Analysis and visualization of phylogenetic xml trees

Robin Paul, Jayendra Khandare and John Leite

Abstract – Biological organisms have evolved from single-celled bacteria to multicellular animals such as humans in a time-span of 4 billion years. During this evolutionary time, millions of different biological species have evolved from a single common ancestor. Such diversity in life has made the understanding and visualization of biological evolution a challenge. Phylogeny and relationship of organisms is in the form of a tree structure which can be represented in the form of an xml tree structure. Here we characterize and visualize a phylogenetic xml file by displaying its various branches using XSLT and converting part of the phylogenetic tree into SQL. Finally, we also visualize the entire phylogenetic tree using the software archaeopteryx.

Introduction

Biological evolution of living organisms has occurred over a span of four billion years giving rise to wide variety of living species. The best data structure to store and analyze this massive amount of biological diversity is in the form of an xml tree.

However, there are several challenges in the construction and visualization of an xml database. First, in many cases, the confidence in delineating a particular lineage of biological species can be difficult. Secondly, in many cases many species are either unnamed or haven't been characterized completely. Such drawbacks make storing of phylogenetic data difficult in traditional SQL databases making NoSQL databases an obvious choice.

We have made use of XSLT to parse out those branches of the XML tree which have confidence above a certain threshold at various levels of the phylogenetic tree. We have also converted part of the XML tree into SQL using a custom python script to enable us firing query in traditional relational databases.

Methods

a) Parsing of xml database with XSLT

An XSLT file was created which prints the branch length, confidence, and scientific name of a particular branch in the xml database. It then uses a series of nested for-loops to print out all the sub-branches in a tree using a threshold confidence of 80%. It prints the output in HTML tabular format, which can be easily viewed using a web browser.

b) Conversion of xml tree into SQL database

For this task, we chose Python language to write a script to perform the necessary task. We used two external Python libraries named BioPython and Psycopg2. Bio.Phylo contains a submodule 'PhyloXML' which was used to handle parsing and tokenizing the XML trees. Psycopg2 was used to connect to the PostgreSQL database. First, we tokenized the XML tree into tags using Phylo.read() function. Then we had to decide which tags are essential and omit the others, so we created a database to store those values. Finally, we connected to the PostgreSQL database and feed it the tokens we generated in the previous phase.

Commented [1]: Briefly describe your custom python code here.

Also talk about the SQL queries.

c) Visualization of phylogenetic tree using archaeopteryx

Commented [2]: Discuss any details here regarding how you generated the diagram.

Results and Discussion

a) XSLT results

b) SQL conversion and query

SQL conversion of XML proved to be a tough feat to achieve. Unlike SQL, XML is not well structured and the XML database we used proved to be impossible to work with. Most of the species identified don't have a name ascribed to them. Some species don't have a confidence level mentioned in the XML. Some have scientific names while some don't. So, to avoid a lot of empty entries we decided to lose some information like 'scientific name' which have very rare occurrence. Converting this data manually is ineffective because it will require more time. Also, if we don't drop a few details, SQL transformation is possible but it will waste a lot of space since most of the columns will be empty. This is the list of the properties which we chose to include:

```
BRANCH_NAME VARCHAR(20),
BRANCH_LENGTH FLOAT(20),
TOTAL_BRANCH_LENGTH FLOAT(20),
CONFIDENCE INTEGER,
NODE_PATH_LENGTH INTEGER,
TERMINALS_NUMBER INT,
PARENT_NODE VARCHAR(20)
```

Commented [3]: Mention the number of records converted into SQL.

Show some sample SQL queries. A screenshot would be great.

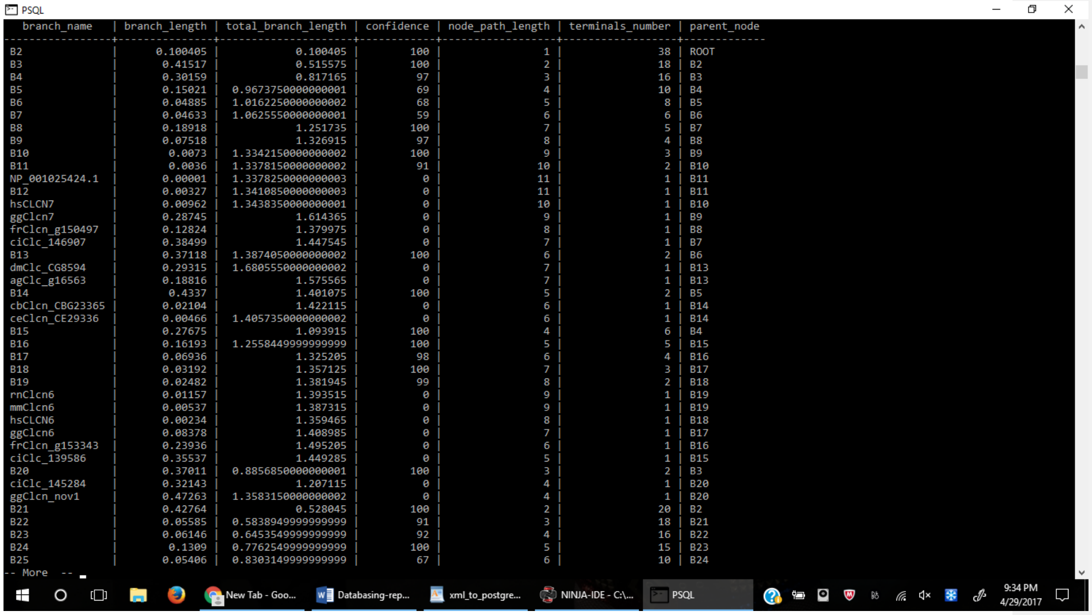
Mention that everything cannot be converted since the XML file does not have a definite schema.

If a species doesn't have a name, we denote it as "Bx" where x stands for an integer value.
We were able to extract data of 134 species from the XML and convert it into SQL format with the above approach.
The output for query "select * from phyloxml" can be seen in the snapshots given at the end of this report. Furthermore, we could fire up a query to find information with specific constraints like (confidence > 50 AND confidence < 75) or descendants of a species by firing simple SQL queries like

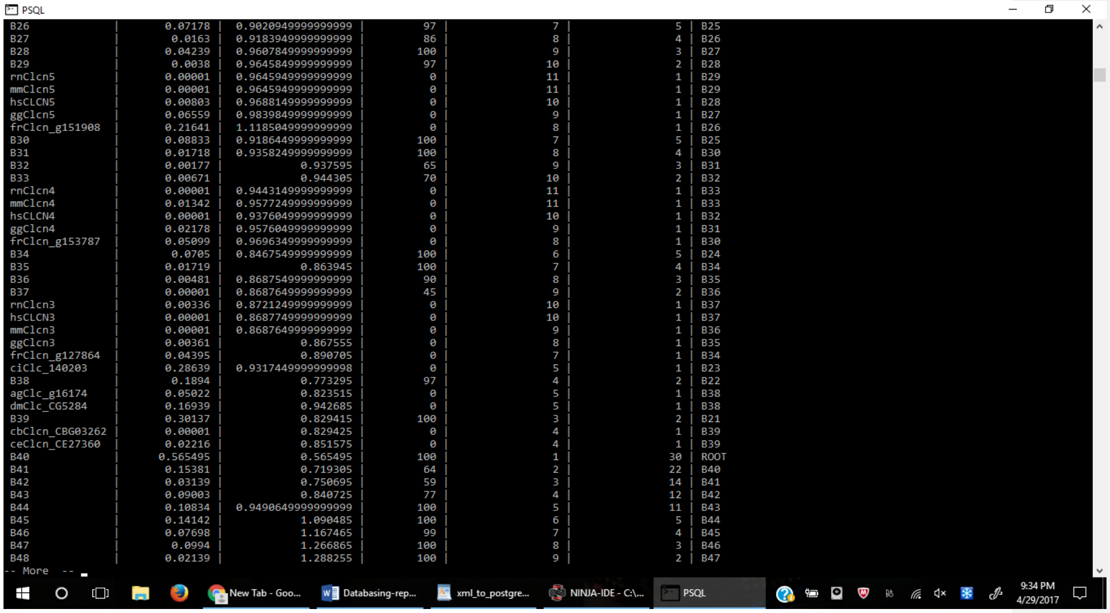
1. "select * from phyloxml
where confidence > 50 AND confidence <75;"
2. "select * from phyloxml
where parent_node = 'B40';"

c) Visualization of xml file

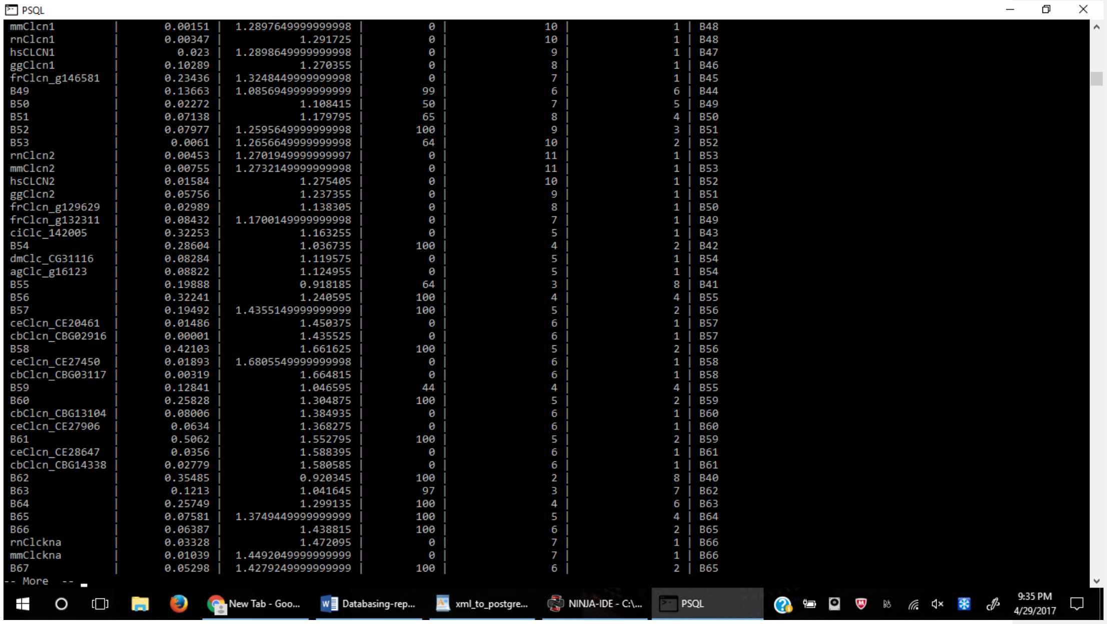
Commented [4]: Add the picture of the phylogenetic tree here. You may want to mention the total numb er of branches in the 1st, 2nd and third levels.



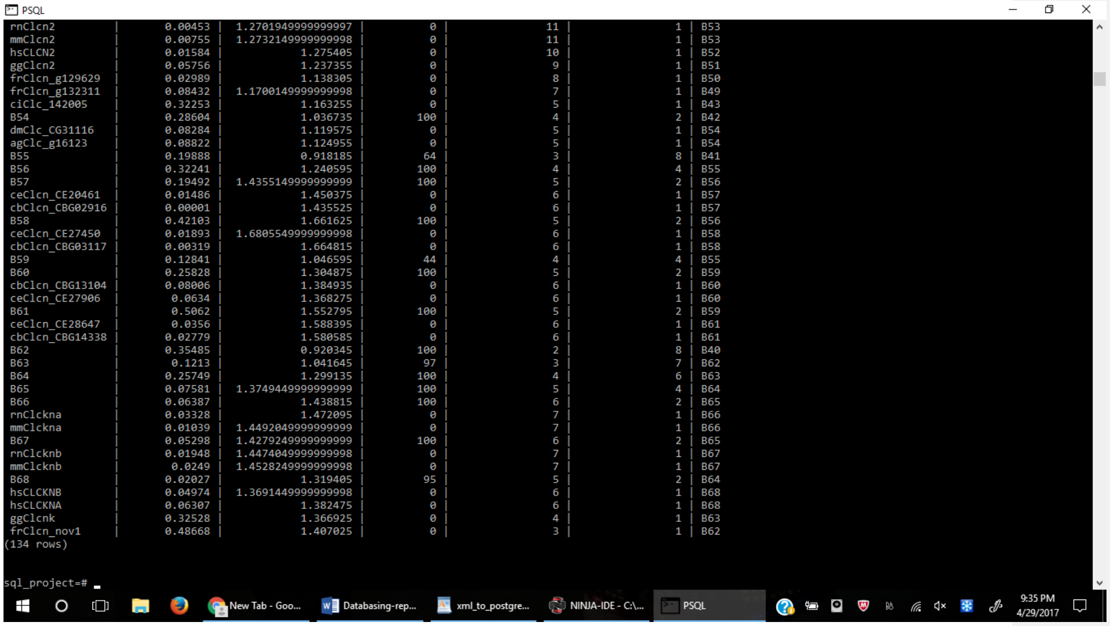
```
   branch_name    | branch_length | total_branch_length | confidence | node_path_length | terminals_number | parent_node
------------------+---------------+---------------------+------------+------------------+------------------+-------------
 B2               |      0.100405 |            0.100405 |        100 |                1 |               38 | ROOT
 B3               |       0.41517 |            0.515575 |        100 |                2 |               18 | B2
 B4               |       0.30159 |            0.817165 |         97 |                3 |               16 | B3
 B5               |       0.15021 |  0.9673750000000001 |         69 |                4 |               10 | B4
 B6               |       0.04885 |  1.0162250000000002 |         68 |                5 |                8 | B5
 B7               |       0.04633 |  1.0625550000000001 |         59 |                6 |                6 | B6
 B8               |       0.18918 |            1.251735 |        100 |                7 |                5 | B7
 B9               |       0.07518 |            1.326915 |         97 |                8 |                4 | B8
 B10              |        0.0073 |  1.3342150000000002 |        100 |                9 |                3 | B9
 B11              |        0.0036 |  1.3378150000000002 |         91 |               10 |                2 | B10
 NP_001025424.1   |       0.00001 |  1.3378250000000003 |          0 |               11 |                1 | B11
 B12              |       0.00327 |  1.3410850000000003 |          0 |               11 |                1 | B11
 hsCLCN7          |       0.00962 |  1.3438350000000001 |          0 |               10 |                1 | B10
 ggClcn7          |       0.28745 |            1.614365 |          0 |                9 |                1 | B9
 frClcn_g150497   |       0.12824 |            1.379975 |          0 |                8 |                1 | B8
 ciClc_146907     |       0.38499 |            1.447545 |          0 |                7 |                1 | B7
 B13              |       0.37118 |  1.3874050000000002 |        100 |                6 |                2 | B6
 dmClc_CG8594     |       0.29315 |  1.6805550000000002 |          0 |                7 |                1 | B13
 agClc_g16563     |       0.18816 |            1.575565 |          0 |                7 |                1 | B13
 B14              |        0.4337 |            1.401075 |        100 |                5 |                2 | B5
 cbClcn_CBG23365  |       0.02104 |            1.422115 |          0 |                6 |                1 | B14
 ceClcn_CE29336   |       0.00466 |  1.4057350000000002 |          0 |                6 |                1 | B14
 B15              |       0.27675 |            1.093915 |        100 |                4 |                6 | B4
 B16              |       0.16193 |  1.2558449999999999 |        100 |                5 |                5 | B15
 B17              |       0.06936 |            1.325205 |         98 |                6 |                4 | B16
 B18              |       0.03192 |            1.357125 |        100 |                7 |                3 | B17
 B19              |       0.02482 |            1.381945 |         99 |                8 |                2 | B18
 rnClcn6          |       0.01157 |            1.393515 |          0 |                9 |                1 | B19
 mmClcn6          |       0.00537 |            1.387315 |          0 |                9 |                1 | B19
 hsCLCN6          |       0.00234 |            1.359465 |          0 |                8 |                1 | B18
 ggClcn6          |       0.08378 |            1.408985 |          0 |                7 |                1 | B17
 frClcn_g153343   |       0.23936 |            1.495205 |          0 |                6 |                1 | B16
 ciClc_139586     |       0.35537 |            1.449285 |          0 |                5 |                1 | B15
 B20              |       0.37011 |  0.8856850000000001 |        100 |                3 |                2 | B3
 ciClc_145284     |       0.32143 |            1.207115 |          0 |                4 |                1 | B20
 ggClcn_nov1      |       0.47263 |  1.3583150000000002 |          0 |                4 |                1 | B20
 B21              |       0.42764 |            0.528045 |        100 |                2 |               20 | B2
 B22              |       0.05585 |  0.5838949999999999 |         91 |                3 |               18 | B21
 B23              |       0.06146 |  0.6453549999999999 |         92 |                4 |               16 | B22
 B24              |        0.1309 |  0.7762549999999999 |        100 |                5 |               15 | B23
 B25              |       0.05406 |  0.8303149999999999 |         67 |                6 |               10 | B24
-- More  --
```



```
 B26              |       0.07178 |  0.9020949999999999 |         97 |                7 |                5 | B25
 B27              |        0.0163 |  0.9183949999999999 |         86 |                8 |                4 | B26
 B28              |       0.04239 |  0.9607849999999999 |        100 |                9 |                3 | B27
 B29              |        0.0038 |  0.9645849999999999 |         97 |               10 |                2 | B28
 rnClcn5          |       0.00001 |  0.9645949999999999 |          0 |               11 |                1 | B29
 mmClcn5          |       0.00001 |  0.9645949999999999 |          0 |               11 |                1 | B29
 hsCLCN5          |       0.00803 |  0.9688149999999999 |          0 |               10 |                1 | B28
 ggClcn5          |       0.06559 |  0.9839849999999999 |          0 |                9 |                1 | B27
 frClcn_g151908   |       0.21641 |  1.1185049999999999 |          0 |                8 |                1 | B26
 B30              |       0.08833 |  0.9186449999999999 |        100 |                7 |                5 | B25
 B31              |       0.01718 |  0.9358249999999999 |        100 |                8 |                4 | B30
 B32              |       0.00177 |            0.937595 |         65 |                9 |                3 | B31
 B33              |       0.00671 |            0.944305 |         70 |               10 |                2 | B32
 rnClcn4          |       0.00001 |  0.9443149999999999 |          0 |               11 |                1 | B33
 mmClcn4          |       0.01342 |  0.9577249999999999 |          0 |               11 |                1 | B33
 hsCLCN4          |       0.00001 |  0.9376049999999999 |          0 |               10 |                1 | B32
 ggClcn4          |       0.02178 |  0.9576049999999999 |          0 |                9 |                1 | B31
 frClcn_g153787   |       0.05099 |  0.9696349999999999 |          0 |                8 |                1 | B30
 B34              |        0.0705 |  0.8467549999999999 |        100 |                6 |                5 | B24
 B35              |       0.01719 |            0.863945 |        100 |                7 |                4 | B34
 B36              |       0.00481 |  0.8687549999999999 |         90 |                8 |                3 | B35
 B37              |       0.00001 |  0.8687649999999999 |         45 |                9 |                2 | B36
 rnClcn3          |       0.00336 |  0.8721249999999999 |          0 |               10 |                1 | B37
 hsCLCN3          |       0.00001 |  0.8687749999999999 |          0 |               10 |                1 | B37
 mmClcn3          |       0.00001 |  0.8687649999999999 |          0 |                9 |                1 | B36
 ggClcn3          |       0.00361 |            0.867555 |          0 |                8 |                1 | B35
 frClcn_g127864   |       0.04395 |            0.890705 |          0 |                7 |                1 | B34
 ciClc_140203     |       0.28639 |  0.9317449999999998 |          0 |                5 |                1 | B23
 B38              |        0.1894 |            0.773295 |         97 |                4 |                2 | B22
 agClc_g16174     |       0.05022 |            0.823515 |          0 |                5 |                1 | B38
 dmClc_CG5284     |       0.16939 |            0.942685 |          0 |                5 |                1 | B38
 B39              |       0.30137 |            0.829415 |        100 |                3 |                2 | B21
 cbClcn_CBG03262  |       0.00001 |            0.829425 |          0 |                4 |                1 | B39
 ceClcn_CE27360   |       0.02216 |            0.851575 |          0 |                4 |                1 | B39
 B40              |      0.565495 |            0.565495 |        100 |                1 |               30 | ROOT
 B41              |       0.15381 |            0.719305 |         64 |                2 |               22 | B40
 B42              |       0.03139 |            0.750695 |         59 |                3 |               14 | B41
 B43              |       0.09003 |            0.840725 |         77 |                4 |               12 | B42
 B44              |       0.10834 |  0.9490649999999999 |        100 |                5 |               11 | B43
 B45              |       0.14142 |            1.090485 |        100 |                6 |                5 | B44
 B46              |       0.07698 |            1.167465 |         99 |                7 |                4 | B45
 B47              |        0.0994 |            1.266865 |        100 |                8 |                3 | B46
 B48              |       0.02139 |            1.288255 |        100 |                9 |                2 | B47
-- More  --
```



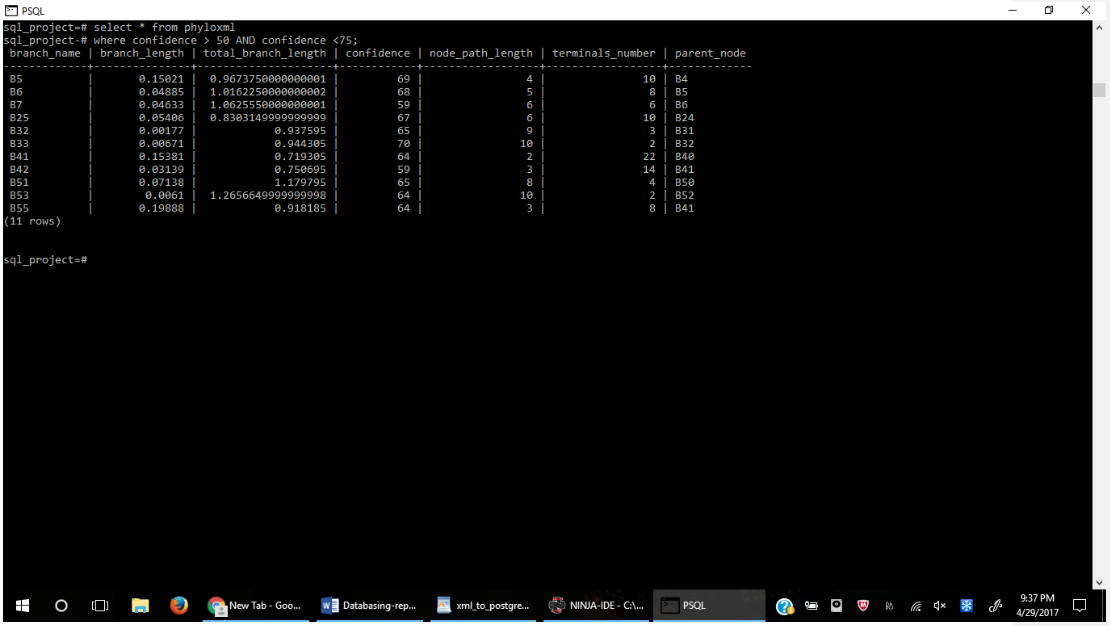
```
 mmClcn1         |     0.00151 | 1.2897649999999998 |        0 |             10 |              1 | B48
 rnClcn1         |     0.00347 |           1.291725 |        0 |             10 |              1 | B48
 hsCLCN1         |       0.023 | 1.2898649999999998 |        0 |              9 |              1 | B47
 ggClcn1         |     0.10289 |           1.270355 |        0 |              8 |              1 | B46
 frClcn_g146581  |     0.23436 | 1.3248449999999998 |        0 |              7 |              1 | B45
 B49             |     0.13663 | 1.0856949999999999 |       99 |              6 |              6 | B44
 B50             |     0.02272 |           1.108415 |       50 |              7 |              5 | B49
 B51             |     0.07138 |           1.179795 |       65 |              8 |              4 | B50
 B52             |     0.07977 | 1.2595649999999998 |      100 |              9 |              3 | B51
 B53             |      0.0061 | 1.2656649999999998 |       64 |             10 |              2 | B52
 rnClcn2         |     0.00453 | 1.2701949999999997 |        0 |             11 |              1 | B53
 mmClcn2         |     0.00755 | 1.2732149999999998 |        0 |             11 |              1 | B53
 hsCLCN2         |     0.01584 |           1.275405 |        0 |             10 |              1 | B52
 ggClcn2         |     0.05756 |           1.237355 |        0 |              9 |              1 | B51
 frClcn_g129629  |     0.02989 |           1.138305 |        0 |              8 |              1 | B50
 frClcn_g132311  |     0.08432 | 1.1700149999999998 |        0 |              7 |              1 | B49
 ciClc_142005    |     0.32253 |           1.163255 |        0 |              5 |              1 | B43
 B54             |     0.28604 |           1.036735 |      100 |              4 |              2 | B42
 dmClc_CG31116   |     0.08284 |           1.119575 |        0 |              5 |              1 | B54
 agClc_g16123    |     0.08822 |           1.124955 |        0 |              5 |              1 | B54
 B55             |     0.19888 |           0.918185 |       64 |              3 |              8 | B41
 B56             |     0.32241 |           1.240595 |      100 |              4 |              4 | B55
 B57             |     0.19492 | 1.4355149999999999 |      100 |              5 |              2 | B56
 ceClcn_CE20461  |     0.01486 |           1.450375 |        0 |              6 |              1 | B57
 cbClcn_CBG02916 |     0.00001 |           1.435525 |        0 |              6 |              1 | B57
 B58             |     0.42103 |           1.661625 |      100 |              5 |              2 | B56
 ceClcn_CE27450  |     0.01893 | 1.6805549999999998 |        0 |              6 |              1 | B58
 cbClcn_CBG03117 |     0.00319 |           1.664815 |        0 |              6 |              1 | B58
 B59             |     0.12841 |           1.046595 |       44 |              4 |              4 | B55
 B60             |     0.25828 |           1.304875 |      100 |              5 |              2 | B59
 cbClcn_CBG13104 |     0.08006 |           1.384935 |        0 |              6 |              1 | B60
 ceClcn_CE27906  |      0.0634 |           1.368275 |        0 |              6 |              1 | B60
 B61             |      0.5062 |           1.552795 |      100 |              5 |              2 | B59
 ceClcn_CE28647  |      0.0356 |           1.588395 |        0 |              6 |              1 | B61
 cbClcn_CBG14338 |     0.02779 |           1.580585 |        0 |              6 |              1 | B61
 B62             |     0.35485 |           0.920345 |      100 |              2 |              8 | B40
 B63             |      0.1213 |           1.041645 |       97 |              3 |              7 | B62
 B64             |     0.25749 |           1.299135 |      100 |              4 |              6 | B63
 B65             |     0.07581 | 1.3749449999999999 |      100 |              5 |              4 | B64
 B66             |     0.06387 |           1.438815 |      100 |              6 |              2 | B65
 rnClckna        |     0.03328 |           1.472095 |        0 |              7 |              1 | B66
 mmClckna        |     0.01039 | 1.4492049999999999 |        0 |              7 |              1 | B66
 B67             |     0.05298 | 1.4279249999999999 |      100 |              6 |              2 | B65
-- More  -- _
```



```
 rnClcn2         |     0.00453 | 1.2701949999999997 |        0 |             11 |              1 | B53
 mmClcn2         |     0.00755 | 1.2732149999999998 |        0 |             11 |              1 | B53
 hsCLCN2         |     0.01584 |           1.275405 |        0 |             10 |              1 | B52
 ggClcn2         |     0.05756 |           1.237355 |        0 |              9 |              1 | B51
 frClcn_g129629  |     0.02989 |           1.138305 |        0 |              8 |              1 | B50
 frClcn_g132311  |     0.08432 | 1.1700149999999998 |        0 |              7 |              1 | B49
 ciClc_142005    |     0.32253 |           1.163255 |        0 |              5 |              1 | B43
 B54             |     0.28604 |           1.036735 |      100 |              4 |              2 | B42
 dmClc_CG31116   |     0.08284 |           1.119575 |        0 |              5 |              1 | B54
 agClc_g16123    |     0.08822 |           1.124955 |        0 |              5 |              1 | B54
 B55             |     0.19888 |           0.918185 |       64 |              3 |              8 | B41
 B56             |     0.32241 |           1.240595 |      100 |              4 |              4 | B55
 B57             |     0.19492 | 1.4355149999999999 |      100 |              5 |              2 | B56
 ceClcn_CE20461  |     0.01486 |           1.450375 |        0 |              6 |              1 | B57
 cbClcn_CBG02916 |     0.00001 |           1.435525 |        0 |              6 |              1 | B57
 B58             |     0.42103 |           1.661625 |      100 |              5 |              2 | B56
 ceClcn_CE27450  |     0.01893 | 1.6805549999999998 |        0 |              6 |              1 | B58
 cbClcn_CBG03117 |     0.00319 |           1.664815 |        0 |              6 |              1 | B58
 B59             |     0.12841 |           1.046595 |       44 |              4 |              4 | B55
 B60             |     0.25828 |           1.304875 |      100 |              5 |              2 | B59
 cbClcn_CBG13104 |     0.08006 |           1.384935 |        0 |              6 |              1 | B60
 ceClcn_CE27906  |      0.0634 |           1.368275 |        0 |              6 |              1 | B60
 B61             |      0.5062 |           1.552795 |      100 |              5 |              2 | B59
 ceClcn_CE28647  |      0.0356 |           1.588395 |        0 |              6 |              1 | B61
 cbClcn_CBG14338 |     0.02779 |           1.580585 |        0 |              6 |              1 | B61
 B62             |     0.35485 |           0.920345 |      100 |              2 |              8 | B40
 B63             |      0.1213 |           1.041645 |       97 |              3 |              7 | B62
 B64             |     0.25749 |           1.299135 |      100 |              4 |              6 | B63
 B65             |     0.07581 | 1.3749449999999999 |      100 |              5 |              4 | B64
 B66             |     0.06387 |           1.438815 |      100 |              6 |              2 | B65
 rnClckna        |     0.03328 |           1.472095 |        0 |              7 |              1 | B66
 mmClckna        |     0.01039 | 1.4492049999999999 |        0 |              7 |              1 | B66
 B67             |     0.05298 | 1.4279249999999999 |      100 |              6 |              2 | B65
```

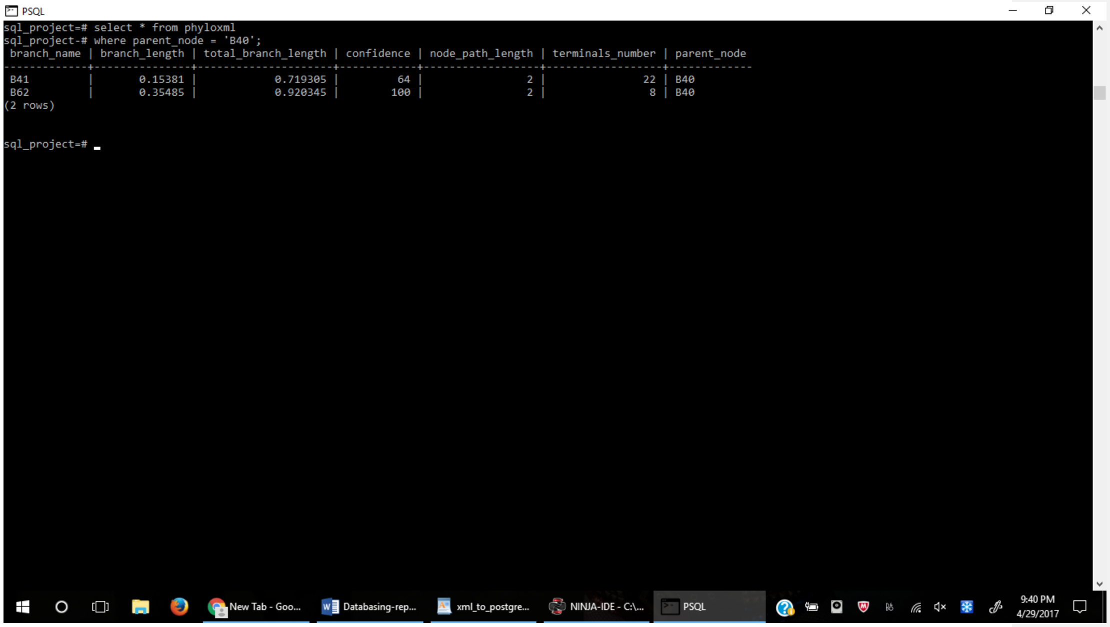
```
 rnClcknb        |     0.01948 | 1.4474049999999998 |        0 |              7 |              1 | B67
 mmClcknb        |      0.0249 | 1.4528249999999998 |        0 |              7 |              1 | B67
 B68             |     0.02027 |           1.319405 |       95 |              5 |              2 | B64
 hsCLCKNB        |     0.04974 | 1.3691449999999998 |        0 |              6 |              1 | B68
 hsCLCKNA        |     0.06307 |           1.382475 |        0 |              6 |              1 | B68
 ggClcnk         |     0.32528 |           1.366925 |        0 |              4 |              1 | B63
 frClcn_nov1     |     0.48668 |           1.407025 |        0 |              3 |              1 | B62
(134 rows)


sql_project=# _
```



```
sql_project=# select * from phyloxml
sql_project-# where confidence > 50 AND confidence <75;
 branch_name | branch_length | total_branch_length | confidence | node_path_length | terminals_number | parent_node
-------------+---------------+---------------------+------------+------------------+------------------+-------------
 B5          |       0.15021 |  0.9673750000000001 |         69 |                4 |               10 | B4
 B6          |       0.04885 |  1.0162250000000002 |         68 |                5 |                8 | B5
 B7          |       0.04633 |  1.0625550000000001 |         59 |                6 |                6 | B6
 B25         |       0.05406 |  0.8303149999999999 |         67 |                6 |               10 | B24
 B32         |       0.00177 |            0.937595 |         65 |                9 |                3 | B31
 B33         |       0.00671 |            0.944305 |         70 |               10 |                2 | B32
 B41         |       0.15381 |            0.719305 |         64 |                2 |               22 | B40
 B42         |       0.03139 |            0.750695 |         59 |                3 |               14 | B41
 B51         |       0.07138 |            1.179795 |         65 |                8 |                4 | B50
 B53         |        0.0061 |  1.2656649999999998 |         64 |               10 |                2 | B52
 B55         |       0.19888 |            0.918185 |         64 |                3 |                8 | B41
(11 rows)


sql_project=#
```



```
sql_project=# select * from phyloxml
sql_project-# where parent_node = 'B40';
 branch_name | branch_length | total_branch_length | confidence | node_path_length | terminals_number | parent_node
-------------+---------------+---------------------+------------+------------------+------------------+-------------
 B41         |       0.15381 |            0.719305 |         64 |                2 |               22 | B40
 B62         |       0.35485 |            0.920345 |        100 |                2 |                8 | B40
(2 rows)


sql_project=# _
```