Representing Protein Domains with PSSMs and HMMs
Biol4230 Tues, February 9, 2016
Bill Pearson wrp@virginia.edu 4-2818 Jordan 6-057

Goals of today’s lecture:
• understand types of domain definitions – folding units, evolutionary (mobile) units; domains vs motifs
• familiarity with InterPro, a "meta"-database of domain databases, and Pfam
• Where do pairwise scoring matrices come from? – the math
• Where do position specific scoring matrices (PSSMs) come from – PSI-BLAST
• What mistakes do Iterative methods (PSI-BLAST) make?

To learn more:
• Domains and InterPro – Pevzner, Part II, Ch. 10
• Scoring Matrices – Pevzner, Part I, Ch. 3
• PSSMs and PSI-BLAST – Pevzner, Part I, Ch. 5, p. 145
• Pick a protein of interest (serine protease, glutathione transferase, your favorite kinase, phosphatase, G-protein)
• Find the protein in interpro. Do the different domain databases find the same domains in the same places?
  – Compare your protein to SwissProt using PSI-BLAST
    • after 3 iterations, look at the domain structure of the five lowest scoring significant (E(<0.001) hits.
    • Are they all homologous (do they have the same domains)?
  – Find the protein in Pfam. What domains are found in the protein?
Finding domains with domain models I: from scoring matrices to PSSMs

- Domains are structurally compact, evolutionarily mobile, protein building blocks
  - atomic, they have a characteristic length
  - often repeated, or found in different sequence contexts
  - essential for building detection systems (PSSMs, HMMs), because they focus on the homologous region (a full length protein can be a mixture of domains)
  - Interpro provides large-scale summary
  - Pfam most comprehensive single resource
- Position independent scoring matrices can be built from a simple evolutionary model: $PAM^{(n)} = PAM(n)$
- Position Specific Scoring Matrices (PSSMs) generalize frequency data for a single position
- PSI-BLAST increases sensitivity with PSSMs

Representing Protein Domains

- Protein domains can be defined structurally, functionally, or based on evolutionary mobility
  - Mobile domains can be identified by duplication: mobile within protein (calmodulin), and alignment context: mobile among proteins
- Multiple-sequence based protein models (PSSMs, HMMs) are extend pair-wise scoring methods to sites on a protein model
- PSI-BLAST and HMMER build sensitive domain models
  - Position-Specific-Scoring Matrix (PSSM) from multiple sequence alignment
- InterPro provides integrated access to most domain annotations on a protein
- PFAM is a high-quality (curated) domain database
- ALL model/domain/sequence methods miss homologs
  - positives are correct, but negatives more ambiguous
Defining Domains

Domain definitions:
- Structural
  - (molecular scissors)
- Evolutionary
  - (alternate contexts)

Often repetitive

Sequence-based protein domains are evolutionarily mobile

protein (sequence)
>ref|NP_000552.2| GSTM1 (human)
MPNGLQYIGLGAAILLLDTSVSTSEKTN
CDAEVGQNLKQFGLGTPFSDLQGAKV1
7Q8HA1LYCTAABBHCGZEEZKIKVVELEQYN
DNNKLQGNCYQVPEPGLSFLALRDSQYE
FLKKGFWAGWVTTFDLYTLDWGSRFPFRCL
DAPWHKLRNSREGLKKSAYKNEKRLPFPVTS

protein (structure, 1XW6)

fasta.bioch.virginia.edu/biol4230
(Evolutionary) Domains vs complete proteins

- Many proteins are made up of multiple domains – structural/sequence units that evolve independently and may fold independently
- For multi-domain proteins, it is the domain, not the protein, that is the “atomic” unit of homology
- For multi-domain proteins, a significant similarity (homology) may apply only to one domain
- Domains are common, >40% of proteins contain more than one domain
- Unlike complete proteins, which have a beginning and end, domain boundaries can be more difficult to determine

Domain abundance (Pfam 27, 2014)
### Identifying mobile domains:

**mobile (duplicated) domains in local alignments**

<table>
<thead>
<tr>
<th>s-w bits</th>
<th>E(1)</th>
<th>%_id</th>
<th>%_sim</th>
<th>alen</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kat2b_Human Histone acetyltransferase Kat2b</td>
<td>3820</td>
<td>1456.0</td>
<td>0.100</td>
<td>1.000</td>
</tr>
<tr>
<td>Kat2a_Human Histone acetyltransferase Kat2a</td>
<td>2747</td>
<td>1049.0</td>
<td>0.721</td>
<td>0.870</td>
</tr>
<tr>
<td>GCN5_SCHPO Histone acetyltransferase gcn5</td>
<td>867</td>
<td>334.7</td>
<td>3e-90</td>
<td>0.483</td>
</tr>
<tr>
<td>GCN5_YEAST Histone acetyltransferase GCN5</td>
<td>792</td>
<td>306.2</td>
<td>1.1e-81</td>
<td>0.469</td>
</tr>
<tr>
<td>GCN5_ORYSJ Histone acetyltransferase GCN5</td>
<td>719</td>
<td>278.4</td>
<td>3.3e-73</td>
<td>0.436</td>
</tr>
<tr>
<td>GCN5_ARATH Histone acetyltransferase GCN5;</td>
<td>286</td>
<td>113.6</td>
<td>7.6e-23</td>
<td>0.495</td>
</tr>
<tr>
<td>NstF_HUMAN Nucleosome-remodeling factor sub</td>
<td>276</td>
<td>109.8</td>
<td>9.1e-17</td>
<td>0.371</td>
</tr>
<tr>
<td>CECR2_HUMAN Cat eye syndrome critical regio (1484)</td>
<td>232</td>
<td>93.2</td>
<td>5e-17</td>
<td>0.371</td>
</tr>
<tr>
<td>BRD4_HUMAN Bromodomain-containing protein 4 (1362)</td>
<td>214</td>
<td>86.4</td>
<td>5.2e-15</td>
<td>0.379</td>
</tr>
<tr>
<td>BBR4_MOUSE Bromodomain-containing protein 4 (1400)</td>
<td>214</td>
<td>86.4</td>
<td>5.3e-15</td>
<td>0.379</td>
</tr>
<tr>
<td>BAZ2A_HUMAN Bromodomain adjacent to zinc f1 (1905)</td>
<td>211</td>
<td>85.2</td>
<td>1.7e-14</td>
<td>0.382</td>
</tr>
<tr>
<td>BAZ2A_XENLA Bromodomain adjacent to zinc f1 (1698)</td>
<td>206</td>
<td>83.3</td>
<td>5.5e-14</td>
<td>0.350</td>
</tr>
<tr>
<td>FSH_DROME Homeotic protein female sterile;  (2038)</td>
<td>205</td>
<td>82.9</td>
<td>8.8e-14</td>
<td>0.341</td>
</tr>
<tr>
<td>BAZ2A_MOUSE Bromodomain adjacent to zinc f1 (1889)</td>
<td>204</td>
<td>82.5</td>
<td>1e-13</td>
<td>0.368</td>
</tr>
<tr>
<td>BRD7_XACFA Bromodomain testis-specific prot (947)</td>
<td>197</td>
<td>80.0</td>
<td>3e-13</td>
<td>0.367</td>
</tr>
<tr>
<td>BRD3_HUMAN Bromodomain-containing protein 3 (726)</td>
<td>194</td>
<td>78.9</td>
<td>4.9e-13</td>
<td>0.362</td>
</tr>
</tbody>
</table>
Homology and Domains – Histone acetyltransferase KAT2B

<table>
<thead>
<tr>
<th>Protein</th>
<th>Description</th>
<th>E-value</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>KAT2B_HUMAN</td>
<td>Histone acetyltransferase</td>
<td>E()&lt;0</td>
<td>832</td>
</tr>
<tr>
<td>KAT2A_HUMAN</td>
<td>Histone acetyltransferase</td>
<td>E()&lt;0</td>
<td>813</td>
</tr>
<tr>
<td>GCN5_YEAST</td>
<td>Histone acetyltransferase</td>
<td>1.1e-81</td>
<td>354</td>
</tr>
<tr>
<td>GCN5_ARATH</td>
<td>Histone acetyltransferase</td>
<td>3.3e-73</td>
<td>369</td>
</tr>
<tr>
<td>BPTF_HUMAN</td>
<td>Nucleosome-remodeling factor</td>
<td>7.6e-23</td>
<td>97</td>
</tr>
</tbody>
</table>

Identifying mobile domains

Like homologous proteins, homologous domains share statistically significant structural or sequence similarity

• Many domain family members share significant sequence similarity (BLAST), and produce partial sequence alignments
• Internally repeated domains can be identified with lalign
  – Domain boundaries may depend on the scoring matrix
• To find all (or most) domain family members, more sensitive methods are used:
  – PSSMs (Position Specific Scoring Matrices) PSI-BLAST, RPS-BLAST
  – HMMs (Hidden Markov Models) HMMER3 (Pfam)
Protein Motif and Domain Databases

RNA sequence databases
Protein sequence databases
   General sequence databases
   Protein properties
   Protein localization and targeting
   Protein sequence motifs and active sites
      ASC - Active Sequence Collection
      Blocks
      COMe - Co-Ordination
      CSA - Catalytic Site Atlas
      eF-site - Electrostatic surface
      eMOTIF
      InterPro
      Metalloprotein Database and Browser
      O-GLYCBASE
      PhosphoBase
      PRINTS
      PROMISE
      PROSITE

Protein domain databases; protein classification
   BAliBASE
   CDD
   CluSTr - Clusters of Swiss-Prot and TrEMBL
   COG - Clusters of Orthologous Groups
   DomIns - Database of Domain Insertions
   FusionDB
   Hits
   HSSP
   InterDom
   InterPro
   iProClass
   MetaFam
   PALI
   Pfam
   PIR-ALN
   PIRSF
   ProClass
   ProDom
   ProtoMap
   ProtoNet
   SBASE
   SMART
   SUPFAM
   TIGRFAMs

InterPro, PFAM, and Prosite

InterPro – The database of Protein databases  [www.ebi.ac.uk/interpro](http://www.ebi.ac.uk/interpro)

PFAM – a “domain” database  [pfam.sanger.ac.uk](http://pfam.sanger.ac.uk)
   • Complete domain alignments. Definition of domains.
   • Example of searching PFAM on-line; what scores mean.
   • Caveats: structural rather than functional classification

   • Patterns and regular expressions
   • The information content of a PROSITE pattern
   • Examples of searching PROSITE on-line
   • Caveats: missing patterns; low-information patterns

Always do control experiments: never trust a server
   • Positive controls -- submit sequences for which you know the right answer.
   • Negative controls -- random or shuffled sequences.
Representations of domains

- Consensus sequences (finger prints) – aligned unweighted PSSMs without gaps – PRINTS
- Regular expressions – exact match to regular expression (good for absolutely conserved motifs, active sites) – ProSite patterns
- HMM/PSSM/Profile (Hidden Markov Model/Position Specific Scoring matrix/Profile) – HMM most flexible, provides statistical significance estimates
  - Pfam, Tigrfam, SuperFamily, Panther, ProSite profiles, HaMap profiles

InterPro analysis of GSTM1_HUMAN
Pfam domains on GSTM1_HUMAN

**Protein: GSTM1_HUMAN (P09488)**

<table>
<thead>
<tr>
<th>Protein</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSTM1_HUMAN</td>
<td>This is the summary of UniProt entry GSTM1_HUMAN (P09488).</td>
</tr>
</tbody>
</table>

**Description:** Glutathione S-transferase Mu 1 EC:2.5.1.18

**Source organism:** Homo sapiens (Human) (NCBI taxonomy ID: 9606)

**Pfam data**

- **Length:** 218 amino acids

**Please note:** When we start each new Pfam data release, we take a copy of the UniProt sequence database. This snapshot of UniProt forms the basis of the overview that you see here. It is important to note that, although some Pfam entries may be removed after a Pfam release, these entries will not be removed from UniProt until the next UniProt database release.

**Pfam domains**

This image shows the arrangement of the Pfam domains that we found on this sequence. Clicking on a domain will take you to the page describing that Pfam entry. The table below gives the domain boundaries for each of the domains.

<table>
<thead>
<tr>
<th>Source</th>
<th>Domain</th>
<th>Start</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pfam A</td>
<td>GST_A</td>
<td>3</td>
<td>82</td>
</tr>
<tr>
<td>Pfam A</td>
<td>GST_C</td>
<td>104</td>
<td>192</td>
</tr>
<tr>
<td>low complexity</td>
<td></td>
<td>118</td>
<td>137</td>
</tr>
</tbody>
</table>
Pfam domain descriptions – GST_N

Family: GST_N (PF02798)

Summary
Domain organisation
Alignments
HMM logo
Trees
Curation & models
Species
Interactions
Structures
Jump to...

Glutathione S-transferase, N-terminal domain

Function: conjugation of reduced glutathione to a variety of targets. Also included in the alignment, but are not GSTs: * S-crystallins from squid. Similarly to GST previously noted. * Eukaryotic elongation factors 1-gamma. Not known to have GST activity; similarity not previously recognised. * HSP70 family of stress-related proteins, including auxin-regulated proteins in plants and stringent starvation proteins in S. cerevisiae. Not known to have GST activity. Similarity not previously recognised. The glutathione molecule binds in a cleft between N- and C-terminal domains: the catalytically important residues are proposed to reside in the N-terminal domain.

Literature references

Pfam GST_N architectures

Family: GST_N (PF02798)

Summary
Domain organisation
Alignments
HMM logo
Trees
Curation & models
Species
Interactions
Structures
Jump to...

There are 6556 sequences with the following architecture: GST_N, GST_C, DMT_M, GST (Pheleophthia sp. [phore DMI]) Dihydroneurothele demethylase EC=4.5.1.3 (397 residues)

There are 1741 sequences with the following architecture: GST_N, GST_C, D71G, FSH2 (GAM) [Achromobacter thelicola (mesa-aer crease)] Probable elongation factor-1-gamma 1 (454 residues)

There are 1254 sequences with the following architecture: GST_N, GST_C, ERNA-antl, Anticodon, 1

There are 11 sequences with the following architecture: FLYWCH x 4, GST_N, GST_C

There are 4 sequences with the following architecture: GST_N, GST_C, IRNA-antl, 1

There are 4 sequences with the following architecture: GST_N, GST_C, DMT_M, GST (Pheleophthia sp. [phore DMI]) Dihydroneurothele demethylase (100 residues)

There are 4 sequences with the following architecture: GST_N, GST_C, ERNA-antl, 1

There are 4 sequences with the following architecture: GST_N, GST_C, IRNA-antl, 1

There are 4 sequences with the following architecture: GST_N, GST_C, DMT_M, GST (Pheleophthia sp. [phore DMI]) Dihydroneurothele demethylase EC=4.5.1.3 (397 residues)

fasbio.ac.virginia.edu/biol4230
Improving search sensitivity with Protein family models (PSSMs and HMMs)

• Where do scoring matrices come from
  – Transition probabilities and PAMs
  – Scoring matrices as log-odds values (log(p[related]/p[chance]))
• From non-position-specific (PAM250, BLOSUM62) to position-specific – PSI-BLAST

DNA transition probabilities – 1 PAM

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>c</th>
<th>g</th>
<th>t</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0.99</td>
<td>0.001</td>
<td>0.008</td>
<td>0.001</td>
<td>= 1.0</td>
</tr>
<tr>
<td>c</td>
<td>0.001</td>
<td>0.99</td>
<td>0.001</td>
<td>0.008</td>
<td>= 1.0</td>
</tr>
<tr>
<td>g</td>
<td>0.008</td>
<td>0.001</td>
<td>0.99</td>
<td>0.001</td>
<td>= 1.0</td>
</tr>
<tr>
<td>t</td>
<td>0.001</td>
<td>0.008</td>
<td>0.001</td>
<td>0.99</td>
<td>= 1.0</td>
</tr>
</tbody>
</table>
### Matrix multiples

\( M^2 = \begin{pmatrix} 0.980 & 0.002 & 0.016 & 0.002 \\ 0.002 & 0.980 & 0.002 & 0.016 \\ 0.016 & 0.002 & 0.980 & 0.002 \\ 0.002 & 0.016 & 0.002 & 0.980 \end{pmatrix} \)

\( M^5 = \begin{pmatrix} 0.952 & 0.005 & 0.038 & 0.005 \\ 0.005 & 0.951 & 0.005 & 0.038 \\ 0.038 & 0.005 & 0.952 & 0.005 \\ 0.005 & 0.038 & 0.005 & 0.952 \end{pmatrix} \)

\( M^{10} = \begin{pmatrix} 0.907 & 0.010 & 0.073 & 0.010 \\ 0.010 & 0.907 & 0.010 & 0.073 \\ 0.073 & 0.010 & 0.907 & 0.010 \\ 0.010 & 0.073 & 0.010 & 0.907 \end{pmatrix} \)

\( M^{100} = \begin{pmatrix} 0.499 & 0.083 & 0.336 & 0.083 \\ 0.083 & 0.499 & 0.083 & 0.336 \\ 0.336 & 0.083 & 0.499 & 0.083 \\ 0.083 & 0.336 & 0.083 & 0.499 \end{pmatrix} \)

\( M^{1000} = \begin{pmatrix} 0.255 & 0.245 & 0.255 & 0.245 \\ 0.245 & 0.255 & 0.245 & 0.255 \\ 0.255 & 0.245 & 0.255 & 0.245 \\ 0.245 & 0.255 & 0.245 & 0.255 \end{pmatrix} \)

Where do scoring matrices come from?

\[ \lambda_S = \log \left( \frac{q_{ij}}{p_j} \right) \]

\( q_{ij} = M^{20} \)  PAM20
\( \{0.828, 0.019, 0.133, 0.019\}, \)
\( \{0.019, 0.828, 0.019, 0.133\}, \)
\( \{0.133, 0.019, 0.828, 0.019\}, \)
\( \{0.019, 0.133, 0.019, 0.828\} \)

\( p_j \) (a,c,g,t) = \( \frac{1}{4} \)

\( \lambda_S = 10 \log \left( \frac{0.828}{0.25} \right) = 5.2 \)

\( \lambda_S = 10 \log \left( \frac{0.019}{0.25} \right) = -11.2 \)

\( \lambda_S = \frac{\log(2)}{10} = 0.33 \)
Two expressions for $S_{ij}$

**Transition frequency** (probability)
- Durbin et al.

$$\lambda S = \log \left( \frac{q_{ij}}{p_j} \right)$$

**Alignment frequency** (probability)
- Altschul

$$\lambda S = \log \left( \frac{q_{ij}^a}{p_ip_j} \right)$$

Altschul $q_{ij}^a = p_i \times$ Durbin $q_{ij}^i$

$$\lambda S = \log \left( \frac{q_{ij}^a = p_i q_{ij}^i}{p_ip_j} \right)$$

---

Scoring matrices at DNA PAMs - ratios

**blastn (DNA)**

PAM1={
  ratio=1/3.13=+1/-3 H=1.90
  (1.99, -6.23, -6.23, -6.22),
  (6.23, 1.99, -6.23, -6.23),
  (-6.23, -6.23, 1.99, -6.23),
  (-6.23, -6.23, -6.23, 1.99)}

PAM2={
  ratio=1/2.65=+2/-5 H=1.82
  (1.97, -5.24, -5.24, -5.24),
  (-5.24, 1.98, -5.24, -5.24),
  (-5.24, -5.24, 1.98, -5.24),
  (-5.24, -5.24, -5.24, 1.98)}

PAM30={
  ratio=1/1.21=+4/-5 H=1.05
  (1.72, -2.09, -2.09, -2.09),
  (-2.09, 1.72, -2.09, -2.09),
  (-2.09, -2.09, 1.72, -2.09),
  (-2.09, -2.09, -2.09, 1.72)}

fasta (DNA)

PAM10={
  ratio=1/1.61=+2/-3 H=1.40
  (1.86, -3.00, -3.00, -3.00),
  (-3.00, 1.86, -3.00, -3.00),
  (-3.00, -3.00, 1.86, -3.00),
  (-3.00, -3.00, -3.00, 1.86)}

PAM45={
  ratio=1.23/1+=5/-4 H=0.54
  (1.40, -1.14, -1.40, -1.14),
  (-1.14, 1.40, -1.14, -1.14),
  (-1.14, -1.14, 1.40, -1.14),
  (-1.14, -1.14, -1.14, 1.40)}

---

fasta.bioch.virginia.edu/biol4230
Where do scoring matrices come from?

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>8</td>
<td>-9</td>
<td>12</td>
<td>-4</td>
<td>-7</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>-4</td>
<td>-13</td>
<td>3</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>-3</td>
<td>-11</td>
<td>-2</td>
<td>4</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>I</td>
<td>L</td>
<td></td>
<td></td>
<td></td>
<td></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>6</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>0</td>
<td>-1</td>
<td>2</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>E</td>
<td>I</td>
<td>L</td>
<td></td>
<td></td>
<td></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 $p_N = 0.043$

$l_2 S_{ij} = \log_2 \left( \frac{q_{ij}}{p_i p_j} \right)$ \quad $l_2 S_{RN (40)} = 0.000435/0.00219 = -2.333$

$l_2 S_{RN (250)} = 0.002193/0.00219 = 0$

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

Shallow matrices reduce evolutionary look-back
Glutathione Transferases (gstm1_human)
### Shallow matrices reduce evolutionary look-back

**Glutathione Transferases (gstm1_human)**

<table>
<thead>
<tr>
<th>Name</th>
<th>Species</th>
<th>E(-50K) %</th>
<th>E(-50K) %</th>
<th>E(-50K) %</th>
<th>E(-50K) %</th>
<th>E(-50K) %</th>
<th>E(-50K) %</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSTM1_HUMAN</td>
<td></td>
<td>8e-92</td>
<td>100.0</td>
<td>1e-116</td>
<td>100.0</td>
<td>1e-173</td>
<td>100.0</td>
</tr>
<tr>
<td>GSTM2_MOUSE</td>
<td></td>
<td>1e-88</td>
<td>83.9</td>
<td>5e-113</td>
<td>83.9</td>
<td>5e-163</td>
<td>84.7</td>
</tr>
<tr>
<td>GSTM2_HUMAN</td>
<td></td>
<td>3e-88</td>
<td>84.4</td>
<td>1e-112</td>
<td>84.4</td>
<td>5e-162</td>
<td>84.4</td>
</tr>
<tr>
<td>GSTM4_HUMAN</td>
<td></td>
<td>2e-86</td>
<td>86.7</td>
<td>8e-107</td>
<td>86.7</td>
<td>2e-155</td>
<td>86.7</td>
</tr>
<tr>
<td>GSTM1_MOUSE</td>
<td></td>
<td>3e-84</td>
<td>78.0</td>
<td>5e-107</td>
<td>78.0</td>
<td>2e-153</td>
<td>78.0</td>
</tr>
<tr>
<td>GSTM3_HUMAN</td>
<td></td>
<td>2e-77</td>
<td>87.6</td>
<td>6e-99</td>
<td>87.6</td>
<td>2e-144</td>
<td>87.6</td>
</tr>
<tr>
<td>GSTM4_MOUSE</td>
<td></td>
<td>4e-74</td>
<td>72.7</td>
<td>6e-94</td>
<td>72.7</td>
<td>2e-132</td>
<td>72.7</td>
</tr>
<tr>
<td>GSTM3_MOUSE</td>
<td></td>
<td>5e-74</td>
<td>73.2</td>
<td>8e-95</td>
<td>73.2</td>
<td>2e-132</td>
<td>72.7</td>
</tr>
<tr>
<td>GSTM2_DERPT</td>
<td></td>
<td>3e-44</td>
<td>49.5</td>
<td>5e-53</td>
<td>49.5</td>
<td>5e-71</td>
<td>50.0</td>
</tr>
<tr>
<td>GSTM1_DERPT</td>
<td></td>
<td>1e-42</td>
<td>46.7</td>
<td>2e-53</td>
<td>46.7</td>
<td>5e-47</td>
<td>49.5</td>
</tr>
<tr>
<td>GSTM2_SCHJA</td>
<td></td>
<td>1e-31</td>
<td>43.5</td>
<td>6e-40</td>
<td>42.6</td>
<td>2e-46</td>
<td>43.5</td>
</tr>
<tr>
<td>GSTM1_FASHE</td>
<td></td>
<td>3e-17</td>
<td>28.9</td>
<td>2e-29</td>
<td>28.9</td>
<td>2e-13</td>
<td>30.1</td>
</tr>
<tr>
<td>GSTM2_HUMAN</td>
<td></td>
<td>4e-17</td>
<td>29.4</td>
<td>2e-20</td>
<td>29.4</td>
<td>7e-16</td>
<td>32.0</td>
</tr>
<tr>
<td>GSTM1_CAEEL</td>
<td></td>
<td>6e-17</td>
<td>32.8</td>
<td>5e-21</td>
<td>31.9</td>
<td>1e-15</td>
<td>32.5</td>
</tr>
<tr>
<td>GSTP1_CSRMI</td>
<td></td>
<td>7e-17</td>
<td>29.4</td>
<td>6e-20</td>
<td>29.4</td>
<td>2e-14</td>
<td>31.4</td>
</tr>
<tr>
<td>GSTM1_HUMAN</td>
<td></td>
<td>2e-16</td>
<td>28.8</td>
<td>3e-20</td>
<td>28.8</td>
<td>2e-11</td>
<td>30.1</td>
</tr>
<tr>
<td>GSTM1_HUMAN</td>
<td></td>
<td>2e-14</td>
<td>29.6</td>
<td>6e-17</td>
<td>29.6</td>
<td>2e-09</td>
<td>29.8</td>
</tr>
<tr>
<td>HPGDS_MUS</td>
<td></td>
<td>2e-12</td>
<td>30.2</td>
<td>6e-14</td>
<td>29.3</td>
<td>2e-04</td>
<td>31.2</td>
</tr>
<tr>
<td>HPGDS_HUMAN</td>
<td></td>
<td>7e-12</td>
<td>29.2</td>
<td>5e-13</td>
<td>28.1</td>
<td>7e-03</td>
<td>31.4</td>
</tr>
<tr>
<td>GSTP1_HUMAN</td>
<td></td>
<td>1e-11</td>
<td>28.5</td>
<td>2e-13</td>
<td>27.2</td>
<td>2e-05</td>
<td>31.2</td>
</tr>
<tr>
<td>HPGDS_CHICK</td>
<td></td>
<td>7e-11</td>
<td>30.4</td>
<td>6e-13</td>
<td>30.2</td>
<td>7e-04</td>
<td>31.9</td>
</tr>
<tr>
<td>GSTP2_HUMAN</td>
<td></td>
<td>1e-10</td>
<td>28.9</td>
<td>6e-12</td>
<td>30.7</td>
<td>5e-04</td>
<td>35.5</td>
</tr>
<tr>
<td>GSTA1_HUMAN</td>
<td></td>
<td>1e-08</td>
<td>27.9</td>
<td>6e-10</td>
<td>26.4</td>
<td>6e-02</td>
<td>28.5</td>
</tr>
<tr>
<td>GSTA2_HUMAN</td>
<td></td>
<td>1e-07</td>
<td>27.3</td>
<td>5e-09</td>
<td>25.9</td>
<td>7e-02</td>
<td>28.2</td>
</tr>
<tr>
<td>GSTA3_HUMAN</td>
<td></td>
<td>1e-06</td>
<td>27.0</td>
<td>5e-08</td>
<td>25.8</td>
<td>7e-02</td>
<td>28.2</td>
</tr>
<tr>
<td>GSTA4_HUMAN</td>
<td></td>
<td>1e-05</td>
<td>26.9</td>
<td>5e-07</td>
<td>25.7</td>
<td>7e-02</td>
<td>28.1</td>
</tr>
<tr>
<td>GSTM1_MOUSE</td>
<td></td>
<td>2e-08</td>
<td>26.5</td>
<td>6e-09</td>
<td>27.2</td>
<td>5e-10</td>
<td>38.7</td>
</tr>
<tr>
<td>GSTA2_MOUSE</td>
<td></td>
<td>4e-08</td>
<td>27.0</td>
<td>6e-09</td>
<td>25.9</td>
<td>5e-02</td>
<td>41.2</td>
</tr>
<tr>
<td>GSTA3_MOUSE</td>
<td></td>
<td>1e-07</td>
<td>27.1</td>
<td>6e-08</td>
<td>25.8</td>
<td>7e-02</td>
<td>41.2</td>
</tr>
<tr>
<td>GSTM1_MOUSE</td>
<td></td>
<td>2e-07</td>
<td>27.1</td>
<td>6e-07</td>
<td>25.7</td>
<td>3e-02</td>
<td>50.0</td>
</tr>
<tr>
<td>GSTM1_MOUSE</td>
<td></td>
<td>3e-07</td>
<td>26.1</td>
<td>6e-07</td>
<td>44.4</td>
<td>9e-00</td>
<td>56.0</td>
</tr>
<tr>
<td>GSTM1_MOUSE</td>
<td></td>
<td>4e-07</td>
<td>26.1</td>
<td>6e-09</td>
<td>37.9</td>
<td>2e-02</td>
<td>41.2</td>
</tr>
<tr>
<td>GSTM1_MOUSE</td>
<td></td>
<td>5e-07</td>
<td>26.3</td>
<td>6e-10</td>
<td>41.2</td>
<td>2e-01</td>
<td>60.0</td>
</tr>
<tr>
<td>GSTM1_MOUSE</td>
<td></td>
<td>6e-07</td>
<td>26.3</td>
<td>7e-10</td>
<td>41.2</td>
<td>2e-01</td>
<td>60.0</td>
</tr>
<tr>
<td>GSTM1_MOUSE</td>
<td></td>
<td>7e-07</td>
<td>25.9</td>
<td>6e-10</td>
<td>42.1</td>
<td>2e-01</td>
<td>60.0</td>
</tr>
<tr>
<td>GSTM1_MOUSE</td>
<td></td>
<td>8e-07</td>
<td>25.9</td>
<td>6e-10</td>
<td>42.1</td>
<td>2e-01</td>
<td>60.0</td>
</tr>
<tr>
<td>GSTM1_MOUSE</td>
<td></td>
<td>9e-07</td>
<td>25.9</td>
<td>6e-10</td>
<td>42.1</td>
<td>2e-01</td>
<td>60.0</td>
</tr>
<tr>
<td>GSTM1_MOUSE</td>
<td></td>
<td>1e-06</td>
<td>26.0</td>
<td>6e-09</td>
<td>26.0</td>
<td>2e-02</td>
<td>44.2</td>
</tr>
<tr>
<td>GSTM1_MOUSE</td>
<td></td>
<td>1e-05</td>
<td>26.0</td>
<td>6e-08</td>
<td>25.7</td>
<td>2e-02</td>
<td>44.2</td>
</tr>
<tr>
<td>GSTM1_MOUSE</td>
<td></td>
<td>1e-04</td>
<td>26.0</td>
<td>6e-07</td>
<td>24.9</td>
<td>2e-02</td>
<td>44.2</td>
</tr>
<tr>
<td>GSTM1_MOUSE</td>
<td></td>
<td>1e-03</td>
<td>26.0</td>
<td>6e-06</td>
<td>24.8</td>
<td>2e-02</td>
<td>44.2</td>
</tr>
<tr>
<td>GSTM1_MOUSE</td>
<td></td>
<td>1e-02</td>
<td>26.0</td>
<td>6e-05</td>
<td>24.8</td>
<td>2e-02</td>
<td>44.2</td>
</tr>
<tr>
<td>GSTM1_MOUSE</td>
<td></td>
<td>1e-01</td>
<td>26.0</td>
<td>6e-04</td>
<td>24.8</td>
<td>2e-02</td>
<td>44.2</td>
</tr>
<tr>
<td>GSTM1_MOUSE</td>
<td></td>
<td>1e-00</td>
<td>26.0</td>
<td>6e-03</td>
<td>24.8</td>
<td>2e-02</td>
<td>44.2</td>
</tr>
</tbody>
</table>
Improving sensitivity with protein/domain family models

- Shallower scoring matrices (dialing back the $q_{ij}$ from the evolutionary model) reduces look-back time
  - VT20: 80% identity; VT80: 35% id; BL50: 25% id
  - reduced look-back = reduced sensitivity
- How to increase look-back time (more sensitivity)
  - Position Specific Scoring Matrices (PSSMs)
  - Hidden Markov Models (HMMs)

Pairwise Alignment

score matrices: 20x20, 210 parameters
position-independent
Profile Alignment

RU1A_HUMAN rrm1  SSATNAL
RU1A_HUMAN rrm2  VQAGAA
SFR1_HUMAN rrm1  RDAEDAV
SXLF_DROME rrm1  MSQRAI
PABP_DROME rrm3  EAAEAVV  query
+3 +4

profile: 20 scores per column
position-dependent

Where pairwise scores come from –

score(AA)=\log \frac{P(A|A)}{f(A)}

"probability of A given an A"
the observed probability of seeing an A
aligned to an A in real alignments

Sc(AA) = \log \frac{0.64}{0.04} = +4

"frequency of A"
the expected frequency of A in any sequence

Sc(AE) = \log \frac{0.01}{0.04} = -2
Two expressions for $S_{ij}$

Transition frequency (probability) 
- Durbin et al.

Alignment frequency (probability) 
- Altschul

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

Altschul $q_{ij}^a = p_i \times$ Durbin $q_{ij}^t$

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

Where profile scores (should) come from

"probability of A at position x" 
the observed probability of seeing an A 
in the consensus column x

\[ \text{score}(A|x) = \log \frac{P(A|\text{position } x)}{f(A)} \]

$\text{Sc}(A|6) = \log \frac{1.00}{0.04} = +4.6$

$\text{Sc}(A|5) = \log \frac{0.04}{0.04} = 0$

$\text{Sc}(N|6) = \log \frac{0.00}{0.06} = -\infty$

$\text{Sc}(N|5) = \log \frac{0.06}{0.06} = 0$

1. what about position-specific gap penalties?
2. how to estimate parameters from small numbers of observations?
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

PSI-BLAST iteratively builds a model (PSSM) of an ancient ATP synthase
PSI-BLAST ATP6_HUMAN - 4 iterations
Threshold 10^-20 for demo, use 10^-3

Results from round: (1) (2) (3) (4)

<table>
<thead>
<tr>
<th>Polypeptide</th>
<th>Score</th>
<th>E-value</th>
<th>Score</th>
<th>E-value</th>
<th>Score</th>
<th>E-value</th>
<th>Score</th>
<th>E-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATP6_HUMAN</td>
<td>296</td>
<td>3e-81</td>
<td>257</td>
<td>1e-49</td>
<td>241</td>
<td>2e-62</td>
<td>222</td>
<td>5e-59</td>
</tr>
<tr>
<td>ATP6_BOVIN</td>
<td>253</td>
<td>2e-48</td>
<td>257</td>
<td>2e-49</td>
<td>239</td>
<td>8e-45</td>
<td>230</td>
<td>2e-41</td>
</tr>
<tr>
<td>ATP6_MOUSE</td>
<td>245</td>
<td>5e-46</td>
<td>247</td>
<td>3e-46</td>
<td>234</td>
<td>4e-46</td>
<td>225</td>
<td>6e-40</td>
</tr>
<tr>
<td>ATP6_XENLA</td>
<td>142</td>
<td>9e-25</td>
<td>227</td>
<td>1e-40</td>
<td>189</td>
<td>3e-49</td>
<td>177</td>
<td>2e-49</td>
</tr>
<tr>
<td>ATP6_OHDA</td>
<td>101</td>
<td>2e-22</td>
<td>204</td>
<td>3e-54</td>
<td>209</td>
<td>5e-35</td>
<td>196</td>
<td>4e-31</td>
</tr>
</tbody>
</table>

Notes:
- ATP6_BOVIN: 1e-49, 8e-45, 2e-41
- ATP6_MOUSE: 3e-46, 4e-46, 6e-40
- ATP6_XENLA: 1e-40, 3e-49, 2e-49
- ATP6_OHDA: 3e-54, 5e-35, 4e-31

Multiple sequence alignment:
Metazoan ATP Synthases

<table>
<thead>
<tr>
<th>Fasta ID</th>
<th>Description</th>
<th>Score</th>
<th>E-value</th>
<th>Score</th>
<th>E-value</th>
<th>Score</th>
<th>E-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATP6_BOVIN</td>
<td>Myoglobin precursor (ATPase domain)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATP6_MOUSE</td>
<td>Myoglobin precursor (ATPase domain)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATP6_HUMAN</td>
<td>Myoglobin precursor (ATPase domain)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATP6_XENLA</td>
<td>Myoglobin precursor (ATPase domain)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATP6_OHDA</td>
<td>Myoglobin precursor (ATPase domain)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Multiple sequence alignment:
Capsid proteins

<table>
<thead>
<tr>
<th>Fasta ID</th>
<th>Description</th>
<th>Score</th>
<th>E-value</th>
<th>Score</th>
<th>E-value</th>
<th>Score</th>
<th>E-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATP6_BOVIN</td>
<td>Myoglobin precursor (ATPase domain)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATP6_MOUSE</td>
<td>Myoglobin precursor (ATPase domain)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATP6_HUMAN</td>
<td>Myoglobin precursor (ATPase domain)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATP6_XENLA</td>
<td>Myoglobin precursor (ATPase domain)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATP6_OHDA</td>
<td>Myoglobin precursor (ATPase domain)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Multiple sequence alignment:
Laminin

<table>
<thead>
<tr>
<th>Fasta ID</th>
<th>Description</th>
<th>Score</th>
<th>E-value</th>
<th>Score</th>
<th>E-value</th>
<th>Score</th>
<th>E-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATP6_BOVIN</td>
<td>Myoglobin precursor (ATPase domain)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATP6_MOUSE</td>
<td>Myoglobin precursor (ATPase domain)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATP6_HUMAN</td>
<td>Myoglobin precursor (ATPase domain)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATP6_XENLA</td>
<td>Myoglobin precursor (ATPase domain)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATP6_OHDA</td>
<td>Myoglobin precursor (ATPase domain)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Multiple sequence alignment:
Laminin

<table>
<thead>
<tr>
<th>Fasta ID</th>
<th>Description</th>
<th>Score</th>
<th>E-value</th>
<th>Score</th>
<th>E-value</th>
<th>Score</th>
<th>E-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATP6_BOVIN</td>
<td>Myoglobin precursor (ATPase domain)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATP6_MOUSE</td>
<td>Myoglobin precursor (ATPase domain)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATP6_HUMAN</td>
<td>Myoglobin precursor (ATPase domain)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATP6_XENLA</td>
<td>Myoglobin precursor (ATPase domain)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATP6_OHDA</td>
<td>Myoglobin precursor (ATPase domain)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Multiple sequence alignment:
Laminin

<table>
<thead>
<tr>
<th>Fasta ID</th>
<th>Description</th>
<th>Score</th>
<th>E-value</th>
<th>Score</th>
<th>E-value</th>
<th>Score</th>
<th>E-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATP6_BOVIN</td>
<td>Myoglobin precursor (ATPase domain)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATP6_MOUSE</td>
<td>Myoglobin precursor (ATPase domain)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATP6_HUMAN</td>
<td>Myoglobin precursor (ATPase domain)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATP6_XENLA</td>
<td>Myoglobin precursor (ATPase domain)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATP6_OHDA</td>
<td>Myoglobin precursor (ATPase domain)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Multiple sequence alignment:
Laminin

<table>
<thead>
<tr>
<th>Fasta ID</th>
<th>Description</th>
<th>Score</th>
<th>E-value</th>
<th>Score</th>
<th>E-value</th>
<th>Score</th>
<th>E-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATP6_BOVIN</td>
<td>Myoglobin precursor (ATPase domain)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATP6_MOUSE</td>
<td>Myoglobin precursor (ATPase domain)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATP6_HUMAN</td>
<td>Myoglobin precursor (ATPase domain)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATP6_XENLA</td>
<td>Myoglobin precursor (ATPase domain)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATP6_OHDA</td>
<td>Myoglobin precursor (ATPase domain)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Position-Specific Scores
ATP Synthase, 4 iterations

<table>
<thead>
<tr>
<th>A</th>
<th>R</th>
<th>N</th>
<th>D</th>
<th>C</th>
<th>Q</th>
<th>E</th>
<th>G</th>
<th>S</th>
<th>T</th>
<th>W</th>
<th>Y</th>
<th>V</th>
<th>bits/pos</th>
</tr>
</thead>
<tbody>
<tr>
<td>BL62</td>
<td>Q</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-3</td>
<td>5</td>
<td>2</td>
<td>-2</td>
<td>0</td>
<td>-3</td>
<td>-2</td>
<td>1</td>
</tr>
<tr>
<td>46</td>
<td>Q</td>
<td>-2</td>
<td>-1</td>
<td>-2</td>
<td>-4</td>
<td>6</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>-4</td>
<td>-3</td>
<td>-1</td>
<td>-2</td>
</tr>
<tr>
<td>47</td>
<td>Q</td>
<td>-1</td>
<td>-1</td>
<td>3</td>
<td>3</td>
<td>-3</td>
<td>3</td>
<td>3</td>
<td>-2</td>
<td>2</td>
<td>-4</td>
<td>-1</td>
<td>-3</td>
</tr>
<tr>
<td>56</td>
<td>Q</td>
<td>-2</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>5</td>
<td>2</td>
<td>-4</td>
<td>1</td>
<td>4</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
</tr>
<tr>
<td>97</td>
<td>Q</td>
<td>-2</td>
<td>-1</td>
<td>0</td>
<td>-2</td>
<td>-4</td>
<td>6</td>
<td>0</td>
<td>-3</td>
<td>8</td>
<td>-4</td>
<td>-1</td>
<td>-2</td>
</tr>
<tr>
<td>131</td>
<td>Q</td>
<td>3</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
<td>5</td>
<td>2</td>
<td>-2</td>
<td>-1</td>
<td>-3</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>152</td>
<td>Q</td>
<td>-2</td>
<td>6</td>
<td>-1</td>
<td>-2</td>
<td>-4</td>
<td>4</td>
<td>0</td>
<td>-3</td>
<td>-1</td>
<td>-3</td>
<td>1</td>
<td>-2</td>
</tr>
<tr>
<td>210</td>
<td>Q</td>
<td>-2</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>-4</td>
<td>7</td>
<td>1</td>
<td>-3</td>
<td>0</td>
<td>-4</td>
<td>-3</td>
<td>1</td>
</tr>
</tbody>
</table>

How much improvement with PSSMs?
- Structure comparison:
  - DALI, VAST, MATRAS, CE, STRUCTAL, SGM
- Pairwise sequence comparison:
  - SSEARCH
- Model-based sequence comparison:
  - PSI-BLAST


B. homologs and non-topolog errors

Errors per Query

0 0.1 1 10

Coverage

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

Structal
Dalii
CE
VAST
MATRAS
SSEARCH
PSI-BLAST
SGM
Sensitive searches with PSI-BLAST

- PSI-BLAST improves sensitivity by building a Position Specific Scoring Matrix (PSSM)
  - models ancestral sequence (consensus distribution)
  - similar to PFAM HMM (but less sophisticated weights, gaps)
- PSI-BLAST likes larger databases (more data)
- Sensitivity improves with additional iterations
  - model moves to base of tree
- Statistical estimates are difficult
  - once a sequence is in, it is “significant” - validation must be done before a sequence is included
- Very diverse families may not produce a well defined PSSM
  - similar problems with HMMs have led to “clans”

Finding domains with domain models I: from scoring matrices to PSSMs

- Domains are structurally compact, evolutionarily mobile, protein building blocks
  - atomic, they have a characteristic length
  - often repeated, or found in different sequence contexts
  - essential for building detection systems (PSSMs, HMMs), because they focus on the homologous region (a full length protein can be a mixture of domains)
  - Interpro provides large-scale summary
  - Pfam most comprehensive single resource
- Position independent scoring matrices can be built from a simple evolutionary model: $PAM^{1(n)} = PAM(n)$
- Position Specific Scoring Matrices (PSSMs) generalize frequency data for a single position
- PSI-BLAST increases sensitivity with PSSMs