

CSE 421  
Intro to Algorithms  
Winter 2004

Sequence Alignment

CSE 421, W '04, Ruzzo 1

## Sequence Alignment

- What
- Why
- A Simple Algorithm
- Complexity Analysis
- A better Algorithm: "Dynamic Programming"

CSE 421, W '04, Ruzzo 2

## Sequence Similarity: What

```

GGACCA
TACTAAG
|:|:|:|:
TCC-AAT
  
```

CSE 421, W '04, Ruzzo 3

## Sequence Similarity: Why

- Diff
- RCS
- Molecular Bio
  - Similar sequences often have similar origin or function
  - Similarity often recognizable after  $10^8 - 10^9$  years

CSE 421, W '04, Ruzzo 4

## Terminology

- **String**: ordered list of letters TATAAG
- **Prefix**: consecutive letters from front  
empty, T, TA, TAT, ...
- **Suffix**: ... from end  
empty, G, AG, AAG, ...
- **Substring**: ... from ends or middle  
empty, TAT, AA, ...
- **Subsequence**: ordered, nonconsecutive  
TT, AAA, TAG, ...

CSE 421, W '04, Ruzzo 5

## Sequence Alignment

```

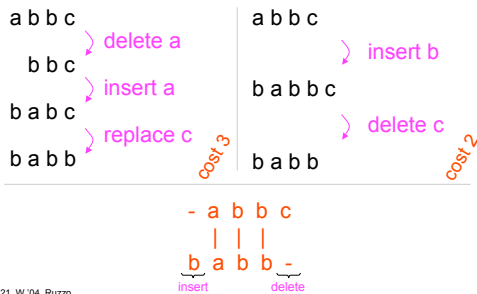
a c b c d b      a c - - b c d b
 / \ | | | |      | | | |
c a d b d      - c a d b - d -
  
```

**Defn:** An *alignment* of strings S, T is a pair of strings S', T' (with spaces) s.t.

- (1)  $|S'| = |T'|$ , and  $(|S| = \text{"length of S"})$
- (2) removing all spaces leaves S, T

CSE 421, W '04, Ruzzo 6

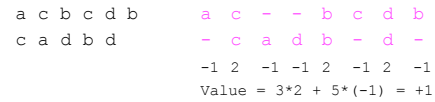
## 6.8: "Min\_Edit\_Distance"



CSE 421, W '04, Ruzzo

7

## Alignment Scoring



- The *score* of aligning (characters or spaces)  $x$  &  $y$  is  $\sigma(x,y)$ .
- *Value* of an alignment =  $\sum_{i=1}^{|S|} \sigma(S^i[i], T^i[i])$
- An *optimal alignment*: one of max value

CSE 421, W '04, Ruzzo

8

## Optimal Alignment: A Simple Algorithm

**for all** subseqs  $A$  of  $S$ ,  $B$  of  $T$  s.t.  $|A| = |B|$  **do**  
 align  $A[i]$  with  $B[i]$ ,  $1 \leq i \leq |A|$   
 align all other chars to spaces  
 compute its value  
 retain the max  
**end**  
 output the retained alignment

$S = abcd$     $A = cd$   
 $T = wxyz$     $B = xz$   
 $-abc-d$     $a-bc-d$   
 $w--xyz$     $-w-xyz$

CSE 421, W '04, Ruzzo

9

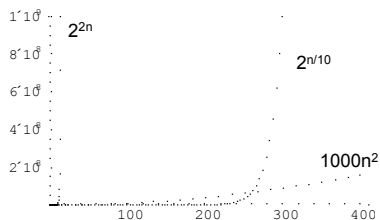
## Analysis

- Assume  $|S| = |T| = n$
- Cost of evaluating one alignment:  $\geq n$
- How many alignments are there:  $\geq \binom{2n}{n}$   
 pick  $n$  chars of  $S, T$  together  
 say  $k$  of them are in  $S$   
 match these  $k$  to the  $k$  unpicked chars of  $T$
- Total time:  $\geq n \binom{2n}{n} > 2^{2n}$ , for  $n > 3$
- E.g., for  $n = 20$ , time is  $> 2^{40}$  operations

CSE 421, W '04, Ruzzo

10

## Polynomial vs Exponential Growth



CSE 421, W '04, Ruzzo

11

## Candidate for Dynamic Programming?

- Common Subproblems?  
 » Plausible: probably re-considering alignments of various small substrings unless we're careful.
- Optimal Substructure?  
 » Plausible: left and right "halves" of an optimal alignment probably should be optimally aligned (though they obviously interact a bit at the interface).

CSE 421, W '04, Ruzzo

12

## Optimal Substructure (In More Detail)

- Optimal alignment ends in 1 of 3 ways:
  - last chars of S & T aligned with each other
  - last char of S aligned with space in T
  - last char of T aligned with space in S
  - ( never align space with space;  $\sigma(-, -) \geq 0$  )
- In each case, the **rest** of S & T should be optimally aligned to each other

CSE 421, W '04, Ruzzo

13

## Optimal Alignment in $O(n^2)$ via "Dynamic Programming"

- Input: S, T,  $|S| = n$ ,  $|T| = m$
- Output: **value** of optimal alignment

Easier to solve a "harder" problem:

$$V(i,j) = \text{value of optimal alignment of } S[1], \dots, S[i] \text{ with } T[1], \dots, T[j] \text{ for all } 0 \leq i \leq n, 0 \leq j \leq m.$$

CSE 421, W '04, Ruzzo

14

## General Case

Opt align of  $S[1], \dots, S[i]$  vs  $T[1], \dots, T[j]$ :

Opt align of  $S_{1..i}$  &  $T_{1..j}$

$$V(i,j) = \max \left\{ \begin{array}{l} V(i-1,j-1) + \sigma(S[i], T[j]) \\ V(i-1,j) + \sigma(S[i], -) \\ V(i,j-1) + \sigma(-, T[j]) \end{array} \right\}$$

for all  $1 \leq i \leq n, 1 \leq j \leq m$ .

CSE 421, W '04, Ruzzo

15

## Base Cases

- $V(i,0)$ : first  $i$  chars of S; all match spaces
 
$$V(i,0) = \sum_{k=1}^i \sigma(S[k], -)$$
- $V(0,j)$ : first  $j$  chars of T; all match spaces
 
$$V(0,j) = \sum_{k=1}^j \sigma(-, T[k])$$

CSE 421, W '04, Ruzzo

16

## Example

Mismatch = -1  
Match = 2

	j	0	1	2	3	4	5	
i			c	a	d	b	d	←T
0		0	-1	-2	-3	-4	-5	
1	a	-1	-1	1				
2	c	-2	1					
3	b	-3						
4	c	-4						
5	d	-5						
6	b	-6						

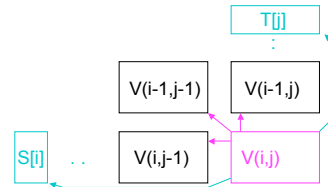
Time =  $O(mn)$

CSE 421, W '04, Ruzzo

17

## Calculating One Entry

$$V(i,j) = \max \left\{ \begin{array}{l} V(i-1,j-1) + \sigma(S[i], T[j]) \\ V(i-1,j) + \sigma(S[i], -) \\ V(i,j-1) + \sigma(-, T[j]) \end{array} \right\}$$



CSE 421, W '04, Ruzzo

18

## Finding Alignments: Trace Back

j	0	1	2	3	4	5	
i		c	a	d	b	d	←T
0	0	-1	-2	-3	-4	-5	
1	a	-1	1	0	-1	-2	
2	c	-2	1	0	0	-1	
3	b	-3	0	0	-1	2	
4	c	-4	-1	-1	-1	1	
5	d	-5	-2	-2	1	0	3
6	b	-6	-3	-3	0	3	2

↑ S

CSE 421, W '04, Ruzzo

19

## Complexity Notes

- Time =  $O(mn)$ , (value and alignment)
- Space =  $O(mn)$
- Easy to get **value** in Time =  $O(mn)$  and Space =  $O(\min(m,n))$
- Possible to get value **and alignment** in Time =  $O(mn)$  and Space =  $O(\min(m,n))$  but tricky.

CSE 421, W '04, Ruzzo

20

### Taxonomy Report

```

root ..... 64 hits 16 orgs
. Eukaryota ..... 62 hits 14 orgs [cellular organisms]
. . Fungi/Metazoa group ..... 57 hits 11 orgs

. . . . . Tetrapoda ..... 26 hits 5 orgs [;;; Vertebrata;;; Sarcopterygii]
. . . . . Eutheria ..... 24 hits 4 orgs [Amniota; Mammalia; Theria]
. . . . . Homo sapiens ..... 20 hits 1 orgs [Primates;; Hominidae; Homo]
. . . . . Murinae ..... 3 hits 2 orgs [Rodentia; Sciurognathi; Muridae]
. . . . . Rattus norvegicus ..... 2 hits 1 orgs [Rattus]
. . . . . Mus musculus ..... 1 hits 1 orgs [Mus]
. . . . . Sus scrofa ..... 1 hits 1 orgs [Cetartiodactyla; Suina; Suidae; Sus]
. . . . . Xenopus laevis ..... 2 hits 1 orgs [Amphibia; ;;; Xenopodinae; Xenopus]
. . . . . Drosophila melanogaster ..... 10 hits 1 orgs [Protostomia; ;;; Drosophila; ;]
. . . . . Caenorhabditis elegans ..... 2 hits 1 orgs [; Nematoda; ;;;; Caenorhabditis]
. . . . . Ascomycota ..... 19 hits 4 orgs [Fungi]
. . . . . Schizosaccharomyces pombe ..... 10 hits 1 orgs [;;; Schizosaccharomyces]
. . . . . Saccharomycetales ..... 9 hits 2 orgs [Saccharomycotina; Saccharomycetes]
. . . . . Saccharomyces ..... 8 hits 2 orgs [Saccharomycetaceae]
. . . . . Saccharomyces cerevisiae ..... 7 hits 1 orgs
. . . . . Saccharomyces kluyveri ..... 1 hits 1 orgs
. . . . . Candida albicans ..... 1 hits 1 orgs [mitosporic Saccharomycetales;]
. . . . . Arabidopsis thaliana ..... 2 hits 1 orgs [Viridiplantae; Brassicaceae;]
. . . . . Apicomplexa ..... 3 hits 2 orgs [Alveolata]
. . . . . Plasmodium falciparum ..... 2 hits 1 orgs [Haemosporida; Plasmodium]
. . . . . Toxoplasma gondii ..... 1 hits 1 orgs [Coccidia; Eimeriida; Sarcocystidae;]
. synthetic construct ..... 1 hits 1 orgs [other; artificial sequence]
. Ivmehocovitis disease virus ..... 1 hits 1 orgs [Viruses; dsDNA viruses, no RNA _]

```

21