Noise of DNA word design is not stochastic

Luca Bortolussi[a,b], Andrea Sgarro[a,b]

[a] *DMI, University of Trieste, Italy*
[b] *CBM, Area Science Park, Trieste, Italy*

Abstract

Usual error-correcting codes foil the stochastic noise of symmetric channels. In this paper we show that nothing like this holds for codes of DNA word-design, as used for error-correction in biological computation; however, such codes do foil a different form of non-stochastic noise based on similarity/dissimilarity of structure between input and output strings, as definable within the framework of possibilistic information theory, which is a belated offshot of Shannon's 1956 zero-error information theory.

Keywords: channel noise, DNA word design, possibilistic information theory, zero-error information theory

1. Introduction, with a reminder on distance-based codes

Think of strings of length n, e.g. binary strings, or quaternary strings as are DNA strings, and assume that a "distance" is given which measures the dissimilarity between each two of those strings. Error-correcting codes (cf. e.g. [10, 13]) are constructed by fixing an integer threshold τ and taking as many strings as possible, with the constraint that each two strings (*codewords*) must be at distance $\geq \tau$. While *standard codes*, as are those of algebraic coding theory, are constructed by constraining the usual Hamming distance $d_H(x, y)$, in DNA word design[1] one resorts to "odd" string distances of biological interest, as is $d_B(x, y)$ to be defined in (1).

Email addresses: luca@dmi.units.it (Luca Bortolussi), sgarro@units.it (Andrea Sgarro)

[1] The codes of DNA word design are error-correcting codes used in biological computation, whose codewords are DNA strings; cf. e.g. [5].

Below, we shall make the following convenient[2] assumptions, which are amply met by codes of interest:

- the space of input strings is (possibly weakly) included in the space of output strings, and so an input string may be received unscathed
- the distance is a pseudo-metric on the output space (in practice: a metric after "gluing" together strings which are at zero distance); without real restriction[3] we assume that the distances are consecutive integers.

An example, is the biological distance, used in DNA word-design when the alphabet is quaternary ($\wedge$ denotes minimum, x and y are strings of length n):

$$\begin{aligned} d_B(x,y) &= d_H(x,y) \wedge d_H(x,y^*) \\ &= d_H(x,y) \wedge d_H(x,y^*) \wedge d_H(x^*,y) \wedge d_H(x^*,y^*) \end{aligned} \tag{1}$$

here y^* is the *reverse complement* (complemented mirror-image) of y; in a way, a string is at the same time itself and its reverse complement. We recall that a *complementation* is a permutation of the letter alphabet with cycles all of size 2, e.g. $(A,T),(C,G)$ in the DNA case on the alphabet $\{A,C,G,T\}$. E.g. $d_B(ATTTA,GAGAT) = d_H(ATTTA,GAGAT) \wedge d_H(ATTTA,ATCTC) = 5 \wedge 2 = 2$. The metric distance d_B can be defined on any alphabet of even size endowed with a complement: it is precisely distance-codes based on such a d_B that are dealt with in this paper.

In the case of standard codes based on Hamming distances, sometimes a stochastic matrix is put forward to explain why these codes can be used to foil the corresponding stochastic noise: the matrix entries are the transition probabilities from inputs to outputs which describe a memoryless stationary *symmetric channel*, cf. e.g. [10, 13]. Decoding by maximum likelihood or by minimum Hamming distance turns out to be the same, since the stochastic matrix is *equivalent* to the distance matrix, in the sense that it is a *strictly*

[2]Actually, this framework might be readily enlarged, e.g. to include non-metric distances as is the "bad" distance hinted at in Section 3; cf. [11], [12]. To avoid trivial specifications, we assume that the letter-alphabet has size at least 2, and that all our matrices are not constant and have at least two rows.

[3]As will become apparent, only order matters, rather than actual values, and so, basically, our tools are defined only up to monotone transformations.

monotone decreasing function thereof (a transition probability increases if and only if the corresponding string distance decreases). The stochastic matrix describing the symmetric channel, which is of course equivalent to the matrix of Hamming-distances, is usually passed off as an exhaustive description of the relevant *noisy channel* over which codewords are sent: in a way, the stochastic matrix *is* the channel noise.

In Section 2 we readily show that *no* equivalent stochastic matrix exists for the error-correcting codes of DNA word design based on constraining d_B. However, in Section 3 we show that this does not imply that one has to give up the notion of channel noise as required in a full-fledged Shannon-theoretic approach to coding, provided one is ready to relinquish stochastic noise in favour of *possibilistic* noise, a form of noise, as explained below, Section 3, based on the (dis)similarity of the strings involved. A short comment follows in Section 4 on new approaches to information theory which are based on structure and patterns rather than randomness.

2. The negative theorem

Equivalence as specified above is basically a problem of linear programming, which it has been dealt with in [2]. In the sequel the following obvious necessary condition for equivalence will do, whose proof is straightforward (a vector strictly dominates another when in each position it has a larger entry, and at least one inequality is strict):

Lemma 1. *Take two rows of a distance matrix and re-arrange their respective entries in non-decreasing order. If one of them strictly dominates the other, the two rows do not have any stochastic equivalent.*

In the case of DNA word design, the input space is formed only by strings constrained to have a *degree of self-hybridisation* $d_H(x, x^*)$ larger than a specified threshold, and so is properly included in the output space of all n-length strings.

Lemma 2. *The maximal possible value achieved by the degree of self-hybridisation is the string-length n.*

Proof: For n even, take any x made up of two halves which are the mirror image of each other; for n odd proceed in the same way, just insert any letter at position $(n+1)/2$.

Theorem 1. *Take $d_B(x, y)$ as in (1), $n \geq 2$. There are two strings x and y with no equivalent couple of stochastic rows; x and y can be constrained to have degree of self-hybridisation $\geq n - 2$.*

Proof: Say (a, t) is a cycle of the complementation. For $n = 2$ take $x = aa$ and $y = at$. One has $d_B(x, z) = 0, 1, 1, 0$ and $d_B(y, z) = d_H(y, z) = 1, 0, 2, 1$ for $z = aa, at, ta, tt$. So, for $n = 2$ one may resort to Lemma 1. For $n > 2$ just take $x = aua$ and $y = aut$, where u is any infix-sequence of maximal degree of self-hybridisation $n - 2$ (use Lemma 2).

In practice, all of this means that we have to give up the Shannon-theoretic notion of channel noise in the case of DNA word design, at least if we insist that channel noise be constrained to have a stochastic nature. A few additional comments are found in the provisos at the end of Section 3.

3. Possibilistic vs. probabilistic noise

Rather than giving up the Shannon-theoretic notion of channel noise in the DNA case, we shall resort to an alternative *non-stochastic* approach to coding, namely to a generalisation of Shannon's zero-error information theory, dubbed *possibilistic information theory* for reasons explained below. The possibilistic approach is based on *similarities/dissimilarities* of structure between input and output strings, rather than transition probabilities.

The basic elements of possibilistic error-correcting codes follow below. Preliminarily, observe that a distance d as in Section 1 serves two distinct purposes: *i)* it is used in decoding, where one minimises *input-output* distances, and *ii)* it is also used in the construction of the codebook when one selects *input* codewords which are distant enough from each other. At least for the moment being, we will have to keep apart these two roles, and so, to avoid confusion, we shall use the term *dissimilarity* for role *i)* and *distinguishability* for role *ii)*. When discussing below the *inverse problem*, we shall make it clear when and how these notions fall back to one, as is often the case in practice. We use the letter d to denote dissimilarities, since these and distances as in Section 1 are quite akin; the notational ambiguity will turn out to be venial.

- *The noisy channel*, which is entirely specified by a *dissimilarity matrix*, whose entries $d(x, z)$ are *dissimilarities* between *input strings* x (row

headings) and *output strings* z (column headings). We shall safely assume that any string x can be received unscathed over the channel, and that the dissimilarity between x and itself is zero; to no real restriction, we shall assume that dissimilarities are consecutive integers.

- A *codebook* $\mathcal{C}$, or simply a *code*, i.e. a non-void subset of input strings called *codewords*.

- *The distinguishability* $\delta(x, y)$ between inputs x and y:

$$\delta(x, y) \doteq \min_{z} \max\left[d(x, z), d(y, z)\right] \tag{2}$$

  (the minimum is taken over all outputs z) and the *the minimum distinguishability* of the codebook $\mathcal{C}$

$$\delta(\mathcal{C}) \doteq \min_{x \neq y,\ x,y \in \mathcal{C}} \delta(x, y) \tag{3}$$

  The latter minimum singles out the most "critical" codeword couple in $\mathcal{C}$; as done in combinatorics, any minimising z as in (2) will be called a *centre* of the set $\{x, y\}$.

- *Optimal codes*: once the integer threshold ρ is chosen, construct maximum-size codes with guaranteed minimum distinguishability ρ, i.e. with $\delta(\mathcal{C}) \geq \rho$.
  Equivalently, one might resort to the complementary notions of *similarities* and *confusabilities*, rather than dissimilarities and distinguishabilities, cf. [8], [12]. One soon proves the following reliability criterion, cf. [12]:

- **Reliability criterion.** *Once the output string z is received, decode to an input codeword*[4] *x which minimises the dissimilarity $d(x, z)$ between input and output: if the input string x was such that $d(x, z) < \delta(\mathcal{C})$ the decoding is successful; instead, there is at least an input/output couple with $d(x, z) = \delta(\mathcal{C})$ which brings about a decoding error.*

The criterion will certainly look familiar to coding theorists, but we stress that there is an important novelty with respect to distance-based codes as in

[4] In case of ties, one might decide to declare a *detected error*, instead; cf. Addendum 5.

Section 1: codebooks are now constructed *by checking the distinguishability* $\delta(x, y)$ between inputs, the latter having been *computed from the dissimilarities* $d(x, z)$ between inputs and outputs. If one takes binary dissimilarities, $0 =$ similar, $1 =$ dissimilar, one re-obtains Shannon's zero-error codes as in [8], where, as well-known, one has to carefully keep apart the two roles of dissimilarities and distinguishabilities (analogously: of similarities and *confusabilities*). In the original Shannon's probabilistic approach an input is similar to an output when the corresponding transition probability is positive, however small it may be.

When one wants to accommodate distance-based codes as in Section 1 into the possibilistic framework, the following *inverse problem* arises, since one has to make sure that constraining distances or constraining distinguishabilities amounts in practice to the same thing:

Inverse problem: *When a codebook has been constructed using a generic distance d as in Section 1, can this be interpreted to be a dissimilarity as in this Section, i.e. can one provide a possibilistic noisy channel (a dissimilarity matrix between inputs and outputs) such that the corresponding distinguishability function between inputs, which serves purpose ii),* is *that distance d, or at least is a monotonic function thereof?*

Solving the inverse problem amounts to understand whether one is dealing with error-correcting codes such as to comply with the Reliability criterion; if the solution is negative, the codebook is just a nice combinatorial object.

In DNA word design one decodes by minimising the distance $d_B(x, y)$; taking this as our dissimilarity, the distinguishability $\delta_B(x, y)$ is soon computed to be $\lceil \frac{1}{2} d_B(x, y) \rceil$, a friendly expression which is analogous to the one for Hamming distances, $\delta_H(x, y) = \lceil \frac{1}{2} d_H(x, y) \rceil$, and which largely trivialises the distinction between distance, dissimilarity and distinguishability in standard coding and in DNA coding, at least *a posteriori*, after having computed explicitly δ. By constraining the distance d_B as in Section 1 or the distinguishability δ_B as in the possibilistic frame one obtains the *same* family of error-correcting codebooks when τ is odd (just take $\rho = \lceil \frac{\tau}{2} \rceil$). Now, codebooks with $d_H(\mathcal{C})$ or $d_B(\mathcal{C}) \geq \tau$ with τ even are *not* needed if one insists on *hard* decoding, as done so far, when giving up decoding in critical situations is prohibited: the thresholds τ and $\tau + 1$ give optimal codes with the same error-correction capabilities, but the latter constraint on size is looser. However, no distance-codebooks get lost if one takes into account also *soft*

decoding, cf. Addendum 5.

In general, the distinguishability δ derived from an integer pseudo-metric dissimilarity d over the output space is readily shown (cf. [12]) to belong[5] to the interval:

$$\left\lceil \frac{d(x,y)}{2} \right\rceil \ \leq \ \delta(x,y) \ \leq \ d(x,y) \tag{4}$$

We stress once more that having a generic "distance" between strings as in Section 1 is *not* enough to have a corresponding possibilistic noisy channel, since the inverse problem might have no solution. In [1] we have taken into account the "non-metric distance" $d_C(x,y) = d_H(x,y^*)$ for $x \neq y$, else 0, since it had been used[6] in the literature on DNA word design: the answer to the inverse problem is *no* in the case of these artificial codebooks based on d_C, which are so devoid of error-correcting capabilities. To avoid misunderstandings, this is *not* due to the non-metric nature of d_C: already in Shannon's zero-error theory one has examples galore of spotless possibilistic channels based on distances/dissimilarities which are definitely unruly from a metric point of view, and crassly violate the triangle inequality, cf. the case of Shannon's (and Lovász') pentagon [8].

While the situation with the usual Hamming distance d_H and the artificial non-metric distance d_C is clear-cut (d_H works, d_C does not), the situation with our d_B is intriguing. The corresponding distance-based codes, as dealt with in the literature on DNA word design, do not appear to admit of stochastic noisy channels such as to support their use, but they perfectly fit into the more "easy-going" possibilistic framework, and so foil possibilistic noise.

MV-logics versus probabilities.. Why choose the attribute "possibilistic", rather than, say, "[dis]similarity-based" or "multi-step"? *Possibility theory* is a form of multi-valued (MV) logic, based on specifying degrees of possibility of events, and it is considered an adequate way of dealing with incomplete knowledge in several situations when the usual probabilistic tools appear to miss the mark. The reader is referred to standard texts on possibility theory, e.g. [7], while he is referred to [11] *for possibilistic information theory.* Even if arisen only as a formal game where probabilistic tools are replaced by the

[5]The upper bound holds uniformly if the metric distance is a *ultrametric*, cf. [12]. Cf. Section 4 for an unruly case.

[6]Even if, fair to say, for mere reasons of combinatorial bounding, since distance-codes for d_C are subsets of proper DNA distance-codes based on d_B.

corresponding ones as available in possibility theory, possibilistic information theory has proven to properly accommodate all the relevant Shannon-theoretic notions of source and channel coding, inclusive of (possibilistic) source entropy and (possibilistic) channel capacity; reliability criteria as ours above may be re-phrased in terms of *error possibilities*. Here we shall simply recall that stochastic matrices of transition (conditional) probabilities from an input string to an output string are replaced by *possibility matrices* of *transition possibilities*. While in a stochastic matrix the sum of each row is equal to 1, in a possibility transition matrix it is the *maximum* in each row which is constrained to be 1. In practice, if one starts from dissimilarity matrices as ours (in each row there is at least a 0), one soon constructs an equivalent possibility matrix of normalised similarities, which specifies the "possibilistic noise" of the channel. In our case one may resort to the transition possibilities from input x to output z (we are mimicking the notation for conditional probabilities):

$$\text{Poss}\{z|x\} \;=\; 1 \;-\; \frac{1}{n} d_B(x,z) \tag{5}$$

The "degree of possibility" (5) specifies that the possibility of the transition decreases as long as the dissimilarity between input and output (between their corresponding "patterns") increases.

Possibilistic information theory turns out to be just a multi-step generalisation of Shannon's zero-error information theory, for which cf. [8]; the latter admits of two steps only, 0 = the transition is impossible, 1 = the transition is possible, without intermediate degrees of possibility. All the relevant notions of the enlarged theory, from codeword distinguishability (or, complementarily, codeword *confusability*) to possibilistic capacity (a multi-step generalisation of zero-error capacity) go back to Shannon; in Shannon's two-step case the distinguishability between two inputs is 0 if and only if there is a common output accessible by both of them, else is 1: use formula (2). Take e.g. the famous Shannon-Lovász pentagon [8] for $n = 1$: inputs and outputs are the vertices of a pentagon, transition possibilities between distinct vertices are 1 or 0 according whether they are adjacent or not; the corresponding dissimilarities are the 1-complements of the transition possibilities; distinguishabilities are equal to dissimilarities, save that they are 0 when two non-adjacent vertices have a third vertex adjacent to both of them. In this time-honoured example both dissimilarities and distinguishabilities (the two "competitors" for the role of distances as in Section 1) sorely violate the

triangle inequality.

A few provisos.. Going back to Section 1, some might object that matrix equivalence is too strict a requirement: actually, decoding by maximum likelihood or minimum distance (d_H or d_B, respectively) depends on each single column of the stochastic matrix, or of the distance matrix, while comparisons between entries in two distinct columns are not needed for decoding. Even if the mathematical problem of understanding when such "weakly" equivalent matrices exist is of interest, and might add further light to the problem of stochastic noise, we deem that a mere column-wise equivalence would be rather artificial, since it would not be robust with respect to alternative decoding rules which might be to the point e.g. when the channel output is so damaged, so "poorly" observable, that one has to deal with several observable outputs at the same time. Instead, fully equivalent matrices always return the same answer to each possible decoding rule which depends only on the relative ordering between distances/dissimilarities.
Fair to say, already in the usual Hamming case what really matters is not transition probabilities as such (who ever cares to estimate them?), but only *their strict monotonic dependence on Hamming distances*; the symmetric channel just serves to show that *in principle* a stochastic model for standard Hamming-distance codes does exist.
Other distances, which are more complex and perhaps biologically more significant, have been used in the code constructions of DNA word design; however, d_B as here is complex enough, due to (complemented) mirror-imaging as in definition (1), to exhibit strong domination as required to apply Lemma 1.

4. Conclusion

The expression for dissimilarities is not always so user-friendly as is the case for d_H and d_B: cf. [6] for the case of an important string metric, called *Spearman footrule* or *rank distance*, where the distinguishability takes its values in the whole interval (4), inclusive of its two extremes. This shows that keeping apart the roles of distinguishabilities and distances/dissimilarities may be of paramount importance also outside Shannon's zero-error theory: the multi-step possibilistic theory proves to be a very general Shannon-theoretic frame for dealing with usual and unusual forms of coding.

Quantification of structural information *is one of the three great challenges for half-century-old computer science*, as was pointed out by Fr.P. Brooks jr. in [3]. Possibilistic information theory, both source coding and channel coding, based as it is on structures and patterns rather than probabilities, which might be hard or impossible to estimate, or even meaningless, may prove to be a small step towards this objective. The possibility degree (5) simply expresses the fact that a transition is *easier to occur* when the input and the output string are similar according to a criterion of structural similarity considered to be adequate for DNA strings. The temptation to replace "easier to occur" or "having a higher degree of possibility" by "more likely" or even by "more probable" is strong, but then one would fall back into the shackles of probability theory.

5. An addendum on detected errors

In this addendum, the decoder will declare a *detected error* in case of ties (*soft decoding*); up to now, we had been dealing with *un*detected errors only (*hard decoding*). Dealing with detected errors may be rather awkward in Shannon's zero-error theory (which actually does not mention error-detection), and so it is *a fortiori* in the multi-step possibilistic theory, at least in the absence of suitable metric assumptions.

Say $\delta(\mathcal{C})$ is the minimum distinguishability (3) of codebook $\mathcal{C}$ and go back to the Reliability criterion of Section 3: if one adopts soft decoding, nothing new happens with respect to hard decoding if the dissimilarity $d(x, z)$ of the observed output z from the actual input x is $< \delta(\mathcal{C})$. Now, assume that the codebook $\mathcal{C}$ has the following property:
• whatever the codewords x and y with "critical" distinguishability $\delta(x, y) = \delta(\mathcal{C})$ and whatever their centre z as in (2) one has $d(x, z) = d(y, z)[= \delta(\mathcal{C})]$
If it is so, whenever $d(x, z) = \delta(\mathcal{C})$ the soft decoder declares a detected error.

In a metric, or pseudo-metric, space the triangular inequality soon implies the following property:

Lemma 3. *Given x and y in a pseudo-metric space, if a centre z as in (2) is equidistant from x and y, so is any other centre.*

In the topology based on the pseudo-metric d_B a convenient property holds, which is shared with Hamming distances d_H:

Proposition 1. *Given x and y, their centres z are equidistant from x and y if and only if $d_B(x, y)$ is an even integer.*

Proof: For $d_B(x, y)$ even, assume without real restriction $d_B(x, y) = d_H(x, y)$ and take z centre of x and y with respect to d_H: $d_H(x, y)$ being even, z has the same Hamming distance $\frac{1}{2}d_H(x, y)$ from both x and y. Were it not $d_B(x, z) = d_H(x, z)$ and $d_B(z, y) = d_H(z, y)$ one would violate the triangular inequality for d_B; recall that $d_B \leq d_H$ and that $d_B(x, y) = d_H(x, y)$. Proceed in a similar way for $d_B(x, y)$ odd to find a centre z which is not equidistant. Use lemma 3.

In practice,[7] when τ is an even integer, soft decoding allows to detect errors of "weight" $d(x, z) = \tau/2$ by constraining the minimum distance d_B between codewords to be $\leq \tau$, precisely as happens in the standard Hamming case. This implies that no codebooks go lost, as happens if one insists on hard decoding, cf. above footnote 5.

[1] L. Bortolussi, A. Sgarro. Possibilistic channels for DNA word design. in *Soft Methods for Integrated Uncertainty Modelling*, ed. by J. Lawry et alt., Advances in Soft Computing, Springer Verlag (2006), pp.327-335 .

[2] L. Bortolussi, A. Sgarro. A criterion for the stochasticity of matrices with specified order relations. *Rend. Ist. Mat. Univ. Trieste*, Vol. XL, pp. 55-64 (2009).

[3] Fr.P. Brooks jr.˙ Three great challanges for half-century-old Computer Science. *Jrnl of the ACM*, Vol.50, No.1, Jan. 2003, pp. 25-26.

[4] A. Condon and A. Brenneman. Strand design for bio-molecular computation. *Theoretical Computer Science*, 287(1):39–58, 2002.

[5] A. Condon, R.M. Corn, and A. Marathe. On combinatorial dna word design. *Journal of Computational Biology*, 8(3):201–220, November 2001.

[7] One might think of a modified distinguishability function for error detection; we do not pursue this point, since it would be rather artificial in a general (non-metric) possibilistic frame, where one can easily exhibit couples of inputs whose centres sometimes are equidistant, sometimes are not.

[6] L.P. Dinu, A. Sgarro. Codeword distinguishability vs. codeword distance: the case of rank coding. *preliminary version available at* `http://www.dmi.units.it/~sgarro/rankCODES.pdf`.

[7] D. Dubois and H. Prade, eds. *Fundamentals of Fuzzy Sets*. Kluwer, 2000.

[8] J. Körner, A. Orlitsky. Zero-error information theory. *IEEE Trans. Inform. Th.*, 44 (6) pp. 2207-2229 (1998).

[9] G. Mauri and C. Ferretti. Word design for molecular computing: A survey. In *Proceedings of 9th Int. Workshop on DNA Based Computers, DNA 2003*, pages 37–46, 2003.

[10] R.M. Roth. *Introduction to Coding Theory*. ambridge University Press, Cambridge, UK, 2006.

[11] A. Sgarro. Possibilistic information theory: a coding-theoretic approach. *Fuzzy Sets and Systems*, 132-1, 11–32, 2002.

[12] A. Sgarro and L. Bortolussi. Codeword distinguishability in minimum diversity decoding. *J. of Discrete Mathematical Sciences & Cryptography*, (2006) Vol.9, N.3, pp. 487-502.

[13] J. van Lint. *Introduction to Coding Theory*. Springer Verlag, Berlin, 1999.