Lecture outline

- Sequence alignment
  - Why do we need to align sequences?
  - Evolutionary relationships
- Gaps and scoring matrices
- Dynamic programming
  - Global alignment (Needleman & Wunsch)
  - Local alignment (Smith & Waterman)
- Database searches
  - BLAST
  - FASTA

21

Dot Plot

Put 1's where characters are identical.

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1		1			
Y					1								
N				1									
R						1					1		
C			1					1		1			
K													
C			1					1		1			
R						1					1		
B		1											
P												1	

22

A More Interesting Dot Matrix

23

Edit Distance

Levenshtein (1966) introduced edit distance between two strings as the minimum number of elementary operations (insertions, deletions, and substitutions) to transform one string into the other

d(**v**,**w**) = MIN number of elementary operations to transform **v** → **w**

24

Edit Distance: Example

TGCATAT → ATCCGAT in 5 steps

TGCATAT → (delete last T)
TGCATA → (delete last A)
TGCAT → (insert A at front)
ATGCAT → (substitute C for 3rd G)
ATCCAT → (insert G before last A)
ATCCGAT (Done)

25

Edit Distance: Example

TGCATAT → ATCCGAT in 5 steps

TGCATAT → (delete last T)
TGCATA → (delete last A)
TGCAT → (insert A at front)
ATGCAT → (substitute C for 3rd G)
ATCCAT → (insert G before last A)
ATCCGAT (Done)

What is the edit distance? 5?

26

Edit Distance: Example (cont'd)

TGCATAT → ATCCGAT in 4 steps

TGCATAT → (insert A at front)
ATGCATAT → (delete 6th T)
ATGCATA → (substitute G for 5th A)
ATGCGTA → (substitute C for 3rd G)
ATCCGAT (Done)

27

Edit Distance: Example (cont'd)

TGCATAT → ATCCGAT in 4 steps

TGCATAT → (insert A at front)
ATGCATAT → (delete 6th T)
ATGCATA → (substitute G for 5th A)
ATGCGTA → (substitute C for 3rd G)
ATCCGAT (Done)

Can it be done in 3 steps???

28

Types of alignment

- Global (Needleman & Wunsch)
  - Strings of similar size
    - Genes with a similar structure
    - Larger regions with a preserved order (syntenic regions)
- Local (Smith & Waterman)
  - Finding similar regions among
    - Dissimilar regions
    - Sequences of different lengths

29

Dynamic programming

- Instead of evaluating every possible alignment, we can create a table of partial scores by breaking the alignment problem into subproblems.
- Consider two sequences CACGA and CGA
  - we have three possibilities for the first position of the alignment

First position	Score	Remaining seqs.
C C	+1	ACGA GA
- C	-1	CACGA GA
C -	-1	ACGA CGA

30

Example

score(H,P) = -2, gap penalty=-8 (linear)

	-	H	E	A	G	A	W	G	H	E	E
-	0	-8	-16	-24	-32	-40	-48	-56	-64	-72	-80
P	-8	-2									
A	-16										
W	-24										
H	-32										
E	-40										
A	-48										
E	-56										

31

The DP recurrence relation

- s(a,b) = score of aligning a and b
- S(i,j) = optimal similarity of A(1:i) and B(1:j)
- Recurrence relation
  - $S(i,0) = \Sigma\, s(A(k),-),\ 0 <= k <= i$
  - $S(0,j) = \Sigma\, s(-,B(k)),\ 0 <= k <= j$
  - $S(i,j) = \max [S(i,j-1) + s(-,B(j)),\ S(i-1,j) + s(A(i),-),\ S(i-1,j-1) + s(A(i),B(j))]$
  - Assume linear gap penalty

32

Example contd.

score(E,P) = 0, score(E,A) = -1, score(H,A) = -2

	-	H	E	A	G	A	W	G	H	E	E
-	0	-8	-16	-24	-32	-40	-48	-56	-64	-72	-80
P	-8	-2	-8								
A	-16	-10	-3								
W	-24										
H	-32										
E	-40										
A	-48										
E	-56										

33

Optimal alignment:

H E A G A W G H E - E
- P - - A W - H E A E

		H	E	A	G	A	W	G	H	E	E
	0	-8	-16	-24	-32	-40	-48	-56	-64	-72	-80
P	-8	-2	-8	-16	-24	-33	-42	-49	-57	-65	-73
A	-16	-10	-3	-4	-12	-19	-28	-36	-44	-52	-60
W	-24	-18	-11	-6	-7	-15	-4	-12	-21	-29	-37
H	-32	-14	-18	-13	-8	-9	-12	-6	-2	-11	-19
E	-40	-22	-8	-16	-16	-9	-12	-14	-6	-4	-5
A	-48	-30	-16	-3	-11	-11	-12	-12	-14	-4	2
E	-56	-38	-24	-11	-6	-12	-14	-15	-12	-8	2

The value in the final cell is the best score for the alignment

34

More examples

- Sequence alignment applet at:

  http://www.iro.umontreal.ca/~casagran/baba.html

35

Semi-global alignment

- In Needleman&Wunsch DP algorithm the gap penalty is assessed regardless of whether gaps are located internally or at the terminal ends.
- Terminal gaps may not be biologically significant

```
AATCTATA
--TCT---
```

- Treat terminal gaps differently than internal gaps → semi-global alignment
- What modifications should be made to the original DP?

36

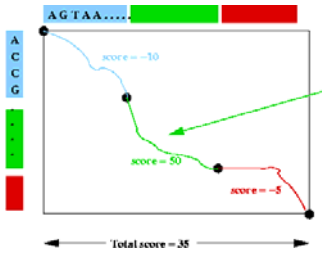
Local sequence alignment

- Suppose, we have a long DNA sequence (e.g., 4000 bp) and we want to compare it with the complete yeast genome (12.5M bp).
- What if only a portion of our query, say 200 bp length, has strong similarity to a gene in yeast.
  - Can we find this 200 bp portion using (semi) global alignment?

    Probably not. Because, we are trying to align the complete 4000 bp sequence, thus a random alignment may get a better score than the one that aligns 200 bp portion to the similar gene in yeast.

37

Local alignment



local alignment may have higher score than overall global alignment

38

Local sequence alignment (Smith-Waterman)

- S(i,j) = optimal local similarity among suffixes of A(1:i) and B(1:j)
- Recurrence relation
  - S(i,0) = 0
  - S(0,j) = 0
  - S(i,j) = max [0,
    S(i,j-1) + s(-,B(j)),
    S(i-1,j) + s(A(i),-),
    S(i-1,j-1) + s(A(i),B(j))]
  - Assume linear gap model

39

Example

Q: E Q L L K A L E F K L
P: K V L E F G Y

Linear gap model
Gap = -1
Match = 4
Mismatch = -2

	-	E	Q	L	L	K	A	L	E	F	K	L
-												
K												
V												
L												
E												
F												
G												
Y												

40

Example

Q: E Q L L K A L E F K L
P: K V L E F G Y

Linear gap model
Gap = -1
Match = 4
Mismatch = -2

	-	E	Q	L	L	K	A	L	E	F	K	L
-	0	0	0	0	0	0	0	0	0	0	0	0
K	0											
V	0											
L	0											
E	0											
F	0											
G	0											
Y	0											

41

Example

Q: E Q L L K A L E F K L
P: K V L E F G Y

Linear gap model
Gap = -1
Match = 4
Mismatch = -2

	-	E	Q	L	L	K	A	L	E	F	K	L
-	0	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0	0	0	4	3	2	1	0	4	3
V	0	0	0	0	0	3	2	1	0	0	3	2
L	0	0	0	4	4	3	2	6	5	4	3	7
E	0	4	3	3	3	2	1	5	10	9	8	7
F	0	3	2	2	2	1	0	4	9	14	13	12
G	0	2	1	1	1	0	0	3	8	13	12	11
Y	0	1	0	0	0	0	0	2	7	12	11	10

42

Example

Q: E Q L L K A L E F K L
P: K V L E F G Y

Linear gap model
Gap = -1
Match = 4
Mismatch = -2

	-	E	Q	L	L	K	A	L	E	F	K	L
-	0	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0	0	0	4	3	2	1	0	4	3
V	0	0	0	0	0	3	2	1	0	0	3	2
L	0	0	0	4	4	3	2	6	5	4	3	7
E	0	4	3	3	3	2	1	5	10	9	8	7
F	0	3	2	2	2	1	0	4	9	14	13	12
G	0	2	1	1	1	0	0	3	8	13	12	11
Y	0	1	0	0	0	0	0	2	7	12	11	10

43

Example

Q: E Q L L K A L E F K L
P: K V L E F G Y

Linear gap model
Gap = -1
Match = 4
Mismatch = -2

	-	E	Q	L	L	K	A	L	E	F	K	L
-	0	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0	0	0	4	3	2	1	0	4	3
V	0	0	0	0	0	3	2	1	0	0	3	2
L	0	0	0	4	4	3	2	6	5	4	3	7
E	0	4	3	3	3	2	1	5	10	9	8	7
F	0	3	2	2	2	1	0	4	9	14	13	12
G	0	2	1	1	1	0	0	3	8	13	12	11
Y	0	1	0	0	0	0	0	2	7	12	11	10

44

Example

Q: E Q L L K A L E F K L
P: K V L E F G Y

Alignment

```
Q:  K A - L E F
P:  K - V L E F
```

	-	E	Q	L	L	K	A	L	E	F	K	L
-	0	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0	0	0	4	3	2	1	0	4	3
V	0	0	0	0	0	3	2	1	0	0	3	2
L	0	0	0	4	4	3	2	6	5	4	3	7
E	0	4	3	3	3	2	1	5	10	9	8	7
F	0	3	2	2	2	1	0	4	9	14	13	12
G	0	2	1	1	1	0	0	3	8	13	12	11
Y	0	1	0	0	0	0	0	2	7	12	11	10

Example

Q: E Q L L K A L E F K L
P: K V L E F G Y

Alignment

Q: K - A L E F
P: K V - L E F

	-	E	Q	L	L	K	A	L	E	F	K	L
-	0	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0	0	0	4	3	2	1	0	4	3
V	0	0	0	0	0	3	2	1	0	0	3	2
L	0	0	0	4	4	3	2	6	5	4	3	7
E	0	4	3	3	3	2	1	5	10	9	8	7
F	0	3	2	2	2	1	0	4	9	14	13	12
G	0	2	1	1	1	0	0	3	8	13	12	11
Y	0	1	0	0	0	0	0	2	7	12	11	10

45

Example

Q: E Q L L K A L E F K L
P: K V L E F G Y

Alignment

Q: K A L E F
P: K V L E F

	-	E	Q	L	L	K	A	L	E	F	K	L
-	0	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0	0	0	4	3	2	1	0	4	3
V	0	0	0	0	0	3	2	1	0	0	3	2
L	0	0	0	4	4	3	2	6	5	4	3	7
E	0	4	3	3	3	2	1	5	10	9	8	7
F	0	3	2	2	2	1	0	4	9	14	13	12
G	0	2	1	1	1	0	0	3	8	13	12	11
Y	0	1	0	0	0	0	0	2	7	12	11	10

46

Another Example

Find the local alignment between:

Q: G C T G G A A G G C A T
P: G C A G A G C A C G

Linear gap model
Gap = -4
Match = +5
Mismatch = -4

Q

P	--	G	C	T	G	G	A	A	G	G	C	A	T
--	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0												
C	0												
A	0												
G	0												
A	0												
G	0												
C	0												
A	0												
C	0												
G	0												

47

Another Example

Q's subsequence: G A A G - G C A
P's subsequence: G C A G A G C A

Q

P	--	G	C	T	G	G	A	A	G	G	C	A	T
--	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	5	1	0	5	5	1	0	5	5	1	0	0
C	0	1	10	6	2	1	1	0	1	1	10	6	2
A	0	0	6	6	2	0	6	6	2	0	6	15	11
G	0	5	2	2	11	7	3	2	11	7	3	11	11
A	0	1	1	0	7	7	11	8	7	7	3	8	7
G	0	5	1	0	5	11	7	7	13	12	8	4	4
C	0	0	10	6	2	7	7	3	9	8	17	13	9
A	0	0	6	6	2	3	11	12	8	5	13	22	18
C	0	0	5	2	2	0	7	8	8	4	18	18	18
G	0	5	1	1	7	7	5	4	13	13	14	14	14

48

Local vs. Global alignment

PIR Entry				Similarity Score: Global, End Penalty	Global, No End Penalty	Local
HBHU	vs	HBHU	Hemoglobin beta-chain—human	725	725	725
		HAHU	Hemoglobin alpha-chain—human	314	320	322
		MYHU	Myoglobin—Human	121	164	166
		GPYL	Leghemoglobin—Yellow lupin	8	28	43
		LZCH	Lysozyme precursor—Chicken	−107	16	32
		NRBO	Pancreatic ribonuclease—Bovine	−124	16	31
		CCHU	Cytochrome c—Human	−160	10	26
MCHU	vs	MCHU	Calmodulin—Human	671	671	671
		TPHUCS	Troponin C, skeletal muscle	395	430	438
		PVPK2	Parvalbumin beta—Pike	−57	103	115
		CBHUH	Calpain heavy chain—Human	−2085	89	100
		AQJFNV	Aequorin precursor—Jelly fish	−65	48	76
		KLSWM	Calcium binding protein—Scallop	−89	45	52
QRHULD	vs	EGMSMG	Epidermal growth factor precursor	−591	475	655

49

Complexity

- O(mn) time
- O(mn) space
  - O(max(m,n)) if only distance value is needed
- More complicated "divide-and-conquer" algorithm that doubles time complexity and uses O(min(m,n)) space [Hirschberg, JACM 1977]

50

Time and space bottlenecks

- Comparing two one-megabase genomes.
- Space:
  An entry: 4 bytes;
  Table: 4 * 10^6 * 10^6= 4 T bytes memory.
- Time:
  1000 MHz CPU: 1M entries/second;
  10^{12} entries: 1M seconds = 10 days.

51

Affine Gap Penalties

- -d for gap opening
- -e for gap extension
- *H(i,j)* for gaps in horizontal seq.
- *V(i,j)* for gaps in vertical seq.

$$S(i,j) = \max\begin{cases} S(i-1,j-1) + s(x_i, y_j), \\ H(i-1,j-1), \\ V(i-1,j-1). \end{cases}$$

$$H(i,j) = \max\begin{cases} F(i-1,j) - d, \\ H(i-1,j) - e. \end{cases}$$

$$V(i,j) = \max\begin{cases} F(i,j-1) - d, \\ V(i,j-1) - e. \end{cases}$$

52