（別紙様式４）

山梨県若手研究者奨励事業　研究成果概要書

所属機関名　国立大学法人山梨大学

職名・氏名　特任准教授 クップル ドミニク ㊞

１　研究テーマ

一次元データの効率的な処理手法の開発

２　研究の目的

遺伝子データ，自然言語テキスト，プログラムコードなどの一次元データ処理において，既存技術では文字データ量に制限があり，実際の問題に対して応用は難しい．この課題に対して，文字数を問わない新たな手法を開発し，より迅速かつ効果的な方法を提案する．

３　研究の方法

$$\widetilde{T} = \mathtt{A}\left\{\begin{matrix}\mathtt{A}\\ \mathtt{C}\end{matrix}\right\}\mathtt{C}\left\{\begin{matrix}\mathtt{G}\\ \mathtt{T}\end{matrix}\right\}, \quad \mathcal{L}(\widetilde{T}) = \{\mathtt{AACG}, \mathtt{AACT}, \mathtt{ACCG}, \mathtt{ACCT}\}.$$

現代のバイオインフォマティクスにおいて、遺伝子配列は遺伝性疾患や遺伝的特性を明らかにするための重要な情報源とみなされている。そのため遺伝子配列が有する欠如単語（Absent Words）や一意単語（Unique Words)といった固有の特徴に関心が向けられている。それに伴い個々の遺伝子データが膨大に膨らむ一方で、それらを保存・検索する方法には制限があり困難が生じている。 遺伝子配列を単純な文字列で表現する場合、余分な領域が生じることでデータ量が膨らむ。それに対して類似した遺伝子配列を簡潔に表現することでデータ量の減縮を図る手法が求められている。 その表現の一つが未決定文字列である。 未決定文字列は、文字（すなわち DNA の塩基対）に対して複数の選択肢を与えることで柔軟な表現を可能にする。例えば、遺伝子配列決定時の誤りや、遺伝的変異を考慮する場合に未決定文字列で表現できる。 しかし，未決定文字列上で欠如単語や一意単語をどのように計算するかは，これまで明確にされていなかった．例は図１で示される．

1: $\widetilde{T}$ $\widetilde{T}$ $\mathcal{L}(\widetilde{T})$

1 c

$\{c\}$ c

AC, T, CC AT, AAA, CCC

本研究では，未決定文字列における欠如単語・一意単語の検索問題を解析し，その計算困難性を明らかにするために，以下の手法を採用した． まず，未決定文字列の定式化を行い，単純な文字列における欠如単語・一意単語の概念を拡張した．この問題の計算量を評

留意事項

①３枚程度で作成してください。

②特許の出願中等の理由により、一定期間公表を見合わせる必要がある箇所がある場合であっても、所定の期日までに公表可能な範囲で作成・提出してください。当該箇所については、後日公表可能となった際に追記して再提出してください。

価するために，NP 困難性の証明を構築した． 具体的には，既知の NP 困難問題から取り組みたい問題への帰着が成り立つ場合，取り組みたい問題を多項式時間で解くことが困難であることを示すことができる． その一方で，実際の応用可能性を検討するために，未決定文字列に対する探索アルゴリズムを設計し，特定の制約条件のもとでの計算効率を評価した．未決定文字列から欠如語や一意語を抽出するために必要な制約を定式化する手段として，解集合プログラミング（ASP: Answer Set Programming）を用いた．

４　研究の成果

$$\widetilde{T} = \mathrm{A}\left\{\begin{array}{c}\mathrm{AG}\\ \epsilon\end{array}\right\}\mathrm{C}\left\{\begin{array}{c}\mathrm{G}\\ \mathrm{T}\\ \mathrm{CCT}\end{array}\right\}$$

本研究の主な成果として，未決定文字列における欠如単語と最短一意単語，それぞれの検索問題が「NP 困難」であることを示した点が挙げられる． これにより，単純な文字列では線形時間で解決可能であった問題が，未決定文字列のもとでは大幅に計算量が増加することが明らかになった． また，未決定文字列に対する類似性測定のアルゴリズム設計にも影響を与える可能性がある． つまり、これらの結果は未決定文字列だけでなく、一般化された遺伝子配列表現に対しても計算量の増加を示している． 特に，未決定文字列の一般化である弾性退化文字列（elastic degenerate strings）は，バイオインフォマティクスにおいて盛んに研究されている．この弾性退化文字列は、各文字位置において代替文字を選択できるだけでなく，任意の長さの文字列を選択できる点で，より柔軟な表現を可能にしている． 任意の長さの文字列の挿入や欠失を引き起こす遺伝的変異を表現することができ，単一塩基多型（SNP: Single-nucleotide polymorphism）のみを扱う未決定文字列よりも，遺伝的変異をより適切にモデル化できる．図２は例を示す．

$$\mathcal{L}(\widetilde{T}) = \{\ \mathrm{AAGCG}, \mathrm{AAGCT}, \mathrm{AAGCCCT}, \mathrm{ACG}, \mathrm{ACT}, \mathrm{ACCCT}\ \}$$

$$2: \qquad \widetilde{T} \qquad \widetilde{T} \qquad \mathcal{L}(\widetilde{T})$$

$$\epsilon$$

他方で、本研究では特定の条件下での計算手法についても検討し，未決定文字列に適用可能なアルゴリズムの方向性を示した． 特に，ASP を用いた効率的な実装を開発し，未決定文字列上での欠如単語および一意単語の計算が現実的な時間内で実行可能であることを示した． この ASP を用いた手法により，未決定文字列を用いた情報検索や遺伝子データ解析の分野における実用化が期待される．

５　今後の展望

本研究の成果を踏まえ，今後は以下の課題に取り組む予定である． まず，現在の ASP による符号化は素朴なものであり，さらなる最適化が必要である． 特に，生物的な特徴や組合せ論的な性質を活用することで，計算効率を高められる可能性があるかどうかを検討していく．

次に，未決定文字列の可能性については未解明の点が多く，今後さらなる研究が必要である． 従来のパターンマッチングに関する結果は存在するが，より一般化されたマッチン

グ，すなわち文字の完全一致以外も許容するような緩やかな条件下でのマッチングに関する研究は十分とは言えない．従って，そのような一般化されたマッチングを効率的に行う手法の確立に取り組んでいる． さらに未決定文字列同士を比較するための手法として，パラメータ化マッチングなどの一般化されたパターンマッチングの応用を検討している．この種のマッチングは，DNA 配列中に現れる相補的塩基対の対応関係を一般化したものであり，バイオインフォマティクスにおける応用が期待される．
また，本研究から派生した興味深い問題として，「2 つの弾性退化文字列の均質性を多項式時間で判定できるか」という課題が挙げられる． 未決定文字列については多項式時間での判定が可能であるが，より一般的な弾性退化文字列に対しては，それが不可能だと証明でき，理論的にも興味深い課題である．
本研究で開発される未決定文字列を活用した効率的な遺伝子データ処理技術は、個人の遺伝的特徴に基づくゲノム医療の発展に貢献できる。 山梨県が取り組む地域包括ケアや高齢者医療において、個別化された治療方針の立案や予防医療の推進に役立つ可能性がある。さらには、医療データの管理においても応用が期待される。

６　研究成果の発信方法（予定を含む）

本研究課題への関心を喚起し，また研究内容に対する反応を得ることを目的として，以下の 2 つの国内学会にて部分的な研究成果を発表した． これにより，国際的な背景を持つ複数の研究者から注目を集め，今後共同研究を進めていく可能性が生まれている． 現在，研究は一定の進展を見せており，計算生物学および文字列処理を対象とする国際会議（たとえば SPIRE 2025 など）への論文投稿を計画している．

Dominik Köppl, Jannik Olbrich: 未決定文字列における一意単語の検索の困難さ. 情報処理学会, 研究報告アルゴリズム 巻 2024-AL-201, 号 4, ページ 1 - 5. (2025)

Dominik Köppl, Jannik Olbrich: 未決定文字列における欠如単語の検索の困難さ. 情報処理学会, 研究報告アルゴリズム 巻 2024-AL-200, 号 15, ページ 1 - 5. (2024)

留意事項
①３枚程度で作成してください。
②特許の出願中等の理由により、一定期間公表を見合わせる必要がある箇所がある場合であっても、所定の期日までに公表可能な範囲で作成・提出してください。当該箇所については、後日公表可能となった際に追記して再提出してください。