Protein Based Link Prediction and Evolutionary Analysis

Kai Thomas

Introduction

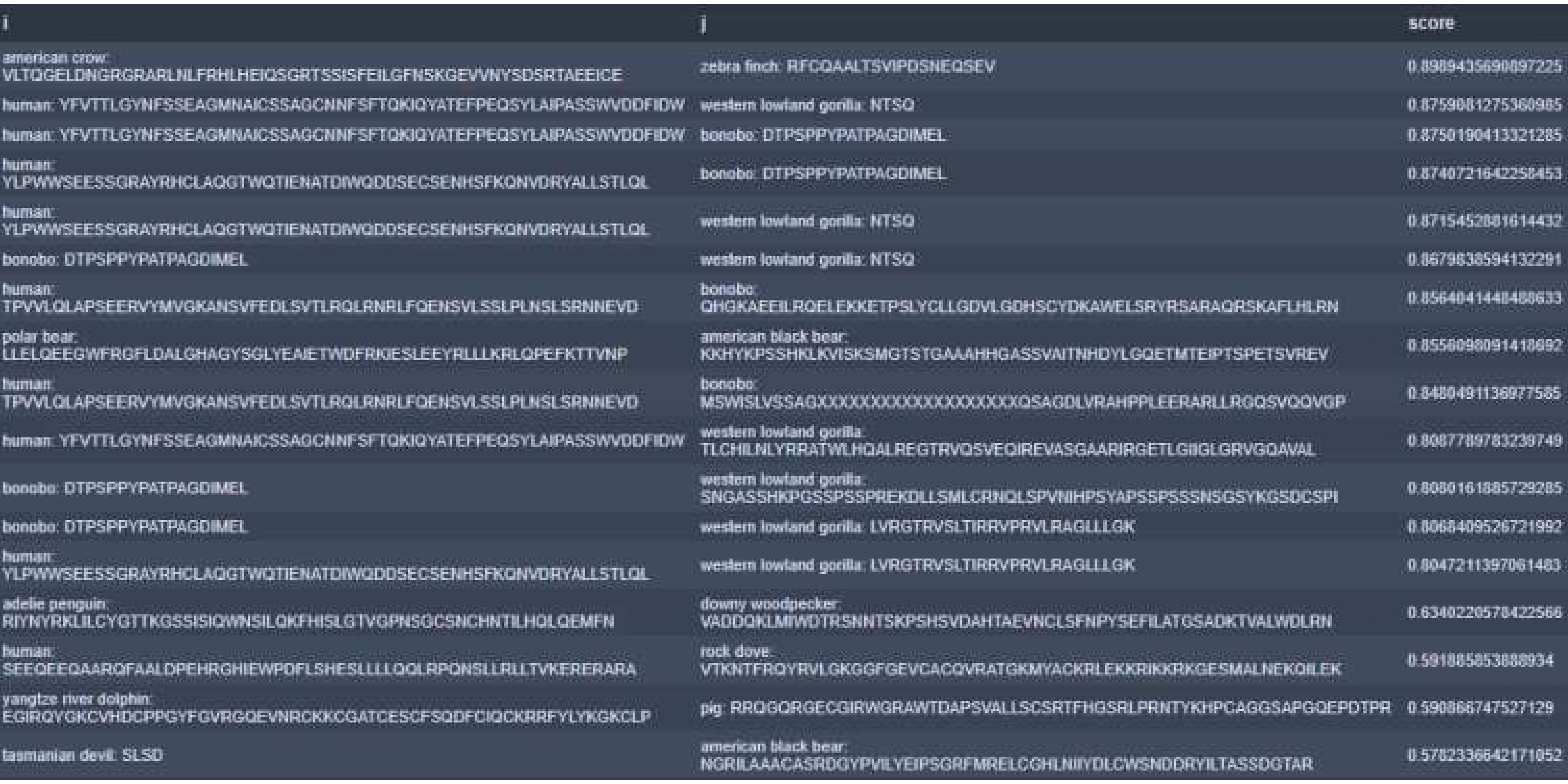
I wanted to see if it would be possible to predict evolutionary connections between species given protein data. This project has no method of quantitative accuracy but it could be a good tool for actual evolutionary biologists to know where to look.

Methods

I started by manually combing through pages and pages of UniProt's reference proteomes to find species that I thought would be interesting and would fit the scope of my project. I ended up choosing 24 species from the UniProt data and imported the .gz files into Python. It was a bit challenging at first to try to interact with the .gz files, as it was a format I had never encountered before, but plenty of research later I found a Python library called gzip, which simplified the process a lot. After importing the files I began processing the data. I chose four (4) random proteins from each of the 24 species in my dataset, which left me with 24 species nodes and 96 protein nodes in my network, for a total of 120 nodes. I then looped over all the nodes and started building my bipartite network. If a protein existed exactly within a species, then an edge was created between that species node and that protein node. This ensured that each species node would have at least four edges, and each protein node would have at least one. After creating the species-protein bipartite network, making the projections was straightforward. As I approached the point of making the projections I was planning on looking at the lecture notes because I couldn't remember exactly how to make projections, but then in the car one day I had a small breakthrough. I remembered that an edge exists in a projection network if (in the original bipartite network) the two nodes had had a common neighbor of the other type. After that breakthrough, it was simple enough to glance at the lecture notes and confirm, then implement a projection network function. Once I was done I had three networks, the species-protein bipartite network, a species-species projection network (in which an edge exists between two species if the two species have a common protein in their proteomes), and the protein-protein projection network (in which an edge exists between two proteins if there exists a species in the network which has both proteins in its proteome). I opted to only do missing link prediction on the protein-protein network because it allows me to predict a species which is not in my network which could have both proteins in question. This prediction allowed me to manually dig into the evolutionary history of the species which those proteins came from and see if there was any basis for a connection, or if I could find a specific species those predictions might point to. The species-species projection would instead point me towards a protein which both species might have, which seems a lot drier and harder to research, and also wouldn't answer the questions I was trying to answer. Then I started getting into missing link prediction and decided to use two different topological predictors and compare their results. I chose both Jaccard and degree product prediction but later decided that degree product was not worth my time. I had no way to derive a quantitative AUC for my predictions but, after researching, I am able to report on a more qualitative measure of accuracy. I later narrowed down the scope of my species selection using the information that Jaccard predicted and the things I had learned through my research. I repeated basically all of the same steps as above on a more specific dataset.

Results

These are some of the most interesting/highest scoring predictions from Jaccard. I hand picked the last few of these data and added them here because the majority of the top 100 or so predictions were just between human and primate proteins. Also, it was frequently predicting a connection between two proteins from the same species. I thought that was kind of boring, and also it didn't give me enough relationships to research, so I chose some interesting ones with a score greater than 0.5 to further analyze. Since I'm not doing any qualitative accuracy measures, this cherry-picking of the data is not a problem, but it is important to be open about it.



i	j	score
american crow: VLTQGELDNGRGRARLNLFRHLHEIQSGRTSSISFEILGFNSKGEVVNYSDSRTAEEICE	zebra finch: RFCQAALTSVIPDSNEQSEV	0.8989435690897225
human: YFVTTLGYNFSSEAGMNAICSSAGCNNFSFTQKIQYATEFPEQSYLAIPASSWVDDFIDW	western lowland gorilla: NTSQ	0.8759081275360985
human: YFVTTLGYNFSSEAGMNAICSSAGCNNFSFTQKIQYATEFPEQSYLAIPASSWVDDFIDW	bonobo: DTPSPPYPATPAGDIMEL	0.8750190413321285
human: YLPWWSEESSGRAYRHCLAQGTWQTIENATDIWQDDSECSENHSFKQNVDRYALLSTLQL	bonobo: DTPSPPYPATPAGDIMEL	0.8740721642258453
human: YLPWWSEESSGRAYRHCLAQGTWQTIENATDIWQDDSECSENHSFKQNVDRYALLSTLQL	western lowland gorilla: NTSQ	0.8715452881614432
bonobo: DTPSPPYPATPAGDIMEL	western lowland gorilla: NTSQ	0.8679838594132291
human: TPVVLQLAPSEERVYMVGKANSVFEDLSVTLRQLRNRLFQENSVLSSLPLNSLSRNNEVD	bonobo: QHGKAEEILRQELEKKETPSLYCLLGDVLGDHSCYDKAWELSRYRSARAQRSKAFLHLRN	0.8564041448488633
polar bear: LLELQEEGWFRGFLDALGHAGYSGLYEAIETWDFRKIESLEEYRLLLKRLQPEFKTTVNP	american black bear: KKHYKPSSHKLKVISKSMGTSTGAAAHHGASSVAITNHDYLGQETMTEIPTSPETSVREV	0.8556098091418692
human: TPVVLQLAPSEERVYMVGKANSVFEDLSVTLRQLRNRLFQENSVLSSLPLNSLSRNNEVD	bonobo: MSWISLVSSAGXXXXXXXXXXXXXXXXXXXXQSAGDLVRAHPPLEERARLLRGQSVQQVGP	0.8480491136977585
human: YFVTTLGYNFSSEAGMNAICSSAGCNNFSFTQKIQYATEFPEQSYLAIPASSWVDDFIDW	western lowland gorilla: TLCHILNLYRRATWLHQALREGTRVQSVEQIREVASGAARIRGETLGIIGLGRVGQAVAL	0.8087789783239749
bonobo: DTPSPPYPATPAGDIMEL	western lowland gorilla: SNGASSHKPGSSPSSPREKDLLSMLCRNQLSPVNIHPSYAPSSPSSSNSGSYKGSDCSPI	0.8080161885729285
bonobo: DTPSPPYPATPAGDIMEL	western lowland gorilla: LVRGTRVSLTIRRVPRVLRAGLLLGK	0.8068409526721992
human: YLPWWSEESSGRAYRHCLAQGTWQTIENATDIWQDDSECSENHSFKQNVDRYALLSTLQL	western lowland gorilla: LVRGTRVSLTIRRVPRVLRAGLLLGK	0.8047211397061483
adelie penguin: RIYNYRKLILCYGTTKGSSISIQWNSILQKFHISLGTVGPNSGCSNCHNTILHQLQEMFN	downy woodpecker: VADDQKLMIWDTRSNNTSKPSHSVDAHTAEVNCLSFNPYSEFILATGSADKTVALWDLRN	0.6340220578422566
human: SEEQEEQAARQFAALDPEHRGHIEWPDFLSHESLLLLQQLRPQNSLLRLLTVKERERARA	rock dove: VTKNTFRQYRVLGKGGFGEVCACQVRATGKMYACKRLEKKRIKKRKGESMALNEKQILEK	0.591885853888934
yangtze river dolphin: EGIRQYGKCVHDCPPGYFGVRGQEVNRCKKCGATCESCFSQDFCIQCKRRFYLYKGKCLP	pig: RRQGQRGECGIRWGRAWTDAPSVALLSCSRTFHGSRLPRNTYKHPCAGGSAPGQEPDTPR	0.590866747527129
tasmanian devil: SLSD	american black bear: NGRILAAACASRDGYPVILYEIPSGRFMRELCGHLNIIYDLCWSNDDRYILTASSDGTAR	0.5782336642171052

American Crow and Zebra Finch

This is actually the most likely edge prediction my program made. The American Crow and the

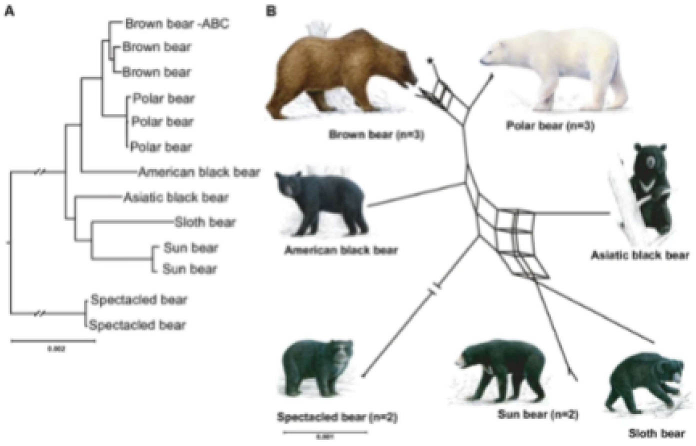
Zebra Finch are both passerines, aka 'perching birds' or 'song-birds'. The crow family is thought to have evolved near what is now Australia and has spread throughout the world. The volcanic islands of the Philippines and Papua New Guinea likely did not exist during the early history of these birds, but as they started to form I imagine they allowed some bird species to ‘island hop’ to other continents. Perhaps the larger size of crows allowed some members of the family to travel across large swaths of water and leave the Australian continent, where the Zebra Finch could not. Birds apparently speciated so rapidly after the dinosaurs that it has been difficult for evolutionary biologists to establish phylogenetic trees.

Essentially, because of their similar location of evolutionary origin and their joint classification as song-birds, I'm going to call this a win for Jaccard prediction. This edge could point to basically any other passerines which evolved in the same area/time-frame

Polar Bear and American Black Bear

The polar bear is a fairly recent species and shares a common ancestor with the brown bear just one speciation back. The brown bear and the polar bear are very closely related, to the point where they can breed and produce fertile offspring. There is actually a subspecies of the brown bear called the ABC bear, whose genome most closely resembles that of the brown bear, but whose mitochondrial DNA very closely resembles the polar bear. Research from 2017 suggests that there is gene flow across more bears than just brown bears and polar bears and includes Asiatic bears. This gene flow points to a time of the Bering strait, when these bears would have been able to mingle with each other. That same research finds that the polar bear and the American black bear are only separated by two speciation events (one which separated the polar and brown bear, and one which separated the black bear from the common ancestor of the polar and brown bears). In Wapusk, Manitoba all three species of North American bear have been found living in the same location/environment. Their ability to survive in the same environment points to the similarity of all three species. I think this predicted edge could easily point toward the brown bear, or the common ancestor of all three. Since these are both North American bears and are evolutionarily close I'm going to call this another win for Jaccard.

Figure 2: A coalescent species tree and a split network analysis from 18,621 GF ML trees.



Extra Species

Yangtze River Dolphin and Pig

When this connection popped up I was initially disbelieving, thinking that the only connection they could really have was both being mammals. I am aware of the innate connection of all species on this Earth (if you go back far enough) and the vestigial bones whales and dolphins have led me to believe that they descended from a terrestrial mammal. I was just surprised that it would be a pig. When I started looking into it though, I learned that there is actually much more basis for a connection that I had anticipated. Around 50 million years ago in south Asia, there was a small semi-aquatic species of Artiodactyl (the same order that modern pigs and deer are a part of), and it is likely that modern day Cetaceans evolved from a species like this. One such family of the Artiodactyls were the Raoellidae, thought to be the closest ancestor of modern whales and dolphins. The Yangtze river dolphin and the pig likely came from a divergence of these early Artiodactyls, so I'm going to say that Jaccard got this one too.

Tasmanian Devil and American Black Bear

The Tasmanian devil is a marsupial, a type of metatheria, which do not have placentas and for which the young are born just a few days after conception and must survive inside an external pouch in the parents body. Marsupials were once a much more dominant group on Earth and filled similar ecological niches to their placental counterparts, but in almost every situation where placental mammals and non-placental mammals interacted, the non-placental mammals were wiped out. The Tasmnaian devil is native to Australia. (This is unrelated, but the Thylacine, which was my extra credit animal from the beginning of the semester, is a marsupial which filled an ecological role very similar to that of carnivorous placental mammals, before it went extinct in the 1980s). I have not been able to find any evidence for a serious connection between the Tasmanian devil and the American Black Bear. You can see in the figure below that the order Carnivora (bears, third from the left in blue) and the order Dasyuromorphia (devils, second from the right in purple) are very far apart, and separated by many speciation events. This edge may point to their most recent common ancestor which probably existed around 160 million years ago. I'm actually going to call this one a failure for Jaccard.

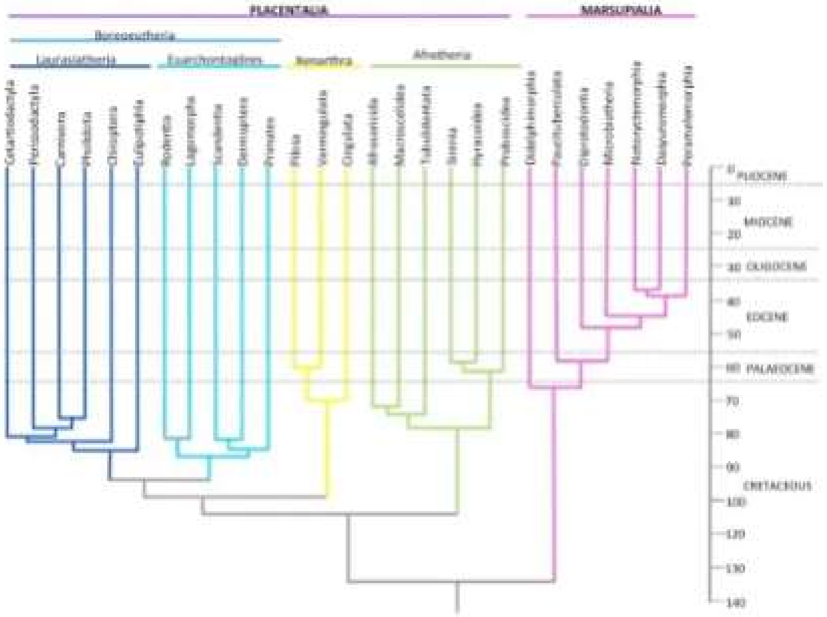
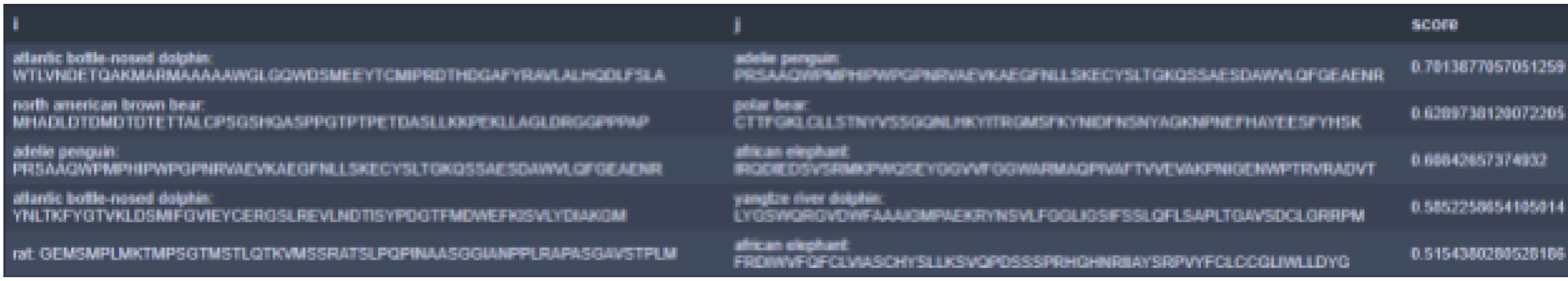


Figure 2 — Cladogram representing the relationships of the major orders of living eutherian and metatherian orders (positions of eutherian orders based on Madsen et al., 2001 and Murphy et al., 2001; metatherian orders based on Nilsson et al., 2010).

After creating my more specific network based on what I had learned from the first one, these are some of my most interesting results. As you can hopefully see, this rerun of the program was able to back up my predictions. Unfortunately I took it too specific and the majority of the predictions were between proteins of the same species. I removed almost all birds, the platypus and the sea turtle, as well as humans and primates. I added in the North American brown bear and the bottlenose dolphin. I hope you can read this, I can’t get it to be any clearer.

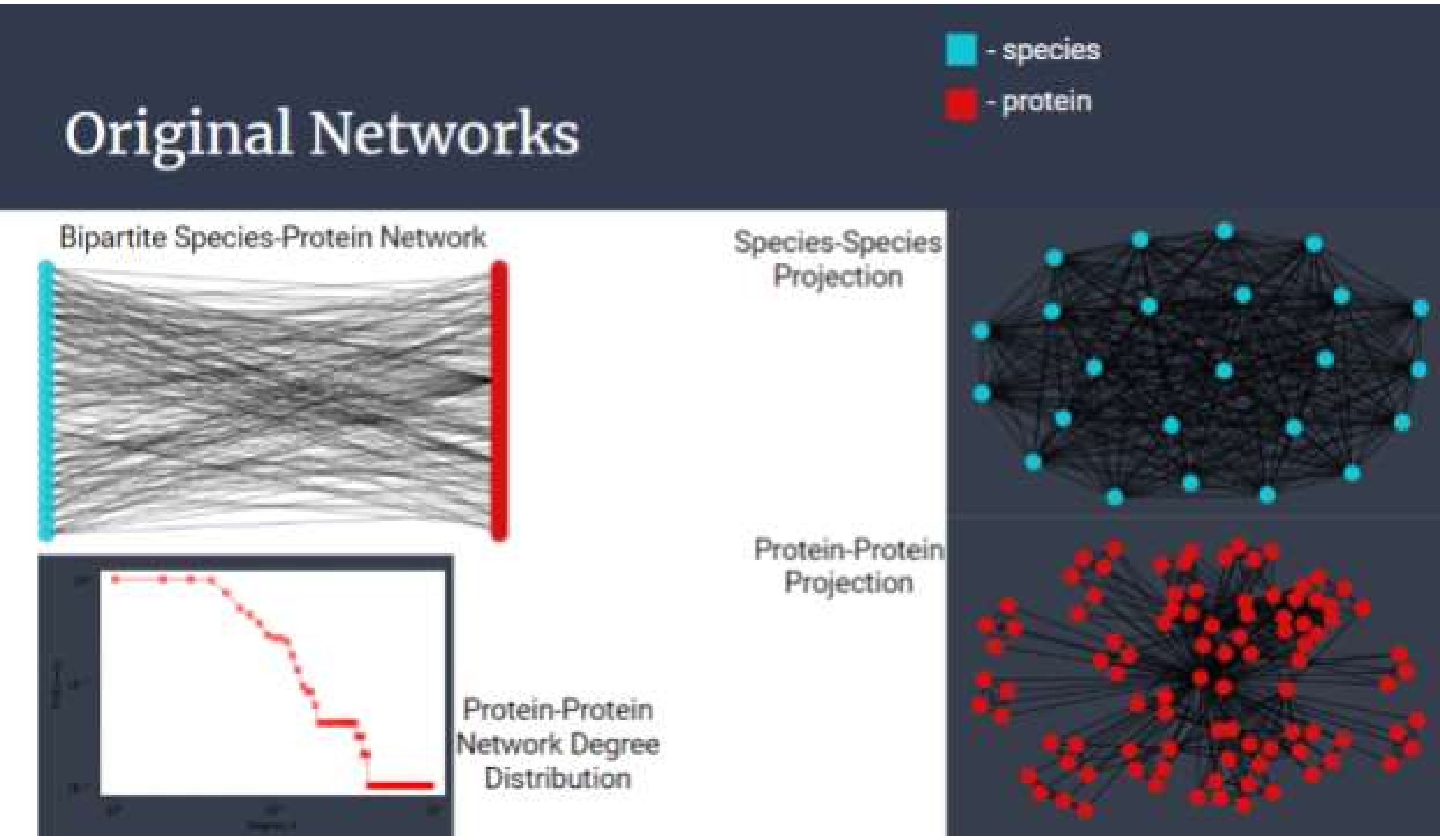


i	j	score
atlantic bottle-nosed dolphin: WTLVNDETQAKMARMAAAAAWGLGQWDSMEEYTCMIPRDTHDGAFYRAVLALHQDLFSLA	adelie penguin: PRSAAQWPMPHIPWPGPNRVAEVKAEGFNLLSKECYSLTGKQSSAESDAWVLQFGEAENR	0.7013877057051259
north american brown bear: MHADLDTDMDTDTETTALCPSGSHGASPPGTPTPETDASLLKKPEKLLAGLDRGGPPPAP	polar bear: CTTFGKLGLLSTNYVSSGQNLHKYITRGMSFKYNIDFNSNYAGKNPNEFHAYEESFYHSK	0.6289738120072205
adelie penguin: PRSAAQWPMPHIPWPGPNRVAEVKAEGFNLLSKECYSLTGKQSSAESDAWVLQFGEAENR	african elephant: IRQDIEDSVSRMKPWQSEYGGVVFGGWARMAQPIVAFTVVEVAKPNIGENWPTRVRADVT	0.60842657374932
atlantic bottle-nosed dolphin: YNLTKFYGTVKLDSMIFGVIEYCERGSLREVLNDTISYPDGTFMDWEFKISVLYDIAKGM	yangtze river dolphin: LYGSWQRGVDWFAAAIGMPAEKRYNSVLFGGLIGSIFSSLQFLSAPLTGAVSDCLGRRPM	0.5852258654105014
rat: GEMSMPLMKTMPSGTMSTLQTKVMSSRATSLPQPINAASGGIANPPLRAPASGAVSTPLM	african elephant: FRDIWVFQFCLVIASCHYSLLKSVQPDSSSPRHGHNRIIAYSRPVYFCLCCGLIWLLDYG	0.5154380280528186

Conclusion

I think Jaccard did pretty well actually. It was able to make (mostly) realistic connections between some of the random species I decided to put in my network. It also taught me things I didn’t already know about evolutionary connections, which is pretty cool. Jaccard outperformed degree product because the protein-protein network displays some level of community structure which Jaccard can take advantage of in a way degree product cannot.

Degree product was terrible and the pigeon basically took over. It has a seemingly high degree and just overpowers everything else. Overall though, missing link prediction on protein information can enable us to make connections between species and point us in the right direction when it comes to studying evolutionary relationships.



Original Networks

🟦 - species

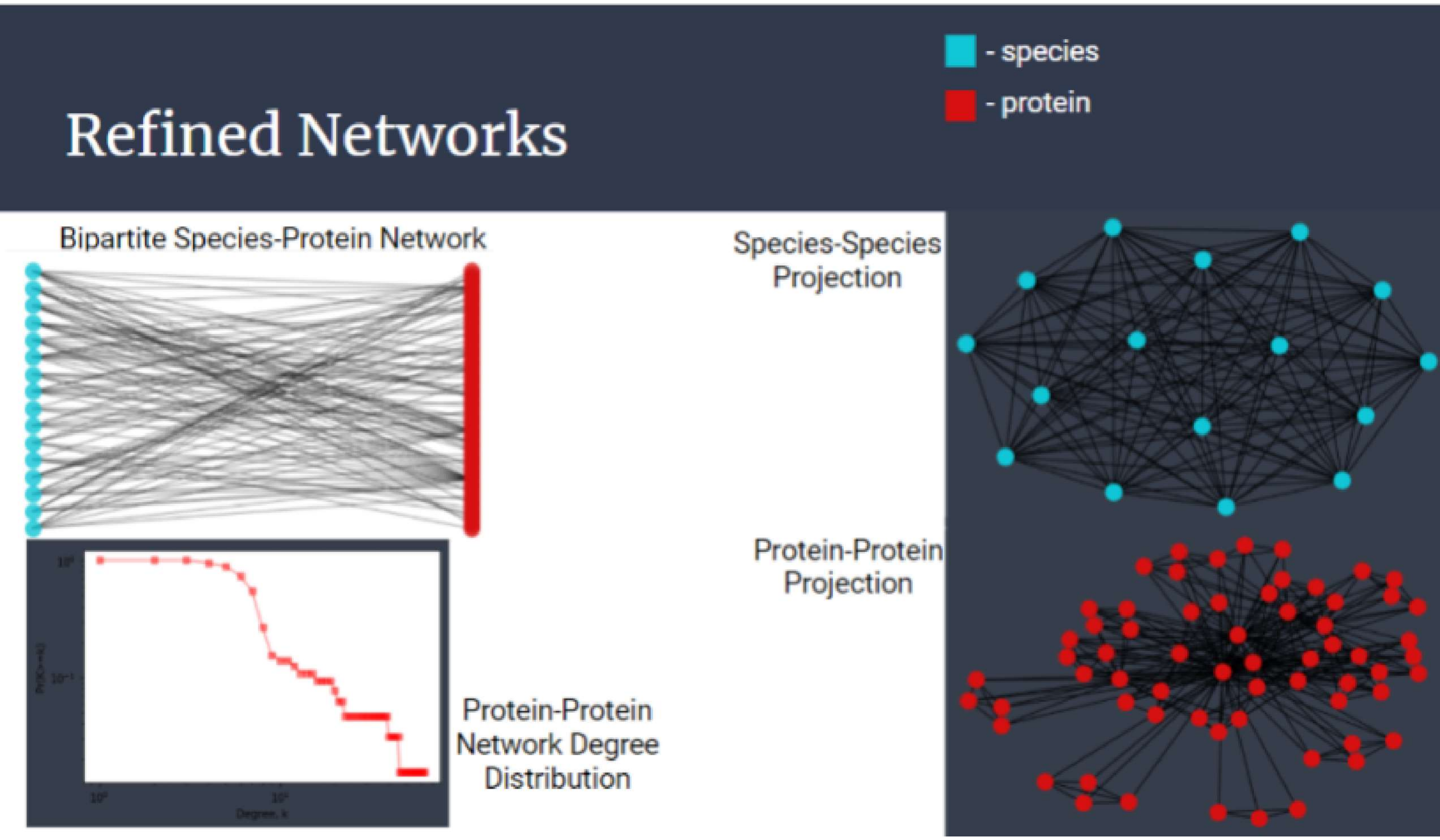
🟥 - protein

Bipartite Species-Protein Network

Species-Species Projection

Protein-Protein Projection

Protein-Protein Network Degree Distribution



Refined Networks

🟦 - species

🟥 - protein

Bipartite Species-Protein Network

Species-Species Projection

Protein-Protein Projection

Protein-Protein Network Degree Distribution

Kumar, V., Lammers, F., Bidon, T., Pfenninger, M., Kolter, L., Nilsson, M. A., & Janke, A. (2017, April 19). The evolutionary history of bears is characterized by gene flow across species. Retrieved from https://www.nature.com/articles/srep46487

Clark, D. (2018, November 24). We found grizzly, black and polar bears together for the first time. Retrieved from https://www.pbs.org/newshour/science/we-found-grizzly-black-and-polar-bears-together-for-the-first-time

Evolutionary history of bears: It's complicated. (2014, June 11). Retrieved from https://www.sciencedaily.com/releases/2014/06/140611093447.htm

Hecht, J. (2011, January 26). Lost islands of the crows revealed in DNA study. Retrieved from https://www.newscientist.com/article/dn20035-lost-islands-of-the-crows-revealed-in-dna-study/

Emery, et al. “Brains, Tools, Innovation and Biogeography in Crows and Ravens.” BMC Evolutionary Biology, BioMed Central, bmcevolbiol.biomedcentral.com/articles/10.1186/1471-2148-12-72.

Thewissen, J. G. M. (2009, April 16). From Land to Water: the Origin of Whales, Dolphins, and Porpoises. Retrieved from https://evolution-outreach.biomedcentral.com/articles/10.1007/s12052-009-0135-2#Sec1

Yong, E. (2010, May 7). Whales evolved from small aquatic hoofed ancestors. Retrieved from https://www.nationalgeographic.com/science/phenomena/2010/05/07/whales-evolved-from-small-aquatic-hoofed-ancestors/

Article: Fossil Focus > Fossil Focus: Marsupial evolution - A limited story? (2012, October 2). Retrieved from https://www.palaeontologyonline.com/articles/2012/fossil-focus-marsupials/?doing_wp_cron=1588583598.8141539096832275390625

ORIGIN AND EVOLUTION OF MARSUPIALS. (n.d.). Retrieved from http://www.nhc.ed.ac.uk/index.php?page=493.168.256

UniProt ConsortiumEuropean Bioinformatics InstituteProtein Information ResourceSIB Swiss Institute of Bioinformatics. (n.d.). UniProt Consortium. Retrieved from https://www.uniprot.org/proteomes/

gzip - Support for gzip files. (n.d.). Retrieved from https://docs.python.org/3/library/gzip.html

Aaron Clauset’s Lecture Notes for CSCI 3352, Biological Networks

Aaron Clauset’s DrawGz function from the Problem Sets