Practicum 1(b) WORKSHEET Due 4pm Fri, October 20

Your name:

Submit your assignment by 4pm in MI646 or electronically to comp‐bio@cs.cmu.edu.

***What to hand in:***

* A fasta formatfile, *<yourname>-prac1b.pylC.fasta*, as described in *PylC*: Part C, step (e)
* A fasta format file, <*yourname>-prac1b.pylB.fasta*, as described in *PylB* Part C, step (f)
* The completed Practicum 1b worksheet with answers to questions in the assignment.

***The PylC enzyme family***

**Part A. Use Blast to retrieve amino acid sequences for members of the 3-methylornithine--L-lysine ligase (PylC) family.**

**Give the accession id of the PylC sequence used as the query sequence in all four searches:**

a) **Summarize the conserved domains in PylC:**

* + Name:
  + Accession
  + E value
  + Coverage

1. * Name
   * Accession
   * E value
   * Coverage

*Note:* Expand this list as needed to enter all the significant matches to CDD domains found in the query sequence.

**Table 1:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Search | Database: Number of letters | Expect  value | Hitlist size | Number of hits | Least significant hit | | |
| Score | E value | Accession |
| 1 |  |  |  |  |  |  |  |
| 2 |  |  |  |  |  |  |  |
| 3 |  |  |  |  |  |  |  |
| 4 |  |  |  |  |  |  |  |

**INTERPRETATION:**

1. Compare **Search 1** and **Search 2**:
   1. What parameter values differed between these two searches?
   2. Did the number of sequences retrieved (i.e., the size of the hitlist) increase or decrease in Search 2, compared with Search 1?
   3. Did the least significant hit in Search 2 have a higher or lower bit score, compared with Search 1?
   4. Did the least significant hit in Search 2 have a more significant or less significant E value, compared with Search 1?
   5. Explain these changes in terms of what you know about the behavior of the BLAST program.
   6. Do you think that Search 2 retrieved all sequences in *refseq* that are significantly similar to the query at the current E value threshold? If so, why? If you think it is possible that Search 2 did not retrieve all significant matches, what BLAST parameters would you change to find out if there are more?
2. Compare **Search 2** and **Search 3**:
   1. What parameter values differed between these two searches?
   2. Did the number of sequences retrieved (i.e., the size of the hitlist) increase or decrease in Search 3, compared with Search 2?
   3. Did the least significant hit in Search 3 have a higher or lower bit score, compared with Search 2?
   4. Did the least significant hit in Search 3 have a more significant or less significant E value, compared with Search 2?
   5. Explain these changes in terms of what you know about the behavior of the BLAST program.
   6. Do you think that Search 3 retrieved all sequences in *refseq* that are significantly similar to the query at the current E value threshold? If so, why? If not, what BLAST parameters would you change to find out if there are more?.
3. Compare **Search 3** and **Search 4**:
   1. What parameter values differed between these two searches?
   2. Did the number of sequences retrieved (i.e., the size of the hitlist) increase or decrease in Search 4, compared with Search 3?
   3. Did the least significant hit in Search 4 have a higher or lower bit score, compared with Search 3?
   4. Did the least significant hit in Search 4 have a more significant or less significant E value, compared with Search 3?
   5. Explain these changes in terms of what you know about the behavior of the BLAST program.
4. Find the Accession of the least significant retrieved in **Search 4.** Search for that Accession in the output to **Search 3**.
   1. What is the bit score of this sequence in Search 3? Has it increased or decreased?
   2. What is the E value of that sequence in Search 3? Has it changed?
   3. Explain these changes in terms of what you know about the behavior of the BLAST program.

.

**Part B. The taxonomic distribution of all retrieved sequences.**

**Table 2:**

|  |  |  |
| --- | --- | --- |
|  | Hits | Organisms |
| Archaea (Euryarchaeota) |  |  |
| Bacteria |  |  |
| Eukaryota |  |  |

**Part C. Which of the retrieved sequences are members of the PylC family?**

**Table 3:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Subset |  | Accession | Max Score | Diff | E value | Query cover | Percent Identity |
| 1 | First |  |  |  |  |  |  |
| Last |  |  |  |  |  |  |
| 2 | First |  |  |  |  |  |  |
| Last |  |  |  |  |  |  |
| 3 | First |  |  |  |  |  |  |
| Last |  |  |  |  |  |  |
| 4 | First |  |  |  |  |  |  |
| Last |  |  |  |  |  |  |
| 5 | First |  |  |  |  |  |  |
| Last |  |  |  |  |  |  |

*Note:* You may not need every row in this table. Feel free to add more rows if you need them

* 1. **Briefly describe the specific characteristics of each of the subsets in Table 3 (one or two sentences**.)
  2. **Based on your observations, select an E value threshold that separates members of the PylC family from similar sequences in other families.**

1. Scroll to the top of the **Search 3** BLAST results tab in your browser. Enter the E value threshold you chose in the previous step into the box marked “Expect Max” in the "Formatting Options" table and click “Reformat”.

Scroll down to the Graphics Summary. **Record the number of hits after this additional E value threshold has been applied.**

Scroll down to the Descriptions section. On the left, immediately below the section labeled “Sequences producing significant alignments”, you have the option to select sequences. Click “All”. Click on the arrow immediately to the right of the word "Download". Verify that "Fasta (compete sequence)" is selected and click "Continue". Select "save file" and click "OK". **Save the file as <yourname>-prac1b.pylC.fasta and hand it in with this assignment**.

**Part D**. **The taxonomic distribution of retrieved pylC sequences**.

**Table 4**:

|  |  |  |
| --- | --- | --- |
|  | Hits | Organisms |
| Archaea (Euryarchaeota) |  |  |
| Bacteria |  |  |
| Eukaryota |  |  |

**Table 5**:

|  |  |  |  |
| --- | --- | --- | --- |
| Domain | Taxon | Hits | Organisms |
| Archaea (Eukarchaeota) |  |  |  |
|  |  |  |  |
|  |  |  |  |
| Bacteria |  |  |  |
|  |  |  |  |
|  |  |  |  |
| Eukaryota |  |  |  |
|  |  |  |  |
|  |  |  |  |

*Note:* You may not need every row in this table. Feel free to add more rows if you need them

***The PylB enzyme family***

**Part A. Retrieve amino acid sequences for all members of the methylornithine synthase (PylB) enzyme family.**

1. **Conserved domains in PylB**
   * Name:
   * Accession
   * E value
   * Coverage
2. * Name
   * Accession
   * E value
   * Coverage

*Note:* Expand this list as needed to enter all the significant matches to CDD domains found in the query sequence.

1. **Record the number of sequences retrieved in this search**

**Part B. The taxonomic distribution of all retrieved sequences.**

**Table 6:**

|  |  |  |
| --- | --- | --- |
|  | Hits | Organisms |
| Archaea (Euryarchaeota) |  |  |
| Bacteria |  |  |
| Eukaryota |  |  |

**Part C. Which of the retrieved sequences are members of the PylB family?**

1. **For PylB, how big a drop in score do you suggest for identifying breakpoints?**

**Table 7:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Subset |  | Accession | Max Score | Diff | E value | Query cover | Percent Identity |
| 1 | First |  |  |  |  |  |  |
| Last |  |  |  |  |  |  |
| 2 | First |  |  |  |  |  |  |
| Last |  |  |  |  |  |  |
| 3 | First |  |  |  |  |  |  |
| Last |  |  |  |  |  |  |
| 4 | First |  |  |  |  |  |  |
| Last |  |  |  |  |  |  |
| 5 | First |  |  |  |  |  |  |
| Last |  |  |  |  |  |  |

*Note:* You may not need every row in this table. Feel free to add more rows if you need them

* 1. **Briefly describe the specific characteristics of each of the subsets in Table 3 (one or two sentences**.)
  2. **Based on your observations, select an E value threshold that separates members of the PylB family from similar sequences in other families.**

1. Scroll to the top of the **Search 3** BLAST results tab in your browser. Enter the E value threshold you chose in the previous step into the box marked “Expect Max” in the "Formatting Options" table and click “Reformat”.

Scroll down to the Graphics Summary. **Record the number of hits after this additional E value threshold has been applied.**

Scroll down to the Descriptions section. On the left, immediately below the section labeled “Sequences producing significant alignments”, you have the option to select sequences. Click “All”. Click on the arrow immediately to the right of the word "Download". Verify that "Fasta (compete sequence)" is selected and click "Continue". Select "save file" and click "OK". **Save the file as <yourname>-prac1b.pylB.fasta and hand it in with this assignment**.

**Part D: The taxonomic distribution of retrieved pylB sequences**

**Table 8:**

|  |  |  |
| --- | --- | --- |
|  | Hits | Organisms |
| Archaea (Euryarchaeota) |  |  |
| Bacteria |  |  |
| Eukaryota |  |  |

**Table 9**:

|  |  |  |  |
| --- | --- | --- | --- |
| Domain | Taxon | Hits | Organisms |
| Archaea (Eukarchaeota) |  |  |  |
|  |  |  |  |
|  |  |  |  |
| Bacteria |  |  |  |
|  |  |  |  |
|  |  |  |  |
| Eukaryota |  |  |  |
|  |  |  |  |
|  |  |  |  |

*Note:* You may not need every row in this table. Feel free to add more rows if you need them

**INTERPRETATION**

Compare the taxonomic distributions of sequences retrieved using PylB as a query with the distribution of sequences retrieved using PylC as a query.

1. Tables 2 and 6 reflect sequences of all types (not just within the gene family) retrieved by PylC and PylB, respectively, when "Max target sequences” is set to 1000 and “Expect threshold” is 0.01.
   1. Overall, compared with PylC, did PylB retrieve more sequences, fewer sequences, or roughly the same number (within 30%)?
   2. Compared with PylC, did PylB retrieve more sequences, fewer sequences, or roughly the same number (within 30%) within the Archaea?
   3. Compared with PylC, did PylB retrieve more sequences, fewer sequences, or roughly the same number (within 30%) in Bacterial species?
   4. Compared with PylC, did PylB retrieve more sequences, fewer sequences, or roughly the same number (within 30%) in Eukaryotes?
2. Tables 4 and 8 reflect sequences of gene family members retrieved by PylC and PylB, respectively.
   1. Overall, compared with PylC, did PylB retrieve more sequences, fewer sequences, or roughly the same number (within 30%)?
   2. Compared with PylC, did PylB retrieve more sequences, fewer sequences, or roughly the same number (within 30%) within the Archaea?
   3. Compared with PylC, did PylB retrieve more sequences, fewer sequences, or roughly the same number (within 30%) in Bacterial species?
   4. Compared with PylC, did PylB retrieve more sequences, fewer sequences, or roughly the same number (within 30%) in Eukaryotes?
3. Tables 5 and 9 show how members of the PylC and PylB family are distributed across taxa within the three taxonomic domains.
   1. Are there any taxa represented in Table 5 that are not present in Table 9, or vice versa? If so, what are they?
   2. For taxonomic groups that possess both PylB sequences and PylC sequences, list the taxa in which PylB is overrepresented compared to PylC.
   3. For taxonomic groups that possess both PylB sequences and PylC sequences, list the taxa in which PylB is underrepresented compared to PylC.
   4. For taxonomic groups that possess both PylB sequences and PylC sequences, list the taxa in which PylB and PylC occur with roughly the same frequency (within 30%).