Goals of today’s lecture:
• what is the probability of an alignment score?
  – given two sequences
  – after a database search
  – after many database searches
• Hidden Markov Models
  – transition state models
  – profile HMMs

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:
  
<table>
<thead>
<tr>
<th>Expectation</th>
<th>poisson probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 flips</td>
<td>( (1/2)^{10} = 0.001 )</td>
</tr>
<tr>
<td>100 flips</td>
<td>( 91(1/2)^{10} \sim 0.1 )</td>
</tr>
<tr>
<td>1000 flips</td>
<td>( 991 (1/2)^{10} \sim 1 )</td>
</tr>
<tr>
<td>1,000,000 flips</td>
<td>( 999,991 (1/2)^{10} \sim 1000 )</td>
</tr>
</tbody>
</table>

- Probability (0 \( \leq p \leq 1 \)) vs Expectation (0 \( \leq E() \leq \) number of trials)
  
  \[ E(x) = p(x) \times N \]

Given an expectation, what is its probability?

*The Poisson Distribution:*

probabilities of counts of random events (radioactive decay, high similarity scores)

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

\( \mu \) = mean expectation of event

\( i \) = number of events
Distribution of solitaire wins

- I play iphone solitaire compulsively when waiting
- I win about 25% of games
- If I have played 2,000 games, how many have I won? how often have I won 2 in a row, 3 in a row, etc.

<table>
<thead>
<tr>
<th>in a row</th>
<th>p()</th>
<th>E(2000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.2</td>
<td>400</td>
</tr>
<tr>
<td>2</td>
<td>0.025</td>
<td>50</td>
</tr>
<tr>
<td>3</td>
<td>0.002</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>1e-4</td>
<td>0.3</td>
</tr>
<tr>
<td>5</td>
<td>6e-6</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Poisson distribution for ranges of events (one or more)

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

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

<table>
<thead>
<tr>
<th>( \mu )</th>
<th>( p(x &gt; 0) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>0.01</td>
<td>0.010</td>
</tr>
<tr>
<td>0.1</td>
<td>0.095</td>
</tr>
<tr>
<td>1.0</td>
<td>0.632</td>
</tr>
<tr>
<td>2.0</td>
<td>0.865</td>
</tr>
</tbody>
</table>

\[ 1 - \exp^{-\mu} \sim \mu \]

for \( \mu < 0.1 \)
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

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

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

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 \( R_n \) 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:

\[ E(s_{i,j}) = \sum_{i,j} p_i p_j s_{i,j} < 0 \] (local alignments)

Then:

\[ E(S \geq x) = Kmne^{-\lambda S} \]

\[ K < 1 \] (space correction)

\[ \lambda \text{ solution of: } \sum_{i,j} p_i p_j e^{\lambda s_{i,j}} \]

\[ E(S \geq x) \] 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

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

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

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>4x10^9/0.001</td>
<td>42</td>
</tr>
<tr>
<td>450</td>
<td>450</td>
<td>100,000</td>
<td>2x10^10/0.001</td>
<td>44</td>
</tr>
<tr>
<td>450</td>
<td>450</td>
<td>10,000,000</td>
<td>2x10^13/0.001</td>
<td>51</td>
</tr>
</tbody>
</table>
How many “bits” do I need?

\[ E(p \mid D) = p(40 \text{ bits}) \times \text{database size} \]

- \[ E(40 \mid 4,000) = 10^8 \times 4,000 = 4 \times 10^8 \] (significant)
- \[ E(40 \mid 40,000) = 10^8 \times 4 \times 10^4 = 4 \times 10^4 \] (significant)
- \[ E(40 \mid 400,000) = 10^8 \times 4 \times 10^3 = 4 \times 10^3 \] (not significant)

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) \[ p \sim 10^{-3} / 10^7 = 10^{-10} / 160,000 = 50 \text{ bits} \]

Statistics, validation, HMMs

- what is the probability of an alignment score?
  - given two sequences
  - after a database search
  - after many database searches
- Hidden Markov Models
  - transition state models
  - profile HMMs
  - HMMER3
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() ~ 1.0
• Statistical estimates can also be confirmed by searches against shuffled sequences

Smith-Waterman (ssearch36)
– highest scoring unrelated from domains

The highest scoring unrelated sequence should have an E()-value ~ 1
In one search.

What about after 10 searches?
After 100?
After 10,000?

Expectations are turned into probabilities using: 1 – exp(-E)
Highest unrelated E() values decrease with more searches

correct for multiple searches

Detectable homologs to human enzymes varying E()-value threshold
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, validation, HMMs

• what is the probability of an alignment score?
  – given two sequences
  – after a database search
  – after many database searches

• Hidden Markov Models
  – transition state models
  – profile HMMs
  – HMMER2
Why HMMs (Hidden Markov Models) ?

- HMMs provide a general purpose strategy for fitting models with adjacent features to data
  - gene models: genscan/twinscan
  - conserved regions: phastcons
  - protein domain families: profile HMMs
    - profile HMMs – Used by Pfam
      - Anders Krogh in David Haussler’s group.
      - Takes the “standard” profiles and uses HMM based “standard” mathematics to solve two problems
        - Profile-HMM scores are comparable (*)
        - Setting gap costs
      - Theoretical framework for what we are doing.
      - (* this is not really true. see later)
Figure 1 A simple hidden Markov model. A two-state HMM describing DNA sequence with a heterogeneous base composition is shown, following work by Churchill [10]. (a) State 1 (top left) generates AT-rich sequence, and state 2 (top right) generates CG-rich sequence. State transitions and their associated probabilities are indicated by arrows, and symbol emission probabilities for A, C, G and T for each state are indicated below the states. (For clarity, the begin and end states and associated state transitions necessary to model sequences of finite length have been omitted.) (b) This model generates a state sequence as a Markov chain and each state generates a symbol according to its own emission probability distribution (c). The probability of the sequence is the product of the state transitions and the symbol emissions. For a given observed DNA sequence, we are interested in inferring the hidden state sequence that ‘generated’ it, that is, whether this position is in a CG-rich segment or an AT-rich segment.


HMM transitions and emissions are probabilities

\[
\begin{array}{ccc}
HMM & = & p(B)p(M1|B)p(a|M1)p(M2|M1)p(t|M2)p(M3|M2)p(g|M3)p(E|M3) \\
& = & 1.0 \times 1.0 \times 1.0 \times 0.2 \times 0.8 \times 0.8 \times 0.25 \times 1.0 = 0.064
\end{array}
\]

\[
\begin{array}{ccc}
p(\text{atgg}|HMM) & = & p(B)p(M1|B)p(a|M1)p(M2|M1)p(t|M2)p(M3|M2)p(g|M3)p(E|M3) \\
& = & 1.0 \times 1.0 \times 1.0 \times 0.0 \times 0.0 \times 0.2 \times 0.8 \times 0.25 \times 1.0 = 0.004
\end{array}
\]
HMM – finding the best alignment
dynamic programming

HMM – alignment with dynamic programming
HMM Algorithms

1. The scoring problem: $P(\text{seq} \mid \text{model})$
   "Forward" algorithm
   (sums over all alignments)
2. The alignment problem: $\max P(\text{seq}, \text{statepath} \mid \text{model})$
   "Viterbi" algorithm
3. The training problem:
   "Forward-backward" algorithm and
   Baum-Welch expectation maximization

For profile HMMs, all three algorithms use $O(MN)$ dynamic programming -- same as "standard" Smith/Waterman and Needleman/Wunsch.
HMM Alignment

Needleman-Wunsch max log likelihood
HMM Viterbi alignment

HMM Forward (score) \[ \Sigma \] probabilities

fasta.bioch.virginia.edu/biol4230

hmmbuild – from multiple sequence alignment to hmm

fasta.bioch.virginia.edu/biol4230
HMMR3.1 – jackhmmer: psiblast with HMMs

HMMER 3.1b2 (February 2015); http://hmmer.org/
Copyright (C) 2015 Howard Hughes Medical Institute.
Freely distributed under the GNU General Public License (GPLv3).

query sequence file:             mgstm1.aa
target sequence database:        /slib2/fa_dbs/pir1.lseg

Description: Glutathione S-transferase Mu 1; GST 1-1; GST class mu 1;
Scores for complete sequences (score includes all domains):

```
<table>
<thead>
<tr>
<th></th>
<th>full sequence</th>
<th>best 1 domain</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>E-value</td>
<td>score</td>
<td>bias</td>
<td>E-value</td>
</tr>
<tr>
<td>1.4e-112</td>
<td>413.3</td>
<td>1.7</td>
<td>1.6e-124</td>
</tr>
<tr>
<td>8.5e-25</td>
<td>87.1</td>
<td>0.0</td>
<td>1.2e-24</td>
</tr>
<tr>
<td>4.6e-14</td>
<td>53.5</td>
<td>0.3</td>
<td>3.1e-14</td>
</tr>
<tr>
<td>1.0e-08</td>
<td>34.5</td>
<td>0.0</td>
<td>1.5e-08</td>
</tr>
<tr>
<td>0.00028</td>
<td>20.0</td>
<td>0.0</td>
<td>0.15</td>
</tr>
</tbody>
</table>
```

Inclusion threshold:
0.0031| 16.6| 0.0| 0.0061| 15.6| 0.0| 1.5| 1| sp|P12653|GSTF1_MAIZE |

HMMR3.1 – jackhmmer: iteration 2

```
@@ Round:                  2
@@ Included in MSA:        7 subsequences (query + 6 subseqs from 6 targets)
@@ Model size:             218 positions
@@
Scores for complete sequences (score includes all domains):

```
<table>
<thead>
<tr>
<th></th>
<th>full sequence</th>
<th>best 1 domain</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>E-value</td>
<td>score</td>
<td>bias</td>
<td>E-value</td>
</tr>
<tr>
<td>1.5e-111</td>
<td>370.7</td>
<td>0.2</td>
<td>1.7e-111</td>
</tr>
<tr>
<td>8.5e-92</td>
<td>306.1</td>
<td>0.0</td>
<td>1.1e-91</td>
</tr>
<tr>
<td>3.1e-90</td>
<td>301.0</td>
<td>0.0</td>
<td>4.2e-90</td>
</tr>
<tr>
<td>3.1e-84</td>
<td>281.4</td>
<td>0.5</td>
<td>3.6e-84</td>
</tr>
<tr>
<td>2.2e-74</td>
<td>249.2</td>
<td>0.0</td>
<td>2.8e-74</td>
</tr>
<tr>
<td>1.9e-17</td>
<td>63.0</td>
<td>0.0</td>
<td>2.3e-17</td>
</tr>
<tr>
<td>2.7e-17</td>
<td>62.6</td>
<td>0.0</td>
<td>3.5e-17</td>
</tr>
<tr>
<td>3.6e-08</td>
<td>32.7</td>
<td>0.0</td>
<td>4.5e-08</td>
</tr>
<tr>
<td>0.00016</td>
<td>20.8</td>
<td>0.0</td>
<td>0.00111</td>
</tr>
</tbody>
</table>
```

Inclusion threshold:
0.078| 12.0| 0.1| 11| 5.0| 0.0| 3.4| 2| sp|P07814|SYEP_HUMAN |
HMMER3.1 alignments w/ confidence limits

<table>
<thead>
<tr>
<th>#</th>
<th>score</th>
<th>bias</th>
<th>c-Evalue</th>
<th>i-Evalue</th>
<th>hmmfrom to</th>
<th>alifrom to</th>
<th>envfrom to</th>
<th>acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>32.4</td>
<td>0.0</td>
<td>3.4e-11</td>
<td>4.5e-08</td>
<td>54</td>
<td>169</td>
<td>47</td>
<td>169</td>
</tr>
</tbody>
</table>

Alignments for each domain:
- **domain 1** score: 32.4 bits; conditional E-value: 3.4e-11

```
sp|P20432|GSTT1_DROME  Glutathione S-transferase 1-1; DDT-dehydrochlorinase; GST class-theta 1; DDT-dehydrochlorinase; GST class-theta
```

```
HMMER3.1 – domain output
```

<table>
<thead>
<tr>
<th>#</th>
<th>score</th>
<th>bias</th>
<th>c-Evalue</th>
<th>i-Evalue</th>
<th>hmmfrom to</th>
<th>alifrom to</th>
<th>envfrom to</th>
<th>acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>43.2</td>
<td>0.0</td>
<td>1.8e-14</td>
<td>2.3e-11</td>
<td>40</td>
<td>91</td>
<td>35</td>
<td>86</td>
</tr>
<tr>
<td>2</td>
<td>17.9</td>
<td>0.0</td>
<td>9.2e-07</td>
<td>0.0012</td>
<td>127</td>
<td>196</td>
<td>136</td>
<td>207</td>
</tr>
</tbody>
</table>

Alignments for each domain:
- **domain 1** score: 43.2 bits; conditional E-value: 1.8e-14

```
sp|P04907|GSTF3_MAIZE  Glutathione S-transferase 3; GST class-phi member 3; GST-III
```

```
HMMER3.1 – domain output
```
Improving sensitivity with protein/domain family models

- HMMER3 – jackhmmer – method
  1. do HMMER (Hidden Markov Model, HMM) search with single sequence
  2. use query-HMM-based implied multiple sequence alignment to more accurate HMM
  3. repeat steps 1 and 2 with HMM

- HMMER3 – results:
  1. Less over-extension because of probabilistic alignment
  2. Used to construct Pfam domain database
     - Many protein families are too diverse for one HMM. Pfam divides families into multiple HMMs and groups in Clans
  3. Clearly homologous sequences are still missed

Missing homology beyond the HMM model

```
>tr|Q8LNM4|Q8LNM4_ORYSJ Eukaryotic aspartyl protease family protein vs
>tr|Q2QSI0|Q2QSI0_ORYSJ Glycosyl hydrolase family 9 protein, expressed OS=O (694 aa)
QRegion: 134-277:171-311 : score=508; bits=240.8; LPr=67.0 : Aspartyl protease
s-w opt: 508 E-score: 1248.7 bits: 240.8 E(1): 9.6e-68
Smith–Waterman score: 508; 62.5% identity (79.2% similar) in 144 aa overlap

130 140 150 160 170 180 190 200
Q8LNM4 TDAKSEPTCSSSNCTMVFTSSELGGTLGIVDATQGATSLGFGCVVSGIDTMQPSGQSLGQAPSSLVQ
Q2QSI0 LCESSHNHCSGGNGVCAEAEYTRA---GDGGCVVTDFGCVAFAGEAPGCGVASHTPHDGSSGVGLRKFWSLVT
170 180 190 200 210 220 230
210 220 230 240 250 260 270 280
Q8LNM4 QMNITKFSYCLTPHDSGKNSRLLLGSSAKLAGGGNSTTTPFVKTSPGDDMSQYYPIQLDGIKAGDAAIALPPSGNTVLVQ
Q2QSI0 QTGVAAFSYCLAPHDAGKNNALFLGSTAKLAGGGKTASTPFVNS
240 250 260 270 280 290 300 310
Q8LNM4 Q2QSI0
```

<table>
<thead>
<tr>
<th>Source</th>
<th>Domain</th>
<th>Start</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pfam</td>
<td>Asp</td>
<td>66</td>
<td>411</td>
</tr>
<tr>
<td>seq</td>
<td>Asp</td>
<td>1</td>
<td>22</td>
</tr>
</tbody>
</table>
Pfam misses/mis-aligns proteins distant from the model

- For diverse families, a single model can find, and miss, closely related homologs
- Even if homologs are found, alignments may be short

How much improvement with PSSMs/HMMs?

Statistics, validation, HMMs

- what is the probability of an alignment score?
  - given two sequences
    - probability of match, times number of match run starts: extreme value
  - after a database search
    - Bonferroni correction for database size
  - after many database searches
    - Bonferroni correction for number of searches (?)
    - what happens to false negatives?

- Hidden Markov Models
  - transition state models
  - profile HMMs
  - HMMER3
    - better, but sometimes missed
    - How might one find “missing” homologs?