Marine Microbial Biodiversity, Bioinformatics & Biotechnology

Grant agreement n°287589

Acronym : Micro B3
Start date of project: 01/01/2012, funded for 48 month

Deliverable 5.5

Data structures and retrieval services for georeference- and sample variable-oriented ENA data access

Expected Submission Date: 31.12.2012
Actual submission Date: 19.12.2012

Lead Party for Deliverable: Guy Cochrane, EMBL-EBI, Hinxton, UK

Mail: cochrane@ebi.ac.uk **Tel.: +44-1223-492564**

Dissemination level:	
Public (PU)	X
Restricted to other programme participants (including the Commission Services) (PP)	
Restricted to a group specified by the consortium (including the Commission Services) (RE)	
Confidential, only for members of the consortium (including the Commission Services) (CO)	

The Micro B3 project is funded from the European Union's Seventh Framework Programme (Joint Call OCEAN.2011-2: Marine microbial diversity – new insights into marine ecosystems functioning and its biotechnological potential) under the grant agreement no 287589. The Micro B3 project is solely responsible for this publication. It does not represent the opinion of the EU. The EU is not responsible for any use that might be made of data appearing herein.

GENERALIST SUMMARY

Under the Micro-B3 project, EMBL-EBI's European Nucleotide Archive (ENA) has developed and launched three important new public programmatic services to improve searchability and access to marine-related sequence data. These programmatic (machine to machine) services allow software applications beyond ENA (such as MEGX/MEGDB) directly to discover and retrieve data from this comprehensive molecular resource to include in analysis and visualisation. The three services launched are:

- Discovery and retrieval of sequence data records by georeference information
- Discovery and retrieval of sequence data records by environmental description and/or political place name
- Discovery by sequence similarity of georeferenced sequence data records

Work on these services will continue to provide enhancements and extensions over 2013.

SUMMARY

Micro-B3 workpackage 5, 'Bioinformatics and Data Integration', concerns the design, implementation and presentation of modular software components that will underlie the Micro-B3 Information System (MB3-IS). An important thread within this work is the presentation of primary sequence data from data repositories to MB3-IS through the implementation of appropriate programmatic services upon the EMBL-EBI's European Nucleotide Archive (ENA). Under task 5-3, we have designed, developed and deployed three services: Discovery and retrieval of sequence data records by georeference information (RESTful), discovery and retrieval of sequence data records by environmental description and/or political place name (RESTful) and discovery by sequence similarity of georeferenced sequence data records (SOAP).

INTRODUCTION

Utility and scope of the three services

Discovery and retrieval of sequence data records by georeference information

Since the addition of specific granular georeference fields to INSDC records several years ago, the use of these annotation structures has grown in popularity, to the extent that some 16 million sequence records now carry latitude and longitude information. Querying the ENA for sequences derived from samples taken within given geographical ranges or transects provides a filter appropriate for many bioinformatic analyses.

Discovery and retrieval of sequence data records by environmental description and/or political place name

Text descriptions of environmental context provided at the time of submission of sequence data provide a valuable route for the discovery of records that may be appropriate for retrieval and inclusion in a bioinformatic analysis. In addition, political place names reported in text fields provide support for geographically-oriented queries into ENA content for those records that do not have formal georeference information.

Discovery by sequence similarity of georeferenced sequence data records

Sequence similarity search, a foundation for many bioinformatics analyses, comprises the comparison of a query sequence against a large library of pre-existing sequences. Given that only 6% of sequence-searchable ENA records (~270 million) have formal georeference information, a specialist sequence similarity search query collection has been created that excludes all non-georeferenced records so as to support those uses that require a sequence similarity output containing only georeferenced records.

An extension to D5.5

It has been possible to deliver in full the three expected services that were laid out at the time of writing the Micro-B3 Description of Work with a substantially reduced staff cost. This report describes these three services. We plan to continue to work on this task in 2013 to enhance and extend the services provided here. Our future work will include the following:

1. Investigation of the WMS and OpenSearch protocols for integration of our georeference discovery services
2. Investigation of queries that combine environmental or georeference search constraints with sequence similarity search (eg. 'return all sequence records within a 50 kilometre radius of 52.205°N 0.119°E that have sequence similarity to my query sequence')
3. geo_triangle() index function to add triangles to the existing rectangles and circles - this will provide an efficient basis for all (straight-edged) polygon look-up functions
4. Enhancement of environment description look-up, including synonym addition and ontology awareness

TECHNICAL SPECIFICATIONS

ENA warehouse RESTful services

The URL syntax for discovery and retrieval of records from the ENA warehouse is:

```
http://www.ebi.ac.uk/ena/data/warehouse/search?query=<query
string>&[domain=<domain>]
```

or

```
http://www.ebi.ac.uk/ena/data/warehouse/search?query=<query
string>&result=<result>[Pagination options][Display options][Download options]
```

By default, the whole ENA is searched using the query string. If a domain or result is specified then only this sub-section of ENA is subject to the search. Please note that the Pagination options, Display options and Download options are only supported when the result parameter is specified.

For full documentation of the ENA warehouse RESTful service, please refer to http://www.ebi.ac.uk/ena/about/browser#data_warehouse.

ENA search SOAP services

A Simple Object Access Protocol (SOAP) service for the ENA Sequence Similarity Search tool is available from the following endpoint:

```
http://www.ebi.ac.uk/ena/web-service/search/services/SearchService?wsdl
```

Discovery and retrieval of sequence data records by georeference information

Latitude and longitude

Performing queries on latitude and longitude require the use of one of six special geospatial functions. For all of these, latitude and longitude must be given as decimal degrees.

Function	Description	Parameters
geo_box1	All locations within a box defined by the lower left (SW) and upper right (NE) points.	south-west latitude, south-west longitude, north-east latitude, north-east longitude
geo_box2	All locations within a box defined by a centre point and a radius in km.	latitude, lontitude, radius (km)
geo_circ	All locations within a circle defined by a centre point and a radius in km.	latitude, longitude, radius (km)
geo_lat	All locations within a latitude range given by a latitude and a radius in km.	latitude, radius (km)
geo_north	All locations north of a given latitude (inclusive).	latitude
geo_south	All locations south of a given latitude (inclusive).	latitude

Examples of use:

1. Find all entries in and around the borders of Russia, Mongolia, China, Kyrgyzstan and Kazakhstan, within a quadrant with a south-west point of 40 N 72 E and a north-east point of 55 N 100 E:

   geo_box1(40, 72, 55, 100)

   (as URL: http://www.ebi.ac.uk/ena/data/warehouse/search?query=%22geo_box1(40,72,55,100)%22&domain=sequence.)

2. Find all entries within a 300km radius of 51.398 N, 84.675 E:

   geo_box2(51. 398, 84.675, 300)

   geo_circ(51. 398, 84.675, 300)

3. Find all entries within 1000km of the equator:

   geo_lat(0, 1000)

4. Find all entries within the Arctic Circle:

   geo_north(66.5622)

5. Find all entries within the Antarctic Circle:

   geo_south(-66.5622)

Altitude

Altitude is set in metres and can be searched against as you would for any numeric field. The operators available are: =, !=, <, <=, >, >=.

Examples of use:

1. Find all entries collected over 5000m:
   altitude > 5000

Collection date

Dates must be queried in the DD-MM-YYYY or DD-MON-YYYY format.

Examples of use:

1. Find all entries collected in 2008:
   collection_date >= 01-01-2008 and collection_date <= 31-12-2008
2. Find all entries collected since the beginning of 2010:
   collection_date >= 01-01-2010
3. Final all entries collected on 24th July, 2007:
   collection_date = 24-07-2007

Discovery and retrieval of sequence data records by environmental description and/or political place name

Isolation source

The isolation source is stored as text and can be search in a case insensitive manner, using the asterisk as a wild card. This wild card character is allowed at the start and end of the query text.

Examples of use:

1. Find all entries isolated from sea water:
   isolation_source = "*sea water*" or isolation_source = "*seawater*"
2. Find all entries isolated from faeces:
   isolation_source = "*feces*" or isolation_source = "*faeces*"

Country

The country is stored as text and can be searched in a case insensitive manner, using the asterisk as a wild card. This wild card character is allowed at the start and end of the query text.

Examples of use:

1. Find all entries from USA:
   country = “USA*”
2. Find entries from Antarctica:
   country = “antarctica*”
3. Final entries that are not from Australia:
   country != “Australia*”

Discovery by sequence similarity of georeferenced sequence data records

A search is initiated by executing the following operation :

```
<wsdl:operation name="submitSearch">
<wsdl:input message="tns:submitSearchRequest"></wsdl:input>
<wsdl:output message="tns:submitSearchResponse"></wsdl:output>
</wsdl:operation>
```

The request parameters map to the following type :

```
<xsd:complexType name="SearchParameter">
<xsd:sequence>
<xsd:element name="queryCollection" nillable="true" type="tns:QueryCollection"/>
<xsd:element name="exonerateClientParameterName" nillable="true" type="tns:Exonerate
ClientParameterName"/>
<xsd:element name="maskingOption" nillable="true" type="xsd:string"/>
<xsd:element name="sequence" type="xsd:string"/>
</xsd:sequence>
</xsd:complexType>
```

By specifying the “Georeferenced” query collection a search can be initiated using the web service that will just search georeferenced data.

Alignments which represent hits from georeferenced sequences can be retrieved using the following endpoint :

```
 <wsdl:operation name="getNewAlignments">
<wsdl:input message="tns:getNewAlignmentsRequest"></wsdl:input>
<wsdl:output message="tns:getNewAlignmentsResponse"></wsdl:output>
</wsdl:operation>
```