Statistics I: Homology, Statistical Significance, and Multiple Tests

• homology and statistical significance
  – how do we measure “significance?”
• what is the probability of an alignment score?
  – given two sequences
  – after a database search
• validating local alignment statistics
  – when do things go wrong
  – confirmation with random shuffles

Statistics and excess similarity – Correcting for multiple tests I

• We recognize excess similarity when we know what happens by chance – models for “random” similarity
  – what is the distribution of random (local, global) similarity scores?
  – how do we correct for multiple tests
  – what do we miss
• How can we test the statistics
  – E()-value of highest scoring unrelated sequence
  – shuffled sequence comparison
• E()-values are good for significance, not for evolutionary distance
Statistics and excess similarity – Correcting for multiple tests I

- Local similarity scores are *NOT* normally distributed:

- We derive the probability by estimating:
  - the probability of matches, correcting for the number of starting locations:
    - the extreme value distribution:
  - correcting for the number of alignment scores

---

Winter et al (1968) Science 162:1433

**Homology as Applied to Proteins**

"Do cats eat bats? Do cats eat bats?" and sometimes "Do bats eat cats?" for you see, as she couldn't answer either question, it didn't much matter which way she put it (7).

Our article entitled “Evolution of structure and function of proteases” dealing with the biochemical approach to the subject of evolution as exemplified by studies of proteolytic enzymes (2) put forth a definition of the term “homology” as it applies to similarities in protein structures. This word has been much bandied about and generally used by many to represent a host of ill-defined concepts. We proposed that the word be taken to connote the occurrence of a degree of structural similarity among proteins greater than might be anticipated by chance alone.

This definition has been criticized by Margoliash (3). His position is that since evolution is traditionally the province of the classical biologist, the classical biologist’s definition of “homology” should prevail. This would add to our definition the additional qualification that the protein structures in question must have evolved from a common ancestral gene. The problem with this restrictive definition is that the word, although precisely defined, can seldom be used in a precise sense. For example, did ancestral genes common to divergent populations give rise to “homologous” proteins, or does the occurrence of “homologous” proteins mean that they arose from genes having a common ancestor? It really doesn’t matter how we put it because like Lewis Carroll’s *Alice*, we do not know the answer to either question. The perishable nature of the gene prevents us from obtaining concrete and objective evidence on the nature or existence of ancestral genes.

significant similarity

+ (precise) common ancestry
Measuring Similarity –
sorting, searching, and statistics

sequences. Cytochrome c and cytochrome c551” J. Mol. Biol. 61:409

An improved method for testing similarities or repeats in protein sequences is
described. It includes three features:

- a measure of similarity for amino acids, based on observed substitutions in homologous proteins;
- a search procedure which compares all pairs of segments of two proteins;
- new statistical tests which estimate the probabilities that observed correlations could have occurred by chance.

![Graph of related and random distributions](image)

- Gaussian random related
- What is the correct distribution?

A. D. McLachlan, 1971

**Similarity – homology (divergence)
or convergence?**

The fundamental assumption of the present approach is that if the amino-acid
sequences of two proteins are so alike that their similarity is very unlikely to have
happened by chance, then they will have the same three-dimensional structure and be
ancestrally related. This is based on the finding from X-ray studies that homologous
proteins have very similar three-dimensional structures, so that observed amino-acid
substitutions usually conserve the folding of the peptide chain. Thus, related proteins
remain structurally similar even if the mutation distances are large. ...

One could object to the fundamental assumption, on the grounds that convergent
evolution is likely to lead to precisely these kinds of accidental similarities between
unrelated proteins. There is not sufficient evidence yet to exclude this possibility.
However, no example is yet known where convergent evolution has led to similarities of
structure or sequence which approach those found repeatedly in homologous proteins.
Rather, the existence of unrelated lysozymes or nucleases, the irregular and apparently
random structural features of many proteins, and the large variety of amino-acid
substitutions in homologous families of proteins, all suggest that the number of
conceivable ways of evolving an enzyme to perform a given function is astronomically
large. Thus, convergent evolution is unlikely to repeat more than a few of the many fine
details of structure and sequence in any pair of proteins.

McLachlan, 1971
Inferring Homology from Statistical Significance

- Real **UNRELATED** sequences have similarity scores that are indistinguishable from **RANDOM** sequences
- If a similarity is NOT **RANDOM**, then it must be NOT **UNRELATED**
- Therefore, NOT **RANDOM** (statistically significant) similarity must reflect **RELATED** sequences

How often do things happen by chance? statistics of coin tosses - expectation

- \(p(H) = p(T) = 0.5\)
- \(p(HHHTH)=p(HTTTH)=p(HHHHH) = (1/2)^5\)
- how many times do we expect a run of 10 heads (by chance) in:
  - 10 flips \(1 (1/2)^{10} \approx 0.001\)
  - 100 flips \(91(1/2)^{10} \approx 0.1\)
  - 1000 flips \(991 (1/2)^{10} \approx 1\)
  - 1,000,000 flips \(999,991 (1/2)^{10} \approx 1000\)

- Probability \((0 \leq p \leq 1)\) vs Expectation \((0 \leq E() \leq \text{number of trials})\)
  \[ E(x) = p(x) \times N \]
Given an expectation, what is its probability?

The Poisson Distribution: probabilities of counts of events (radioactive decay, high similarity scores, RNAseq reads?)

\[ p(\mu, i) = \frac{\mu^i e^{-\mu}}{i!} \]

\( \mu \) = mean expectation of event

\( i \) = number of events

Poisson distribution for ranges of events (one or more)

\[ p(x \geq 1) = \sum_{i=1}^{\infty} \frac{\mu^i e^{-\mu}}{i!} = \mu^1 e^{-\mu} / 1! + \mu^2 e^{-\mu} / 2! + ... \]

\[ p(x \geq 1) = 1 - p(0) = 1 - \mu^0 e^{-\mu} / 0! = 1 - e^{-\mu} \]

\[ \mu \quad p(x>0) \]

\| 0.001 | 0.001 |
\| 0.01 | 0.010 |
\| 0.1  | 0.095 |
\| 1.0  | 0.632 |
\| 2.0  | 0.865 |

1-exp^\mu \sim \mu for \( \mu < 0.1 \)
Statistics of “Head” runs

\[ E(I) = n \cdot p^l \]

Results from tossing a coin 14 times; black circles indicate heads. The probability of 5 heads in a row is \( p(5) = (1/2)^5 = 1/32 \), but since there were 10 places that one could have obtained 5 heads in a row, the expected number of times that 5 heads occurs by chance is \( E(5H) = 10 \times 1/32 = 0.31 \).

Alignment scores as coin tosses

- \( E(# \; \text{of H of length m}) \sim np^m \)
- if the longest run is unique, \( 1 = np^{R_n} \)
  \[ 1/n = p^{R_n} \]
  \[ -\log_e(n) = R_n \cdot \log_e(p) \]
  \[ -\log_e(n)/\log_e(p) = R_n \]
  \[ R_n = \log_{(1/p)}(n) \]

Converting logarithms:
\[ 10^x = B^y \]
\[ x \cdot \log_{10} 10 = y \cdot \log_{10} B \]
\[ x = y \cdot \log_{10} B \]
\[ x/ \log_{10} B = y \]

Thus, the expected length of the longest run \( R_n \) increases as \( \log(n) \) of the run length
Statistics of “Head” alignments

\[ E(l) = m \times n \times p \]

*The expected length of the longest run \( R_p \) increases as \( \log(mn) \).*

Comparison of two protein sequences, with identities indicated as black circles. Assuming the residues were drawn from a population of 20, each with the same probability, the probability of an identical match is \( p = 0.05 \). In this example, there are \( m = 10 \times n = 8 \) boxes, so \( E(l) = m \times n \times p = 80 \times 0.05 = 4 \) matches are expected by chance. The probability of two successive matches is \( p^2 = (1/20)^2 \) so a run of two matches is expected about \( n \times m \times p^2 = 8 \times 10 \times (1/20)^2 = 0.2 \) times by chance.

From “Head” runs to scores

The longest “Head” run is equivalent to the “longest hydrophobic stretch” using a scoring matrix that assigns positive values \( S_i \) for some residues \( i \) and \(-\infty\) for all other residues. Then:

\[ p(S) = \sum p(s_i) \text{ for residues } i \text{ with } s_i > 0 \]

The same analogy can be made for alignment scores between \( i,j \) where \( s_{ij} \) the score for aligning residues \( i,j \) is either + with \( p(s_{ij}) \) or \(-\infty\). Now the score for the longest positive alignment score is:

\[ E(S \geq x) \propto mnp^x \]
\[ E(S \geq x) \propto mne^{x \ln p} \]
\[ E(S \geq x) \propto mne^{-\lambda x} \text{ where } \lambda = -\ln p \]
Karlin-Altschul statistics for alignments without gaps

Given:
\[ E(s_{i,j}) = \sum_{i,j} p_ip_js_{i,j} < 0 \text{ (local alignments)} \]

Then:
\[ E(S \geq x) = Kmne^{-\lambda S} \]
\[ K < 1 \text{ (space correction)} \]
\[ \lambda \text{ solution of: } \sum_{i,j} p_ip_j e^{\lambda s_{i,j}} \]

\[ E(S \geq x) \] is the **Expectation** (average # of times) of seeing score \( S \) in an alignment. so, we apply the Poisson conversion:
\[ p(x) = 1 - \exp(-x) \Rightarrow \]
\[ p(S > x) = 1 - \exp(-Kmne^{-\lambda S}) \]

The Similarity Statistics Mantra…

- Find the **Probability** of a rare event (e.g. a high score) in a cluster of residues
  \[ p^n \propto e^{-\lambda S} \]
- Find the **Expectation** of this event by correcting for all the places it could have happened
  \[ Kmne^{-\lambda S} \]
- Convert that into a **Probability** using the Poisson formula:
  \[ 1 - \exp(-Kmne^{-\lambda S}) \]
- Convert that **Probability** into an **Expectation** for the number of sequences in the database
  \[ E(S > x) = P \cdot D = (1 - \exp(-Kmne^{-\lambda S})) \cdot D \]
Extreme value distribution

\[ S' = \lambda S_{raw} - \ln K \    m \    n \]
\[ S_{bit} = (\lambda S_{raw} - \ln K)/\ln(2) \]
\[ P(S' > x) = 1 - \exp(-e^{-x}) \]
\[ P(S_{bit} > x) = 1 - \exp(-mn2^{-x}) \]
\[ E(S' > x | D) = P D \]

\[ P(B \text{ bits}) = mn 2^{-B} \]
\[ P(40 \text{ bits}) = 1.5 \times 10^{-7} \]
\[ \lambda s \quad E(40 | D=4000) = 6 \times 10^{-4} \]
\[ \text{bit} \quad E(40 | D=50E6) = 7.5 \]

How many bits do I need?

\[ P(S_b > x_b) = mn2^{-x_b} = \frac{mn}{2^{x_b}}, S_b \text{ is a score in "bits"} \]

<table>
<thead>
<tr>
<th>Query size ( m )</th>
<th>Lib. seq. size: ( n )</th>
<th>DB Entries ( D )</th>
<th>( mnD/0.01 )</th>
<th>Bit threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>200</td>
<td>100,000</td>
<td>4 \times 10^9/0.001</td>
<td>42</td>
</tr>
<tr>
<td>450</td>
<td>450</td>
<td>100,000</td>
<td>2 \times 10^9/0.001</td>
<td>44</td>
</tr>
<tr>
<td>450</td>
<td>450</td>
<td>10,000,000</td>
<td>2 \times 10^{13}/0.001</td>
<td>51</td>
</tr>
</tbody>
</table>

BLAST statistics report effective search space (mnD)

Effective search space used: 12,967,374,320
Database: Non-redundant UniProtKB/SwissProt sequences
Number of letters in database: 174,324,904
Number of sequences in database: 464,207
What is a “bit” score?

- Scoring matrices (PAM250, BLOSUM62, VTML40) contain “log-odds” scores:
  \( s_{ij} \) (bits) = \( \log_2 \left( \frac{q_{ij}}{p_ip_j} \right) \) (\( q_{ij} \) freq. in homologs/ \( p_ip_j \) freq. by chance)
  \( s_{ij} \) (bits) = 2 -> a residue is 2\(^4\)=4-times more likely to occur by homology compared with chance (at one residue)
  \( s_{ij} \) (bits) = -1 -> a residue is 2\(^-1\)=1/2 as likely to occur by homology compared with chance (at one residue)

- An alignment score is the maximum sum of \( s_{ij} \) bit scores across the aligned residues. A 40-bit score is 2\(^{40}\) more likely to occur by homology than by chance.

- How often should a score occur by chance? In a 400 * 400 alignment, there are \(~160,000\) places where the alignment could start by chance, so we expect a score of 40 bits would occur:
  \( P(S_{\text{bit}} > x) = 1 - \exp(-mn^2x) \sim mn^2x^{400}x^{2^{40}} = 1.6 \times 10^7 \) times
  Thus, the probability of a 40 bit score in ONE alignment is \(~10^{-7}\)

- But we did not ONE alignment, we did 4,000, 40,000, 400,000, or 16 million alignments when we searched the database:
  \( E(S_{\text{bit}} | D) = p(40 \text{ bits}) \times \text{database size} \)
  \( E(40 | 4,000) = 10^{-8} \times 4,000 = 4 \times 10^{-4} \) (significant)
  \( E(40 | 40,000) = 10^{-7} \times 4 \times 10^4 = 4 \times 10^{-3} \) (significant)
  \( E(40 | 400,000) = 10^{-7} \times 4 \times 10^5 = 4 \times 10^{-2} \) (not significant)
  \( E(40 | 16 \text{ million}) = 10^{-7} \times 1.6 \times 10^7 = 1.6 \) (not significant)

How many “bits” do I need?

- To get \( E() \sim 10^{-3} \):
  genome (10,000) \( p \sim 10^{-3}/10^4 = 10^{-7}/160,000 = 40 \text{ bits} \)
  SwissProt (500,000) \( p \sim 10^{-3}/10^6 = 10^{-9}/160,000 = 47 \text{ bits} \)
  Uniprot/NR (10\(^7\)) \( p \sim 10^{-3}/10^7 = 10^{-10}/160,000 = 50 \text{ bits} \)

<table>
<thead>
<tr>
<th>Query Database Size</th>
<th>Query Database Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;40</td>
<td>40-69</td>
</tr>
<tr>
<td>40-69</td>
<td>50-82</td>
</tr>
<tr>
<td>50-82</td>
<td>80-200</td>
</tr>
<tr>
<td>80-200</td>
<td>&gt;=200</td>
</tr>
</tbody>
</table>

**Color key for alignment scores**

- very significant \( 10^{-60} \)
- significant \( 10^{-6} \)
- significant \( 10^{-3} \)
- not significant
Should you trust the E()-value??

- The inference of homology from statistically significant similarity depends on the observation that unrelated sequences look like random sequences
  - Is this ALWAYS true?
  - How can we recognize when it is not true?

- If unrelated==random, then the E()-value of the highest scoring unrelated sequence should be $E() \sim 1.0$

- Statistical estimates can also be confirmed by searches against shuffled sequences

Smith-Waterman (ssearch36)
- highest scoring unrelated from domains
Removing low complexity regions with “pseg”

Protein Sequence Comparison

Statistics are Accurate
Statistical estimates from random shuffles

- BLAST estimates statistical significance from simulations of “normal” (average composition) proteins
- FASTA estimates statistical significance from the distribution of similarity scores obtained during the database search (selects 60,000 unrelated sequence scores from the database of real proteins)
- What if the sequences are different from most proteins, but similar to each other, e.g. membrane proteins?
- FASTA/SSSEARCH can also estimate statistical significance by producing hundreds of shuffled (random) sequences with the same length and composition, and then estimate $\lambda$ and $K$ from comparisons against those proteins

Two ways to shuffle: uniform and window

| PRSS34 | 1000 shuffles; uniform shuffle
| Unshuffled s-w score: 178; bits$(n=178|n=271) = 34.8$ p$(178) < 2.005e^{-06}$  | For 10000 sequences, a score $\geq 178$ is expected $0.02005$ times |

| LWECS_E. coli-transporting ATP synthase (EC 3.6.1.34) protein 6 - Escherichia coli |
| BASEMATP FQGHNVLQ QLARLFSNLQ PQHPPHMTT HNDDSFSS VLGGFLVLFL |
| ESKPPRH YPGQFALE VQPGYVNQV EKTHVREK EIPLALLTFV WFFLMNSDL |
| IFQGPP QEYKFGALT VQPGADQTVT EHLGAOVTIL FLFSTIKRKG IGFPTNAIL |
| QPFLIPFIP VWLFEGVL LKPSLVFLRL LFQNYAAGEL IFILLAGILLP WWQWILNVP |
| WAHSLKST LQAFT ETIVYLSASEE H |

| LWECS_6 shuffled window: 10 |
| Unshuffled s-w score: 178; bits$(n=178|n=271) = 34.5$ p$(178) < 2.601e^{-06}$  | For 10000 sequences, a score $\geq 178$ is expected $0.02602$ times |

| PRSS34 | 1000 shuffles; window shuffle, window size: 20 |
| Unshuffled s-w score: 178; bits$(n=178|n=271) = 34.5$ p$(178) < 2.601e^{-06}$  | For 10000 sequences, a score $\geq 178$ is expected $0.02602$ times |
### Statistical estimates from searches

**(gtr1_human, 12 transmembrane domains)**

<table>
<thead>
<tr>
<th></th>
<th>qutd_qmensi</th>
<th>citi_ecoli</th>
<th>kgtp_ecoli</th>
<th>yb8g_yeast</th>
<th>λ</th>
</tr>
</thead>
<tbody>
<tr>
<td>blast2.0</td>
<td>2.0e-25</td>
<td>1.0e-05</td>
<td>0.077</td>
<td>2.0</td>
<td>0.2700</td>
</tr>
<tr>
<td>ssearch BL50</td>
<td>1.6e-28</td>
<td>6.1e-05</td>
<td>0.014</td>
<td>0.72</td>
<td>0.1544</td>
</tr>
<tr>
<td>raw-score</td>
<td>536</td>
<td>199</td>
<td>148</td>
<td>123</td>
<td></td>
</tr>
<tr>
<td>bit-score</td>
<td>127</td>
<td>48</td>
<td>40</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>%identity</td>
<td>27.1</td>
<td>22.1</td>
<td>24.1</td>
<td>22.1</td>
<td></td>
</tr>
<tr>
<td>ssearch BL62</td>
<td>4.7e-32</td>
<td>1.2e-4</td>
<td>1.3</td>
<td>3.1</td>
<td>0.2584</td>
</tr>
<tr>
<td>raw-score</td>
<td>356</td>
<td>120</td>
<td>75</td>
<td>72</td>
<td></td>
</tr>
<tr>
<td>bit-score</td>
<td>138</td>
<td>47</td>
<td>34</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>%identity</td>
<td>26.9</td>
<td>21.0</td>
<td>27.9</td>
<td>24.1</td>
<td></td>
</tr>
</tbody>
</table>

... and shuffles

<table>
<thead>
<tr>
<th></th>
<th>shuffled BL50</th>
<th>BL50</th>
<th>0.0539</th>
<th>92</th>
</tr>
</thead>
<tbody>
<tr>
<td>λ</td>
<td>0.1375</td>
<td>0.1317</td>
<td>0.1263</td>
<td></td>
</tr>
<tr>
<td>window 20</td>
<td>3.9e-09</td>
<td>0.097</td>
<td>0.21</td>
<td>361</td>
</tr>
<tr>
<td>λ</td>
<td>0.0653</td>
<td>0.1064</td>
<td>0.1206</td>
<td>0.1110</td>
</tr>
<tr>
<td>BL62</td>
<td>6.6e-30</td>
<td>8.5e-04</td>
<td>0.36</td>
<td>49.</td>
</tr>
<tr>
<td>λ</td>
<td>0.2108</td>
<td>0.2011</td>
<td>0.2256</td>
<td>0.2172</td>
</tr>
</tbody>
</table>

### E()-values when??

- **E()-values (BLAST expect)** provide accurate statistical estimates of similarity by chance
  - non-random -> not unrelated (homologous)
  - E()-values are accurate (0.001 happens 1/1000 by chance)
  - E()-values factor in (and depend on) sequence lengths and database size
- **E()-values are NOT** a good proxy for evolutionary distance
  - doubling the length/score SQUARES the E()-value
  - percent identity (corrected) reflects distance (given homology)
Statistics and excess similarity – Correcting for multiple tests I

• Similarity by chance: extreme value distribution
  – correct for multiple tests using Bonferroni
• Testing the statistics:
  – $E()$-value of highest scoring unrelated sequence
  – shuffled sequence comparison
• $E()$-values good for significance, not for evolutionary distance