Statistical Significance of Local Alignments

Zhiping Weng
Boston University

Erdos-Renyi Law for the Longest Run of Hs

• A \(N\)-amino-acid-long random sequence is made up of two types of amino acids: hydrophobic (H) and polar (P). The occurrence of Hs or Ps at each position of the sequence is random, with probabilities \(P(H) = p\) and \(P(P) = 1-p\).

• What is the expected length \(K\) of the longest run of Hs in this random sequence?
An Intuitive explanation for the Expected length of Longest Run of Hs

Randomly pick a position at the random sequence (N=31 in the above example). The probability that K residues (K=5 in the above example) from that point are all Hs is: \( p^K \). There are \( (N-(K-1)) \) ways such picking can be done, therefore the frequency of observing such a run is \( p^K \times (N-(K-1)) \). A theory that is very hard to prove indicates when a frequency is a small value, it can be treated as a probability. Since probability can at most be 1, by setting \( p^K \times (N-(K-1)) \) to 1, we solve \( K \) to be:

\[
K = \log_{1/p} N \quad \text{For a large } N
\]

The Erdos-Renyi law says:

\[
P(\lim_{N \to \infty} \frac{K}{\log_{1/p} N} = 1) = 1
\]

\( P(\quad ) \): The probability of an event

Or:

\[
\lim_{N \to \infty} (\log_{1/p} N) = K
\]
The expected length $K$ of the longest run of matches between two random sequences

Let us make a dot plot. A match has a score of 1, and a mismatch has a score of 0. The probability of having a single match is $p$. A run of 3 consecutive matches is shown in the above plot.

If the lengths of the two sequences are $N_1$ and $N_2$, the frequency of a $K$-long match is $p^K(N_1-(K-1))(N_2-(K-1))$. By setting it to 1, we obtain:

$$K = \log_{1/p} (N_1 \ast N_2) \quad \text{For a large } N_1, N_2$$
Why is percent identity (%id) not a good indicator for the significance of sequence comparison?

\[ N_1 = N_2 = 100; \quad p = 0.5 \quad \text{and} \quad N_1 = N_2 = 1000; \quad p = 0.5 \]

\[ K = \log_{1/p} N_1 N_2 = 13 \quad \text{and} \quad K = \log_{1/p} N_1 N_2 = 20 \]

When you compare two 100-amino-acid-long random sequences, a run of 13 perfect matches happens by chance. For two 1000-amino-acid-long random sequences, a run of 20 perfect matches happens by chance. Therefore a run of 18 perfect matches is (somewhat) significant for the former but insignificant by any mean for the latter. However %id does not take into account such a (log)-length-dependence (and many other factors).

---

**Length & Alignment Score**

Local alignment scores of human cytochrome C against the Swiss-Prot database
Equivalent Questions:

- I obtained a score $S$ for my sequence analysis. How significant is this value $S$?
- What is the probability of obtaining a score $S$ for a random sequence?

Assessing the Statistical Significance of Any Region with High Aggregate Score (Measured by General Scoring Schemes): Defining the known parameters

A N-letter long random sequence completely defined by the following properties:

- $\{a_1, a_2, ..., a_r\}$ Alphabet
- $\{p_1, p_2, ..., p_r\}$ Letters of a sequences are generated in an independent fashion such that letter $a_j$ is selected with probability $p_j$
- $\{s_1, s_2, ..., s_r\}$ $s_j$ is the score for the letter $a_j$. It describes the desirable property we would like to detect, such as hydrophobicity, secondary structure propensity, homology to a target sequence. Very conveniently we can express $s_i$ as a “log-likelihood” score of a set of target frequencies $\{q_1, q_2, ..., q_r\}$: $s_i = \log \left( \frac{q_i}{p_i} \right)$
Assessing the Statistical Significance of Any Region with High Aggregate Score (Measured by General Scoring Schemes): What is the question?

We are interested in the segment of a sequence with the greatest aggregate (additive) score. This segment is called “maximal segment”. Its score is called “maximal segment score”, denoted as $M$.

We want to solve for the Probability Distribution Function (p.d.f) and the Cumulative Distribution Function (c.d.f) of $M$ in random sequences.

Assessing the Statistical Significance of Any Region with High Aggregate Score (Measured by General Scoring Schemes): Constraints on $\{s_i\}$

- At least one of the $s_i$ is positive. Otherwise you will end up with (on average) 0-length maximal segment.
- The expected score per letter is negative. Otherwise you will end up with (on average) the whole sequence as the maximal segment.

$$\sum_{i=1}^{r} p_i s_i < 0$$

For protein sequences, $r = 20$

- Log-likelihood scores naturally satisfy the above two constraints.
Assessing the Statistical Significance of Any Region with High Aggregate Score (Measured by General Scoring Schemes): What is the question?

\[ F(m) = c.d.f(M, m) \equiv P(M \leq m) \]

\[ f(m) = p.d.f(M, m) \equiv \lim_{\Delta x \to 0} \frac{P(m - \Delta m \leq M \leq m + \Delta m)}{2 \cdot \Delta m} \]

\( P(\ ) \): The probability of an event

\[ F(m) = \int_{x' = -\infty}^{x} f(m') \, dm' \]

\[ f(m) = \frac{d}{dm}(F(m)) \]

\[ F(m) = c.d.f(M, m) \] gives us a measurement for the statistical significance of score \( M \).

Review for the Binomial Distribution

Assume the probability of getting a head for one coin tossing is \( p \). We do \( N \) tossing. The probability of getting \( L \) heads (the remaining \( N-L \) are tails) is called the binomial distribution:

\[ P(L\, \text{heads}) = C_L^N p^L (1 - p)^{N-L} \]

\[ C_L^N = \frac{N!}{(N-L)!L!} \]
Review for the Poisson Distribution

Poisson Distribution is closely connected to the Binomial distribution. It is the distribution of the number of occurrences of rare events, that is, an event with small probability \( p \), in \( N \) independent trials.

**Binomial:**

\[
P(L \text{ heads}) = C_L^N \, p^L \, (1 - p)^{N-L}
\]

\[\begin{align*}
p & \to 0 \\
N & \to \infty \\
NP & \to \lambda
\end{align*}\]

**Poisson:**

\[
P(L \text{ heads}) = e^{-\lambda} \frac{\lambda^L}{L!}
\]

The Number of Unrelated Maximal Segments

The number of unrelated matches \( I \) with score \( M \geq m \) is approximately Poisson distributed, with mean:

\[
E(N) = K \cdot Ne^{-2\lambda}
\]

\[
P(I = i) = \frac{(E(N))^i \, e^{-E(N)}}{i!}
\]

\( \lambda \) and \( K \) are constants that can be computed from \( \{p_1, p_2, ..., p_r\} \) and \( \{s_1, s_2, ..., s_r\} \).

\( \lambda \) is the solution of the equation \( \sum_{i=1}^{r} p_i e^{i\lambda} = 1 \). It is a scale parameter to convert aggregate scores into a natural scale.

\( K \) corrects for the non-independence of possible starting points for matches.

The way to compute \( K \) can be found in Karlin & Altschul (1990), PNAS, 87:2264.
The Probability of Observing at Least One Maximal Segment with Score $M \geq m$

$$P(M \geq m) = P(I \geq 1)$$
$$= 1 - P(I = 0)$$
$$= 1 - \left(\sum_{i=0}^{\infty} e^{-E(N)} \frac{E(N)^i}{i!}\right)^0$$
$$= 1 - e^{-E(N)}$$
$$= 1 - e^{-KNm}$$

---

**EVD**

*Figure 2.13* Left, a scatter plot of the distribution of local match scores obtained from comparing human cytochrome C (SWISS-PROT accession code P000001) against the SWISS-PROT34 protein database with the Smith–Waterman implementation SSEARCH [Pearson 1996]. Right, the corresponding length-normalised distribution of scores, showing the fit to an EVD distribution.
Intermission

*M and S Satisfies the Extreme Value Distribution (EVD)*

A normalized score: $S = \lambda * M - \ln(K * N) $

$$f(x) = p d.f(S, x) = e^{-x} e^{-x}$$

$$F(x) = c d.f(S, x) = P(S \leq x) = e^{-e^{-x}}$$
The Extreme Value Distribution

For large value of $x$:

$p.d.f (S, x) = e^{-x} e^{-e^{-x}}$

$\approx e^{-x} (1 - e^{-x} + \frac{e^{-2x}}{2!} - \frac{e^{-3x}}{3!} + ...) \approx e^{-x}$

Compare to a Gaussian Distribution with mean 0 and standard deviation 1:

$p.d.f (T, x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$
Example

Question: For a N-residue-long random protein sequence, how high does the maximal segment score $M$ need to be in order to achieve the 0.99-level significance?

Solution:

\[
\text{c.d.f. } (S, x) = P(S \leq x) = e^{-e^{-x}} = 0.99 \\
x = -\ln(-\ln 0.99) \\
M = \frac{S + \ln(K \cdot N)}{\lambda} > -\frac{\ln(-\ln 0.99) + \ln(K \cdot N)}{\lambda}
\]
The Statistical Significance of Pairwise Sequence Comparison: Very Slight Difference From the Single Sequence Case.

- \{a_1, a_2, \ldots, a_r\} Alphabet
- \{p_1, p_2, \ldots, p_r\} & \{p'_1, p'_2, \ldots, p'_r\}
- \{s_{11}, s_{12}, \ldots, s_{1r}, s_{21}, s_{22}, \ldots, s_{2r}, \ldots, s_{r1}, s_{r2}, \ldots, s_{rr}\} s_{ij} is the similarity score between letter \(a_i\) and \(a_j\). It is frequently the elements of a “log-likelihood” scoring matrix (such as PAM and BLOSUM): \(s_{ij} = \log (\frac{p_{ij}}{p_ip_j})\)

The Statistical Significance of Pairwise Sequence Comparison: \(\lambda\) and \(K\).

\(\lambda\) and \(K\) are constants that can be computed from \(\{p_1, p_2, \ldots, p_r\}, \{p'_1, p'_2, \ldots, p'_r\}\) and \(\{s_{11}, s_{12}, \ldots, s_{rr}\}\).

\(\lambda\) is the solution of the equation \(\sum_{i=1}^{r} \sum_{j=1}^{r} p_i p_j e^{\lambda s_{ij}} = 1\)

The way to compute \(K\) can be found in Karlin & Altschul (1990), PNAS, 87:2264
The Statistical Significance of Pairwise Sequence Comparison: p.d.f and c.d.f

A normalized score:
\[ S = \lambda * M - \ln(K * N_1 N_2) \]

\( M \) and \( S \) satisfy the extreme value distribution

\[ p.d.f (S, x) = e^{-x} e^{-e^{-x}} \]

\[ c.d.f (S, x) = e^{-e^{-x}} \]

What if we want to consider several “separate” high-scoring segments?

The probability of finding 1 segment (in one comparison trial) with score > x is:

\[ P(S > x) = 1 - c.d.f (S, x) = 1 - e^{-x} \approx e^{-x} \text{ (for large } x) \]

Since this is a rare event, we expect \( P(S > x) \) to be small.
The number of finding “separate” high-scoring segments is closely approximated by a Poisson distribution:

\[ P(\text{exactly } K \text{ segments having score } S > x) = e^{-y} \frac{y^K}{K!} \]

\( y = E - \text{value} \) of \( (S > x) \approx e^{-x} \text{ for large } x. \)
What if we want to consider several “separate” high-scoring segments?

The probability of finding $K$ or more distinct segments with score $S > x$ is:

$$P(\text{K or more segments having score } S > x)$$

$$= 1 - \sum_{i=0}^{K-1} P(\text{exactly } i \text{ segments having score } S > x)$$

$$= 1 - e^{-x} \sum_{i=0}^{K-1} \frac{y^i}{i!}$$

One caveat: You are only considering the worst performer in your L maximal segments!

Let us say you compare two sequences and obtain 3 segments with good scores: 40, 45, and 50. You can compute the significance of having at least 3 segments with scores better than 40 from previous discussions. However, the fact that exactly how much better is ignored.

Solution: You can compute the probability of having at least one segment better than 50 first. Then you compute the probability of having at least 2 segments better than 45. If the latter is smaller than the former, then it means the 45-scoring segment contributes the significance. Then you can go ahead and compute the probability of having at least 3 segments better than 40. Otherwise you stop there.
What if we want to do database searching when multiple comparisons need to be performed?

Database search is like comparing the query sequence (N letter long) with every sequence in your database (let us assume there are D of them). Since the probability of achieving good score for one comparison is a rare event, we can apply the almighty Poisson distribution again. Now \( \lambda \) is \( D \cdot P(S > x) \), where \( P(S > x) \) is the probability of a single event.

\[
P(\text{one or more segments with } S > x; D) = 1 - P(\text{zero segment with } S > x; D)
= 1 - e^{-\lambda} \frac{\lambda^L}{L!} \bigg|_{L=0}
= 1 - e^{-\lambda} \\
= 1 - e^{-D \cdot P(S > x)} \approx DP(S > x)
\]

What if sequences in the database have varying lengths?

\[
P(\text{one or more segments with } S > x; D) \approx DP(S > x) = \frac{N_D}{N} \ast P(S > x)
\]

where \( N_D \) is the total residue in the database, 
\( N \) is the length of the query sequence.
References

- Karlin, S., et al., 1983, PNAS, 80: 5660-5664
- Altschul, SF et al., 1994, Nature Genetics, 6:119-129