PhyloclassTalk v 1.0

User's Guide

Hernán Morales

26/11/2015

Contents

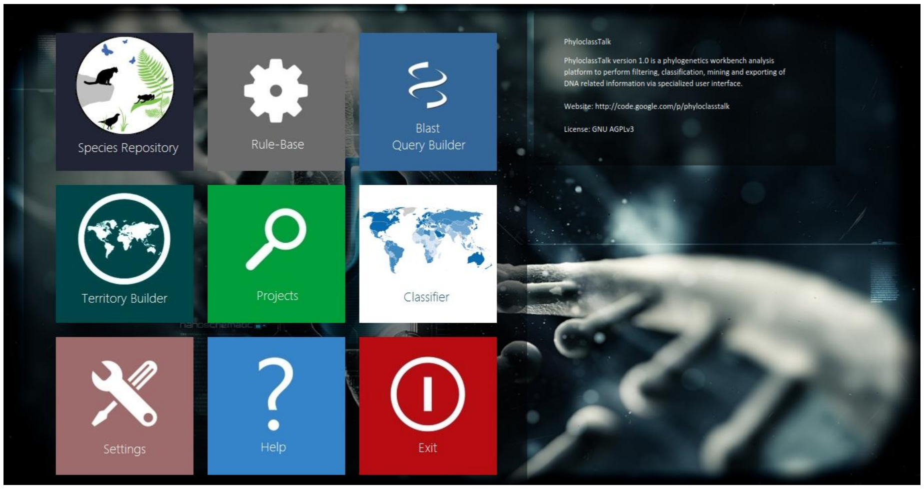
Introduction

PhyloclassTalk is an open source bioinformatics platform that aims to help in phylogenetics workflows. PhyloclassTalk can provide information about wealth and geographical distribution of studied species. Its main features are:

- A novel graphical user-interface to build BLAST queries (Blast Query Builder).
- A territorial builder framework based in a new geopolitical library (Territorial).
- A species repository module enabling species name search and selection for later recognition.
- Geographical classification of retrieved sequences through feature extraction and text-mining from the GenBank database.

PhyloclassTalk is developed using the following technologies:

- Pharo Smalltalk (http://www.pharo.org)
- Spec UI Framework (http://spec.st)
- Roassal Visualization Engine (http://objectprofile.com)
- Fuel Serialization Engine
- BioSmalltalk (http://biosmalltalk.github.io/web/), a pure object bioinformatics library.



PhyloclassTalk

PhyloclassTalk version 1.0 is a phylogenetics workbench analysis platform to perform filtering, classification, mining and exporting of DNA related information via specialized user interface.

Website: http://code.google.com/p/phyloclasstalk

License: GNU AGPLv3

Figure 1: PhyloclassTalk Main Window

Platforms

PhyloclassTalk is distributed as a desktop application, developped, maintained and tested on Pharo Smalltalk (http://pharo.org/), and therefore it should be able to execute under Microsoft Windows, GNU/Linux and MacOS X platforms. To access latest fixes and features, click in the Configuration icon and the button “Update PhyloclassTalk” from the settings list. PhyloclassTalk has been developed and tested in the following Operating Systems

- Microsoft Windows XP (SP2, SP3), 7, 8.0, 8.1
- GNU/Linux CentOS 6, 7
- GNU/Linux Debian

Privacy

Project data and user working data sets are not sent to any on-line server.

Workflows

The PhyloclassTalk application’s architecture and behavior can be broken up into two separate function-groups: Configuring and running a classifier over downloaded information, and curating the classified results while updating dictionaries. These tasks will be covered in more detail later in this document. PhyloclassTalk enables you to easily filter BLAST hits, associate sequences with GenBank annotations, and perform classifications on these sequences.

Blast Query Builder

The Blast Query Builder is an user interface to create and execute queries dynamically against a blast result dataset previously downloaded in XML format. Such input file can be obtained from the NCBI Blast Web Site, exporting the resulting alignment in XML.

Loading a Data set

To build a query first load XML results exported from the NCBI website. Select File -> Open BLAST XML from the main menu.

If Blast XML results loading is successful, you should see an information dialog as confirmation. A basic validation check is performed to assure a Blast XML file was actually selected. However, XML files which are truncated or not in the expected format (NCBI's BLAST XML) will not be detected until query execution. Blast Query Builder validation process only reads the first bytes of the loaded file, but does not process the entire file until necessary.

You must create and execute at least an initial query filter to work with the results.

> Note: Sample XML data sets are included in the PhyloclassTalk files sub-directory.

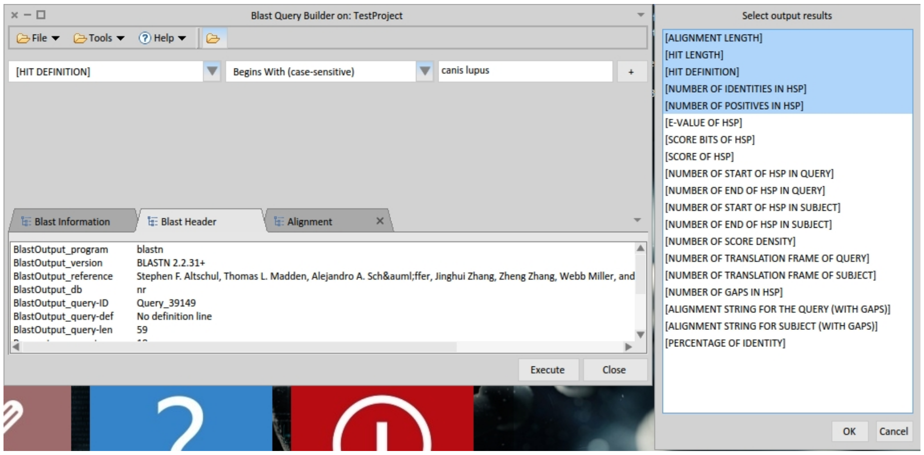
Building the query

The BLAST query builder works by selecting BLAST properties of interest along with an operator an a query value. Such "triplet" forms a filter. A query cannot be executed until at least a filter is provided. The steps for creating and executing a query are:

- Select property, operator and value to create a filter: Oerators are associated with the property selected. Selecting a numeric property (like "Alignment Length") will present corresponding comparators in the operator list. Selecting a string property (like "Hit Definition") will present appropriate operations for comparing Strings.
- Hit "+" button to add a new filter (optional).
- Click the "Execute" button.
- Select nodes of interest in the resulting palette: Blast hits information is taken from the DTD definition at http://www.ncbi.nlm.nih.gov/dtd/NCBI_BlastOutput.mod.dtd

Supported properties of BLAST results are:

- Alignment Length: Length of the alignment used



[HIT DEFINITION] | Begins With (case-sensitive) | canis lupus

```
BlastOutput_program     blastn
BlastOutput_version     BLASTN 2.2.31+
BlastOutput_reference   Stephen F. Altschul, Thomas L. Madden, Alejandro A. Sch&auml;ffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and
BlastOutput_db          nr
BlastOutput_query-ID    Query_39149
BlastOutput_query-def   No definition line
BlastOutput_query-len   59
```

[ALIGNMENT LENGTH]
[HIT LENGTH]
[HIT DEFINITION]
[NUMBER OF IDENTITIES IN HSP]
[NUMBER OF POSITIVES IN HSP]
[E-VALUE OF HSP]
[SCORE BITS OF HSP]
[SCORE OF HSP]
[NUMBER OF START OF HSP IN QUERY]
[NUMBER OF END OF HSP IN QUERY]
[NUMBER OF START OF HSP IN SUBJECT]
[NUMBER OF END OF HSP IN SUBJECT]
[NUMBER OF SCORE DENSITY]
[NUMBER OF TRANSLATION FRAME OF QUERY]
[NUMBER OF TRANSLATION FRAME OF SUBJECT]
[NUMBER OF GAPS IN HSP]
[ALIGNMENT STRING FOR THE QUERY (WITH GAPS)]
[ALIGNMENT STRING FOR SUBJECT (WITH GAPS)]
[PERCENTAGE OF IDENTITY]

Figure 2: Blast Query Results Selection

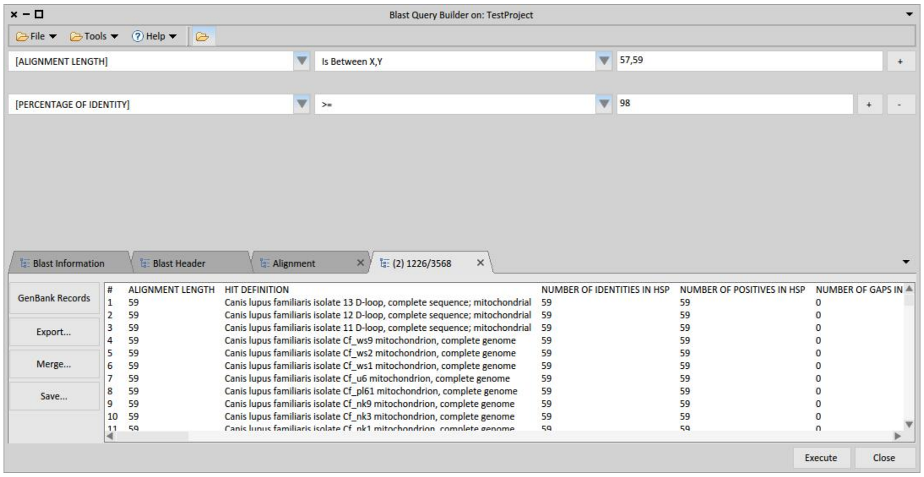
- Hit Definition: Scan the definition line for subject sequence.
- Hit Length: Length of the subject sequence.
- Number of Identities in HSP (high-scoring segment pair)
- Number of Positives in HSP
- E-Value of HSP: Expect value of the HSP
- Score Bits of HSP: A Bit-Score is a normalized log-scaled version of a score expressed in bits. From the NCBI web site: *The bit score, S', is derived from the raw alignment score, S, taking the statistical properties of the scoring system into account. Because bit scores are normalized with respect to the scoring system, they can be used to compare alignment scores from different searches.*
- Score of HSP: The HSP score represents the overall quality of the alignment. Higher numbers correspond to higher similarity. The score scale depends on the scoring system used (substitution matrix, gap penalty)
- Number of Start of HSP in Query
- Number of End of HSP in Query
- Number of Start of HSP in Subject
- Number of End of HSP in Subject
- Number of Score Density
- Number of Translation Frame of Query
- Number of Translation Frame in Subject
- Number of Gaps in HSP
- Alignment String for the Query (with gaps)
- Alignment String for the Subject (with gaps)

Executing Queries

Query execution works by walking the user filters and executing each one in turn. Execution can be cumulative (default behavior) or non-cumulative. In the cumulative configuration, each filter is executed on the previous filtered results. Non-cumulative execution generates a result set for each filter, which is useful to compare filters performance.

Cumulative execution and other useful settings can be changed by selecting Tools -> Options from the toolbar menu.

Each execution result is displayed in a new palette in the Results pane, however it is not automatically added to the user project. To add a blast results in a palette click in the Save button. All saved BLAST filtered results can be accessed through Tools -> Saved Blast Results window.



[ALIGNMENT LENGTH]	Is Between X,Y	57,59
[PERCENTAGE OF IDENTITY]	>=	98

#	ALIGNMENT LENGTH	HIT DEFINITION	NUMBER OF IDENTITIES IN HSP	NUMBER OF POSITIVES IN HSP	NUMBER OF GAPS IN
1	59	Canis lupus familiaris isolate 13 D-loop, complete sequence; mitochondrial	59	59	0
2	59	Canis lupus familiaris isolate 12 D-loop, complete sequence; mitochondrial	59	59	0
3	59	Canis lupus familiaris isolate 11 D-loop, complete sequence; mitochondrial	59	59	0
4	59	Canis lupus familiaris isolate Cf_ws9 mitochondrion, complete genome	59	59	0
5	59	Canis lupus familiaris isolate Cf_ws2 mitochondrion, complete genome	59	59	0
6	59	Canis lupus familiaris isolate Cf_ws1 mitochondrion, complete genome	59	59	0
7	59	Canis lupus familiaris isolate Cf_u6 mitochondrion, complete genome	59	59	0
8	59	Canis lupus familiaris isolate Cf_pl61 mitochondrion, complete genome	59	59	0
9	59	Canis lupus familiaris isolate Cf_nk9 mitochondrion, complete genome	59	59	0
10	59	Canis lupus familiaris isolate Cf_nk3 mitochondrion, complete genome	59	59	0
11	59	Canis lupus familiaris isolate Cf_nk1 mitochondrion, complete genome	59	59	0

Figure 3: Blast Query Results

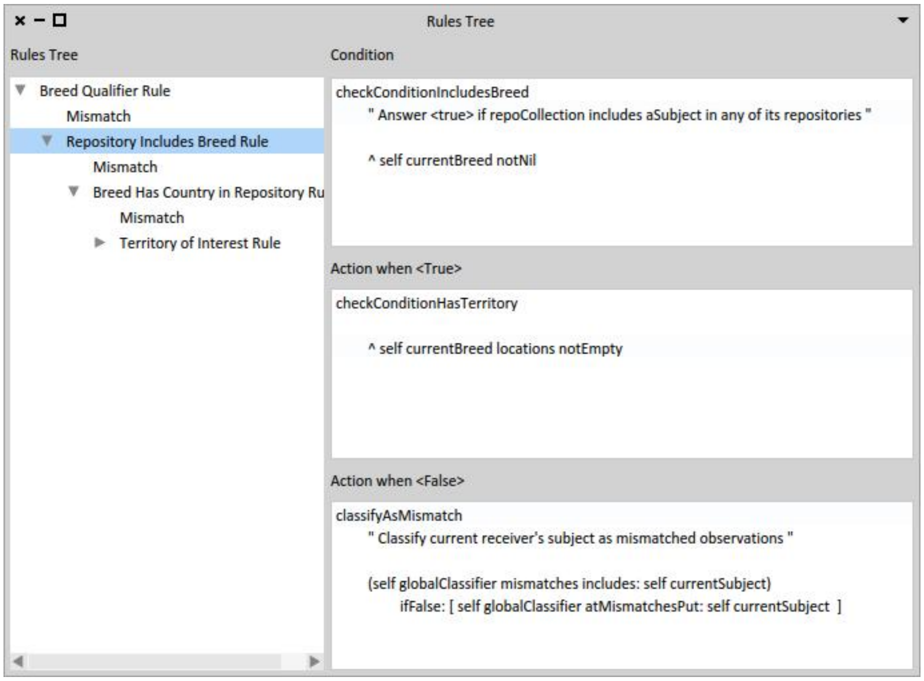
> Note: If you do not save your filtered results clicking the Save button, they will be lost if you close the Blast Query Builder window.

Exporting Results

Results can be saved and exported to FASTA and CSV formats. Exporting as CSV report includes all the available information previously selected in the resulting nodes window. To export to FASTA format, select, in the output result window, both the HIT_DEFINITION and the aligned sequences items.

Rule-Base

The Rule-Base Module is an UI for browsing and editing rules in a hierarchical fashion. Rules are displayed as a binary tree enabling to explore how the classification algorithm is evaluated. Each selected rule is displayed along with the Smalltalk code which contains its condition, the true and false branches.


Rules Tree
Rules Tree
Breed Qualifier Rule
Mismatch
Repository Includes Breed Rule
Mismatch
Breed Has Country in Repository Ru
Mismatch
Territory of Interest Rule
Condition

```
checkConditionIncludesBreed
	" Answer <true> if repoCollection includes aSubject in any of its repositories "

	^ self currentBreed notNil
```

```
checkConditionHasTerritory

	^ self currentBreed locations notEmpty
```

```
classifyAsMismatch
	" Classify current receiver's subject as mismatched observations "

	(self globalClassifier mismatches includes: self currentSubject)
		ifFalse: [ self globalClassifier atMismatchesPut: self currentSubject  ]
```

Figure 4: PhyloclassTalk Rules Tree

Species Repository

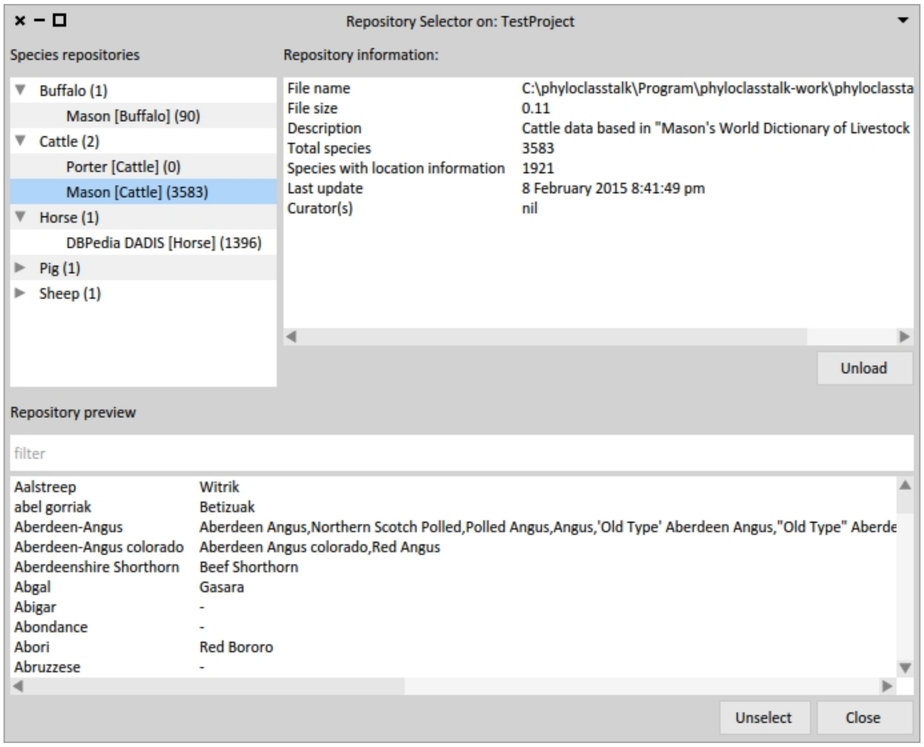
Browsing Repository Data

The species repository contains curated dictionaries of species names, synonyms, locations and additional information like date of extinction (if extinct). Each species is shown in a tree, and could contain several repositories. Our species model does not integrate species data in an unique dictionary. This is done on purpose to be able to measure performance of specific dictionaries, and because different species/pedigrees dictionaries have different models (and would require an ontology disambiguation). The classifier module allows to combine different dictionaries but this feature is experimental and currently has not been tested.

A species item is displayed along with the number of repositories currently imported. Opening a species tree item will show a list of repositories for that species, with the corresponding size of the dictionary.

The repository information panel displays metadata about the currently selected item in the species repository tree.

The repository preview panel displays the selected dictionary, with the possibility of filtering by specific term.



File name	C:\phyloclasstalk\Program\phyloclasstalk-work\phyloclassta
File size	0.11
Description	Cattle data based in "Mason's World Dictionary of Livestock
Total species	3583
Species with location information	1921
Last update	8 February 2015 8:41:49 pm
Curator(s)	nil

Aalstreep	Witrik
abel gorriak	Betizuak
Aberdeen-Angus	Aberdeen Angus,Northern Scotch Polled,Polled Angus,Angus,'Old Type' Aberdeen Angus,"Old Type" Aberde
Aberdeen-Angus colorado	Aberdeen Angus colorado,Red Angus
Aberdeenshire Shorthorn	Beef Shorthorn
Abgal	Gasara
Abigar	-
Abondance	-
Abori	Red Bororo
Abruzzese	-

Figure 5: Repository Selector

-

Updating Repository Data

Dictionary data can be updated from remote (through Google Drive currently) repositories automatically. If you curate a dictionary, please register in the project page to be sure your changes will be reviewed and integrated in the next software version.

A species dictionary format has a CSV format with the following specification:

- Column 1: Unique name of pedigree (mandatory) in English language.
- Column 2: Synonyms (if available) delimited by commas (,) in English language. If no synonyms are entered, this field should be empty.
- Column 3: Territory of origin in English language. If several territories are available, they must be delimited by commas (,). If no territories are entered, this field should be empty. Territory of origin could be, for example, a country name.
- Column 4: Sub-territory or sub-region of territory of origin.
- Column 5: If extinct, date of extinction.
- Column 6: Unique name of pedigree in Spanish language.
- Column 7: Unique name of pedigree in French language.

If a pedigree name or synonym is duplicated, only the first detected (from the beginning of file) will be used during a classification.

Territory Builder

Introduction

The Territory Builder supports building custom geopolitical territories for population studies. Any containment relationship could be defined, for example, you may want to define your territory of interest by creating artificial groupings like “Developed Countries” and “Developing Countries”.

The usage of the Territory Builder is driven by a user-interface (UI) wizard. You can open the Territory Builder wizard by clicking the “Territory Builder” icon in the main application screen. When the process finishes, an “Explorer” window is opened. In the Explorer you may see the properties in your new territory, and save it to a file.

Creating a new territory

To create a new territory, select the “New Territory” option in the wizard, and click the Next button. A territory name could be composed of any character supported in the host file system, this usually means that characters as , / are not allowed. The entered territory name will be used as the file name when saving the territory to a file. A territory is a file with .tty extension.

A new territory could be “Composite” or “Simple”:

- A “Simple Territory” consist in a single atomic location, which will not be subdivided logically in your study. This means it could be a Country or even a Continent, but it will not be composed of other sublocations.

- A “Composite Territory” is composed by one or several Simple or Composite territories itself.

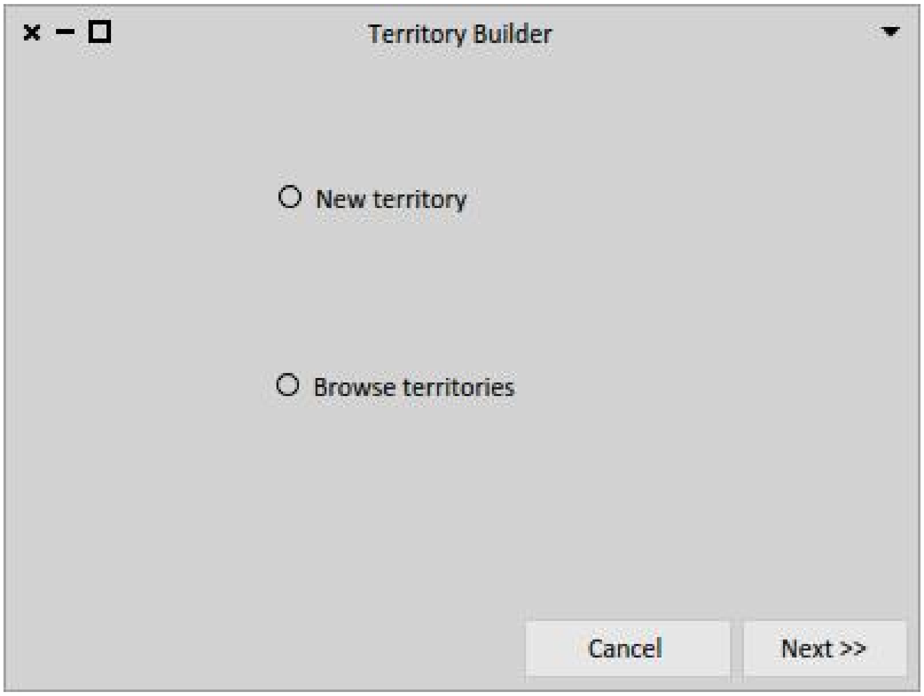


Figure 6: Territory Builder

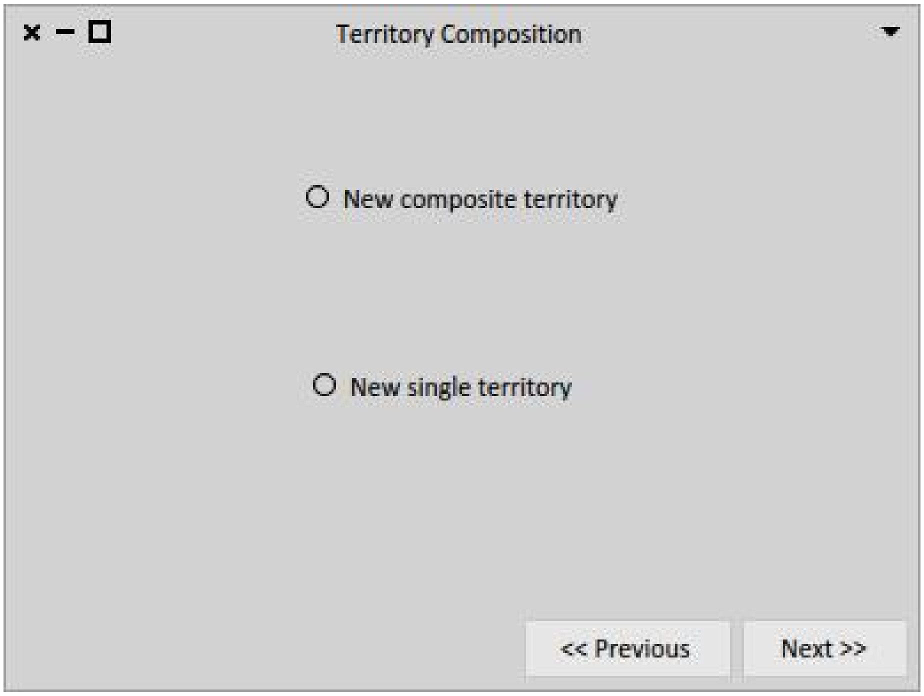
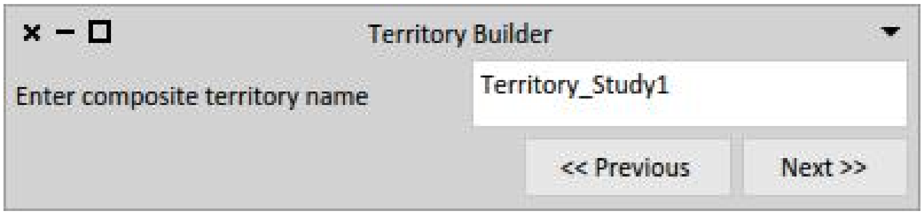


Figure 7: Territory Composition

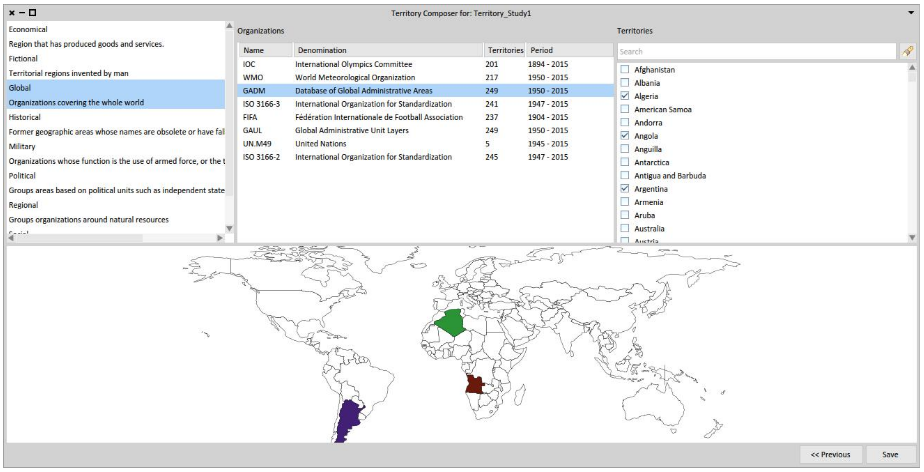
Enter a territory name, and hit Next.



Enter composite territory name: Territory_Study1

Figure 8: Territory Name

Next window will allow to compose territories based on different types of common organizations in the world. New territory groups can be created by using the Territorial Application Programming Interface.



Economical
Region that has produced goods and services.
Fictional
Territorial regions invented by man
Global
Organizations covering the whole world
Historical
Former geographic areas whose names are obsolete or have fal
Military
Organizations whose function is the use of armed force, or the t
Political
Groups areas based on political units such as independent state
Regional
Groups organizations around natural resources

Name	Denomination	Territories	Period
IOC	International Olympics Committee	201	1894 - 2015
WMO	World Meteorological Organization	217	1950 - 2015
GADM	Database of Global Administrative Areas	249	1950 - 2015
ISO 3166-3	International Organization for Standardization	241	1947 - 2015
FIFA	Fédération Internationale de Football Association	237	1904 - 2015
GAUL	Global Administrative Unit Layers	249	1950 - 2015
UN.M49	United Nations	5	1945 - 2015
ISO 3166-2	International Organization for Standardization	245	1947 - 2015

- □ Afghanistan
- □ Albania
- ☑ Algeria
- □ American Samoa
- □ Andorra
- ☑ Angola
- □ Anguilla
- □ Antarctica
- □ Antigua and Barbuda
- ☑ Argentina
- □ Armenia
- □ Aruba
- □ Australia

Figure 9: Territory Name

Once finished, the wizard will popup an Explorer window to review or make modifications to the new territory. To view the possible operations over the territory, select the top node (identified by the Territory name) and right-click to bring the contextual menu.

To add territories to a "Composite Territory", select the option "Add territory…" in the menu item.

Open territory files

A territory is a file with .tty extension. You can open a previously created territory by:

- Click in the Territory Builder icon and select Territory Viewer. A window with list of saved territories is displayed at the top. The bottom of the window displays a map for the currently selected territory.
- Open the "World menu", then Tools, and File Browser. Navigate to your project folder, and selecting the territory file. Once found, select the .tty file and then right-click to bring the contextual menu. Select the menu item "BioSmalltalk: Materialize". An explorer window will open displaying the territory containment structure.

Classifier

Introduction

The PhyloclassTalk Classifier is a frendly user-interface window to perform rule-based classification of observations (or instances). Currently observations are based in GenBank Records but other type of records (for example, EMBL) could be possible, although not currently implemented.

The classifier window is divided in three main panels. Left-most panel - named Classifier Parameters - contains the classifier settings stored in the current project. Both the middle and right-panels contains the classification results once finished.

Classifier Parameters

The Classifier Parameters includes all necessary settings to change the Classifier behavior. All settings are required to have non-empty values.

Classifier Rules

Rules are the core of the classifier, they define the conditions and actions to be applied to the project observations. Such rules can be configured from the Edit button in the Classifier Parameters panel, and different rules can be selected in the drop list. The currently selected rule name correspond to a *Smalltalk class name* which is an entry point for specifying the actual rule code.

PhyloclassTalk includes two pre-built rules:

- **PCTBreedRuleTree**: Contains rules for recognizing pedigree names, including synonyms.
- **PCTTerritorialRuleTree**: Contains rules for recognizing territorial names, including demonyms and synonyms.

Observations

Observations are the user instances to be classified. Observations can be filtered by clicking the Change… button in the Observations field. The current implementation supports browsing GenBank Records. We refer to the GenBank Browser Help to learn how to use the UI.

The Selected Species settings is intended to display a list of repositories selected in the current project. The Selected Species UI is described in the Species Repository window.

Classification Matches

Classification Mismatches

Mismatches can be queried from the “Mismatches View” by selecting an item, a feature of interest, and launching a search provider through an external web browser. The search provider enables to narrow the query to a specific search engine. For example, selecting the feature for DOI will search using the http://doi.org/ search engine, selecting an accession number feature (GBSeq_accession-version), will launch a search through Entrez e-Utils service.

Notice the features available to select are generated dynamically from the currently available features for the selected item in the Mismatches View data grid.

GenBank Browser

The GenBank Browser allows to browse downloaded GenBank records from the NCBI GBSeq (http://www.ncbi.nlm.nih.gov/dtd/NCBI_GBSeq.dtd). From the GenBank DTD comment:

> GBSeq represents the elements in a GenBank style report of a sequence with some small additions to structure and support for protein (GenPept) versions of GenBank format as seen in Entrez. While this represents the simplification, reduction of detail, and flattening to a single sequence perspective of GenBank format (compared with the full ASN.1 or XML from which GenBank and this format is derived at NCBI), it is presented in ASN.1 or XML for automated parsing and processing. It is hoped that this compromise will be useful for those bulk processing at the GenBank format level of detail today. Since it is a compromise, a number of pragmatic decisions have been made.

The GenBank Browser works by parsing the nodes found in a downloaded GenBank XML tree. You can download such XML tree through the Blast Query

Module (each results palette includes a button to download the GenBank records for the retrieved/filtered hits).

GenBank Feature Tables

Notice that a data set will actually contain additional and/or different columns (features) names than those listed in the Profile List. This is because GenBank/EMBL/DDBJ provides a Feature Table (see http://www.insdc.org/files/feature_table.html for details) to include additional extensible features. Columns like “organism”, “isolate”, “country”, “breed”, etc. are contained in GBQualifier_name and GBQualifier_value nodes in the NCBIGBSeq DTD. It means that to view such “hidden” features, you should select GBQualifier_name and GBQualifier_value in the list of features.

The GenBank Browser contains a list of features to filter results in the Data View grid. Once a feature is selected, features (columns) are updated for the current loaded data set. This is useful to narrow further which nodes are significant for your needs by improving space and processing time.

Credits

PhyloclassTalk has been originally developed by Hernán Morales Durand and Guillermo Giovambattista. The work is sponsored by the Institute of Veterinary Genetics (IGEVET) - National Scientific and Technical Research Council (CONICET) in Argentina.

Contributing

If you have an interest in PhyloclassTalk, https://github.com/PhyloclassTalk is the place you want to go right away. The main developer of PhyloclassTalk may be reached by sending email to hernan.morales@gmail.com

Troubleshooting

Before you contact mailing list, gather the background information that you will need to describe your problem. When describing a problem, be as specific as possible and include all relevant background information. To save time, know the answers to these questions:

- What PhyloclassTalk version were you running when the problem occurred?
- Do you have logs, traces, or messages that are related to the problem?

- Can you reproduce the problem? If so, what steps do you take to reproduce it?
- Is there a workaround for the problem? If so, be prepared to describe the workaround.