TECNOLOGIA BARREIRO

ESCOLA SUPERIOR
POLITÉCNICO SETÚBAL

Knee deep into phylogenetics

Análise de Sequências Biológicas

Licenciatura em Bioinformática

Docente: Francisco Martins

Data: 17/05/2024

201901365 Rendrick Carreira

Índice

Índice de figuras

1. Introdução

Os tardígrados, igualmente conhecidos como ursos-d'água, constituem um filo de micro-organismos extremófilos, notáveis pela sua capacidade excecional de resistir a condições ambientais adversas. Caracterizam-se pelo seu tamanho diminuto, geralmente não ultrapassando 1 mm de comprimento, apresentando um corpo segmentado e equipados com oito patas. No artigo (*What Are Tardigrades and Why Are They Nearly Indestructible?* | *Live Science*, n.d.) refere que este micro-organismos foram descobertos no século XVIII por Johann August Ephraim Goeze, e têm suscitado um interesse científico crescente devido à sua capacidade de sobreviver a extremos de temperatura, pressão, radiação e desidratação, recorrendo a um processo denominado criptobiose.

Num artigo relativamente recente (Kayastha et al., 2023) revela a descoberta de uma nova espécie de tardígrado, Paramacrobiotus gadabouti sp. nov. Este estudo foca-se na análise das características morfológicas, morfométricas e genéticas desta nova espécie, aplicando métodos de taxonomia integrativa. Ao identificar e caracterizar esta nova espécie, os investigadores almejam aprofundar o entendimento sobre a diversidade, evolução e distribuição dos tardígrados, com especial enfoque na família Macrobiotidae.

Nesse artigo realizaram uma árvore filogenética, onde os autores recorreram ao *software* MAFFT (Katoh & Standley, 2014) versão 7 para o alinhamento das sequências utilizando o método AUTO (para os marcadores COI e ITS2), e o método Q-INS-I (para os marcadores ribossomais: 18S rRNA e 28S rRNA). Em seguida o artigo menciona que foi feita um corte das sequências, que foi realizada manualmente para tamanhos específicos: 994 pb (18S rRNA), 811 pb (28S rRNA), 487 pb (ITS-2), 658 pb (COI), sendo depois concatenadas através do software *SequenceMatrix* (Vaidya et al., 2011).

Outro *software* utilizado foi *PartitionFinder* (Vaidya et al., 2011), para selecionar o melhor esquema de particionamento e modelos de substituição, conforme o Critério de Informação de Akaike (AIC). A inferência filogenética foi feita através do programa MrBayes v3.2 (Lanfear et al., 2017), realizando-se 10 milhões de gerações, com

amostragem da cadeia de Markov a cada 1000 gerações. Um desvio padrão médio das frequências de divisão inferior a 0,01 serviu como critério para assegurar a convergência das análises independentes. O *software* Tracer v1.6 foi utilizado para verificar se as cadeias de Markov atingiram a estacionaridade e para determinar o "burn-in" adequado (10% das gerações).

A árvore de *maximum likelihood* (em português, máxima verossimilhança) foi calculada usando o programa RAxML v8.0.19 (Stamatakis, 2014), com 1000 réplicas de *bootstrap* rápido para avaliar o suporte dos nós internos. A árvore de consenso resultante foi visualizada e analisada através do software FigTree v.1.4.3.

Os autores do artigo concluíram que a descrição da nova espécie de tardígrado Paramacrobiotus gadabouti sp. nov., identificada numa população na Madeira, representa um avanço significativo para o conhecimento da biodiversidade destes organismos. A abordagem integrativa adotada, que combina análises morfológicas e genéticas detalhadas, revelou-se fundamental para a caracterização precisa da nova espécie. Adicionalmente, a reconstrução filogenética do superclade II da família *Macrobiotidae* permitiu posicionar a nova espécie dentro do contexto evolutivo do grupo, destacando a influência da dispersão humana na distribuição geográfica de espécies unissexuais em comparação com táxons bissexuais do gênero *Paramacrobiotus*.

2. Objetivos

O objetivo deste projeto consiste na reprodução dos resultados apresentados no artigo (Kayastha et al., 2023), que aborda a árvore filogenética. Pretende-se desenvolver uma pipeline que facilite a replicação da árvore filogenética por parte de outros investigadores e permita a interpretação dos dados gerados.

Além disso, a pipeline desenvolvida deverá possibilitar a comparação dos resultados obtidos com os reportados no artigo original. Este processo visa garantir a validade e a fiabilidade das técnicas utilizadas, assim como aferir a replicabilidade dos resultados no contexto da investigação filogenética.

3. Métodos e Materiais

Para a replicação da arvore filogenética do artigo foi desenvolvida uma pipeline (PhyloFlow) no snakemake versão 8.11.4, que continha todas etapas desde o download dos ficheiros fasta ate a criação da arvore, como mostra a Figura 3.1

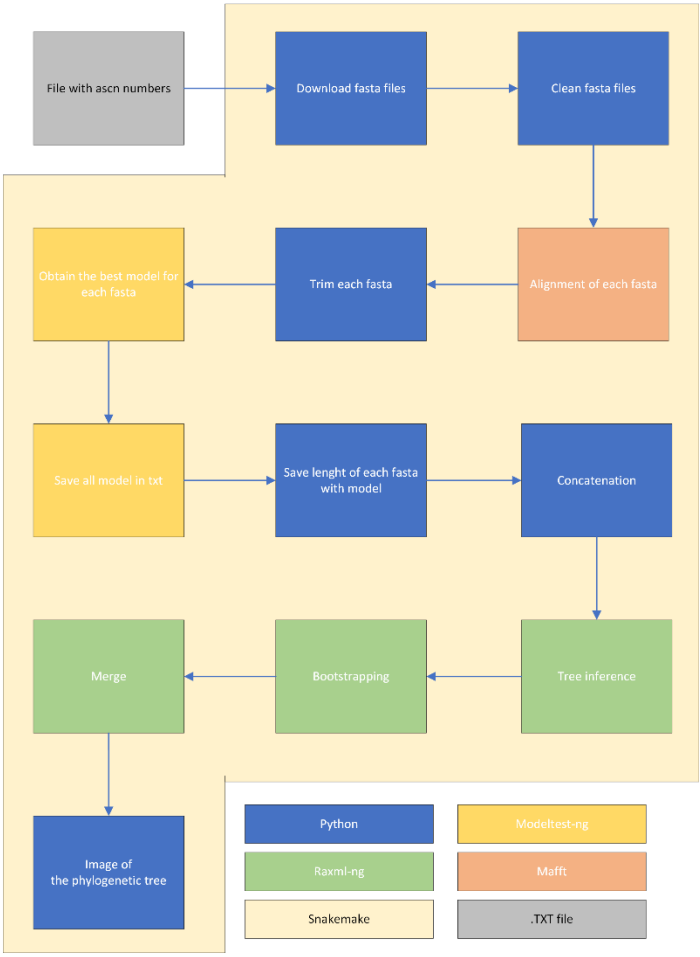

Figura 3.1 - *Workflow* da *pipeline* contendo todos os processos

3.1. Dataset

Para a replicação do estudo em análise, empregou-se o mesmo conjunto de dados genéticos que foi utilizado no artigo original, como mostra a Tabela 3.1.

Tabela 3.1 – Sequencias usadas para analise filogenética no artigo (Kayastha et al., 2023)

Taxon	18S rRNA	28S rRNA	COI	ITS–2
Paramacrobiotus gadabouti sp. nov. MD50.1	OP394210		OP394113	
Paramacrobiotus gadabouti sp. nov. MD50.2	OP394211	OP394209		
Paramacrobiotus gadabouti sp. nov. MD50.4	OP394212		OP394114	
Macrobiotus rybaki	MW588029	MW588034	MW593931	MW588022
Mesobiotus datanlanicus	MK584659	MK584658	MK578905	MK584657
Minibiotus furcatus	FJ435746	FJ435760	FJ435802	
Minibiotus gumersindoi	FJ435748	FJ435761	FJ435803	
Minibiotus intermedius	ON005189	ON005195	ON005160	
Minibiotus ioculator	MT023998	MT024041	MT023412	MT024000
Minibiotus pentannulatus 1	MT023999	MT024042	MT023413	MT024001
Minibiotus pentannulatus 2	MT023999	MT024043	MT023414	MT024001
Minibiotus sp.	OK663227	OK663238		OK663216
Paramacrobiotus aff. richtersi AU	MH664932	MH664949	MH675999	MH666081
Paramacrobiotus aff. richtersi BR 1	MH664934	MH664952	MH676000	MH666082
Paramacrobiotus aff. richtersi BR 2			MH676001	
Paramacrobiotus aff. richtersi BR 3			MH676002	
Paramacrobiotus aff. richtersi FR 1	MH664935	MH664953	MH676003	MH666083
Paramacrobiotus aff. richtersi FR 2			MH676004	
Paramacrobiotus aff. richtersi HU 1	MH664936	MH664954	MH676005	MH666084
Paramacrobiotus aff. richtersi HU 2			MH676006	
Paramacrobiotus aff. richtersi MG 1	MH664938	MH664956	MH676008	MH666086
Paramacrobiotus aff. richtersi MG 2				MH666087
Paramacrobiotus aff. richtersi NO	MH664939	MH664957	MH676009	MH666088
Paramacrobiotus aff. richtersi NZ	MH664940	MH664958	MH676010	MH666089
Paramacrobiotus aff. richtersi PT 1	MH664944	MH664961	MH676014	MH666093
Paramacrobiotus aff. richtersi PT 2			MH676015	
Paramacrobiotus aff. richtersi TN	MH664945	MH664962	MH676016	MH666094
Paramacrobiotus aff. richtersi TZ	MH664933	MH664951	MH676017	MH666095
Paramacrobiotus arduus	MK041032		MK041020	
Paramacrobiotus areolatus	MH664931	MH664948	MH675998	MH666080
Paramacrobiotus celsus	MK041031		MK041019	
Paramacrobiotus cf. klymenki IT	MH664937	MH664955	MH676007	MH666085
Paramacrobiotus cf. klymenki PT	MH664943	MH664960	MH676013	MH666092
Paramacrobiotus depressus	MK041030		MK041015	
Paramacrobiotus experimentalis	MN073468	MN073465	MN097837	MN073464
Paramacrobiotus fairbanksi PL	MH664941	MH664950	MH676011	MH666090
Paramacrobiotus filipi 1	MT261913	MT261904	MT260372	
Paramacrobiotus filipi 2			MT260373	
Paramacrobiotus lachowskae	MF568532	MF568533	MF568534	MF568535
Paramacrobiotus metropolitanus	LC637243	LC649795	LC637242	LC649794
Paramacrobiotus richtersi	MK041023		MK040994	
Paramacrobiotus richtersi S38 1	OK663224	OK663236	OK662995	OK663213
Paramacrobiotus spatialis	MK041024		MK040996	
Paramacrobiotus spatialis S107 1	OK663225	OK663236	OK662996	OK663214
Paramacrobiotus tonolli US	MH664946	MH664963	MH676018	MH666096
Sisubiotus spectabilis	MN888371	MN888357	MN888322	MN888331
Tenuibiotus cf. ciprianoi	MN888376	MN888361	MN888328	MN888348
Tenuibiotus danilovi	MN888377	MN888362	MN888329	MN888349
Tenuibiotus tenuiformis	MN888378	MN888363	MN888330	MN888350
Tenuibiotus voronkovi	KX810045	KX810049	KX810042	KX810046
Tenuibiotus zandrae	MN443040	MN443035	MN444827	MN443038

No entanto, verificou-se que alguns genes exibiam características peculiares, particularmente os genes da espécie *Paramacrobiotus experimentalis*. Durante a fase de alinhamento, estes genes revelaram numerosas discrepâncias em relação aos genes das outras espécies analisadas. Devido a essas inconsistências, optou-se pela exclusão completa desta espécie do conjunto de dados inicial. Além disso, procedeu-se à remoção de outras duas espécies (*Sisubiotus spectabilis* e *Mesobiotus datanlanicus*), que correspondiam aos *outgroup* (em português, grupos externos), a fim de prevenir problemas relacionados com a não monofilia do grupo externo.

Para o download das sequências, desenvolveu-se um *script* em Python 3.12 (*Python*, 2006) (down_ascn.py), utilizando a versão 1.83 do BioPython especificamente os módulos Entrez e SeqIO. Este *script* permitiu o *download* das sequências de cada espécie e o subsequente armazenamento em arquivos no formato FASTA, organizados por gene, contudo é necessário que ficheiros terem o formato apresentado na Figura 3.2, para o script funcionar.

```
Paramacrobiotus_gadabouti_sp._nov._MD50.1;OP394210
Paramacrobiotus_gadabouti_sp._nov._MD50.2;OP394211
Paramacrobiotus_gadabouti_sp._nov._MD50.4;OP394212
Macrobiotus_rybaki:MW588029
```

Figura 3.2 - Formato do ficheiro de texto por gene

Limpeza do *dataset*

Após a obtenção das sequências, constatou-se a presença de dados ausentes, indicados pelo caractere "N" em algumas sequencias. Para resolver esta questão, implementou-se um outro *script* em Python (fasta_cleaner.py), que eliminava todos os caracteres "N" presentes nas sequências, garantindo assim a integridade e a completude dos dados para análises subsequentes.

3.2. Alinhamento das sequências

Utilizando sequências "limpas", iniciou-se o alinhamento dos ficheiros FASTA através da utilização do MAFFT versão 7.520. O procedimento foi executado empregando-se o comando AUTO, que recorreu ao método L-INS-i (*Local pairwise alignment information*, ou em português Informação de alinhamento local entre pares) para todos os ficheiros FASTA.

Adicionalmente, verificou-se a necessidade de eliminar as extremidades de cada ficheiro FASTA. Para tal, utilizou-se um script em Python (trim_fasta_edges.py), que efetuou um corte, conhecido por "trim", nas pontas das sequências, até que estas contivessem 80% de nucleótidos, como mostra na Figura 3.3

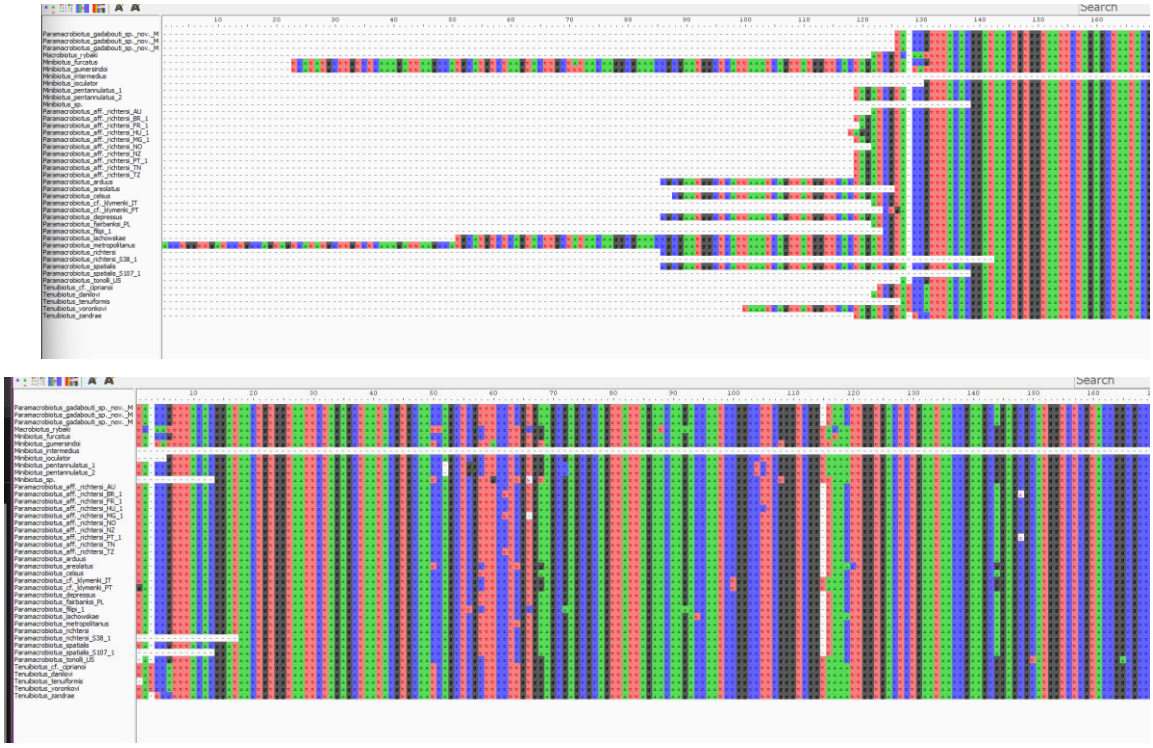

Figura 3.3 – (Cima) sequencias alinhadas do gene 18sRNA; (Baixo) sequencias alinhadas do gene 18sRNA após realizar o *trim*

Modelo estatístico

Nesta fase, foi também determinado o modelo estatístico mais adequado segundo o Critério de Informação de Akaike Corrigido (AICc) para cada ficheiro alinhado. Para

isso, empregou-se a ferramenta modeltest-ng (Darriba et al., 2020) versão 0.1.7, a fim de identificar os melhores modelos estatísticos para cada ficheiro. Paralelamente, através de um script em Python (sequence_model_processor.py), gerou-se um ficheiro contendo os modelos selecionados para cada ficheiro alinhado, juntamente com o tamanho da sua sequência, como mostra a Figura 3.4.

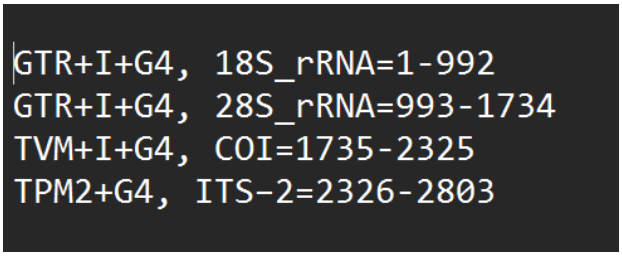

Figura 3.4 – Formato do ficheiro com os modelos estatístico juntamente com o tamanho das sequências

Concatenação

Por último, recorreu-se ao uso de Python (concatenate_fasta.py), para concatenar todos os ficheiros FASTA por ordem alfabética, correspondendo esta à mesma sequência apresentada na figura Figura 3.4.

3.3. Raxml-ng

A utilização do RAxML-NG (Kozlov et al., 2019) versão 1.2.0, foi empregada em três contextos distintos, cada um com um objetivo específico, conforme descrito a seguir:

1 - Na primeira etapa, referente à construção da árvore filogenética, procede-se à inferência da árvore de *maximum likelihood* a partir de sequências alinhadas. Esta inferência é realizada utilizando-se 100 árvores geradas aleatoriamente e 100 árvores de parcimónia, selecionando-se a topologia que apresenta maior adequação.

2 - Posteriormente, na segunda etapa, a avaliação da robustez das inferências é conduzida por meio de uma análise de *bootstrap*. Este procedimento tem por finalidade avaliar a robustez dos ramos da árvore filogenética. Para tal, são geradas

até 10000 árvores a partir de amostras dos dados alinhados, contudo o processo termina mais cedo se arvores derem converge.

3 - Na terceira etapa, denominada análise de suporte, os resultados obtidos na análise de *bootstrap* são utilizados para anotar a árvore de *maximum likelihood* com valores de suporte. Estes valores fornecem indicativos quantitativos da confiança nas relações filogenéticas estabelecidas entre as espécies estudadas.

Representação da árvore

Finalmente, a representação gráfica da árvore filogenética é realizada utilizando o módulo Toytree (Eaton, 2020) num python script (newick_tree_visualizer.py).

4. Resultados

Com todos os passos feitos foi obtido a arvore filogenética apresentada na Figura 4.1, onde é possível verificar 3 grandes clades bem estruturados (*Paramacrobiotus, Tenuibiotus* e *Minibiotus*).

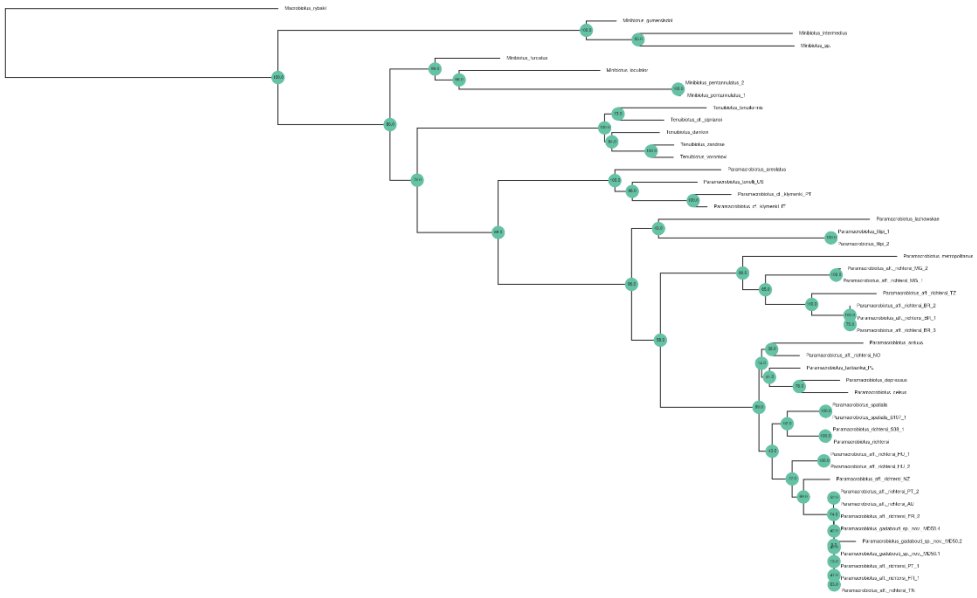

Figura 4.1 – Árvore ML contruída partir do ficheiro concatenado usando RAXML-NG com diferentes modelos (GTR+I+G4 para 18S_rRNA, GTR+I+G4 para 28S_rRNA, TVM+I+G4 para COI, e

TPM2+G4para ITS–2). Os números representam os valores de bootstrap, e foram feitas 450 replicações. *Macrobiotus rybaki* representa a espécie *outgroup. (ver imagem com maior qualidade)*

Verifica-se que *Paramacrobiotus gadabouti sp. nov*. (amostras MD50.1 e MD50.2) constitui um agrupamento próximo com várias outras espécies e estirpes de *Paramacrobiotus aff. Richtersi,* e outras espécies relacionadas. Os valores de *bootstrap* para o agrupamento das amostras de *P. gadabouti* apresentam-se de moderados a altos, sugerindo uma confiança razoável na proximidade evolutiva entre estas amostras e outras espécies do mesmo clade.

Adicionalmente, *P. gadabouti sp. nov*. encontra-se proximamente relacionado com diversos *nodes* de P. aff. richtersi, o que indica que estes agrupamentos partilham um ancestral comum recente. O *clade* que inclui *P. gadabouti* é robustamente suportado por valores elevados de *bootstrap*, conferindo uma forte confiança nas relações evolutivas inferidas.

No que concerne à diversidade e especiação, a inserção de *P. gadabouti sp. nov.* num *subclade* bem suportado dentro do género *Paramacrobiotus* sugere a ocorrência de eventos de diversificação recentes que culminaram em especiação.

5. Discussão

A posição taxonómica de *Paramacrobiotus gadabouti* sugere que esta espécie pode ter-se diversificado recentemente de um ancestral comum partilhado com espécies proximamente relacionadas, como *Paramacrobiotus aff. richtersi*, indicando uma possível recente radiação adaptativa neste *clade*. Além disso, a inclusão de *Paramacrobiotus gadabouti* no estudo reflete a elevada complexidade e diversidade observadas no género, que podem denotar uma adaptabilidade ecológica significativa e uma frequência elevada de eventos de especiação, sublinhando a dinâmica evolutiva que caracteriza este grupo.

5.1. Comparação entre árvores

Em relação a árvore apresentada no artigo (Kayastha et al., 2023), que é possível ver na Figura 5.1 são semelhantes com a árvore apresentada na Figura 4.1.

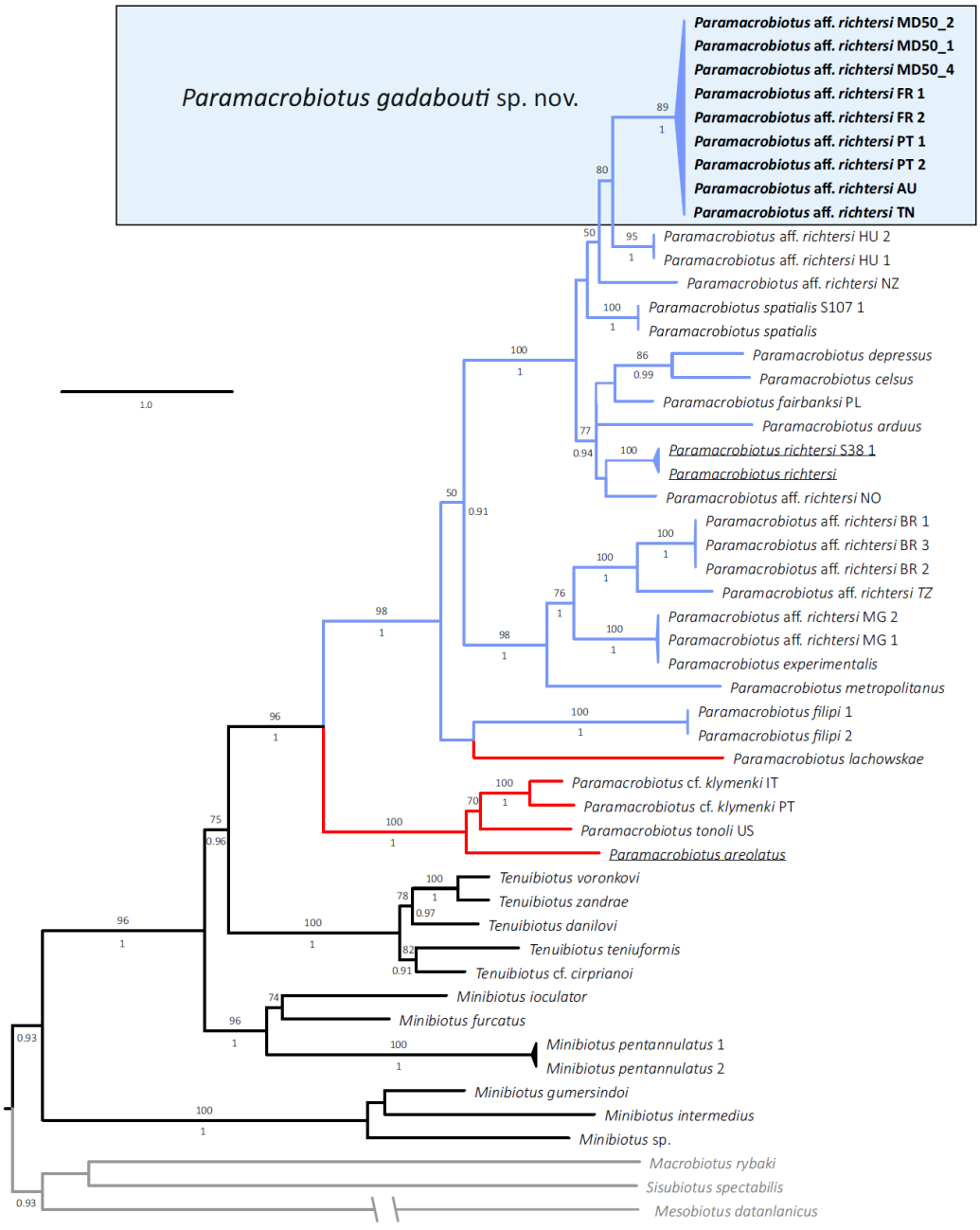

Figura 5.1 - Árvore filogenética presente no artigo (Kayastha et al., 2023)

6. Bibliografia

Darriba, Di., Posada, D., Kozlov, A. M., Stamatakis, A., Morel, B., & Flouri, T. (2020). ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models. *Molecular Biology and Evolution*, *37*(1), 291–294. https://doi.org/10.1093/MOLBEV/MSZ189

Eaton, D. A. R. (2020). Toytree: A minimalist tree visualization and manipulation library for Python. *Methods in Ecology and Evolution*, *11*(1), 187–191. https://doi.org/10.1111/2041-210X.13313

Katoh, K., & Standley, D. M. (2014). MAFFT: Iterative Refinement and Additional Methods. *Methods in Molecular Biology*, *1079*, 131–146. https://doi.org/10.1007/978-1-62703-646-7_8

Kayastha, P., Stec, D., Sługocki, Ł., Gawlak, M., Mioduchowska, M., & Kaczmarek, Ł. (2023). Integrative taxonomy reveals new, widely distributed tardigrade species of the genus Paramacrobiotus (Eutardigrada: Macrobiotidae). *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-28714-w

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., & Stamatakis, A. (2019). RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, *35*(21), 4453–4455. https://doi.org/10.1093/BIOINFORMATICS/BTZ305

Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T., & Calcott, B. (2017). PartitionFinder 2: New Methods for Selecting Partitioned Models of Evolution for Molecular and Morphological Phylogenetic Analyses. *Molecular Biology and Evolution*, *34*(3), 772–773. https://doi.org/10.1093/MOLBEV/MSW260

Python. (2006).

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. https://doi.org/10.1093/BIOINFORMATICS/BTU033

Vaidya, G., Lohman, D. J., & Meier, R. (2011). SequenceMatrix: concatenation software for the fast assembly of multi-gene datasets with character set and codon

information. *Cladistics*, *27*(2), 171–180. https://doi.org/10.1111/J.1096-0031.2010.00329.X

What are tardigrades and why are they nearly indestructible? | *Live Science*. (n.d.). Retrieved May 16, 2024, from https://www.livescience.com/57985-tardigrade-facts.html