p-value, E-value and BLAST

Paola Castro
May 29, 2012

This handout explains the concept of a p-value\(^1\) and an E-value by applying it to a simple example and then to the Basic Local Alignment Search Tool (BLAST).

How many hours does a bioinformatics student sleep?

Let’s start by asking ... Does a bioinformatics student sleep less than 8 hours?

Our null hypothesis will be that a bioinformatics student sleeps as much as someone in the general population. The general population follows a distribution with a mean of 8 hours and a standard deviation of 1.5 hours, shown in figure 1.

![Null distribution for number of hours of sleep](image)

The alternative hypothesis will be that a bioinformatics student sleeps less than 8 hours. Consequently, a bioinformatics student does not fall in a distribution with a mean of 8 hours and a standard deviation of 1.5 hours.

Let’s use the p-value to test the null hypothesis first with Maria and Jose.

Maria, sleeps 7 hours. Jose, sleeps 5 hours.

Given the null hypothesis, the probability observing a student like Maria or more extreme, that is, a student that sleeps 7 hours or less...
is 0.25. 0.25 is the p-value for an observation of 7 hours. The p-value is the area under the null hypothesis curve representing the sum of the probabilities of observations as and more extreme as 7 hours shown in figure 2.

Given the null hypothesis, the probability observing a student like Jose or more extreme, that is, a student that sleeps 5 hours or less is 0.02. 0.02 is the p-value for an observation of 5 hours. The p-value is the area under the null hypothesis curve representing the sum of of the probabilities of observations as and more extreme as 5 hours in figure 3.

A biologist traditionally rejects the null hypothesis if the p-value is less than a significance level of 0.05. Maria’s p-value is greater than 0.05, so a biologist would not reject the null hypothesis in this case and conclude that the number of hours Maria sleeps follows the null distribution Jose’s p-value is less than 0.05, so a biologist would reject the null hypothesis in this case and conclude that the number of hours Jose sleeps does not follow the null distribution and accept the alternative hypothesis.

**The null hypothesis can also be challenged using E-values**

Like with p-values, with E-values there is also a cut-off. For this example, the E-value cut-off is:

(significance level) * (# of experiments) = 0.05 * 2

Maria’s E-value = (Maria’s p-value) * (# of experiments) = (0.25) * (2)

Jose’s E-value = (Jose’s p-value) * (# of experiments) = (0.02) * (2)

Maria’s E-value < E-value cut-off, so the number of hours Maria sleeps is significant. Jose’s E-value > E-value cut-off, so the number of hours Jose sleeps is not significant.

**BLAST example**

The question you ask when using BLAST is, does the alignment between my query sequence and arbitrary sequence from the database reflect biological significance?

Can I infer homology between my query sequence and arbitrary sequence from the database based on the alignment score?

**BLAST’s Null Hypothesis** is that the alignment score for my query sequence and arbitrary sequence from the database is random, that is, it follows the null distribution of random scores.

**BLAST’s alternative hypothesis** is that the alignment score for my query sequence and arbitrary sequence from the database is
not random, that is, it does not follow the null distribution of random scores.

To test whether an alignment score reflects biological relatedness BLAST uses the E-value\(^4\). Since the E-value reflects the number of times you expect to see alignment score \(X\) by chance, the lower the E-value, the fewer times you expect to see alignment score \(X\) by chance. In p-value terms, the p-value in this case would reflect the probability of observing score \(X\). If the alignment is not due to chance, then it may be due to a biological relationship between the two sequences.

For this class, we will conclude that an E-value of less than 1\(e^{-4}\) is considered significant, that is it reflects biological significance. In other words, we can infer homology between two sequences with an E-value lower than 1\(e^{-4}\). The higher the score, the lower the E-value.

**BLAST’s E-value calculation**

E-values are proportional to the query and database lengths as shown in equation 1. Where \(m\) is the length of the sequence, \(n\) is the length of the database and \(mn = \text{effective search space}\). The Bit score or alignment score, \(S’\), is shown in equation 2.

\[
E = mn2^{S’} \tag{1}
\]

\[
S’ = \frac{\lambda S - \ln K}{\ln 2} \tag{2}
\]

Since the E-value depends on database length, the E-value will be different with different size database searches. The larger the database used the larger the E-value, in other words, if you increase your search space you have more opportunities of observing a specific alignment score. The Bit score will always be the same for two given sequences.

**What does an E-value of 0.0 mean?** A very low calculated E-value. If the calculated E-value is \(< 1e^{-179}\) then it will be reported as 0.0.

**Acknowledgements**

I would like to thank David Bernick, Brandon Rice, Olga Botvinnik for our fruitful discussions.
p-value and E-value Worksheet

What is the question that we’re asking when we do a BLAST search or sequence alignment? Question?

What is the null hypothesis?
H0 -

What is the alternate hypothesis?
H1 -

Where would the score of a homologous sequence fall in this distribution?

What does the p-value mean in this case?
The p-value in this case means:

What does the E-value mean in this case?
The e-value in this case means:

What are examples of a change of size in the database?

How would the E-value change if the database increases?

What is the range of a p-value?
For this example, what is the range of a p-value?

What is the range of the E-value in this experiment?
For this example, what is the range of an e-value?