Tratamento e Integração de Metadados Genômicos Mitocondriais em Filogenômica

Gustavo Saboia[1], Jose de Souza[1], Ana Tereza de Vasconcelos[2], André Elias Rodrigues Soares[2], Kary Ocaña[2]

[1]Faculdade de Educação Tecnológica do Estado do Rio de Janeiro (FAETERJ)
[2]Laboratório Nacional de Computação Científica (LNCC). Petrópolis – RJ – Brasil

{gsaboia,jjunior}@faeterj-petropolis.edu.br,
{atrv,asoares,karyann}@lncc.br

Abstract. *This paper presents the tool MPCreator, a workflow that allows the control and automation in the treatment and integration of genomic metadata for later analyzes of phylogenomics. Despite being a work in progress, MPCreator has already been tested and validated by scientists regarding support in the analytical capacity of metadata. MPCreator is available at https://github.com/gustavoSaboia97/MPCreator and the next steps lead to the use of parallel environments and task distribution.*

Resumo. *Este artigo apresenta a ferramenta MPCreator, um workflow que permite o controle e automação no tratamento e integração de metadados genômicos para posteriores análises de filogenômica. Apesar de ser um trabalho em andamento, o MPCreator já foi testado e validado por cientistas no que tange ao apoio na capacidade analítica dos metadados. Ele está disponível em https://github.com/gustavoSaboia97/MPCreator e os próximos passos conduzem ao uso de ambientes paralelos e distribuição de tarefas.*

1. Introdução

Dados genômicos, conhecidos como *biological big data* [Navale e Bourne 2018], são volumosos, heterogêneos e distribuídos em bancos de dados como o GenBank[1]. O processo de obtenção, tratamento e análise da informação (metadados) contida nesses dados é uma fase crítica na Bioinformática e um desafio atual para a Ciência da Computação e Ciência de Dados [Attwood *et al.* 2017; Fingert 2018]. Embora o processo de acesso a essas bases de dados possa ser automatizado computacionalmente [Yin *et al.* 2017], não existe um algoritmo padrão definido para esta tarefa. Este cenário leva ao pesquisador a realizá-lo total ou parcialmente de forma manual, o que é tedioso, demorado e propenso a erros sistemáticos e falhas na reprodutibilidade.

A filogenética é o estudo das relações evolutivas entre organismos através de informação genética. Os genomas mitocondriais são uma fonte altamente viável de informação genética, uma vez que são abundantes nas células, com uma taxa evolutiva acelerada em relação ao genoma nuclear, além de possuir uma estrutura genômica bem conservada [Wang e Wu 2015]. Embora a utilização de material genético da mitocôndria não seja apropriada em análises de grupos muito distintos, é altamente utilizado em diversos campos das ciências da vida, como genética de populações,

[1] https://www.ncbi.nlm.nih.gov/genbank/

biogeografia, demografia histórica, análises forenses e médicas [Rubinoff e Holland 2005].

Para realizar análises filogenômicas utilizando mitocôndrias é preciso organizar e agrupar os dados a serem usados por programas de análise filogenética como RAxML/ExaML, PhyML, IQ-TREE e BEAST. Nesse artigo a ferramenta proposta MPCreator é um *workflow* que oferece a gerência e automação transparente para a obtenção, tratamento e controle no processo de padronização de dados genômicos mitocondriais.

Atualmente existem *workflows* de código aberto baseados na *web* como Tavaxy[2] (Taverna & Galaxy), BioExtract[3] e FeatureExtract [Wernersson 2005]. Eles são uma alternativa para os usuários iniciantes e oferecem plataformas que permitem o uso *naive* destas análises, além de fornecer maior reprodutibilidade em análises em escala genômica. No entanto, especialistas e especialistas na área necessitam de ferramentas altamente escaláveis, acessíveis por interface de texto e que possam ser integrados em seus *workflows* customizados, que geralmente utilizam APIs do BioPerl[4], Bioconductor[5] e Ensembl[6].

O MPCreator é escrito em Python e as atividades foram modeladas na forma de um arcabouço independente. Ele pode ser acoplado como *subworkflow* em outros *workflows* em filogenética e adaptado a diversos ambientes e sistemas de gerência. O MPCreator é uma ferramenta transparente, eficiente e de fácil uso. Ele formata, minera e organiza os metadados das principais *features* do GenBank: CDS, D-loop, rRNA e tRNA e permite ter um melhor domínio no tratamento das anotações e metadados. Ele está disponível em https://github.com/gustavoSaboia97/MPCreator.

Este artigo está organizado em 4 seções, além desta introdução. A Seção 2 apresenta a motivação. A Seção 3 apresenta o *workflow* MPCreator e Seção 4 apresenta os resultados. Finalmente, a Seção 5 conclui este artigo.

2. Motivação

O processo de tratamento e integração de dados genômicos é um passo crítico que muitas vezes demanda esforço manual do cientista, o que pode levar dias e induzir erros sistemáticos *e.g.*, torna-se inviável encontrar a posição exata no meio de uma sequência nucleotídica com mais de 10 mil caracteres. Embora os dados genéticos se apresentem de maneira simplificada, com longas sequências de caracteres que representam a sequência de nucleotídeos presente na molécula de DNA, no contexto biológico esses dados interagem com o organismo de diversas maneiras. As células acessam o material genético baseado na sua sequência, e podem utilizar determinados trechos como base para a construção de proteínas, mas também simplesmente como regiões que irão indicar outros processos biológicos. Esse tipo de informação se encontra nas bases de dados na forma de metadados que irão informar ao pesquisador onde começam e terminam genes, regiões controladoras e demais *features* do genoma.

---

2 http://www.tavaxy.org/

3 http://bioextract.org

4 https://bioperl.org/

5 https://www.bioconductor.org/

6 https://www.ensembl.org/

A base de dados mais comum para armazenamento de dados genômicos, o NCBI, utiliza um formato padrão denominado GenBank. Os registros do GenBank apresentam três seções: um *header* contém identificadores (ID) e versões; *features* com informações de começo, fim, tipo, *etc.,* de cada região gênica e *sequence* com a sequência nucleotídica em si. O MPCreator atua minerando os metadados presentes nesses registros, identificando cada região da sequência nucleotídica, e permitindo ao pesquisador separar essa regiões em suas análises. Esse processo permite que diversas informações *a priori* sobre as sequências possam ser incorporadas nas análises filogenéticas, além de permitir a recuperação de *features* específicas de interesse.

O MPCreator aceita como entrada um grupo de ID, obtém os arquivos GenBank do NCBI (pelo *header*), minera os metadados *source*, *localization*, *organism*, *gene*, *etc.*, (das *features*) e organiza as sequências gerando como resultado arquivos estruturados e organizados por *features* que podem ser usados como entrada por programas de filogenia como RAxML/ExaML, PhyML, IQ-TREE e BEAST.

3. MPCreator: *workflow* para o tratamento de metadados genômicos

O MPCreator é um *workflow* escrito em Python formado por 7 atividades e pode ser acoplado antes da execução de um *workflow* de filogenômica (Figura 1).

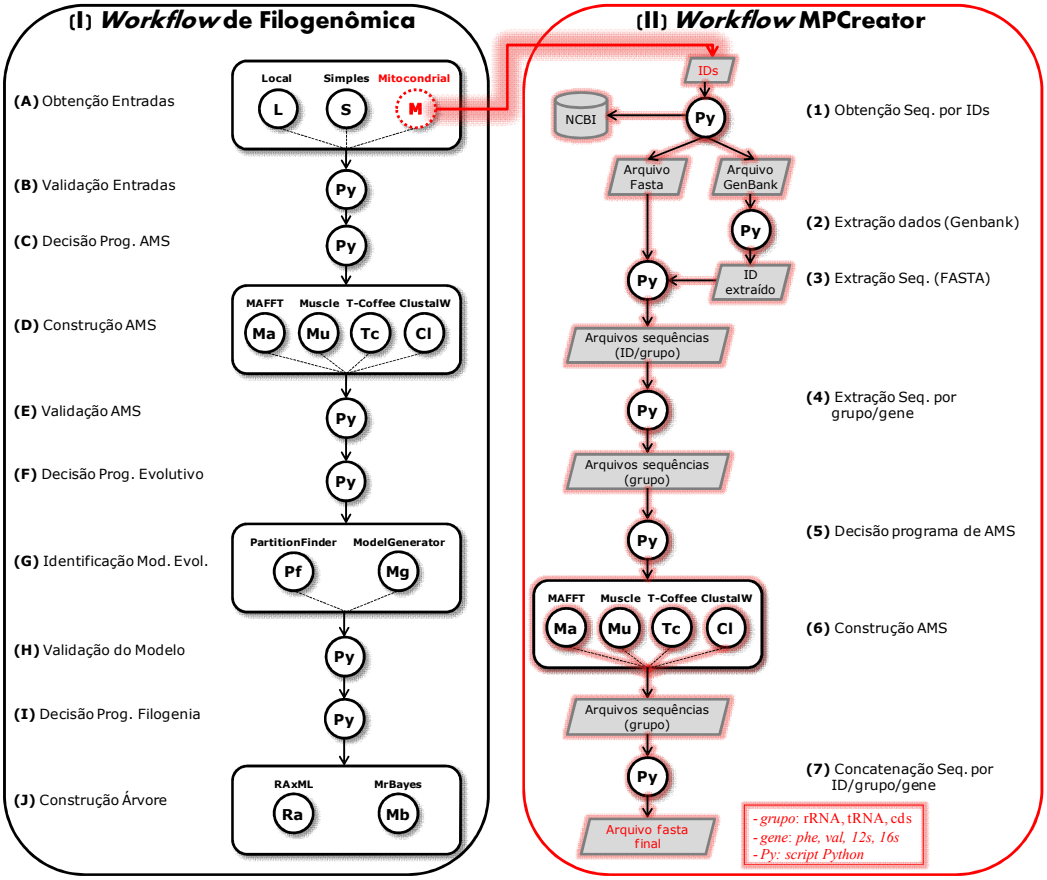

Figura 1. Vista conceitual de um experimento de filogenômica. Na esquerda (I) *Workflow* de filogenômica e na direita (II) *Workflow* MPCreator

O MPCreator recebe como entrada um grupo de ID de genomas mitocondriais, cria acesso ao NCBI (*web* ou local), obtém arquivos, trata e integra dados e gera arquivos agrupados por nomenclatura de *features*. Por questões didáticas nesse artigo, o MPCreator segue 2 níveis de *features* usados para minerar dados seguindo a nomenclatura dos arquivos (Figura 2): por *grupo* (CDS, D-loop, rRNA, tRNA) e por *subgrupo* ("/product" do GenBank *e.g.*, tRNA-Phe, rRNA12s).

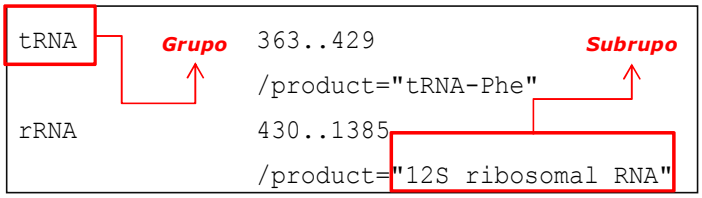

Figura 2. *Grupos* e *subgrupos* no formato GenBank

A atividade 1 obtém os arquivos formatos GenBank e FASTA das sequências dos genomas mitocondriais do NCBI, para cada ID fornecido, usando a API do NCBI - *Entrez Direct: E-utilities*[7]. A atividade 2 extrai para cada ID do arquivo GenBank, os metadados *grupo*, *subgrupo* e índice ou posição da sequência nucleotídica *e.g.*, 363...429.

A atividade 3 extrai, para cada ID, as sequências do arquivo FASTA e o *grupo* e gera arquivos formato FASTA com as sequências agrupadas por ID (*e.g.*, KM926619, KM146616) e *grupo* (Figura 3 A). A atividade 4 extrai as sequências FASTA e as agrupa cruzando com informações do *grupo* e *subgrupo* ao qual pertencem e gera arquivos FASTA (Figura 3 B) a serem alinhados na próxima atividade.

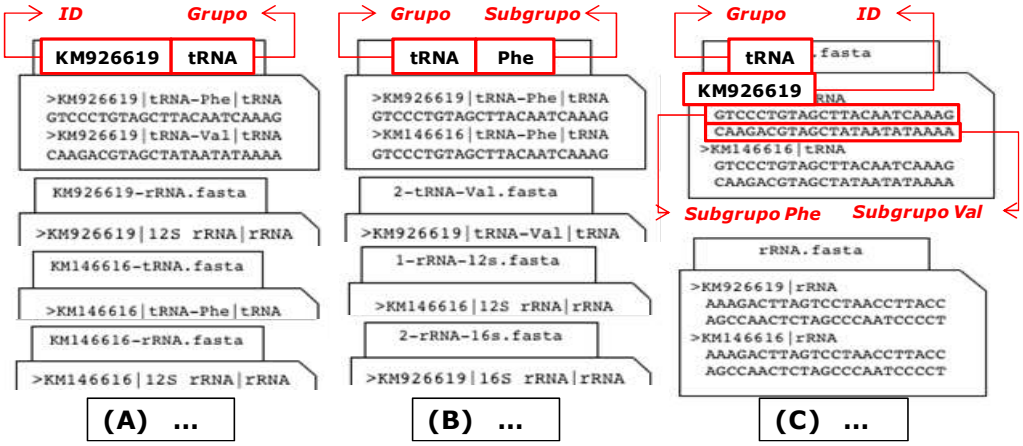

Figura 3. Geração de arquivos FASTA organizados por: (A) ID e *grupos*, (B) *grupos* e *subgrupos* e (C) *subgrupos* concatenados

A atividade 5 toma a decisão sobre qual programa de alinhamento múltiplo de sequências (AMS) será usado (MAFFT, Muscle, ClustalW, T_Coffee). A atividade 6 executa o programa de AMS escolhido pela atividade 5 usando os arquivos FASTA gerados pela atividade 4. A atividade 7 concatena os AMS de forma ordenada por ID, *grupo* e *subgrupo* (Figura 3 A, B) e concatena arquivos com os *subgrupos* (Figura 3 C).

7 https://www.ncbi.nlm.nih.gov/home/develop/api/

4. Resultados

O MPCreator exige apenas como entrada os índices dos arquivos, programa de AMS e diretório de saída. *Transparente*: de código aberto e escrito em Python. *Independente*: pode ser acoplado na composição de qualquer *workflow*. *Eficiente*: executando 1 arquivo (i) na análise quantitativa, o MPCreator levou 1 minuto *versus* 30 minutos pela forma habitual/manual do especialista e (ii) na análise qualitativa, as sequências produzidas por ambos, o MPCreator e o especialista, geraram composições de base e comprimentos idênticos. O processador usado nas execuções é Ryzen 3, geração 2.200, 8 *cores*, 8 GB de RAM e 240 Gb de armazenamento SSD.

O MPCreator é executado via linha de comando com 2 opções e gera as telas da Figura 4: (1) *Automática*, com o comando "*python3 MPCreator.py IDFile.txt*", sendo *IDFile.txt* o arquivo que contém os ID. (2) *Iterativa com Usuário*, o comando "*python3 MPCreator.py*" requer que o usuário forneça as informações: (i) ID ou arquivo contendo ID; (ii) diretório de saída que armazena genomas mitocondriais do GenBank e resultados gerados e (iii) programa de AMS que realizará o alinhamento.

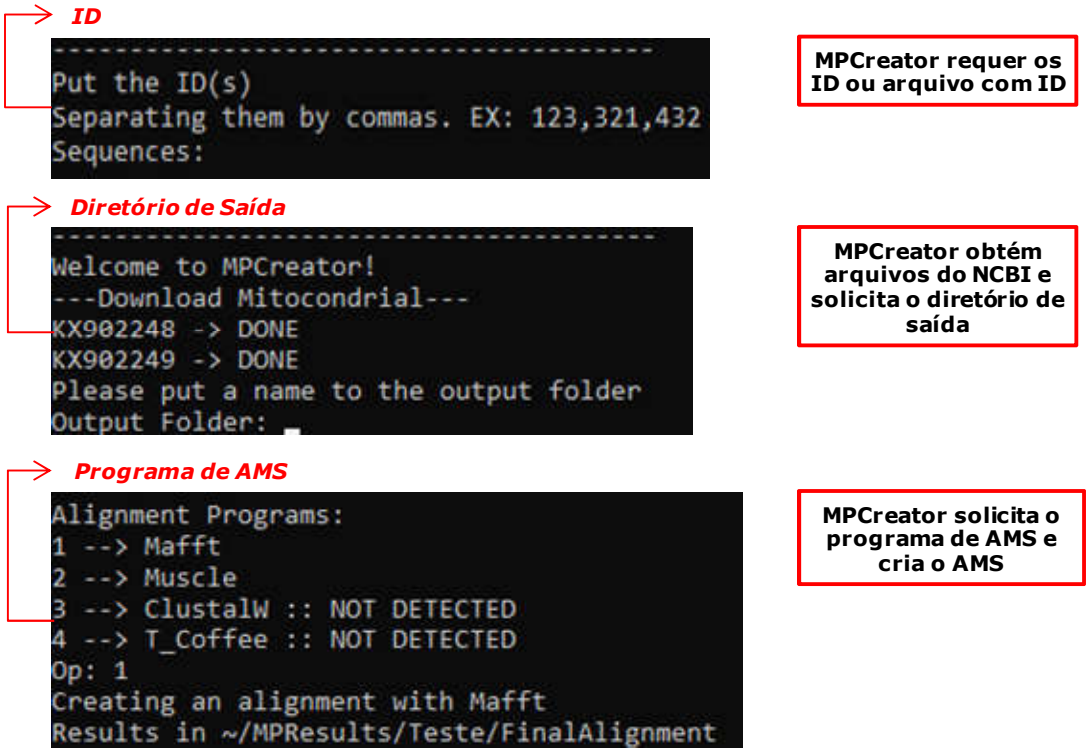

Figura 4. MPCreator requer os ID de interesse

O MPCreator foi testado com 120 genomas mitocondriais da família Columbidae (aves, pombos). A Figura 5 mostra que o tempo de execução é crescente e dependente ao número de entradas. Os algoritmos usados na construção de AMS possuem complexidade de tempo e espaço, o que leva a um alto tempo de execução e uso de memória, pelo que técnicas de paralelismo e distribuição de tarefas são necessárias.

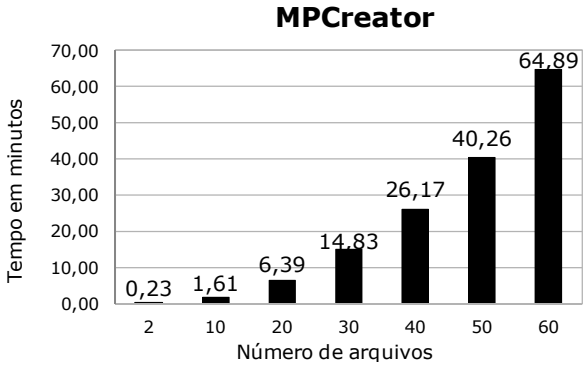

Figura 5. Tempo de execução do MPCreator

5. Conclusões e Trabalhos Futuros

Técnicas de processamento paralelo em plataformas híbridas de GPU/*multicores* e a alocação distribuída de tarefas serão acopladas ao MPCreator para reduzir a demanda de tempo e alocação de memória. O MPCreator v2 está migrando para o sistema de gerência para SAMbA[8], uma extensão do Apache Spark para ambientes de processamento de alto desempenho. Funcionalidades do AnnotationBustR mostram-se também interessantes e potencialmente úteis para serem testadas.

Agradecimentos

Este trabalho foi parcialmente financiado pelos projetos Universal MCTI/CNPq nº 01/2016 processo 429328/2016-8 e JCNE/FAPERJ nº 03/2017 processo 232985. A.E.R.S. é financiado pela CAPES com uma bolsa PNPD.

Referências

Attwood, T. K., Blackford, S., Brazas, M. D., Davies, A. e Schneider, M. V. (2017). A global perspective on evolving bioinformatics and data science training needs. *Briefings in Bioinformatics*, p. bbx100–bbx100.

Fingert, H. J. (2018). Expanding Role of Data Science and Bioinformatics in Drug Discovery and Development. *Clin. Pharmac. & Therap.*, v. 103, n. 1, p. 47–49.

Navale, V. e Bourne, P. E. (2018). Cloud computing applications for biomedical science: A perspective. *PLOS Computational Biology*, v. 14, n. 6, p. e1006144.

Rubinoff, D. e Holland, B. S. (2005). Between Two Extremes: Mitochondrial DNA is neither the Panacea nor the Nemesis of Phylogenetic and Taxonomic Inference. *Systematic Biology*, v. 54, n. 6, p. 952–961.

Wang, Z. e Wu, M. (2015). An integrated phylogenomic approach toward pinpointing the origin of mitochondria. *Scientific Reports*, v. 5, n. 1.

Wernersson, R. (2005). FeatureExtract--extraction of sequence annotation made easy. *Nucleic Acids Research*, v. 33, n. Web Server, p. W567–W569.

Yin, Z., Lan, H., Tan, G., *et al*. (2017). Computing Platforms for Big Biological Data Analytics: Perspectives and Challenges. *Comp. Struct. Biotech. J.*, v. 15, p. 403–411.

[8] https://github.com/UFFeScience/SAMbA