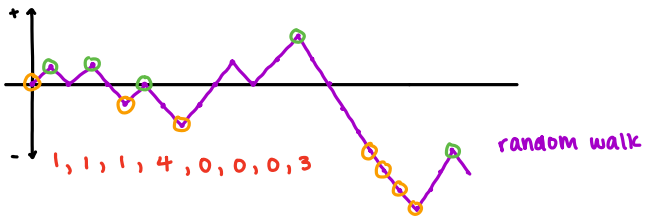
Ch 1 : The BLAST Algorithm

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

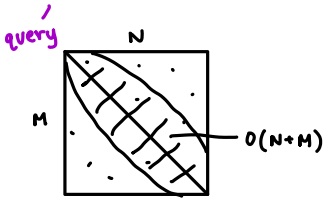
def: Excursion *
- highest point in the walk from the ladder point before the next ladder point

---

BLAST
- $O(N+M)$

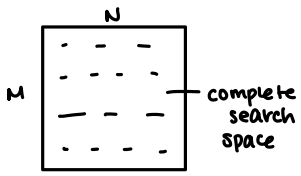
linear time
empirical
approximation

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:

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

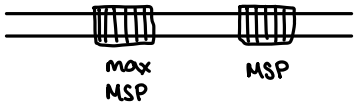
The Alternative Hypothesis

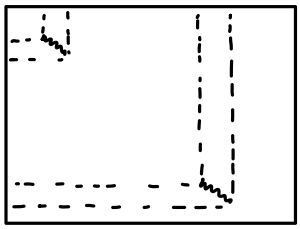
$$\text{prob}(j,k) = q(j,k)$$

* $q(j,k)$ 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}

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

Do the two sequences have a common ancestor? Homology

- random match, high score alignment is just by chance
- how long the alignment needs to be to be statistically significant?

RANDOM ≡ NOISE

De-noising : $\ell_{denoising} <$ align score

The Karlin-Altschul Theory of BLAST

- BLAST paper : S. Altschul, G. Myers, D. Lipman
  - ↳ Brown alumni : Applied Math, Biology
    Director of NCBI / NIH
    Engineering BLAST tool

5 axioms

AX1: The scoring matrix must have a positive entry
AX2: The expected score of the scoring matrix needs to be negative
] reasonable constraints for the used data

AX3: The letters of the random sequences in this model are independently and identically distributed (iid)
AX4: The sequences are infinitely long
AX5: Alignments do not contain gaps
] false

The Karlin-Altschul Equation

$$E = k \cdot m \cdot n \cdot e^{-\lambda S}$$

→ E = number of alignments expected by chance during a sequence database search, is a function of the size of the search space ($m \cdot n$), and the normalized score (λs), and constant (k)

→ m = size of the query

→ n = size of the database of sequences

NOTES

1. The relation step between the search space ($m+n$) and E = expected number of alignments by chance is linear
2. If the size of the search space is doubled, then the expected number of alignments E with a particular score S is expected to double
3. The relationship between E and the score S is exponential
   Small changes in score can lead to large differences in E

The Karlin-Altschul statistics provides a way to calculate just how long a sequence must be before it can produce an alignment with statistical significance

The minimum length is usually referred to as the expected HSP length

$$\ell = \frac{\ln(k \cdot m \cdot n)}{H}$$

H = relative entropy of the scoring matrix

Information Theory

BLAST Algorithm

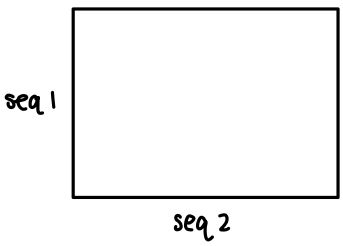
BLAST Program	DB	query
BLASTN	DNA	DNA
BLASTP	proteins	protein
BLASTX	proteins	DNA
TBLASTN	DNA	protein
TBLASTX	DNA → proteins	DNA

3 stages

I. SEEDING

II. EXTENSION

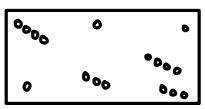
III. EVALUATION



SEEDING

Def : word of length k (k-mer)

word hit : ∘



Def : hit

- two $=$ just identical strings not good enough
- A concept : Neighborhood

Def : The neighborhood of a word contains the word w itself and all of the words of equal length over the input alphabet Σ that have an alignment score with w at least T (score in the scoring matrix)

$\underbrace{=}_{\text{score 1}}\ \underbrace{=}_{\text{score 2}}\ \underbrace{=}_{\text{score 3}}\ \ldots\ =$

score $i > T \Rightarrow \alpha_i$ in the neighborhood of N

- adjusting T controls the size of the neighborhood

BLASTN : T never used

α = size of word (= 7)

BLASTP : α = 2 or 3

T = 999 for α = 3

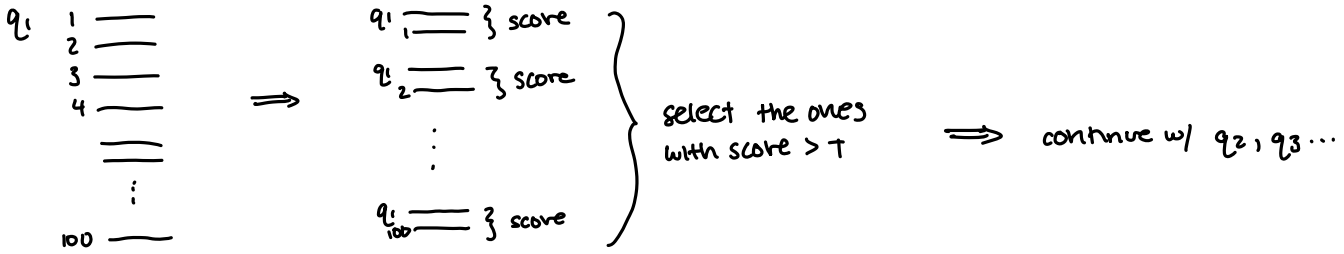
A database query algorithm:

Q = query sequence "local alignments between Q and sequences from DB"

DB = sequences

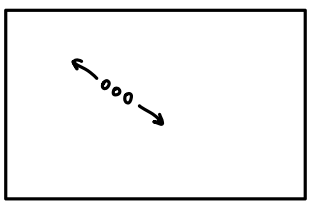
- Q all the k-mers of Q

  $q_1, q_2, \dots q_m$



All k-mers from Σ^k with > T score of alignment to a k-mer in Q = hits

EXTENSION



* no gap extensions

- how much negative score do you tolerate? Z bound — lowest score you allow
- heuristic clusterings of hits clusters
- want as many high-scoring segment pairs (HSPs) as possible

EVALUATION

The seeds are extended in both directions to create ungapped alignments that are evaluated to be statistically significant by using the Karlin-Altschul statistic

We use a threshold S to sort alignments into low and high scoring

- S, E are related to each other in the Karlin-Altschul Eq.

  So a score threshold is synonymous to a statistical threshold

- However, there are complex dependencies

Random Walk Theory

- Special cases of Markov chains
- Random walk theory supplies the basic probability theorem of BLAST

BLAST: searches for high-scoring local alignments between two sequences and then tests for significance of the scores found via p-values

- The p-value calculation takes into account the lengths of the sequences in value since longer sequences are more likely to be significant by chance

The Simple Random Walk

A random process: starts at an arbitrary point h and moves inaependently of the past history

a step down every unit of time with probability q, or

a step up with probability p

$p+q=1$

h = initial position of the walk

This walk is restricted to the interval $[a, b]$, where a and b are integers with

$$a < h < b,$$

and the walk stops when it reaches a or b

Two fundemental questions:

Q1: what is the probability that eventually the walk finishes at b rather than a?

Q2: what is the mean number of steps until the walk stops?

Difference Equations

Def absorption probability

let w_h = the prob that the simple random walk eventually finishes, or it is absorbed at point b rather than a given the initial point h

$$w_h = pw_{h+1} + qw_{h-1}$$
$$w_a = 0, \quad w_b = 1$$

a homogeneous difference eq. with boundary conditions

A solution of the eq is a set of values for w_h that satisfy the eq. for all integer values of h

One solution of the eq is of the form:

$$w_h = e^{\theta h}$$

for some constant θ

To solve for θ, we substitute in the eq

$$e^{\theta n} = pe^{\theta(n+1)} + qe^{\theta(n-1)} \qquad \text{multiply both sides by } e^{\theta - \theta n}$$

$$e^{\theta - \theta n} e^{\theta n} = e^{\theta - \theta n} pe^{\theta(n+1)} + e^{\theta - \theta n} qe^{\theta(n-1)}$$

$$e^{\theta} = pe^{2\theta} + q$$

$$p^{2\theta} - p^{\theta} + q = 0$$

$$x = p^{\theta} \implies px^2 - x + q = 0 \begin{cases} x_1 = 1 \\ x_2 = \frac{q}{p} \end{cases}$$

$$p \neq q$$

$$e^{\theta} = 1, \quad e^{\theta} = \frac{q}{p}$$

$$\Downarrow$$

$$\theta = 0, \quad \theta = \log\left(\frac{q}{p}\right)$$

General solution is any given combination of the two solutions

$$\underline{A1}: \quad w_n = \frac{e^{\theta * n} - e^{\theta * a}}{e^{\theta * b} - e^{\theta * a}} \qquad \text{finishes at } b$$

$$\underline{A'1}: \quad u_n = \frac{e^{\theta * b} - e^{\theta * n}}{e^{\theta * b} - e^{\theta * a}} \qquad \text{finishes at } a$$

$$w_n + u_n = 1$$

$$\underline{A2}: \quad m_n = \frac{n-a}{q-p} - \left(\frac{b-a}{q-p}\right)\left(\frac{e^{\theta * n} - e^{\theta * a}}{e^{\theta * b} - e^{\theta * a}}\right)$$

Extension (continued)

+1 match
−1 mismatch

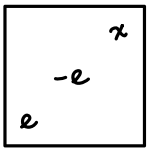
THE QUICK BROWN FOX JUMPS OVER THE LAZY DOG
THE QUIET BROWN CAT PURRS WHEN SHE SEES HIM

score → 123 45654 56789 876 56545 4343 234 3210 -1 -2 -3
dropoff score → 000 00012 10000 123 40120 ...

$x = 5$ (tolerate in the negative decreasing score)

Evaluate the significance of the HSP scores

GUMBEL extreme value distribution



$Y_1, Y_2, \ldots, Y_N \qquad Y_i = ce^{-\lambda x}$ ← exponentially distributed

$Y_{MAX} = MAX[Y_1, \ldots Y_N]$

$$P(S \geq x) = 1 - e^{-e^{-\lambda(x-u)}}$$

$$u = \frac{\log(k m' n')}{\lambda}, \qquad \lambda, k \text{ constants}$$

m = size of query

n = size of DB

$m' = m - \frac{\ln(Kmn)}{H}$

$n' = n - \frac{\ln(Kmn)}{H}$

H = average expected score per aligned pairs of aa of two random sequences

Making two or more HSPs into a longer alignment

- two methods : Poisson method and sum of squares method

Karlin-Altschul Theory

1. single sequences of numbers

2. pair of aligned sequences

The single sequence case

-2, 20, 13, -5, -100, 31, -1, 21, -5, -10

- charge, volume, hydrophobicity, secondary structure α-helices

① Introduce a statistical model that provides precise formulas for accessing statistical significance of any subsequence with high aggregate score (aggregate = summing)

② set of results describing the statistical properties of the high scoring segments

* Line of research is focusing on the fundamental question :
Which scoring schemes are "optimal" for distinguishing biologically relevant patterns?

Theory

Given an alphabet $\Sigma = \{a_1, a_2, \dots a_n\}$ we define a Random model m_0 as follows :

- a random sequence consists of letters sampled independently from Σ with respective prob. $p_1, p_2, \dots p_n$

  $a_1, a_2, \dots a_n$

  associated with each letter a_i there is a score s_i

PB1: Study the segments of the sequence with greatest aggregate additive score

"max segment" score

Two conditions on the scores

① at least one score is positive

② the expected score per letter $E = \sum_{i=1}^{n} p_i s_i < 0$

Two Theorems

To assess the stat. significance of high-scoring segments (subsequences), we need to know the probability distribution for max-segment scores from random sequences of length n

The theorems use a key parameter λ^*, which is the unique positive solution of the equation:

$$\sum_{i=1}^{n} p_i e^{\lambda s_i} = 1$$

* observe 0 is a solution also

let $M(n)$ denote the length of the max score segment

$$\tilde{M}(n) = M(n) - \frac{\ln(n)}{\lambda^*}$$

$\tilde{M}(n)$ = centered max segment score

THEOREM 1

The random variable $\tilde{M}(n)$ has the close approximating distribution

$$P[\tilde{M}(n) > x] \cong 1 - e^{-Ke^{-\lambda^* x}}$$

THEOREM 2

As the length of the random sequence grows $n \to \infty$, the frequency of letter a_i in any sufficiently high scoring segment approaches

$$p_i e^{\lambda^* s_i}$$

with probability = 1

Dayhoff

PAM

1PAM = 1% aa substitution

LOD scores (logs odds ratio)

$$\text{lod(pair of aa)} = \log_2\left(\frac{\text{obs. freq.}}{\text{expected freq.}}\right)$$

↑ random model

score $s_{ij} = \log \frac{q_{ij}}{p_i p_j}$

$p_i p_j$ freq of $a_i a_j$ aa

* If obs. freq = exp. freq

  $\Rightarrow$ LOD = 0

* If LOD > 0 then $\frac{a_i}{a_j}$ is common
* If LOD < 0 then $\frac{a_i}{a_j}$ is unlikely

Converting raw scores to normalized scores requires matrix specific constant λ

Sum of target frequencies

$$\sum_{i=1}^{n}\sum_{j=1}^{i} q_{ij} = 1 \quad \Rightarrow \quad \lambda s_{ij} = \log_e\left(\frac{q_{ij}}{p_i p_j}\right)$$

To solve for λ :

$$\Sigma\Sigma\, q_{ij} = \Sigma\Sigma\, p_i p_j e^{\lambda s_{ij}} = 1$$

λ : expectation score for every HSP

Relative entropy

The expected score of a scoring matrix is the sum of its raw scores multiplied by the frequencies $\boxed{E = \sum_{i=1}^{20}\sum_{j=1}^{i} p_i p_j s_{ij}}$ ← raw scores

Relative entropy of a scoring matrix (H) summarizes the general properties of the matrix

$$\boxed{H = -\sum_{i=1}^{20}\sum_{j=1}^{i} q_{ij}\lambda s_{ij}}$$ ← normalized score

H = average number of bits per position in an alignment and it is always positive