様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科研費
KAKENHI

科学研究費助成事業　　研究成果報告書

令和　６年　６月１２日現在

機関番号：１１４０１
研究種目：基盤研究(C)（一般）
研究期間：2019～2023
課題番号：１９Ｋ１１８１５

研究課題名（和文）Clusters of repetition roots

研究課題名（英文）Clusters of repetition roots

研究代表者
Ｆａｚｅｋａｓ　Ｓｚｉｌａｒｄ（Fazekas, Szilard Zsolt）

秋田大学・理工学研究科・准教授

研究者番号：７０７２５３８２

交付決定額（研究期間全体）：（直接経費）　3,100,000円

研究成果の概要（和文）：目標は、シーケンス内で発生する x...x 形式の異なる繰り返しの数に対するより良い上限を得ることでした。私たちは、そのルート x がシーケンス内で現れる位置の集合、すなわち繰り返しのクラスタを通じてその数を研究する新しいアプローチを導入しました。各クラスタがそれに含まれる他のクラスタの数より大きいことを示すことを目指しました。

まず、いくつかの部分的な結果を示し、最終的に Rauzy グラフを用いた最適な上限を示した最近の結果を拡張して、一般的な仮説を証明しました。私たちの結果は、繰り返しルートのネストされたクラスタ構造を考慮することで、文字列の繰り返しを調査する新しい方向性を開きます。

研究成果の学術的意義や社会的意義
The significance of our results is that now we have better tools to study sequences containing many repetitions, which can lead to a better understanding of compression and pattern matching algorithms, which are of critical importance to our web infrastructure and computing in general.

研究成果の概要（英文）：The research goal was to obtain better upper bounds on the number of distinct repetitions of the form xx...x that can occur in a sequence. We introduced a new approach to study the number of such repetitions through the set of positions their root x occurs in the sequence, called the cluster of the repetition. We aimed to show that each cluster must be larger than the number of other clusters included in it.
During the project we first proved that our conjecture about clusters in some special cases. In the final year we worked on extending a recent result by Brlek and Li that proved the upper bound on such repetitions equal to the length of the string divided by the exponent minus one, using Rauzy graphs. We managed to extend the approach to prove our conjecture regarding the clusters of distinct repetition roots. Our result opens up new directions for investigating repetitions in strings by considering the nested cluster structures of the repetition roots.

研究分野：Computer science and combinatorics

キーワード：distinct repetitions　combinatorics　compressibility

１．研究開始当初の背景

Repetitions and periodicities are fundamental topics in the combinatorics of character sequences (words), and their study goes back to the founding of the field by Axel Thue. The most basic repetitive structure is xx, where x is a word. These are called *squares*, due to their form $xx = x^2$. We talk about cubes, such as $xxx = x^3$, and in general k-repetitions x^k, where x is the root and k is the exponent. Examples exist in natural languages, e.g., $hotshots = (hots)^2$, but binary sequences which store most of existing data generally contain significantly more repetitions.

The last two decades of research on repetitions focused in particular on maximal repetitions (runs) and distinct repetitions. Efficient algorithms for finding them were proposed and bounds on their maximal number have been extensively investigated, see [Crochemore, Ilie, Rytter, Theor. Comp. Sci, 2009]. In the case of runs, Bannai et al. gave an elegant proof that the upper bound is less than the length of the word [SIAM J. Comput., 2017], very close to the best known lower bound. **The case of distinct repetitions, however, was still wide open. Fraenkel and Simpson proved that the number of distinct squares in a word of length n is at most $2n$ and conjectured that the bound is less than n** [*J. Comb. Theory, Ser. A, 1998*]. The best known upper bound at the start of the prject was $11n/6$ [Deza et al., Discr. App. Math., 2015], whereas the best known lower bound is smaller than n. We proved earlier (Fazekas et al., 2008) that the bounds for distinct repetitions x^k are quadratic, $\theta(n^2)$, for all exponents $1 < k < 2$, highlighting the significance of squares, the smallest exponent where the upper bound is linear. For integer exponents $k > 2$, we showed earlier (Fazekas et al, 2011) that the number of distinct k-repetitions is always less than $n/(k-2)$.

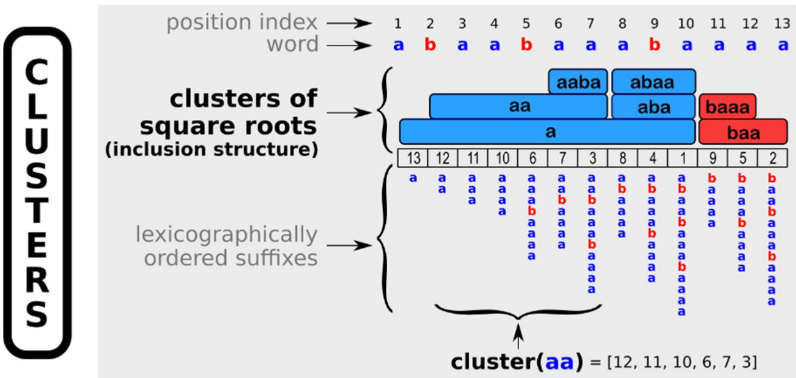
２．研究の目的

The purpose was to **improve the upper and lower bounds on the maximal number of distinct repetitions** in strings, searching to prove the then 20-year old conjecture of Fraenkel and Simpson. In order to achieve this, I aimed to **develop the theory of clusters of repetition roots**, which additionally could lead to general theorems about k-repetitions rather than fragmented results for specific values of k.

３．研究の方法

The approach I proposed has several original features, **relating distinct repetitions to number of occurrences** of factors, which allows fundamentally different reasoning than before:

- **Represent distinct repetitions by the clusters of their root** (see CLUSTERS) instead of the previously used distinguished position of a particular occurrence (e.g. start of the rightmost occurrence, etc.) and **group the repetitions lexicographically** (see CLUSTERS);
- **Count distinct repetitions with a common prefix as a function depending on the number of occurrences of the common prefix**. This brings a new strategy in proving upper bounds: show that the size of a cluster is greater than the number of clusters it includes (see GOALS).



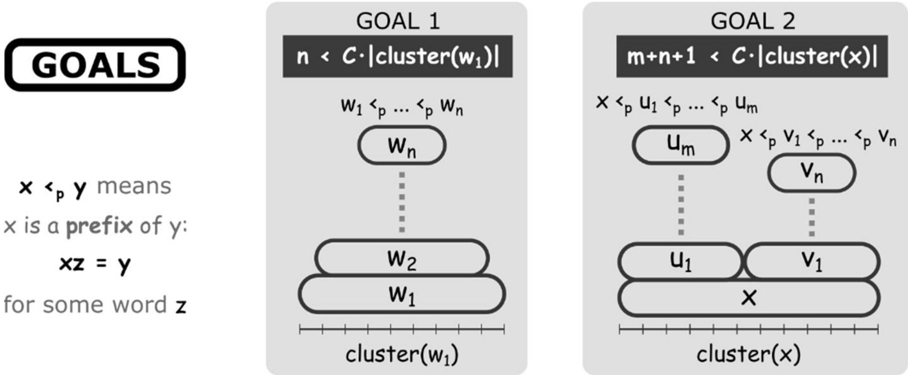
４．研究成果

According to the original plan I worked mainly with research collaborators Robert Mercas (Loughborough University, UK) and Shinnosuke Seki (UEC, Tokyo), both of whom published important results previously on the topic.

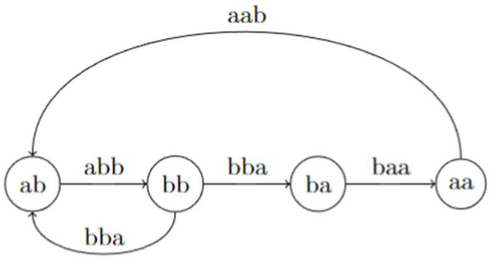
First, we worked on bounds for a single inclusion chain of clusters (GOAL 1), where the roots are totally ordered by the prefix ordering. We showed [1] that the conjecture holds for exponent 2, by considering so called anchors of distinct squares.

Next, we worked on generalizing the result on single inclusion chains to higher exponent and showing that the lower bound on the size of clusters is optimal by an effective construction [2].

The final piece of the puzzle was to prove the conjecture in general, that is in cases when two or more



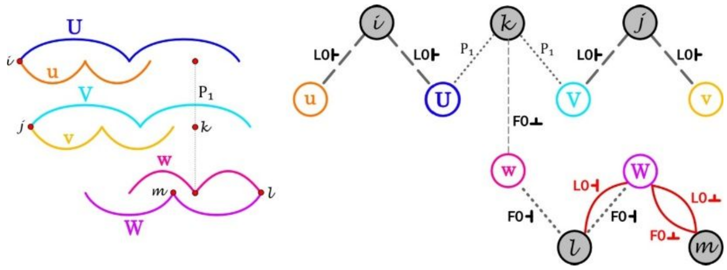
incomparable chains are included in a cluster (GOAL 2). This proof came together in the final year by collaboration with Shuo Li (Université du Québec à Montréal). The manuscript describing the final result is still in preparation with Li and Mercas. The approach used there is an extension of the method used by Brlek and Li to prove the Fraenkel and Simpson conjecture in 2023. It uses so called Rauzy graphs (see below), which we generalized to X-Rauzy graphs, that model the relationships among factors sharing a



common prefix X in a given word. By giving upper bounds on the number of short cycles in those graphs we showed that the cluster of each root U is strictly larger than the number of distinct repetition roots that have U as a prefix. This is a stronger result than the one conjectured by Fraenkel and Simpson, which implies the upper bound on the number of distinct repetitions proved by Brlek and Li.

Our results open up new directions for investigating repetitions in strings by considering the nested cluster structures of the repetition roots, and studying what structures allow for high repetition density in the strings. We showed earlier that our lower bound on cluster sizes is optimal when the roots are linearly ordered by the prefix relation. An interesting question to pursue is whether the lower bound is optimal when the roots form a non-linear partial order under the prefix relation.

I also tried an alternative approach to investigate the dense packing of squares while working with Seki. We introduced square networks (see example below) on words [3] which are bipartite graphs that model the relationships between various kinds of distinguished positions of distinct squares and their absolute position in a given packing word. We showed that certain structural restrictions on square networks translate directly to upper bounds on the number of distinct squares.



[1] Szilárd Zsolt Fazekas, Robert Mercas: **Clusters of Repetition Roots: Single Chains. SOFSEM 2021, Lecture Notes in Computer Science, Vol. 12607, pp. 400-409**

[2] Szilárd Zsolt Fazekas, Robert Mercas: **Clusters of Repetition Roots Forming Prefix Chains. DCFS 2022, Lecture Notes in Computer Science, Vol. 13439, pp. 43-56**

[3] Szilárd Zsolt Fazekas, Shinnosuke Seki: **Square network on a word. Theoretical Computer Science 894: 121-134 (2021)**

[4] **Szilard Zsolt Fazekas, Shuo Li, Robert Mercas: Clusters of repetition roots: multiple chains, manuscript in preparation**

５．主な発表論文等

〔雑誌論文〕　計6件（うち査読付論文　6件／うち国際共著　6件／うちオープンアクセス　1件）

１．著者名 Fazekas Szilard Zsolt、Mercas Robert	４．巻 13439
２．論文標題 Clusters of Repetition Roots Forming Prefix Chains	５．発行年 2022年
３．雑誌名 Lecture Notes in Computer Science	６．最初と最後の頁 43～56
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-031-13257-5_4	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

１．著者名 Fazekas Szilard Zsolt、Seki Shinnosuke	４．巻 894
２．論文標題 Square network on a word	５．発行年 2021年
３．雑誌名 Theoretical Computer Science	６．最初と最後の頁 121～134
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.tcs.2021.08.004	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

１．著者名 Fazekas Szilard Zsolt、Kim Hwee、Matsuoka Ryuichi、Morita Reoto、Seki Shinnosuke	４．巻 32
２．論文標題 Linear Bounds on the Size of Conformations in Greedy Deterministic Oritatami	５．発行年 2021年
３．雑誌名 International Journal of Foundations of Computer Science	６．最初と最後の頁 575～596
掲載論文のDOI（デジタルオブジェクト識別子） 10.1142/S0129054121410082	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

１．著者名 Fazekas Szilard Zsolt, Mercas Robert	４．巻 12607
２．論文標題 Clusters of Repetition Roots: Single Chains	５．発行年 2021年
３．雑誌名 Lecture Notes in Computer Science	６．最初と最後の頁 400～409
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-030-67731-2_29	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

〔学会発表〕　計6件（うち招待講演　0件／うち国際学会　0件）

1．発表者名 Szilard Fazekas
2．発表標題 Clusters of Repetition Roots Forming Prefix Chains
3．学会等名 24th IFIP WG 1.02 International Conference on Descriptional Complexity of Formal Systems
4．発表年 2022年

1．発表者名 Szilard Fazekas
2．発表標題 The general case of the clusters conjecture
3．学会等名 RIMS workshop on "Group, Ring, Language and Related Areas in Computer Science"
4．発表年 2023年

1．発表者名 Szilard Zsolt Fazekas
2．発表標題 Clusters of Repetition Roots: Single Chains
3．学会等名 SOFSEM 2021: 47th International Conference on Current Trends in Theory and Practice of Informatics
4．発表年 2021年

1．発表者名 Szilard Zsolt Fazekas
2．発表標題 Chains of repetition roots
3．学会等名 Algebraic system, Logic, Language and Related Areas in Computer Science II
4．発表年 2020年

1．発表者名 Szilard Zsolt Fazekas
2．発表標題 Clusters of repetition roots: generalization and optimality
3．学会等名 Nagaokakyo Seminar
4．発表年 2020年

1．発表者名 Szilard Zsolt Fazekas
2．発表標題 New upper bounds on chains of square root clusters
3．学会等名 Nagaokakyo Seminar
4．発表年 2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

－

6．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究協力者	メルカス　ロバート (Mercas Robert)	ラフバラー大学・Department of Computer Science・Senior Lecturer	
研究協力者	関　新之助 (Seki Shinnosuke)	電気通信大学・院情報理工学研究科・准教授	

7．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関			
英国	Loughborough University			