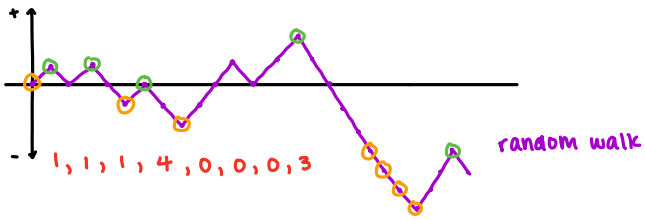
Ch 1 : The BLAST Algorithm

1.1 Random Walks

- ungapped alignment
  - ↳ match +1
    mismatch -1 } for DNA



def: Ladder point *
- points on the walk lower than any previously reached point

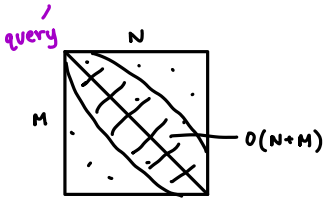
def: Excursion *
- highest point in the walk from the ladder point before the next ladder point

---

BLAST
- $O(N+M)$

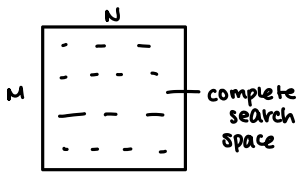
linear time
empirical
approximation

J = DB — database
$|J| = M$
$|Q| = N$
query



Smith-Waterman
- $O(NM)$

quadratic $N = M$
$O(N^2)$



* both local alignment algorithms

Protein Sequences :

T Q L A A W C R ...
R H L D D W R R ...
-1 1 5 -2 1 15 -4 7

BLOSUM scoring matrix

- consider an ungapped alignment of two protein sequences both of length N

The Null Hypothesis to be tested is that for each alignment pair of amino acids, the two amino acids were generated by a random process independently such that if amino acid j occurs with probability p_j at any position in the first sequence, and amino acid k occurs at any position in the second sequence with probability p_k, then the probability that the occur together in the alignment is:

$$\text{prob}(j,k) = p_j * p_k$$

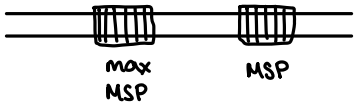
The Alternative Hypothesis

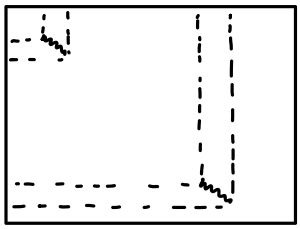
$$\text{prob}(j,k) = q(j,k)$$

* $q(j,k)$ function to be determined

$q(j,k)$ related to substitution matrices (BLOSUM, PAM) (scoring)

Maximum Segment Pair (MSP)





- highest scoring subsequences

BLAST Random Walk

alignment ⇒ score: cummulative

Score ⇒ random walk

N = number of positions in alignment

S(j,k) = score of aligning amino acid (aa) j with aa k

Scoring Matrix :

two axioms :

AX1 : at least one positive score

AX2 : The Null Hypothesis score to have negative mean

$$\sum_{j,k} p_j p_k S(j,k) < 0$$

↳ when the Null Hypothesis is true, the random walk has a negative drift and will go through a succession of increasingly negative ladder point

Let $Y_1, Y_2, \ldots$ be the highs of the excursions of this walk

Let Y_{MAX} be the max of these excursions

Y_{MAX} = the test stastistic of BLAST

It is necessary to find the Null Hypothesis of Y_{MAX}

The variables $Y_1, Y_2, \ldots$ are identically distributed random variables

The asymptotic distribution of the Y_i is a geometric-like distribution

$$P_r(Y \geq y) \approx Ce^{-y\lambda}$$

* constants C, λ depend on the substitution matrix

p_j, p_k' frequencies of aa

m = # ladder points

$$\boxed{P_r(Y_{MAX} \geq y) \approx 1 - e^{-mCe^{-\lambda y}}}$$

$$e^{-mCe^{-\lambda y}} \leq P_r(Y_{MAX} \leq y) \leq e^{-mCe^{-\lambda(y-1)}}$$

P-values for Y_{MAX}

$$1 - e^{-mCe^{-\lambda y}} \leq \text{P-value} \leq 1 - e^{-mCe^{-\lambda(y-1)}}$$