

Sequence Alignment

Sequence alignment

Bioinformatics Fall 2011

1

Motivation

- Sequence the most prevalent biological data
 - Much of the techniques also apply to text/string analysis
- Sequence alignment has multiple uses
 - Detecting orthologs
 - Predicting function
 - Sequencing
- Lecture(s) plan
 - Global and local alignment
 - Genetic variations
 - Score matrices
 - Alignment statistics
 - Multiple sequence alignment
 - Database searching
 - Next generation sequencing

Sequence alignment

Bioinformatics Fall 2011

2

A simple alignment

- Let us try to align two short nucleotide sequences:
 - AATCTATA and AAGATA
- Without considering any gaps (insertions/deletions) there are 3 possible ways to align these sequences

AATCTATA	AATCTATA	AATCTATA
AAGATA	AAGATA	AAGATA

- Which one is better?

Sequence alignment

Bioinformatics Fall 2011

3

Scoring the alignments

- We need to have a scoring mechanism to evaluate alignments
 - match score
 - mismatch score
- We can have the total score as:

$$\sum_{i=1}^n \text{match or mismatch score at position } i$$

- For the simple example, assume a match score of 1 and a mismatch score of 0:

AATCTATA	AATCTATA	AATCTATA
AAGATA	AAGATA	AAGATA
4	1	3

Sequence alignment

Bioinformatics Fall 2011

4

Good alignments require gaps

- Maximal consecutive run of spaces in alignment
 - Matching mRNA (cDNA) to DNA
 - Shortening of DNA during replication
 - Slippage during replication
 - Unequal cross-over during meiosis
 - ...
- Need a scoring function that considers gaps

Sequence alignment

Bioinformatics Fall 2011

5

Simple alignment with gaps

- Considering gapped alignments vastly increases the number of possible alignments:

AATCTATA	AATCTATA	AATCTATA	more?
AAG-AT-A	AA-G-ATA	AA--GATA	
1	3	3	

$$f(m,n) = \text{number of alignments of two strings of length } m,n$$
$$= f(m,n-1) + f(m-1,n) + f(m-1,n-1)$$

- If gap penalty is -1 what will be the new scores?

Sequence alignment

Bioinformatics Fall 2011

6

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

		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										

Sequence alignment Bioinformatics Fall 2011 13

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										

Sequence alignment Bioinformatics Fall 2011 14

Optimal alignment: HEAGAWGHE - E
- P - - AW - HEAE

		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	-1	-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

Score of the best alignment

Sequence alignment Bioinformatics Fall 2011 15

Local sequence alignment (Smith-Waterman)

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

Sequence alignment Bioinformatics Fall 2011 16

Example

α 's subsequence: G C A G A G C A
Match = +5
 β 's subsequence: G A A G - G C A
Mismatch = -4

Linear gap model

		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
G	0	0	5	2	2	0	7	8	8	4	18	18	18
C	0	5	1	1	7	7	5	4	13	13	14	14	14

Sequence alignment Bioinformatics Fall 2011 17

Affine gaps: d (open), e (extend)

- Four dynamic programming terms
 - $G(i,j)$ is the optimal local similarity when $\alpha(i)$ is aligned with $\beta(j)$
 - $E(i,j)$ is the optimal local similarity when $\alpha(i)$ is aligned to the left of $\beta(j)$
 - $F(i,j)$ is the optimal local similarity when $\alpha(i)$ is aligned to the right of $\beta(j)$
 - $S(i,j)$ is the optimal score of aligning $\alpha(1:i)$ and $\beta(1:j)$
- Recurrences
 - $S(i,j) = \max [E(i,j), F(i,j), G(i,j)]$
 - $G(i,j) = S(i-1,j-1) + s(\alpha(i), \beta(j))$
 - $E(i,j) = \max [F(i,j-1) + d, G(i,j-1) + d, E(i,j-1) + e]$
 - $F(i,j) = \max [E(i-1,j) + d, G(i-1,j) + d, F(i-1,j) + e]$

If $|d| > |e|$, these simplify to
 $E(i,j) = \max [S(i,j-1) + d, E(i,j-1) + e]$
 $F(i,j) = \max [S(i-1,j) + d, F(i-1,j) + e]$
- Base case
 - $S(0,0) = 0, S(i,0) = E(i,0) = -d + (i-1)*e$
 - $S(0,j) = F(0,j) = d + (j-1)*e$

Sequence alignment Bioinformatics Fall 2011 18

Typical parameters

- DNA
 - Match = +2
 - Mismatch = -3
 - Gaps: open = -5, extension = -2
- Protein
 - BLOSUM/PAM
 - Gaps: open = -11, extension = -1

Sequence alignment Bioinformatics Fall 2011 19

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]
- Convex gaps: $O(mn \log(m+n))$ time
- Arbitrary gaps: $O(mn(m+n))$ time

Sequence alignment Bioinformatics Fall 2011 20

Linear time global alignment

- $S^r(i,j)$ = similarity score of last i characters of α with last j characters of β = similarity score of first i characters of α^r with the first j characters of β^r
- $S(m,n) = \max_k [S(m/2,k) + S^r(m/2,n-k)]$

Sequence alignment Bioinformatics Fall 2011 21

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:
 - 1 GHz CPU: 10 M entries/second;
 - 10^{12} entries: 10^5 seconds > 1 day.

Sequence alignment Bioinformatics Fall 2011 22

Banded global alignment

- Two sequences differ by at most w ($w \ll n$).
- w -band algorithm: $O(wn)$ time and space.
- Example: $w = 3$
 - Linear gap penalty
 - Match = +1
 - Mismatch = -1

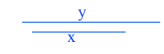
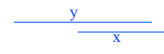
		A	C	C	A	C	A	C	A
	0	-1	-2	-3					
A	-1	1	0	-1	-2				
C	-2	0	2	1	0	-1			
A	-3	-1	1	1	2	1	0		
C	-2	0	2	1	3	2	1		
C	-1	1	1	1	2	2	3	2	
A			0	1	1	2	2	4	
T				0	0	1	1	3	
A					-1	1	0	2	

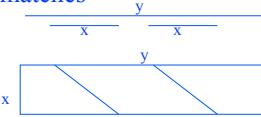
What if k is not known?

Sequence alignment Bioinformatics Fall 2011 23

Other kinds of alignments

- Overlaps



- Repeated matches



Sequence alignment Bioinformatics Fall 2011 24

Statistics of local alignment

- Expected score should be negative.
- $\sum q_a q_b s(a,b) < 0$, assume no gaps
- $\sum q_a q_b \log (p_{ab} / q_a q_b) < 0$, for log odds score
- $-\sum q_a q_b \log (q_a q_b / p_{ab}) < 0$ H = 0 only when $q^2 = p$
- $-H(q^2 \parallel p) < 0$, since H is positive
 - H is the relative entropy between the product distribution and the joint distribution
 - Also called the Kullback-Leibler distance