Sequence Comparison: Significance of similarity scores

Genome 373
Genomic Informatics
Elhanan Borenstein
Quick review: Local alignment

Find the optimal **local** alignment of AAG and GAAGGC.
Use a gap penalty of $d = -5$.
Summary

Global alignment algorithm: **Needleman-Wunsch.**

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>A</th>
<th>A</th>
<th>T</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0</td>
<td>-4</td>
<td>-8</td>
<td>-12</td>
<td>-16</td>
</tr>
<tr>
<td>A</td>
<td>-4</td>
<td>-5</td>
<td>-9</td>
<td>-13</td>
<td>-12</td>
</tr>
<tr>
<td>T</td>
<td>-12</td>
<td>-8</td>
<td>1</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>A</td>
<td>-16</td>
<td>-12</td>
<td>2</td>
<td>11</td>
<td>7</td>
</tr>
</tbody>
</table>

Local alignment algorithm: **Smith-Waterman.**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>
Significance of scores

Alignment algorithm

HPDKKAHSIHAWILSKSKVLEGNTKEVVVDNVLK
LENENQGKCTIAEYKYGKKASVYNSFVSNGVKE

45
Low score = unrelated
High score = related

But ... how high is high enough?

Subjective
Problem specific
Parameter specific
We want to know how surprising a given score is, ... assuming that the two sequences are not related.

This assumption is called the null hypothesis.

The purpose of most statistical tests is to determine whether the observed result provides a reason to reject the null hypothesis.

Put differently, we want to determine how likely is it to obtain a specific score (or higher) under the null hypothesis.
P-values as a representation of surprise

- The probability of observing a score $\geq X$ is the area under the curve to the right of $X$.
- This probability is called a $p$-value.
- $p$-value = $\Pr(\text{data} | \text{null})$
Sequence similarity score distribution

- Search a database of unrelated sequences using a given query sequence.
- What will be the form of the resulting distribution of pairwise alignment scores?
Empirical null score distribution

- This shows the distribution of scores from a real database search using BLAST.
Empirical null score distribution

• This shows the distribution of scores from a real database search using BLAST.

• This distribution contains scores from a few related and lots of unrelated pairs.

(see graph) High scores from related sequences (note - there are lots of lower scoring alignments not reported)
The distribution of scores obtained from aligning a given sequence to a database of random sequences (note - there are lots of lower scoring alignments not reported)
Empirical null score distribution

• The distribution of scores obtained from aligning a given sequence to a database of random sequences

• But .... How will be generate a database of random sequences??

(note - there are lots of lower scoring alignments not reported)
Computing an empirical p-value

- P-value = The probability of observing a score $\geq X$ is the area under the curve to the right of X.

e.g. out of 1,685 scores, 28 received a score of 20 or better. Thus, the p-value associated with a score of 20 is $\sim \frac{28}{1685} = 0.0166$. 
Problems with empirical distributions

• We are interested in very small probabilities.

• These are computed from the *tail* of the null distribution.

• Estimating a distribution with an accurate tail is feasible but computationally very expensive because we have to make a very large number of alignments.
A solution

• Characterize the form of the score distribution mathematically.

• Fit the parameters of the distribution empirically (or compute them analytically).

• Use the resulting distribution to compute accurate p-values.

(first solved by Karlin and Altschul)
This distribution is roughly normal near the peak, but characterized by a larger tail on the right.

- **For an Unscaled EVD:**

\[
P(S \geq x) = 1 - e^{(-e^{-x})}
\]
What p-value is significant?
What p-value is significant?

• The most common thresholds are 0.01 and 0.05.
• A threshold of 0.05 means that even if the null hypothesis is correct you will still get such score (or higher) in 5% of cases.
• Why 0.05? It depends upon the cost associated with making a mistake.

Examples of costs:
– Doing extensive wet lab validation (expensive)
– Making clinical treatment decisions (very expensive)
– Misleading the scientific community (very expensive)
– Doing further simple computational tests (cheap)
– Telling your grandmother (very cheap)
Multiple testing
Multiple testing

• Say that you perform a statistical test with a 0.05 threshold, but you repeat the test on twenty different observations (e.g. 20 different alignments)

• Assume that all of the observations are explainable by the null hypothesis.

• What is the chance that at least one of the observations will receive a p-value < 0.05?
Multiple testing

• Say that you perform a statistical test with a 0.05 threshold, but you repeat the test on twenty different observations (e.g. 20 different alignments)

• Assume that all of the observations are explainable by the null hypothesis.

• What is the chance that at least one of the observations will receive a p-value < 0.05?

\[1 - 0.95^{20} = 0.6415\]
Bonferroni correction

• Assume that individual tests are *independent*.

• Divide the desired p-value threshold by the number of tests performed.

• In the example about, a Bonferroni correction would suggest using a p-value threshold of $0.05 / 20 = 0.0025$. 
Database searching

• Say that you search the non-redundant protein database at NCBI, containing roughly one million sequences (i.e. you are doing $10^6$ pairwise tests).

• and ... you want to use a p-value of 0.01.

• Recall that you would observe such a p-value by chance approximately every 100 times in a random database.

• That is, without correcting for multiple testing you will get $\sim$10,000 false positives!!

• A Bonferroni correction would suggest using a p-value threshold of $0.01 / 10^6 = 10^{-8}$. 
E-values

- An E-value is the expected number of times that the given score would appear in a random database of the given size.
- One simple way to compute the E-value is to multiply the p-value times the size of the database.
- Thus, for a p-value of 0.01 and a database of 1,000,000 sequences, the corresponding E-value is $0.01 \times 1,000,000 = 10,000$.

(BLAST actually calculates E-values in a more complex way, but they mean the same thing)
<table>
<thead>
<tr>
<th>Accession</th>
<th>Description</th>
<th>Score (bits)</th>
<th>E Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>gi</td>
<td>112670</td>
<td>sp</td>
<td>P15711</td>
</tr>
<tr>
<td>gi</td>
<td>14268530</td>
<td>gb</td>
<td>AAK56556.1</td>
</tr>
<tr>
<td>gi</td>
<td>14268528</td>
<td>gb</td>
<td>AAK56555.1</td>
</tr>
<tr>
<td>gi</td>
<td>14268526</td>
<td>gb</td>
<td>AAK56554.1</td>
</tr>
<tr>
<td>gi</td>
<td>31210185</td>
<td>ref</td>
<td>XP_314059.1</td>
</tr>
<tr>
<td>gi</td>
<td>22971724</td>
<td>ref</td>
<td>ZP_00018655.1</td>
</tr>
<tr>
<td>gi</td>
<td>32403566</td>
<td>ref</td>
<td>XP_322396.1</td>
</tr>
<tr>
<td>gi</td>
<td>24639766</td>
<td>ref</td>
<td>NP_572189.1</td>
</tr>
<tr>
<td>gi</td>
<td>30348569</td>
<td>emb</td>
<td>CAC84361.1</td>
</tr>
<tr>
<td>gi</td>
<td>9629342</td>
<td>ref</td>
<td>NP_044542.1</td>
</tr>
<tr>
<td>gi</td>
<td>24639768</td>
<td>ref</td>
<td>NP_726958.1</td>
</tr>
<tr>
<td>gi</td>
<td>4757118</td>
<td>emb</td>
<td>CAB42096.1</td>
</tr>
<tr>
<td>gi</td>
<td>17534529</td>
<td>ref</td>
<td>NP_495288.1</td>
</tr>
<tr>
<td>gi</td>
<td>15241089</td>
<td>ref</td>
<td>NP_195809.1</td>
</tr>
<tr>
<td>gi</td>
<td>43489677</td>
<td>gb</td>
<td>EAD99646.1</td>
</tr>
<tr>
<td>gi</td>
<td>44419062</td>
<td>gb</td>
<td>EAJ13596.1</td>
</tr>
<tr>
<td>gi</td>
<td>43969222</td>
<td>gb</td>
<td>EAG41329.1</td>
</tr>
<tr>
<td>gi</td>
<td>15792145</td>
<td>ref</td>
<td>NP_281968.1</td>
</tr>
<tr>
<td>gi</td>
<td>43926327</td>
<td>gb</td>
<td>EAG18073.1</td>
</tr>
<tr>
<td>gi</td>
<td>39595869</td>
<td>emb</td>
<td>CAE67372.1</td>
</tr>
<tr>
<td>gi</td>
<td>30020082</td>
<td>ref</td>
<td>NP_831713.1</td>
</tr>
<tr>
<td>gi</td>
<td>43723946</td>
<td>gb</td>
<td>EAF16931.1</td>
</tr>
<tr>
<td>gi</td>
<td>11545212</td>
<td>gb</td>
<td>AAG37800.1</td>
</tr>
<tr>
<td>gi</td>
<td>40788024</td>
<td>emb</td>
<td>CAE47751.1</td>
</tr>
<tr>
<td>gi</td>
<td>42656951</td>
<td>ref</td>
<td>XP_052597.6</td>
</tr>
<tr>
<td>gi</td>
<td>32698642</td>
<td>ref</td>
<td>NP_872557.1</td>
</tr>
<tr>
<td>gi</td>
<td>12840300</td>
<td>dbj</td>
<td>BAB24814.1</td>
</tr>
</tbody>
</table>
Take home message

- A **distribution** plots the frequencies of types of observation.
- The area under the distribution curve is 1.
- Most statistical tests compare observed data to the expected result according to a **null hypothesis**.
- Sequence similarity scores follow an **extreme value distribution**, which is characterized by a long tail.
- The **p-value** associated with a score is the area under the curve to the right of that score.
- Selecting a **significance threshold** requires evaluating the cost of making a mistake.
- **Bonferroni correction**: Divide the desired p-value threshold by the number of statistical tests performed.
- The **E-value** is the expected number of times that a given score would appear in a random database of the given size.
Computing a p-value

- The probability of observing a score $\geq 4$ is the area under the curve to the right of 4.

For an *Unscaled EVD*:

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

$$P(S \geq 4) = 1 - e^{-e^{-4}} = 0.018149$$