

Chapter 2 Pairwise Alignment [Main textbook]

- A common problem is to find out if two sequences are related
 - ⇒ This is done by aligning the sequences & determining if this is statistically significant

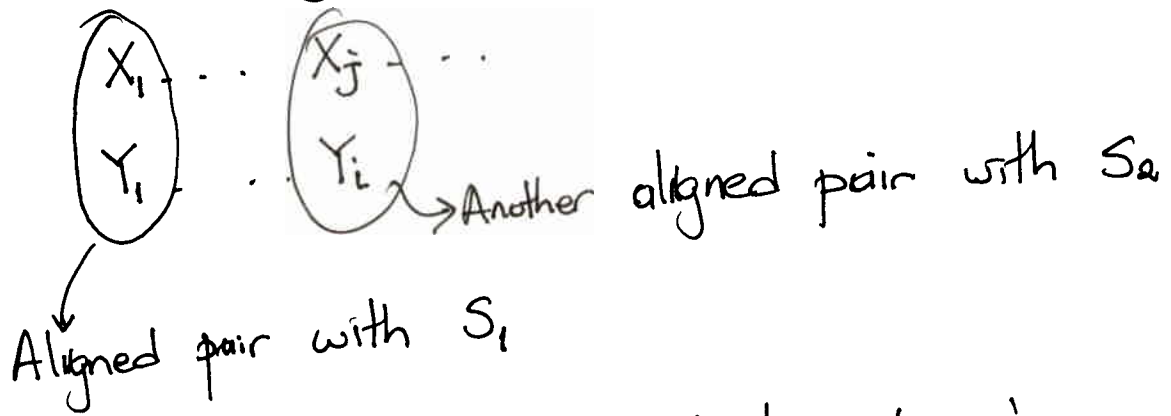
• Issues:

- ① What sort of alignment should be considered
e.g. local, global
- ② The scoring system to rank alignments
- ③ Algorithm to find the optimal alignment
- ④ Statistical methods to evaluate the significance of the match.

Scoring Model

- The sequences derived from a common ancestor via
 - Substitutions
 - Insertions
 - Deletions } Gaps

- Additive scoring assumption



$$\text{Score} = \sum \text{Cost/Score of all aligned pairs.}$$

$$= S_1 + S_2 + \dots$$

- Additive scoring implies independence of mutations
 - Reasonable for DNA/Protein sequences
 - Not accurate for RNAs

DNA sequence ATCGAAG - - -

$\mathcal{A} = \{ A, G, C, T \} \rightarrow \text{Bases/Residues}$

↳ Alphabet size = 4

Protein sequence GSAQVKG - - - -

$\mathcal{A} = \{ \underbrace{A, \dots}_{20 \text{ Amino acid}} \}$

- x, y two sequences

$$P(x, y | \underbrace{\text{Random Model}}_{\text{Call } R}) = \prod_i q_{x_i} \prod_j q_{y_j}$$

$x: (x_1) x_2 x_3 \dots$ & $x_i \in \mathcal{A}$
 $P(x_i) = q_{x_i}$ (Occurance frequency)

Example: $x: \text{ATCCA GCATATCCG}$
 $x_1 \ x_2 \ x_3 \ \dots \ \uparrow$
 x_{13}
 & q_A, q_T, q_C, q_G

$$P(x, y | R) = P(x_1) P(x_2) P(x_3) P(x_4) \dots P(x_{13}) \cdot P(y_1) \dots P(y_{13})$$

$$= q_{x_1} q_{x_2} \dots q_{x_{13}} q_{y_1} \dots q_{y_{13}}$$

In the example

$$= q_A q_T q_C q_C q_A \dots$$

- Match model M

$$P(x, y | M) = \prod_i P_{x_i y_i}$$

Note the independence assumption

Odds Ratio: $\frac{P(x,y|M)}{P(x,y|R)} = \frac{\prod_i P_{x_i,y_i}}{\prod_i q_{x_i} \prod_j q_{y_j}} = \prod \frac{P_{x_i,y_i}}{q_{x_i} \cdot q_{y_i}}$

log odds ratio = $\sum_i \underbrace{S(x_i, y_i)}_{\text{log likelihood ratio}}, S(x_i, y_i) = \log \left(\frac{P_{x_i, y_i}}{q_{x_i} \cdot q_{y_i}} \right)$

- $S(a,b) = 0$ a, b do ~~not~~ ^{have} preference toward each other.
- > 0 (a,b) pairing occurs more freq. than chance
i.e. a, b favors each other
- < 0 The opposite.

	A	T	C	G
A	10	-3	-	.
T		5	-4	.
C			12	
G				8

Some numbers

↳ Score matrix
Scoring matrix
Substitution matrix

Example:

D, E, K are charged
V, I, L hydrophobic

$$s(D, E) =$$

$$s(D, K) =$$

$$s(E, K) =$$

$$s(V, I) =$$

$$s(V, L) =$$

$$s(I, L) =$$

$$s(D, V) =$$

$$s(D, I) =$$

$$s(D, L) =$$

$$s(E, V) =$$

$$s(E, I) =$$

$$s(E, L) =$$

$$s(K, V) =$$

$$s(K, I) =$$

$$s(K, L) =$$

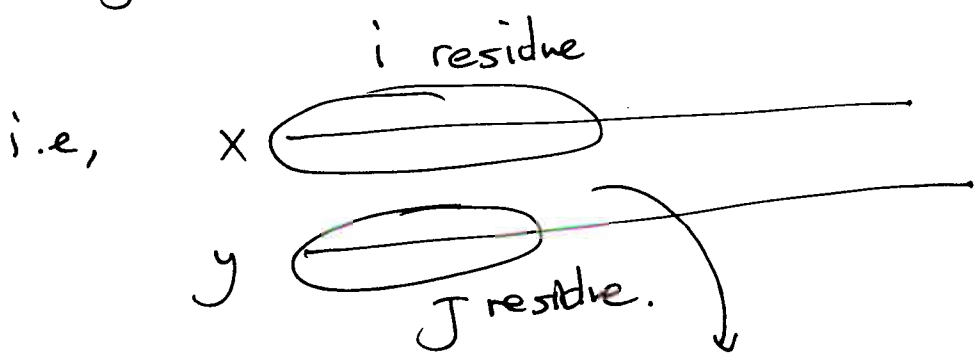
Alignment Algorithms

- When two seq of the same length,
 ⇒ One possible global alignment (if gaps not allowed)
- x is of length n
 y is of length n
 Global alignment
 Gaps are allowed } How many global alignments are possible?
 $C(2n, n) = \frac{(2n)!}{(n!)^2} \approx \frac{2^{2n}}{\sqrt{\pi n}}$
- Brute-force approach not possible
- Dynamic programming are fast & optimal
 ⇒ Build on previous/smaller solutions
- Heuristic algorithms are super fast, & nonoptimal.

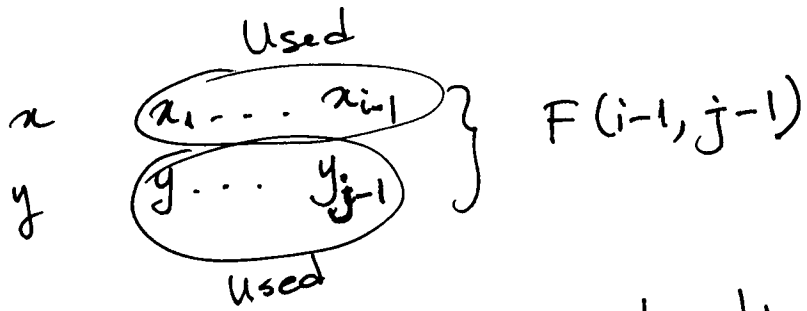
Global Alignments

x, y , x_i : i th residue in x

$F(i, j)$ = the best scoring match of $x_1 \dots i$ & $y_1 \dots j$



The diff is the gaps.

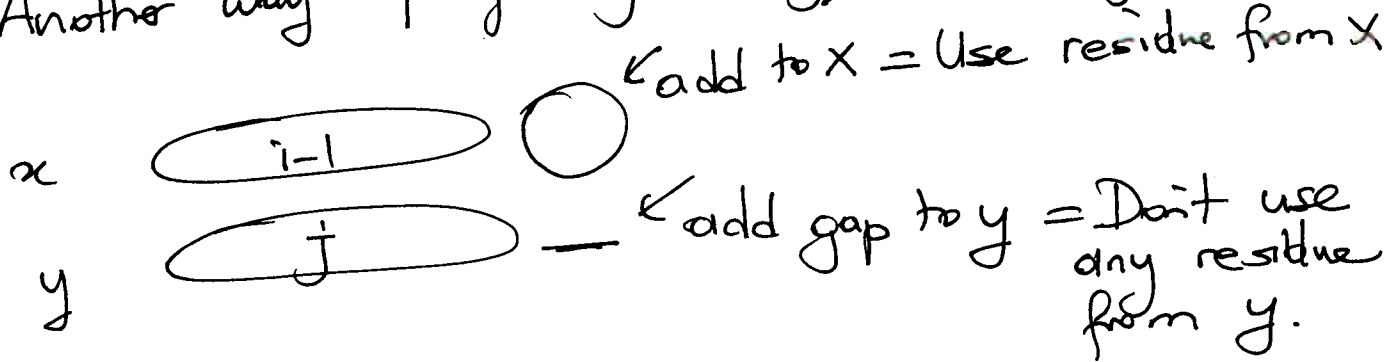


Note x_{i-1} does not necessarily align with y_{j-1} .

- To get $F(i, j)$, $F(i-1, j-1) + S(x_i, y_j)$
 - \Rightarrow Use one residue from x
 - Use one residue from y
 - Align them
 - \Rightarrow Align x_i & y_j

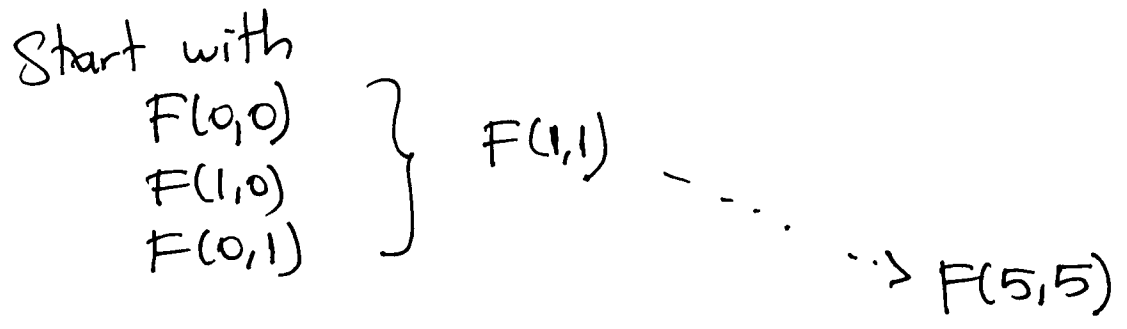
$$\Rightarrow F(i, j) = F(i-1, j-1) + S(x_i, y_j)$$

- Another way of getting $F(i, j)$ is through $F(i-1, j)$



$$\begin{aligned}
 F(i, j) &= F(i-1, j) + S(x_i, -) \\
 &= F(i-1, j) - d
 \end{aligned}$$

- Is this the best way of implementing?
- Bottom-up approach



- Recursive eqn's are sufficient ??
⇒ No, initial conditions
 $F(0,0) = 0$
 $F(i,0) = -id$
 $F(0,j) = -jd$