Sequence comparison: Significance of alignment scores

http://faculty.washington.edu/jht/GS559_2014/

Genome 559: Introduction to Statistical and Computational Genomics
Prof. James H. Thomas
Unscaled EVD equation

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

S is data score, x is test score

(FYI this is 1 minus the cumulative density function or CDF)
Scaling the EVD

- An EVD derived from, e.g., the Smith-Waterman algorithm with a given substitution matrix and gap penalties has a characteristic mode $\mu$ and scale (width) parameter $\lambda$.

\[
P(S \geq x) = 1 - e^{-e^{-x}} \quad \text{scaled:} \quad P(S \geq x) = 1 - e^{-e^{-\lambda(x-\mu)}}
\]

$\lambda$ and $\mu$ depend on the substitution matrix and the gap penalties.
Similar to scaling the standard normal

\[ PDF_{\text{snormal}} = Ce^{-x^2/2} \]

where \( C = 1/\sqrt{2\pi} \)

\[ PDF_{\text{gnormal}} = Ce^{-(x-\mu)^2/2\nu} \]

where \( C = 1/\sqrt{2\pi\nu} \)

\( \nu \) is variance, \( \mu \) is mean

PDF = probability density function

(\( \mu \) moves peak and \( \nu \) adjusts width)
An example

You run BLAST and get a maximum match score of 45. You then run BLAST on a shuffled version of the database, and fit an EVD to the resulting empirical distribution. The parameters of the EVD are $\mu = 25$ and $\lambda = 0.693$. What is the p-value associated with score 45?

$$P(S \geq 45) = 1 - e^{(-e^{-0.693(45-25)})}$$

$$= 1 - e^{(-e^{-13.86})}$$

$$= 1 - e^{-9.565 \times 10^{-7}}$$

$$= 1 - 0.9999999043$$

$$= 9.565 \times 10^{-7}$$

BLAST has precomputed values of $\mu$ and $\lambda$ for common matrices and gap penalties.
What p-value is significant?

• The most common thresholds are 0.01 and 0.05.
• A threshold of 0.05 means you are 95% sure that the result is significant.
• Is 95% enough? 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

- 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 blast runs).

- 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 of 0.05 or less?

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

• Assume that individual tests are *independent*.

• Multiply the p-values by the number of tests performed.
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). What p-value threshold should you use?

- Say that you want to use a conservative p-value of 0.001.

- Recall that you would observe such a p-value by chance approximately every 1000 times in a random database.
E-values

• A p-value is the probability of making a mistake.
• 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 by the number of sequences in the database.
• Thus, for a p-value of 0.001 and a database of 1,000,000 sequences, the corresponding E-value is 0.001 × 1,000,000 = 1,000.

(BLAST actually calculates E-values in a different way, but they mean about the same thing)
>104K_THEPA 104 KD MICRONEME-RHOPTRY ANTIGEN
MKFLILLFNILCLFPVLADNHGVGPQGASGVDPTTFDNSNQTGPAFLTAEMAGVKYLQ
HRLVEGNVVIWENASTPLYTGATIVTNNDGPYMAVEVLDGDPLQLFFIKSGDAWVTLEHEY
AVHIESVFSLNMAFQLLENKKYEVEATHAKNGANMVTFIPRNGHIKCMVYHKNVRITYKATGD
RGLRLLLINVFSIDDNGMMSNRYFQHVDDKYVPISQKNYESGTIVKLKDHYKHAYHPVDLDIK
### Sequences producing significant alignments:

| gi[112670] | sp|P15711|104K_THEPA | 104 KD MICRONEME-RHOPTRY ANT... | 1352 | 0.0 |
| gi[14268530] | gb|AAK56556.1 | 104 kDa microneme-rhoptry antige... | 243 | 1e-62 |
| gi[14268528] | gb|AAK56555.1 | 104 kDa microneme-rhoptry antige... | 242 | 4e-62 |
| gi[14268526] | gb|AAK56554.1 | 104 kDa microneme-rhoptry antige... | 238 | 7e-62 |
| gi[3210185] | ref|XP_314059.1 | ENSANGP00000015608 [Anopheles ... | 37 | 2.1 |
| gi[22971724] | ref|ZP_00018655.1 | hypothetical protein [Chloro... | 35 | 9.7 |
| gi[32403566] | ref|XP_322396.1 | hypothetical protein [Neurosp... | 35 | 12 |
| gi[24639766] | ref|NP_572189.1 | CG2861-PA [Drosophila melanoga... | 34 | 17 |
| gi[30348569] | emb|CAC84361.1 | hypothetical protein [Saimiriin... | 34 | 19 |
| gi[6492132] | gb|AAFL14193.1 | spherical body protein 3 [Babesia... | 34 | 20 |
| gi[9629342] | ref|NP_044542.1 | virion protein [Human herpesvir... | 34 | 21 |
| gi[24639768] | ref|NP_726958.1 | CG2861-PB [Drosophila melanoga... | 34 | 21 |
| gi[4757118] | emb|CAB42096.1 | TashAT2 protein [Theileria annul... | 34 | 22 |
| gi[17534529] | ref|NP_495288.1 | putative protein (2G676) [Caen... | 33 | 22 |
| gi[15241089] | ref|NP_195809.1 | leucine-rich repeat transmembr... | 33 | 23 |
| gi[43489677] | gb|EAD99646.1 | unknown [environmental sequence] | 33 | 23 |
| gi[44419062] | gb|EAL13596.1 | unknown [environmental sequence] | 33 | 25 |
| gi[43969222] | gb|EAG14329.1 | unknown [environmental sequence] | 33 | 29 |
| gi[15792145] | ref|NP_281968.1 | putative oxidoreductase [Campy... | 33 | 34 |
| gi[43926327] | gb|EAG18073.1 | unknown [environmental sequence] | 33 | 37 |
| gi[39595869] | emb|CAE67372.1 | Hypothetical protein CBG12848 [... | 33 | 38 |
| gi[30020082] | ref|NP_831713.1 | Glycosyltransferase [Bacillus ... | 33 | 40 |
| gi[43723946] | gb|EAF16931.1 | unknown [environmental sequence] | 33 | 41 |
| gi[11545212] | gb|AAFL37800.1 | hypothetical telomeric SfiI frag... | 33 | 44 |
| gi[40788024] | emb|CAE47751.1 | ubiquitin specific proteinine 5... | 32 | 51 |
| gi[42656951] | ref|XP_052597.6 | ubiquitin specific protease 53... | 32 | 51 |
| gi[32698642] | ref|NP_872557.1 | DNA-ligase [Adoxophyes orana g... | 32 | 52 |
| gi[12840300] | dbj|BAB24814.1 | unnamed protein product [Mus mu... | 32 | 54 |
| gi[28899333] | ref|NP_798938.1 | 4-diphosphocytidyl-2C-methyl-D... | 32 | 55 |
| gi[7243081] | dbj|BAA92588.1 | KIAA1350 protein [Homo sapiens] | 32 | 62 |
Summary

• 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 alignment scores for unrelated sequences 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**: Multiply the p-value by the number of statistical tests performed.
• The **E-value** is the expected number of times that a given score would appear in a randomized database.