The file format for
X! series search engines

Introduction

The output file format used by X! series search engines is based on the XML language BIOML (www.bioml.com), which was designed to store annotation information about biopolymers. The files are in a valid XML format and they can be parsed using normal XML parsing software, such as EXPAT.

The file format makes no attempt to be concise: some of the parameters stored in the files could be determined on the fly by the data display software. The choice of which data to store and which data to recalculate was made to try to make display software as efficient as possible.

The main body of the XML is separated into a series of data objects, surrounded by `<group> </group>` tags. These group elements act as separators to break up the data into logical sections. The main types of groups are as follows:

> `<group type="model">` - these groups contain all of the information about a single peptide identification, including the original mass spectrum, histograms about the statistics of an identification, the peptide sequences that match to the spectrum and the protein sequences containing those peptides.
>
> `<group type="parameters">` - these groups contain parameters about the search performed, such as the input parameters for the search engine and the performance statistics associated with the whole search.
>
> `<group type="support">` - these groups include any histograms that are needed to describe results and they are only found inside of `<group type="model">` elements.

Header

The file header is a standard XML statement which is necessary to conform to the XML standard. The header looks like:

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="/tandem/tandem-style.xsl"?>
<bioml xmlns:GAML="http://www.bioml.com/gaml/" label="test">
```

The first line is required by the XML standard. The second line allows the specification of a stylesheet, if XLST display is to be used. The third line indicates that it is a BIOML document, with a namespace set aside for GAML tags (www.gaml.org). GAML is used to represent any histograms, such as mass spectra.

Model groups

The elements contained within `<group type="model"> </group>` tags contain all of the information about an individual identification. The group element tag looks like the following:

```
<group id="46" mh="1955.165047" z="3" expect="2.4e-005"
label="ENSP00000306469" type="model" sumI="3.15"
maxI="340.56" fI="3.44" >
```

The parameters are:

- `id` – the number associated with the mass spectrum that was identified. This number usually represents the 1-based position of this spectrum in the original data file.
- `mh` – the parent ion mass (plus a proton) from the spectrum.
- `z` – the parent ion charge from the spectrum.
- `expect` – the expectation value for the top ranked protein identified with this spectrum.
- `label` – the text from the protein sequence FASTA file description line for the top ranked protein identified
- `sumI` – the $\log_{10}$ value of the sum of all of the fragment ion intensities
- `maxI` – the maximum fragment ion intensity
- `fI` – a multiplier to convert the normalized spectrum contained in this group back to the original intensity values

The proteins identified that correspond to the spectrum associated with this group are then listed sequentially, from most confident to least confident. The protein expectation values are calculated based on all of the peptides that fit with that protein from the entire set of spectra tested. The general format is:

```
<protein>
<note></note>
<file />
<peptide>
<domain>
<aa />
</domain>
</peptide>
</protein>
```

Each model group may contain many protein elements. Each of the elements contained with a protein element contain specific information:

- `<note>` - one or more descriptive notes about the protein (usually only one derived from the FASTA file description line).
- `<file>` - the original FASTA file path name.

`<peptide>` - contains the peptide sequence of the protein

`<domain>` - describes the region of the protein's sequence that was identified.

`<aa>` - describes specific modifications to residues within a domain.

The details of each element's parameters are as follows:

1. <protein expect="-4.6" id="46.1" uid="30707" label="ENSP00000306469" sumI="3.46" >

`expect` – the log10 value of the expectation value for the protein

`id` – the identifier for this particular identification (spectrum #).(id #)

`uid` – a unique number for this protein, calculated by the search engine

`label` – the description line from the FASTA file

`sumI` – the sum of all of the fragment ions that identify this protein

2. <file type="peptide" URL="../fasta/human_e.fasta.pro"/>

`type` – peptide is the only value

`URL` – the path used to the original FASTA file

3. <peptide start="1" end="376">

`start` – the number associated with the beginning of the protein's peptide sequence.

`end` – the number associated with the end of the protein's peptide sequence

4. <domain id="46.1.1" start="97" end="114" expect="2.4e-005" mh="1954.064" delta="1.101" hyperscore="1.101" peak_count="16" pre="NELR" post="INRE" seq="VAPDEHPILLTEAPLNPK" missed_cleavages="0">

`id` – the identifier for this particular identified domain (spectrum #).(id #).(domain#)

`start` – the first residue of the domain

`end` – the last residue of the domain

`expect` – the expectation value for the peptide identification

`mh` – the calculated peptide mass + a proton

`delta` – the spectrum mh minus the calculated mh

`hyperscore` – Tandem's score for the identification

`peak_count` – the number of peaks that matched between the theoretical and the test mass spectrum

`pre` – the four residues preceding the domain

`post` – the four residues following the domain

`seq` – the sequence of the domain

missed_cleavages – the number of potential cleavage sites in this peptide sequence.

5. <aa type="C" at="247" modified="57.01" />

type – the single letter abbreviation for the modified residue
at – the residue number (in the protein’s peptide sequence)
modified – the residue mass change caused by the modification

Following the protein elements are a set of group elements containing histograms that contain supporting information relevant to the identification. These histograms are represented in GAML. The overall format is punctuated with group elements as follows:

```
<group label="supporting data" type="support">

<GAML:trace label="46.hyper" type="hyperscore expectation
function">
</GAML:trace>

<GAML:trace label="46.b" type="b ion histogram">
</GAML:trace>

<GAML:trace label="46.y" type="y ion histogram">
</GAML:trace>
</group>

<group type="support" label="fragment ion mass spectrum">

<note label="Description"></note>
<GAML:trace id="46" label="46.spectrum" type="tandem mass
spectrum">
</GAML:trace>

</group>
```

The first group (label =”supporting data”) contain a set of histograms that were calculated during the identification process.

The second group (label=”fragment ion mass spectrum”) contains the histogram representing the peaks in the mass spectrum that were actually used by the search engine to perform the identification.

GAML histograms are recorded using the standard GAML format. The data is broken up into x and y coordinates, listed as ASCII numbers separated by white space. For example:

```
<GAML:trace>
<GAML:Xdata >
<GAML:values format="ASCII" numvalues="4">
173.465 175.114 177.928 201.136
</GAML:values>
```

```
</GAML:Xdata>
<GAML:Ydata>
<GAML:values format="ASCII" numvalues="4">
2 7 2 8
</GAML:values>
</GAML:Ydata>
</GAML:trace>
```

Parameters

The group elements specified as by `type=”parameters”` contain information of interest regarding the entire identification process. Typically, there are three such groups at the end of a file.

`1. <group label="input parameters" type="parameters">`

This group contains all of the parameters used by the search engine.

`2. <group label="unused input parameters" type="parameters">`

This group contains all of the input parameters given to the search engine that it could not interpret. The values recorded here could be caused by

1. spelling errors,
2. specific interface commands that were not meant for the search engine, or
3. commands meant for other search engines in the X! series.

`3. <group label="performance parameters" type="parameters">`

This group contains information recorded by the search engine about it’s operations, such as the date and time stamp for the run, the number of identifications found and the number of spectra used.

All of the data in these groups are recorded using `<note>` elements. The format of the elements is as follows:

```
<note type="input" label="spectrum, path">t.mgf</note>
```

where:

> `type` – specifies how the information was used (`input` is reserved for use by the search engine as in input parameter),

	label – specifies a descriptive identifier for the parameter The data inside of the element (in this case “t.mgf”) is the value for the parameter named by its label. A full description of all of the input parameters available can be found at http://www.thegpm.org/TANDEM/api/index.html.