令和 6 年度 山梨県若手研究者奨励事業 研究成果報告書

所属機関名　　国立大学法人 山梨大学 大学院　総合研究部
職名・氏名　　特任准教授・クップル ドミニク

概要

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．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．未決定文字列の中に，各位置に 1 つの文字が格納されているだけでなく，複数の文字の選択肢が認められている．さらにテキスト位置ですべての可能性を列挙するモデル化に適用できる．実用化のため，パターン照合や未決定文字列の類似性を測定するなど，単純な文字列に対して既知の技術を未決定文字列に応用し，様々なクエリの種類に応じることに関心が向けられているが，遺伝子データの解析において頻繁に利用されている一意単語・欠如単語の検索に関しては未検証である．本研究では単純な文字列において，最短の一意単語・欠如単語の検索は既に線形時間で出力可能であるが，未決定文字列上での計算は NP 困難であることを明らかにした．

キーワード：NP 困難・未決定文字列・一意単語・欠如単語・解集合プログラミング

1　はじめに

近年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．遺伝子配列が有する欠如単語や一意単語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．未決定文字列では，各文字位置において 1 文字ではなく複数の代替文字（すなわち DNA の塩基）を許容することで，柔軟な表現が可能となる．たとえば，配列決定時の読み取りエラーや，個体差に由来する遺伝的変異（英語：SNP）を考慮する場面で，未決定文字列を用いることで現実的なデータの表現が可能となる．言い換えれば，任意の位置に複数の候補文字を持つことができる文字列を定式化・解析することが本研究の出発点である．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への貢献に期待される．

未決定文字列の各要素は入力アルファベットの部分集合となり，これを**記号**と呼ぶ．記号は，DNA や RNA 配列における未決定の核酸を表現するために IUPAC 表記法 [34] を一般化したものである．図 1 は，未決定文字列の例を示している．

2　先行研究

本論のテーマに関連する研究は，未決定文字列，極小一意単語，極小欠如単語に特化した研究である．以下では先行研究を紹介する．

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

図 1: 未決定文字列 $\widetilde{T}$ の例．$\widetilde{T}$ の言語 $\mathcal{L}(\widetilde{T})$ は右に示されている．この言語は，未決定文字列によって表される文字列の集合である．1 つの文字 c のみを格納する記号は，$\{c\}$ または単に c として書かれる．極小一意単語は AC, T, CC など．極小欠如単語は AT, AAA, CCC など．

2.1 未決定文字列

未決定文字列に関する研究の大部分は，パターン照合，構造的特性や規則性，必ずしも自己索引ではない索引からの未決定文字列の再構築，および 2 つの未決定文字列の比較に関心が向けられている．

パターン照合　　未決定文字列に対しては，Boyer–Moore の適応版 [30]，ShiftAnd と Boyer–Moore–Sunday の組み合わせ [49]，および KMP を元した方法 [43] が提案されている．

構造的特性　　未決定文字列の構造的特性については，カバーおよび/またはシードの計算 [9, 12]，Lyndon 分解の拡張 [22] が知られている．[37] は未決定文字列の新しい表現モデルを提案した．

再構築　　データ構造からの未決定文字列の再構築も活発な研究分野である．再構築は，ボーダー配列，接尾辞配列，および LCP 配列から構築が可能である [42]．接頭辞配列に基づくグラフから [7]，または頂点がテキスト位置で辺が一致する文字を持つテキスト位置であるグラフからも再構築が行われている [29]．[18] は，配列が単純な文字列または未決定文字列の接頭辞配列であるかどうかの特徴を示した．最後に，[15] は，未決定文字列の接頭辞配列と無向グラフおよびボーダー配列との関係を研究した．

2.2 極小一意単語

極小一意単語（MUS）の計算問題は Pei らにより導入された [46]．この問題は，シーケンスアラインメント [4]，ゲノム比較 [27]，および系統樹の構築 [16] などに応用されてきた．より一般的な極短一意部分文字列（SUS）は，局所的な極小性のみを求めるもので，テキスト位置または区間をカバーすることで表現できる．SUS の計算に対する研究注目は，時間量および領域量に関する改善につながった [32, 50, 33, 38, 40]．両方の量を均衡する解決策も提供された [26, 11]．また，スライディングウィンドウ内での計算 [41] や連超圧縮された文字列上での計算 [38] など，異なる設定が提案されている．理論的な保証を得るために，[39] は，与えられたクエリ位置をカバーする SUS の最大数を研究した．近年，SUS 計算に関するサーベイ記事が公開されている [3]．最後に，与えられた範囲内で SUS を計算するための変形が存在する [1, 2] または k ミスマッチを持つ SUS を計算する [31, 48, 8]．

2.3 極小欠如単語

極小欠如単語（MAW: minimal absent word）は，潜在的な予防および治療的医療応用のためのバイオマーカーとして [47] によって導入された．MAW は，系統学 [17]，シーケンス比較 [20]，音楽コンテンツの情報検索 [19]，および円形バイナリ文字列の再構築 [45] において価値があることが示されている．MAW を計算するために，接尾辞配列 [13]，有向非巡回単語グラフ（DAWG: directed acyclic

word graph） [25, 24]，または最大繰り返し [10] を使用するアルゴリズムが提案されている．並列処理 [14] や外部メモリ [28] で動作するアルゴリズムも提案されている．拡張として，連長圧縮文字列の MAW の計算 [6]，複数の文字列に共通する MAW の計算 [44]，または木上の MAW の計算 [23] がある．もう一つの拡張は，スライディングウィンドウ内での計算 [21, 41] であり，スライディングの動きに基づく解答集合の変化の数に関する境界が分析されている [5].

3 背景知識

まず，文字列の基本概念を紹介し，その後，未決定文字列への一般化について表現する．

文字列　　アルファベットを Σ とする．Σ^* の要素は（単純な）**文字列**と呼ばれる．文字列 T が与えられたとき，T の i 番目の文字は $T[i]$ と表される（整数 $i \in [1..|T|]$ の場合），ここで $|T|$ は T の長さを表す．整数 i と j が $1 \leq i \leq j \leq |T|$ を満たすとき，位置 i から始まり位置 j で終わる T の部分文字列は $T[i..j]$ と表される．すなわち，$T[i..j] = T[i]T[i+1]\cdots T[j]$ である．T の部分文字列 P は，$P \neq T$ の場合に**真の部分文字列**と呼ばれる．

未決定文字列　　単純な文字列の以下の拡張を研究する．そのために，アルファベット Σ の**文字**と，次の 3 種類の一般化のいずれかに属する文字列の**記号**を区別する．**未決定文字列**は，記号が Σ の空でない部分集合から引かれる文字列 $\widetilde{S}[1..n]$ である．すなわち，$\emptyset \neq \widetilde{S}[i] \subset \Sigma$ である．$r = r(\widetilde{S}) = \max_i |\widetilde{S}[i]|$ は最大の記号のサイズを示す．

4 研究結果

本研究では，未決定文字列における一意単語・欠如単語の計算量を解析する．そのため，単純な文字列に基づく一意単語・欠如単語の形式的な定義を行い，未決定文字列へ拡張する．

4.1 一意単語

単純な文字列 T の中にある部分文字列 U は**一意単語**であり，T 中の出現が一つしかない．つまり，T 中に U と一致する部分文字列が他には存在しない．特に，すべての U の部分文字列は，T の一意単語ではない場合，U を**極小一意部分文字列** (*MUS*) と呼ぶ．未決定文字列 $\widetilde{S}$ において，文字列 $P \in \Sigma^*$ が，$\widetilde{S}$ のテキスト位置 i から出現するとは，$\widetilde{S}$ の言語に属する文字列 X が存在し，$X[i..i+|P|-1] = P$ であることを意味する．P が $\widetilde{S}$ で一意であるとは，その出現の位置が一意に決定されるということである．さらに，P が**極小一意部分文字列** *(MUS)* であるとは，P の任意の真部分文字列 X が $\widetilde{S}$ において少なくとも 2 回出現するということである． この定義を踏まえて，以下の問題に取り組む.

問題 1 (一意単語判断問題)．未決定文字列 $\widetilde{S}$ と整数 k が与えられたとき，k-一意単語判断問題は，長さが最大 k の一意単語が $\widetilde{S}$ に存在するかどうかを決定することである．

定理 1 ([36])．一意単語判断問題は $\sigma \geq 3$ および $r \geq 2$ の場合に NP 困難である．

4.2 欠如単語

単純な文字列 T の**欠如単語** S とは，T の部分文字列ではない文字列を指す，ただし S の長さは 1 以上である．T の欠如単語 X は，X のすべての真の部分文字列が T に出現する場合，**極小欠如単語** (MAW) と呼ばれる．より一般的には，未決定文字列 $\widetilde{S}$ において，出現しない文字列を**欠如**していると言う．欠如文字列 X は，そのすべての真の部分文字列が $\widetilde{S}$ に出現する場合，$\widetilde{S}$ において**極小**である．次に，特定の長さ以下の欠如単語を見つける決定問題が NP 困難である検討する．

問題 2 (欠如単語判断問題)**.** 未決定文字列 $\widetilde{S}$ と整数 k が与えられたとき，k-欠如単語判断問題は，長さが最大 k の欠如単語が $\widetilde{S}$ に存在するかどうかを決定することである．

定理 2 ([35])**.** 欠如単語判断問題は $\sigma \geq 3$ および $r \geq 3$ の場合に NP 困難である．

結果として，両問題は NP 困難であるため，高速化されたアルゴリズムの存在の確率は低いことが明らかとなった．この結果を踏まえて，未決定文字列において指定された長さの一意単語または欠如単語を計算するための SAT 定式化を作成した．SAT (充足可能性問題, satisfiability problem) とは連言標準形の式で表現された問題である．ただし，その式は一連の節を連言で結合されている．SAT 定式化を解集合プログラミング（ASP: answer set programming）で符号化し，両問題に新たな手法で取り組むことができる．ASP 符号化を以下のリンクで提供している．`https://github.com/koeppl/edstringcharacteristics`

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．

参考文献

[1] Paniz Abedin, Arnab Ganguly, Solon P. Pissis, and Sharma V. Thankachan. Range shortest unique substring queries. In *Proc. SPIRE*, volume 11811 of *LNCS*, pages 258–266, 2019.

[2] Paniz Abedin, Arnab Ganguly, Solon P. Pissis, and Sharma V. Thankachan. Efficient data structures for range shortest unique substring queries. *Algorithms*, 13(11):276, 2020.

[3] Paniz Abedin, M. Oguzhan Külekci, and Sharma V. Thankachan. A survey on shortest unique substring queries. *Algorithms*, 13(9):224, 2020.

[4] Boran Adas, Ersin Bayraktar, Simone Faro, Ibraheem Elsayed Moustafa, and M. Oguzhan Külekci. Nucleotide sequence alignment and compression via shortest unique substring. In *Proc. IWBBIO*, volume 9044 of *LNCS*, pages 363–374, 2015.

[5] Tooru Akagi, Yuki Kuhara, Takuya Mieno, Yuto Nakashima, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda. Combinatorics of minimal absent words for a sliding window. *Theor. Comput. Sci.*, 927:109–119, 2022.

[6] Tooru Akagi, Kouta Okabe, Takuya Mieno, Yuto Nakashima, and Shunsuke Inenaga. Minimal absent words on run-length encoded strings. In *Proc. CPM*, volume 223 of *LIPIcs*, pages 27:1–27:17, 2022.

[7] Ali Alatabbi, M. Sohel Rahman, and William F. Smyth. Inferring an indeterminate string from a prefix graph. *J. Discrete Algorithms*, 32:6–13, 2015.

[8] Daniel R. Allen, Sharma V. Thankachan, and Bojian Xu. An ultra-fast and parallelizable algorithm for finding k-mismatch shortest unique substrings. *IEEE ACM Trans. Comput. Biol. Bioinform.*, 18(1):138–148, 2021.

[9] Pavlos Antoniou, Maxime Crochemore, Costas S. Iliopoulos, Inuka Jayasekera, and Gad M. Landau. Conservative string covering of indeterminate strings. In *Proc. PSC*, pages 108–115, 2008.

[10] Aqil M. Azmi. On identifying minimal absent and unique words: An efficient scheme. *Cogn. Comput.*, 8(4):603–613, 2016.

[11] Hideo Bannai, Travis Gagie, Gary Hoppenworth, Simon J. Puglisi, and Luís M. S. Russo. More time-space tradeoffs for finding a shortest unique substring. *Algorithms*, 13(9):234, 2020.

[12] Md. Faizul Bari, Mohammad Sohel Rahman, and Rifat Shahriyar. Finding all covers of an indeterminate string in $o(n)$ time on average. In *Proc. PSC*, pages 263–271, 2009.

[13] Carl Barton, Alice Héliou, Laurent Mouchard, and Solon P. Pissis. Linear-time computation of minimal absent words using suffix array. *BMC Bioinform.*, 15:388, 2014.

[14] Carl Barton, Alice Héliou, Laurent Mouchard, and Solon P. Pissis. Parallelising the computation of minimal absent words. In *Proc. PPAM*, volume 9574 of *LNCS*, pages 243–253, 2015.

[15] Francine Blanchet-Sadri, Michelle Bodnar, and Benjamin De Winkle. New bounds and extended relations between prefix arrays, border arrays, undirected graphs, and indeterminate strings. *Theory Comput. Syst.*, 60(3):473–497, 2017.

[16] Supaporn Chairungsee. A new approach for phylogenetic tree construction based on minimal absent

words. In *Proc. DEXA*, pages 15–19, 2014.

[17] Supaporn Chairungsee and Maxime Crochemore. Using minimal absent words to build phylogeny. *Theor. Comput. Sci.*, 450:109–116, 2012.

[18] Manolis Christodoulakis, Patrick J. Ryan, William F. Smyth, and Shu Wang. Indeterminate strings, prefix arrays & undirected graphs. *Theor. Comput. Sci.*, 600:34–48, 2015.

[19] Tim Crawford, Golnaz Badkobeh, and David Lewis. Searching page-images of early music scanned with OMR: A scalable solution using minimal absent words. In *Proc. ISMIR*, pages 233–239, 2018.

[20] Maxime Crochemore, Gabriele Fici, Robert Mercas, and Solon P. Pissis. Linear-time sequence comparison using minimal absent words & applications. In *Proc. LATIN*, volume 9644 of *LNCS*, pages 334–346, 2016.

[21] Maxime Crochemore, Alice Héliou, Gregory Kucherov, Laurent Mouchard, Solon P. Pissis, and Yann Ramusat. Minimal absent words in a sliding window and applications to on-line pattern matching. In *Proc. FCT*, volume 10472 of *LNCS*, pages 164–176, 2017.

[22] Jacqueline W. Daykin and Bruce W. Watson. Indeterminate string factorizations and degenerate text transformations. *Math. Comput. Sci.*, 11(2):209–218, 2017.

[23] Gabriele Fici and Pawel Gawrychowski. Minimal absent words in rooted and unrooted trees. In *Proc. SPIRE*, volume 11811 of *LNCS*, pages 152–161, 2019.

[24] Yuta Fujishige, Takuya Takagi, and Diptarama Hendrian. Truncated DAWGs and their application to minimal absent word problem. In *Proc. SPIRE*, volume 11147 of *LNCS*, pages 139–152, 2018.

[25] Yuta Fujishige, Yuki Tsujimaru, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda. Computing DAWGs and minimal absent words in linear time for integer alphabets. In *Proc. MFCS*, volume 58 of *LIPIcs*, pages 38:1–38:14, 2016.

[26] Arnab Ganguly, Wing-Kai Hon, Rahul Shah, and Sharma V. Thankachan. Space-time tradeoffs for finding shortest unique substrings and maximal unique matches. *Theor. Comput. Sci.*, 700:75–88, 2017.

[27] Bernhard Haubold, Nora Pierstorff, Friedrich Möller, and Thomas Wiehe. Genome comparison without alignment using shortest unique substrings. *BMC Bioinform.*, 6:123, 2005.

[28] Alice Héliou, Solon P. Pissis, and Simon J. Puglisi. emMAW: computing minimal absent words in external memory. *Bioinform.*, 33(17):2746–2749, 2017.

[29] Joel Helling, Patrick J. Ryan, W. F. Smyth, and Michael Soltys. Constructing an indeterminate string from its associated graph. *Theor. Comput. Sci.*, 710:88–96, 2018.

[30] Jan Holub, William F. Smyth, and Shu Wang. Fast pattern-matching on indeterminate

strings. *J. Discrete Algorithms*, 6(1):37–50, 2008.

[31] Wing-Kai Hon, Sharma V. Thankachan, and Bojian Xu. In-place algorithms for exact and approximate shortest unique substring problems. *Theor. Comput. Sci.*, 690:12–25, 2017.

[32] Xiaocheng Hu, Jian Pei, and Yufei Tao. Shortest unique queries on strings. In *Proc. SPIRE*, volume 8799 of *LNCS*, pages 161–172, 2014.

[33] Atalay Mert Ileri, M. Oguzhan Külekci, and Bojian Xu. A simple yet time-optimal and linear-space algorithm for shortest unique substring queries. *Theor. Comput. Sci.*, 562:621–633, 2015.

[34] IUPAC-IUB Commission on Biochemical Nomenclature (CBN). Abbreviations and symbols for nucleic acids, polynucleotides and their constituents. *Journal of Molecular Biology*, 55(3):299–310, 1971.

[35] Dominik Köppl and Jannik Olbrich. 未決定文字列における欠如単語の検索の困難さ. Technical Report 15, Local Proceedings of the 200th アルゴリズム研究会, 11 2024.

[36] Dominik Köppl and Jannik Olbrich. 未決定文字列における一意単語の検索の困難さ. Technical Report 4, Local Proceedings of the 201th アルゴリズム研究会, 1 2025.

[37] Felipe A. Louza, Neerja Mhaskar, and W. F. Smyth. A new approach to regular & indeterminate strings. *Theor. Comput. Sci.*, 854:105–115, 2021.

[38] Takuya Mieno, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda. Shortest unique substring queries on run-length encoded strings. In *Proc. MFCS*, volume 58 of *LIPIcs*, pages 69:1–69:11, 2016.

[39] Takuya Mieno, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda. Tight bounds on the maximum number of shortest unique substrings. In *Proc. CPM*, volume 78 of *LIPIcs*, pages 24:1–24:11, 2017.

[40] Takuya Mieno, Dominik Köppl, Yuto Nakashima, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda. Space-efficient algorithms for computing minimal/shortest unique substrings. *Theor. Comput. Sci.*, 845:230–242, 2020.

[41] Takuya Mieno, Yuki Kuhara, Tooru Akagi, Yuta Fujishige, Yuto Nakashima, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda. Minimal unique substrings and minimal absent words in a sliding window. In *Proc. SOFSEM*, volume 12011 of *LNCS*, pages 148–160, 2020.

[42] Sumaiya Nazeen, M. Sohel Rahman, and Rezwana Reaz. Indeterminate string inference algorithms. *J. Discrete Algorithms*, 10:23–34, 2012.

[43] Mhaskar Neerja and William F. Smyth. Simple KMP pattern-matching on indeterminate strings. In *Proc. PSC*, pages 125–133, 2020.

[44] Kouta Okabe, Takuya Mieno, Yuto Nakashima, Shunsuke Inenaga, and Hideo Bannai. Linear-time computation of generalized minimal absent words for multiple strings. In *Proc. SPIRE*, volume 14240 of *LNCS*, pages 331–344, 2023.

[45] Takahiro Ota and Akiko Manada. A reconstruction of circular binary string using substrings and minimal absent words. *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.*, 107(3):409–416, 2024.

[46] Jian Pei, Wush Chi-Hsuan Wu, and Mi-Yen Yeh. On shortest unique substring queries. In *Proc. ICDE*, pages 937–948, 2013.

[47] Armando J. Pinho, Paulo Jorge S. G. Ferreira, Sara P. Garcia, and João M. O. S. Rodrigues. On finding minimal absent words. *BMC Bioinform.*, 10, 2009.

[48] Daniel W. Schultz and Bojian Xu. Parallel methods for finding k-mismatch shortest unique substrings using GPU. *IEEE ACM Trans. Comput. Biol. Bioinform.*, 18(1):386–395, 2021.

[49] William F. Smyth and Shu Wang. An adaptive hybrid pattern-matching algorithm on indeterminate strings. *Int. J. Found. Comput. Sci.*, 20(6):985–1004, 2009.

[50] Kazuya Tsuruta, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda. Shortest unique substrings queries in optimal time. In *Proc. SOFSEM*, volume 8327 of *LNCS*, pages 503–513, 2014.