Combinatorics on words in DNA computing

Š. Starosta

17. & 20. 5.2011

Outline

Outline

1. DNA computing
   - Introduction
   - DNA molecule
   - The first DNA computation
   - Operations in DNA computing
   - DNA computing revisited
   - What to avoid in a test tube

2. Combinatorics on Words
   - Setting in CoW
   - Results
   - More results
   - If there is some time left...

Introduction

barriers for “traditional” computers

1. HUP
2. von Neumann bottleneck

DNA/biomolecular computers: 1994 Leonard Adleman

Introduction

barriers for "traditional" computers

1. HUP
2. von Neumann bottleneck

DNA/biomolecular computers: 1994 Leonard Adleman

DNA - basics

deoxyribonucleic acid

1950s: DNA carries genetic information (double helix model)

structure: polymer chains - strands consisting of bases attached to sugar phosphate backbone

4 bases: A, G, T, C

Watson-Crick complementarity: A likes T, C likes G

bases are at a distance of $3,4$ ångströms, resulting in 18 Mbits per inch

DNA - basics

deoxyribonucleic acid

1950s: DNA carries genetic information (double helix model)

structure: polymer chains - strands consisting of bases attached to sugar phosphate backbone

4 bases: A, G, T, C

Watson-Crick complementarity: A likes T, C likes G

bases are at a distance of $3,4$ ångströms, resulting in 18 Mbits per inch

DNA - basics

deoxyribonucleic acid

1950s: DNA carries genetic information (double helix model)

structure: polymer chains - strands consisting of bases attached to sugar phosphate backbone

4 bases: A, G, T, C

Watson-Crick complementarity: A likes T, C likes G

bases are at a distance of $3,4$ ångströms, resulting in 18 Mbits per inch

DNA - basics

deoxyribonucleic acid

1950s: DNA carries genetic information (double helix model)

structure: polymer chains - strands consisting of bases attached to sugar phosphate backbone

4 bases: A, G, T, C

Watson-Crick complementarity: A likes T, C likes G

bases are at a distance of $3,4$ ångströms, resulting in 18 Mbits per inch

DNA - basics

deoxyribonucleic acid

1950s: DNA carries genetic information (double helix model)

structure: polymer chains - strands consisting of bases attached to sugar phosphate backbone

4 bases: A, G, T, C

Watson-Crick complementarity: A likes T, C likes G

bases are at a distance of $3,4$ ångströms, resulting in 18 Mbits per inch

DNA - basics

deoxyribonucleic acid

1950s: DNA carries genetic information (double helix model)

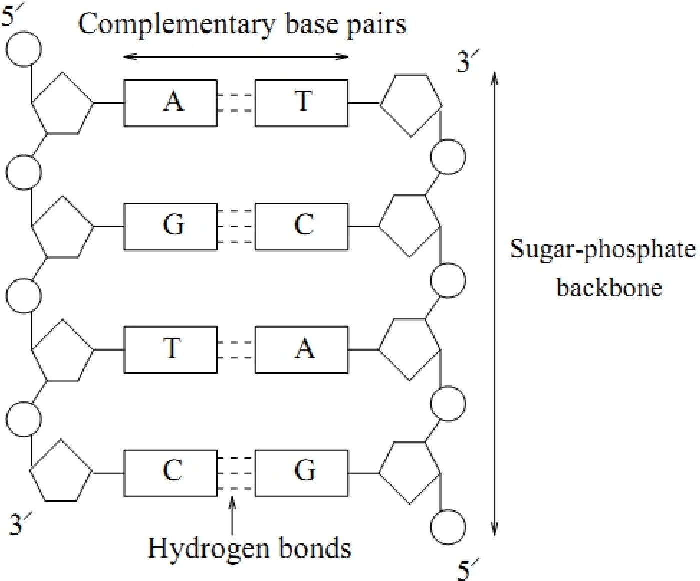
structure: polymer chains - strands consisting of bases attached to sugar phosphate backbone

4 bases: A, G, T, C

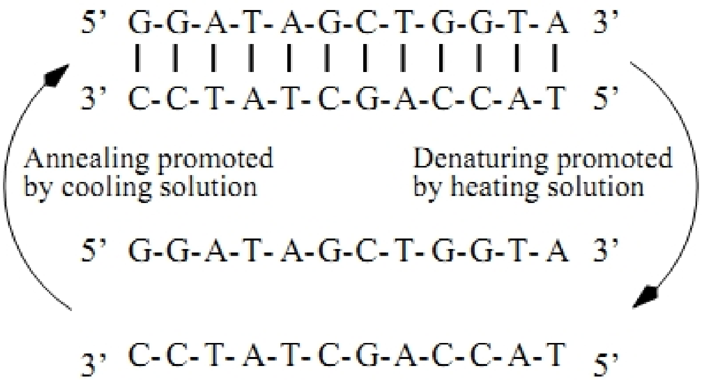
Watson-Crick complementarity: A likes T, C likes G

bases are at a distance of $3,4$ ångströms, resulting in 18 Mbits per inch

Structure of DNA

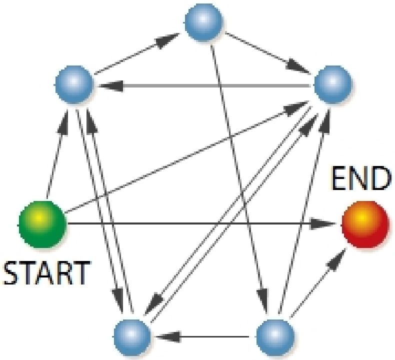


Watson-Crick complementarity and WK-palindromes



Hamiltonian Path Problem

Leonard Adleman in 1994: Hamiltonian Path Problem (NP-complete) on a graph with 7 vertices



Adleman's DNA computation I

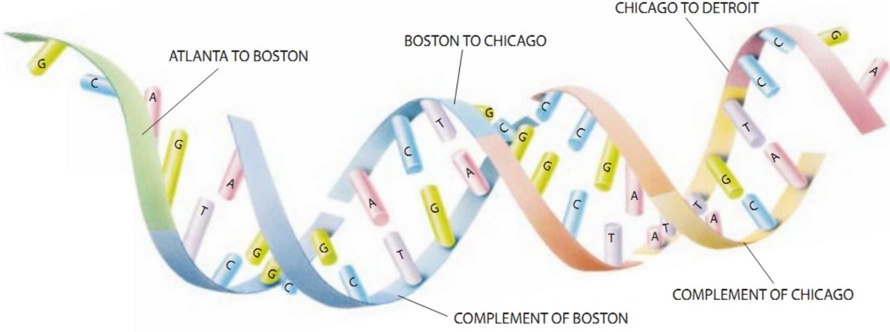
CITY	DNA NAME	COMPLEMENT
ATLANTA	ACTTGCAG	TGAACGTC
BOSTON	TCGGACTG	AGCCTGAC
CHICAGO	GGCTATGT	CCGATACA
DETROIT	CCGAGCAA	GGCTCGTT

Adleman's DNA computation I

CITY	DNA NAME	COMPLEMENT
ATLANTA	ACTTGCAG	TGAACGTC
BOSTON	TCGGACTG	AGCCTGAC
CHICAGO	GGCTATGT	CCGATACA
DETROIT	CCGAGCAA	GGCTCGTT

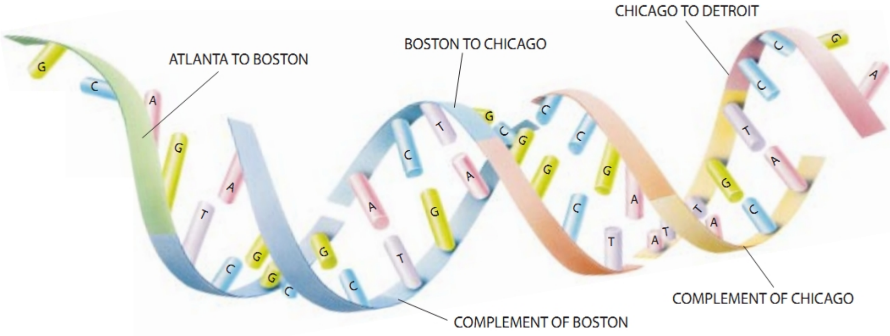
FLIGHT	DNA FLIGHT NUMBER
ATLANTA - BOSTON	GCAGTCGG
ATLANTA - DETROIT	GCAGCCGA
BOSTON - CHICAGO	ACTGGGCT
BOSTON - DETROIT	ACTGCCGA
BOSTON - ATLANTA	ACTGACTT
CHICAGO - DETROIT	ATGTCCGA

Adleman's DNA computation II



200-node instance would require 10^{24} Earth masses of DNA

Adleman's DNA computation II



200-node instance would require 10^{24} Earth masses of DNA

The likely frame

The scale of this ligation reaction far exceeded what was necessary for the graph under consideration. For each edge in the graph, approximately 3×10^{13} copies of the associated oligonucleotides were added to the ligation reaction. Hence it is likely that vast numbers of DNA molecules encoding the Hamiltonian path were created. In theory the creation of a single such molecule would be sufficient. Hence, for this graph, sub-attomol quantities of oligonucleotides would probably have been sufficient. Alternatively, a much larger graph could have been processed with the pmol quantities employed here.

Manipulation with DNA

- synthesis
- denaturing, annealing and ligation
- separation - affinity purification
- detect
- gel electrophoresis
- PCR - polymerase chain reaction
- cutting (using restriction enzymes)

Very general model of a DNA calculation

1. select the language - code
2. do the computation
3. decode

Advantages and drawbacks

Advantages:

- Size: the information density could go up to 1 bit per cube nm
- High parallelism: 10^9 calculations per ml of DNA per second
- Energy efficiency: 1019 operations per Joule

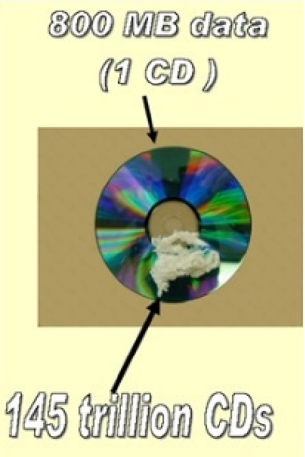
Drawbacks:

- required mass of DNA
- reagents
- need of manipulation by a human
- ...

Advantages and drawbacks

Advantages:

- Size: the information density could go up to 1 bit per cube nm
- High parallelism: 10^9 calculations per ml of DNA per second
- Energy efficiency: 1019 operations per Joule



Drawbacks:

- required mass of DNA
- reagents
- need of manipulation by a human
- ...

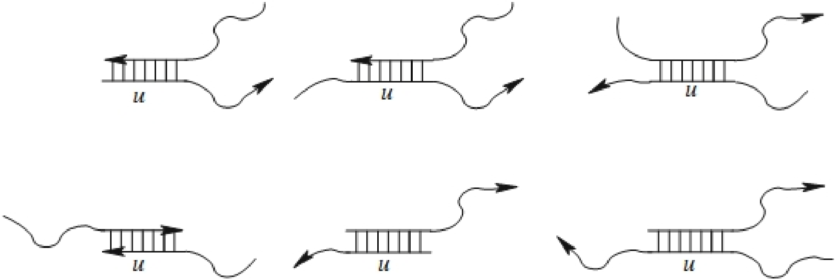
What problems can be solved by a DNA computer

Problems solved that can be found in literature:

- TSP - travelling salesman problem
- addition
- SAT - satisfiability problem
- DES cracking
- maximal clique problem
- ...

Intramolecular hybridization - hairpins

Intermolecular hybridization



$|u| = k$

Outline

Representing DNA strands

DNA strands are considered in their $5' \to 3'$ orientation as finite words over $\Delta = \{A, G, C, T\}$

involutive antimorphism WK: $A \leftrightarrow T$, $C \leftrightarrow G$

an encoding (i.e. initial set of words) for a computation is a language $L \subset \Delta^*$

Representing DNA strands

DNA strands are considered in their $5' \to 3'$ orientation as finite words over $\Delta = \{A, G, C, T\}$

involutive antimorphism WK: $A \leftrightarrow T$, $C \leftrightarrow G$

an encoding (i.e. initial set of words) for a computation is a language $L \subset \Delta^*$

Representing DNA strands

DNA strands are considered in their $5' \to 3'$ orientation as finite words over $\Delta = \{A, G, C, T\}$

involutive antimorphism WK: $A \leftrightarrow T$, $C \leftrightarrow G$

an encoding (i.e. initial set of words) for a computation is a language $L \subset \Delta^*$

Representing DNA strands

DNA strands are considered in their $5' \to 3'$ orientation as finite words over $\Delta = \{A, G, C, T\}$

involutive antimorphism WK: $A \leftrightarrow T$, $C \leftrightarrow G$

an encoding (i.e. initial set of words) for a computation is a language $L \subset \Delta^*$

Questions

How to choose the initial coding set L?

1. simulation
2. algorithm
3. theory

Bio-operations vs. L: after each bio-operation the language changes - do the properties hold?

Questions

How to choose the initial coding set L?

1. simulation
2. algorithm
3. theory

Bio-operations vs. L: after each bio-operation the language changes - do the properties hold?

Constraints on L

to avoid possible hybridizations, we impose several constraints on L

- L is Θ-k-m-subword code if $\forall u \in \mathcal{A}^k, 1 \leq i \leq m, \mathcal{A}^* u \mathcal{A}^i \Theta(u) \mathcal{A}^* \cap L = \emptyset$
- L is Θ-k-code if $\forall u, v \in \mathcal{A}^k \cap L, u \neq \Theta(v)$
- L is bond-free if $\forall u, v \in \mathcal{A}^k \cap L, H(u, \Theta(v)) > d$, where H is the Hamming distance
- frequency of C and G in each word should be $\frac{1}{2}$
- ...

Constraints on L

to avoid possible hybridizations, we impose several constraints on L

- L is Θ-k-m-subword code if $\forall u \in \mathcal{A}^k, 1 \leq i \leq m, \mathcal{A}^* u \mathcal{A}^i \Theta(u) \mathcal{A}^* \cap L = \emptyset$
- L is Θ-k-code if $\forall u, v \in \mathcal{A}^k \cap L, u \neq \Theta(v)$
- L is bond-free if $\forall u, v \in \mathcal{A}^k \cap L, H(u, \Theta(v)) > d$, where H is the Hamming distance
- frequency of C and G in each word should be $\frac{1}{2}$
- ...

Constraints on L

to avoid possible hybridizations, we impose several constraints on L

- L is Θ-k-m-subword code if $\forall u \in \mathcal{A}^k, 1 \leq i \leq m, \mathcal{A}^* u \mathcal{A}^i \Theta(u) \mathcal{A}^* \cap L = \emptyset$
- L is Θ-k-code if $\forall u, v \in \mathcal{A}^k \cap L, u \neq \Theta(v)$
- L is bond-free if $\forall u, v \in \mathcal{A}^k \cap L, H(u, \Theta(v)) > d$, where H is the Hamming distance
- frequency of C and G in each word should be $\frac{1}{2}$
- ...

Constraints on L

to avoid possible hybridizations, we impose several constraints on L

- L is Θ-k-m-subword code if $\forall u \in \mathcal{A}^k, 1 \leq i \leq m, \mathcal{A}^* u \mathcal{A}^i \Theta(u) \mathcal{A}^* \cap L = \emptyset$
- L is Θ-k-code if $\forall u, v \in \mathcal{A}^k \cap L, u \neq \Theta(v)$
- L is bond-free if $\forall u, v \in \mathcal{A}^k \cap L, H(u, \Theta(v)) > d$, where H is the Hamming distance
- frequency of C and G in each word should be $\frac{1}{2}$
- ...

Constraints on L

to avoid possible hybridizations, we impose several constraints on L

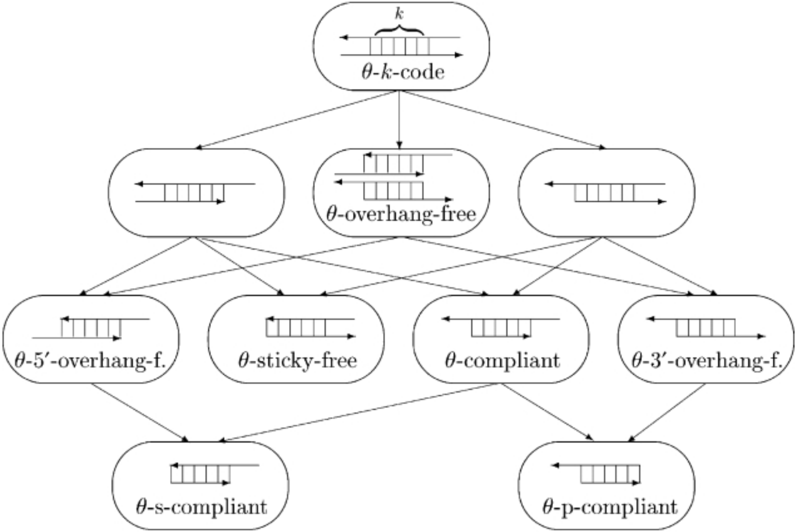
- L is Θ-k-m-subword code if $\forall u \in \mathcal{A}^k, 1 \leq i \leq m, \mathcal{A}^* u \mathcal{A}^i \Theta(u) \mathcal{A}^* \cap L = \emptyset$
- L is Θ-k-code if $\forall u, v \in \mathcal{A}^k \cap L, u \neq \Theta(v)$
- L is bond-free if $\forall u, v \in \mathcal{A}^k \cap L, H(u, \Theta(v)) > d$, where H is the Hamming distance
- frequency of C and G in each word should be $\frac{1}{2}$
- ...

Constraints on L

to avoid possible hybridizations, we impose several constraints on L

- L is Θ-k-m-subword code if $\forall u \in \mathcal{A}^k, 1 \leq i \leq m, \mathcal{A}^* u \mathcal{A}^i \Theta(u) \mathcal{A}^* \cap L = \emptyset$
- L is Θ-k-code if $\forall u, v \in \mathcal{A}^k \cap L, u \neq \Theta(v)$
- L is bond-free if $\forall u, v \in \mathcal{A}^k \cap L, H(u, \Theta(v)) > d$, where H is the Hamming distance
- frequency of C and G in each word should be $\frac{1}{2}$
- ...

Constraints on L ...



Conjugates and Θ-conjugates

Proposition (Lyndon and Schützenberger, 1962)

Let $u, v, w \in \mathcal{A}^*$ such that $uv = vw$. Then there exist $p, q \in \mathcal{A}^+$ such that $u = pq$, $w = qp$ and $v = p(qp)^i$ for $i \geq 0$.

Proposition (Kari and Mahalingam, 2007)

Let $u, v, w \in \mathcal{A}^+$ such that $uv = \Theta(v)w$, where Θ is an involutive morphism or antimorphism over $\mathcal{A}$.

1. *If Θ is an antimorphism, then there exist $x, y \in \mathcal{A}^*$ such that either*
   - $u = xy$, $w = y\Theta(x)$ and $v = \Theta(x)$; or
   - $u = x = \Theta(w)$, $y = \Theta(y)$ and $v = y$.
2. *If Θ is a morphism, then there exist $x, y \in \mathcal{A}^*$ such that $u = xy$ and one of the following hold:*
   - $w = y\Theta(x)$ and $v = (\Theta(xy)xy)^i\Theta(x)$ for some $i \geq 0$;
   - $w = \Theta(y)x$ and $v = (\Theta(xy)xy)^i\Theta(xy)x$ for some $i \geq 0$.

Conjugates and Θ-conjugates

Proposition (Lyndon and Schützenberger, 1962)

Let $u, v, w \in \mathcal{A}^*$ such that $uv = vw$. Then there exist $p, q \in \mathcal{A}^+$ such that $u = pq$, $w = qp$ and $v = p(qp)^i$ for $i \geq 0$.

Proposition (Kari and Mahalingam, 2007)

Let $u, v, w \in \mathcal{A}^+$ such that $uv = \Theta(v)w$, where Θ is an involutive morphism or antimorphism over $\mathcal{A}$.

1. *If Θ is an antimorphism, then there exist $x, y \in \mathcal{A}^*$ such that either*
   - $u = xy$, $w = y\Theta(x)$ *and* $v = \Theta(x)$; *or*
   - $u = x = \Theta(w)$, $y = \Theta(y)$ *and* $v = y$.
2. *If Θ is a morphism, then there exist $x, y \in \mathcal{A}^*$ such that $u = xy$ and one of the following hold:*
   - $w = y\Theta(x)$ *and* $v = (\Theta(xy)xy)^i\Theta(x)$ *for some* $i \geq 0$;
   - $w = \Theta(y)x$ *and* $v = (\Theta(xy)xy)^i\Theta(xy)x$ *for some* $i \geq 0$.

Commutativity and Θ-commutativity

Proposition (Lyndon and Schützenberger, 1962)

Let $u, v \in \mathcal{A}^*$ such that $uv = vu$. Then there exist $p \in \mathcal{A}^+$ such that $u = p^i$ and $w = p^j$ for $i, j > 0$ (i.e. it has cyclic solution).

Proposition (Kari and Mahalingam, 2007)

Let $u, v \in \mathcal{A}^+$ such that $uv = \Theta(v)u$, where Θ is an involutive morphism or antimorphism over $\mathcal{A}$.

1. *If Θ is an antimorphism, then there exist $x, y \in \mathcal{A}^*$ such that $u = x(yx)^n$, $v = (yx)^m$, $x = \Theta(x)$, $y = \Theta(y)$, $m \geq 1$ and $n \geq 0$.*
2. *If Θ is a morphism, then there exists $x \in \mathcal{A}^+$ such that:*
   - *$x = \Theta(x)$, $u = x^i$ and $v = x^j$ for some $i, j \geq 1$;*
   - *$u = x(\Theta(x)x)^i$ and $v = (x\Theta(x))^j$ for some $i \geq 0$ and $j \geq 1$.*

Commutativity and Θ-commutativity

Proposition (Lyndon and Schützenberger, 1962)

Let $u, v \in \mathcal{A}^*$ such that $uv = vu$. Then there exist $p \in \mathcal{A}^+$ such that $u = p^i$ and $w = p^j$ for $i, j > 0$ (i.e. it has cyclic solution).

Proposition (Kari and Mahalingam, 2007)

Let $u, v \in \mathcal{A}^+$ such that $uv = \Theta(v)u$, where Θ is an involutive morphism or antimorphism over $\mathcal{A}$.

1. *If Θ is an antimorphism, then there exist $x, y \in \mathcal{A}^*$ such that $u = x(yx)^n$, $v = (yx)^m$, $x = \Theta(x)$, $y = \Theta(y)$, $m \geq 1$ and $n \geq 0$.*
2. *If Θ is a morphism, then there exists $x \in \mathcal{A}^+$ such that:*
   - *$x = \Theta(x)$, $u = x^i$ and $v = x^j$ for some $i, j \geq 1$;*
   - *$u = x(\Theta(x)x)^i$ and $v = (x\Theta(x))^j$ for some $i \geq 0$ and $j \geq 1$.*

Set of Θ-palindromes

Proposition

The set of all Θ-palindromes (Θ being an involutive antimorphism) is not regular.

Lemma (Pumping lemma for regular languages)

Let L be a regular language. Then there exists an integer $p \geq 1$ depending only on L such that every $w \in L$ of length at least p can be written as $w = xyz$ satisfying the following conditions:

1. $|y| \geq 1$,
2. $|xy| \geq p$,
3. for all $i \geq 0$, $xy^iz \in L$.

Proposition

The set of all Θ-palindromes is context-free.

Set of Θ-palindromes

Proposition

The set of all Θ-palindromes (Θ being an involutive antimorphism) is not regular.

Lemma (Pumping lemma for regular languages)

Let L be a regular language. Then there exists an integer $p \geq 1$ depending only on L such that every $w \in L$ of length at least p can be written as $w = xyz$ satisfying the following conditions:

1. $|y| \geq 1$,
2. $|xy| \geq p$,
3. *for all* $i \geq 0$, $xy^iz \in L$.

Proposition

The set of all Θ-palindromes is context-free.

Set of Θ-palindromes

Proposition

The set of all Θ-palindromes (Θ being an involutive antimorphism) is not regular.

Lemma (Pumping lemma for regular languages)

Let L be a regular language. Then there exists an integer $p \geq 1$ depending only on L such that every $w \in L$ of length at least p can be written as $w = xyz$ satisfying the following conditions:

1. $|y| \geq 1$,
2. $|xy| \geq p$,
3. *for all* $i \geq 0$, $xy^iz \in L$.

Proposition

The set of all Θ-palindromes is context-free.

Untitled frame

Proposition

Let Θ be either an involutive morphism or antimorphism and let $w, v, u \in \mathcal{A}^+$ such that $wu = \Theta(u)v$ and $w\Theta(u) = uv$. Then $w = (xy)^m$, $y = (yx)^m$ and $u = (xy)^n x$ for some $x, y \in \mathcal{A}^*$, $x = \Theta(x)$, $y = \Theta(y)$, $m \geq 1$, $n \geq 0$.

Proposition

Let $u \in \mathcal{A}$ be a non-empty word such that $u \neq \Theta(u)$. Then the following statements are equivalent

1. *$\sqrt{u}$ is the product of two non-empty Θ-palindromes.*
2. *There exists a non-empty Θ-palindrome v such that Θ-commutes with u.*
3. *u is a product of two non-empty Θ-palindromes.*

Untitled frame

Proposition

Let Θ be either an involutive morphism or antimorphism and let $w, v, u \in \mathcal{A}^+$ such that $wu = \Theta(u)v$ and $w\Theta(u) = uv$. Then $w = (xy)^m$, $y = (yx)^m$ and $u = (xy)^n x$ for some $x, y \in \mathcal{A}^*$, $x = \Theta(x)$, $y = \Theta(y)$, $m \geq 1$, $n \geq 0$.

Proposition

Let $u \in \mathcal{A}$ be a non-empty word such that $u \neq \Theta(u)$. Then the following statements are equivalent

1. *$\sqrt{u}$ is the product of two non-empty Θ-palindromes.*
2. *There exists a non-empty Θ-palindrome v such that Θ-commutes with u.*
3. *u is a product of two non-empty Θ-palindromes.*

Second untitled frame

Proposition

Let Θ be either a morphic or an antimorphic involution and let Σ be such that for all $a \in \Sigma$, $a \neq \Theta(a)$.
(...)

References

Lila Kari

Thank you.