

マルチプルアライメントで利用する PC 仮想基盤の評価

安沢伸二¹

概要: ARM アーキテクチャーPC のリリースといった個人が利用するコンピューティング環境が向上しており、Windows や Linux のアプリケーションを PC 仮想環境上で動作させることが可能になっている。また、バイオインフォマティクスにおいては、インターネットから解析アプリケーションや公共 DB の大量データを入手して、目的とする多様な解析を行うことが可能である。本報告では、公共 DB から入手したウイルスゲノム配列情報を Windows アプリケーションに実装されたツールを使ってマルチプルアライメントを行い、解析で利用した各 PC 仮想基盤の実装方式比較やパフォーマンス比較を行った。

キーワード: ARM, Virtual server, COVID-19

Evaluation of PC virtual infrastructure used for multiple alignment

SHINJI YASUZAWA^{†1}

Abstract: The computing environment used by the personal, such as the release of ARM architecture PCs, is improving, and it is now possible to run Windows and Linux applications in a virtual environment. In bioinformatics, it is possible to perform various analysis with downloading a large amount of data from public databases, windows applications from the internet site. In this paper, multiple alignment is performed using the virus genome sequence information obtained from the public DB using the tool implemented in the Windows application. and Implementation methods and performance evaluation of each PC virtual infrastructure used in this analysis are discussed.

Keywords: ARM, Virtual server, COVID-19

1. はじめに

昨今、ARM アーキテクチャーPC のリリースにより個人が利用するコンピューティング環境が向上しており、Windows や Linux のアプリケーションを PC 仮想環境上で動作させることが可能になっている。現時点、利用可能である仮想基盤を調査し、CPU は Intel Core i9 と M1, Host OS は Windows 10 pro と MacOS Big Sur, 仮想基盤は VirtualBox を使用した。ARM アーキテクチャーの M1 については、現時点、唯一利用可能な仮想環境である Parallels desktop 16 for M1 Mac Technical Preview^[1](Parallels_Desktop16-1.1-49141, Parallels Desktop 16*と表記)を使用した。バイオインフォマティクスにおいては、インターネットを介して解析アプリケーションや公共 DB の大量データを入手し、多様な解析を行うことが可能である。大量データには NCBI ウイルスから COVID-19 のゲノム配列情報を 133 件、Influenza A を 56 件、Influenza B を 8 件、Influenza C 14 件をダウンロードした。ゲノム配列情報を Windows アプリケーション UGENE^[2]に用意されたマルチプルアライメント ツールの MAFFT コマンドを利用してマルチプルアライメントを行い、この解析に要する時間、仮想基盤環境の構築の容易さ、コストを対象として仮想基盤を評価した。

2. 評価環境と評価データ

(1) 評価環境

評価対象 PC についての仕様を下表に示す。

表 1 評価対象 PC

	Host	CPU	Mem	DISK	仮想基盤 (Guest OS)
T1	Windows 10 Pro	Intel Core i9 8 コア	64GB	512GB	VirtualBox 6.1 (Windows 2010 Eval)
T2	MacOS Big Sur	Intel Core i9 8 コア	64GB	2TB	VirtualBox 6.1 (Windows 2010 Eval)
T3	MacOS Big Sur	Apple M1 8 コア	16GB	1TB	Parallels Desktop 16* (Windows 10 on ARM Insider Preview)

マルチプルアライメントには、オープンソース クロスプラットフォーム バイオインフォマティクス ソフトウェア UGENE と MAFFT コマンドを使用した。下表に示す。

表 2 マルチプルアライメントソフトウェア

解析 S/W	Version	解析ツール
UGENE	V36.0	MAFFT

表 3 UGENE リソース

Type	Optimize for CPU	Memory limit	Threads limit
T1	8	1157MB	1000
T2	1	2047MB	1000
T3	2	1536MB	1000

¹ 三菱電機株式会社

(2) 評価データ

マルチプルアライメントの大量データには、参考文献^[3]にある COVID-19 を対象とした。さらに、評価用にマルチプルアライメントの計算量を増加させるため、Influenza A, Influenza B, Influenza C を対象に加えた。操作は、NCBI ウイルスにアクセスし、データをダウンロードする。COVID-19 は Taxonomy ID 2697049, オプションとして全長配列 complete を指定し Web ページ単位に 1 件毎ダウンロードした。Influenza A, Influenza B, Influenza C は、Refseq のデータを選択しダウンロードした。ウイルスゲノム配列データについて下表に示す。

表 3 ウイルスゲノム配列データ

COVID-19	133 件	(例) NC_045512.2	4094KB (FASTA)
Influenza A	56 件	(例) NC_026431.1	
Influenza B	8 件	(例) NC_002206.1	
Influenza C	14 件	(例) NC_006306.2	

3. 評価

評価対象 PC T1, T2, T3 を使い以下の手順で評価データのマルチプルアライメントを行い、処理時間を測定した。

- ① UGENE を起動し、マルチプルアライメントツールメニューから MAFFT を選択する。
- ② 評価データの入力先設定、解析結果データ出力先を設定する。
- ③ Log 出力から MAFFT コマンド起動時間と終了時間を確認する。

測定結果を以下に示す。

表 4 測定結果

Type	仮想基盤 (Guest OS)	処理時間
T1	VirtualBox 6.1 (Windows2010Eval)	2 分
T2	VirtualBox 6.1 (Windows2010Eval)	1 分
T3	Parallels Desktop 16* (Windows10 on ARM Insider Preview)	3 分

マルチプルアライメント結果の画面 HC を以下に示す。

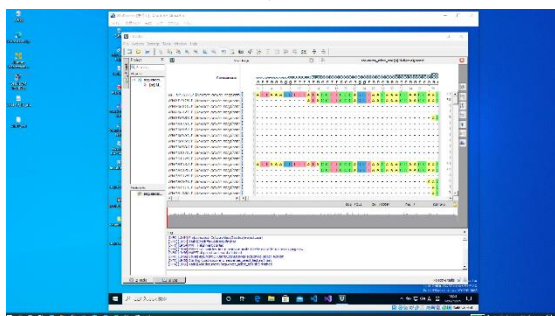


図 1 T1 マルチプルアライメント結果

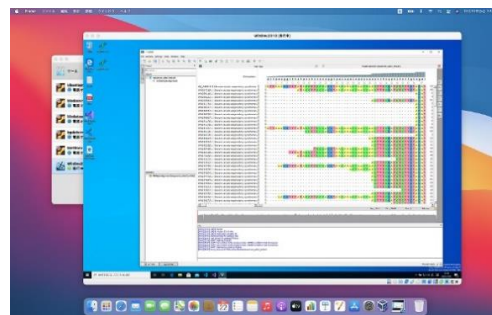


図 2 T2 マルチプルアライメント結果

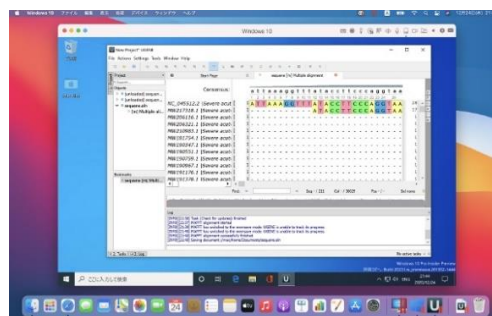


図 3 T3 マルチプルアライメント結果

解析処理時間、仮想基盤環境の構築の容易さ、コストを下表にまとめる。

表 5 仮想基盤の評価

Type	処理時間	構築容易性	コスト
T1	2	○	△
T2	1	○	△
T3	3	△	○

構築容易性△ 専用の仮想基盤と Guest OS の取得が必要

コスト△ > 約\$3,500

T3 はコストの評価が高く、将来性もあるが、仮想基盤のリリースが 1 ベンダーであり、今後、他の仮想基盤や Guest OS のリリースに期間を要することが想定される。

4. おわりに

マルチプルアライメントを行い、解析に要する処理時間、仮想基盤構築の容易性、コストを評価した。今後は、T2 の Parallels Desktop 16 for Mac による評価、Guest OS を Linux として評価を続けたい。

謝辞 公共 DB, マルチプルアライメントソフトウェア、仮想基盤, Guest OS をリリースされている皆様^{[4][5]}, および執筆にご協力いただいた皆様に謹んで感謝の意を表する。

参考文献

- [1] “Apple M1 チップを搭載した Mac への Parallels Desktop 対応状況について”. <https://www.parallels.com/jp/blogs/apple-m1-apple-m1-チップを搭載した-Mac-への-Parallels-Desktop-対応-状況-について/>, (参照 2021-01-06).
- [2] “Unipro UGENE”. <http://ugene.net/>, (参照 2021-01-06).
- [3] “ウイルス配列の入手とアライメント”. https://www.thermofisher.com/blog/learning-at-the-bench/virusalignment_ngs_dpcr_gsd_ts_1/, (参照 2021-01-06).
- [4] Tiago Antao(著), 阿久津達也(翻訳), 竹本和弘(翻訳). バイオインフォマティクス: python による実践レシピ. 2020.
- [5] 坊農秀雄(著), Dr.Bono の生命科学データ解析. 2017.