Protein Sequence Comparison
and Protein Evolution
(What BLAST does/Why BLAST works

William R. Pearson
faculty.virginia.edu/wrpearson
pearson@virginia.edu

Effective Similarity Searching in Practice

1. Always search protein databases (possibly with translated DNA)
2. Use E()-values, not percent identity, to infer homology
   - E() < 0.001 is significant in a single search
3. Search smaller (comprehensive) databases
4. Change the scoring matrix for:
   - short sequences (exons, reads)
   - short evolutionary distances (mammals, vertebrates, a-proteobacteria)
   - high identity (>50% alignments) to reduce over-extension
5. All methods (pairwise, HMM, PSSM) miss homologs, and find homologs the other methods miss
Homology Fundamentals

• Homologous sequences are unexpectedly similar (excess similarity)
  – excess compared to ??? – random similarity (similarity by chance, E()-value)

• Non-significant similarity IS NOT evidence for non-homology
  – significant similarity to a protein of different structure shows non-homology

• Homology at the (entire) sequence level is different from homology at the residue level
  – sequence homology is inferred from statistics
  – residue homology REQUIRES sequence homology

Establishing homology from statistically significant similarity

Why BLAST works

• For most proteins, homologs are easily found over long evolutionary distances (500 My – 2 By) using standard approaches (BLAST, FASTA)
• Difficult for distant relationships or very short domains
• Most default search parameters are optimized for distant relationships and work well
• Not every aligned residue is homologous
  – but with significant similarity, there is an homologous domain
This talk is not about:

- **Alignment**
  - Alignment quality may be more sensitive to parameter choice
  - Multiple sequences for biologically accurate alignments
- **Inferring Protein Function**
  - Homology (common ancestry) implies common structure (guaranteed), not necessarily common function
  - Homologs have different functions
  - Non-homologs have similar (or identical) functions
- **The best sequences for building trees**
  - Protein sequences are clearly best for establishing homology, but DNA sequences may be better for resolving recent divergence

Homologues share a common ancestor
When do we infer homology?

Homology <=> structural similarity  
?  sequence similarity

Bovine trypsin (5ptp)
Structure:  
E()<10^{-23}; RMSD 0.0 Å
Sequence:  
E()<10^{-84} 100% 223/223

S. griseus trypsin (1sgt)
E()<10^{-14} RMSD 1.6 Å
E()<10^{-19} 36%; 226/223

S. griseus protease A (2sga)
E()<10^{-4}; RMSD 2.6 Å
E()<2.6 25%; 199/181

When can we infer non-homology?

Non-homologous proteins have different structures

Bovine trypsin (5ptp)
Structure:  
E()<10^{-23} RMSD 0.0 Å
Sequence:  
E()<10^{-84} 100% 223/223

Subtilisin (1sbt)
E() >100
E()<280; 25% 159/275

Cytochrome c4 (1etp)
E() >100
E()<5.5; 23% 171/190
Homology and *EXCESS* similarity

- Why are proteins (sequences or structures) similar – alternative hypotheses?
  1. common ancestry – *homology*
  2. convergence from independent origins due to functional constraints
- We infer *homology* from *excess* similarity
  - if no excess, could happen by chance (independently)
- To recognize *excess* similarity we need to know what *random* similarity looks like

---

Query: *atp6_human.aa ATP synthase a chain – 226 aa*
Library: PIR1 Annotated (rel. 66)
5190103 residues in 13351 sequences
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
The PAM250 matrix

<table>
<thead>
<tr>
<th></th>
<th>Cys</th>
<th>Ser</th>
<th>Thr</th>
<th>Thr</th>
<th>Pro</th>
<th>Ala</th>
<th>Gly</th>
<th>Asp</th>
<th>Glu</th>
<th>Gln</th>
<th>His</th>
<th>Ile</th>
<th>Leu</th>
<th>Lys</th>
<th>Met</th>
<th>Phe</th>
<th>Tyr</th>
<th>Trp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cys</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Ser</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Thr</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Thr</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Pro</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Ala</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Gly</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Asp</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Glu</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Gln</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>His</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Ile</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Leu</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Lys</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Met</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Phe</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Tyr</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Trp</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

The PAM250 matrix is used to calculate the similarity between two sequences based on the number of matches and mismatches. The matrix assigns a score to each possible alignment of two residues, with higher scores indicating greater similarity.
Where do scoring matrices come from?

- Scoring matrices can be designed for different evolutionary distances (less=shallow; more=deep)
- Deep matrices allow more substitution

\[ \lambda S = \log \left( \frac{q_{ij}}{p_i p_j} \right) \]

frequency of alignment by chance

<table>
<thead>
<tr>
<th>Protein Accession</th>
<th>Description</th>
<th>Score</th>
<th>% Alignment</th>
<th>PDB Score</th>
<th>Sim. Score</th>
<th>E Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>sp</td>
<td>P00846</td>
<td>ATP6_HUMN ATP synthase a chain</td>
<td>( 226) 1409 325.8 5.8e-9 1.000 1.226</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sp</td>
<td>P00847</td>
<td>ATP6_BOVIN ATP synthase a chain</td>
<td>( 226) 1157 270.5 2.5e-7 3.079 0.951 226</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sp</td>
<td>P00848</td>
<td>ATP6_MOGE ATP synthase a chain</td>
<td>( 226) 1118 261.7 1.2e-7 0.075 0.916 226</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sp</td>
<td>P00849</td>
<td>ATP6_KDHA ATP synthase a chain</td>
<td>( 226) 745 176.4 4.0e-5 0.533 0.847 229</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sp</td>
<td>P00851</td>
<td>ATP6_DROFY ATP synthase a chain</td>
<td>( 224) 473 115.0 1.7e-6 0.378 0.721 222</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sp</td>
<td>P00854</td>
<td>ATP6_YEAST ATP synthase a chain pre</td>
<td>( 259) 428 104.7 2.3e-6 0.353 0.694 232</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sp</td>
<td>P00855</td>
<td>ATP6_EMBRI ATP synthase a chain pre</td>
<td>( 256) 365 90.4 4.8e-5 0.304 0.691 230</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sp</td>
<td>P14862</td>
<td>ATP6_KOCHI ATP synthase a chain</td>
<td>( 257) 353 87.7 3.2e-8 0.313 0.650 214</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sp</td>
<td>P06289</td>
<td>ATP6_MARPO ATP synthase a chain</td>
<td>( 248) 126 36.0 0.01 0.221 0.571 231</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sp</td>
<td>P06452</td>
<td>ATP6_PEA Chloroplast ATP synthase a</td>
<td>( 247) 144 40.1 0.00062 0.242 0.580 231</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sp</td>
<td>P06457</td>
<td>ATP6_ECOLI ATP synthase a chain</td>
<td>( 291) 283 71.7 2.3e-10 0.311 0.667 180</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sp</td>
<td>P08289</td>
<td>ATP6_MAIZE Chloroplast ATP synthase a</td>
<td>( 247) 134 39.9 0.00072 0.250 0.586 232</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sp</td>
<td>P08290</td>
<td>ATP6_MARPO Chloroplast ATP synthase a</td>
<td>( 247) 138 38.8 0.0016 0.242 0.580 231</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sp</td>
<td>P08444</td>
<td>ATP6_MARPO Chloroplast ATP synthase a</td>
<td>( 247) 138 38.8 0.0016 0.242 0.580 231</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sp</td>
<td>P02718</td>
<td>ATP6_SYN3 ATP synthase a chain</td>
<td>( 276) 142 39.7 0.00952 0.265 0.571 170</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sp</td>
<td>P02651</td>
<td>ATP6_SYNL Chloroplast ATP synthase</td>
<td>( 247) 138 38.8 0.0016 0.242 0.580 231</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sp</td>
<td>P02717</td>
<td>ATP6_SYN3 ATP synthase a chain</td>
<td>( 276) 142 39.7 0.00952 0.265 0.571 170</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sp</td>
<td>P02651</td>
<td>ATP6_SYNL Chloroplast ATP synthase</td>
<td>( 247) 138 38.8 0.0016 0.242 0.580 231</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The best scores are:

| Query: atp6_human.aa ATP synthase a chain - 226 aa |
| Library: 5190103 residues in 13351 sequences |

<table>
<thead>
<tr>
<th>Score</th>
<th>PDB Score</th>
<th>Sim. Score</th>
<th>E Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>7/16/16</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accession</td>
<td>Description</td>
<td>Length</td>
<td>E-value</td>
</tr>
<tr>
<td>-----------</td>
<td>-------------</td>
<td>--------</td>
<td>---------</td>
</tr>
<tr>
<td>sp</td>
<td>P0AB98</td>
<td>ATP6_ECOLI ATP synthase a chain (ATP6_ECOLI)</td>
<td>271</td>
</tr>
<tr>
<td>sp</td>
<td>P06451</td>
<td>ATP6_SPIOL Chloroplast ATP synthase a chain (ATP6_SPIOL)</td>
<td>274</td>
</tr>
<tr>
<td>sp</td>
<td>P69371</td>
<td>ATP6_ATRBE Chloroplast ATP synthase a chain (ATP6_ATRBE)</td>
<td>271</td>
</tr>
<tr>
<td>sp</td>
<td>P08444</td>
<td>ATP6_SYNP6 ATP synthase a chain (ATP6_SYNP6)</td>
<td>271</td>
</tr>
<tr>
<td>sp</td>
<td>P06452</td>
<td>ATP6_PEA Chloroplast ATP synthase a chain (ATP6_PEA)</td>
<td>266</td>
</tr>
<tr>
<td>sp</td>
<td>P30391</td>
<td>ATP6_EUGGR Chloroplast ATP synthase a chain (ATP6_EUGGR)</td>
<td>265</td>
</tr>
<tr>
<td>sp</td>
<td>P0C2Y5</td>
<td>ATP6_ORYSA Chloroplast ATP synthase a chain (ATP6_ORYSA)</td>
<td>260</td>
</tr>
<tr>
<td>sp</td>
<td>P27178</td>
<td>ATP6_SYNY3 ATP synthase a chain (ATP6_SYNY3)</td>
<td>260</td>
</tr>
<tr>
<td>sp</td>
<td>P06289</td>
<td>ATP6_MARPO Chloroplast ATP synthase a chain (ATP6_MARPO)</td>
<td>250</td>
</tr>
<tr>
<td>sp</td>
<td>P07925</td>
<td>ATP6_MAIZE ATP synthase a chain (ATP6_MAIZE)</td>
<td>215</td>
</tr>
<tr>
<td>sp</td>
<td>P68526</td>
<td>ATP6_TRITI ATP synthase a chain (ATP6_TRITI)</td>
<td>209</td>
</tr>
<tr>
<td>sp</td>
<td>P00854</td>
<td>ATP6_YEAST ATP synthase a chain (ATP6_YEAST)</td>
<td>204</td>
</tr>
<tr>
<td>sp</td>
<td>P00852</td>
<td>ATP6_EMENI ATP synthase a chain (ATP6_EMENI)</td>
<td>178</td>
</tr>
<tr>
<td>sp</td>
<td>P00849</td>
<td>ATP6_XENLA ATP synthase a chain (ATP6_XENLA)</td>
<td>173</td>
</tr>
<tr>
<td>sp</td>
<td>P00847</td>
<td>ATP6_BOVIN ATP synthase a chain (ATP6_BOVIN)</td>
<td>172</td>
</tr>
<tr>
<td>sp</td>
<td>P14862</td>
<td>ATP6_COCHE ATP synthase a chain (ATP6_COCHE)</td>
<td>171</td>
</tr>
<tr>
<td>sp</td>
<td>P00848</td>
<td>ATP6_MOUSE ATP synthase a chain (ATP6_MOUSE)</td>
<td>166</td>
</tr>
</tbody>
</table>

**Query:** ATP6(ecoli).aa

**Library:** 5190103 residues in 13351 sequences

**Homology is Transitive (on domains)**
Homology and *EXCESS* similarity

The importance of perspective

- We use the E()-value to infer homology based on excess similarity *BETWEEN the QUERY and the SUBJECT* sequences.
- Two homologous sequences may not share excess similarity in a BLAST search with one or the other as query, but share significant similarity to a third sequence (or to a PSSM or HMM).
- If two sequences share significant similarity to the same sequence in the same region, we can infer homology.

As always, non-excess similarity does not imply non-homology.

Homology and Domains – Histone acetyltransferase KAT2B

The best scores are:

<table>
<thead>
<tr>
<th>Protein (Organism)</th>
<th>E-value</th>
<th>bits</th>
<th>Identities</th>
<th>Similarity</th>
<th>Alignment Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>KAT2B_HUMAN Histone acetyltransferase KAT2B(Human)</td>
<td>3.82e-145.6</td>
<td>0.1400 1.000</td>
<td>832</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KAT2A_HUMAN Histone acetyltransferase KAT2A(Human)</td>
<td>8.28e-1044.9</td>
<td>0.0721 0.870</td>
<td>313</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GCN5_SCHPO Histone acetyltransferase gm5(Schizosaccharomyces pombe)</td>
<td>3.87e-334.7</td>
<td>0.3690 0.769</td>
<td>353</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GCN5_YEAST Histone acetyltransferase GCN5(Saccharomyces cerevisiae)</td>
<td>3.52e-306.2</td>
<td>0.4690 0.760</td>
<td>353</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GCN5_ORYSJ Histone acetyltransferase GCN5(Octopus Howardensis)</td>
<td>1.35e-867.3</td>
<td>0.4750 0.769</td>
<td>353</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GCN5_ARATH Histone acetyltransferase GCN5(Plantae)</td>
<td>1.6e-792.4</td>
<td>0.4530 0.756</td>
<td>369</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BPTF_HUMAN Nucleosome-remodeling factor sub(Bovine)</td>
<td>8.03e-286.6</td>
<td>0.5070 0.804</td>
<td>97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NU301_DROME Nucleosome-remodeling factor su(Drosophila melanogaster)</td>
<td>7.70e-276.9</td>
<td>0.5130 0.819</td>
<td>94</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CECR2_HUMAN Cat eye syndrome critical regio(Human)</td>
<td>9.23e-232.1</td>
<td>0.5000 0.790</td>
<td>105</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BRD4_HUMAN Bromodomain-containing protein 4(Human)</td>
<td>1.84e-214.5</td>
<td>0.5000 0.790</td>
<td>116</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BRD4_MOUSE Bromodomain-containing protein 4(Mouse)</td>
<td>1.84e-214.5</td>
<td>0.5000 0.790</td>
<td>116</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BA2A_HUMAN Bromodomain adjacent to zinc f1(Human)</td>
<td>8.05e-211.7</td>
<td>0.5000 0.790</td>
<td>123</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BA2A_XENLA Bromodomain adjacent to zinc f1(Xenopus)</td>
<td>3.83e-206.3</td>
<td>0.5000 0.790</td>
<td>123</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FSH_DROME Homotic protein female sterile(Drosophila melanogaster)</td>
<td>2.05e-205.8</td>
<td>0.5000 0.790</td>
<td>129</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BA2A_MOUSE Bromodomain adjacent to zinc f1(Mouse)</td>
<td>1.82e-204.8</td>
<td>0.5000 0.790</td>
<td>125</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BRDT_MACPA Bromodomain testis-specific prot(Schizosaccharomyces pombe)</td>
<td>1.97e-197.8</td>
<td>0.5000 0.790</td>
<td>109</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BRD3_HUMAN Bromodomain-containing protein 3(Human)</td>
<td>7.81e-194.9</td>
<td>0.5000 0.790</td>
<td>116</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Homology and Domains – Histone acetyltransferase KAT2B**

### Similarity searches for homology

- Homologous sequences are unexpectedly similar (excess similarity)
  - excess compared to ??? – random similarity (similarity by chance, E()-value)
- In a similarity search, excess similarity reflects the perspective of the query sequence
  - different queries can reveal excess similarity
  - homology in the aligned region
- Non-significant similarity IS NOT evidence for non-homology
  - significant similarity to a protein of different structure shows non-homology
Effective similarity searching

- Use protein/translated DNA comparisons
- Modern sequence similarity searching is highly efficient, sensitive, and reliable – homologs are homologs
  - similarity statistics are accurate
  - databases are large
  - most queries will find a significant match
- Improving similarity searches
  - protein (translated DNA) similarity searches
  - smaller databases
  - appropriate scoring matrices for short reads/assemblies
  - appropriate alignment boundaries
- Extracting more information from annotations
  - homologous over extension
  - scoring sub-alignments to identify homologous domains
- All methods (pairwise, HMM, PSSM) miss homologs
  - all methods find genuine homologs the other methods miss

DNA vs protein sequence comparison

The best scores are:

<table>
<thead>
<tr>
<th>Protein</th>
<th>DNA</th>
<th>tfastx3</th>
<th>prot.</th>
</tr>
</thead>
<tbody>
<tr>
<td>DMGST</td>
<td>D.melanogaster GST-1</td>
<td>1.3e-164</td>
<td>4.1e-109</td>
</tr>
<tr>
<td>MDGST1</td>
<td>M.domestica GST-1 gene</td>
<td>2e-77</td>
<td>3.0e-95</td>
</tr>
<tr>
<td>LUCGLTR</td>
<td>Lucilia cuprina GST</td>
<td>1.5e-72</td>
<td>5.2e-91</td>
</tr>
<tr>
<td>MDGST2A</td>
<td>M.domesticus GST-2 mRNA</td>
<td>9.3e-53</td>
<td>1.4e-77</td>
</tr>
<tr>
<td>MDNF1</td>
<td>M.domestica nfi gene. 10</td>
<td>4.6e-51</td>
<td>2.8e-77</td>
</tr>
<tr>
<td>MDNF6</td>
<td>M.domestica nfi gene. 10</td>
<td>2.8e-51</td>
<td>4.2e-77</td>
</tr>
<tr>
<td>MDNF7</td>
<td>M.domestica nfi7 gene. 10</td>
<td>6.1e-47</td>
<td>9.2e-77</td>
</tr>
<tr>
<td>AGGST15</td>
<td>A.gambiae GST mRNA</td>
<td>3.1e-58</td>
<td>4.2e-76</td>
</tr>
<tr>
<td>CYUB7958</td>
<td>Culicoides GST</td>
<td>1.8e-41</td>
<td>4.0e-73</td>
</tr>
<tr>
<td>AGG368211</td>
<td>A.gambiae GST-1 mRNA</td>
<td>1.5e-46</td>
<td>2.8e-55</td>
</tr>
<tr>
<td>BMO6502</td>
<td>Bombyx mori GST mRNA</td>
<td>1.1e-23</td>
<td>8.8e-50</td>
</tr>
<tr>
<td>AGSSXGT12</td>
<td>A.gambiae GST-1 gene</td>
<td>2.3e-16</td>
<td>4.5e-46</td>
</tr>
<tr>
<td>NDZGLUGTRA</td>
<td>Manduca sexta GST</td>
<td>5.7e-07</td>
<td>2.5e-30</td>
</tr>
<tr>
<td>RLGTGARNH</td>
<td>R. legominosarum gztA</td>
<td>0.0029</td>
<td>3.2e-13</td>
</tr>
<tr>
<td>HUNGST2A</td>
<td>H. sapiens GSTT2</td>
<td>0.32</td>
<td>3.3e-10</td>
</tr>
<tr>
<td>HSGSTT1</td>
<td>H.sapiens GSTT1 mRNA</td>
<td>7.2</td>
<td>8.4e-13</td>
</tr>
<tr>
<td>ECAD000319</td>
<td>E. coli hypoth.ppt.</td>
<td>—</td>
<td>4.7e-10</td>
</tr>
<tr>
<td>MNDOCHNA</td>
<td>Methyl. dichlorometh. DH</td>
<td>—</td>
<td>1.1e-09</td>
</tr>
<tr>
<td>BCU19883</td>
<td>Burkholderia maleylactate rd.</td>
<td>—</td>
<td>1.2e-09</td>
</tr>
<tr>
<td>NPU43126</td>
<td>Naegleri fowleri GST</td>
<td>—</td>
<td>3.2e-07</td>
</tr>
<tr>
<td>SP505GST</td>
<td>Sphingomonas paucim</td>
<td>—</td>
<td>1.8e-06</td>
</tr>
<tr>
<td>EN1838</td>
<td>H. sapiens maleylacto. iso.</td>
<td>—</td>
<td>2.1e-06</td>
</tr>
<tr>
<td>HSUE6529</td>
<td>Human GSTZ1</td>
<td>—</td>
<td>3.0e-06</td>
</tr>
<tr>
<td>SYCCPNC</td>
<td>Synechocystis GST</td>
<td>—</td>
<td>1.2e-05</td>
</tr>
<tr>
<td>HSEF1GMR</td>
<td>H.sapiens EF1g mRNA</td>
<td>—</td>
<td>9.0e-05</td>
</tr>
</tbody>
</table>
Why smaller databases produce more sensitive searches – statistics

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

What is a “bit” score (I)?

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

2. 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.

3. How often should a score occur by chance? In a 400 * 400 alignment, there are \( \sim 160,000 \) places where the alignment could start by chance, so we expect a score of 40 bits would occur:
   - \[ P(S_{\text{bit}} > x) = 1 - \exp(-mn2^{-x}) \sim mn2^{-x} \]
   - \( 400 \times 400 \times 2^{-40} = 160,000 / 2^{40} (10^{13.3}) = 1.5 \times 10^{-7} \) times
   - Thus, the probability of a 40 bit score in ONE alignment is \( \sim 10^{-7} \)
What is a “bit” score (I)?

4. But we did not ONE alignment, we did 4,000, 40,000, 500,000, or 20 million alignments when we searched the database:
   - \( E(S_{\text{bit}} \mid D) = p(40 \text{ bits}) \times \text{database size} \)
   - \( E(40 \mid 4,000) = 10^{-7} \times 4,000 = 4 \times 10^{-4} \) (significant)
   - \( E(40 \mid 40,000) = 10^{-7} \times 4 \times 10^{4} = 4 \times 10^{-3} \) (not significant)
   - \( E(40 \mid 500,000) = 10^{-7} \times 5 \times 10^{5} = 0.05 \) (not significant)
   - \( E(40 \mid 20 \text{ million}) = 10^{-7} \times 2.0 \times 10^{7} = 2.0 \) (not significant)

Bonferroni correction for multiple tests – each alignment is a test

Not significant does not mean not-homologous

E()-values are conservative frequentist estimates that similarity occurred by chance

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

Bonferroni correction:
\[
E(S' > x \mid D) = P(D(\text{# of tests})
\]

With modern sequence databases (thousands of complete proteomes), \( E() < 10^{-10} \) is routine for sequences >25% identical, after correcting for 10,000,000 sequences (tests)
How many “bits” do I need?

E() = p() x database size

E(40 | 4,000) = 10^-7 x 4,000 = 4 x 10^-4 (significant)
E(40 | 40,000) = 10^-7 x 4 x 10^4 = 4 x 10^-3 (not significant)
E(40 | 500,000) = 10^-7 x 5 x 10^6 = 0.05 (not significant)

To get E() ~ 10^-3, how many bits do I need? p = m n 2^-bits

bits = -log2(p/(m n)) = -log2(E/((database_size m n))
genome (10,000) p ~ 10^-9/10^4 = 10^-7/160,000 = 40 bits
SwissProt (500,000) p ~ 10^-3/10^6 = 10^-9/160,000 = 47 bits
Uniprot/NR (10^7) p ~ 10^-3/10^9 = 10^-10/160,000 = 50 bits

What database to search?

- Search the smallest comprehensive database likely to contain your protein
  - vertebrates – human proteins (40,000)
  - fungi – S. cerevisiae (6,000)
  - bacteria – E. coli, gram positive, etc. (<100,000)
- Search a richly annotated protein set (SwissProt, >500,000)
- Always search NR (> 50 million) LAST
- Never Search “GenBank” (DNA)
Scoring matrices

- Scoring matrices can set the evolutionary look-back time for a search
  - Lower PAM (PAM10/MDM10 ... PAM60) for closer (10% ... 50% identity)
  - Higher BLOSUM for higher conservation (BLOSUM50 distant, BLOSUM80 conserved)
- Shallow scoring matrices for short domains/short queries (metagenomics)
  - Matrices have “bits/position” (score/position), 40 aa at 0.45 bits/position (BLOSUM62) means 18 bit ave. score (50 bits significant)
- Deep scoring matrices allow alignments to continue, possibly outside the homologous region

Where do scoring matrices come from?

<table>
<thead>
<tr>
<th>Pam40</th>
<th>Pam250</th>
</tr>
</thead>
<tbody>
<tr>
<td>A  8</td>
<td>A  2</td>
</tr>
<tr>
<td>R -9</td>
<td>R -2</td>
</tr>
<tr>
<td>N -4</td>
<td>N  0</td>
</tr>
<tr>
<td>D -4</td>
<td>D  0</td>
</tr>
<tr>
<td>E -3</td>
<td>E  0</td>
</tr>
<tr>
<td>I -6</td>
<td>I -1</td>
</tr>
<tr>
<td>L -8</td>
<td>L -2</td>
</tr>
</tbody>
</table>

\[ S_{i,j} = \log_b \left( \frac{q_{i,j}}{p_i p_j} \right) \]

\[ q_i : \text{replacement frequency at PAM40, 250} \]
\[ q_{RN(40)} = 0.000435 \quad p_R = 0.051 \]
\[ q_{RN(250)} = 0.002193 \quad p_N = 0.043 \]
\[ \lambda = \log_2 \left( \frac{q}{p p} \right) \]
\[ \lambda_S = -2.333 \]
\[ \lambda_S = -2.333 / \lambda = -7 \]
\[ \lambda_S = 0 \]
PAM matrices and alignment length

Short domains require “shallow” scoring matrices

Empirical matrix performance
(median results from random alignments)

<table>
<thead>
<tr>
<th>Matrix</th>
<th>target % ident</th>
<th>bits/position</th>
<th>aln len (50 bits)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VT160 -12/-2</td>
<td>23.8</td>
<td>0.26</td>
<td>192</td>
</tr>
<tr>
<td>BLOSUM50 -10/-2</td>
<td>25.3</td>
<td>0.23</td>
<td>217</td>
</tr>
<tr>
<td>BLOSUM62* -11/-1</td>
<td>28.9</td>
<td>0.45</td>
<td>111</td>
</tr>
<tr>
<td>VT120 -11/-1</td>
<td>27.4</td>
<td>1.03</td>
<td>48</td>
</tr>
<tr>
<td>VT80 -11/-1</td>
<td>51.9</td>
<td>1.55</td>
<td>32</td>
</tr>
<tr>
<td>PAM70* -10/-1</td>
<td>33.8</td>
<td>0.64</td>
<td>78</td>
</tr>
<tr>
<td>PAM30* -9/-1</td>
<td>45.5</td>
<td>1.06</td>
<td>47</td>
</tr>
<tr>
<td>VT40 -12/-1</td>
<td>72.7</td>
<td>2.76</td>
<td>18</td>
</tr>
<tr>
<td>VT20 -15/-2</td>
<td>84.6</td>
<td>3.62</td>
<td>13</td>
</tr>
<tr>
<td>VT10 /16/-2</td>
<td>90.9</td>
<td>4.32</td>
<td>12</td>
</tr>
</tbody>
</table>

HMMs can be very "deep"
Scoring matrices affect alignment boundaries (homologous over-extension)

Scoring Matrices - Summary

- PAM and BLOSUM matrices greatly improve the sensitivity of protein sequence comparison – low identity with significant similarity
- PAM matrices have an evolutionary model - lower number, less divergence – lower=closer; higher=more distant
- BLOSUM matrices are sampled from conserved regions at different average identity – higher=more conservation
- Short alignments (domains, exons, reads) require shallow (higher information content) matrices
- Shallow matrices set maximum look-back time
Effective similarity searching

- Modern sequence similarity searching is highly efficient, sensitive, and reliable – homologs are homologs
  - similarity statistics are accurate
  - databases are large
  - most queries will find a significant match
- Improving similarity searches
  - smaller databases
  - appropriate scoring matrices for short reads/assemblies
  - appropriate alignment boundaries
- Extracting more information from annotations
  - homologous over extension
  - scoring sub-alignments to identify homologous domains
- All methods (pairwise, HMM, PSSM) miss homologs
  - all methods find genuine homologs the other methods miss

Overextension into random sequence
Can homologous proteins have different structures?

Scoring matrices affect alignment boundaries (homologous over-extension)
Overextension into random sequence

Sub-alignment scoring detects overextension

> g726|15075250|E6SGT6|E6SGT6_THEM7 Heavy metal translocating P-type ATPase EC=3.6.3.4
Length=888
Score = 299 bits (766), Expect = 1e-90, Method: Compositional matrix adjust.
Identities = 178/341 (50%), Positives = 224/341 (66%), Gaps = 19/341 (6%)

Query 84
FLFWFAALFNYWPTEGKILMFGKLEKVLITLILLGKTLEAVAKGRTSEAIKKLMGLKA

Sbjct 312
WLYSTVAVAFPQIFPSMALAEVFYDVTAVVVALVNLGLALELRARGRTSEAIKKLIGLQA

Query 144
KRARVIRGGRELDIPVEAVLAGDLVVVRPGEKIPVDGVVEEGASAVDESMLTGESLPVDK

Sbjct 372
RTARVVRDGTEVDIPVEEVLVGDIVVVRPGEKIPVDGVVIEGTSSVDESMITGESIPVEM

Query 204
QPGDTVIGATLNKQGSFKFRATKVGRDTALAQIISVVEEAQGSKAPIQRLADTISGYFVP

Sbjct 432
KPGDEVIGATINQTGSFRFRATKVGKDTALSQIIRLVQDAQGSKAPIQRIVDRVSHYFVP

Query 264
VVVSATIFPWAVAPFTRALLFTANLVAICPLCALATLPSMTVDGKAGAEK

Sbjct 552
AVLIALAVAVVVFPEPAPAYLYVFTTILACPCALATLPSLTVIGKAGAEK

Query 324
VNNGLAVITFFWYFAVAPENFTRALLNFTAVLVIACPCALGLATPTSIMVGTGKGAEKG

Sbjct 684
AVLILAIVAAVVWVFGPEPAYIYALIVFVTTLIIACPCALGLATPTSLTVGIGKGAEQG

Query 375
TVAFQKNTGFKLKIPIGQAQLQREVAASESIVISAYPIVGV

Sbjct 688
---AERKSEPHALTAYVEGALARGLAPEGSFARLPIGHOV

Smith-Waterman score: 765; 49.7% identity (73.3% similar) in 344 aa overlap (81-415:309-642)
Scoring domains highlights over extension

<table>
<thead>
<tr>
<th>Seq</th>
<th>Score</th>
<th>E-value</th>
<th>Id</th>
<th>Q-value</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>sp</td>
<td>SRC8_HUMAN Src substrate cortactin (550 aa)</td>
<td>84.7</td>
<td>1.2e-159</td>
<td>84.7%</td>
<td>1-550:11-563</td>
</tr>
<tr>
<td>sp</td>
<td>HCLS1_MOUSE Hematopoietic cell-sp (486 aa)</td>
<td>44.1</td>
<td>4.1e-61</td>
<td>44.1%</td>
<td>1-548:1-485</td>
</tr>
</tbody>
</table>

**Q** = -10 log(p)  
Q > 30.0 -> p < 0.001

---

### Homology, non-homology, and over-extension

- **Sequences that share statistically significant sequence similarity** are homologous (simplest explanation)
- But not all regions of the alignment contribute uniformly to the score:
  - lower identity/Q-value because of non-homology (over-extension)?
  - lower identity/Q-value because more distant relationship (domains have different ages)?
- **Test** by searching with isolated region:
  - can the distant domain (?) find closer (significant) homologs?
- **Similar** (homology) or distinct (non-homology) structure is the gold standard
- **Multiple** sequence alignment can obscure over-extension:
  - if the alignment is over-extended, part of the alignment is NOT homologous
Improving sensitivity with protein/domain family models

• PSI-BLAST - method
  1. do BLAST search
  2. use query-based implied multiple sequence alignment to build Position Specific Scoring Matrix (PSSM)
  3. repeat steps 1 and 2 with PSSM, for 5 – 10 iterations

• PSI-BLAST – results:
  1. Typically 2X as sensitive as single sequence methods
  2. Over-extension can cause PSSM contamination

• 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

For diverse families, a single model can find, and miss, closely related homologs.

Even if homologs are found, alignments may be short.
**Homology Fundamentals**

- Homologous sequences are unexpectedly similar (excess similarity)
  - excess compared to ?? – random similarity (similarity by chance, E()-value)
- Non-significant similarity IS NOT evidence for non-homology
  - significant similarity to a protein of different structure shows non-homology
- Homology at the (entire) sequence level is different from homology at the residue level
  - sequence homology is inferred from statistics
  - residue homology REQUIRES sequence homology

**Effective Similarity Searching in Practice**

1. Always search protein databases (possibly with translated DNA)
2. Use E()-values, not percent identity, to infer homology
   - E() < 0.001 is significant in a single search
3. Search smaller (comprehensive) databases
4. Change the scoring matrix for:
   - short sequences (exons, reads)
   - short evolutionary distances (mammals, vertebrates, a-proteobacteria)
   - high identity (>50% alignments) to reduce over-extension
5. All methods (pairwise, HMM, PSSM) miss homologs, and find homologs the other methods miss