String Attractors of Pseudostandard and Rote sequences

Veronika Hendrychová

Supervisor: L'ubomíra Dvořáková (FNSPE, CTU, Czechia)
Consultant: Karel Břinda (Inria/IRISA Rennes, France)

December 1, 2023

Dictionary compression

- Motivation: handling huge **highly repetitive** text collections
  - Example: genomic databases (currently, data growing faster than computational capacities)

Dictionary compression

- Motivation: handling huge **highly repetitive** text collections
  - Example: genomic databases (currently, data growing faster than computational capacities)
- One of the classes of state-of-the-art compressors: dictionary compression
- **General principle:** storing a dictionary of repeated phrases
- Includes Lempel-Ziv methods (gzip), methods using Burrows-Wheeler transform (bzip2), grammars, ...

Repetitiveness measures of dictionary compressors

- Dictionary compressors induce **measures** of the word's repetitiveness
  - Can be used to assess the **word's complexity**
  - Specific to individual compression method
  - Obtained after the compression process

Repetitiveness measures of dictionary compressors

- Dictionary compressors induce **measures** of the word's repetitiveness
  - Can be used to assess the **word's complexity**
  - Specific to individual compression method
  - Obtained after the compression process

Question

Could be the measures based on dictionary compressors unified using a single combinatorial concept?

Preliminaries

- Alphabet: $\boxed{\mathcal{A} = \{\mathtt{0}, \mathtt{1}, \cdots \mathtt{m}\}}$
- Word (string) over the alphabet: $\boxed{w}$
  - e.g. $w = 1101323...$
- k-th letter in the word: $\boxed{w[k]}$
  - e.g. $w[2] = 0$
- Factor (substring) of the word: $\boxed{w[i..j] = w[i]w[i+1]...w[j]}$
  - e.g. $w[3..5] = 132$

String attractors

Definition of a string attractor [Prezza, ICTCS 2017]

A string attractor of a finite string w over alphabet $\mathcal{A}$ is a set of positions $\Gamma = \{j_1, ..., j_{|\Gamma|}\}$ such that every substring $w[i...j]$ has an occurrence $w[i'...j'] = w[i...j]$ with $j_k \in [i', ..., j']$, for some $j_k \in \Gamma$.

String attractors

Definition of a string attractor [Prezza, ICTCS 2017]

A string attractor of a finite string w over alphabet $\mathcal{A}$ is a set of positions $\Gamma = \{j_1, ..., j_{|\Gamma|}\}$ such that every substring $w[i...j]$ has an occurrence $w[i'...j'] = w[i...j]$ with $j_k \in [i', ..., j']$, for some $j_k \in \Gamma$.

Example:

$$w = 012300123012$$

$$\Gamma = \{2, 3, 4, 8, 10\} \leftrightarrow w = 01\underline{230}012\underline{3}0\underline{1}2$$

String attractors

Definition of a string attractor [Prezza, ICTCS 2017]

A string attractor of a finite string w over alphabet $\mathcal{A}$ is a set of positions $\Gamma = \{j_1, ..., j_{|\Gamma|}\}$ such that every substring $w[i...j]$ has an occurrence $w[i'...j'] = w[i...j]$ with $j_k \in [i', ..., j']$, for some $j_k \in \Gamma$.

Example:

$$w = 012300123012$$

$$\Gamma = \{2, 3, 4, 8, 10\} \leftrightarrow w = 01\boxed{230}012\boxed{3}0\boxed{1}2$$

$$\Gamma^* = \{2, 3, 4, 10\} \leftrightarrow w = 01\boxed{230}0123 0\boxed{1}2$$

String attractors

Definition of a string attractor [Prezza, ICTCS 2017]

A string attractor of a finite string w over alphabet $\mathcal{A}$ is a set of positions $\Gamma = \{j_1, ..., j_{|\Gamma|}\}$ such that every substring $w[i...j]$ has an occurrence $w[i'...j'] = w[i...j]$ with $j_k \in [i', ..., j']$, for some $j_k \in \Gamma$.

Example:

$$w = 012300123012$$

$$\Gamma = \{2, 3, 4, 8, 10\} \leftrightarrow w = 01\boxed{230}012\boxed{3}0\boxed{1}2$$

$$\Gamma^* = \{2, 3, 4, 10\} \leftrightarrow w = 01\boxed{230}01230\boxed{1}2$$

$$\Gamma^* = \{3, 5, 7, 10\} \leftrightarrow w = 012\boxed{3}0\boxed{0}1\boxed{2}30\boxed{1}2$$

Γ^* = some attractor with the minimum length

The key property of string attractors

[Kempa & Prezza, STOC 2018]

Dictionary compressors can be interpreted as approximation algorithms for the smallest string attractor.

The key property of string attractors

[Kempa & Prezza, STOC 2018]

Dictionary compressors can be interpreted as approximation algorithms for the smallest string attractor.

- Enable expressing **lower and upper bounds** for dictionary compressors, and comparing them
- **Direct stringological measure** (rather than the result of a specific compression method)
- To find the smallest attractor size is (generally) NP-hard problem

The key property of string attractors

[Kempa & Prezza, STOC 2018]

Dictionary compressors can be interpreted as approximation algorithms for the smallest string attractor.

- Enable expressing **lower and upper bounds** for dictionary compressors, and comparing them
- **Direct stringological measure** (rather than the result of a specific compression method)
- To find the smallest attractor size is (generally) NP-hard problem
  - However, adding **structural assumptions** on the individual data types (e.g., special classes of words) may make the computation tractable

Our work on string attractors

Overview

1. Study of the connection between attractors and dictionary compressors
2. Determination of attractors of specific sequences, and formal proofs of their form and minimality
3. Programs to support or disprove conjectures during the process

Our interest: pseudostandard and Rote sequences

- **Infinite sequences** over binary alphabet
  - Studying attractors of finite prefixes of various lengths
- **Low complexity** among aperiodic sequences
- Obtained by **palindromic and antipalindromic closures**
- **e.g.** 0110010110010110011001011001011001....

How to generate these sequences?

Palindromes

$|w| = n : \forall\, i \in \{0, .., n-1\} :$
$w[i] = w[n-i-1]$

e.g. $1001, 11011, 10101$

How to generate these sequences?

Palindromes

$|w| = n : \forall\, i \in \{0, .., n-1\} :$
$w[i] = w[n-i-1]$

e.g. $1001, 11011, 10101$

Palindromic closures

Pal. closure of w = the shortest palindrome having w as a prefix

e.g. $100 \rightarrow 1001$, $1011 \rightarrow 101101$

How to generate these sequences?

Palindromes

$|w| = n : \forall\, i \in \{0, .., n-1\}$:
$w[i] = w[n-i-1]$

e.g. $1001, 11011, 10101$

Antipalindromes

$|w| = n : \forall\, i \in \{0, .., n-1\}$:
$w[i] = \overline{w[n-i+1]}$,
$\overline{0} = 1$ and $\overline{1} = 0$

e.g. $111000, 1010, 110100$

Palindromic closures

Pal. closure of w = the shortest palindrome having w as a prefix

e.g. $100 \rightarrow 1001$, $1011 \rightarrow 101101$

How to generate these sequences?

Palindromes

$|w| = n : \forall\, i \in \{0, .., n-1\}:$
$w[i] = w[n - i - 1]$

e.g. $1001, 11011, 10101$

Palindromic closures

Pal. closure of w = the shortest palindrome having w as a prefix

e.g. $100 \rightarrow 1001$, $1011 \rightarrow 101101$

Antipalindromes

$|w| = n : \forall\, i \in \{0, .., n-1\}:$
$w[i] = \overline{w[n - i + 1]}$,
$\overline{0} = 1$ and $\overline{1} = 0$

e.g. $111000, 1010, 110100$

Antipalindromic closures

Antipal. closure of w = the shortest antipalindrome having w as a prefix

e.g. $10 \rightarrow 1010$, $1011 \rightarrow 10110010$

How to generate these sequences?

Palindromes
$|w| = n : \forall\, i \in \{0, .., n-1\}$:
$w[i] = w[n - i - 1]$

e.g. $1001, 11011, 10101$

Antipalindromes
$|w| = n : \forall\, i \in \{0, .., n-1\}$:
$w[i] = \overline{w[n - i + 1]}$,
$\overline{0} = 1$ and $\overline{1} = 0$

e.g. $111000, 1010, 110100$

Palindromic closures
Pal. closure of w = the shortest palindrome having w as a prefix

e.g. $100 \rightarrow 1001$, $1011 \rightarrow 101101$

Antipalindromic closures
Antipal. closure of w = the shortest antipalindrome having w as a prefix

e.g. $10 \rightarrow 1010$, $1011 \rightarrow 10110010$

Algorithm: Given any directive sequence $\rightarrow$ adding letter one by one and creating (anti)palindromic closures in each step

Result #1: Attractors of pseudostandard sequences

Example:

Directive sequence $\Delta = \boxed{0}\ \boxed{1}\ \boxed{0}\ \boxed{0}\ \boxed{1}\ \ldots$ + antipalindromic closures

Result #1: Attractors of pseudostandard sequences

Example:

Directive sequence $\Delta = 0\ 1\ 0\ 0\ 1\ \ldots$ + antipalindromic closures

$w_1 = 0\mathbf{1}$

Example:

Directive sequence $\Delta = 0\ 1\ 0\ 0\ 1\ \ldots$ + antipalindromic closures

$w_1 = 0\mathbf{1}$

$w_2 = 011\mathbf{001}$

Result #1: Attractors of pseudostandard sequences

Example:

Directive sequence $\Delta = \boxed{0}\ \boxed{1}\ \boxed{0}\ \boxed{0}\ \boxed{1}$ ... + antipalindromic closures

$w_1 = \boxed{0}\mathbf{1}$

$w_2 = \boxed{0}1\boxed{1}\mathbf{001}$

$w_3 = 01\boxed{1}001\boxed{0}\mathbf{11001}$

Result #1: Attractors of pseudostandard sequences

Example:

Directive sequence $\Delta = 0\ 1\ 0\ 0\ 1\ \ldots$ + antipalindromic closures

$$w_1 = 0\mathbf{1}$$

$$w_2 = 011\mathbf{001}$$

$$w_3 = 0110010\mathbf{11001}$$

$$w_4 = 011001011001 0\mathbf{11001}$$

Result #1: Attractors of pseudostandard sequences

Example:

Directive sequence $\Delta = 0\ 1\ 0\ 0\ 1\ \ldots$ + antipalindromic closures

$w_1 = 0\mathbf{1}$

$w_2 = 01 1\mathbf{001}$

$w_3 = 011001 0\mathbf{11001}$

$w_4 = 0110010110010\mathbf{11001}$

$w_5 = 01100101100101100 1\mathbf{0010110010110 01}$

Result #1: Attractors of pseudostandard sequences

Example:

Directive sequence $\Delta = 0\ 1\ 0\ 0\ 1\ \ldots$ + antipalindromic closures

$w_1 = 0\mathbf{1}$

$w_2 = 0\,1\,1\,\mathbf{0\,0\,1}$

$w_3 = 0\,1\,1\,0\,0\,1\,0\,\mathbf{1\,1\,0\,0\,1}$

$w_4 = 0\,1\,1\,0\,0\,1\,0\,1\,1\,0\,0\,1\,0\,\mathbf{1\,1\,0\,0\,1}$

$w_5 = 0\,1\,1\,0\,0\,1\,0\,1\,1\,0\,0\,1\,0\,1\,1\,0\,0\,1\,1\,\mathbf{0\,0\,1\,0\,1\,1\,0\,0\,1\,0\,1\,1\,0\,0\,1}$

Theorem [Dvořáková L., Hendrychová V., 2023]

Let w_n be a non-empty antipalindromic prefix of $w(\Delta, E^\omega)$ where the prefix of Δ of length n contains at least two 0's and one 1, and $\Delta = 0\cdots$. Then, when indexing from 0, a minimum size attractor is equal to $\{m_0, m_1, |w_n| - |m_1| - 1\}$, where
$m_\gamma = \max\{|q| : q \text{ is antipalindromic and } q\gamma \text{ is prefix of } w_n\}$.

Result #1: Attractors of pseudostandard sequences

Example:

Directive sequence $\Delta = 0\ 1\ 0\ 0\ 1 \ldots$ + antipalindromic closures

$w_1 = 0\mathbf{1}$

$w_2 = 011\mathbf{001}$

$w_3 = 011001\mathbf{011001}$

$w_4 = 011001011001\mathbf{011001}$

$w_5 = 011001011001011001\mathbf{100101100101011001}$

Theorem [Dvořáková L., Hendrychová V., 2023]

Let w_n be a non-empty antipalindromic prefix of $w(\Delta, E^\omega)$ where the prefix of Δ of length n contains at least two 0's and one 1, and $\Delta = 0\cdots$. Then, when indexing from 0, a minimum size attractor is equal to $\{m_0, m_1, |w_n| - |m_1| - 1\}$, where $m_\gamma = \max\{|q| : q \text{ is antipalindromic and } q\gamma \text{ is prefix of } w_n\}$.

Result #2: Attractors of Rote sequences

- Complementary-symmetric Rote sequences = combination of (anti)palindromic closures avoiding the following patterns: $\{(ab, EE) : a, b \in \{0,1\}\} \cup \{(a\bar{a}, RR) : a \in \{0,1\}\} \cup \{(aa, RE) : a \in \{0,1\}\}$.

Result #2: Attractors of Rote sequences

- Complementary-symmetric Rote sequences = combination of (anti)palindromic closures avoiding the following patterns: $\{(ab, EE) : a, b \in \{0,1\}\} \cup \{(a\bar{a}, RR) : a \in \{0,1\}\} \cup \{(aa, RE) : a \in \{0,1\}\}$.
- **Example directive bisequence**
  $\Delta = 0\ 0\ 1\ 1\ 0\ 0\ \ldots.$
  $\Theta = R\ R\ E\ R\ E\ R\ \ldots.$

Result #2: Attractors of Rote sequences

- Complementary-symmetric Rote sequences = combination of (anti)palindromic closures avoiding the following patterns: $\{(ab, EE) : a, b \in \{0,1\}\} \cup \{(a\bar{a}, RR) : a \in \{0,1\}\} \cup \{(aa, RE) : a \in \{0,1\}\}$.
- **Example directive bisequence**

$\Delta = 0\ 0\ 1\ 1\ 0\ 0\ \ldots.$

$\Theta = R\ R\ E\ R\ E\ R\ \ldots.$

$$w_1 = 0$$
$$w_2 = 00$$
$$w_3 = 001\mathbf{1}$$
$$w_4 = 00111\mathbf{00}$$
$$w_5 = 0011100\mathbf{011}$$
$$w_6 = 0011100011\mathbf{00011100}$$

Theorem [Dvořáková L., Hendrychová V., 2023]

Let w be a standard CS Rote sequence, then the size of the minimal attractor of any pseudopalindromic prefix equals the number of letters contained in the prefix. More precisely, if the directive bi-sequence (Δ, Θ) has $(0, R)$ as the first element, then the minimal attractors of the pseudopalindromic prefixes of w containing at least two letters are of the following form:

1. If $w_n = E(w_n)$, $\delta_n = a$, and w_i is the longest antipalindromic prefix of w_n followed by $\overline{a}$, then

$$\begin{aligned} \Gamma_1 &= \{|w_i|, |w_{n-1}|\}\,; \\ \Gamma_2 &= \{|w_{n-1}| - |w_i| - 1, |w_n| - |w_i| - 1\} \end{aligned}$$

are attractors of w_n.

Result #2: Attractors of Rote sequences

Theorem (continuation)

2. If $w_n = R(w_n)$, $\delta_n = a$, $\vartheta_{n-1} = E$, and w_j is the longest palindromic prefix of w_n followed by $\overline{a}$, then

$$\Gamma = \{|w_j|, |w_{n-1}|\}$$

is an attractor of w_n.

3. If $w_n = R(w_n)$, $\delta_n = a$, $\vartheta_{n-1} = R$, and m is the minimum index satisfying that $\vartheta_i = R$ for all $i \in \{m, \ldots, n\}$, then the attractor of w_m from Item 2 is simultaneously an attractor of w_n.

Algorithms and software implementation

- Programs to study attractors on **practical examples**
  - observations, manually unmanageable verification, disproving conjectures...

Algorithms and software implementation

- Programs to study attractors on **practical examples**
  - observations, manually unmanageable verification, disproving conjectures...
- **Outcome:** Four algorithms to work with sequences implemented in Python
  - Generator of prefixes of episturmian sequences and their attractors
  - Generator of prefixes of pseudostandard sequences and their attractors
  - General attractor verifier
  - General attractor generator

Summary and new questions

- Explained the string attractors' connection to dictionary compressors
- Newly discovered attractors of special prefixes of pseudostandard and CS Rote sequences, and proved their form and minimality
  - Paper available at https://arxiv.org/abs/2308.00850
- Implemented useful algorithms to work with episturmian, pseudostandard and general sequences

Summary and new questions

- Explained the string attractors' connection to dictionary compressors
- Newly discovered attractors of special prefixes of pseudostandard and CS Rote sequences, and proved their form and minimality
  - Paper available at https://arxiv.org/abs/2308.00850
- Implemented useful algorithms to work with episturmian, pseudostandard and general sequences

- What is the form of attractors of **generalized pseudostandard sequences**?
- How does the minimum attractor size affect the **form of examined words compressed** by dictionary compressors? Do they also remain constant?

Thank you for your attention!