DNA codebooks and DNA channels

Luca Bortolussi\textsuperscript{1} \hspace{1cm} Andrea Sgarro\textsuperscript{2}

\textsuperscript{1}Department of Mathematics and Computer Science  
University of Udine, Italy.

\textsuperscript{2}Department of Mathematics and Computer Science  
University of Trieste, Italy.

GNCS meeting 2006, Milano, 15\textsuperscript{nd} February 2006
Outline

1. Introduction
   - Biological Background
   - DNA Word Design

2. DNA codebooks
   - Looking for Maximum Codes
   - Our Algorithm

3. DNA channels
   - A General Framework for Channel Coding
   - What is noise in DNA Word Design?
Outline

1 Introduction
   - Biological Background
   - DNA Word Design

2 DNA codebooks
   - Looking for Maximum Codes
   - Our Algorithm

3 DNA channels
   - A General Framework for Channel Coding
   - 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

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

L. Bortolussi, A. Sgarro

DNA codebooks and DNA channels
Two problems

Find methods to construct DNA codes of good quality, hopefully of maximal size.

Understand better these DNA codes from the point of view of Channel Coding Theory.
Outline

1. Introduction
   - Biological Background
   - DNA Word Design

2. DNA codebooks
   - Looking for Maximum Codes
   - Our Algorithm

3. DNA channels
   - A General Framework for Channel Coding
   - What is noise in DNA Word Design?
Looking for Maximum Codes

Construction of DNA codes

The Problem
We want to construct DNA codes of maximum size.

Approaches
- All approaches to construct DNA codes are based on Stochastic Local Search Algorithms.
- We studied the feasibility of a Branch and Bound schema.
Looking for Maximum Codes

Branch and Bound approach

Branch and Bound

- *Branch and Bound* methods tackle constrained optimization problems by constructing a search tree covering the whole space,
- Variables are *instantiated one by one*.
- Branching steps (on different values of a variable) are followed by propagation steps (narrowing domains of non-instantiated variables).
- The tree is visited by means of *backtracking*, and areas that are guaranteed not to improve current solution are pruned.
Our Algorithm

Branch and Bound for DNA Word Design

- One variable for each word in the codebook.
- **Propagation** is done by keeping track of the set of strings satisfying all distance constraints from the words currently in the codebook (*global consistency*).
- **Pruning** is done when it is not possible to find a bigger code of the best one found so far.
Our Algorithm

Symbolic Representation of Domains

- The dimension of the domain of a single variable is *exponential in the length of the words*.
- We use a symbolic representation using a variant of Ordered Binary Decision Diagrams (OBDD), capable of representing boolean functions.
- Propagation is performed at OBDD level (OBDD intersection), and it is globally consistent.
Results

- Currently we are not able to explore the whole search space, due to its gigantic dimension.
- The algorithm, however, finds good solutions quickly.

Future Work

- Putting more constraints to the problem, in order to reduce the search space.
- Find more powerful bounds, to prune more regions of the tree.
- Introduce stochastic ingredients (randomized backtracking)
Outline

1. Introduction
   - Biological Background
   - DNA Word Design

2. DNA codebooks
   - Looking for Maximum Codes
   - Our Algorithm

3. DNA channels
   - A General Framework for Channel Coding
   - What is noise in DNA Word Design?
A General Framework for Channel Coding

Distinguishability

- $\mathcal{A}$ is the input space
- $\mathcal{B}$ is the output space
- $d : \mathcal{A} \times \mathcal{B} \to \mathbb{R}$ is a diversity.
- $\delta : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$ is the distinguishability, given by
  \[ \delta(x, y) = \min_{z \in \mathcal{B}} d(x, z) \lor d(y, z). \]
- Given a codebook $\mathcal{C} \subseteq \mathcal{A}$, the distinguishability of $\mathcal{C}$ is
  \[ \delta_{\mathcal{C}} = \min_{x, y \in \mathcal{C}, x \neq y} \delta(x, y). \]

Theorem

All errors of diversity $\leq \tau$ are corrected iff $\delta_{\mathcal{C}} > \tau$. 
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.
What is noise in DNA Word Design?

Where is the channel?

Target

Find a channel and a decoding mechanism to justify DNA codes in the light of error-correction.

Consider the following mechanism of information transmission:

- (noisy) channel $\rightarrow$ DNA duplication (amplification)
- decoding $\rightarrow$ hybridization
Introduction

DNA codebooks

DNA channels

What is noise in DNA Word Design?

Does $d_{RC}$ alone works?

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

Theorem

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

DNA code construction w.r.t. Reverse Complement Hamming distance only are NOT justified.
What is noise in DNA Word Design?

The Right Diversity

The input space is \( \mathcal{A} = \{ x \mid d_{RC}(x, x) \geq D \} \)

The output space is \( \mathcal{B} = \{ a, c, g, t \}^n \)

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 \( \delta_C \).
What is noise in DNA Word Design?

THANKS FOR THE ATTENTION!

QUESTIONS?