

Local Alignment

	_	H	E	A	G	A	W	G	H	E	E
_	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0
A	0	0	0	5d	0	5d	0	0	0	0	0
W	0	0	0	0	2d	0	20d	12h	4h	0	0
H	0	10d	2h	0	0	0	12u	18d	22d	14h	6h
E	0	2u	16d	8h	0	0	4u	10u	18d	28d	20d
A	0	0	8u	21d	13h	5d	0	4d	10u	20u	27d
E	0	0	6d	13u	18d	12d	4h	0	4d	16d	26d

■ EHGWA

■ EH--WA



■ AWGHE

■ AW--HE



Local Alignment

- Expected score for a random match < 0

$$\sum_{a,b} q_a q_b s(a,b) < 0$$

- Also called Smith-Waterman



Repeated matches

- Previously seen best **single** local match between two sequences
- If two sequences are long, there are many subsequences with a high alignment score, all of which may be of importance
 - Repeated domain or motif in a protein
- **Problem:** Find one or more copies of sections of one sequence in the other
 - Asymmetric
 - Non-overlapping copies

Repeated Matches



- The matching regions in the first sequence do not overlap
- The first sequence is partitioned into matching/unmatching regions.
- The matching subsequences in S2 may overlap
- We need a significance threshold T



Redefine $F(i,j)$

$$F(i, j) = \begin{cases} \text{Best matching score } x_{1..i} \text{ that } x_i \& y_j & x_i \in M \\ F(i,0) & x_i \notin M \end{cases}$$



Recursive equations...

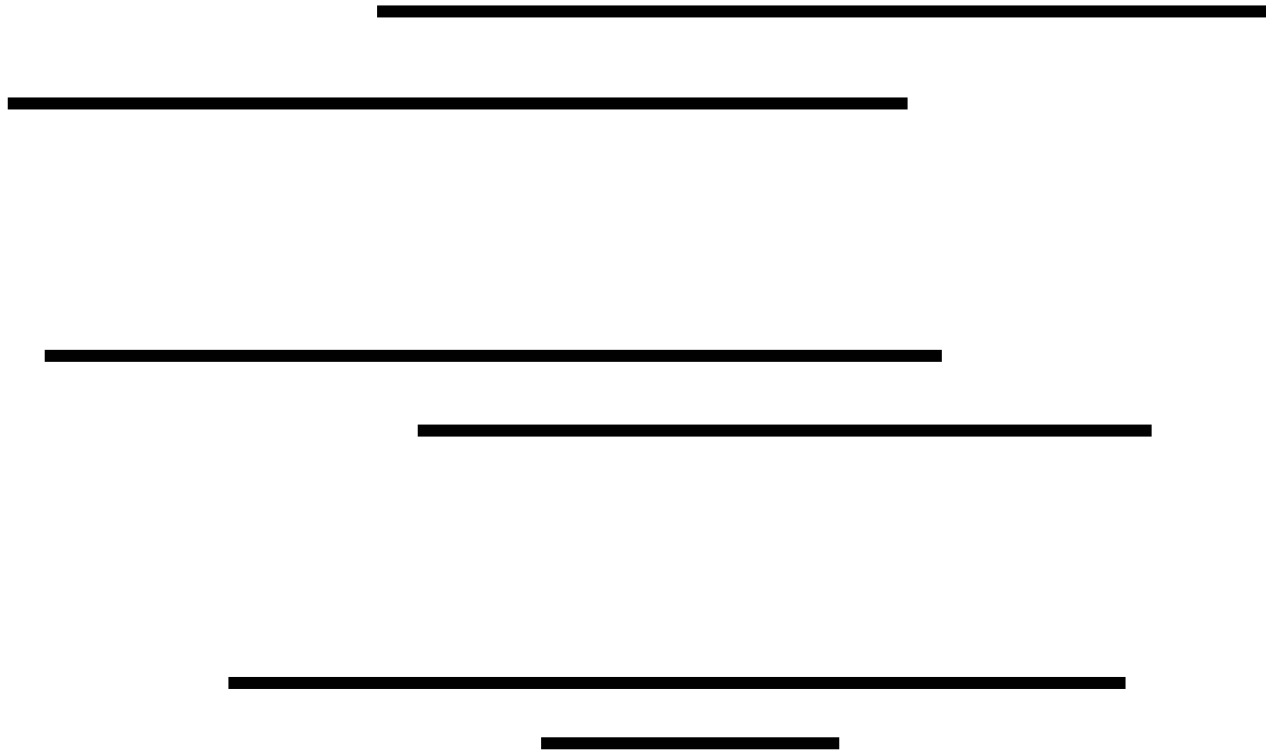
$$F(i,0) = \begin{cases} F(i-1,0) \\ F(i-1,j) - T \quad j = 1, \dots, m \end{cases}$$

$$F(i,j) = \max \begin{cases} F(i-1,0) \\ F(i-1,j-1) - s(x_i, y_i) \\ F(i-1,j) - d \\ F(i,j-1) - d \end{cases}$$



End-space free alignment

- Also called overlap match



- Gaps at the ends are not punished



Overlap match

- Same as the global alignment with small differences
- The backtrace does not have to start from $F(m,n)$ and $F(0,0)$.
 - Starts at the maximum value over last column/row
 - End when we reach 0, not $F(0,0)$
- Boundary conditions are different
 - $F(i,0)=0$
 - $F(0,j)=0$

Overlap match example

	_	H	E	A	G	A	W	G	H	E	E
_	0	0	0	0	0	0	0	0	0	0	0
P	0	-2	-1	-1	-2	-1	-4	-2	-2	-1	-1
A	0	-2	-2	4	-1	3	-4	-4	-4	-3	-2
W	0	-3	-5	-4	1	-4	18	10	2	6	-6
H	0	10	2	6	-6	-1	10	16	20	12	4
E	0	2	16	8	0	7	2	8	16	26	18
A	0	-2	8	21	13	5	3	2	8	18	25
E	0	0	4	13	18	12	4	4	2	4	24

■ EEHWAG

■ AEH...WAP



■ GAWGHEE

■ PAW...HEA



Dynamic programming with complex models

- Considered the simplest gap model

$$\gamma(g) = -gd$$

- May not be ideal for the application at hand
- For more general models, modify recursion

$$F(i, j) = \max \begin{cases} F(i-1, j-1) + s(x_i, y_j) \\ F(k, j) + \gamma(i-k) & k = 0, \dots, i-1 \\ F(i, k) + \gamma(j-k) & k = 0, \dots, j-1 \end{cases}$$

- $O(n \times n) \rightarrow O(n \times n \times n)$
 - Have to check $i+j+1$ cells



Alignment with affine gap score

$$\gamma(g) = -d - (g - 1)e$$

- Affine gap score yield $O(n^2)$ implementation.
- But, we have to keep track of 3 variables in each cell

$$M(i, j) = \max \begin{cases} M(i-1, j-1) + s(x_i, y_j) \\ I_x(i-1, j-1) + s(x_i, y_j) \\ I_y(i-1, j-1) + s(x_i, y_j) \end{cases}$$

$$I_x(i, j) = \max \begin{cases} M(i-1, j) - d \\ I_x(i-1, j) - e \end{cases}$$

$$I_y(i, j) = \max \begin{cases} M(i, j-1) - d \\ I_y(i, j-1) - e \end{cases}$$



Comments

- Insertion (gap in y) is not followed by deletion (gap in x)
- This would be true on the optimal path if
 - $-(d+e) < \text{minimum mismatch score}$
- These recursive equations can be implemented with FSA (Finite State Automata)



Example with Affine Gap Model



Homework #1

- Implement “Global Match”
- Implement “Local Match”
- Implement “Overlap Match”
- Implement “Affine Gap Score Match”
 - Input
 - Two sequences (entered by typing)
 - Parameters (d, e) etc.
 - Assume BLOSUM score matrix
 - Output
 - Aligned sequences as shown in the book/notes



Significance of scores

- How do we know the best score yields an alignment that is biologically meaningful?
- Mathematically, given any two sequence x and y , what is the probability that they are biologically related?



A little math...

$$\begin{aligned}P(M | x, y) &= \frac{P(x, y | M)P(M)}{P(x, y)} \\&= \frac{P(x, y | M)P(M)}{P(x, y | M)P(M) + P(x, y | R)P(R)} \\&= \frac{P(x, y | M)P(M)/P(x, y | R)P(R)}{P(x, y | M)P(M)/P(x, y | R)P(R) + P(x, y | R)P(R)/P(x, y | R)P(R)} \\&= \frac{P(x, y | M)P(M)/P(x, y | R)P(R)}{P(x, y | M)P(M)/P(x, y | R)P(R) + 1}\end{aligned}$$

- Now, define

$$\frac{P(x, y | M)P(M)}{P(x, y | R)P(R)} = \frac{P(x, y | M) P(M)}{\underbrace{P(x, y | R)}_{e^S} \underbrace{P(R)}_{e^{S''}}}$$

Cont...

$$S = \log \frac{P(x, y | M)}{P(x, y | R)}$$

$$S'' = \log \frac{P(M)}{P(R)}$$

$$S' = S + S''$$

$$S = \log \frac{P(x, y | M)}{P(x, y | R)} + \log \frac{P(M)}{P(R)}$$

$$P(M | x, y) = \frac{P(x, y | M)P(M)/P(x, y | R)P(R)}{P(x, y | M)P(M)/P(x, y | R)P(R) + 1}$$

$$= \frac{e^{S'}}{1 + e^{S'}}$$

$$= \sigma(S')$$

$$\sigma(x) = \frac{e^x}{1 + e^x}$$



Comments

- Sigma function is called “logistic” function or “sigmoid” function.
 - $\rightarrow 1$ as $x \rightarrow \text{infinity}$
 - $\rightarrow 0$ as $x \rightarrow -\text{infinity}$
 - At $X=0$, its value 0.5
- Large database search
 - A large number of sequences will increase random match chance
 - Should compare S' with $\log(N)$ rather than 0



The classical approach

$$S = S(x_1, y_1) + \dots + S(x_l, y_l) \\ = \sum_l IID$$

- S is Normal distributed (sum of IID's)

$$M_N = \max \{ \underbrace{S^{(1)}, S^{(2)}, \dots, S^{(N)}}_{N \text{ random matches}} \}$$

$$M_N \sim EVD$$



Extreme value distribution

- Take N samples from $g(x)$

$$P(\max(S_i) < x) = P(S_1 < x)P(S_2 < x)\dots P(S_N < x)$$

$$= \left(\int_{-\infty}^x g(u) du \right)^N$$

$$= G(x)^N$$

- Differentiating would yield pdf

$$EVD_N(x) = NG(x)^{N-1} g(x)$$

$$EVD(x) = \lim_{N \rightarrow \infty} NG(x)^{N-1} g(x)$$

EVD for $g(x)$ =exponential distr.

$$g(x) = \lambda e^{-\lambda x}$$

$$EVD(x) = \lim N \lambda e^{-\lambda x} \left(1 - e^{-\lambda x}\right)^{N-1}$$

$$= \lim N \lambda e^{-\lambda \left[z + \frac{1}{\lambda} \log N\right]} \left(1 - e^{-\lambda \left[z + \frac{1}{\lambda} \log N\right]}\right)^{N-1}$$

$$= \lim N \lambda e^{-\lambda z} \left(1 - \frac{e^{-\lambda z}}{N}\right)^{N-1}$$

$$= \lambda e^{-\lambda z} e^{\lambda e^{-\lambda z}}$$

$$= \lambda e^{-\lambda \left[x - \frac{1}{\lambda} \log N\right]} e^{\lambda e^{-\lambda \left[x - \frac{1}{\lambda} \log N\right]}}$$

$$= N \lambda e^{-\lambda x} e^{N e^{-\lambda x}}$$



Cont...

$$\begin{aligned}P(M_N < S) &= \int_0^S N\lambda e^{-\lambda x} e^{Ne^{-\lambda x}} dx \\&= \frac{1}{\lambda} e^{-Ne^{-\lambda S}} \\&\approx e^{-KNe^{-\lambda S}}\end{aligned}$$

- If this probability is small, say 0.01, then it is likely that x and y are related



Local ungapped alignments

- The number of unrelated matches with score greater than S is Poisson distributed

$$P([\#match > S] = a) = e^{-E} \frac{E^a}{a!}$$

- With mean

$$E(S) = Kmne^{-\lambda S}$$

- Probability that at least one match with a greater score is

$$P(x > S) = 1 - e^{-E(S)}$$

