DNA Watson-Crick complementarity in computer science

Shinnosuke Seki
(Supervisor) Dr. Lila Kari

Department of Computer Science, University of Western Ontario

Ph.D. thesis defense public lecture
August 9, 2010
Path to the understanding of DNA information processing

1953 Watson and Crick discovered the DNA double helix

discoveries of fundamental information processing mechanisms in living organisms like DNA replication.

1994 Adleman’s first experiment of molecular computing (DNA computing)

2003 Completion of Human Genome Project

8/9/2010 one small step to the understanding of DNA information encoding
In molecular computing, DNA strands are often employed as media for information processing, storage, and transmission.

**DNA strand design**

Good DNA strands for molecular computing $\approx$
Encoding information into DNA

In molecular computing, DNA strands are often employed as media for information processing, storage, and transmission.

**DNA strand design**

Good DNA strands for molecular computing ≈

Knowledge in computer science
- coding theory
- information theory
- etc.
In molecular computing, DNA strands are often employed as media for information processing, storage, and transmission.

**DNA strand design**

Good DNA strands for molecular computing ≈

- Knowledge in computer science
  - coding theory
  - information theory
  - etc.

+ Particularities of DNA information
  - 4-letters
  - Watson-Crick complementarity
Criteria of DNA strand design [Sager & Stefanovic 2006]

1. no strand forms any undesired intra-molecular structure
2. no strand hybridizes with a strand in any undesirable manner
3. no strand hybridizes with the WK-complement of a strand in any undesirable manner
## Criteria of DNA strand design [Sager & Stefanovic 2006]

1. no strand forms any undesired intra-molecular structure
2. no strand hybridizes with a strand in any undesirable manner
3. no strand hybridizes with the **WK-complement** of a strand in any undesirable manner
Due to

- Watson-Crick (WK) complementarity: $A \rightarrow T$, $C \rightarrow G$,
- Anti-parallelism,

Two **WK-complementary** DNA single strands with opposite orientation can form a DNA double helix.
DNA replication works in the following way:

Watson strand

Crick strand
DNA replication works in the following way:

Watson strand

Crick strand
DNA replication works in the following way:

Watson strand

Crick strand

Watson strand

Crick strand
DNA replication works in the following way:

Watson strand

Crick strand

Watson strand

Crick strand

N.B., the figure above lacks various important features of DNA replication such as replication fork, Okazaki fragment, leading and lagging strands, and DNA ligase.

**Observation**

Two WK-complementary strands are equivalent w.r.t. the information they encode.
Definition

A mapping $\theta : \Sigma^* \rightarrow \Sigma^*$ is called an **antimorphic involution** if

1. for any $x, y \in \Sigma^*$, $\theta(xy) = \theta(y)\theta(x)$ (antimorphism),

2. $\theta \circ \theta$ is the identity (involutive).

Watson-Crick (WK-) involution

The antimorphic involution $\tau$ defined as $\tau(A) = T$, $\tau(T) = A$, $\tau(C) = G$, $\tau(G) = C$ models WK-complementarity, and hence, called WK-involution. For instance,

$$\tau(ACTG) = \tau(G)\tau(ACT) = \cdots = \tau(G)\tau(T)\tau(C)\tau(A) = CAGT.$$
Formalization of the equivalence

For any word $u$, consider $u$ and its WK-complement $\tau(u)$ equivalent.

Idea for generalization

How about regarding $u$ and $\theta(u)$ equivalent for any word $u$ and any antimorphic involution $\theta$?
Ultimate Goal
Establish a system with notions and theorems in combinatorics on words which can handle a word and its complement in a uniform manner.

The main contribution of this thesis
We will generalize the classical notions of *power*, *repetition*, and even *primitivity* of words in this thesis.

Example
ACTGCAGTCAGT can be considered “repetitive” because it can be written as $ACTG_\tau (ACTG)^2$. 
Recall the above-mentioned criteria for DNA strand design.

Criteria of DNA strand design [Sager & Stefanovic 2006]

1. no strand forms any undesired intra-molecular structure
2. no strand hybridizes with a strand in any undesirable manner
3. no strand hybridizes with the WK-complement of a strand in any undesirable manner
Importance in applications

Recall the above-mentioned criteria for DNA strand design.

**Criteria of DNA strand design [Sager & Stefanovic 2006]**

1. no strand forms any undesired intra-molecular structure
2. no strand hybridizes with a strand in any undesirable manner
3. no strand hybridizes with the WK-complement of a strand in any undesirable manner

Our results in this thesis aim at providing uniform treatment of 2nd and 3rd criteria.
A word $w \in \Sigma^+$ is called a **power of** $u$ if $w \in u^+$.

**Definition**

A word $w \in \Sigma^+$ is **primitive** if $w = u^k$ implies $k = 1$. The (unique) primitive word $u \in \Sigma^+$ with $w \in u^+$ is called the **primitive root** of $w$, denoted by $\rho(w)$.

**Example**

ACGT is primitive, while ACGTACGT is not; $\rho(ACGTACGT) = ACGT$. 
**θ-primitivity of words**

We call a word in the set \( u\{u, \theta(u)\}^* \) a \( \theta \)-power of \( u \).

**Definition**

We say that a word \( w \in \Sigma^+ \) is \( \theta \)-primitive if for any \( u \in \Sigma^+ \), \( w \in \{u, \theta(u)\}^k \) implies \( k = 1 \).

**Example**

For the WK-involution \( \tau \), ACGT is not \( \tau \)-primitive because \( GT = \tau(AC) \), and hence, \( ACGT = AC\tau(AC) \).
Definition of $\theta$-primitive root

Problem

Can we define the $\theta$-primitive root of a word $w$ as the unique $\theta$-primitive word $t$ such that $w \in t\{t, \theta(t)\}^*$?

The answer is YES. The existence of such $t$ is trivial. Its uniqueness will be a corollary of extended Fine and Wilf’s theorem.
A relationship between primitivity and $\theta$-primitivity

Proposition

For any antimorphic involution $\theta$, a $\theta$-primitive word is primitive.

Example

ACGT is primitive but not $\tau$-primitive for the WK-involution $\tau$. 
A well-known example of defect effect

On $(x, y)$-coordinate, the condition $y = x$ decreases the degree of freedom by 1 ($x$ is a free variable, and $y$ becomes a bound variable).
Defect effect II

For given words and a system of equations on them, if the system forces some of the involved words to be powers of same word, then the system is said to possess the defect effect.

Example

$uv = vu$ implies $\rho(u) = \rho(v)$ (defect effect on $u, v$).

Example ([Choffrut & Karhumäki 1997])

If a word in $u\{u, v\}^*$ and a word in $v\{u, v\}^*$ share a prefix of length $|u| + |v|$, then $\rho(u) = \rho(v)$. 
Weak defect effect problem

Does a given system of word equations force two words $u, v$ involved to be in $\{t, \theta(t)\}^*$ for some word $t$?

Theorem ([Czeizler, Kari, & Seki 2010])

\[ \rho(u) = \rho(v) \text{ implies that } u, v \in t\{t, \theta(t)\}^* \text{ for some word } t, \text{ and hence, has a weak defect effect on } u \text{ and } v. \]
We investigate the weak defect effect problem in contexts of

1. (extended) Fine and Wilf’s theorem;
2. (extended) Lyndon-Schützenberger equation.
Main contributions to weak defect effect

We investigate the weak defect effect problem in contexts of

1. (extended) Fine and Wilf’s theorem;
2. (extended) Lyndon-Schützenberger equation.
Theorem ([Fine & Wilf 1965])

If a power of $u$ and a power of $v$ share a prefix of length $|u| + |v| - \gcd(|u|, |v|)$, then $\rho(u) = \rho(v)$. 
An extended Fine and Wilf’s theorem

Problem

How long prefix do $\alpha \in u\{u, \theta(u)\}^*$ and $\beta \in v\{v, \theta(v)\}^*$ have to share to imply that $u, v \in t\{t, \theta(t)\}^*$ for some word $t$?
**exFW theorem**

If a $\theta$-power of $u$ and a $\theta$-power of $v$ share a prefix of length $\text{lcm}(|u|, |v|)$, then $u, v \in t\{t, \theta(t)\}^*$ for some word $t$.

**Definition of $\theta$-primitive root** [Czeizler, Kari, & Seki 2010]

We can define the $\theta$-primitive root of $u$, denoted by $\rho_\theta(u)$, as a unique $\theta$-primitive word $t$ such that $u \in t\{t, \theta(t)\}^*$.
exFW theorem 2 [Czeizler, Kari, & Seki 2010]

Let \( u, v \in \Sigma^+ \) with \( |u| \geq |v| \). If a \( \theta \)-power of \( u \) and a \( \theta \)-power of \( v \) share a prefix of length

\[
2|u| + |v| - \gcd(|u|, |v|),
\]

then \( u, v \in t\{t, \theta(t)\}^* \) for some word \( t \).
For two positive integers $p, q \in \mathbb{N} \ (p \geq q)$, let

$$b(p, q) = \begin{cases} 
\operatorname{lcm}(p, q) & \text{if } q \leq 2 \gcd(p, q) \\
2p + q - \gcd(p, q) & \text{if } q \geq 3 \gcd(p, q).
\end{cases}$$

extended Fine and Wilf’s theorem

If a $\theta$-power of $u$ and a $\theta$-power of $v$ share a prefix of length $b(|u|, |v|)$, then $u, v \in t\{t, \theta(t)\}^*$ for some word $t$. 
The bound for FW theorem is strong optimal [Constantinescu & Ilie 2005] in a sense: for any positive integers $p, q$, one can construct $u, v$ such that

1. $|u| = p$, $|v| = q$, $\rho(u) \neq \rho(v)$,
2. a power of $u$ and a power of $v$ share a prefix of length $|u| + |v| - \gcd(|u|, |v|) - 1$.

**Problem**

Is $b(p, q)$ strongly optimal?

**A partial (trivial?) positive answer**

$b(p, q)$ is optimal for any $p, q$ with $p \geq q = \gcd(p, q)$.

**Negative answer [Kari & Seki 2010]**

$b(p, q)$ is not strongly optimal.
For $p, q$ with $p > q \geq 2 \gcd(p, q)$, let

$$b'(p, q) = b(p, q) - \left\lfloor \frac{\gcd(p, q)}{2} \right\rfloor.$$
Optimality of the new bound

A partial positive answer

For any \( p, q \) with \( p > q = 2 \gcd(p, q) \), \( b'(p, q) \) is optimal.

iff condition for \( b'(p, q) \) to be optimal for \( (p, q) \)

Let \( u, v \) with \( |u| > |v| \geq 3 \gcd(|u|, |v|) \), \( \rho_{\theta}(u) \neq \rho_{\theta}(v) \), and \( \gcd(|u|, |v|) = 1 \). A \( \theta \)-power of \( u \) and a \( \theta \)-power of \( v \) share a prefix of length \( b'(|u|, |v|) - 1 \) iff either

1. \( u = (ab(ba)^i b)^m ab \) and \( v = ab(ba)^i b \), or
2. \( u = [a(ba)^i(ab)^i+1 a]^m a(ba)^i ab \) and \( v = a(ba)^i(ab)^i+1 a \)

for some \( m \geq 1 \), \( i \geq 0 \), and \( a, b \in \Sigma \) s.t. \( a = \theta(a) \) and \( b = \theta(b) \).

Consequently, \( b'(p, q) \) is NOT strongly optimal (e.g., it is not optimal for \( (9, 5) \)).
Optimality of the new bound

A partial positive answer

For any $p, q$ with $p > q = 2 \gcd(p, q)$, $b'(p, q)$ is optimal.

iff condition for $b'(p, q)$ to be optimal for $(p, q)$

Let $u, v$ with $|u| > |v| \geq 3 \gcd(|u|, |v|)$, $\rho_\theta(u) \neq \rho_\theta(v)$, and $\gcd(|u|, |v|) = 1$. A $\theta$-power of $u$ and a $\theta$-power of $v$ share a prefix of length $b'(|u|, |v|) - 1$ iff either

1. $u = (ab(ba)^i b)^m ab$ and $v = ab(ba)^i b$, or
2. $u = [a(ba)^i(ab)^i+1 a]^m a(ba)^i ab$ and $v = a(ba)^i(ab)^i+1 a$ for some $m \geq 1$, $i \geq 0$, and $a, b \in \Sigma$ s.t. $a = \theta(a)$ and $b = \theta(b)$.

Consequently, $b'(p, q)$ is NOT strongly optimal (e.g., it is not optimal for $(9, 5)$).
We investigate the weak defect effect problem in contexts of

1. (extended) Fine and Wilf's theorem;

2. (extended) Lyndon-Schützenberger equation.
Lyndon-Schützenberger equation

Lyndon-Schützenberger equation
[Lyndon & Schützenberger 1962]

For words $u, v, w \in \Sigma^*$ and non-negative integers $\ell, n, m \geq 0$, an equation of the form

$$u^\ell = v^n w^m$$

is called the Lyndon-Schützenberger equation.


For $\ell, n, m \geq 2$, the equation $u^\ell = v^n w^m$ implies $u, v, w \in t^+$ for some word $t$. 
An extended Lyndon Schützenberger equation

Let us extend the LS-equation as:

\[ u_1 \cdots u_\ell = v_1 \cdots v_n w_1 \cdots w_m, \]

where \( u_1, \ldots, u_\ell \in \{u, \theta(u)\} \), \( v_1, \ldots, v_n \in \{v, \theta(v)\} \), and \( w_1, \ldots, w_m \in \{w, \theta(w)\} \).

**Problem**

Find conditions on \( \ell, n, m \) under which the exLS equation \((\ell, n, m)\) possesses the weak defect effect, i.e., implies \( u, v, w \in \{t, \theta(t)\}^+ \) for some word \( t \).
In [Czeizler, Czeizler, Kari, & Seki 2009], we provided examples to verify the following.

Proposition ([Czeizler, Czeizler, Kari, & Seki 2009])

If one of \( \ell, n, m \) is 2, then exLS equations \((\ell, n, m)\) does NOT possess the weak defect effect.

As a result, all of \( \ell, n, m \) must be at least 3 for exLS equations \((\ell, n, m)\) to possess the weak defect effect.
Proposition ([Czeizler, Czeizler, Kari, & Seki 2009])

**ExLS equations** $(\geq 6, \geq 3, \geq 3)$ possess the weak defect effect.

Symmetry enables us to assume that $|v_1 \cdots v_n| \geq |w_1 \cdots w_m|$. Since $\ell \geq 6$, this means $|v_1 \cdots v_n| \geq \frac{1}{2}|u_1 \cdots u_\ell| \geq |u_1 u_2 u_3|$. Thus, $u_1 \cdots u_\ell$ and $v_1 \cdots v_n$ share a prefix of length at least

$$\max(3|u|, 3|v|) \geq 2 \max(|u|, |v|) + \min(|u|, |v|).$$
Proof

The common prefix between $u_1 \cdots u_\ell$ and $v_1 \cdots v_n$ is long enough to apply exFW theorem to obtain $u, v \in \{t, \theta(t)\}^*$ for some $\theta$-primitive word $t$. Hence, $w_1 \cdots w_m \in \{t, \theta(t)\}^*$. Using the lcm-variant of exFW theorem, we can conclude $w \in \{t, \theta(t)\}^*$.

The same proof technique works for exLS equations $(5, \geq 5, \geq 5)$ or $(5, 4, \geq 7)$.

Proposition ([Czeizler, Czeizler, Kari, & Seki 2009])

*ExLS equations $(5, \geq 5, \geq 5)$ and $(5, 4, \geq 7)$ possess the weak defect effect.*
The following stronger statement requires results in combinatorics on words.

**Theorem ([Czeizler, Czeizler, Kari, & Seki 2009])**

*ExLS equations* \((\geq 5, \geq 3, \geq 3)\) *possess the weak defect effect.*

More useful results in combinatorics on words have been obtained in [Kari, Masson, & Seki 2009] to address some cases left open in [Czeizler, Czeizler, Kari, & Seki 2009].

**Theorem [Kari, Masson, & Seki 2009]**

*ExLS equations* \((\geq 4, \geq 3, \geq 3)\) *possess the weak defect effect.*
### Summary on the exLS equation

<table>
<thead>
<tr>
<th>$\ell$</th>
<th>$n$</th>
<th>$m$</th>
<th>weak defect</th>
<th>how to prove</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\geq 6$</td>
<td>$\geq 3$</td>
<td>$\geq 3$</td>
<td>YES</td>
<td>exFW theorem</td>
</tr>
<tr>
<td>5</td>
<td>$\geq 5$</td>
<td>$\geq 5$</td>
<td>YES</td>
<td>[Czeizler, Czeizler, Kari, &amp; Seki 2009]</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>$\geq 4$</td>
<td>YES</td>
<td>combinatorial arguments</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>$\geq 3$</td>
<td>YES</td>
<td>[Czeizler, Czeizler, Kari, &amp; Seki 2009]</td>
</tr>
<tr>
<td>4</td>
<td>$\geq 3$</td>
<td>$\geq 3$</td>
<td>YES</td>
<td>combinatorial arguments</td>
</tr>
<tr>
<td>3</td>
<td>$\geq 3$</td>
<td>$\geq 3$</td>
<td>OPEN</td>
<td>[Kari, Masson, &amp; Seki 2009]</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>NO</td>
<td>examples</td>
</tr>
</tbody>
</table>

[Czeizler, Czeizler, Kari, & Seki 2009]
Open problems I

Open problems about the exFW theorem

1. Find an optimal bound for \((p, q)\) for which \(b'(p, q)\) is not optimal.

2. Extend the problem setting further as: how long prefix do a \(\theta\)-power of \(u_1\), that of \(u_2\), \ldots, that of \(u_n\) have to share to imply that \(u_1, u_2, \ldots, u_n \in t\{t, \theta(t)\}^*\) for some word \(t\)?
Open problems about the exLS equation

1. Solve exLS equations $u_1 u_2 u_3 = v_1 \cdots v_n w_1 \cdots w_m$ with $n, m \geq 3$.

2. Find a condition under which exLS equations with $\ell = 2$ are solved positively.

3. Extend the exLS equation further as

$$u_1 \cdots u_\ell = v_{11} \cdots v_{1n_1} v_{21} \cdots v_{2n_2} \cdots v_{k1} \cdots v_{kn_k},$$

where $k \geq 2$, $\ell, n_1, \ldots, n_k \geq 1$, $u_1, \ldots, u_\ell \in \{u, \theta(t)\}$, and $v_{i1}, \ldots, v_{in_i} \in \{v_i, \theta(v_i)\}$. 

Open problems III

Other open problems

1 ExFW theorem and positive answers to exLS equations are part of weak defect theorem. Investigate and establish the weak defect theorem.
The following is a list of my publications related to the topics introduced in this talk. The list of all of my publications is available on my website:

http://www.csd.uwo.ca/~sseki


Information of my publication II


Combinatorics of words.

An extension of the Lyndon Schützenberger result to pseudoperiodic words.

On a special class of primitive words.

Uniqueness theorem for periodic functions.


References IV


Thank you for your kind attention.