

Binf 6210 Assignment 1

Introduction:

Natural environmental products have always provided a starting framework for new drug developments. The phylum of Bryozoa, found in marine habitats, has demonstrated potential for anticancer drug development and potential benefits to human health (Figuerola and Avila, 2019), (Ciavatta *et al*, 2020). Bryozoa are an aquatic organism, also known as moss animals, that consist of tiny modules, called zooids (Ciavatta, 2020). Found in greater abundance in Canada and United States. Among the marine invertebrates, Bryozoa are an important bioactive compound (Ciavatta, 2020). Currently, two-hundred and thirty compounds have been isolated from twenty-six species of Bryozoa (Ciavatta, 2020). These bioactive compounds, have demonstrated, illustrated in Ciavatta, 2020, how they can complement and outline short comings of existing cancer treatments (Ciavatta, 2020).

The significance of this organism is growing with its potential promise for enhancing cancer treatment, this project poses a research question, how many species in the phylum of Bryozoa, have been barcoded in Canada vs United States, which have contributed to enhancing cancer treatments. According to the BOLD database, bryozoan species are found in greater abundance in North America in comparison to other continents, it is predicated that bryozoan species whom have been found effective for cancer treatment will be found in higher quantities in Canada and United States in comparison to other parts of the world.

Figures:

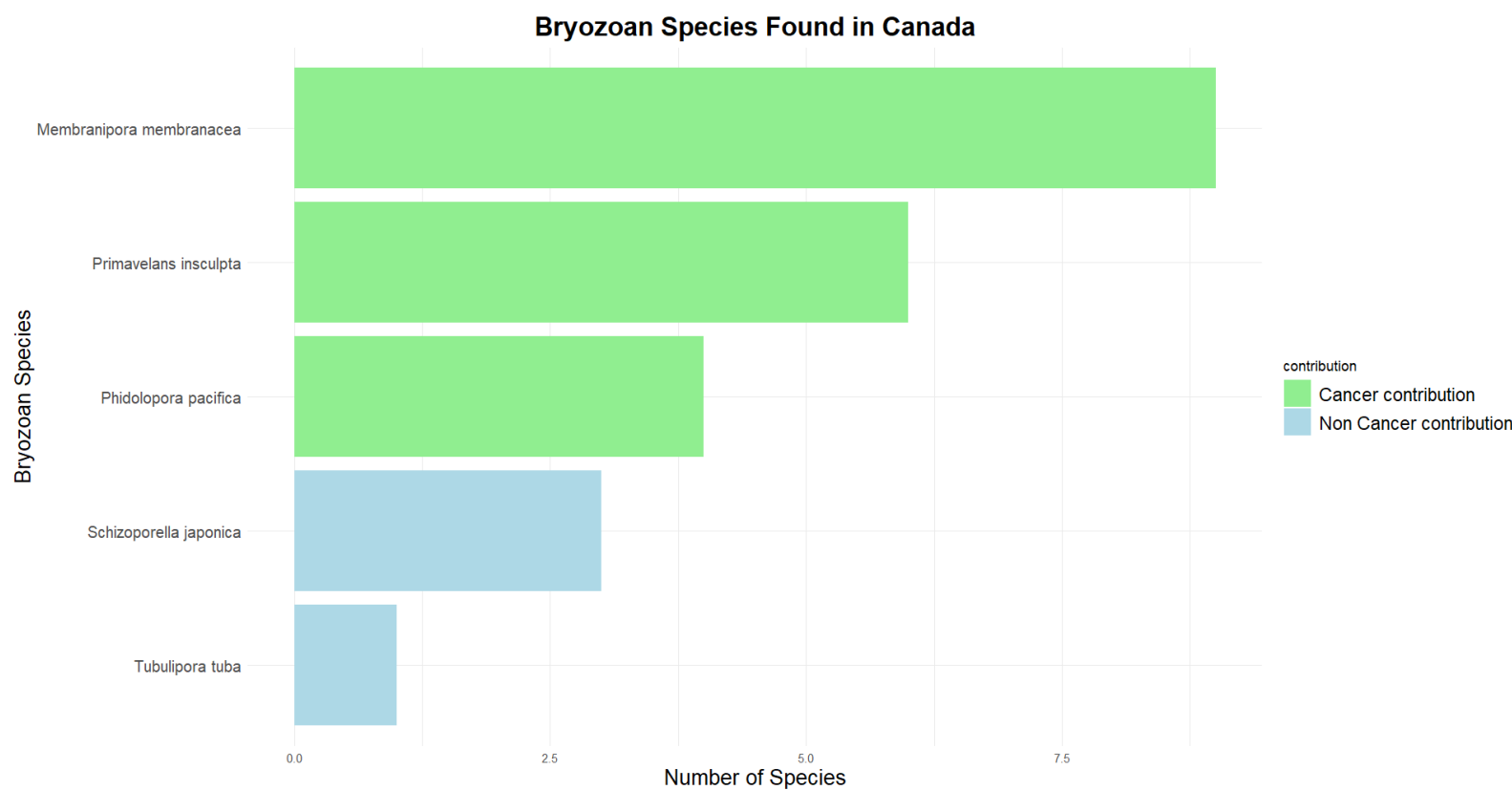


Figure 1:

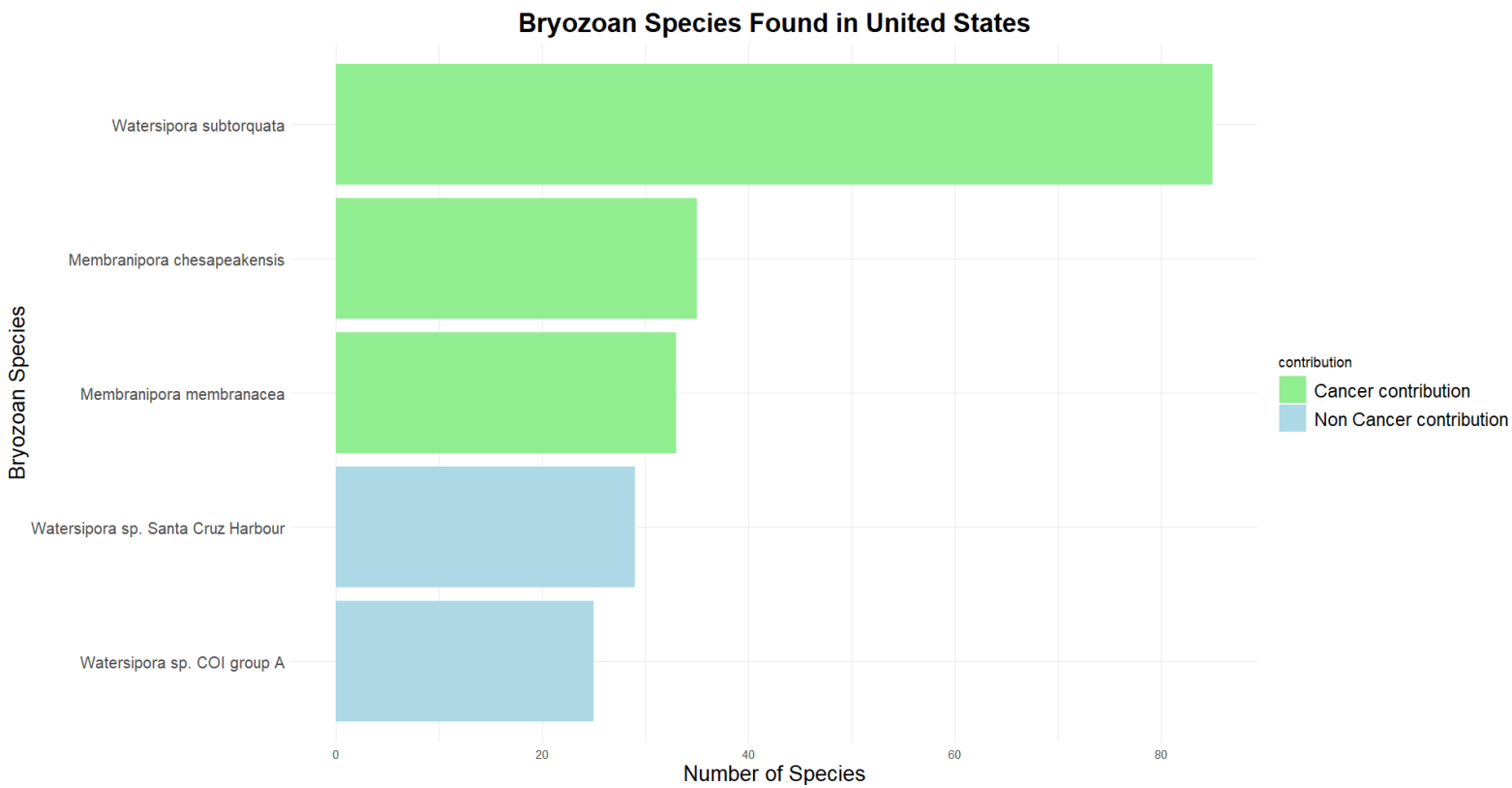


Figure 2:

Distribution of Bryozoan Species in the United States

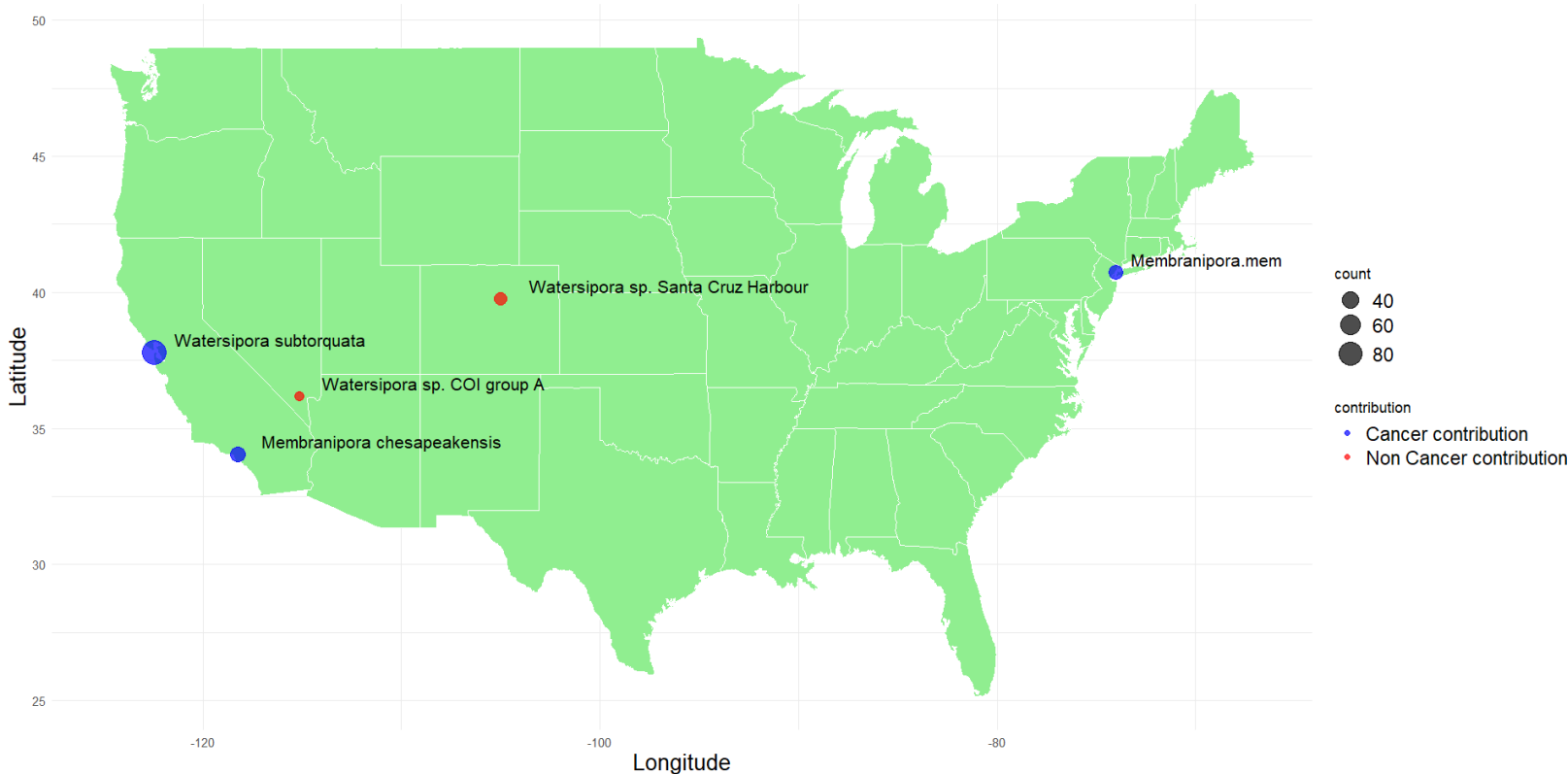
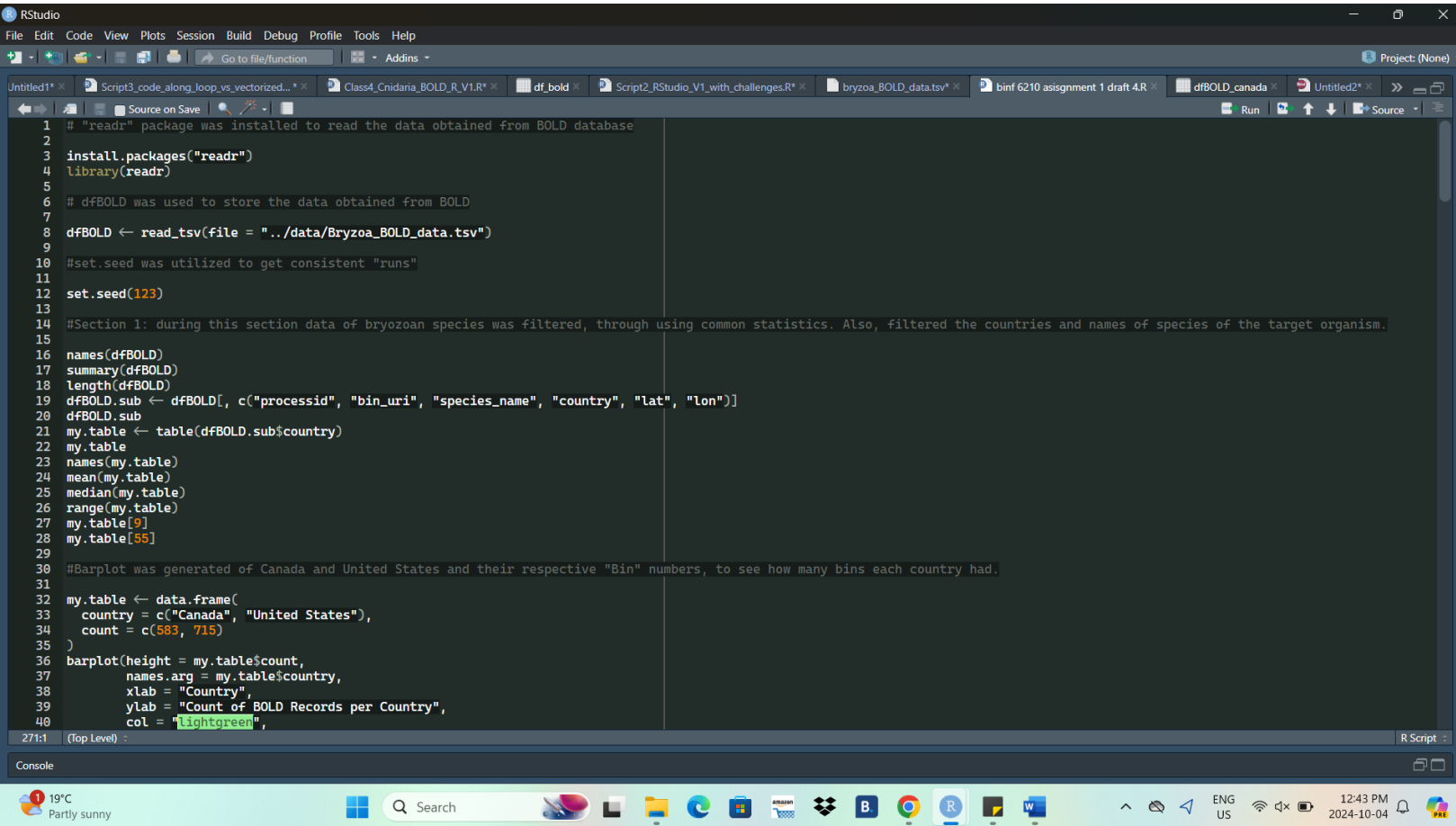


Figure 3:

Coding Section:



The screenshot displays the RStudio interface with a script editor containing R code for data processing and visualization. The code includes package installation, data reading, filtering, summarization, and a barplot.

```
1 # "readr" package was installed to read the data obtained from BOLD database
2
3 install.packages("readr")
4 library(readr)
5
6 # dfBOLD was used to store the data obtained from BOLD
7
8 dfBOLD <- read_tsv(file = "../data/Bryzoa_BOLD_data.tsv")
9
10 #set.seed was utilized to get consistent "runs"
11
12 set.seed(123)
13
14 #Section 1: during this section data of bryozoan species was filtered, through using common statistics. Also, filtered the countries and names of species of the target organism.
15
16 names(dfBOLD)
17 summary(dfBOLD)
18 length(dfBOLD)
19 dfBOLD.sub <- dfBOLD[, c("processid", "bin_uri", "species_name", "country", "lat", "lon")]
20 dfBOLD.sub
21 my.table <- table(dfBOLD.sub$country)
22 my.table
23 names(my.table)
24 mean(my.table)
25 median(my.table)
26 range(my.table)
27 my.table[9]
28 my.table[55]
29
30 #Barplot was generated of Canada and United States and their respective "Bin" numbers, to see how many bins each country had.
31
32 my.table <- data.frame(
33   country = c("Canada", "United States"),
34   count = c(583, 715)
35 )
36 barplot(height = my.table$count,
37         names.arg = my.table$country,
38         xlab = "Country",
39         ylab = "Count of BOLD Records per Country",
40         col = "lightgreen",
```

The console at the bottom shows the command prompt: 271:1 (Top Level) :

The Windows taskbar at the bottom indicates the system temperature is 19°C and it is partly sunny. The date and time are 12:43 PM, 2024-10-04.

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins Project: (None)

Untitled1* Script3_code_along_loop_vs_vectorized... Class4_Cnidaria_BOLD_R_V1.R* df_bold Script2_RStudio_V1_with_challenges.R* bryzoa_BOLD_data.tsv* binf 6210 assignment 1 draft 4.R dfBOLD_canada Untitled2*

Source on Save Run Source

```
40 col = "lightgreen",
41 main = "BOLD Records by Country")
42
43
44 #The Bryozoan species in obtained from the BOLD data base was filtered out and barplot was generated to see the names and amount of particular species in the BOLD data base. Basic
statistics were also conducted.
45
46 my.table <- table(dfBOLD.sub$species_name)
47 my.table
48 barplot(my.table,
49 main = "Bold Record of Bryozoan Species",
50 xlab = "Bryozoan Species",
51 ylab = "Number of Bold Records")
52 names(my.table)
53 mean(my.table)
54 median(my.table)
55 range(my.table)
56
57 #Section 2: During this section latitude and longitude of the data was generated,so we can filter the lat and long of Canada and United States.
58
59
60 my.table <- table(dfBOLD.sub$bin_uri)
61 my.table
62 my.table <-table(dfBOLD.sub$lat)
63 my.table
64 my.table <- table(dfBOLD.sub$lon)
65 my.table
66 library(dplyr)
67
68 dfBOLD_United_States <- dfBOLD.sub %>%
69 filter(country == "United States")
70
71 table_lat <- table(dfBOLD_United_States$lat)
72
73 table_lat
74 library(dplyr)
75
76 dfBOLD_United_States <- dfBOLD.sub %>%
77 filter(country == "United States")
78
```

271:1 (Top Level) R Script

Console

19°C Partly sunny

Search

ENG US

12:44 PM 2024-10-04

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Project: (None)

Untitled1* Script3_code_along_loop_vs_vectorized_* Class4_Cnidaria_BOLD_R_V1.R* df_bold Script2_RStudio_V1_with_challenges.R* bryzoa_BOLD_data.tsv* binf 6210 asignment 1 draft 4.R dfBOLD_canada Untitled2*

```
79 table_lon <- table(dfBOLD_United_States$lon)
80
81 table_lon
82
83 #The BOLD records of Bryozoan species per country, who had high abundace was sorted from highest to lowest, to get a good idea of which country had high abundance. I had figured it
was United States and Canada, however I wanted to make sure.
84
85 plot(sort(table(dfBOLD.sub$country), decreasing = TRUE)[1:5],
86       main = "Bold Records of Bryozoan Species per Country",
87       xlab = "Countries",
88       ylab = "Number of Bold Records")
89
90 # A table was generated in effort to filter the Bryozoan species with their respective conutries to further filter out the data to answer my research question.
91
92 y <- table(dfBOLD.sub$country, dfBOLD.sub$species_name)
93 y
94 dfBOLD_canada <- dfBOLD.sub[dfBOLD.sub$country == "Canada",]
95 library(dplyr)
96 dfBOLD_canada <- dfBOLD.sub %>% filter(country == "Canada")
97 ls()
98 y <- table(dfBOLD_canada$species_name)
99 y
100
101 #The Abundance of North American Bryozoan spcies were graphed.
102
103 barplot(y,
104         main = "The Abundance of Bryozoan Species",
105         xlab = "Bryozoan Species",
106         ylab = "The Abundance")
107 counts <- c(9, 6, 4, 3, 1)
108 names <- c("Membranipora membranacea", "Primavelans insculpta", "Phidolopora pacifica", "Schizoporella japonica", "Tubulipora tuba")
109
110 # Section 3: During this section Bryozoan species found in Canada were plotted using ggplot. This plot highlights the species found in highest abundance in Canada, and also, the
species that contained bioactive compounds, which contribute to cancer. The first three species are highlighted as contribution to cancer treatment.
111
112
113 library(ggplot2)
114 df <- data.frame(
115   species = c("Membranipora membranacea", "Primavelans insculpta", "Phidolopora pacifica", "Schizoporella japonica", "Tubulipora tuba"),
116
```

271:1 (Top Level) R Script

Console

19°C Partly sunny

Search

ENG US 12:44 PM 2024-10-04

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Project: (None)

Untitled1* Script3_code_along_loop_vs_vectorized... Class4_Cnidaria_BOLD_R_V1.R* df_bold Script2_RStudio_V1_with_challenges.R* bryzoa_BOLD_data.tsv* binf 6210 assignment 1 draft 4.R dfBOLD_canada Untitled2*

Source on Save Run

```
116 count = c(9,6,4,3,1),
117 contribution = c("Cancer contribution", "Cancer contribution", "Cancer contribution", "Non Cancer contribution", "Non Cancer contribution")
118 )
119 )
120 ggplot(df,aes(reorder(species, count), y = count, fill = contribution)) +
121   geom_bar(stat = "identity") +
122   labs(
123     title = "Bryozoan Species Found in Canada",
124     x = "Bryozoan Species",
125     y = "Number of Species"
126   ) +
127   coord_flip() +
128   theme_minimal() +
129   theme(
130     plot.title = element_text(size = 20,face = "bold", hjust = 0.5),
131     axis.title.x = element_text(size = 16),
132     axis.title.y = element_text(size = 16),
133     axis.text.y = element_text(size = 12),
134     legend.text = element_text(size = 14),
135     legend.key.size = unit(1.5,"lines")
136   ) +
137   scale_fill_manual(values = c("lightgreen", "lightblue"))
138
139
140 # The Countries were filtered out and then particular data pertaining to United States, such as the species found and their abundance was filtered out. Then barplot was generated to
141 # see the highest abundance of the target species found in United States and also highlighting the species that contribute to cancer treatment. The first three species were highlighted.
142
143 unique(dfBOLD$country)
144 dfBOLD_United_States <- dfBOLD.sub[dfBOLD.sub$country == "United States",]
145 library(dplyr)
146 dfBOLD_United_States <- dfBOLD.sub %>% filter(country == "United States")
147 ls()
148 y <- table(dfBOLD_United_States$species_name)
149 y
150 counts <- c(85,35,33,29,25)
151 names <- c("Watersipora subtorquata", "Membranipora chesapeakeensis", "Membranipora membranacea", "Watersipora sp. COI group A", "Watersipora sp. Santa Cruz Harbour ")
152
153 library(ggplot2)
154 df <- data.frame(
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
```

Console

271:1 (Top Level) :

R Script

19°C Partly sunny

Search

12:44 PM 2024-10-04


```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins
Project
Untitled1* Script3_code_along_loop_vs_vectorized... Class4_Cnidaria_BOLD_R_V1.R* df_bold Script2_RStudio_V1_with_challenges.R* bryzoa_BOLD_data.tsv* binf 6210 assignment 1 draft 4.R* dfBOLD_canada* Untitled2*
Source on Save Run
154 df <- data.frame(
155   species = c("Watersipora subtorquata", "Membranipora chesapeakeensis", "Membranipora membranacea", "Watersipora sp. COI group A", "Watersipora sp. Santa Cruz Harbour"),
156
157   count = c(85, 35, 33, 25, 29),
158   contribution = c("Cancer contribution", "Cancer contribution", "Cancer contribution", "Non Cancer contribution", "Non Cancer contribution")
159 )
160 ggplot(df, aes(reorder(species, count), y = count, fill = contribution)) +
161   geom_bar(stat = "identity") +
162   labs(
163     title = "Bryozoan Species Found in United States",
164     x = "Bryozoan Species",
165     y = "Number of Species"
166   ) +
167   coord_flip() +
168   theme_minimal() +
169   theme(
170     plot.title = element_text(size = 20, face = "bold", hjust = 0.5),
171     axis.title.x = element_text(size = 16),
172     axis.title.y = element_text(size = 16),
173     axis.text.y = element_text(size = 12),
174     legend.text = element_text(size = 14),
175     legend.key.size = unit(1.5, "lines")
176   ) +
177   scale_fill_manual(values = c("lightgreen", "lightblue"))
178
179
180 #Section 4: During this section the Bryozoan species from Canada and United States, which also contribute to cancer treatment as found was the previous graphs plotted, were graphed
181 #together. This was done in effort to compare between the two countries the species in abundance and to answer my research question in a more clearer way. The particular abundance of
182 #target species found in United states and Canada can now be clearly intreperated.
183
184 dfBOLD.sub <- data.frame(
185   species_name = c("Membranipora membranacea", "Membranipora membranacea", "Watersipora subtorquata", "Primavelans insculpta"),
186   country = c("United States", "Canada", "United States", "Canada"),
187   contribution = c(35, 9, 85, 6))
188
189 library(dplyr)
190 df_combined <- dfBOLD.sub %>%
191   filter(country %in% c("United States", "Canada")) %>%
192   group_by(species_name, country) %>%
271:1 (Top Level) :
```

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Project: (None)

Untitled1* Script3_code_along_loop_vs_vectorized... Class4_Cnidaria_BOLD_R_V1.R* df_bold Script2_RStudio_V1_with_challenges.R* bryzoa_BOLD_data.tsv* binf 6210 assignment 1 draft 4.R dfBOLD_canada Untitled2*

```
191 group_by(species_name, country) %>%
192 summarise(contribution = sum(contribution)) %>%
193 ungroup()
194
195 library(ggplot2)
196
197 ggplot(df_combined, aes(x = "species name", y = contribution, fill = country)) +
198   geom_bar(stat = "identity", position = position_dodge(width = 0.8), width = 0.4) +
199   labs(title = "The Abundance of Membranipora membranacea Species Found in United States and Canada",
200        x = "Membranipora membranacea",
201        y = "Number of species") +
202   theme_minimal() +
203   theme(
204     plot.title = element_text(size = 20, face = "bold", hjust = 0.5),
205     axis.title.x = element_text(size = 16),
206     axis.title.y = element_text(size = 16),
207   )
208
209 # A stacked bar graph was generated to make the intreperation even more precie and clear, as the graph indicated United States has a higher abundance of target species, which
# contribute to cancer treatment.
210
211 ggplot(df_combined, aes(x = species_name, y = contribution, fill = country)) +
212   geom_bar(stat = "identity") +
213   labs(
214     title = "Cancer Contributing Bryozoan Species Found in United States and Canada",
215     x = "Bryozoan Species",
216     y = "Number of Species"
217   ) +
218   theme_minimal() +
219   theme(
220     plot.title = element_text(size = 20, face = "bold", hjust = 0.5),
221     axis.title.x = element_text(size = 16),
222     axis.title.y = element_text(size = 16),
223     axis.text.y = element_text(size = 14),
224     legend.text = element_text(size = 14),
225     legend.key.size = unit(1.5, "lines")
226   ) +
227   coord_flip()
228
229
```

271:1 (Top Level) R Script

Console

NASDAQ +0.84%

Search

ENG US 12:44 PM 2024-10-04

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Project: (None)

Script3_code_along_loop_vs_vectorized_... Class4_Cnidaria_BOLD_R_V1.R* df_bold Script2_RStudio_V1_with_challenges.R* bryzoa_BOLD_data.tsv binf 6210 assignment 1 draft 4.R dfBOLD_canada Untitled2*

Source on Save Run Source

```
230 #Section 5: Since we found that United States has a higher abundance in comparison to Canada, map plot was generated to see which part of the country the species are found and in what
    abundance.
231
232
233 library(ggplot2)
234 library(maps)
235
236 us_map <- map_data("state")
237
238 df <- data.frame(
239   species = c("Watersipora subtorquata", "Membranipora chesapeakeensis",
240             "Membranipora.mem", "Watersipora sp. COI group A",
241             "Watersipora sp. Santa Cruz Harbour"),
242   count = c(85, 35, 33, 25, 29),
243   contribution = c("Cancer contribution", "Cancer contribution",
244                  "Cancer contribution", "Non Cancer contribution",
245                  "Non Cancer contribution"),
246   lat = c(37.7749, 34.0522, 40.7128, 36.1699, 39.7392),
247   lon = c(-122.4194, -118.2437, -74.0060, -115.1398, -104.9903)
248 )
249
250 ggplot() +
251   geom_polygon(data = us_map, aes(x = long, y = lat, group = group), fill = "lightgreen", color = "white") +
252   geom_point(data = df, aes(x = lon, y = lat, color = contribution, size = count), alpha = 0.7) +
253   geom_text(data = df, aes(x = lon, y = lat, label = species), hjust = -0.1, vjust = -0.5, size = 4.5) +
254   scale_size(range = c(3, 8)) +
255   labs(
256     title = "Distribution of Bryozoan Species in the United States",
257     x = "Longitude",
258     y = "Latitude"
259   ) +
260   theme_minimal() +
261   theme(
262     plot.title = element_text(size = 20, face = "bold", hjust = 0.5),
263     axis.title.x = element_text(size = 16),
264     axis.title.y = element_text(size = 16),
265     legend.text = element_text(size = 14),
266   ) +
267   scale_color_manual(values = c("blue", "red"))
268
```

231:1 (Top Level) R Script

Console

19°C Partly sunny

Search

1:00 PM 2024-10-04

Results and Discussion:

In the process of answering the research question posed above, it was found many species of Bryozoa have been barcoded in United States and Canada. In figure 1, Bryozoan species, which were found in high abundance in Canada were graphed. The figure depicts abundance of the *Membranipora membranacea*, *Primavelans insculpta*, and *Phidolopora pacifica*. The Bryozoan species mentioned, according to Figuerola and Avila, 2019, have shown compounds and cytotoxic activity against cancer. Similarly, figure 2 illustrates the Bryozoan species found in high abundance in United states. Species such as, *Watersipora subtorquata*, *Membranipora chespeakensis*, and *Membranipora membranacea* were found in high abundance and based on Figuerola and Avila, 2019, these organisms contain compounds and cytotoxic activity against cancer. The results of this project indicates that many Bryozoan species are barcoded in United States and Canada, which have contributed to enhancing cancer treatment, however pertaining to the second question asked in the introduction, relatively high abundance of species were found in United States, in contrast, Canada did not have high abundance of species, which contributed to cancer.

The data obtained from BOLD data base indicates that there may have been sampling bias, as higher abundance of Byrozoan species were found in North America (US-715, Can-583), in comparison to other continents of the world. The reason this is a sampling bias is because countries like Japan (356), New Zealand (224), and Australia (390) also have high bin quantities of the species. To further progress in this project, other countries, which showcase high abundance, should be further sampled and the surrounding countries to limit sampling bias and potentially advance in cancer treatment using Bryozoan species and its compounds.

Acknowledgments:

I would like to thank and acknowledge my classmates who have either helped or gave me either, which completing this assignment. I would like to give a special thank you to the TA, Brittany, for helping me throughout this assignment and give us encourage

References:

1. Ciavatta, M. L., Lefranc, F., Vieira, L. M., Kiss, R., Carbone, M., van Otterlo, W. A. L., Lopanik, N. B., & Waeschenbach, A. (2020). The Phylum Bryozoa: From Biology to Biomedical Potential. *Marine drugs*, 18(4), 200. <https://doi.org/10.3390/md18040200>
2. Figuerola, B., & Avila, C. (2019). The Phylum Bryozoa as a Promising Source of Anticancer Drugs. *Marine drugs*, 17(8), 477. <https://doi.org/10.3390/md17080477>