Similarity Searching II

Algorithms, statistics, scoring matrices

Biol4230 Tues, Feb 2, 2016
Bill Pearson wrp@virginia.edu 4-2818 Jordan 6-057

Goals of today’s lecture:

• Quick overview of alignment algorithms
  – local vs global
  – dynamic programming
  – gaps and alignment graphs
  – non-overlapping local alignments
• Improving search performance - local alignment statistics
  – the extreme value distribution
  – why database size matters
  – evaluating statistical accuracy
• Where scoring matrices come from
  – scoring matrices as log-odds matrices
  – short alignments, shallow matrices
  – shallow matrices, higher identity alignment
  – matrix "depth" and evolutionary look-back

To learn more:

• Alignment algorithms:
  – Bioinformatics and Functional Genomics (BFG), Ch. 3 p 76 – 80
• Search sensitivity:
• Statistical accuracy:
  – BFG Ch. 3, pp 88 – 90
• Scoring matrices part I
  – BFG Ch. 3, pp. 57 – 76
  – Pearson (2013) Curr Protocols Bioinformatics 3.5.1-3.5.9
Similarity searching II – algorithms, statistics, and scoring matrices

• Global and local alignments
  – Global alignments can be more sensitive for globally similar proteins
  – Local alignments are robust to partial sequences, domain homologies
• Local similarity scores are well described by the extreme value distribution
  – E()-value depends on similarity score AND database size
  – A 50 bit score is almost always significant
  – E()-values are not good measures of evolutionary distance
• Scoring matrices can be designed for long (deep) or short (shallow) evolutionary distances (large/small amounts of change)
  – "shallow" matrices provide more statistical significance for each aligned position, but require higher homologs
  – "deep" matrices can find more distant homologs, but require longer alignments

Algorithms for sequence alignment

• How do we get from this:
  >ATP6_HUMAN ATP synthase a chain (ATPase protein 6)
  MNENLFASFIAPTILGLPAAVLIILFFPLLIPTSKYLHLLITQQNLIKLTSQMNTMHTKGRTWSL
  MLVSLIIIFIIAATNLGLLPHFSFTPQTQLSNHLAMIAIPWAVTVIMGFRSKIKNALSALHLFQGPTFPLIFPM
  LVIIETISLLIQPMALAVVLTAMITAGHLMLHGLIGSATLAMSTINLPSLIIISMTILEIAVALIQ
  AYVFIIKSLVLHIDNT

• And this:
  >ATP6_HUMAN ATP synthase a chain (ATPase protein 6)
  MNENLFASFIAPTILGLPAAVLIILFFPLLIPTSKYLHLLITQQNLIKLTSQMNTMHTKGRTWSL
  MLVSLIIIFIIAATNLGLLPHFSFTPQTQLSNHLAMIAIPWAVTVIMGFRSKIKNALSALHLFQGPTFPLIFPM
  LVIIETISLLIQPMALAVVLTAMITAGHLMLHGLIGSATLAMSTINLPSLIIISMTILEIAVALIQ
  AYVFIIKSLVLHIDNT

• To …
Algorithms for sequence alignment

• To this:

To this:

>sp|P0AB98|ATP6_ECOLI ATP synthase subunit a; ATP synthase F0 subunit;
Length=271
Score = 47.9 bits (178), Expect = 3e-06
Identities = 55/199 (27%), Positives = 113/199 (56%), Gaps = 37/199 (18%)

<table>
<thead>
<tr>
<th>Query</th>
<th>SFIAPFIILGLPAAVLLLIFPPFLTIPISTYKLINHRLITITQQML I KLTSKQMWMTMNHTGGTTWSLML</th>
<th>72</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sbjct</td>
<td>SMFFSVVLGL---LFLVLFRSVAKKATS---</td>
<td>105</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Query</th>
<th>VSLIFPIATNMLGLLernational---HSF------TPTQSLNNLAMAIPLWAGTVIMGFRSKI</th>
<th>121</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sbjct</td>
<td>LTIQVWFLMLMLDLPDPFQIAEVLGLPALKVPSADVNLTVLSAGF----</td>
<td>167</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Query</th>
<th>KNAALVLQPGQFTPIPPLHMLVAVLRTANITAGHLMLHIGSATLM</th>
<th>181</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sbjct</td>
<td>MKIGEVKTEFQXPFPNHPPAFTFPVPWNLIGEGVSLKKPVSLGLKFPNMYAGELIFILLIGLPLW</td>
<td>232</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Query</th>
<th>STINLPSTLIFTLLTILEIAVIALQAYVFTILVSLY</th>
<th>222</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sbjct</td>
<td>SQWILNVPHAIFHIIIT---------------LQAFIFMVTIVY</td>
<td>264</td>
</tr>
</tbody>
</table>

Local, global, and "glocal" alignments

• Global alignments go from include the entire length of both sequences (Needleman-Wunsch, 1970)
  – high global similarity = small sequence distance (100% identity = distance 0)
  – similarity scores can be negative
  – scores are (probably) normally distributed
  – single domain, approx. constant length proteins
  – GGSEARCH calculates "global" alignment scores

• Local alignments find the best match, regardless of the length of the match. (Smith-Waterman, 1981)
  – requires similarity scoring matrix with $E(s_i) < 0.0$
  – all similarity scores are $> 0.0$
  – scores are extreme value distributed
  – good for partial sequences, homologous domains with sequences
  – BLASTP, FASTA, and SSEARCH generate "local" alignment scores

• "glocal" alignments are "global" in the query (e.g. a domain), but local in the subject
  – a domain within a protein
  – GLSEARCH
Local, global, and "glocal" alignments

- Local – 26.3% id, $E() < 0.00024$
- Globally similar:
- Global – 20.6% id, $E() < 10^{-7}$
- Locally similar:
- Local – 29.2% id, $E() < 9$
- Global – 15.3% id, $E() < 7000$
- Glocal – 26.8% id, $E() < 0.02$

Dynamic programming for sequence alignment

- Sequence alignments can be *global* – end-to-end, or *local*
- The *Dynamic Programming Algorithm* allows one to examine $2^n$ alignments ($n=100$, $10^{-77}$) in $O(n^2)$ ($n=100$, $O(n^2)=10,000$) time
- Local alignments can also be used to find duplicated domains in proteins
Algorithms for Global and Local Similarity Scores

Global:

\[ S(0,0) = 0 \]

for \( j = 1 \) to \( N \) do

\[ S(0,j) = S(0,j-1) + \alpha(0_i, j) \]

for \( i = 1 \) to \( L \) do

\[ S(i,0) = S(i-1,0) + \alpha(\_i, 0) \]

for \( j = 1 \) to \( N \) do

\[ S(i,j) = \max \{ S(i-1,j-1) + \alpha(0_i, 0_j), S(i-1,j) + \alpha(\_i, 0_j), S(i,j-1) + \alpha(0_i, \_j) \} \]

write “Global similarity score is” \( S(M,N) \)

Local:

\[ \text{best} = 0 \]

for \( i = 1 \) to \( N \) do

\[ \text{best} = 0 \]

for \( i = 1 \) to \( L \) do

\[ \text{best} = 0 \]

for \( j = 1 \) to \( N \) do

\[ \text{best} = \max \{ \text{best}, S(i-1,j-1) + \alpha(0_i, 0_j), S(i-1,j) + \alpha(\_i, 0_j), S(i,j-1) + \alpha(0_i, \_j) \} \]

write “Local similarity score is” best

+1 : match
-1 : mismatch
-2 : gap

align_path2
alignment paths
highlight indels

LALIGN – non-overlapping local alignments can identify mobile domains
Improving Similarity Searching

• What gets missed? / What shouldn't be found
  – comparing sequence and structural similarity
  – what is a "non-homolog"?
• Homology from "significance" – local alignment statistics
  – E()-values and bit-scores
• Use protein databases
  – smaller
  – more sensitive
  – better statistics

How well does BLAST work?

Gold standard – homologous proteins ALWAYS share statistically significant structural similarity
  – databases of structures: SCOP (structural classification of proteins)
  – CATH (Class, Architecture, Topology, Homology)
    • All "Homologs" are "homologous"
    • Some "Topologs" might be homologous
    • Architecture without similar topology, non-homologous
How well are homologs identified?

- Structure comparison:
  - DALI, VAST, MATRAS, CE, STRUCTAL, SGM
- Pairwise sequence comparison:
  - SSEARCH
- Model-based sequence comparison:
  - PSI-BLAST


What is a non-homolog?

Five serine proteases: three trysin like (A, B, C, homologs), subtilisin (E, non-homolog), and ? (D)
Non-homologs have different domains

Improving sensitivity by improving statistical significance

- Local similarity scores follow the "extreme value distribution"
  - unrelated ➔ random, thus:
  - not random ➔ homologous
  - random == extreme value distribution
- improve sensitivity with smaller databases
- can we trust the statistics?
Smaller databases for more sensitive searches
which 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, 450,000)
- Always search NR (> 12 million) LAST
- Never Search “GenBank” (DNA)

Why smaller databases are better – statistics

\[ S' = \lambda S_{\text{raw}} - \ln K \cdot m \cdot n \]
\[ S_{\text{bit}} = (\lambda S_{\text{raw}} - \ln K) / \ln(2) \]
\[ P(S' > x) = 1 - \exp(-e^{-x}) \]
\[ P(S_{\text{bit}} > x) = 1 - \exp(-mn2^{-x}) \]
\[ E(S' > x | D) = P \cdot D \]
\[ P(B \text{ bits}) = m \cdot n \cdot 2^{-B} \]
\[ P(40 \text{ bits}) = 1.5 \times 10^{-7} \]
\[ E(40 | D=4000) = 6 \times 10^{-4} \]
\[ E(40 | D=80E6) = 12 \]
NCBI – selecting sequences with Entrez

Bits and significance

• 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:

$$P(S_{bit} > x) = 1 - \exp(-mn^2-x) \sim mn^2$$

$$400 \times 400 \times 2^{-40} = 1.6 \times 10^5 / 2^{40} (10^{13.3}) = 1.5 \times 10^{-7} \text{ times}$$

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:

$E(S_{bit} | D) = p(40 \text{ bits}) \times \text{database size}$

$E(40 | 4,000) = 10^{-7} \times 4,000 = 4 \times 10^{-4}$ (significant)

$E(40 | 40,000) = 10^{-7} \times 4 \times 10^4 = 4 \times 10^{-3}$ (not 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?

E(p | D) = p(40 bits) x database size

E(40 | 4,000) = $10^8 \times 4,000 = 4 \times 10^8$ (significant)

E(40 | 40,000) = $10^8 \times 4 \times 10^4 = 4 \times 10^4$ (significant)

E(40 | 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 $\sim 10^{-3}/10^4 = 10^{-7}/160,000 \approx 40$ bits
- SwissProt (500,000) p $\sim 10^{-3}/10^6 = 10^{-9}/160,000 \approx 47$ bits
- Uniprot/NR (10^7) p $\sim 10^{-3}/10^7 = 10^{-10}/160,000 \approx 50$ bits

Should you trust the E()-value??
(what is the control for this experiment)

- 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 (ssearch)

The best scores are:

<table>
<thead>
<tr>
<th>Protein</th>
<th>Description</th>
<th>Score</th>
<th>Aln len</th>
<th>E-value</th>
<th>Identity</th>
<th>Query len</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTM1_MOUSE</td>
<td>Glutathione S-trans (218)</td>
<td>1497</td>
<td>363.5</td>
<td>2e-100</td>
<td>1.000</td>
<td>218</td>
</tr>
<tr>
<td>GTM2_CHICK</td>
<td>Glutathione S-trans (210)</td>
<td>958</td>
<td>234.9</td>
<td>1.1e-01</td>
<td>0.619</td>
<td>218</td>
</tr>
<tr>
<td>GTP_HUMAN</td>
<td>Glutathione S-trans (210)</td>
<td>356</td>
<td>91.2</td>
<td>1.0e-18</td>
<td>0.308</td>
<td>211</td>
</tr>
<tr>
<td>PGD2_MOUSE</td>
<td>Glutathione-reg. (199)</td>
<td>262</td>
<td>68.8</td>
<td>9.7e-12</td>
<td>0.319</td>
<td>204</td>
</tr>
<tr>
<td>GTA1_MOUSE</td>
<td>Glutathione S-trans (223)</td>
<td>229</td>
<td>60.9</td>
<td>2.6e-09</td>
<td>0.284</td>
<td>225</td>
</tr>
<tr>
<td>SCL1_OCTDO</td>
<td>S-crystallin 1 (OL1) (215)</td>
<td>228</td>
<td>60.7</td>
<td>3.0e-09</td>
<td>0.269</td>
<td>219</td>
</tr>
<tr>
<td>GT5_MUSDO</td>
<td>Glutathione S-trans (241)</td>
<td>228</td>
<td>60.6</td>
<td>3.4e-09</td>
<td>0.264</td>
<td>201</td>
</tr>
<tr>
<td>GTA1_CAVELE</td>
<td>Prob. Glut. S-trans (210)</td>
<td>220</td>
<td>58.8</td>
<td>1.1e-08</td>
<td>0.284</td>
<td>225</td>
</tr>
<tr>
<td>GTS_OXMSL</td>
<td>Glutathione S-trans (203)</td>
<td>196</td>
<td>53.0</td>
<td>5.5e-07</td>
<td>0.258</td>
<td>209</td>
</tr>
<tr>
<td>GTH1_ARATH</td>
<td>Glutathione S-trans (215)</td>
<td>142</td>
<td>40.1</td>
<td>0.0045</td>
<td>0.310</td>
<td>126</td>
</tr>
<tr>
<td>GTH2_HUMAN</td>
<td>Glutathione S-trans (244)</td>
<td>132</td>
<td>37.7</td>
<td>0.027</td>
<td>0.257</td>
<td>167</td>
</tr>
<tr>
<td>GTH4_DROME</td>
<td>Glutathione S-trans (216)</td>
<td>131</td>
<td>37.5</td>
<td>0.028</td>
<td>0.255</td>
<td>153</td>
</tr>
<tr>
<td>YFCG_ECOLI</td>
<td>Hypothetical GST (215)</td>
<td>112</td>
<td>33.0</td>
<td>0.643</td>
<td>0.235</td>
<td>187</td>
</tr>
<tr>
<td>YJY1_YEAST</td>
<td>Hypothetical 30.5 (261)</td>
<td>110</td>
<td>32.4</td>
<td>1.1e-01</td>
<td>0.248</td>
<td>149</td>
</tr>
<tr>
<td>DCMA_METS1</td>
<td>Dichloromethane DM (267)</td>
<td>103</td>
<td>30.8</td>
<td>0.214</td>
<td>0.214</td>
<td>210</td>
</tr>
<tr>
<td>YA2_HAEIN</td>
<td>Hypothetical prot. (617)</td>
<td>108</td>
<td>31.7</td>
<td>4.6e-07</td>
<td>0.283</td>
<td>120</td>
</tr>
<tr>
<td>GTO1_RAT</td>
<td>Glutathione trans (241)</td>
<td>100</td>
<td>30.1</td>
<td>0.234</td>
<td>0.158</td>
<td>158</td>
</tr>
<tr>
<td>DP41_RABC</td>
<td>DNA polymerase I (413)</td>
<td>104</td>
<td>30.8</td>
<td>5.4e-04</td>
<td>0.234</td>
<td>184</td>
</tr>
<tr>
<td>GTH1_WHEAT</td>
<td>Glutathione S-trans (229)</td>
<td>98</td>
<td>29.6</td>
<td>0.246</td>
<td>0.246</td>
<td>171</td>
</tr>
<tr>
<td>LGUL_SOYBN</td>
<td>Lactoylglutathione (219)</td>
<td>97</td>
<td>29.4</td>
<td>0.200</td>
<td>0.200</td>
<td>190</td>
</tr>
<tr>
<td>VP2_AHSV3</td>
<td>Outer capsid prot. (1057)</td>
<td>108</td>
<td>31.5</td>
<td>8.9e-07</td>
<td>0.205</td>
<td>200</td>
</tr>
<tr>
<td>GTH5_ARATH</td>
<td>Glutathione S-trans (218)</td>
<td>96</td>
<td>29.2</td>
<td>0.258</td>
<td>0.258</td>
<td>166</td>
</tr>
<tr>
<td>DCMA_METS2</td>
<td>Dichloromethane DM (288)</td>
<td>98</td>
<td>29.5</td>
<td>0.195</td>
<td>0.195</td>
<td>200</td>
</tr>
<tr>
<td>GTXA_ARATH</td>
<td>Glutathione S-trans (224)</td>
<td>96</td>
<td>29.1</td>
<td>0.248</td>
<td>0.248</td>
<td>125</td>
</tr>
<tr>
<td>SLT_HAEIN</td>
<td>Putative soluble 1 (593)</td>
<td>103</td>
<td>30.5</td>
<td>9.9e-07</td>
<td>0.227</td>
<td>185</td>
</tr>
</tbody>
</table>

Breaking the statistics: low complexity regions

Search with complete group_drome:

The best scores are:

<table>
<thead>
<tr>
<th>Protein</th>
<th>Description</th>
<th>Score</th>
<th>Aln len</th>
<th>E-value</th>
<th>Identity</th>
<th>Query len</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTHUB1 GTP</td>
<td>GTP-binding regulatory protein</td>
<td>(341)</td>
<td>237</td>
<td>46.6</td>
<td>3.5e-05</td>
<td></td>
</tr>
<tr>
<td>RGBO3 GTP</td>
<td>GTP-binding regulatory protein</td>
<td>(341)</td>
<td>237</td>
<td>46.6</td>
<td>3.5e-05</td>
<td></td>
</tr>
<tr>
<td>GTHUB4 GTP</td>
<td>GTP-binding regulatory protein</td>
<td>(341)</td>
<td>233</td>
<td>46.0</td>
<td>5.2e-05</td>
<td></td>
</tr>
<tr>
<td>RGHB3 GTP</td>
<td>GTP-binding regulatory protein</td>
<td>(341)</td>
<td>232</td>
<td>45.8</td>
<td>5.7e-05</td>
<td></td>
</tr>
<tr>
<td>PINHUFF salivary</td>
<td>proline-rich glycoprotein</td>
<td>(252)</td>
<td>224</td>
<td>44.5</td>
<td>0.00010*</td>
<td></td>
</tr>
<tr>
<td>RGFFB GTP</td>
<td>GTP-binding regulatory protein</td>
<td>(347)</td>
<td>223</td>
<td>44.5</td>
<td>0.00014</td>
<td></td>
</tr>
<tr>
<td>PIR3T acidic</td>
<td>proline-rich protein precursor</td>
<td>(207)</td>
<td>199</td>
<td>40.8</td>
<td>0.0011*</td>
<td></td>
</tr>
<tr>
<td>PINHUB salivary</td>
<td>proline-rich protein precursor</td>
<td>(393)</td>
<td>203</td>
<td>41.6</td>
<td>0.0012*</td>
<td></td>
</tr>
<tr>
<td>COBRO28 collagen alpha 2(1) chain</td>
<td>bovine (fragme</td>
<td>(403)</td>
<td>195</td>
<td>40.5</td>
<td>0.0027*</td>
<td></td>
</tr>
<tr>
<td>WNB25 capsid protein human herpesvirus 1 (strA</td>
<td>(636)</td>
<td>192</td>
<td>40.2</td>
<td>0.0051*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>W4NLB5 E4 protein human papillomavirus type 5b</td>
<td>(246)</td>
<td>170</td>
<td>36.6</td>
<td>0.024*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OZQWQF circumsoralzote protein precursor Plasma</td>
<td>(368)</td>
<td>172</td>
<td>37.1</td>
<td>0.026*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>POMMEE gag polypeptide murine leukemia virus</td>
<td>(537)</td>
<td>161</td>
<td>35.6</td>
<td>0.010*</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Search with seg-ed group_drome: (low complexity regions removed)

The best scores are:

<table>
<thead>
<tr>
<th>Protein</th>
<th>Description</th>
<th>Score</th>
<th>Aln len</th>
<th>E-value</th>
<th>Identity</th>
<th>Query len</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTHUB3 GTP</td>
<td>GTP-binding regulatory protein</td>
<td>(341)</td>
<td>233</td>
<td>56.5</td>
<td>3.6e-08</td>
<td></td>
</tr>
<tr>
<td>GTHUB4 GTP</td>
<td>GTP-binding regulatory protein</td>
<td>(341)</td>
<td>232</td>
<td>56.3</td>
<td>4.1e-08</td>
<td></td>
</tr>
<tr>
<td>RGHB2 GTP</td>
<td>GTP-binding regulatory protein</td>
<td>(341)</td>
<td>232</td>
<td>55.5</td>
<td>7.2e-08</td>
<td></td>
</tr>
<tr>
<td>RGBO1 GTP</td>
<td>GTP-binding regulatory protein</td>
<td>(341)</td>
<td>228</td>
<td>54.9</td>
<td>1.1e-07</td>
<td></td>
</tr>
<tr>
<td>RGFFB GTP</td>
<td>GTP-binding regulatory protein</td>
<td>(347)</td>
<td>223</td>
<td>54.5</td>
<td>1.5e-07</td>
<td></td>
</tr>
<tr>
<td>BVYNSH MS1 protein yeast Saccharomyces cerevis</td>
<td>(423)</td>
<td>135</td>
<td>37.0</td>
<td>0.033*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ERHUAH coatomer complex alpha chain homolog human</td>
<td>(1225)</td>
<td>134</td>
<td>37.1</td>
<td>0.088*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A28468 chromogranin A precursor human</td>
<td>(458)</td>
<td>122</td>
<td>34.4</td>
<td>0.21*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RGOOBE GTP</td>
<td>GTP-binding regulatory protein</td>
<td>(342)</td>
<td>120</td>
<td>33.9</td>
<td>0.22</td>
<td></td>
</tr>
</tbody>
</table>
pseg removes low-complexity regions

>gi|17380405|sp|P16371|GROU_DROME Groucho protein (Enhancer of split M9/10)

1-8    MYPSPVRK
9-19   IKTFT1DLERIKEEPNFQAOYHSIXKLEC
20-131 EXLSNQETNYIQMVYHNYEMSFQLGHVEEN
QTEIARKHLNYIHLFFDQHTSSQQVQDA
VERAKQVTMLNLIQGQHA

132-143
144-281
ALNPPQALGATMLPQPCQCLNEXPEHHR
P1KFGTQLEGPAAEERLRSMSGSPEASEKY
MTSFSLDIEUDSHKREKELQEDRSQDSQ
DLVVDGOAENGSHSPPHGHEYVYQRE
SLNGERLEQPSGSGIKQE

282-297
298-310
311-330
331-351
352-719
DPYQRPPSDPAAYGRPPMPYDHRXVFSTNG
IPHSALDQGDPAYSFHMKEEGLQVPFP
PGAVVCOIPRHRQINTLSCAVCVCVTI
SNFTKYVTQGCVKVIDISQGKXNPVS
QLCSLQDHYIRSLPQGRTLIGCEAS
NLSSKEASPTPPKELTSAAAPACTALAI
SPESYCFDSCCDSHDIAVLINEILRQG
QCHTDASCIDISPPQERLVIALTLDFTVRS
WDLRMROQQLQHDDSQSFSLCYCPTDQWN
AVGCHSHVVLHLSSPQKDLICLLEHSCIY
SLRFAACOXQPVTOCKHHLXAMTPPGAS
IFPSRETSSVL6CD5TDDKIVTDQGKX
ATYVEVTV

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/SSEARCH 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

```bash
>lwec6_H+-transporting ATP synthase (EC 3.6.1.34) protein 6 - Escherichia coli
YGHHLNQLQ LDLRTFSLVD PQHPPTFNT IKIQHFFSV VLGLLFLVF

>lwec6_0 shuffled
GDPSCVGLFKK PPKWAVRLLF YSVYIFEFPAI VEPYVONVRFV AVAGHAIYK

>lwec6 shuffled window: 10
GMPISVLLFK PPEVLLVFLL SVMGTNFPAW GGFIMKGFKI VSFVGWVRFV AVAGHAIYK
```

PRSS34 - 1000 shuffles; uniform shuffle
unshuffled s-w score: 178; bits(s=178|n_l=271): 34.8 p(178) < 2.005e-06
For 10000 sequences, a score >= 178 is expected 0.02005 times

PRSS34 - 1000 shuffles; window shuffle, window size: 20
unshuffled s-w score: 178; bits(s=178|n_l=271): 34.5 p(178) < 2.602e-06
For 10000 sequences, a score >= 178 is expected 0.02602 times
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)

Scoring matrices

- Scoring matrices can set the evolutionary look-back time for a search
  - Lower PAM (PAM10/VT10 ... PAM/VT40) 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>A</th>
<th>R</th>
<th>N</th>
<th>D</th>
<th>E</th>
<th>I</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>-9</td>
<td>-4</td>
<td>-3</td>
<td>-6</td>
<td>-8</td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>12</td>
<td>11</td>
<td>11</td>
<td>12</td>
<td>-10</td>
<td>-16</td>
</tr>
<tr>
<td>N</td>
<td>-7</td>
<td>3</td>
<td>4</td>
<td>-7</td>
<td>-9</td>
<td>-16</td>
</tr>
<tr>
<td>D</td>
<td>11</td>
<td>11</td>
<td>4</td>
<td>-4</td>
<td>-9</td>
<td>-12</td>
</tr>
<tr>
<td>E</td>
<td>11</td>
<td>11</td>
<td>2</td>
<td>-2</td>
<td>2</td>
<td>-1</td>
</tr>
<tr>
<td>I</td>
<td>12</td>
<td>10</td>
<td>4</td>
<td>-7</td>
<td>3</td>
<td>-11</td>
</tr>
<tr>
<td>L</td>
<td>-10</td>
<td>-12</td>
<td>-4</td>
<td>11</td>
<td>-11</td>
<td>-16</td>
</tr>
</tbody>
</table>

Pam40:

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

Pam250:

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

$q_{ij}$: replacement frequency at PAM40, 250

$q_{RN(40)} = 0.000435 \quad p_{R} = 0.051$

$q_{RN(250)} = 0.002193 \quad p_{N} = 0.043$

$l_S = \log_2 \left( \frac{q_{ij}}{p_i p_j} \right)$

$S_{RN(40)} = \log_2 \left( \frac{q_{ij}}{p_i p_j} \right)$

$S_{RN(40)} = \log_2 \left( \frac{0.000435}{0.002193} \right) = -2.333$

$S_{RN(250)} = \log_2 \left( \frac{0.002193}{0.002193} \right) = 0$

Scoring matrices set look back time:
Glutathione Transferases (gstm1_human)

fasta.bioch.virginia.edu/biol4230
### PAM matrices and alignment length

Short domains require “shallow” scoring matrices


`fasta.bioch.virginia.edu`
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"

Pearson (2013) Curr Protoc. Bioinfo 3.5.1-3.5.9

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
- Shallow matrices set maximum look-back time
- Short alignments (domains, exons, reads) require shallow (higher information content) matrices
Similarity searching II – algorithms, statistics, and scoring matrices

- Global and local alignments
  - Global alignments can be more sensitive for globally similar proteins
  - Local alignments are robust to partial sequences, domain homologies
- Local similarity scores are well described by the extreme value distribution
  - $E()$-value depends on similarity score AND database size
  - A 50 bit score is almost always significant
  - $E()$-values are not good measures of evolutionary distance
- Scoring matrices can be designed for long (deep) or short (shallow) evolutionary distances (large/small amounts of change)
  - "shallow" matrices provide more statistical significance for each aligned position, but require higher homologs
  - "deep" matrices can find more distant homologs, but require longer alignments