Significance of local alignments

- When computing local alignments, the first question is often: Is this local alignment occurring by chance or not.
- More precisely, one is interested in quantifying the statistical significance of a local alignment, based on a simple model of random sequences.
- The statistical significance is usually expressed by an expectation value (E-value for short).
- The smaller the E-value, the more significant the local alignment.

Significance of local alignments

- In this section we want to determine an E-value for a local alignment in which the number of mismatches is smaller than some threshold.
- We exclude alignments with indels i.e. focus on ungapped alignments.
- This is because gapped alignments are more difficult to handle statistically, using e.g. the Karlin-Altschul Statistics which we have considered as a black box at the end of the section on Blast.
- In this section, the cost of the local alignment is thus measured in terms of the hamming distance of the aligned pair of substrings.
- We describe all details of an approach to calculate E-values.
- For two strings x and y of equal length, define

$$\mathcal{H}(x, y) = |\{i \mid 1 \leq i \leq |x|, x[i] \neq y[i]\}|.$$

$\mathcal{H}(x, y)$ is the hamming distance of x and y.

Significance of local alignments

- Now consider two sequences u and v of length m and n, respectively.
- We want to determine pairs of substrings in u and v allowing for a maximum number $k \geq 0$ of mismatches in these substrings.
- So k is a *mismatch threshold*.

A *local k-mismatch alignment of u and v* (k-LMA, for short) of length ℓ is a triple (ℓ, i, j) such that the following holds:

- $k < \ell$, i.e. the number of mismatches is smaller than the length of the match.
- $1 \leq i \leq m - \ell + 1$ and $1 \leq j \leq n - \ell + 1$, i.e. $u[i \ldots i + \ell - 1]$ is a substring of u and $v[j \ldots j + \ell - 1]$ is a substring of v, both of length ℓ.
- $\mathcal{H}(u[i \ldots i + \ell - 1], v[j \ldots j + \ell - 1]) \leq k$, i.e. the number of mismatches between $u[i \ldots i + \ell - 1]$ and $v[j \ldots j + \ell - 1]$ is at most k.

Significance of local alignments

- It makes sense to extend each alignment maximally to both sides.
- Hence we define a notion of maximality: (ℓ, i, j) is *left-maximal* if either $i = 1$ or $j = 1$ or $u[i-1] \neq v[j-1]$.
- (ℓ, i, j) is *right maximal* if either $i + \ell = m + 1$ or $j + \ell = n + 1$ or $u[i+\ell] \neq v[j+\ell]$.
- The following figure illustrates these notions:

$$u = \underbrace{\overset{i}{a}\quad\overset{i+\ell-1}{\quad}\overset{}{b}}_{u[i\ldots i+\ell-1]} = \underbrace{\overset{j}{c}\quad\overset{j+\ell-1}{\quad}\overset{}{d}}_{v[j\ldots j+\ell-1]} = v$$

$$\text{left maximal} \iff u[i-1] = a \neq c = v[j-1]$$
$$\text{right maximal} \iff u[i+\ell] = b \neq d = v[j+\ell]$$

- (ℓ, i, j) is *maximal* if it is left-maximal and right-maximal.

Significance of local alignments

– To assess the significance of a k-LMA, one determines its *E-value*, which is the size of the following set:

$$\{(\ell', i', j') \mid (\ell', i', j') \text{ is } k'\text{-LMA in } u' \text{ and } v',\ \ell \leq \ell', k' \leq k\}$$

where u' and v' are random sequences of length $|u'| = |u|$, $|v'| = |v|$.

– That is, the E-value is the number of local alignments of the same length or longer and with the same or a fewer number of mismatches, that occur between two random sequences of the same length as u and v.

– As a model of random sequences, we assume the uniform Bernoulli model.

– That is, we assume that at each position of a random sequence, each of the r characters in $\mathcal{A}$ occurs with the same probability $p = \frac{1}{r}$.

Restricting to exact matches

- Let us first restrict to the case that $k = 0$, in which case we have no mismatches in the local alignments (i.e. they represent *maximal exact matches*).
- We therefore use the term "maximal exact match" instead of k-LMA.
- We first show an important property of maximal exact matches of length $\geq \ell$.

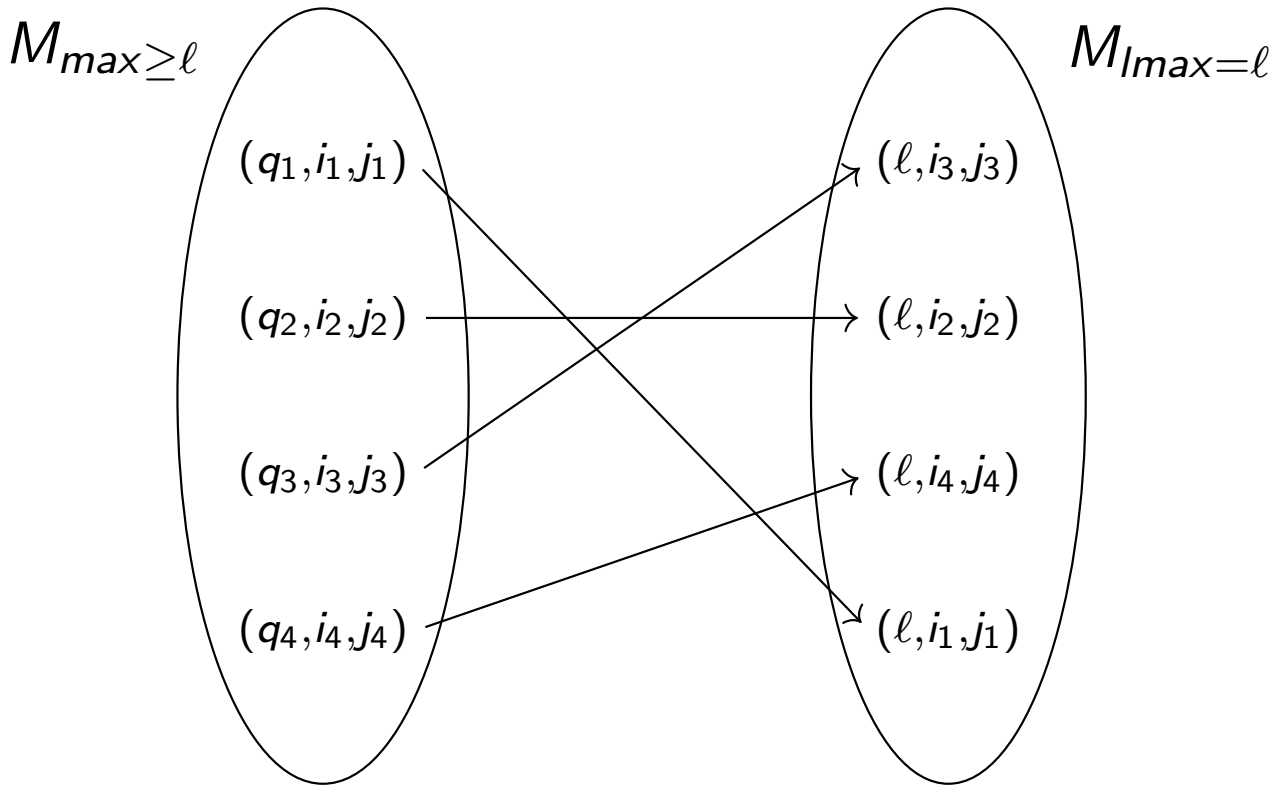
Lemma 1

For any pair of sequences u and v, the number of maximal exact matches of length $\geq \ell$ equals the number of left-maximal exact matches of length exactly ℓ.

Proof

- Let $M_{max \geq \ell}$ be the set of maximal exact matches of length $\geq \ell$ and $M_{lmax=\ell}$ be the set of left-maximal exact matches of length exactly ℓ.
- We have to show that these two sets are of the same size.
- This property is shown by constructing a bijective mapping between the two sets, see the following illustration:

Proof



Bijective means one-by-one, that is, we can map each element in $M_{max \geq \ell}$ to a unique element in $M_{lmax=\ell}$ and vice versa.

Proof

- The mapping is as follows: Each maximal exact match (q, i, j) for $q \geq \ell$ (thus it is an element in $M_{max \geq \ell}$) is mapped to (ℓ, i, j) (which is an element in $M_{lmax=\ell}$).
- (ℓ, i, j) is left maximal, but not right-maximal whenever $q > \ell$.
- Let us denote this mapping by γ (pronounce as gamma).
- To show that γ is bijective, one first needs to show that it is injective, i.e. $\gamma((q, i, j)) \neq \gamma((q', i', j'))$ for all pairs of different maximal exact matches (q, i, j) and (q', i', j').
- In terms of the illustration above, this means that two different arcs never point to the same element in the set on the right.
- So let (q, i, j) and (q', i', j') be different maximal exact matches.
- Suppose that $(i, j) = (i', j')$.

Proof

- As $(q, i, j) \neq (q', i', j')$ we conclude $q \neq q'$.
- This is a contradiction, since the length of a maximal exact match at a given pair of positions (i, j) is uniquely determined.
- Hence $(i, j) \neq (i', j')$ which implies $\gamma((q, i, j)) = (\ell, i, j) \neq (\ell, i', j') = \gamma((q', i', j'))$.
- Next we have to show that γ is surjective, i.e. for each left-maximal match (ℓ, i, j) there must be some maximal match (q, i, j) such that $\gamma((q, i, j)) = (\ell, i, j)$.
- In terms of the illustration above, this means that for any element in the set on the right, there is at least one arc pointing to it.
- But this is of course the case, since each left-maximal match (ℓ, i, j) can be extended on the right to a maximal match (q, i, j) of length $q \geq \ell$.
- The fact that γ is injective and surjective implies that it is bijective.
- As a consequence, $M_{max \geq \ell}$ and $M_{lmax=\ell}$ are of the same size.

Restricting to exact matches

We use the following notions and notations:

- $\mathbb{E}[\#$ events$]$ is the expected number of events
- in our case the events are maximal exact matches of length $\geq \ell$ or left-maximal matches of length $= \ell$.
- the expected number of events can in our case be determined as follows:
  - consider each possible position where an event can occur
  - determine the probability that the event occurs at the considered position
- the probability of an event occurring at position p is denoted by $Pr[$event occurs at position $p]$
- So $\mathbb{E}[\#$ events$] = \sum_{\text{position } p} Pr[$event occurs at position $p]$

Restricting to exact matches

Ignoring the effects of boundary cases (where $i = 1$ or $j = 1$ or $i + \ell - 1 \geq m$ or $j + \ell - 1 \geq n$), we obtain the following equations

$$
\begin{aligned}
&\mathbb{E}[\#\text{ of maximal exact matches of length} \geq \ell] \\
&= \mathbb{E}[\#\text{ of left-maximal exact matches of length } \ell] \\
&= \sum_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}} Pr[(\ell, i, j) \text{ is a left-maximal exact match}] \\
&= \sum_{i,j} Pr[u[i \ldots i + \ell - 1] = v[j \ldots j + \ell - 1], u[i-1] \neq v[j-1]] \\
&= \sum_{i,j} Pr[u[i \ldots i + \ell - 1] = v[j \ldots j + \ell - 1]] \cdot Pr[u[i-1] \neq v[j-1]] \\
&= \sum_{i,j} p^{\ell}(1-p) = m\ n\ p^{\ell}\ (1-p) \text{ where } p = \frac{1}{r}
\end{aligned}
$$

Restricting to exact matches

- Note that the second equality is due to the fact that we can count the number of left-maximal exact matches of length ℓ by looking at each pair of positions i and j, considering the probability of a match at that position and summing up all such probabilities.
- So we have finally shown that

$$\mathbb{E}[\#\text{ of maximal exact matches of length} \geq \ell] = m\ n\ p^{\ell}\ (1-p).$$

Example 1 shows some E-values of maximal exact matches of length $\geq \ell$.

Allowing for mismatches

- E-values for k-LMAs, $k > 0$, can be computed in a similar way.
- First, assume fixed values for ℓ and k.
- To determine the E-values in the general case, we first have to compute the number of choices of k positions with mismatches from ℓ possible positions in the sequences.
- We can apply a standard formula from combinatorics:

Given a set X of size ℓ, there are exactly

$$\binom{\ell}{k} = \frac{\ell!}{k!(\ell - k)!} \tag{1}$$

subsets $Y \subseteq X$ such that Y has exactly k elements.

Allowing for mismatches

- To see this, consider the following: at first, we have ℓ choices for the first element of a subset, $\ell - 1$ choices for the second element of a subset etc.
- In general, we have $(\ell - (q-1))$ choices for the qth element of the subset, for all q, $1 \leq q \leq \ell$.
- All combinations of choices are possible.
- Hence if we make the choices one after the other, we have

$$\prod_{q=1}^{k} (\ell - (q-1)) = \ell \cdot (\ell - 1) \cdot \cdots \cdot (\ell - (k-1)) = \frac{\ell!}{(\ell - k)!} \qquad (2)$$

choices altogether to obtain the subsets.

Allowing for mismatches

- However, many choices lead to the same sets.
- In particular, if we choose the same elements in a different order, we obtain the same subset.
- More precisely, each permutation of each subset of size k is generated.
- Now the number of permutations of a set of size k is $k!$.

We however only want one subset for each such permutation of choices and thus we have to divide (2) $\left(i.e.\frac{\ell!}{(\ell-k)!}\right)$ by $k!$ to obtain (1) $\left(i.e.\frac{\ell!}{k!(\ell-k)!}\right)$.

Allowing for mismatches

- Now we apply the above formula.
- At first note that there are $\binom{\ell}{k}$ choices for k positions out of ℓ possible positions in two strings of length ℓ.
- For each of the k positions, there is a mismatch with probability $1 - p$.
- Hence, since the positions are independent, $(1 - p)^k$ is the probability that there is a mismatch in all k positions.
- Moreover, $p^{\ell-k}$ is the probability that there is a match in all remaining $\ell - k$ positions of the strings.
- Hence, the probability of two independent random sequences x and y, both of length ℓ, to have a hamming distance of exactly k is

$$Pr[\mathcal{H}(x, y) = k] = \binom{\ell}{k} p^{\ell-k}(1 - p)^k.$$

Allowing for mismatches

- To compute the expected number of LMAs of length ℓ or longer and with k or fewer mismatches, one has to sum over all possible $k' \leq k$ and over all lengths $\ell' \geq \ell$.
- The latter is necessary, in contrast to the case of exact substring matches, because for k-LMAs it is no longer true that the number of LMAs of length $\geq \ell$ equals the number of left-maximal LMAs of length exactly ℓ.
- Hence, if we ignore boundary cases, we obtain:

Allowing for mismatches

$$\mathbb{E}[\#\text{ of maximal} \leq k\text{-LMAs of length} \geq \ell]$$

$$= \sum_{k'=0}^{k} \sum_{\ell'=\ell}^{\min(m,n)} \sum_{i\in[1,m],j\in[1,n]} Pr[(\ell', i, j) \text{ is a maximal } k'\text{-LMA}]$$

$$= \sum_{k'=0}^{k} \sum_{\ell'=\ell}^{\min(m,n)} \sum_{i\in[1,m],j\in[1,n]} Pr[\mathcal{H}(u[i \ldots i+\ell'-1], v[j \ldots j+\ell'-1]) = k' \text{ and } u[i-1] \neq v[j-1] \text{ and } u[i+\ell'] \neq v[j+\ell']]$$

$$= \sum_{k'=0}^{k} \sum_{\ell'=\ell}^{\min(m,n)} \sum_{i\in[1,m],j\in[1,n]} Pr[\mathcal{H}(u[i \ldots i+\ell'-1], v[j \ldots j+\ell'-1]) = k'] \cdot Pr[u[i-1] \neq v[j-1]] \cdot Pr[u[i+\ell'] \neq v[j+\ell']]$$

$$= \sum_{k'=0}^{k} \sum_{\ell'=\ell}^{\min(m,n)} \sum_{i\in[1,m],j\in[1,n]} \binom{\ell'}{k'} p^{\ell'-k'} (1-p)^{k'} (1-p)(1-p)$$

$$= \sum_{i\in[1,m],j\in[1,n]} \sum_{k'=0}^{k} \sum_{\ell'=\ell}^{\min(m,n)} \binom{\ell'}{k'} p^{\ell'-k'} (1-p)^{k'+2}$$

$$= m\ n \sum_{k'=0}^{k} \sum_{\ell'=\ell}^{\min(m,n)} \binom{\ell'}{k'} p^{\ell'-k'} (1-p)^{k'+2}.$$

Because the sums are largely dominated by the terms for $k' = k$ and $\ell' = \ell$, this can be approximated by

$$m\ n \binom{\ell}{k} p^{\ell-k}(1-p)^{k+2}$$

Example 1

Consider two DNA sequences of length $m = n = 10^6$. Let $p = \frac{1}{4}$. Then the E-values of an k-LMA of length $\geq \ell$ for different values of $\ell \in \{16, 32, 64, 128\}$ and $k \in \{0, 1, 2, 3\}$ are shown in the following table:

	k			
ℓ	0	1	2	3
16	$1.3 \cdot 10^{2}$	$6.3 \cdot 10^{3}$	$1.4 \cdot 10^{5}$	$2.0 \cdot 10^{6}$
32	$3.0 \cdot 10^{-8}$	$2.9 \cdot 10^{-6}$	$1.4 \cdot 10^{-4}$	$4.1 \cdot 10^{-3}$
64	$1.7 \cdot 10^{-27}$	$3.2 \cdot 10^{-25}$	$3.0 \cdot 10^{-23}$	$1.9 \cdot 10^{-21}$
128	$4.9 \cdot 10^{-66}$	$1.9 \cdot 10^{-63}$	$3.6 \cdot 10^{-61}$	$4.5 \cdot 10^{-59}$