

武漢コロナウイルスのホモログマップ作成法の紹介

天野晃 *

* 無所属

*amano.au1@gmail.com

概要 2020 年初頭、新型コロナウイルスの感染が拡大し、各国で緊急事態宣言が発せられるまでに至った。当該ウイルス（だけではないが）の感染検査には、主に PCR 法が用いられるが、プライマー設計はその精度を左右する大きな要因の一つである。特に False-positive を忌避する場合は他のウイルス/生物のゲノム（断片）のコンタミネーションに対してロバストである必要があるが、ホモログマップによるゲノム特徴の可視化は、その判断の参考となる。本報告では、ホモログマップの作成法について紹介する。また、医療系、生物系が専門でない参加者の方々のために、テクニカルタームの説明を付録として用意する。ポスター閲覧の際の参考にされたい。

Wuhan corona virus homologue mapping

Kou AMANO*

*independent

1 はじめに

ウイルス等の感染検査の一つに、PCR 検査がある。この検査は、文字通り検出対象となる DNA(RNA)断片が存在するかを、PCR 増幅により直接的に検出・確認する方法である。PCR の際には、ターゲットとなるゲノム断片の一部と相同性を持つ、さらに短い DNA(RNA)断片をプライマーとすることによりターゲットを特異的に増幅させるが、当然、ターゲット以外にもプライマーと相同な配列を持つゲノム（断片）は存在する可能性があり、これらがコンタミネーションを起こしている場合は、False-positive を導く。そのような場合も配列解析を行うことにより、正確な検出が可能となるが、コストは大きくなる。

2 目的

PCR プライマーの設計において、ロバストネスの判断の参考となり得る、簡易かつ低コストなホ

モログマップの作成方法を紹介する。

具体的には、(1) 宿主側ゲノムに対するマップ、(2) ウイルスゲノムに対するマップ、(3) 自身のゲノムの特徴化、について述べる。

3 マッピング方法

3.1 宿主側

[DB 側ゲノム] 宿主のゲノムとして、turkey、rock pigeon、pig、rabbit、mouse、human、ferret、dog、cat、camel、beluga、bat を用いた。配列情報の取得は、NCBI のサイト [1] より 2020 年 3 月に行った。完全ゲノム情報を用いた種と、全ゲノムショットガンシーケンシングの結果を用いた種を含む。DB 作成は、makeblastdb コマンドのデフォルトオプションで行った。

[クエリー側ゲノム] 配列データは武漢コロナウイルス完全ゲノム、MN908947.3 を用いた。検索は

megablast を利用した。クエリー条件は、10 塩基以上のマッチを指定し、その他はデフォルトとした。

3.2 ウイルス側

[DB 側ゲノム] 配列情報の取得は、NCBI のサイト [1] より 2020 年 3 月に行った。NCBI のサイトには、ウイルスおよびファージのゲノム配列がまとめられたセクションがあり、これを一括ダウンロードした。

[クエリー側ゲノム] 3.1 に同じ。

3.3 自身のゲノム

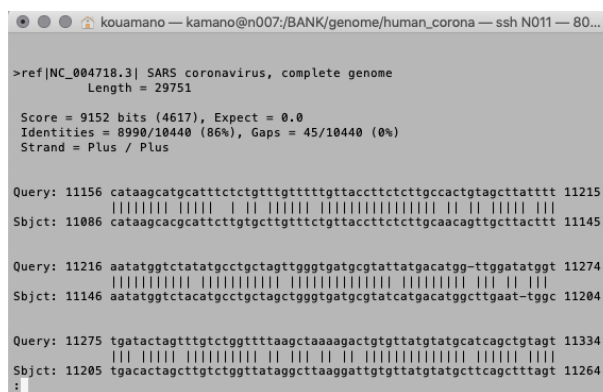
自身のゲノムにおいて、どの領域がどの程度のコピー数を持つか、を示すものである。

[DB 側ゲノム] 3.1 と同じ配列を 25 塩基ごとにオーバーラップなしで分割したデータを作成し、これをもとに前述と同じ方法で DB を作成した。

[クエリー側ゲノム] DB 作成時と同じ配列情報をクエリーとして、3.1 と同様に行った。

[Window-fourier] DNA 配列に対してフーリエ変換を行うことにより、コーディング領域ではコードンの特徴が強調され長さ 3 に相当する周期のピークが検出されることや、その他の特徴を強調できることが知られており [2, 3, 4]、ORF が実際のコーディング領域であるか等の傍証となり得る。次の条件にて、フーリエ変換を行った。

1. 塩基配列情報を次のように数値列に変換する：
A → 1、G → I、C → -I、T → -1
2. 1000 ベースごとに配列を分割する
3. 分割されたそれぞれの配列に対して直接フーリエ変換を行う
4. 分割されたそれぞれの配列に対して変換後の各値の絶対値を取得する



```
>ref|NC_004718.3| SARS coronavirus, complete genome
Length = 29751

Score = 9152 bits (4617), Expect = 0.0
Identities = 8990/10440 (86%), Gaps = 45/10440 (0%)
Strand = Plus / Plus

Query: 11156 cataagcatgcatttctctgtttgtttttgtacattcttgcactgtagcttattt 11215
Sbjct: 11086 cataagcagcatttctgtgcttgtttctgttaccttcttgcacagttgcttacttt 11145

Query: 11216 aatatgggtctatatgcctgtagttgggtgatgctattatgacatgg-ttggatatggt 11274
Sbjct: 11146 aatatgggtctacatgcctgtagttgggtgatgctatcatgacatggctgaat-tggc 11204

Query: 11275 tgatactagttgtctggttttaagctaaaagactgtgttatgtatgcacagctgtagt 11334
Sbjct: 11205 tgacactagctgtctggttataggcttaaggattgtgttatgtatgcacagcttagt 11264
```

図 1: megablast の結果の例。

4 マッピング結果の可視化

4.1 バーマップ

magablast の結果より、ヒット領域が武漢コロナウイルスゲノム上の塩基位置として判明するので (図 1)、この領域をバーで表す。脊椎動物種に対しては種ごとにマップを作成、ウイルス・ファージゲノムに対しては同一のマップ上にバーを表示した。

4.2 頻度マップ

バーマップと同様の原理により、武漢コロナウイルスの 1 塩基ごとにヒットの回数がカウント可能である。この頻度を、横軸を武漢コロナウイルスのゲノムの塩基位置、縦軸を頻度とする折れ線グラフで表す。用いた脊椎動物種全てのゲノム、およびウイルス・ファージゲノムに対して、それぞれ一つのマップで表示した。

4.3 フーリエマップ

1000 ベースごとに作成したフーリエ変換プロットを、武漢コロナウイルスの対象となるゲノム位

置にあわせて表示した。

4.4 相同領域マップ

自身のゲノムを 25 ベースごとに分断して構成した DB とクエリーのヒット結果より、領域（25 ベースの Window）ごとのヒット数を折れ線グラフで表す。これを武漢コロナウイルスの対象となるゲノム位置にあわせて表示した。プライマーに対応する配列が複数箇所が存在すれば、増幅されるコピー数が異なることになる。なお、blast 検索は、asymmetric であり、DB 側から見たヒット領域とクエリー側から見たヒット領域が異なる。

4.5 高頻度ホモログ配列

各生物種およびウイルス・ファージ全ゲノム対武漢コロナウイルス全ゲノムの megablast の結果より、ヒットが高頻度となる領域に対応する塩基配列パターンを抽出した。用いた脊椎動物ゲノム全て、およびウイルス・ファージゲノムに対してそれぞれ頻度を 5 回以上、21 回以上とした。さらに、それぞれ重なり合うホモログ群より、すべてに出現する塩基配列パターン部分を抽出した。

4.6 プライマー位置

マニュアル [5] に示されるプライマーの位置を武漢コロナウイルスのゲノム上にマップした。

附録：テクニカルターム解説

- PCR : Polymerase Chain Reaction の略。DNA ポリメラーゼ（合成酵素）を利用して DNA を複製する系。DNA 鋳型、プライマー、合成酵素、DNA の構成要素であるデオキシヌクレオチド（塩基が、アデニン、グアニン、シトシン、チミンの 4 種）等をバッファに投入し、温度サイクルを作成することにより DNA

複製が可能となる。産物を検出する際は電気泳動を行う。またはリアルタイム定量的（逆転写）PCR を用い、増幅と同時に検出するのが一般的である。[5]

- 電気泳動：DNA やタンパクなど、電荷を持つ分子の分離を行う系。蛍光マーキング等により視覚的に産物の確認を行う。
- リアルタイム定量的（逆転写）PCR : DNA の定量を目的とする PCR。蛍光マーカー等を用い、これを測定することにより産物の量を計測（推測）する。増幅中にリアルタイムに計量を行う。RNA 量を計測する際には逆転写を行うので、このように呼ばれる。
- プライマー：PCR の際、鋳型 DNA に結合し合成開始のプライマーとなる短い DNA 断片。
- 全ゲノムショットガンシーケンシング：配列決定を行う際、chromosome 全体を読み取することは困難なため、ゲノムをある程度の大きさに切断してシーケンシングを行い、後に計算機により可能性の高い配列を（接合）推測することが一般的である。完全に接合されていない状態での配列情報をこう呼ぶ。
- 相同性：特に遺伝子およびアミノ酸の相同性を指す。基本は文字列の相同性を基にしているが、置換を受けやすい／受けにくいペアが判明しており、マッチングには遺伝学の知識が反映されている。
- blast : "Basic Local Alignment Search Tools" の略。DNA (RNA) およびアミノ酸配列の相同部分を検索するシステム。複数のコマンドからなり、主にデータベース作成コマンド、データベース検索コマンドに分かれる。
- ホモログ：検索による相同部分、あるいは相同な遺伝子をこう呼ぶ。
- コーディング領域：DNA においてタンパク質に翻訳される領域。あるいは、そうであると予想される領域。

- ORF : "Open Reading Frame" の略。タンパク質への翻訳は、開始コドンであるメチオニンから始まり、終止コドンである3種のトリプレットで終了することが知られている。この領域をこう呼ぶ。
- コドン : DNA がタンパク質 (アミノ酸配列) に翻訳される際、3塩基が1組でひとつのアミノ酸をコードする。この3塩基のコード (トリプレットコード) をコドンと呼ぶ。

注・文献

- [1] <<https://ftp.ncbi.nlm.nih.gov/>>, (参照:2020-03)
- [2] Sergey V. Petoukhov. The genetic code, 8-dimensional hypercomplex numbers and dyadic shifts. <http://symmetry.hu/isabm/petoukhov.html>.
- [3] GUY DODIN, PIERRE VANDERGHEYNST, PATRICK LEVOIR, CHRISTINE CORDIER, LAURENCE MARCOURT. Fourier and Wavelet Transform Analysis, a Tool for Visualizing Regular Patterns in DNA Sequences. J. theor. Biol. No. 206, 2000, p.323-326.
- [4] V. R. Chechetkina, V.V. Lobzinc aEngelhardt. Large-scale chromosome folding versus genomic DNA sequences: A discrete double Fourier transform technique. DOI:10.1016/j.jtbi.2017.05.0.
- [5] 病原体検出マニュアル 2019-nCoV Ver.2.6. <<https://www.niid.go.jp/niid/images/lab-manual/2019-nCoV20200217.pdf>>. (参照:2020-03)