Proposed BLAST XML changes

Summary

The BLAST XML specification was introduced around 2001 and predated some current BLAST features. One important feature is the "concatenation" of queries, meaning that the BLAST application scans the database once for multiple queries. This feature can reduce the run time of a BLAST search. It should be transparent to the user. As now implemented, this feature can changes how many (XML) documents are produced, and the number can vary from one document for the entire search to one per query. It can also vary depending upon how the results are formatted. Multiple XML documents can now be written into one output file, which is not valid XML. Items 1 and 2 (below) address this issue.

The BLAST XML specification does not list organism information in a fielded manner. Item 3 (below) addresses this.

The new ASN.1 spec, DTD, and Schema are here.

Additionally, the BLAST applications will be able to write the XML data in the JSON format, though this feature will not support the Xinclude mechanism described in item 2.

Proposed Changes

1. Rename the element "Iteration" to "BlastSearch". Under BlastOutput both BlastOutput_iterations and BlastOutput_search will be optional (though one must be chosen). Each BlastSearch will contain results for only one query and result in one XML document. Multiple iteration PSI-BLAST will still list all iterations for one query within one XML document.
2. Write out XML for each query as it is finished to a separate file. BLAST will name the files following some pattern, specified by the user. This will allow users to start processing results as a query is completed. Additionally, BLAST will also write out an Xinclude master document at the end of the run. Users may then use the Xinclude mechanism to bundle results from all queries into one XML document. Information about Xinclude is available at the W3C site. Currently, BLAST writes out XML for all queries to one file. This may also result in more than one XML document residing in the same file.
3. Remove query-ID, query-def, query-len, and query-seq from BlastOutput. Leave query-ID, query-def, and query-len in BlastSearch.
4. Separate each identifier and title into their own element for non-redundant subject entries. One sequence in a non-redundant database may be associated with multiple sequence identifiers and titles. The different identifiers may be from different databases (e.g., UniProt and RefSeq) and different organisms. Currently, the first sequence identifier has its own <Hit_id> and <Hit_accession> elements, but the rest of the entries are all concatenated into the <Hit_def> element, along with the title for the first identifier. In the future, each sequence identifier will have its own <Hit_description> element.
5. Add organism information (NCBI taxid and binomial name) for each sequence identifier in the set. Each <HitDescr> will have zero or more <TaxBlk>'s. The <TaxBlk> contains the fields <TaxBlk_taxid> and <TaxBlk_sciname>. The <TaxBlk_sciname> will list the genus and species of an organism and include strain information if available. The <TaxBlk_taxid> is simply the NCBI taxid (e.g., 9606 for human). A <HitDescr> may have multiple <TaxBlk> to accommodate the recently introduced WP proteins that may annotate a protein on different genomes with different strains or species.
6. Add cbs (composition-based statistics) and genetic-code to <Parameters>.
7. Add a series of pairs of integers indicating query regions masked. The specification uses a one-offset convention, so they will also be one-offset.
8. Add query-strand and subject-strand elements. These will only be shown for BLASTN and megaBLAST searches.
9. Show the frame element (i.e., <Hsp_query-frame> and <Hsp_hit-frame>) only for translated DNA (e.g., the query for BLASTX and the subject for TBLASTN). This also means that BLASTP and BLASTN searches will no longer show frames.
10. Add an optional subject element to support the blast2seq mode. The subject element will be a set of strings, each string listing the identifier for a subject sequence. The database element will also become optional, but either the subject or database element should be chosen.

Feedback

You are invited to provide feedback on this proposal. Please use this link.

Example XML

Below is an example of the current <Hit> element and the corresponding <Hit> element with the changes from points 4.) and 5.) included. The frame elements have also been removed from the <Hit> as this is a BLASTP search.

Current hit:

```
<Hit>
  <Hit_num>3</Hit_num>
  <Hit_id>gi|1384086|dbj|BAA08539.1|</Hit_id>
  <Hit_def>hemoglobin [Paramecium triaurelia] &gt;gi|1384087|dbj|BAA08540.1| hemoglobin [Paramecium jenningsi]</Hit_def>
  <Hit_accession>BAA08539</Hit_accession>
  <Hit_len>117</Hit_len>
  <Hit_hsps>
    <Hsp>
      <Hsp_num>1</Hsp_num>
      <Hsp_bit-score>90.8929</Hsp_bit-score>
      <Hsp_score>224</Hsp_score>
      <Hsp_evalue>2.91687e-22</Hsp_evalue>
      <Hsp_query-from>13</Hsp_query-from>
      <Hsp_query-to>128</Hsp_query-to>
      <Hsp_hit-from>1</Hsp_hit-from>
      <Hsp_hit-to>116</Hsp_hit-to>
      <Hsp_query-frame>0</Hsp_query-frame>
      <Hsp_hit-frame>0</Hsp_hit-frame>
      <Hsp_identity>41</Hsp_identity>
      <Hsp_positive>68</Hsp_positive>
      <Hsp_gaps>0</Hsp_gaps>
      <Hsp_align-len>116</Hsp_align-len>
      <Hsp_qseq>ISIYDKIGGHEAIEVVVEDFYVRVLADDQLSAFFSGTNMSRLKGKQVEFFAAALGGPEPYTGAPMKQVHQGRGITMHHFSLVAGHLADALTAAGVPSETITEILGVIAPLAVDVTS</Hsp_qseq>
      <Hsp_hseq>MTLFEQLGGEAAVTAVTTQFYANIQADATVANFFNGINMADQTNKTASFLCAALGGPKAWGGRNLKEVHANMGVTNAQFTTVIGHLRSALTSAGVAADLVEQTVAVAETVRGDVVT</Hsp_hseq>
      <Hsp_midline>+++++++GG  A+  V   FY  + AD  ++ FF+G NM+    K   F  AALGGP+ + G  +K+VH   G+T   F+ V GHL  ALT+AGV ++ + + + V   +  DV +</Hsp_midline>
    </Hsp>
  </Hit_hsps>
</Hit>
```

Proposed Hit:

```
<Hit>
  <Hit_num>3</Hit_num>
  <Hit_description>
        <HitDescr>
        <HitDescr_id>gi|1384086|dbj|BAA08539.1|</HitDescr_id>
        <HitDescr_title>hemoglobin [Paramecium triaurelia]</HitDescr_title>
        <HitDescr_accession>BAA08539</HitDescr_accession>
        <HitDescr_taxonomy>
                <TaxBlk>
                        <TaxBlk_taxid>44031</TaxBlk_taxid>
                        <TaxBlk_sciname>Paramecium triaurelia</TaxBlk_sciname>
                </TaxBlk>
        </HitDescr_taxonomy>
        </HitDescr>
        <HitDescr>
        <HitDescr_id>gi|1384087|dbj|BAA08540.1|</HitDescr_id>
        <HitDescr_title>hemoglobin [Paramecium jenningsi]</HitDescr_title>
        <HitDescr_accession>BAA08540</HitDescr_accession>
        <HitDescr_taxonomy>
                <TaxBlk>
                        <TaxBlk_taxid>44029</TaxBlk_taxid>
                        <TaxBlk_sciname>Paramecium jenningsi</TaxBlk_sciname>
                </TaxBlk>
        </HitDescr_taxonomy>
        </HitDescr>
  </Hit_description>
  <Hit_len>117</Hit_len>
  <Hit_hsps>
    <Hsp>
      <Hsp_num>1</Hsp_num>
      <Hsp_bit-score>90.8929</Hsp_bit-score>
      <Hsp_score>224</Hsp_score>
      <Hsp_evalue>2.91687e-22</Hsp_evalue>
      <Hsp_query-from>13</Hsp_query-from>
      <Hsp_query-to>128</Hsp_query-to>
      <Hsp_hit-from>1</Hsp_hit-from>
      <Hsp_hit-to>116</Hsp_hit-to>
      <Hsp_identity>41</Hsp_identity>
      <Hsp_positive>68</Hsp_positive>
      <Hsp_gaps>0</Hsp_gaps>
      <Hsp_align-len>116</Hsp_align-len>
      <Hsp_qseq>ISIYDKIGGHEAIEVVVEDFYVRVLADDQLSAFFSGTNMSRLKGKQVEFFAAALGGPEPYTGAPMKQVHQGRGITMHHFSLVAGHLADALTAAGVPSETITEILGVIAPLAVDVTS</Hsp_qseq>
      <Hsp_hseq>MTLFEQLGGEAAVTAVTTQFYANIQADATVANFFNGINMADQTNKTASFLCAALGGPKAWGGRNLKEVHANMGVTNAQFTTVIGHLRSALTSAGVAADLVEQTVAVAETVRGDVVT</Hsp_hseq>
      <Hsp_midline>+++++++GG  A+  V   FY  + AD  ++ FF+G NM+    K   F  AALGGP+ + G  +K+VH   G+T   F+ V GHL  ALT+AGV ++ + + + V   +  DV +</Hsp_midline>
    </Hsp>
  </Hit_hsps>
</Hit>
```