**Journal Club (2024 年 10 月 7 日) まとめ**

担当: 博田　悠斗

**発表論文**:

Sanabria, M., Hirsch, J., Joubert, P.M. et al. DNA language model GROVER learns sequence context in the human genome. Nat Mach Intell 6, 911–923 (2024). https://doi.org/10.1038/s42256-024-00872-0

**研究目的および概要**:

近年、DNA 配列データを系列データとみなし、自然言語モデルを適用して解析する研究が進められている。しかし、DNA 配列データは通常の言語とは異なる特徴を持つ。第一に、DNA 配列には方向性が明確に定義されていない。第二に、DNA 配列には単語が定義されておらず、トークン化が困難である。このため、DNA 配列データに自然言語処理モデルを適用する際には、トークン化の過程が特に重要となる。

現在、DNA 自然言語モデルにおいては k-mer というトークン化手法がある。この手法は、塩基配列を一定の長さ（例えば 4 つや 5 つ、6 つ）に区切るものである。しかし、この方法ではトークンの頻度に偏りが生じやすく、モデルが全体の文脈ではなく頻度に基づいて予測を行うため、性能が低下するという問題がある。そこで、この問題を解決するために、BPE（バイトペアエンコーディング）という技術が導入され、トークンの頻度を均衡させる手法が採用されている (図 1)。

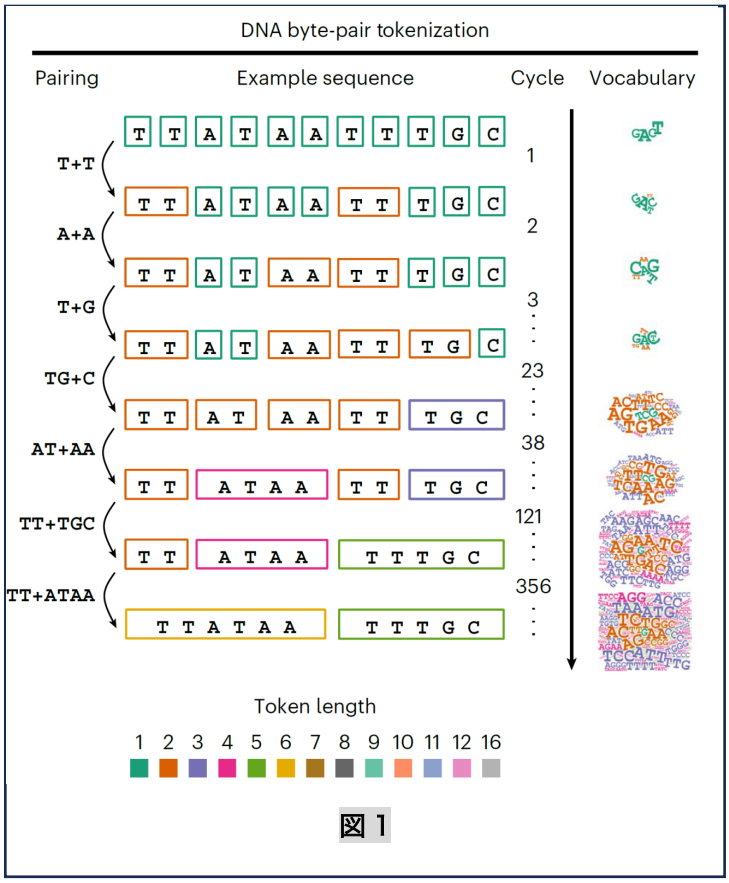


図 1

この技術は元々テキスト圧縮アルゴリズムとして開発されたが、現在では GPT-3 などの自然言語処理における一般的なトークン化戦略として知られている。具体的には、隣接する塩基配列（例えば A と T）を連結して新しいトークン（AT）を形成し、このペアリングの過程を複数回繰り返すことで、徐々に大きなトークン

を生成する。BPE のサイクルを経るごとにトークンの長さは長くなり、その頻度は減少する。これにより、従来のモデルでは困難であった頻度の調整が可能となり、バランスの取れたボキャブラリーを生成できるようになる。
しかし、現状の研究では DNA 配列を入力として最適な BPE サイクルを決定した例は存在せず、どのような設定で最適な学習が可能であるかを十分に把握できていない。そこで、本研究では最適な BPE サイクルを決定することを目的として研究を行った。

**先行研究と比べて何がすごい？ 技術やアプローチのキモはどこ？**:

- 従来は、ヒトゲノムデータに対して BPE（バイトペアエンコーディング）を導入した際、そのトークン化の最適なサイクル数を明らかにいていなかった。本研究では、具体的に BPE の学習サイクルを 100 回から 5000 回までの様々な試行回数で実験を行い、その最適なサイクルを決定した。

**どうやってこの手法/仮説の有効性を検証したのか**:

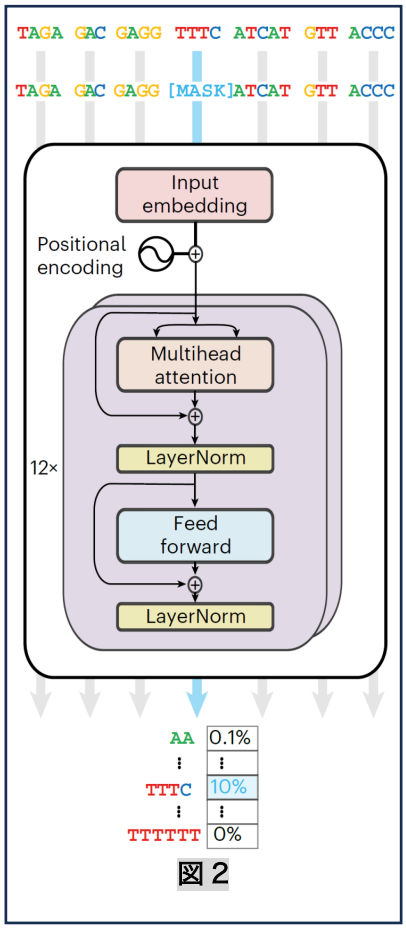

図 2

- 本実験では、BERT のエンコーダモデルを用い、トークン化を行う過程（BPE）を導入した上で、最適なサイクル数を検討するために、まず事前学習で入力されたトークンを特殊なトークンに置き換え、その予測を行うタスクを実行した (図 2)。ファインチューニング時には、トークンの次の k-mer が何であるかを予測するタスクを実行した。学習の際には、BPE の学習サイクル数を変更し、その最適なサイクル数を決定する実験を行うことにより、その最適なサイクル数が 600 前後であることを確認した。
- モデルにおいてマスクされたトークンを正確に予測することが困難な単語の割合を示すパープレキシティを評価指標として設定し、既存の手法との比較を行う実験を実施し、本手法の性能が既存の手法を上回ることを確認した。さらに、トークンの平均長さが 4.07 であることを確認した。
- 本手法の埋め込み手法である GROVER の埋め込みと、周囲の単語から中央の単語を予測する連続的な Bag-of-Words のアプローチを採用する Word2Vec の埋め込み手法を比較し、本モデルが配列の文脈を考慮した学習を行えることを確認した。また、BERT のアーキテクチャ全体を用いて、同じトークン配列から得られる埋め込みのコサイン類似度を導出し、その評価を通じて既存の手法の有効性を上回ることを確認した。
- プロモーターの識別の有無や、CTCF に結合するタンパク質の結合部位を予測する下

流タスクを実行することにより、モデルの性能を確認した。

**その他、議論した内容 (ネガティブコメントや limitation もあれば)**:

- BPE の最適なサイクル数を決定したが、その理由や背後にある生物学、理論的な根拠については十分に解明されておらず、解明する必要があると考えられる。

**この研究をさらに発展させるとしたら**:

- 本モデルのファインチューニングを、遺伝子発現や疾患関連領域の予測などのタスクで行い、個別化医療や新たな治療法の開発に貢献できるようなモデルの構築を目指すことが考えられる。