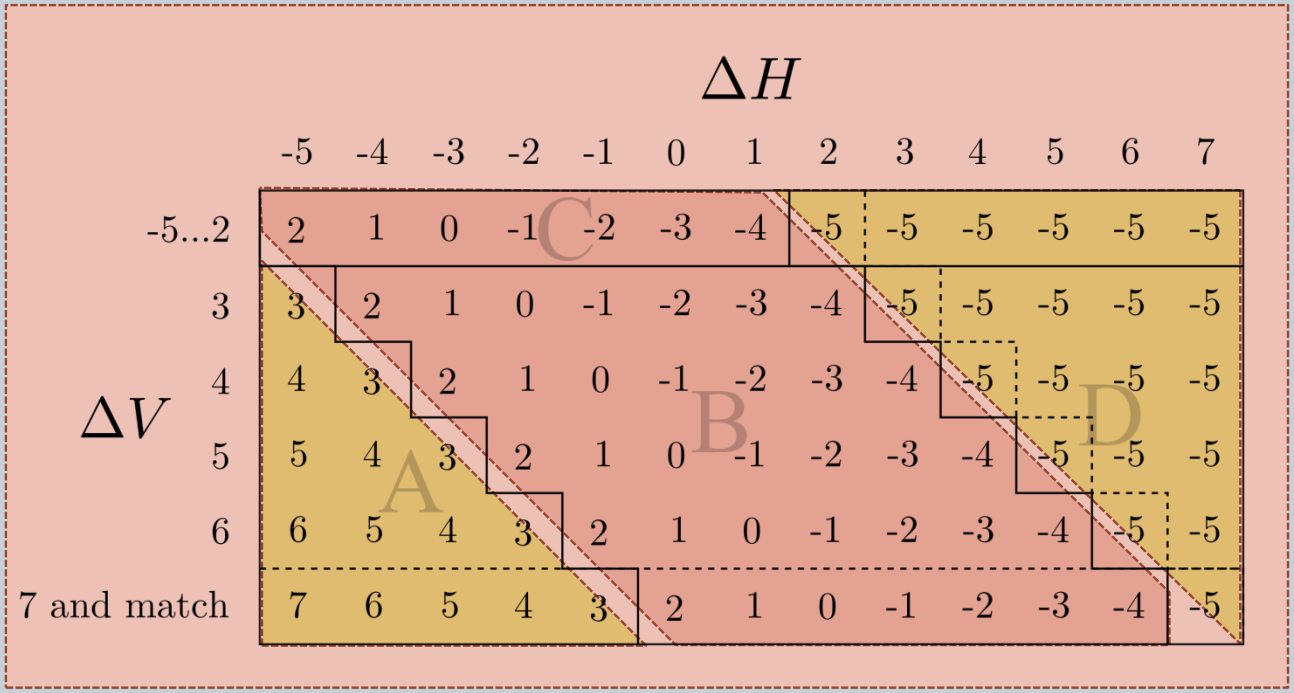
	-	A	C	T	G	C	A	A
-	0	-5	-10	-15	-20	-25	-30	-35
A	-5	2	-3	-8	-13	-18	-23	-28
G	-10	-3	-1	-6	-6	-11	-16	-21
T	-15	-8	-6	1	-4	-9	-14	-19
C	-20	-13	-6	-4	-2	-2	-7	-12
A	-25	-18	-11	-9	-7	-5	0	-5
A	-30	-23	-16	-14	-12	-10	-3	2

Match Vectors

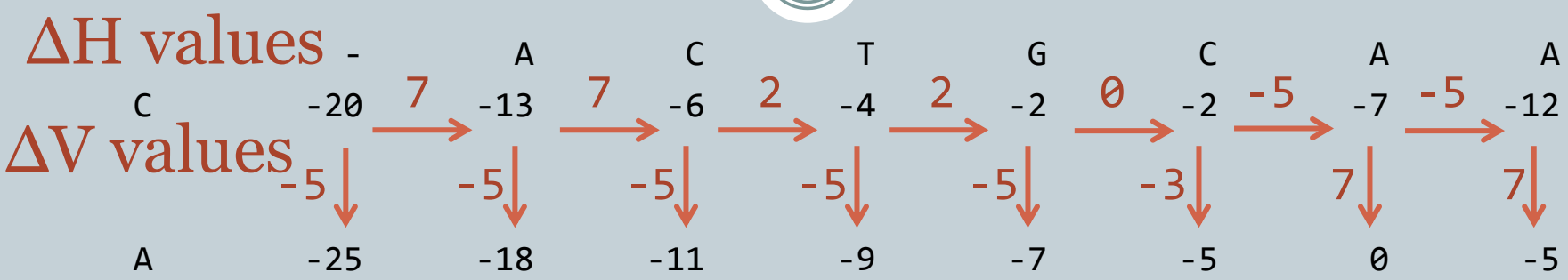
MatchesA	1	0	0	0	0	1	1
MatchesC	0	1	0	0	1	0	0
MatchesT	0	0	1	0	0	0	0
MatchesG	0	0	0	1	0	0	0

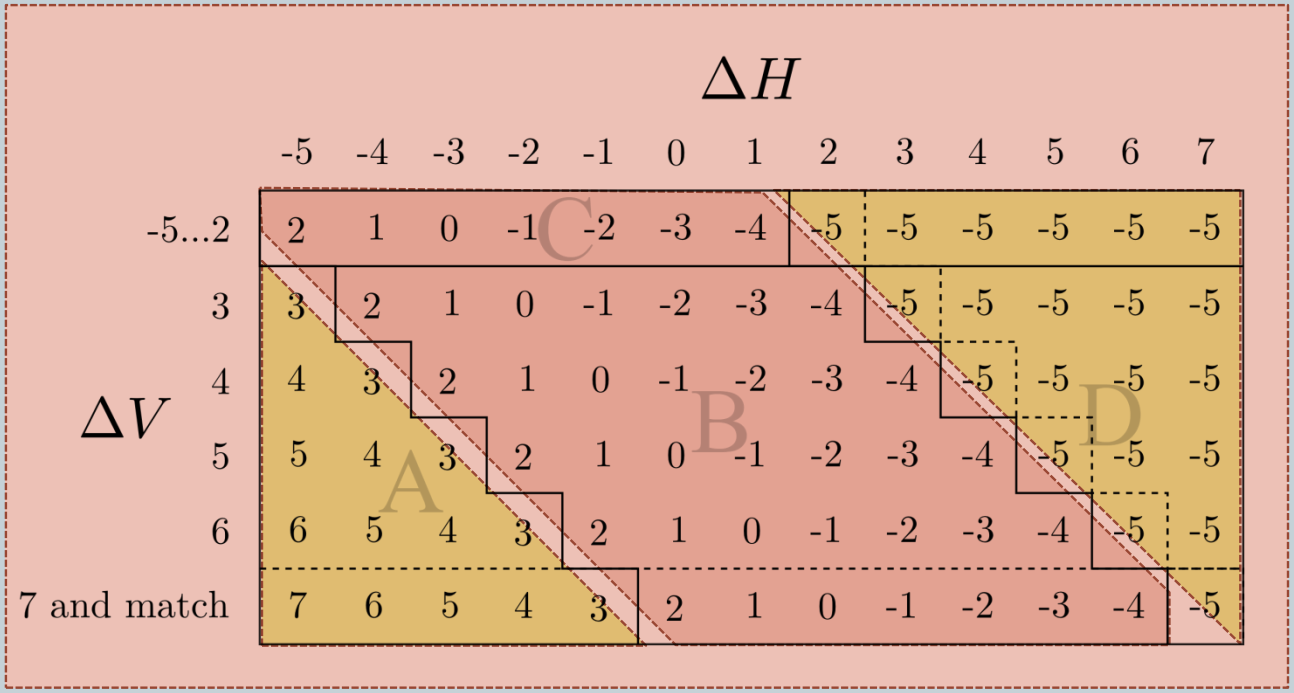
Algorithm

- Start with ΔH values
- Compute ΔV values
- Then compute the new ΔH values



Algorithm: Example





Time Complexity

$$O\left(zn\frac{m}{w}\right)$$

where

n = |Sequence Y|

m = |Sequence X|

w = length of computer word

$$z = \frac{(M - 2G + 1)^2 - (I - 2G)^2}{2}$$

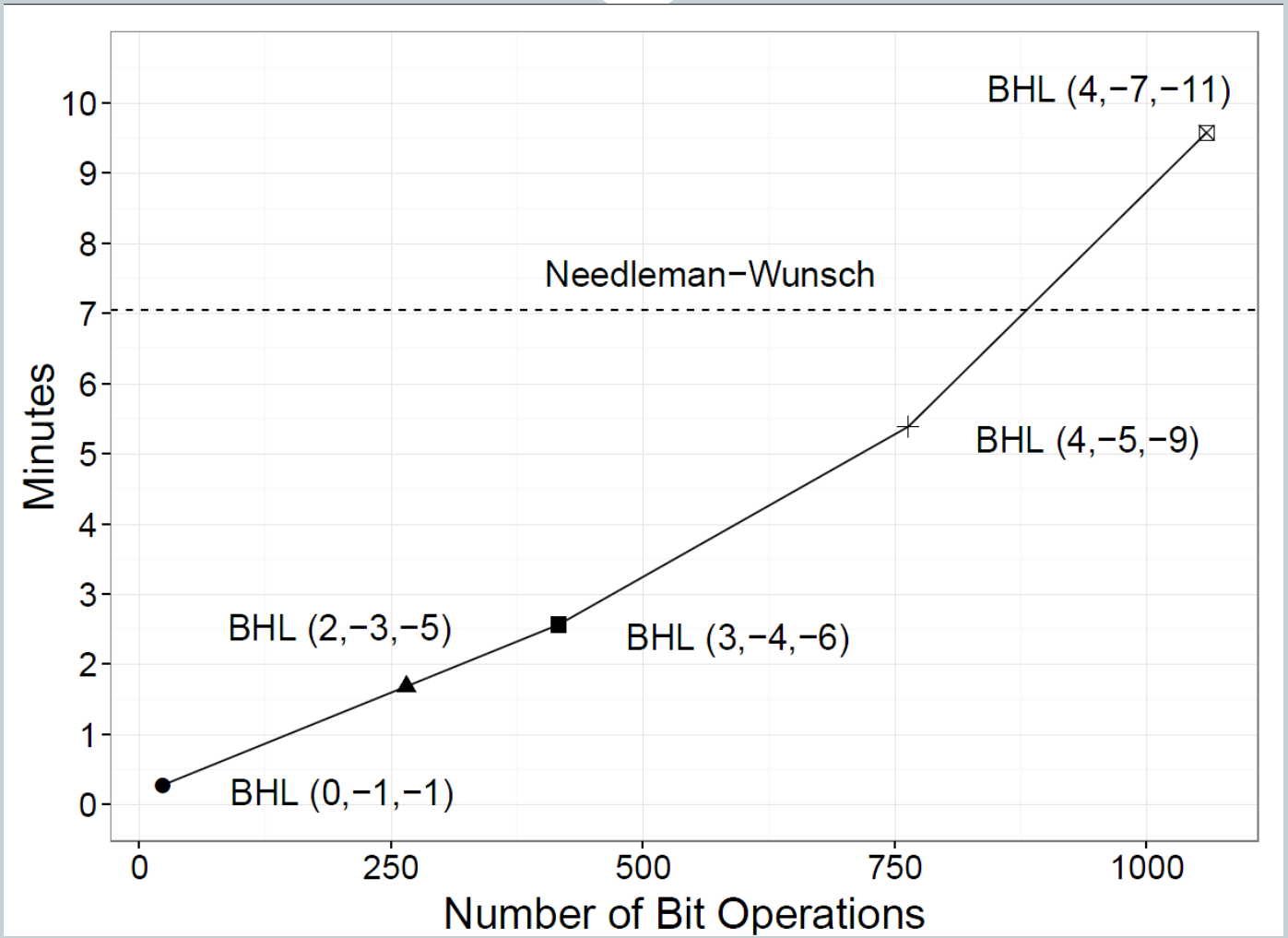
Implementation

- Python script that generates C code based on input parameters (M; I; G)
- Will eventually have web page for download of code

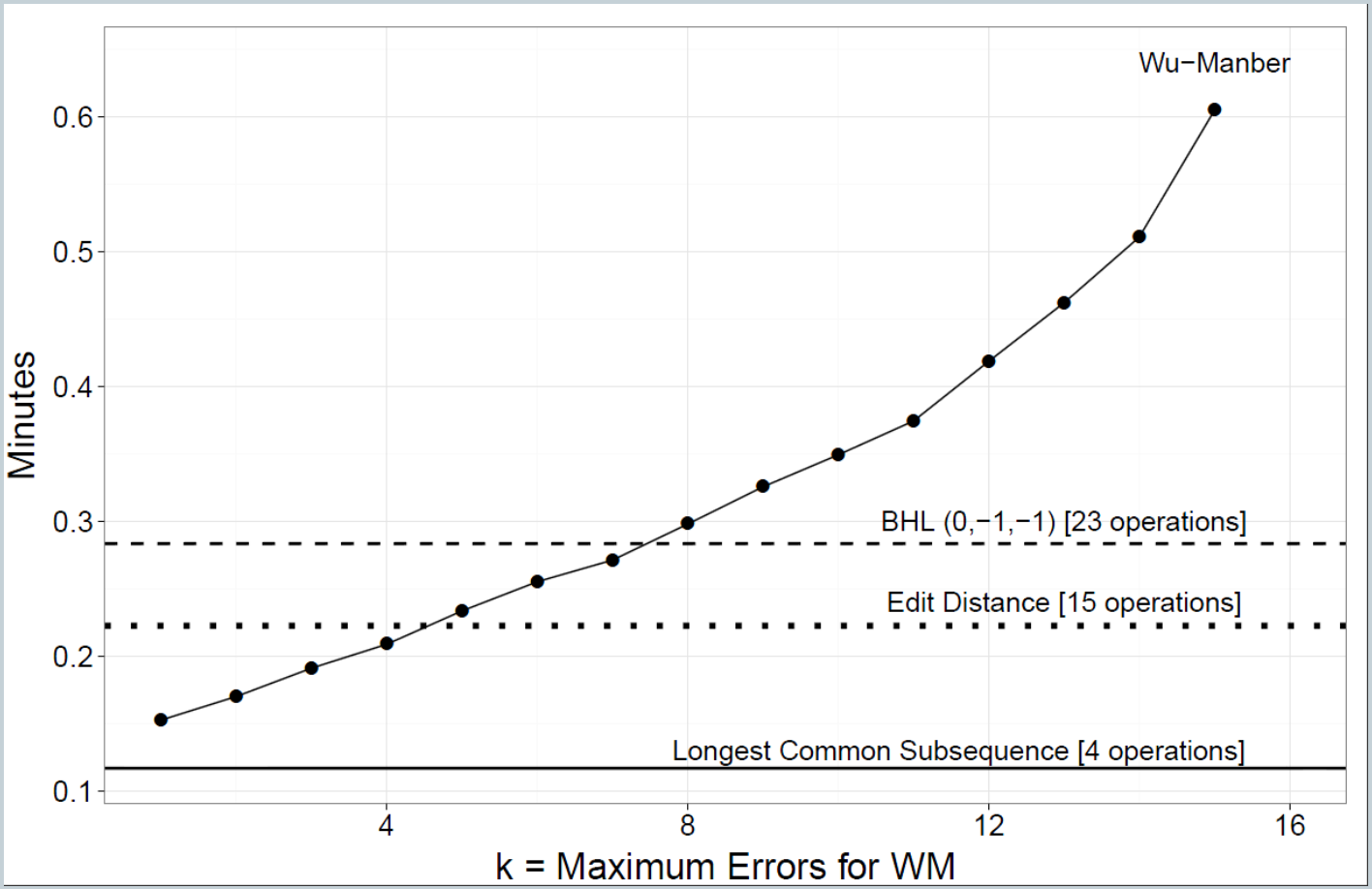
Experimental Analysis

- Compared BHL with
  - Wu-Manber K-differences algorithm
  - Unit cost edit distance bit-parallel algorithm
  - Longest Common Subsequence bit-parallel algorithm
  - Needleman-Wunsch dynamic programming algorithm
- Computed 25 million alignments with each program
- Each DNA sequence was 63 bases long
- All programs compiled using GCC, optimization level O3
- Computation done on a typical desktop computer

Results: Comparison to NW algorithm



Results: comparison to bit-parallel algorithms



Current and Future Work

- Implementation for sequences longer than one word
- Single Instruction Multiple Data (SIMD) implementation
- BLOSUM and PAM type substitution matrix support
- General Purpose Graphics Processing Unit (GPGPU) implementation
- New bit-parallel representations for greater speed and compactness of data

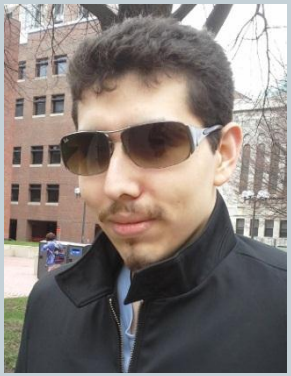
Acknowledgements

My advisor, Dr. Gary Benson

Lab members

Yevgeniy Gelfand

Yozen Hernandez



Funded by the National Science Foundation (NSF)

Questions