Protein sequence analysis

Part 3: patterns, motifs, domains
and iterative BLAST
Why patterns are important

• Newly obtained protein sequence may not be similar to known sequences over entire length – makes functional assignment difficult
• Can gain insight into protein function by identifying conserved subsequences with known functions
• Such patterns, known as motifs and domains, are more conserved than other protein regions and tend to evolve (be gained, lost, or shuffled over time) as units
How patterns help

• Can’t always distinguish functional relationships between proteins using simple BLAST because of evolutionary divergence

• Proteins often perform multiple functions; single annotation in a sequence database may be inadequate to describe this fully

• Identification of motifs & domains useful in resolving these issues
  – serve as diagnostic features for protein family
  – associated functional features in a query sequence can be revealed more rapidly by confirming presence of pattern than by global matching of sequences
Motifs & Domains

• Motif: short conserved sequence pattern
  – often associated with distinct structural/functional site
  – 10-20 amino acids in length

• Domains are larger patterns that have functional significance
  – independent globular unit: keeps its shape when apart from the rest of the structure
  – 50 or more amino acids in a domain, each representing a modular component; average length is 100 AAs
  – Most proteins contain 2 or 3 domains
ProSite

• Can be used to search for motifs that may be present as a result of post-translational modifications

• ProSite contains a repository of such motifs with specific biological properties

• Longer matches are more likely to be reliable; similar short sequences may have similar functions, but such a result is questionable, because the shorter the match, the more likely it could occur by chance
ProSite


- Can type in sequence number from SwissProt/TrEMBL or PDB, or paste in raw sequence (no FASTA heading)
Example: Using ProSite

- Reverse the sequence search parameters, as shown below; for now, we are interested in even short patterns, and we don’t want to take the time required to scan profiles, since patterns are better indicators of post-translational modification
- Click START THE SCAN button (bottom of page)
Example result (from KPC1_DROME)

<table>
<thead>
<tr>
<th>Ruler</th>
<th>1</th>
<th>100</th>
<th>200</th>
<th>300</th>
<th>400</th>
<th>500</th>
<th>600</th>
<th>700</th>
<th>800</th>
<th>900</th>
<th>1000</th>
</tr>
</thead>
</table>

**Hits by patterns:** [3 hits (by 3 distinct patterns) on 1 sequence]

**USERSEQ1**

PS00479  ZF_DAG_PE_1  *Zinc finger phorbol-ester/DAG-type signature*:

<table>
<thead>
<tr>
<th>Start - Stop</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>120 - 171</td>
<td>HnFepftyagptf.CdhCgs1Lygiyhqglk.CsaCdmnvHarCkenvpslCgC</td>
</tr>
</tbody>
</table>

PS00107  PROTEIN_KINASE_ATP  *Protein kinases ATP-binding region signature*:

<table>
<thead>
<tr>
<th>Start - Stop</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>356 - 379</td>
<td>LGKGSFGKVL1Aerkgseel.......YAIK</td>
</tr>
</tbody>
</table>

PS00108  PROTEIN_KINASE_ST  *Serine/Threonine protein kinases active-site signature*:

<table>
<thead>
<tr>
<th>Start - Stop</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>470 - 482</td>
<td>IlYrDLK1dNVLL</td>
</tr>
</tbody>
</table>

**Hits by patterns with a high probability of occurrence or by user-defined patterns:** [31 hits (by 6 distinct patterns) on 1 sequence]

**USERSEQ1**
Notes on ProSite results

• A strong pattern – such as the three shown at the top of the output on the previous slide – suggests a “real” result
• Two relatively weak patterns, if related and in close proximity, might be considered a single strong pattern
• ProSite is not exhaustive: there are post-translational modifications that aren’t listed in its database (leads to false negatives)
• More accurate results may be obtained from programs found at the following URL:
  http://motif.stanford.edu/
  these are not web-based & must be installed on your computer
Domains & Databases

• Several domain collections exist at various sites
• Domain definition and identification are tricky; not everybody agrees as to what constitutes a domain, so there may be considerable discrepancy in results from different databases
• Your best option is to try multiple databases and look for possible consensus
• A good place to start is InterPro: searches many databases simultaneously
Searching InterPro

• URL: http://www.ebi.ac.uk/Tools/InterProScan/
  – paste in raw FASTA (no heading)
  – choose applications
  – submit job … wait

Test sequence shown is KPC1_DROME
This can take awhile ...

Unchecking some of the larger databases (ProDom, in particular, is substantially larger than most) can speed up the process.
Interpreting the results

• First column: indicates type of information provided; e.g. specific to a domain, family, active site, etc.

• The link in each box points to InterPro summary of documentation on the entry from all of the databases where the pattern was found
Interpreting the results

• Each entry includes a hyperlink to the original database and a graphical representation of the match – size is proportional to extent of match; some examples:
Other domain servers: CD server

- CD (conserved domain) search at NCBI can be reached directly at:
  

  or by clicking a link on a BLAST results page:
Using CDServer

Image above shows default options with sequence (KPC1_DROME) again pasted in; next slide shows different options.
Using CDServer

- Several databases to choose from using drop arrow; CDD is the most comprehensive
- We turn off the low-complexity filter to get results that include repeated patterns
- We also made the Expect Threshold less stringent, to widen the field of results
- If we get too much information, we can always tweak any or all of these options
- Click Submit button to run query
Results

Conserved domains on [cl|seqsig_c1677dba754ea2bf8c4695e1b2ed135]

- Graphical summary shows regions of protein that match a domain
- Hit list reports matches in order by E-value (best to worst)
- Clicking the “View Full Results” button yields longer list
PSI-BLAST: Protein-Specific Iterated BLAST

• The CD server is known by another name: reverse PSI-BLAST

• Given that CD server is designed to look for domains, you could logically conclude that the reverse of this is looking for what the domains belong to – in other words, finding all proteins even remotely related to query sequence

• This is what PSI-BLAST is for
PSI-BLAST

• As the name implies, PSI-BLAST works iteratively:
  – repeat query for several rounds
  – first round produces results similar to blastp
  – in next round, algorithm will eliminate results below a threshold for consideration
  – user has option to select/deselect which sequences should be considered
Using PSI-BLAST

• You get to PSI-BLAST via the same links that take you to blastp; in fact, it’s exactly the same page

• Near the bottom of the query screen, you click a radio button to specify which BLAST program to run: choose PSI-BLAST
Results

• Look much like blastp results, but with added features:

• In the first round, all results will be marked NEW; the green dot doesn’t appear until second & subsequent rounds
Results

• Individual results have checkboxes
  – if checked, sequence used in next round
  – if unchecked, sequence is discarded

Box at left shows (arbitrarily) checked & unchecked sequences
Results

Illustration above shows a sample of second-round results