Practical 14. Working with Information

Mikael Bjorklund

IBI1 Semester 2, 2019/20

Learning objective

Learn to extract information from XML file using Python

Background

In this practical, you will write python codes for one problem. You may need to look for online references to complete this assignment, which is also a good practice on how to ask right questions. You are expected to use version control systems (e.g. git) to store your codes. After this practical, you may share your code with your peers and receive peer review for improvement. Use the discussion board if you need help.

Problem to solve: Find GO terms in an XML file

The go_obo.xml file has this kind of general structure:

```
<term>
    <id>GO:XXXXXXX</id> #the X refers to some number
    <name>xxx</name>
    <def>
        <defstr>some_terms</defstr>
        <is_a>GO:XXXXXXX</is_a>
    </def>
</term>
```

You have interest in biological process called 'autophagosome'. In the xml document, if text in <defstr> contains the word 'autophagosome', we suppose this gene ontology class is related to 'autophagosome'. Therefore, find all occurrences of 'autophagosome' in the <defstr> element and return the GO id, term name and the definition string (the text within <id>, <name> and <defstr> elements).

Then you can also attempt to find the number of childNodes for each 'autophagosome' related gene ontology term you found. To do this, the <is_a>tag means subclass, e.g. condensed chromosome 'is-a' chromosome. In figure 1, suppose GO:456 is related to 'autophagosome', the number of child nodes is 6 (count until you reach the bottom = all yellow boxes).

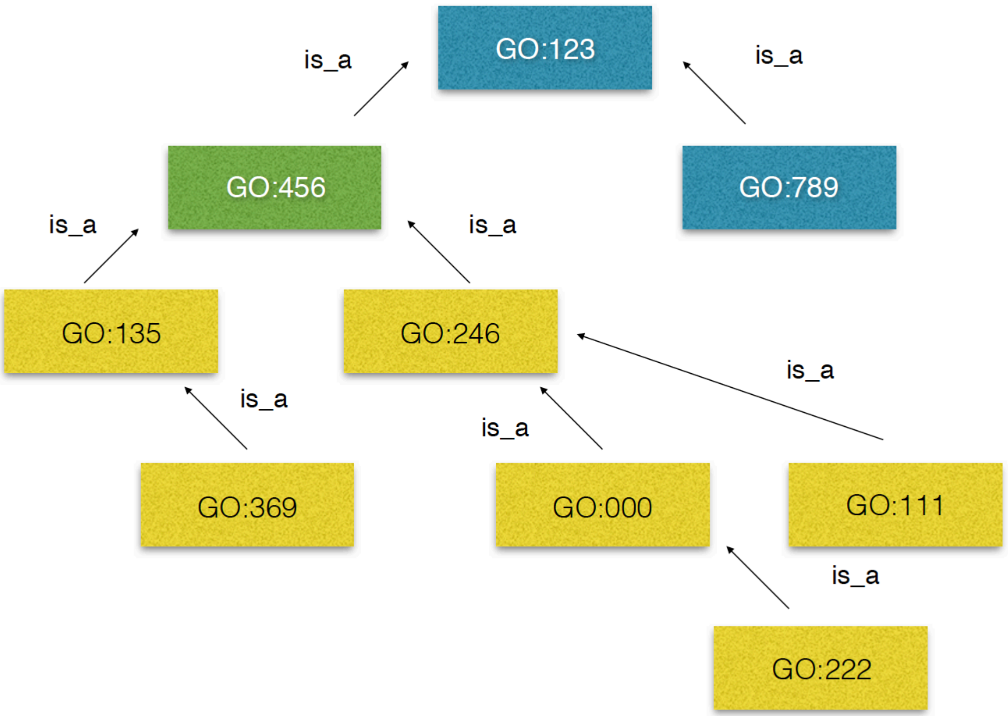


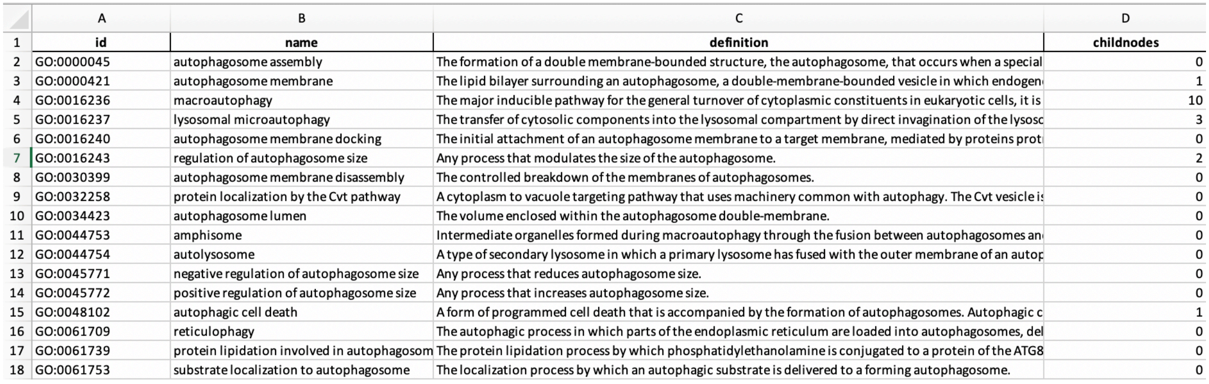
Figure 1. Example GO tree

Summary of the task

You are given: an XML document containing Gene Ontology information named 'go_obo.xml'

You should return: an Excel spreadsheet that contains: GO id, term name, definition string, number of child nodes

Expected results (beginning of the autophagosome.xlsx file)



	A	B	C	D
1	**id**	**name**	**definition**	**childnodes**
2	GO:0000045	autophagosome assembly	The formation of a double membrane-bounded structure, the autophagosome, that occurs when a special	0
3	GO:0000421	autophagosome membrane	The lipid bilayer surrounding an autophagosome, a double-membrane-bounded vesicle in which endogen	1
4	GO:0016236	macroautophagy	The major inducible pathway for the general turnover of cytoplasmic constituents in eukaryotic cells, it is	10
5	GO:0016237	lysosomal microautophagy	The transfer of cytosolic components into the lysosomal compartment by direct invagination of the lysosc	3
6	GO:0016240	autophagosome membrane docking	The initial attachment of an autophagosome membrane to a target membrane, mediated by proteins prot	0
7	GO:0016243	regulation of autophagosome size	Any process that modulates the size of the autophagosome.	2
8	GO:0030399	autophagosome membrane disassembly	The controlled breakdown of the membranes of autophagosomes.	0
9	GO:0032258	protein localization by the Cvt pathway	A cytoplasm to vacuole targeting pathway that uses machinery common with autophagy. The Cvt vesicle is	0
10	GO:0034423	autophagosome lumen	The volume enclosed within the autophagosome double-membrane.	0
11	GO:0044753	amphisome	Intermediate organelles formed during macroautophagy through the fusion between autophagosomes an	0
12	GO:0044754	autolysosome	A type of secondary lysosome in which a primary lysosome has fused with the outer membrane of an autop	0
13	GO:0045771	negative regulation of autophagosome size	Any process that reduces autophagosome size.	0
14	GO:0045772	positive regulation of autophagosome size	Any process that increases autophagosome size.	0
15	GO:0048102	autophagic cell death	A form of programmed cell death that is accompanied by the formation of autophagosomes. Autophagic c	1
16	GO:0061709	reticulophagy	The autophagic process in which parts of the endoplasmic reticulum are loaded into autophagosomes, del	0
17	GO:0061739	protein lipidation involved in autophagosom	The protein lipidation process by which phosphatidylethanolamine is conjugated to a protein of the ATG8	0
18	GO:0061753	substrate localization to autophagosome	The localization process by which an autophagic substrate is delivered to a forming autophagosome.	0

Tips: Use DOM and pandas.DataFrame.to_excel