EECS 730
Introduction to Bioinformatics
BLAST: Basic Local Alignment Search Tools

Luke Huan
Electrical Engineering and Computer Science
http://people.eecs.ku.edu/~jhuan/
Sequence Similarity Searching in Sequence databases

- The premise:
  - One sequence by itself is not informative; it must be analyzed by comparative methods against existing sequence databases to develop hypothesis concerning relatives and function.
Similarity Searches in Sequence Databases

Goals

- Identify all homologs (true positives)
  - infer function, transfer annotations, structure/domain information
- Limit the misidentification of non-homologs (false positives)
- Search large sets of sequences efficiently
Sequence Similarity Comparisons

- Alignments can be global or local (this is algorithm specific)
  - A global alignment is an optimal alignment that includes all characters from each sequence (Clustal generates global alignments)
  - A local alignment is an optimal alignment that includes only the most similar local region or regions (BLAST generates local alignments).

```
>gi|2498170|sp|Q27974|AUXI_BOVIN Auxilin
Length = 910

Score = 107 bits (268), Expect = 4e-23
Identities = 76/275 (27%), Positives = 131/275 (47%), Gaps = 21/275 (7%)

Query: 22 DLDLTYIYPNNIAGFAPPARLIVYRNNIDVVRFLDSKHKNYKIYNYLCAERHYDTAKF 81
       DLD TYI+  II M FP + ++ +RN +DD+ +LDS+H +HY +YNL + + Y TAKF
Sbjct: 60 DLDFTYVTSTIIVMSFPLDSVDIGFRNQVDDIRSFLDSRHLDDHYTVYNL-SPKSYRTAKF 118
```
The BLAST algorithm

- The BLAST programs (Basic Local Alignment Search Tools) are a set of sequence comparison algorithms introduced in 1990 that are used to search sequence databases for optimal local alignments to a query.


### Several different BLAST programs:

<table>
<thead>
<tr>
<th>Program</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>blastp</td>
<td>Compares an amino acid query sequence against a protein sequence database.</td>
</tr>
<tr>
<td>blastn</td>
<td>Compares a nucleotide query sequence against a nucleotide sequence database.</td>
</tr>
<tr>
<td>blastx</td>
<td>Compares a nucleotide query sequence translated in all reading frames against a protein sequence database. You could use this option to find potential translation products of an unknown nucleotide sequence.</td>
</tr>
<tr>
<td>tblastn</td>
<td>Compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames.</td>
</tr>
<tr>
<td>tblastx</td>
<td>Compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database. Please note that the tblastx program cannot be used with the nr database on the BLAST Web page because it is too computationally intensive.</td>
</tr>
</tbody>
</table>
Considerations for choosing a BLAST database

- First consider your research question:
  - Are you looking for a particular gene in a particular species?
    - BLAST against the genome of that species.
  - Are you looking for additional members of a protein family across all species?
    - BLAST against the non-redundant database (nr).
  - Are you looking to annotate genes in your species of interest?
    - BLAST against known genes in a closely related species.
When choosing a database for BLAST…

- **It is important to know your agent.**
  - Changing your choice of database is changing your search space
  - Database size affects the BLAST statistics
    - record BLAST parameters, database choice, database size in your bioinformatics lab book, just as you would for your wet-bench experiments.
  - Databases change rapidly and are updated frequently
    - It may be necessary to repeat your analyses
BLAST protein databases available at through blastp web interface @ NCBI

[Image of BLAST interface with "blastp db" highlighted]
### BLAST Results

**Database:** PDB protein database
- 31,492 sequences; 7,075,060 total letters

**Query:** 1jx4a  **length:** 341

**Length:** 341

<table>
<thead>
<tr>
<th>PDB</th>
<th>Chain</th>
<th>Description</th>
<th>Score</th>
<th>E Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>pdb</td>
<td>2ASD</td>
<td>Chain A, Oxog-Modified Insertion Ternary Complex &gt;...</td>
<td>592</td>
<td>6e-170</td>
</tr>
<tr>
<td>pdb</td>
<td>2BQ3</td>
<td>Chain A, Dna Adduct Bypass Polymerization By Sulfo...</td>
<td>589</td>
<td>4e-169</td>
</tr>
<tr>
<td>pdb</td>
<td>1JXL</td>
<td>Chain A, Crystal Structure Of A Y-Family Dna Polym...</td>
<td>589</td>
<td>4e-169</td>
</tr>
<tr>
<td>pdb</td>
<td>2IMW</td>
<td>Chain P, Mechanism Of Template-Independent Nucleot...</td>
<td>587</td>
<td>1e-163</td>
</tr>
<tr>
<td>pdb</td>
<td>2AG0</td>
<td>Chain A, Fidelity Of Dpo4: Effect Of Metal Ions, N...</td>
<td>587</td>
<td>1e-163</td>
</tr>
<tr>
<td>pdb</td>
<td>2UWR</td>
<td>Chain A, Crystal Structures Of Mutant Dpo4 Dna Pol...</td>
<td>587</td>
<td>1e-163</td>
</tr>
<tr>
<td>pdb</td>
<td>2UWV</td>
<td>Chain A, Crystal Structures Of Mutant Dpo4 Dna Pol...</td>
<td>587</td>
<td>2e-163</td>
</tr>
<tr>
<td>pdb</td>
<td>1JX4</td>
<td>Chain A, Crystal Structure Of A Y-Family Dna Polym...</td>
<td>585</td>
<td>7e-163</td>
</tr>
<tr>
<td>pdb</td>
<td>1K1Q</td>
<td>Chain A, Crystal Structure Of A Dinh Family Error ...</td>
<td>329</td>
<td>8e-91</td>
</tr>
<tr>
<td>pdb</td>
<td>1M4</td>
<td>Chain A, Crystal Structure Of A Dinh Homolog (Dbh)...</td>
<td>243</td>
<td>7e-65</td>
</tr>
<tr>
<td>pdb</td>
<td>2DPI</td>
<td>Chain A, Ternary Complex Of Hpoli With Dna And Dct...</td>
<td>105</td>
<td>2e-23</td>
</tr>
<tr>
<td>pdb</td>
<td>2ALZ</td>
<td>Chain A, Ternary Complex Of Hpoli With Dna And Dctp</td>
<td>104</td>
<td>4e-23</td>
</tr>
</tbody>
</table>
>gi|39998047|ref|NP_953998.1| redox-active disulfide protein 2 [Geobacter sulfurreducens PCA]
>gi|39984992|gb|AAR36348.1| redox-active disulfide protein 2 [Geobacter sulfurreducens PCA]

Length = 78

Score = 58.5 bits (140), Expect = 2e-08
Identities = 31/76 (40%), Positives = 54/76 (71%)

Query: 1
MMKIQIYGTGCANCOQMLEKNAREAVKELGIDAEFKIKEMDQILEAGLTLAPGLAVIDGEL 60
+MKI++ GTGCA C+ L +N ++AV+ G +AE K++E+ +I++ G+ + P L +DG +
Sbjct: 3
IMKIEVLGTCACKTLYENVQKAEMGKEAVVKVEEIQKIMKYGVSTPATIVEGTVV 62

Query: 61
KIMGRVASKEEEKKIL 76
K  G+V + +EIK +L
Sbjct: 63
KFSGKVPAADIEKGLML 78

>gi|34557156|ref|NP_906971.1| hypothetical protein WS0755 [Wolinella succinogenes]
>gi|344882872|emb|CAE09871.1| hypothetical protein [Wolinella succinogenes]

Length = 80

Score = 58.2 bits (139), Expect = 3e-08
Identities = 31/75 (41%), Positives = 52/75 (69%)

Query: 2
MKIQUIYGTGCANCOQMLEKNAREAVKELGIDAEFKIKEMDQILEAGLTLAPGLAVIDGELK 61
MKI+I GTGC C+ L + +EAV + G A+ EK++++ +I+ G+ + P L VDG +K
Sbjct: 1
MKIEILGTCACKLALFATKEAVAQSGKFAQIEKVEDIQKIMGYGVSTPATIVEGTVK 60

Query: 62
IMGRVASKEEEKKIL 76
+ G+V S +E+K++L
Sbjct: 61
LSGKVPSVDELKRVL 75
How does the BLAST algorithm work?

- Three phases
  - Compile a list of word pairs with a fixed length above a threshold $T$
  - Scan the database for entries that match the compiled list of word pairs.
  - Extend a hit in either direction to create a local alignment and compute the significance of the alignment
QUERY sequence(s)

BLAST program

BLAST database

BLAST results

Distribution of 26 Blast Hits on the Query Sequence

Color Key for Alignment Scores

- (0, 40)
- (40, 80)
- (80, 120)
- (120, 160)
- (160, 200)
- (200, 240)
BLAST Algorithm

- Phase 1: compile a list of “words” that are similar to each other
BLAST Algorithm

- Phase 2: search for occurrences of the similar words in sequence database
BLAST Algorithm

- Phase 3: BLAST algorithm extends the initial word-pair (known as “seed”) into high scoring segment pair (HSP)
  - Local optimal alignment
  - More than one HSP can be found

```
KENFDKARFSGTYWAMAKKDPEG  50
MKGLDIPvQKVAGTWYSLAMAASD.  44

RBP (query)
lactoglobulin (hit)
```

extend Hit! extend
Sequence Similarity Searching –
Statistics are important

- Discriminating between true and false matches is done using an estimate of probability that the match might occur by chance.
- BLAST compute the matching scores (S) and the matching e-values (E)
What do the Score and the e-value really mean?

- The **quality** of the alignment is represented by the **Score**.
  - **Score (S)**
  - The score of an alignment is calculated as the sum of substitution and gap scores. Substitution scores are given by a look-up table (PAM, BLOSUM) whereas gap scores are assigned empirically.

- The **significance** of each alignment is computed as an **E value**.
  - **E value (E)**
  - Expectation value. The number of different alignments with scores equivalent to or better than S that are expected to occur in a database search by chance. The lower the E value, the more significant the score.
How to Compute an E-value?

\[ E = Kmn \, e^{-\lambda S} \]

This equation is derived from a description of the extreme value distribution

\( S = \) the score

\( E = \) the expect value = the number of high-scoring segment pairs (HSPs) expected to occur with a score of at least \( S \)

\( m, n = \) the length of two sequences

\( \lambda = \) a parameter that scales the scoring system

\( K = \) a scaling factor for the search space
The interpretation of E-value and P-value?

- The E-value is not a probability; it’s an expect value
  - The BLAST programs report $E$-value rather than $P$-values because it is easier to understand the difference between, for example, $E$-value of 5 and 10 than $P$-values of 0.993 and 0.99995.
  - However, when $E < 0.01$, $P$-values and $E$-value are nearly identical.
  - $P = 1 - e^{-E}$
Other BLAST Algorithms

- PSI-BLAST
- PHI-BLAST
- Find-a-gene
Position specific iterated BLAST: PSI-BLAST

The purpose of PSI-BLAST is to look deeper into the database for matches to your query protein sequence by employing a scoring matrix that is customized to your query.

Finding distantly related proteins
PSI-BLAST is performed in five steps

[1] Select a query and search it against a protein database

[2] PSI-BLAST constructs a multiple sequence alignment then creates a “profile” or specialized position-specific scoring matrix (PSSM)
<table>
<thead>
<tr>
<th>Sequence ID</th>
<th>Count</th>
<th>Amino Acid Sequence</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>730496</td>
<td>66</td>
<td>FTVDENGQMSATAKGRVRLFNNUDVCADMIGSFTDTEDPAKFKMKYUGVASFLQKGNDDH</td>
<td>125</td>
</tr>
<tr>
<td>200679</td>
<td>63</td>
<td>FSVDEKGHMSATAKGRVRLLSNWEVCADMVGTFDTEDPAKFKMKYUGVASFLQQRNGNDH</td>
<td>122</td>
</tr>
<tr>
<td>206589</td>
<td>34</td>
<td>FSVDEKGHMSATAKGRVRLLSNWEVCADMVGTFDTEDPAKFKMKYUGVASFLQQRNGNDH</td>
<td>93</td>
</tr>
<tr>
<td>2136812</td>
<td>2</td>
<td>MSATAKGRVRLLSNWEVCADMVGTFDTEDPAKFKMKYUGVASFLQQRNGNDH</td>
<td>53</td>
</tr>
<tr>
<td>132408</td>
<td>65</td>
<td>FKIEdNGKTTATAKGRVRIILKLELCANMVGTFTETNDAKRYMKYHGAILERGLDDH</td>
<td>124</td>
</tr>
<tr>
<td>267584</td>
<td>44</td>
<td>FSVDESGKVTATAHGRVIIILNNWEMCANMFGTFEDTPDPAKFKMRYUGAASYLQTGNDH</td>
<td>103</td>
</tr>
<tr>
<td>267585</td>
<td>44</td>
<td>FSVDESGKVTATAQGRVIILNNWEMCANMFGTFEDTPDPAKFKMRYUGAAYLQSGNDH</td>
<td>103</td>
</tr>
<tr>
<td>8777608</td>
<td>63</td>
<td>FTVHEDGAMTATAKGRVIILNNWEMCANMDFETTTDPAKFKRMRIUGAASYLQTGNDH</td>
<td>122</td>
</tr>
<tr>
<td>6687453</td>
<td>60</td>
<td>FKVVEDGTMATAIGRVIILNNWEMCANMFGTFEDTEDPAKFKMRYUGAAYLQSGNDH</td>
<td>119</td>
</tr>
<tr>
<td>10697027</td>
<td>81</td>
<td>FKVQEDGTMATAATGRIILNNWEMCANMFGTFEDTEEPARFKMKUYGAAYLQSGYDDH</td>
<td>140</td>
</tr>
<tr>
<td>13645517</td>
<td>1</td>
<td>MGVTFTDTEDPAKFKMKYUGVASFLQKGNDDH</td>
<td>32</td>
</tr>
<tr>
<td>13925316</td>
<td>38</td>
<td>FSVDESGKMTATAQGRVIILNNWEMCANMFGTFEDTPDPAKFKMRYUGAAYLQSGNDH</td>
<td>97</td>
</tr>
<tr>
<td>131649</td>
<td>65</td>
<td>YTVEEDGTMATASSKGRVKLGFWICADMADQYTRGPTTPAKMYMTYQGALSYLGSSGDDNY</td>
<td>126</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>R</td>
<td>N</td>
</tr>
<tr>
<td>1</td>
<td>M</td>
<td>-1</td>
<td>-2</td>
</tr>
<tr>
<td>2</td>
<td>K</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>W</td>
<td>-3</td>
<td>-3</td>
</tr>
<tr>
<td>4</td>
<td>V</td>
<td>0</td>
<td>-3</td>
</tr>
<tr>
<td>5</td>
<td>W</td>
<td>-3</td>
<td>-3</td>
</tr>
<tr>
<td>6</td>
<td>A</td>
<td>5</td>
<td>-2</td>
</tr>
<tr>
<td>7</td>
<td>L</td>
<td>-2</td>
<td>-2</td>
</tr>
<tr>
<td>8</td>
<td>L</td>
<td>-1</td>
<td>-3</td>
</tr>
<tr>
<td>9</td>
<td>L</td>
<td>-1</td>
<td>-3</td>
</tr>
<tr>
<td>10</td>
<td>L</td>
<td>-2</td>
<td>-2</td>
</tr>
<tr>
<td>11</td>
<td>A</td>
<td>5</td>
<td>-2</td>
</tr>
<tr>
<td>12</td>
<td>A</td>
<td>5</td>
<td>-2</td>
</tr>
<tr>
<td>13</td>
<td>W</td>
<td>-2</td>
<td>-3</td>
</tr>
<tr>
<td>14</td>
<td>A</td>
<td>3</td>
<td>-2</td>
</tr>
<tr>
<td>15</td>
<td>A</td>
<td>2</td>
<td>-1</td>
</tr>
<tr>
<td>16</td>
<td>A</td>
<td>4</td>
<td>-2</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>S</td>
<td>2</td>
<td>-1</td>
</tr>
<tr>
<td>38</td>
<td>G</td>
<td>0</td>
<td>-3</td>
</tr>
<tr>
<td>39</td>
<td>T</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>40</td>
<td>W</td>
<td>-3</td>
<td>-3</td>
</tr>
<tr>
<td>41</td>
<td>Y</td>
<td>-2</td>
<td>-2</td>
</tr>
<tr>
<td>42</td>
<td>A</td>
<td>4</td>
<td>-2</td>
</tr>
</tbody>
</table>
PSI-BLAST is performed in five steps

[1] Select a query and search it against a protein database

[2] PSI-BLAST constructs a multiple sequence alignment then creates a “profile” or specialized position-specific scoring matrix (PSSM)

[3] The PSSM is used as a query against the database

[4] PSI-BLAST estimates statistical significance (E values)
apolipoprotein D [Rattus norvegicus]... 147 4e-35
alpha-microglobulin/bikunin... 144 6e-34
apolipoprotein D, apoD [human, plasma, ... 143 8e-34
(X02824) RBP (aa 101-172) [Homo sapiens] 139 1e-32
apolipoprotein D precursor [Homo sapiens] 138 4e-32
APOLIPOPROTEIN D PRECURSOR >gi|482... 134 4e-31
APOLIPOPROTEIN D PRECURSOR >gi|11... 133 7e-31
mutant retinol binding prot... 80 9e-15
apolipoprotein D, apoD (C-ter... 77 8e-14
(AF025335) mutant retinol binding prot... 67 8e-11
LAZARILLO PROTEIN PRECURSOR >gi|... 63 1e-09
AMBPROtein PRECURSOR [CONTAINS... 63 2e-09
AMBPROtein PRECURSOR [CONTAINS... 63 2e-09
alpha 1 microglobulin/bikunin [Mus... 62 2e-09
(ARO4907) putative [Mus musculus] 62 3e-09
alpha-1 microglobulin/bikunin [Rattu... 62 3e-09
AMBPROtein PRECURSOR [CONTAINS: ... 61 8e-09
alpha-1-microglobulin/inter-alpha-trypsin... 60 1e-08
AMBPROtein PRECURSOR [CONTAINS... 59 2e-08
AMBPROtein PRECURSOR - pig 59 2e-08
(X52007) precursor codes for two protein... 59 2e-08
neural Lazariello ... 59 3e-08
(AB006584) NLaz gene product [Drosophi... 58 3e-08
CRUSTACYANIN A2 SUBUNIT >gi|10275... 57 8e-08
AMBPROtein PRECURSOR [CONTAINS... 57 1e-07
insecticyanin A - tobacco hornworm >gi|971... 56 1e-07
AMBPROtein PRECURSOR - cat (fragment) 54 8e-07
(AB085089) gallerin [Galleria mellonella] 56 2e-07
AMBPROtein PRECURSOR [CONTAINS... 55 3e-07
[D83712] Prostaglandin D Synthase [Xe... 54 5e-07
retinol-binding protein - cat (fragment) 54 8e-07
LIPOCALIN PRECURSOR >gi|104284|pi... 53 1e-06
retinol-binding protein, RBP (N-termina... 52 3e-06
gene cpl-1 protein - African clawed frog ... 52 3e-06
OUTER MEMBRANE LIPOPROTEIN BLC PRE... 51 9e-06
PSI-BLAST is performed in five steps

[1] Select a query and search it against a protein database

[2] PSI-BLAST constructs a multiple sequence alignment then creates a “profile” or specialized position-specific scoring matrix (PSSM)

[3] The PSSM is used as a query against the database

[4] PSI-BLAST estimates statistical significance (E values)

## Results of a PSI-BLAST search

<table>
<thead>
<tr>
<th>Iteration</th>
<th># hits</th>
<th>&gt; threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>104</td>
<td>49</td>
</tr>
<tr>
<td>2</td>
<td>173</td>
<td>96</td>
</tr>
<tr>
<td>3</td>
<td>236</td>
<td>178</td>
</tr>
<tr>
<td>4</td>
<td>301</td>
<td>240</td>
</tr>
<tr>
<td>5</td>
<td>344</td>
<td>283</td>
</tr>
<tr>
<td>6</td>
<td>342</td>
<td>298</td>
</tr>
<tr>
<td>7</td>
<td>378</td>
<td>310</td>
</tr>
<tr>
<td>8</td>
<td>382</td>
<td>320</td>
</tr>
</tbody>
</table>
PSI-BLAST alignment of RBP and $\beta$-lactoglobulin:
iteration 1

Score = 46.2 bits (108), Expect = $2e^{-04}$
Identities = 40/150 (26%), Positives = 70/150 (46%), Gaps = 37/150 (24%)

Query: 27  VKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLNNDWVC 86
V+ENFD  ++ G WY + +K P      + I A +S+ E G + K  ++
Sbjct: 33  VQENFVKKYLGRWYEI-EKIPASFEKGNCIQANYSLMENGNIENVK--------ELS 82

Query: 87  ADMVGTF--------TDTEDPAKFKMKYWGVASFLQKGNDDHIVDTDYDTYAVQYSCR 137
D GT      ++ PAK ++++ +  +WIL TDY+ YA+ YSC
Sbjct: 83  PD--GTMQVKGEAKQSNVSEPACLEQVFPPILMP------PAPYWILATDYENYALVYSCT 135

Query: 138  -----LLNLGTCADSYSFVFSRPNGLPPE 163
L ++D     ++ R+P LPPE
Sbjct: 136  TFFWLFHVD--------FFWILGRNPY-LPPE 158
PSI-BLAST alignment of RBP and β-lactoglobulin: iteration 2

Score = 140 bits (353), Expect = 1e-32
Identities = 45/176 (25%), Positives = 78/176 (43%), Gaps = 33/176 (18%)

Query: 4   VWALLLLAAAWAAAERDCRVSF-----------RVKENFDKARFSGTWYAMAKKDPEGLFLQD 55
           V L+ LA A + +F             V+ENFD ++ G WY + +K P 
Sbjct: 2  VTMLMFLATLGLFTTAKGQNFHGLGKCPSSPVQENFDVKKYLGRWYEI-EKIPASFEKGN 60

Query: 56   NIVAEFSVDETGQMSATAKGRVRLLNNDVCADMV---GTFTDTEDPAKFKMKYWGVASF 112
           I A +S+ E G + K + D + V ++ +PAK +++++ +
Sbjct: 61  CIQANYSLMENGNIEVNLKEL----SPDGMNQVKGEAKQSNVSEPALKLEVQFFPL--- 112

Query: 113 LQKGNDDHWIVDTDYDTYAVQYSCR-----LLNLGTCADSYSVFVSRFDPNGLPPEA 164
            +WI+ TDY+ YA+ YSC L ++D + ++ R+P LPPE
Sbjct: 113 --MPPAPYWILATDYENYALVLYSCTTTTFWLFHVD------FFWILGRNPY-LPPET 159
PSI-BLAST alignment of RBP and $\beta$-lactoglobulin: iteration 3

Score = 159 bits (404), Expect = 1e-38
Identities = 41/170 (24%), Positives = 69/170 (40%), Gaps = 19/170 (11%)

Query: 3 WVWALLLLAAWAAAERD--------CRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQ 54
          V L+ LA A + S V+ENFD ++ G WY + K
Sbjct: 1 MVTMLMFLATLAGLFTTAKQNFHLGKCPSPPVQENFDVKYLYEIEKIPASFE-KG 59

Query: 55 DNIVAEFSVDETGQMSATAKGRVRLNNWDCADMVGTFDTDDEPAKFKMKYWGVASFLQ 114
       + I A +S+ E G + K V + ++ +PAK +++++ +
Sbjct: 60 NCIQANYSLMENGNIEVLKELSPDGTMNQVKE--AKQSNVEPAKLEVQFFPL---- 112

Query: 115 KGNDDHWIVTVDYDTVAYVQYSRCLLNLDGTCADSYSFVFSPDPNGLPPEA 164
       +WI+ TDY+ YA+ YSC + ++ R+P LPPE
Sbjct: 113 MPPAPYWILATDYEYALVYSCTTFFWL--FHVDFFWILGRNPY--LPPET 159
1

Score = 46.2 bits (108), Expect = 2e-04
Identities = 40/150 (26%), Positives = 70/150 (46%), Gaps = 37/150 (24%)

Query: 27  
VKENFDKARFSGTWYAMAKKDPEGFLQLQDNIVAEFSVEDTQGMSATAKGRVLLNNWDV 86 
+ G WY + K P + I A + S+ E G + K ++
Sbjct: 33 
VQENFVDVKKYLGRWYEIEKIPASFEKGNCIQANYSLMENGNIEVLNK----------ELS 82

Query: 87 
ADMVGTF----------TDTEDPAKFKMKYWGVASFLQKGNDDHWIVTDYDYTAVQYSCR 137 
D GT ++ PAK ++++ + +WI+ TDY+ YA+ YSC
Sbjct: 83 
PD--GMNQVKGEAKQSNVSEPAPLEVQFFPLMP-----PAPYWILATDYENYALVYSCT 135

Query: 138 ----LLNLGDCTADSYSFVFSRDPNGLPPE 163 
L ++D ++ R+P LPPE
Sbjct: 136 TFFWLFHVD-------FFWILGRNPY-LPPE 158

3

Score = 159 bits (404), Expect = 1e-38
Identities = 41/170 (24%), Positives = 69/170 (40%), Gaps = 19/170 (11%)

Query: 3  
WVWALLLLAAWAAAAERD--------CRVSSFRVKENFDKARFSGTWYAMAKKDPEGFLQ 54 
V L+ LA A + S V+ENFD ++ G WY + K
Sbjct: 1 
MVTMLMFLATLAGEFTTAKGQNFHLGKCPSPPVQENFVDVKKYLGRWYEIEKIPASFE-KG 59

Query: 55 
DNIVAEFSVEDTQGMSATAKGRVLLNNWDVCDMVGTFTTEDPAKFKMKYWGVASFLQ 114 
+ I A + S+ E G + K V ++ PAK ++++ +
Sbjct: 60 
NCIQANYSLMENGNIEVLNKELSPTDGMNQVKGE--AKQSNVSEPAPLEVQFFPL------ 112

Query: 115 
KGNDDHWIVTDYDYTAVQYSCRLLNLGDCTADSYSFVFSRDPNGLPPEA 164 
+WI+ TDY+ YA+ YSC ++ R+P LPPE
Sbjct: 113 MPPAPYWILATDYENYALVYSCTFFWL--FFWILGRNPY-LPPE 159
The universe of lipocalins (each dot is a protein)

- retinol-binding protein
- apolipoprotein D
- odorant-binding protein
Scoring matrices let you focus on the big (or small) picture.

your RBP query

retinol-binding protein
Scoring matrices let you focus on the big (or small) picture.
PSI-BLAST generates scoring matrices more powerful than PAM or BLOSUM

retinol-binding protein
PSI-BLAST: the problem of corruption

- PSI-BLAST is useful to detect weak but biologically meaningful relationships between proteins.

- The main source of false positives is the spurious amplification of sequences not related to the query. For instance, a query with a coiled-coil motif (a motif widely occurs in protein interacting DNAs) may detect thousands of other proteins with this motif that are not homologous.

- Once even a single spurious protein is included in a PSI-BLAST search above threshold, it will not go away.
PHI-BLAST: Pattern hit initiated BLAST

- Use regular expressions as a way to specify constraints in sequence similarity search.
- Given a protein sequence S and a regular expression pattern P occurring in S, PHI-BLAST helps answer the question:
  - What other protein sequences both contain P and are homologous to S in the vicinity of the pattern occurrences?
- PHI-BLAST filters out those cases where the pattern occurrence is probably random and not indicative of homology.
Align three lipocalins (RBP and two bacterial lipocalins)

1

ebcl ecblc MRLPLVAAA TAAFLVVACS SPTPPRGVT V VNNFDKRYL GTWYEIARFD 50
vc vc MRAIFLILCS V...LLNGL G..MPESVDP VSDFELNNYL GKWYEVARLD
hsrpb hsrpb ~~~MKWVWAL LLLAWAAAE RDCRVSSFRV KENFDKARFS GTWYAMAKKD

GXW[YF][EA][IVLM]
Syntax rules for PHI-BLAST

The syntax for patterns in PHI-BLAST follows the conventions of PROSITE (protein lecture, Chapter 8).

When using the stand-alone program, it is permissible to have multiple patterns. When using the Web-page only one pattern is allowed per query.

[ ] means any one of the characters enclosed in the brackets e.g., [LFYT] means one occurrence of L or F or Y or T

- means nothing (spacer character)

x(5) means 5 positions in which any residue is allowed

x(2,4) means 2 to 4 positions where any residue is allowed
Options for advanced blasting

Limit by entrez query

Composition-based statistics

Choose filter □ Low complexity □ Mask for lookup table only □ Mask lower case

Expect 10

Word Size 3

Matrix BLOSUM62 ▼ Gap Costs ▼ Existence: 11 Extension: 1

PSSM

Other advanced

PHI pattern GXW[YF][EA][IVLM]
Alignments

Significant alignments for pattern occurrence 1 at position 38

>gi|2497702|sp|Q46036|BLC_CITFR OUTER MEMBRANE LIPOPROTEIN BLC PRECURSOR
  gi|2121019|pir||I40710 outer membrane lipoprotein - Citrobacter freundii
  gi|717136|gb|AAC46456.1| (U21727) lipocalin precursor [Citrobacter freundii]
  Length = 177

  Score = 17.3 bits (52), Expect = 7e-04
  Identities = 23/82 (28%), Positives = 36/82 (43%), Gaps = 1/82 (1%)

Query:  27  VKENFDKARFSGTWYAMAKKDPDEGLFLQDNIVAEFSVTDETQMSATAKGRVRLNNWDVC  86
  pattern 38  ********
  V  NFD  R+ GTWY  +A  D  D  A  +S+  +G  ++  KG  W

Sbjct:   30  VVNNFDAKRYLGTYWEIARFDHRFERGDLKVTATYSLRDDGGINVINKGYNPDRMNWQK-  88

Query:  87  ADMVGTFDTDPAKFKMKYW 108
  +  FT  A  K+  ++G

Sbjct:   89  TEGKAYFTGDPSTAAALKVSFFG 110
BLAST for gene discovery

Start with the sequence of a known protein

Search a DNA database (e.g. HTGS, dbEST, or genomic sequence from a specific organism)

tblastn

Search your DNA or protein against a protein database (nr) to confirm you have identified a novel gene

blastx or blastp nr

Find matches…
[1] to DNA encoding known proteins
[2] to DNA encoding related (novel!) proteins
[3] to false positives

inspect
Acknowledge

- Some of the images and slides in this PowerPoint presentation are
  - http://bioinformatics.ca/workshop_pages/bioinformatics/