Dynamic Programming

- Sequence Analysis
- The Sequence Alignment Problem
- Dynamic Programming Formulation

Imdad ullah Khan

Sequence Alignment

Sequence Alignment serves as a (constructive) similarity/distance Measure between sequences

It has many applications

- Checking minor difference between two similar sequences
- Analyzing polymorphisms (ex. SNPs) between closely related sequences
- Applications – Prediction of gene function, prediction of protein structure
- Search for common patterns of characters

Pairwise Sequence Alignment

Pairwise Sequence Alignment is a (constructive) similarity/distance Measure between two sequences

- Establish pairwise correspondence between related sequences

  ▷ Constructive

- Pairwise alignment is the basis for database searching (e.g., BLAST)
- Multiple Sequence Alignment

Sequence Alignment: Indels

Sequences can diverge from a common ancestor through various types of mutations Edit Operations:

Let $\Sigma^* = \Sigma \cup \{-\}$ for $- \notin \Sigma$ (a special symbol representing empty string)

An edit operation is a triplet (x, i, y) with $(x, y) \in \Sigma^* \times \Sigma^*$ and integer i

Are Gaps Biologically Justified?

- Indels of various sizes can occur in one sequence relative to the other
- For instance, the shortening of a polypeptide chain in a protein
- Substitution actually are two indels

Sequence Alignment: Definition

An alignment of two strings X and Y is a pair of strings X' and Y' (X and Y but with inserted dashes) such that

1. $|X'| = |Y'|$
2. Removing all dashes leaves X and Y
3. For all $1 \leq i \leq |X'|$, either $X'_i \neq -$ or $Y'_i \neq -$

$$\overset{1}{A}\ \overset{2}{C}\ \overset{3}{C}\ \overset{4}{G}\ \overset{5}{A}\ \overset{6}{T}\ \overset{7}{G} = X \qquad Y = \overset{1}{T}\ \overset{2}{C}\ \overset{3}{A}\ \overset{4}{G}\ \overset{5}{T}\ \overset{6}{G}$$

$A\ C\ -\ C\ G\ A\ T\ G$ $T\ C\ A\ -\ G\ -\ T\ G$	$A\ C\ C\ G\ A\ T\ G$ $T\ C\ A\ G\ -\ T\ G$	$A\ -\ C\ C\ G\ A\ T\ G$ $-\ T\ C\ A\ G\ -\ T\ G$

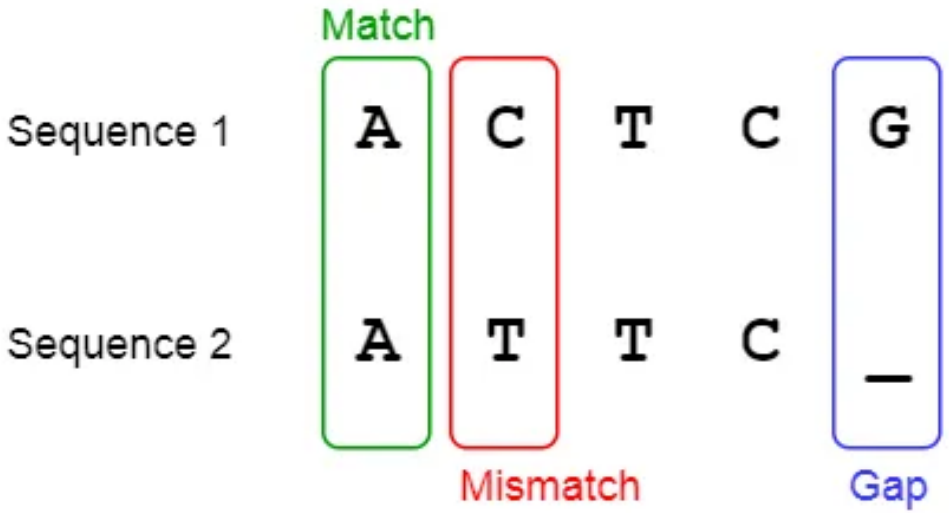
Three different alignment of string X and Y

Sequence Alignment: Definition

An alignment of two strings X and Y is a pair of strings X' and Y' (X and Y but with inserted dashes) such that

1. $|X'| = |Y'|$
2. Removing all dashes leaves X and Y
3. For all $1 \leq i \leq |X'|$, either $X'_i \neq -$ or $Y'_i \neq -$

Display sequence over another (with gaps inserted) to assess similarity

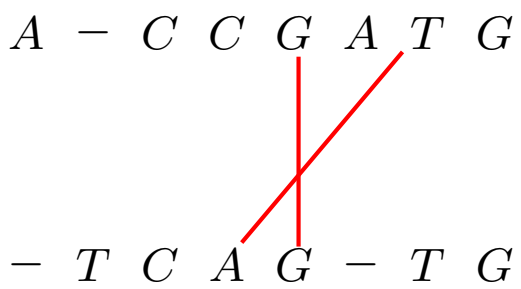


Sequence Alignment: Alternative Definition

An alignment P of two strings X and Y is a set of ordered pairs (x_i, y_j) such that

- each character of X and Y appears in at most one pair
- no two pairs are crossing

The pairs (x_i, y_j) and $(x_{i'}, y_{j'})$ cross if $i < i'$ and $j > j'$



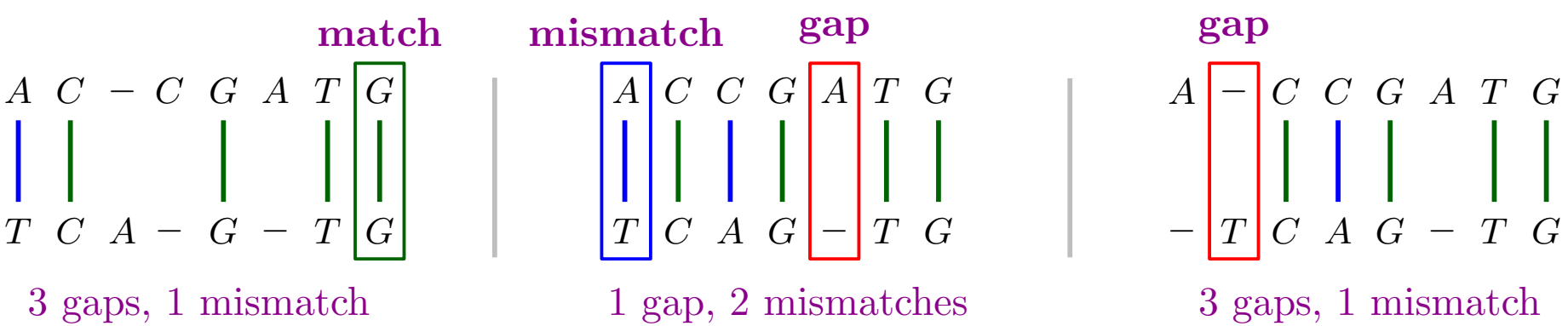
Sequence Alignment: Alternative Definition

An alignment P of two strings X and Y is a set of ordered pairs (x_i, y_j) such that

- each character of X and Y appears in at most one pair
- no two pairs are crossing

The pairs (x_i, y_j) and $(x_{i'}, y_{j'})$ cross if $i < i'$ and $j > j'$

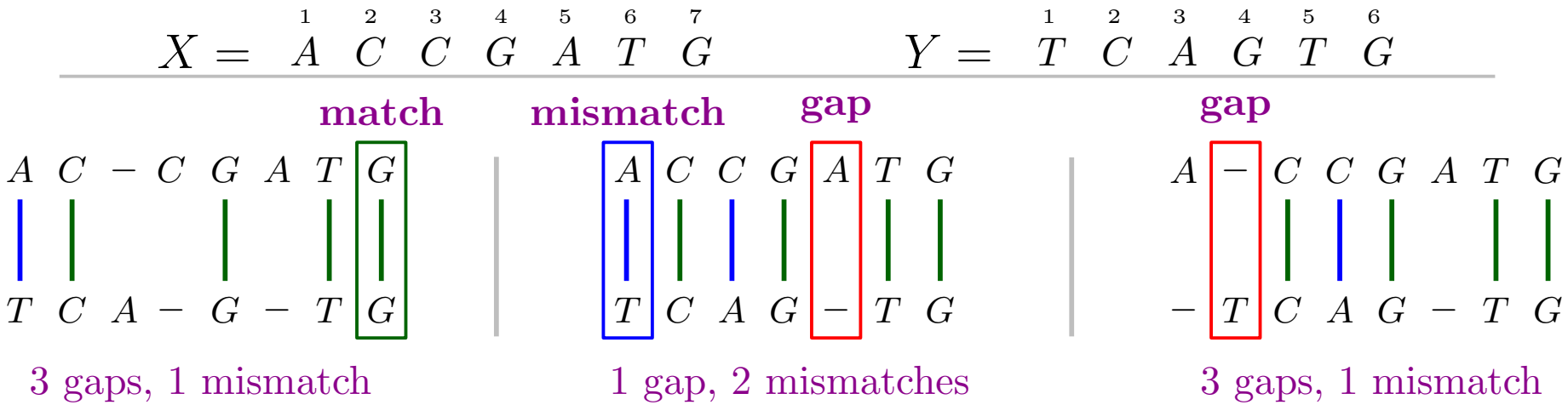
$X = \overset{1}{A}\ \overset{2}{C}\ \overset{3}{C}\ \overset{4}{G}\ \overset{5}{A}\ \overset{6}{T}\ \overset{7}{G}$ $\qquad$ $Y = \overset{1}{T}\ \overset{2}{C}\ \overset{3}{A}\ \overset{4}{G}\ \overset{5}{T}\ \overset{6}{G}$



Sequence Alignment: Goodness Measure

An alignment of two strings X and Y is a pair of strings X' and Y' (X and Y but with inserted dashes) such that

1. $|X'| = |Y'|$
2. Removing all dashes leaves X and Y
3. For all $1 \leq i \leq |X'|$, either $X'_i \neq -$ or $Y'_i \neq -$



Sequence Alignment: : Goodness Measure

An alignment of two strings X and Y is a pair of strings X' and Y' (X and Y but with inserted dashes) such that

1. $|X'| = |Y'|$
2. Removing all dashes leaves X and Y
3. For all $1 \leq i \leq |X'|$, either $X'_i \neq -$ or $Y'_i \neq -$

Introducing too many gaps can generate meaningless alignments

s----e------qu----en--ce
sometimesquipsentice

Need a goodness measure to find alignments making biological sense

A score to positively weigh match and negatively weigh gap and mismatch

For example: Each match gets $+1$, Mismatch gets -1, and gap gets -2

Sequence Alignment: Goodness Measure

- Some amino acids are more "exchangeable" than others (e.g., Serine and Threonine are more similar than Tryptophan and Alanine)
- However, mismatch costs are not usually used in aligning DNA or RNA sequences, because no substitution is "better" than any other (in general)
- A substitution matrix can be used to introduce "mismatch costs" for handling different (or preferred) types of substitutions

Sequence Alignment: Cost

An alignment of two strings X and Y is a pair of strings X' and Y' (X and Y but with inserted dashes) such that

1. $|X'| = |Y'|$
2. Removing all dashes leaves X and Y
3. For all $1 \leq i \leq |X'|$, either $X'_i \neq -$ or $Y'_i \neq -$

Mismatch and gap penalty matrix

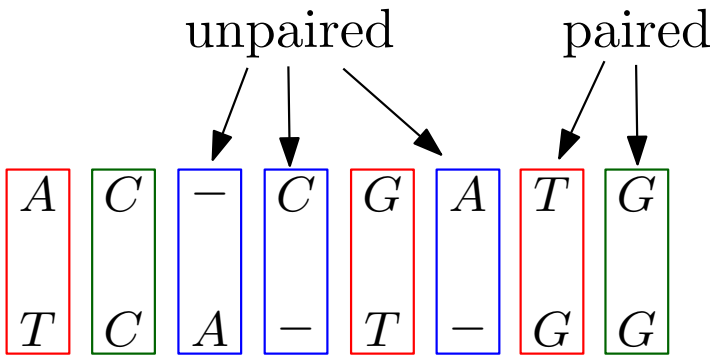
- Matches should be rewarded
- Mismatches could be (un)equally penalized
- Does not have to be symmetric matrix

α	A	C	T	G	$-$
A	0	4	1	4	2
C	4	0	3	2	2
T	1	3	0	3	2
G	4	2	3	0	2
$-$	2	2	2	2	$\ddots$

Sequence Alignment: Cost

An alignment P of two sequences X and Y is a pair of strings X' and Y' (X and Y but with inserted dashes) such that

1. $|X'| = |Y'|$
2. Removing all dashes leaves X and Y
3. For all $1 \leq i \leq |X'|$, either $X'_i \neq -$ or $Y'_i \neq -$



Penalty Matrix α

α	A	C	T	G	$-$
A	0	4	1	4	δ
C	4	0	3	2	δ
T	1	3	0	3	δ
G	4	2	3	0	δ
$-$	δ	δ	δ	δ	$\ddots$

$$Cost(P) = \underbrace{\sum_{(x_i,y_j)\in P} \alpha(x_i, y_j)}_{paired} + \underbrace{\sum_{i:x_i \text{ unpaired}} \delta + \sum_{j:y_j \text{ unpaired}} \delta}_{gaps}$$

Spelling errors confusion matrix

Dan Jurafsky

Confusion matrix for spelling errors

sub[X, Y] = Substitution of X (incorrect) for Y (correct)

X	Y (correct)																									
	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
a	0	0	7	1	342	0	0	2	118	0	1	0	0	3	76	0	0	1	35	9	9	0	1	0	5	0
b	0	0	9	9	2	2	3	1	0	0	0	5	11	5	0	10	0	0	2	1	0	0	8	0	0	0
c	6	5	0	16	0	9	5	0	0	0	1	0	7	9	1	10	2	5	39	40	1	3	7	1	1	0
d	1	10	13	0	12	0	5	5	0	0	2	3	7	3	0	1	0	43	30	22	0	0	4	0	2	0
e	388	0	3	11	0	2	2	0	89	0	0	3	0	5	93	0	0	14	12	6	15	0	1	0	18	0
f	0	15	0	3	1	0	5	2	0	0	0	3	4	1	0	0	0	6	4	12	0	0	2	0	0	0
g	4	1	11	11	9	2	0	0	0	1	1	3	0	0	2	1	3	5	13	21	0	0	1	0	3	0
h	1	8	0	3	0	0	0	0	0	0	2	0	12	14	2	3	0	3	1	11	0	0	2	0	0	0
i	103	0	0	0	146	0	1	0	0	0	0	6	0	0	49	0	0	0	2	1	47	0	2	1	15	0
j	0	1	1	9	0	0	1	0	0	0	0	2	1	0	0	0	0	0	5	0	0	0	0	0	0	0
k	1	2	8	4	1	1	2	5	0	0	0	0	5	0	2	0	0	0	6	0	0	0	4	0	0	3
l	2	10	1	4	0	4	5	6	13	0	1	0	0	14	2	5	0	11	10	2	0	0	0	0	0	0
m	1	3	7	8	0	2	0	6	0	0	4	4	0	180	0	6	0	0	9	15	13	3	2	2	3	0
n	2	7	6	5	3	0	1	19	1	0	4	35	78	0	0	7	0	28	5	7	0	0	1	2	0	2
o	91	1	1	3	116	0	0	0	25	0	2	0	0	0	0	14	0	2	4	14	39	0	0	0	18	0
p	0	11	1	2	0	6	5	0	2	9	0	2	7	6	15	0	0	1	3	6	0	4	1	0	0	0
q	0	0	1	0	0	0	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r	0	14	0	30	12	2	2	8	2	0	5	8	4	20	1	14	0	0	12	22	4	0	0	1	0	0
s	11	8	27	33	35	4	0	1	0	1	0	27	0	6	1	7	0	14	0	15	0	0	5	3	20	1
t	3	4	9	42	7	5	19	5	0	1	0	14	9	5	5	6	0	11	37	0	0	2	19	0	7	6
u	20	0	0	0	44	0	0	0	64	0	0	0	0	2	43	0	0	4	0	0	0	0	2	0	8	0
v	0	0	7	0	0	3	0	0	0	0	0	1	0	0	1	0	0	0	8	3	0	0	0	0	0	0
w	2	2	1	0	1	0	0	2	0	0	1	0	0	0	0	7	0	6	3	3	1	0	0	0	0	0
x	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0
y	0	0	2	0	15	0	1	7	15	0	0	0	2	0	6	1	0	7	36	8	5	0	0	1	0	0
z	0	0	0	7	0	0	0	0	0	0	0	7	5	0	0	0	0	2	21	3	0	0	0	0	3	0

Sequence Alignment: Problem

Input: Two sequences $X = x_1, \ldots, x_m$ and $Y = y_1, \ldots, y_n$ over Σ and a score/penalty matrix

Output: Minimum cost alignment between X and Y

Cost of the optimal alignment is the alignment distance between X and Y

Theorem: If all penalties are 1, alignment distance = edit distance (otherwise it is equal to the weighted edit distance)

- It implies that alignment distance is a metric
- Edit distance does not specify the edits (it just counts)
- Optimal alignment does not only compute edit distance but also specify the edits