Protein Structure Analysis

Iosif Vaisman

2007

Regular Expressions

<table>
<thead>
<tr>
<th>Operation</th>
<th>Regular Expression</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concatenation</td>
<td>DLIIV</td>
<td>DLIV</td>
</tr>
<tr>
<td>Alternation</td>
<td>D[LIV]K</td>
<td>DLK</td>
</tr>
<tr>
<td>Replication</td>
<td>DL(2,5)K</td>
<td>DLLK</td>
</tr>
</tbody>
</table>

Patterns described in a standard way are known as regular expressions

[AC]-x-V-x(4)-{ED}.  
[Ala or Cys]-any-Val-any-any-any-any-{any but Glu or Asp}

...LKVHAYVFGALIYWIK...
...AVEMASVRYLQVQHGS...
...LYTQAIVNNQPGYMA...
...KEYRCKKEKELTDCN...

Regular Expressions

Sequence patterns

KKFAQSTNLKSHILT
KQFSHSAQLRAHIST
GKFDSNQLKSHMLV
KDISSSESRLTHMFK
KRFSHSGSYSSHIISS
KRFSHSGSFSHMTS
KTLSDRLEYQQHMLK

PROSITE Database

Current version contains 1079 documentation entries that describe 1459 different patterns, rules and profiles/matrices

[ST]-x(2)-{DE}
Casein kinase II phosphorylation site

[AG]-x(4)-G-K-[ST]
ATP/GTP-binding site motif A (P-loop)

Y-x-[NOH]-K-{DE}-[IVA]-F-[LM]-R-[ED]
Heat shock hsp90 proteins family signature

http://www.expasy.ch/prosite
Blocks Database

Blocks are multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins

N-6 Adenine-specific DNA methylases proteins

<table>
<thead>
<tr>
<th>Block ID</th>
<th>Start</th>
<th>Stop</th>
<th>DA block</th>
</tr>
</thead>
<tbody>
<tr>
<td>DMA_VIBCH</td>
<td>85</td>
<td>77</td>
<td>CTQWWPPF</td>
</tr>
<tr>
<td>HEMK_MYCLE</td>
<td>181</td>
<td>100</td>
<td>DLFWAQXPTL</td>
</tr>
<tr>
<td>MT57_ECOLI</td>
<td>111</td>
<td>13</td>
<td>DGALGNPPF</td>
</tr>
<tr>
<td>MTC1_CHVN1</td>
<td>172</td>
<td>8</td>
<td>NFVFLDPPY</td>
</tr>
<tr>
<td>MTC1_COREQ</td>
<td>71</td>
<td>49</td>
<td>QLSFSCPPF</td>
</tr>
<tr>
<td>MTH2_HAEHA</td>
<td>32</td>
<td>52</td>
<td>KIAFFDPQY</td>
</tr>
<tr>
<td>MTH3_HAEIN</td>
<td>23</td>
<td>73</td>
<td>HAIISDIPY</td>
</tr>
<tr>
<td>MTM1_MICAM</td>
<td>306</td>
<td>14</td>
<td>AAVLTPPPFP</td>
</tr>
<tr>
<td>MTM2_MORBO</td>
<td>25</td>
<td>14</td>
<td>QLAVIDPPY</td>
</tr>
<tr>
<td>MTMU_MYCSP</td>
<td>37</td>
<td>13</td>
<td>QVIYADPPW</td>
</tr>
<tr>
<td>MTR1_RHOSH</td>
<td>60</td>
<td>8</td>
<td>QLIICDPPY</td>
</tr>
</tbody>
</table>

http://www.blocks.fhcrc.org/

Pfam Database

Pfam is a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains

Zinc finger, C2H2 type

<table>
<thead>
<tr>
<th>Protein ID</th>
<th>Start</th>
<th>Stop</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>TYY1_HUMAN</td>
<td>383-407</td>
<td>YVCPF.DGCN...KKFAQSTNLKSHILT...H</td>
<td></td>
</tr>
<tr>
<td>ZG52_XENLA</td>
<td>61-83</td>
<td>YTCT...QCN...KQFSHSAQLRARIST...H</td>
<td></td>
</tr>
<tr>
<td>KRUP_DROME</td>
<td>306-328</td>
<td>YTCE...ICD...GKFSDSNQLKSHMLV...H</td>
<td></td>
</tr>
<tr>
<td>YKQ8_CAEEL</td>
<td>78-102</td>
<td>YKCT...VCR...KDISSSESLRTHMFKQ.HH</td>
<td></td>
</tr>
<tr>
<td>DEF1_CHICK</td>
<td>269-292</td>
<td>YECP...NCK...KFPKNSGGSHSSFKEK.KC</td>
<td></td>
</tr>
<tr>
<td>ZFH1_DROME</td>
<td>389-413</td>
<td>FGCD...NCG...KRFSHSGSFSSHMTSK.KC</td>
<td></td>
</tr>
<tr>
<td>YL57_CAEEL</td>
<td>42-65</td>
<td>YLCY...YCG...KTLSDRLEYQQHMLK..VH</td>
<td></td>
</tr>
<tr>
<td>ZFA_MOUSE</td>
<td>542-564</td>
<td>FKCD...ICL...LTFSDTKEVQQHALV...H</td>
<td></td>
</tr>
<tr>
<td>BASO_HUMAN</td>
<td>719-742</td>
<td>FQCD...ICK...KTFKNACSVKIHHKN..MH</td>
<td></td>
</tr>
<tr>
<td>HUNB_DROME</td>
<td>297-319</td>
<td>FQCD...KCS...YTCVNKSMLNSHRKS...H</td>
<td></td>
</tr>
<tr>
<td>SFP1_YEAST</td>
<td>598-623</td>
<td>FKCPV.IGCE...KTYKNQNGLKYHRLH..GH</td>
<td></td>
</tr>
<tr>
<td>ZG29_XENLA</td>
<td>62-84</td>
<td>FVCT...VCG...KTYKYKHGLNTHLHS...H</td>
<td></td>
</tr>
</tbody>
</table>

http://pfam.wustl.edu/

Other Motif Databases

PRINTS : a compendium of protein fingerprints. A fingerprint is a group of conserved motifs used to characterise a protein family
http://bioinf.man.ac.uk/dbbrowser/PRINTS/

DOMO : a protein domain database
http://www.infobiogen.fr/~gracy/domo/home.htm

ProDom : a protein domain database
http://protein.toulouse.inra.fr/prodom.html

InterPro Database

InterPro : integrated resource for the commonly used signature databases - Pfam, PRINTS, PROSITE, ProDom and SWISS-PROT + TrEMBL.

Current release of InterPro (3.2) contains 3939 entries, representing 1009 domains, 2850 families, 65 repeats and 15 post-translational modification sites.

http://www.ebi.ac.uk/interpro

Sequence-structure patterns

Bradley et al., PNAS, 2002
Sequence-structure patterns

Structural motifs of PROSITE patterns

Distribution of rmsd values for the true hits. The rmsd was calculated from all true hits eliminating false and unidentified hits for each of the 466 patterns having more than one true hit.

Distribution of number of hits per pattern. Each column represents frequency of patterns having sequence matches in the 3D-sequence library. The bin width is 10. There were 1058 patterns having numbers of hits smaller than 10, including 712 patterns with no hits.

Protein representation (Crambin)

Neighbor identification in proteins

10 neighbors

7 neighbors

2 neighbors
Voronoi Tessellation

Delaunay simplex is defined by points, whose Voronoi polyhedra have common vertex.

Delaunay simplex is always a triangle in a 2D space and a tetrahedron in a 3D space.

Neighbor identification in proteins: Voronoi/Delaunay Tessellation in 2D

Delaunay tessellation of Crambin

Voronoi Tessellation
Delaunay Tessellation

6
7
6

Delaunay simplices classification

\[ d_{ij}, d_{jk}, d_{kl} \]
Three views at one object:

Topological: simplex Delaunay
Geometrical: tetrahedron
Compositional: quadruplet of points

Classification of Delaunay simplices by sequential proximity

Types of Delaunay simplices in Crambin

Correlations between protein structure family assignment and relative content of classes of Delaunay simplices

Tetrahedrality of Delaunay simplices

Tetrahedrality distribution of Delaunay simplices

\[ T = \frac{\sum_{ij} (d_{ij} - t)^2}{15T^2} \]