Channel models for DNA Word Design

Luca Bortolussi$^1$ Andrea Sgarro$^2$

$^1$Department of Mathematics and Computer Science
University of Udine, Italy.

$^2$Department of Mathematics and Computer Science
University of Trieste, Italy.

SIMAI 2006, Baia Samuele, Ragusa, 21$^{st}$-26$^{th}$ May 2006
Outline

1. Biological Background
   - DNA computing
   - DNA Word Design

2. Preliminaries
   - A General Framework for Coding Theory

3. DNA channels
   - What is noise in DNA Word Design?
Outline

1. Biological Background
   - DNA computing
   - DNA Word Design

2. Preliminaries
   - A General Framework for Coding Theory

3. DNA channels
   - What is noise in DNA Word Design?
DNA

- DNA is a complex molecule storing all information necessary to build a living being.
- It is composed by two oriented complementary strands, that can be seen as long sequences over the alphabet \( \{a, c, g, t\} \), one the reverse complement of the other.
- Hybridization is the process leading to the formation of double helix from complementary strings. It is error prone.
DNA stores all the information needed to build a complex living being.

Since 1994, it has been used as a computational entity: DNA computing.

The basic idea of DNA computing is to encode a given problem in a set of DNA strings, let these strings interact together and “read” the output.

The main operation happening between those strings is the hybridization.

One needs good set of strings (that don’t self-hybridize): DNA word design.
DNA strings are oriented strings over $\Sigma = \{a, c, g, t\}$. Fixing the length $n$, and taking $x, y \in \Sigma^n$ we consider the following distances:

- $d_H(x, y)$, the usual *Hamming distance* (number of positions in which $x$ and $y$ differ);
- $d_{RC}(x, y) = d_H(x, y^{RC})$, the *reverse complement distance*, which is the distance between $x$ and the reverse complement of $y$.

A subset $C \subset \Sigma^n$ is a *DNA code* with threshold $D$ iff

$$\forall x \neq y \in C, \ d_H(x, y) \geq D \text{ and } d_{RC}(x, y) \geq D,$$

and also

$$d_{RC}(x, x) \geq D.$$
The problem

The question
DNA codes are constructed forcing constraints between codewords in terms of distances $d_{RC}$ and $d_H$. Have they any error correction capability?

How to
To answer this question, we have to solve an “inverse problem”: find a channel model and an error-fighting decoding mechanism that explains this code construction in terms of error correction.
Outline

1 Biological Background
   - DNA computing
   - DNA Word Design

2 Preliminaries
   - A General Framework for Coding Theory

3 DNA channels
   - What is noise in DNA Word Design?
Distinguishability

$\delta : \mathcal{A} \times \mathcal{A} \rightarrow \mathbb{R}$ is given by

$$\delta(x, y) = \min_{z \in \mathcal{B}} d(x, z) \lor d(y, z).$$

Let $\Sigma = \{a_1, \ldots, a_k\}$ be an alphabet, $\mathcal{A} = \mathcal{B} = \Sigma^n$ and $d = d_H$, the usual Hamming distance. Then $\delta_H(x, y) = \left\lfloor \frac{d_H(x, y)}{2} \right\rfloor$. 

$\mathcal{A}$ is the input space

$\mathcal{B}$ is the output space

$d : \mathcal{A} \times \mathcal{B} \rightarrow \mathbb{R}$ is a diversity measure between input and output objects.
Distinguishability II

Given a codebook $C \subseteq A$, the distinguishability of $C$ is $\delta_C = \min_{x, y \in C, x \neq y} \delta(x, y)$. The decoding mechanism is by minimum diversity.

Intuition

The more the codewords are scattered apart, the greater the errors that the codebook will “tolerate” (i.e. correct).

Theorem

For a codebook $C$, all diversities $\leq \tau$ are corrected iff $\delta_C > \tau$. 
Distinguishability III

In this case, the code has no error correction capabilities.

In this case the code can correct up to one error.
Error-Correcting Codes

Constructing Error-Correcting Codes

- Fix a error-correction capability $\tau$ for a diversity $d$.
- Consider codebooks $C$ such that $\forall x, y \in C, \delta(x, y) > \tau$
- Maximize the size of $C$.

Code constructions w.r.t. distinguishability and w.r.t diversity
DO NOT generally coincide.

Theorem

If distinguishability is a monotonic non-decreasing function of diversity, code constructions w.r.t distinguishability and w.r.t diversity coincide (This is the case for Hamming distance).
Outline

1. Biological Background
   - DNA computing
   - DNA Word Design

2. Preliminaries
   - A General Framework for Coding Theory

3. DNA channels
   - What is noise in DNA Word Design?
Where is the channel?

Target

Finding a channel and a decoding mechanism for DNA codes means, in our framework, to identify a diversity measure such that its distinguishability is a non-decreasing function of the DNA metric.

As DNA metrics, we can use \(d_{RC}\) alone or \(d_{RC} \land d_H\). Therefore, we look for two diversities \(\xi_1\) and \(\xi_2\) such that:

\[
\delta_{\xi_1} = f(d_{RC});
\]
\[
\delta_{\xi_2} = f(d_{RC} \land d_H);
\]

where \(f : \mathbb{R} \rightarrow \mathbb{R}\) is a monotonic non-decreasing function.
Does $d_{RC}$ alone work?

- Can we construct DNA codes w.r.t. reverse complement Hamming distance only?
- Is there any diversity $\xi_1$ such that its distinguishability $\delta_{\xi_1}$ is a monotonic non-decreasing function $f(d_{RC})$ of $d_{RC}$?

**Theorem**

*If $\xi_1$ satisfy condition above, then $\xi_1 \equiv 0$.***

DNA code construction w.r.t. Reverse Complement Hamming distance only are NOT justified.
The input space is $\mathcal{A} = \{x \in \Sigma^n \mid d_{RC}(x, x) \geq D\}$
The output space is $\mathcal{B} = \Sigma^n$
The diversity is $d(x, y) = \min\{d_H(x, y), d_{RC}(x, y)\}$

**Theorem**

The distinguishability for diversity $d$ is

$$\delta(x, y) = \left\lceil \frac{d(x, y)}{2} \right\rceil.$$

DNA code constructions (done w.r.t diversity $d$) have error correction capabilities given by $\delta_C$. 
The End

THANKS FOR THE ATTENTION!

QUESTIONS?