CBIS'24
XX Congresso Brasileiro de Informática em Saúde
08/10 a 11/10 de 2024 - Belo Horizonte/MG - Brasil

Um *workflow* para análise evolutiva acelerada de sequências genéticas

A workflow for accelerated evolutionary analysis of genetic sequences

Un *workflow* para el análisis evolutivo acelerado de secuencias genéticas

Felipe Santiago Carraro Eduardo[1], Igor dos Santos Rosa da Silva [1], Renan Pereira Souza[1], Alexandre da Costa Sena[2]

1 Aluno de Mestrado, Instituto de Matemática e Estatística (IME), Universidade do Estado do Rio de Janeiro (UERJ), Rio de Janeiro (RJ), Brasil.
2 Prof. Dr., Instituto de Matemática e Estatística (IME), Universidade do Estado do Rio de Janeiro (UERJ), Rio de Janeiro (RJ), Brasil.

Autor correspondente: (Prof. Dr.) Alexandre da Costa Sena
E-mail: asena@ime.uerj.br

Resumo

Por natureza, os vírus estão em constante mutação. Apesar de uma grande parte das mutações não alterar o comportamento de um vírus, algumas dessas mutações podem gerar novas variantes que, por exemplo, podem fazer um vírus se espalhar mais rapidamente. Uma maneira de verificar essa evolução é através de modelos evolutivos. Assim, o objetivo deste trabalho é avaliar a evolução genética dos vírus. O método usado é o alinhamento par a par das cadeias do vírus, seguido do cálculo da distância genética. Ainda, para permitir a avaliação de uma grande quantidade de sequência, essas duas etapas são implementadas através de um *Workflow*. Os resultados obtidos através de dois estudos de casos utilizando os vírus SARS-COV-2 e monkeypox, mostraram não só o excelente desempenho do *workflow*, diminuindo consideravelmente o tempo de execução das análises, mas também a evolução das suas sequências genéticas.

Descritores: Workflow; Distância Genética; Big Data

Abstract

By nature, viruses are constantly mutating. Although most mutations do not change the behavior of a virus, some of these mutations can generate new variants, which, for example, can make a virus spread more quickly. One way to verify this evolution is through

evolutive models. Therefore, the objective of this work is to evaluate the genetic evolution of viruses. The method used is pairwise alignment of the virus sequences, followed by calculation of genetic distance. Furthermore, to allow the evaluation of a large amount of sequence, these two steps are implemented through a Workflow. Results obtained through two case studies using the SARS-COV-2 and monkeypox viruses, showed not only the excellent performance of the workflow, considerably reducing the analysis execution time, but also the evolution of their genetic sequences.
Keywords: Workflow; Genetic Distance; Big Data

Resumen

Por naturaleza, los virus están en constante mutación. Aunque la mayoría de las mutaciones no cambian el comportamiento de un virus, algunas de estas mutaciones pueden generar nuevas variantes. Una forma de comprobar esta evolución es a través de modelos evolutivos. Por tanto, el objetivo de este trabajo es evaluar la evolución genética de los virus. El método utilizado es el alineamiento por pares de las secuencias del virus, seguido del cálculo de la distancia genética. Además, para permitir la evaluación de una gran cantidad de secuencia, estos dos pasos se implementan a través de un *Workflow*. Los resultados muestran el excelente rendimiento del *Workflow*, capaz de aprovechar los múltiples núcleos de procesamiento disponibles en las nuevas arquitecturas, sino también la evolución de sus secuencias genéticas.
Descriptores: Workflow; Distancia genética; Grandes datos

Introdução

Nos últimos anos, o mundo tem sido profundamente impactado por surtos epidemiológicos causados por vírus. Por exemplo, é possível destacar o surto de varíola dos macacos em 2022 [(1)] e a pandemia de COVID-19 em 2019 [(2)]. Uma das razões que torna o processo de controle do surto ou pandemia ainda mais complexo é a alta taxa de mutação dos vírus, podendo ser até um milhão de vezes maior que a dos seus hospedeiros [(3)]. Apesar de em grande parte dos casos uma mutação não alterar o comportamento dos vírus, em uma parte significativa pode acarretar a criação de

variantes que podem fazer com que o vírus se espalhe mais rapidamente ou torná-lo resistente à vacina.

Uma forma de verificar a evolução dos vírus é através de modelos evolutivos, como a análise filogenética e distância genética. Esses modelos exigem uma interpretação tanto biológica quanto matemática e são baseadas em distância evolucionária entre sequências de RNA/DNA que podem ser obtidas a partir do alinhamento genético [4]. Entretanto, apesar das cadeias de RNA/DNA dos vírus não serem muito grandes, para avaliar a evolução de grandes quantidades de sequências (e.g. o site do NCBI já contém quase 9 milhões de cadeias do vírus SARS-CoV-2) é necessário aproveitar ao máximo os recursos computacionais disponíveis, por se tratar de um processo computacional muito custoso.

Nesse contexto, o objetivo deste trabalho é propor e implementar um *workflow* para avaliar a evolução de sequências genéticas. O *workflow* é composto de duas fases. Enquanto na primeira fase são realizados alinhamentos par a par das cadeias de DNA, na segunda fase os resultados dos alinhamentos são utilizados para calcular a distância genética entre as cadeias. Através dessa proposta é esperado não só conseguir métricas para medir a evolução das cadeias, mas principalmente realizar esses cálculos de forma eficiente, permitindo a avaliação de uma grande quantidade de cadeias.

A avaliação do *workflow* proposto é realizada através de dois estudos de caso que avaliam não só o desempenho computacional do *workflow* proposto, mas também a evolução genética dos vírus SARS-CoV-2 e *Monkeypox*. Os resultados mostram que a solução proposta é muito eficiente, acelerando consideravelmente o tempo de execução. Por exemplo, a avaliação da evolução de 50.000 sequências do vírus SARS-CoV-2 foi reduzida de ≈ 1 dia para apenas ≈ 45 minutos, em uma máquina com 36 processadores. Além disso, as análises das cadeias genéticas dos vírus SARS-CoV-2 e *monkeypox* mostraram a viabilidade do *workflow* para avaliar a evolução de vírus, mas também de qualquer sequência genética.

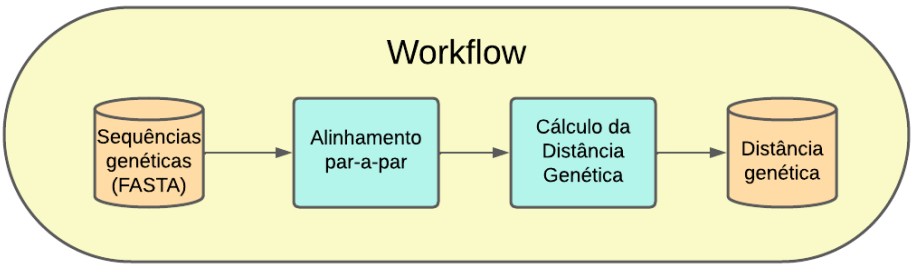
No contexto do desenvolvimento de *workflows* para aplicações de bioinformática, foi possível identificar o trabalho apresentado em [5], que propôs e implementou uma metaferramenta para alinhamento múltiplo. Em função da granularidade distinta das tarefas, o desempenho ficou limitado pela tarefa mais longa. Por sua vez, a ferramenta

ViralFlow [6] apresenta um *workflow* para vigilância genômica do vírus SARS-CoV-2, automatizando todos os passos necessários para análise genômica. Embora esta ferramenta esteja preparada para ser executada em vários ambientes computacionais, ela é limitada ao vírus SARS-CoV-2. Ainda, o trabalho em [7] apresenta um pipeline fácil de usar e customizável para detectar genomas virais a partir do sequenciamento Nanopore. Apesar deste trabalho ser muito importante para a identificação de vírus, ele só é aplicável para genomas provenientes de sequenciamento Nanopore ou Illumina.

Métodos

Uma visão geral da metodologia proposta para análise evolutiva acelerada de sequências genéticas pode ser vista na Figura 1.

Figura 1 - Visão geral da metodologia adotada neste trabalho.



Inicialmente, a entrada da solução proposta são cadeias genéticas (i.e. RNA, DNA ou Proteínas) a serem avaliadas em arquivos do tipo FASTA, que é o formato padrão na área da bioinformática. A primeira etapa da análise evolutiva é o alinhamento par a par da sequência de referência com as sequências a serem analisadas. Em seguida, a partir dos alinhamentos gerados, as distâncias genéticas são calculadas e armazenadas em um arquivo. Porém, para prover uma execução acelerada, escalável e capaz de lidar com uma grande quantidade de dados, toda a abordagem foi implementada na forma de um *workflow*. As subseções a seguir detalham cada etapa da metodologia adotada neste trabalho.

Base de Dados

Neste trabalho foram selecionadas sequências de RNA e DNA dos vírus *monkeypox* e SARS-CoV-2 que foram submetidas no sítio do *National Center for Biotechnology Information* (NCBI)[1]. Para obter sequências mais homogêneas e de maior qualidade, foram utilizados filtros para selecionar a faixa de tamanho das sequências, além de considerar apenas hospedeiros humanos e com nucleotídeos completos e, por fim, relativos a um intervalo de datas específico.

Para o vírus *monkeypox* foram escolhidas apenas sequências com comprimento entre 197.150 e 197.250 nucleotídeos no período de 2001 a 2023. Por sua vez, para o vírus SARS-CoV-2 foram escolhidas sequências com comprimento entre 29.900 e 29.903 nucleotídeos no período de 2020 a 2023. Em função da quantidade de sequências disponíveis no NCBI e principalmente do tamanho das sequências, foram selecionadas 1.000 sequências do vírus *monkeypox* e 50.000 sequências do vírus SARS-CoV-2. Na Seção Resultados e Discussões os conjuntos de sequências avaliados são descritos.

Workflow

Um *Workflow* pode ser definido como uma série ordenada de tarefas que devem ser realizadas para que um determinado objetivo seja atingido. Por sua vez, neste trabalho é utilizado um *Workflow* computacional que é usado para descrever os métodos complexos de várias etapas usadas para coleta de dados, preparação de dados, análise, modelagem preditiva e simulação para gerar novos dados.

Existem várias ferramentas para implementação e gerenciamento de *Workflows* computacionais. Em função das etapas do Workflow envolverem aplicações da área de bioinformática e, ao mesmo tempo, tratar grande quantidade de dados, optou-se por uma ferramenta capaz de lidar com esse tipo de aplicação, realizando as etapas do *Workflow* o mais rápido possível. Quando comparadas com os scripts ou códigos em linguagens de programação, as ferramentas para gerenciamento de Workflows fornecem proveniência e rastreabilidade dos dados. Além disso, seu código é mais portável e escalável, além de mais fácil manutenção.

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

Neste trabalho, dentre as várias ferramentas que seguem os padrões estabelecidos pela *Workflows Community Initiative*[2], foi utilizada a ferramenta *Nextflow* [8], em função dela ter sido desenvolvida para gerenciar *workflows* na área de bioinformática, provendo a execução paralela eficiente e tolerante a falhas. Além disso, *Nextflow* permite suporte nativo para a execução de *containers* em ambientes de nuvens computacionais.

O *workflow* proposto e implementado neste trabalho é composto por dois processos distintos (retângulos azuis na Figura 1.) responsáveis, respectivamente, pelo alinhamento par a par e cálculo da distância genética. Embora a ferramenta *Nextflow* facilite a criação de um Workflow, esta tarefa é semelhante ao desenvolvimento de códigos em linguagens de programação ou *script*. O processo *Nextflow* responsável pelo alinhamento das cadeias genéticas foi programado para realizar esta tarefa em paralelo de acordo com a quantidade de processadores disponíveis no ambiente computacional utilizado. Por sua vez, o segundo processo calcula a distância genética a partir dos dados gerados pelos alinhamentos. Esses dois processos são detalhados a seguir.

O alinhamento genético é um processo fundamental na biologia, permitindo a comparação minuciosa entre duas ou mais sequências de DNA, RNA ou proteínas para identificar similaridades e divergências. Neste trabalho, para realizar os alinhamentos par a par entre a cadeia de referência e as cadeias analisadas foi utilizado o MASA (*Multi-Platform Architecture for Sequence Aligner*), que é um *framework* para realizar o alinhamento ótimo de duas sequências. Ele apresenta um ambiente de alinhamento de sequências flexível e altamente customizável, capaz de ser executado em diferentes plataformas de hardware e software [9]. É importante destacar que o processo que calcula o alinhamento par a par entre a cadeia de referência e todas as sequências a serem avaliadas é muito custoso computacionalmente. Porém, cada alinhamento entre a cadeia de referência e uma das sequências é independente. Assim, a ferramenta *Nextflow* cria um processo para cada processador disponível no ambiente computacional, otimizando a execução.

Ao final deste processo, é gerado um arquivo contendo os resultados dos alinhamentos. Assim, o segundo processo recebe este arquivo como entrada e avalia o

[2] https://workflows.community/

processo evolutivo das sequências. Dentre os fenômenos evolutivos, o processo de substituição é o mais utilizado pelos modelos para explicar o processo evolutivo, uma vez que os mecanismos de substituição apresentam uma relevância no processo evolutivo e existe uma maior quantidade de modelos probabilísticos e estatísticos que explicam este processo. O número de substituições de nucleotídeos que se acumularam nas sequências desde a divergência pode ser calculado através da distância genética [4], que pode ser vista na equação 1.

$$d_{ij} = \frac{-3}{4} ln\left(1 - \frac{4}{3} p_{ij}\right) \quad (1)$$

A proporção p_{ij} representa a probabilidade de diferenças nucleotídicas entre as sequências i e j, indicando a chance de dois nucleotídeos aleatórios, um de cada sequência, serem diferentes. n_p é o número de nucleotídeos distintos entre i e j, calculado pela contagem de pares de nucleotídeos que diferem nessas sequências. n representa o comprimento total das sequências i e j, obtido pela soma do número total de nucleotídeos em cada sequência. O resultado final do *workflow* é um arquivo csv contendo as descrições das cadeias e suas distâncias genéticas.

Resultados e Discussões

A análise do *workflow* proposto e implementado neste trabalho foi realizada através de dois estudos de casos distintos, utilizando os vírus SARS-CoV-2 e *monkeypox*. A escolha desses dois vírus se deve a relevância recente em função da Pandemia de COVID-19 [2] e o surto de varíola dos macacos em 2022 [1]. Além disso, outra razão importante para escolha desses dois vírus foi a diferença do seu tipo de genoma e quantidade de nucleotídeos. Enquanto no SARS-CoV-2 o material genético é o RNA e sua quantidade de nucleotídeos é ≈ 29.900, o *monkeypox* é um vírus de DNA e sua quantidade de nucleotídeos é ≈ 197.000. A análise foi dividida em duas etapas distintas. A eficiência do Workflow é avaliada na subseção Análise do Desempenho do *Workflow*, enquanto que a evolução dos dois vírus é avaliada na subseção Análise Evolutiva dos Vírus SARS-CoV-2 e *monkeypox*.

Análise do Desempenho do *Workflow*

Para avaliar o desempenho do *Workflow* foram selecionadas 50.000 sequências de RNA do vírus SARS-CoV-2 e 1.000 sequências de DNA do vírus *monkeypox*, disponíveis no sítio do NCBI. A maior quantidade de sequências do vírus SARS-CoV-2 tem duas razões: quantidade de amostras disponíveis no sítio do NCBI e diferença na quantidade de nucleotídeos das duas sequências. Destes conjuntos, foram selecionadas amostras de 100, 500 e 1.000 sequências do vírus *monkeypox* e 1.000, 10.000, 20.000, 30.000, 40.000 e 50.000 sequências do vírus SARS-CoV-2 para avaliar o impacto da quantidade de sequências no desempenho do *Workflow*.

Os experimentos foram realizados em um servidor composto de 2 processadores *Intel® Xeon® Gold 5220* com 18 núcleos físicos em cada processador, com 128 GB de memória. Além disso, possui tecnologia *Hyper-Threading Intel®*, permitindo a execução simultânea de 72 alinhamentos genéticos. Para avaliar a escalabilidade, o *Workflow* foi executado com 1, 2, 4, 9, 18, 36 e 72 processos. Os resultados dos tempos de execução para os diversos conjuntos podem ser observados nas Tabelas 1 e 2. Verifica-se que o comportamento do *Workflow* em um ambiente paralelo com memória compartilhada foi bastante consistente para as duas análises, onde os tempos de execução diminuem proporcionalmente à medida que se aumenta a quantidade de processos (linhas das tabelas). Ainda, é possível observar que o tempo aumenta proporcionalmente à medida que se aumenta a quantidade de sequências analisada (colunas das tabelas).

Tabela 1 - Tempos de execução do workflow em segundos (SARS-CoV-2).

SARS-Cov2							
Tamanho das sequências: 29900 a 29903							
	Quantidade de Processos						
nº de sequências	**1**	**2**	**4**	**9**	**18**	**36**	**72**
1000	1790	908	474	224	131	76	61
10000	17874	9104	4815	2216	1279	731	548
20000	35880	18184	9697	4424	2551	1471	1094
30000	53910	27775	14542	6628	3810	2200	1644
40000	71872	36465	19453	8845	5070	2935	2194
50000	89739	45620	24124	11092	6351	3668	2749

Comparando os tempos de execução das duas tabelas, como esperado, o tempo para realizar a análise genética do vírus *monkeypox* é muito maior, uma vez que o tamanho da cadeia desse vírus é muito maior do que o tamanho do vírus SARS-CoV-2. Por exemplo, enquanto que o tempo para calcular a evolução de 1.000 sequências do SARS-CoV-2 com apenas 1 processo foi de apenas 1790 segundos, o tempo para calcular a evolução de 1.000 sequências do *monkeypox* com apenas 1 processo foi de 56.024 segundos (mais de 15 horas), ou seja, aproximadamente 31 vezes maior.

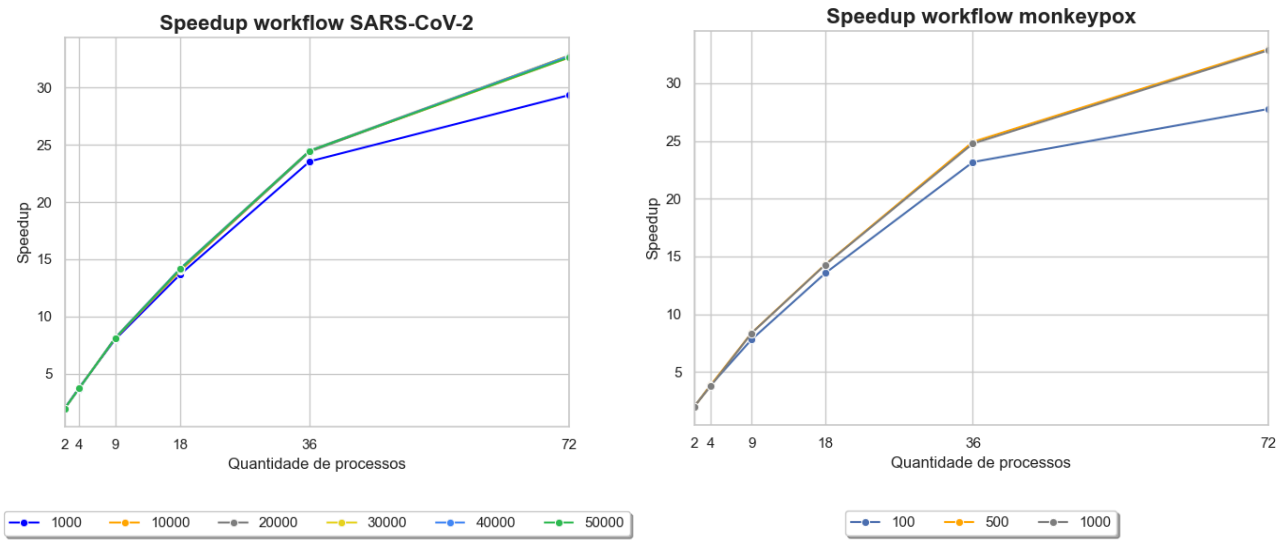
Tabela 2 - Tempos de execução do workflow em segundos (monkeypox).

Monkeypox							
Tamanho das sequências: 197000 a 197400							
	Quantidade de Processos						
nº de sequências	**1**	**2**	**4**	**9**	**18**	**36**	**72**
100	5608	2825	1459	719	414	242	202
500	28206	14332	7350	3368	1971	1132	856
1000	56024	28222	14909	6709	3922	2262	1706

A explicação para esse aumento 31 vezes maior no tempo de execução, enquanto que o tamanho da cadeia do vírus *monkeypox* é apenas aproximadamente 6 vezes maior do que a cadeia do vírus SARS-CoV-2 está na complexidade do alinhamento par a par. O alinhamento gerado pelo MASA (Seção Métodos) é ótimo, ou seja, produz a melhor solução computacionalmente possível. Em função disso, a complexidade do algoritmo é O (n^2), onde n é o tamanho da sequência. Por outro lado, ao avaliar a evolução de uma quantidade grande de cadeias, o tempo pode aumentar bastante, mesmo para sequências do vírus SARS-CoV-2, que são bem menores do que as sequências do *monkeypox*. Por exemplo, o tempo para calcular a evolução de 50.000 sequências do SARS-CoV-2 com apenas 1 processo foi de aproximadamente 1 dia (24,9 horas) e foi reduzido pela solução proposta para apenas ≈ 45 minutos, executando com 72 processos.

Para avaliar a escalabilidade do *workflow*, foi utilizada a métrica de $speedup = \frac{t(1)}{t(p)}$, onde $t(1)$ é o tempo de execução em 1 processador e $t(p)$ é o tempo de execução em p processadores. O gráfico com os *speedups* pode ser visualizado na Figura 2.

Figura 2 - Speedup dos workflows dos vírus SARS-CoV-2 (a esquerda) e monkeypox (a direita).



O comportamento das execuções para as análises das evoluções genéticas dos dois vírus foi bem similar mostrando claramente a escalabilidade da solução proposta. As execuções apresentaram um aumento linear, tanto para a análise do vírus *monkeypox* (a direita), como também para o vírus SARS-CoV-2 (a esquerda). Duas oscilações no desempenho podem ser observadas. Em primeiro lugar, a piora do desempenho da execução com 72 processos ocorre em função da utilização de *threads* e não processadores, em função da máquina possuir apenas 36 núcleos de processamento. Apesar da tecnologia *hyper-threading* permitir a execução de duas *threads* para cada núcleo físico real, o desempenho é inferior. Em segundo lugar, o desempenho ligeiramente pior para as execuções com as menores quantidades de sequências se deve simplesmente a menor quantidade de instâncias, o que fez com que a diferença na distribuição dos processos do *workflow* resultasse em um ligeiro aumento no tempo de execução.

Análise Evolutiva dos Vírus SARS-CoV-2 e *monkeypox*

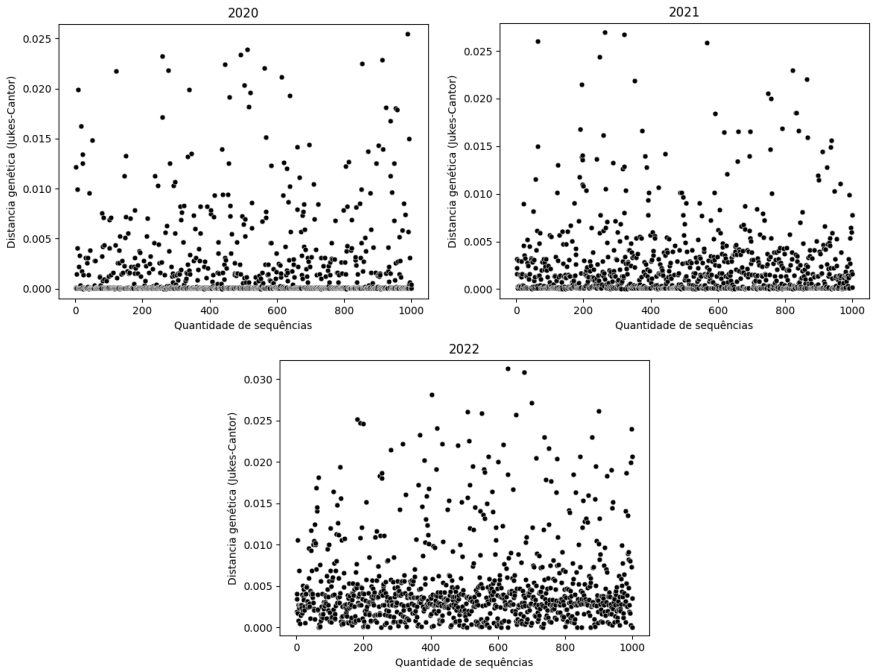
As duas subseções a seguir apresentam as evoluções dos Vírus SARS-CoV-2 e *monkeypox* utilizando a distância genética. As duas análises utilizaram conjuntos compostos de 1.000 sequências, porém o tipo de avaliação realizado para cada um dos

vírus foi bem diferente. A análise do vírus SARS-CoV-2 foi feita comparando a cadeia de referência do vírus com 3 conjuntos de sequências relativas aos anos de 2020, 2021 e 2022. Por sua vez, a evolução do vírus *monkeypox* foi feita comparando o único conjunto de 1.000 sequências disponível com as duas cadeias de referência existentes.

Distância Genética do Vírus SARS-CoV-2

A distância genética do vírus SARS-CoV-2 foi calculada utilizando a sequência de referência NC_045512 encontrada em Wuhan, na China, disponível no sítio do NCBI. Para mensurar a evolução do RNA, foi calculada a distância genética de 3 conjuntos de 1.000 sequências relativas aos anos de 2020, 2021 e 2022, em relação a sequência de referência. O objetivo é avaliar o aumento da distância genética a cada ano da pandemia. A Figura 3 apresenta as distâncias genéticas para os anos 2020, 2021 e 2022.

Figura 3 - Distância genética de 1.000 sequências do vírus SARS-CoV-2 dos anos de 2020, 2021 e 2022 para a cadeia de referência.



Analisando os gráficos, o comportamento é bem similar, porém é possível verificar que a medida que o ano aumenta as distâncias genéticas vão se distanciando do 0 (zero). Para uma análise mais detalhada, foram calculadas as médias e medianas das distâncias genéticas das 1.000 sequências para cada ano. Enquanto a média aritmética representa o

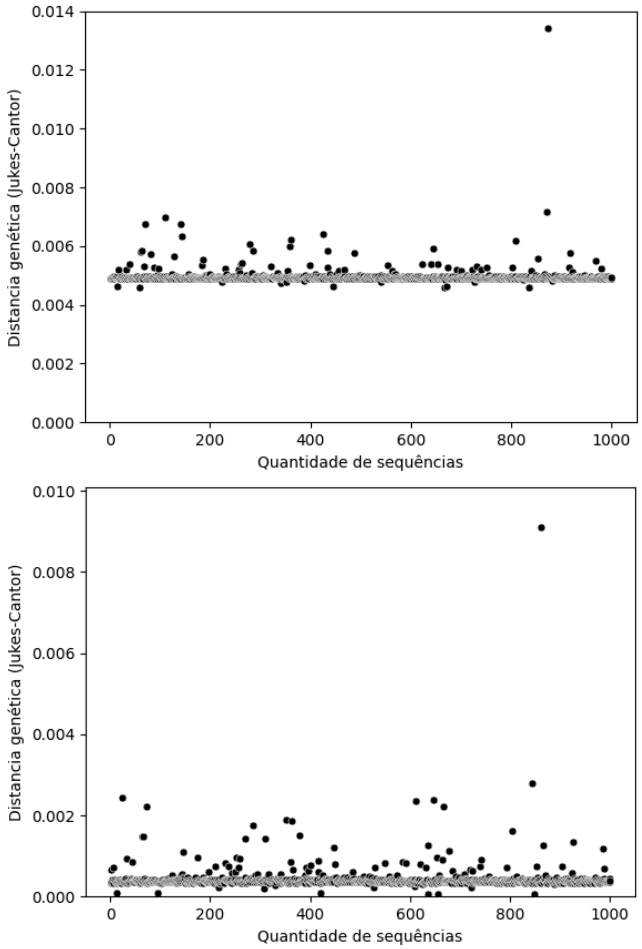
valor média e pode ser influenciado por valores *outliers*, a mediana representa o valor central da amostra que sofre menos influência dos *outliers*. As médias e medianas para os 3 conjuntos de sequências para os anos 2020, 2021 e 2022, respectivamente, foram 0,001985 e 0,0001064, 0,002472 e 0,001197, e, por fim, 0,004940 e 0,003323.

Os valores das médias e medianas das distâncias genéticas aumentam a cada ano. Com relação, as médias aritméticas é possível notar um aumento grande do ano de 2020 para o ano de 2022, aumentando de 0,001985 para 0,004940. Por sua vez, apesar da mediana também apresentar esse mesmo comportamento, os valores são menos afetados por valores *outliers*. Por exemplo, para o ano de 2020, enquanto a média foi quase 0.002, a mediana ficou próxima de 0.0001, representando que metade das amostras desse ano ficaram abaixo desse valor e a outra metade acima. Ou seja, o valor da média foi aproximadamente 20 vezes maior, em função dos *outliers*. Além disso, é possível notar uma evolução nas amostras do ano de 2022, através do aumento significativo tanto da média como da mediana. Uma razão para esta diferença acentuada pode ser em função de uma grande parte das amostras do ano de 2022 ser da *ômicron* (identificada em novembro de 2021) que possui características distintas de outras variantes.

Distância Genética do Vírus *monkeypox*

Enquanto que no sítio do NCBI existem quase 9 milhões de sequências de RNA do vírus SARS-CoV-2, o que permitiu criar diferentes grupos de sequências de vários tamanhos, divididos por ano, não foi possível realizar esse mesmo tipo de análise para o vírus *monkeypox* devido ao número reduzido de amostras, onde, na sua maioria, são relativas ao ano de 2022. Assim, foi criado apenas um conjunto de 1.000 sequências do vírus *monkeypox*, que foi comparado com as duas sequências de referência disponíveis no NCBI, a NC_003310.1 do ano de 2001 encontrada no Zaire e a NC_063383.1 do ano de 2022 encontrada na Nigéria. Ou seja, foi calculada a distância genética das 1.000 sequências em relação as duas cadeias de referência, utilizando o *workflow* implementado neste trabalho. Os valores das distâncias genéticas podem ser observados na Figura 4.

Figura 4 - Distância genética de 1.000 sequências do vírus monkeypox para cadeia de referência do ano 2001 (a esquerda) e 2022 (a direita).



Como esperado, ao analisar os dois gráficos fica muito clara a diferença da evolução do vírus em relação as duas cadeias de referência. Enquanto que a distância genética para a sequência do ano de 2022 fica próxima do 0, ela aumenta significativamente em relação a amostra do ano de 2001, ficando próxima de 0,005. A razão para este comportamento é a grande diferença de tempo entre as duas cadeias de referência. Enquanto a cadeia de referência de 2022 é a sequência que deu origem ao surto de 2022 e, consequentemente, muito próximas das 1.000 sequências avaliadas neste trabalho, a cadeia de 2001 é referente a um surto que aconteceu no Zaire e mais distante do vírus *monkeypox* atual.

Para uma análise mais detalhada foram calculadas as médias e medianas para as cadeias de referência de 2001 e 2022 que foram, respectivamente, 0,00496985 e 0,00492973, e 0,0004357 e 0,0003854. Uma característica distinta das outras médias e medianas apresentas neste trabalho, é que a média e mediana para a cadeia de referência de 2001 ficaram muito próximas. A principal razão para tal comportamento é a grande diferença de tempo entre a cadeia de referência e as sequências comparadas o que faz com que a diferença dos valores médios e os *outliers* seja menor quando comparada com valores onde a diferença temporal seja menor. Outra observação importante é com relação a diferença entre a distância genética do SARS-CoV-2, apresentada na subseção anterior, e a distância genética do *monkeypox*. Por exemplo, a média das distâncias genéticas das sequências do vírus SARS-CoV-2 de 2022 para a sequência de referência de Wuhan de 2019 foi de 0,004940, enquanto que a média das distâncias genéticas das sequências do vírus *monkeypox* de 2022 para a sequência de referência da Nigéria de 2022 foi de apenas 0,0004357. A principal razão para explicar esse comportamento é que a base do vírus *monkeypox* é o DNA (e não o RNA) que tem uma taxa de mutação muito mais lenta [(3)].

Conclusão

Os modelos evolutivos são ferramentas importantes para os cientistas avaliarem as mutações genéticas, em especial, dos vírus. Embora os tamanhos das sequências desses organismos não sejam grandes, a avaliação de uma quantidade grande de sequências é uma tarefa computacionalmente custosa. Assim, este trabalho propôs e implementou um *workflow* para análise evolutiva acelerada de sequências genéticas. Os resultados obtidos mostram claramente o excelente desempenho do *workflow*. Por exemplo, o tempo da avaliação de 1.000 sequências do vírus *monkeypox* foi reduzido de ≈ 15 horas para apenas ≈ 28 minutos em uma máquina com 36 processadores, o que demonstra a capacidade da solução proposta ser escalável e, assim, poder lidar com uma grande quantidade de dados. Além disso, dois estudos de caso distintos mostraram a evolução dos vírus *monkeypox* e SARS-CoV-2, onde foi possível perceber a evolução mais rápida do vírus SARS-CoV-2, possivelmente em função da sua base genética ser o RNA.

Agradecimentos

Os autores agradecem a CAPES e ao CNPq pelo apoio e, também, a FAPERJ através do edital APQ1 26/2021 (Número do processo E-26/211.798/2021).

Referências

1. Farahat RA, Sah R, El-Sakka AA, others. Human monkeypox disease (MPX). Infez Med. 2022; 30: p. 372-391.

2. Hu B, Guo H, Zhou P, others. Characteristics of SARS-CoV-2 and COVID-19. Nature Reviews Microbiology. 2021; 19: p. 141-154.

3. Duffy S. Why are RNA virus mutation rates so damn high? PLOS Biology. 2018 August; 16: p. 1-6.

4. Verli H. Bioinformática: da Biologia à Flexibilidade Molecular. 1st ed.: Sociedade Brasileira de Bioquímica e Biologia Molecular - SBBq; 2014.

5. Junior MJ, Sena A, Rebello V. Fragmentando o DNA de Ferramentas de Alinhamento Progressivo: uma Metaferramenta Eficiente. Anais do XXIV Simp. em Sist. Comp. de Alto Desempenho; 2023; Porto Alegre, Brasil. p. 349–360.

6. Dezordi FZ, Neto AMD, Campos TL, Jeronimo PMC, Wallau GL. ViralFlow: A Versatile Automated Workflow for SARS-CoV-2 Genome Assembly, Lineage Assignment, Mutations and Intrahost Variant Detection. Viruses. 2022; 14: p. 217.

7. Kim K, Park K, Lee S, Baek SH, Lim TH, Kim J, et al. VirPipe: an easy-to-use and customizable pipeline for detecting viral genomes from Nanopore sequencing. Bioinformatics. 2023 May; 39: p. btad293.

8. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nature Biotechnology. 2017; 35: p. 316–319.

9. De O. Sandes EF, Miranda G, Martorell X, Ayguade E, Teodoro G, De Melo ACMA. MASA: A Multiplatform Architecture for Sequence Aligners with Block Pruning. ACM Trans. Parallel Comput. 2016; 2 (4): p. 1-31.