

2025 年度年次報告書

ライフサイエンスデータベース統合推進事業
統合化推進プログラム

【研究課題名】「大規模言語モデルを活用した病的スプライシング変異データベースの自律的構築」

研究代表者：

白石 友一（国立がん研究センター 研究所 ゲノム解析基盤開発分野 分野長）

研究分担者：

なし



©2025 白石 友一(国立がん研究センター) licensed under CC表示4.0国際

1. 研究進捗状況と成果

本年度は、スプライスサイト生成変異 (Splice-Site Creating Variants, SSCV) データベース (SSCV DB) の機能拡充と、大規模言語モデルを用いた病的変異予測ワークフローのプロトタイプ構築を進めた。

まず、SSCV DB については、データベース構成のリファクタリングを行うとともに、Alu 配列に関連する SSCV の解析結果を新たに統合した。Alu 配列はイントロン内に多数存在し、特にアンチセンス方向の Alu は潜在的なスプライスサイトモチーフを含むため、変異による cryptic exon inclusion の原因となり得る。本年度は、全データセットに対して Alu 解析パイプラインを実行し、各 SSCV 周辺 100 bp と RepeatMasker による SINE/Alu アノテーションとの重なりを判定した。さらに、Alu コンセンサス配列に対するローカルアラインメントにより、SSCV およびスプライスサイトの Alu 内相対位置、Alu の方向性、サブファミリー、新規エクソン化との関係を算出し、変異テーブルに統合した。これにより、利用者は SSCV と Alu 配列との位置関係をデータベース上で直接参照できるようになった。また、外部リファレンスデータベースとの連携も更新・拡充した。ClinGen、NHGRI Clinical Genomic Database、NCBI gene_info、GENCODE の情報を最新版に更新し、さらに OMIM mim2gene を新規に追加した。これにより、遺伝子の疾患関連性、dosage sensitivity、遺伝子名別名、ゲノム座標、OMIM 疾患 ID との対応関係を含むアノテーションを強化した。

次に、大規模言語モデルを用いた病的変異予測システム「VUS Guesser」のプロトタイプを構築した。本システムは、問い合わせ変異に対して、近傍領域に存在する既知の病的類似変異を探索し、その情報に基づいて機能的影響を推定するものである。Claude Agent SDK を基盤とし、類似変異探索を担う Finder Agent と、収集された証拠を統合・解釈する Summarizer Agent からなる二段構成のワークフローを実装した。複数変異のバッチ処理、構造化 JSON による入出力、変異ごとの根拠ファイルおよび遺伝子単位のサマリ出力も整備した。LLM による誤った類似変異選択を抑制するため、「真に類似する変異」の判定基準を明確化した。具体的には、同一遺伝子上の変異であること、スプライシング異常タイプおよびタンパク質帰結カテゴリが対応すること、問い合わせ変異周辺の定義済み探索範囲内に存在すること、さらにヒト患者での報告または機能的証拠を伴うことを必須条件とした。これにより、単にデータベースに登録されているだけで病原性根拠が不十分な変異の混入を抑える設計とした。さらに、LLM エンジンおよびプロンプトの比較評価を行い、コストと出力品質のバランスを検討した。その結果、メインエンジンに Opus 4.5、サブエンジンに Sonnet 4.6 を用いる構成が最も実用的であると判断した。最終構成では、平均約 1.7 ドル/変異、7.5 分/変異での解析を達成し、遺伝子単位でまとめて問い合わせることで、1 変異あたり 40-47% のコスト削減が可能であることも確認した。加えて、NOTCH1 変異を含む既知事例をもとに陽性対照を拡充し、最終的に陽性対照 25 変異、陰性対照 5 変異からなるゴールデンセット 30 変異を整備した。初期評価として 100 変異規模のテスト実行を行い、類似変異が検出され機能予測の根拠が得られたものが 46 変異、検出されなかったものが 54 変異であった。一部の既知陽性例では期待通り検出できた一方、見落とし事例も確認され、今後はゴールデンセットを用いた定量評価とワークフローの安定化を進める必要があることが明らかとなった。

2. 代表的な成果事例

該当なし

以上