Channel Models for DNA Word Design

Luca Bortolussi
Universita' di Udine
Via delle Scienze 206, 33100 Udine
luca.bortolussi@dimi.uniud.it

Andrea Sgarro
Universita' di Trieste
Via Valerio,12/b, 34127 Trieste
sgarro@units.it

Introduction. We deal with DNA word design problem, i.e. the construction of codes of DNA strings under some biological combinatorial constraints. The point of view taken here is that of channel coding theory, meaning that we try to explain DNA coding by identifying a suitable channel and a suitable decoding mechanism. The coding theoretic framework used is based on the rather general concept of distinguishability, as developed in [4]. In the following, we fist give a brief overview of DNA word design and of the distinguishability framework, and then we focus on the definition and the analysis of DNA channels. A provably equivalent point of view, based on possibilistic channel, is underpinned in [1].

A short reminder on DNA word design. In the last ten years, a new computational paradigm emerged from a very uncommon place, i.e. wet labs of biologists. The fact that DNA contains all the basic information necessary to build very complex living organisms convinced Adlemann that it could also be used as a computational entity. In 1994 he proposed a computational model based on very simple manipulations of DNA that can be performed in a wet lab. This model is Turing-complete and bases its power on the massive parallelism achievable by using DNA. Moreover, one of the basic operations performed is the hybridization of complementary DNA strings. Specifically, DNA strings are oriented strings over the alphabet $\Sigma = \{a, c, g, t\}$, where *a*-*t* and *c*-*g* are complementary letters. Two such strings are said to be complementary if they have the same length and if one can be generated by reversing the other and complementing each of its letters. Physically, complementary DNA strings can hybridize, i.e. they can attach one to the other, forming the famous double helix. Actually, hybridization can occur also between strings that are not perfect complements, but close to it. In DNA computations, data is coded by short strings of DNA in such a way that hybridizations occurring determine the output of the "algorithm". Therefore, one of the main concerns is to avoid that "spurious" hybridizations occur, leading straight to the so-called *DNA word design* problem.

DNA word design (cf. [2]) consists of identifying maximal sets of DNA strings of a given length, called *DNA codes*, satisfying some constraints, usually related to distances between codewords.

A Framework for Channel Coding. We shortly revise the material of [4] (which is actually rather more general). One considers n-length sequences over the alphabet $\mathcal{A} = \{a_1, a_2, \ldots, a_K\}$. To each ordered couple of sequences x, y a non-negative number $d(x, y)$ is assigned called their *diversity*. One chooses a subset $\mathcal{C} \subset \mathcal{A}^n$ called the *code*, whose sequences are called *codewords*. One sends one such sequence through a noisy channel. The received n-length sequence $z \in \mathcal{A}^n$ is decoded by *minimum diversity*, i.e. the decoder gives back a codeword c such that $d(c, z)$ be minimum; the underlying assumption is that the higher the diversity, the less "likely" it is to occur (in a very broad sense of the word "likely", cf [4]). The *distinguishability* between two sequences is defined as:

$$\delta(x, y) \;=\; \min_z \, d(x, z) \vee d(y, z)$$

(By the way, the distinguishability δ is always symmetric, even if the diversity d is not). The *minimum distinguishability* $\delta_{\mathcal{C}}$ of the code $\mathcal{C}$ is the minimum distinguishability between any two distinct codewords. The operational meaning of $\delta_{\mathcal{C}}$ is given by the following *reliability criterion*:

Theorem 1 *The minimum distinguishability $\delta_{\mathcal{C}}$ is the lowest diversity which is not always corrected when decoding by minimum diversity; diversities $< \delta_{\mathcal{C}}$ are always corrected.*

The classical optimization problem of channel coding is maximizing its size (which is the same as maximizing its transmission rate) subject to a specified reliability constraint. In the case when the diversity is *Hamming distance* the distinguishability is soon found to be [4]:

$$\delta_H(x, y) \;=\; \left\lceil \frac{d_H(x, y)}{2} \right\rceil$$

Since the distinguishability is a non-decreasing function f of the Hamming distance, one can construct reliable codes with respect to reliability constraints $d_{\mathcal{C}} \geq \lambda$ expressed in terms of the minimum Hamming distance between distinct codewords $d_{\mathcal{C}}$ (as one usually does in the Hamming case), rather than constraints on distinguishability $\delta_{\mathcal{C}} \geq \tau$, as one may do in in full generality; cf. again [4]. Actually, the constraints $\delta_{\mathcal{C}} \geq \tau$ and $d_{\mathcal{C}} \geq f^{-1}(\tau)$ are equivalent, with $f^{-1}(\tau)$ equal to the smallest diversity λ for which $f(\lambda) = \tau$.

The inverse problem of channel noise in DNA word design. DNA word design is an "odd" form of coding used in molecular computation, where, based on biological facts, one exhibits maximum-size code constructions relative to constraints of the form $\xi(x, y) \geq \lambda$ for a suitable DNA string distance ξ. An information-theoretic problem arises: what is the nature of the biological channel one is implicitly envisaging, or, equivalently: what sort of biological "noise" are we fighting against when we use these code constructions? Thinking of the above arguments, we can re-formulate the question as follows: can $\xi(x, y)$ be interpreted as a "pseudo-distinguishability", i.e.: can one exhibit a distortion measure $d(x, y)$ between inputs and outputs such that the corresponding distinguishability function $\delta(x, y)$ is a non-trivial and non-decreasing function of $\xi(x, y)$? We shall discuss two types of code constructions found in the literature: the answer will be positive in one case, which is better justified also from the biological point of view, and negative in the other.

We shall deal only with two DNA "distortions"[1], which however are very representative, the *reverse Hamming distance* and a variation thereof:

$$d_R(x, y) \quad \text{and} \quad d_{H \wedge R}(x, y) \; = \; d_H(x, y) \wedge d_R(x, y)$$

Here $d_H(x, y)$ is the usual Hamming distance, while the reverse Hamming distance is $d_R(x, y) = d_H(x, y^*)$, with y^* mirror image of y. In practice, in the case of d_R, codewords in a good code should have a large reverse Hamming distance, while they should have *both* a large Hamming distance *and* a large reverse Hamming distance in the case of $d_{H \wedge R}$. We recall that $d_{H \wedge R}(x, y)$ is a *pseudometric*; one has $d_{H \wedge R}(x, y) = 0$ when $x = y$ or when x and y are mirror images of each other. Nothing so tame happens in the case of d_R, which violates the triangle inequality.

Below we shall try to "explain" the corresponding DNA code constructions by exhibiting a suitable possibilistic noisy channel and a suitable noise-fighting decoder. To achieve this, let us begin by the "friendlier case", and let us compute the distinguishability $\delta_{H \wedge R}$ corresponding to the string distance $d_{H \wedge R}$ taken as the distortion between inputs and outputs. We decode the output z by minimum distortion, and so we are implicitly assuming that it is "unlikely" (i.e. possible only to a small degree) that z has both a large Hamming distance and a large reverse Hamming distance from the codeword c actually sent over the channel.

[1]DNA complementarity has been forgotten out of simplicity, since it does not really change the problem, but makes notations and formulations heavier; cf. also [3]. In addition, the constraint about self-hybridization can be easily dealt with by restricting the input space to those sequences satisfying it.

Theorem 2 *Decode the output z by minimizing $d_{H\wedge R}(c,z)$, $c \in \mathcal{C}$; the corresponding distinguishability function is:* $\delta_{H\wedge R}(x,y) = \left\lceil \frac{d_{H\wedge R}(x,y)}{2} \right\rceil$.

This is exactly the same situation as found with usual Hamming distances and the codes of algebraic coding. In practice, this means that a channel based on the distortion $d_{H\wedge R}(x,z)$ and the corresponding "noise" quite adequately "explain" the code constructions based on checking the pseudometric $d_{H\wedge R}$, as are those found in the literature.

Now, let us think of a DNA word design construction where one controls *only* the minimum reverse Hamming distance between codewords. The situation is less friendly, because if we decide to decode by minimum reverse Hamming distance, the corresponding distinguishability function turns out to be a non-decreasing function of the *usual* Hamming distance, and *not* of the reverse Hamming distance, as a simple computation shows. In other words, against this sort of noise one would need the *usual* codes of coding theory, and *not* the codes of DNA word design which we are trying to "explain". So, the following problem is relevant:
Problem: Exhibit a non-trivial distortion $\eta(x,z)$ with distinguishability function $\Xi(x,y)$ such as to be a non-decreasing function of the reverse Hamming distance.
Unfortunately this problem has a negative answer (cf. [1]), meaning that, at least within the distinguishability framework, ample as it may be, code constructions based on checking reverse Hamming distances have *no* counterparts in terms of noisy channels and channel decoders; *no* distortion $\eta(x,z)$ exists which would adequately support those constructions.

REFERENCES

1. L. Bortolussi and A. Sgarro, Possibilistic channels for DNA word design. *To be presented at SMPS 2006*, Bristol, September 2006.

2. A. Condon and A. Brenneman, Strand design for bio-molecular computation. *Theoretical Computer Science*, 287(1):3958, 2002. 5.

3. A. Condon, R.M. Corn, and A. Marathe, On combinatorial DNA word design. *Journal of Computational Biology*, 8(3):201220, November 2001.

4. A. Sgarro and L. Bortolussi, Codeword Distinguishability in Minimum Diversity Decoding. To appear in *J. of Discrete Math. Sc. and Cryptography.*