Nanopore DNA sequencing channel modeling

Laura Conde-Canencia, Lara Dolecek

Article accepté et présenté à la conférence IEEE SiPS en Octobre 2018

Nanopore DNA sequencing channel modeling

Laura Conde-Canencia
Lab-STICC, CNRS UMR 6285
Université Bretagne-Sud
Lorient, France
laura.conde-canencia@univ-ubs.fr

Lara Dolecek
ECE Department
University of California
Los Angeles, USA
dolecek@ee.ucla.edu

***Abstract*—In this paper we first present a simplified DNA storage chain (partially inspired by the classical digital communication principles) and review the last advances on error-correcting codes for DNA storage. We then introduce a novel DNA channel model for a new generation of nanopore sequencers. For this, we analyze the reported experimental results to define and characterize an original DNA sequencing channel model. This contribution opens new directions in the design of efficient error-correcting codes to improve the performance of DNA-based data storage.**

I. Introduction

We have entered an era where the explosion of data is a reality –ranging from engineering and science to business and biology. Ninety percent of the data in the world today has been created in the last two years alone and our current output of data is roughly 2.5 quintillion bytes a day [1]. Moreover, with a world that becomes more connected through an ever-increasing number of digital devices, this trend can only grow over the coming years. Consequently, novel challenges in data storage technologies continue to arise.

Data archives (or archival storage) serve as a way of reducing primary storage consumption and related costs. This kind of storage protects older information that is not needed for everyday operations but may have to be accessed occasionally. Devices for archival storage then require a large capacity and excellent durability. These features are difficult to find in many modern memories, leading to the study of alternative emerging technologies based on molecular storage.

DNA storage is much more compact than any existing storage device due to the data density of the DNA. Moreover, the capability for longevity and resistance to obsolescence of DNA is undeniable: DNA is a universal and fundamental data storage mechanism in biology. As claimed by the authors in [2]: with information retention times that range from thousands to millions of years, volumetric density 103 times greater than flash memory and energy of operation 108 times less, DNA used as a memory-storage material in nucleic acid memory products promises a viable and compelling alternative to electronic memory. In other terms, the extremely high storage capacity is such that a human cell, with a mass of roughly 3 picograms, hosts DNA encoding 6.4 GB of information.

Thanks to the lastest advances of biotechnological systems for DNA synthesis (writing data as bases/nucleotides) and sequencing (reading or retrieving the data), DNA-based storage is becoming feasible. Two different teams outlined practical architectures for DNA data storage. The first approach [2] achieved a density of 700 TB/gram, while the second architecture [3] increased the density to 2 PB/gram. This density gain was achieved thanks in part to the introduction of coding schemes at different levels of the architecture [3]: at the source coding/compression step, at the mapping step (differential encoding) and at an error control (channel coding) level. For the first step, Huffman coding (i.e., fixed-to-variable length entropy coding/compression method) was used. Differential encoding was used to represent the information based on the differences of consecutive symbols or the difference between a sequence and a given template. Finally, a single parity-check coding (i.e., a single symbol indicating the parity of the string) was considered for error control. This kind of channel coding was replaced by Reed-Solomon codes [4] in a more recent work [5]. An outstanding advance in DNA storage technologies was the rewritable storage architecture with random access capabilities presented in [6]. Their new scheme encompassed a number of coding features, including constrained coding, prefix synchronized coding and Low-Density Parity-Check (LDPC) coding for combating rewrite errors in stored data.

The process of reading the stored data in these storage systems is known as *DNA sequencing* and it raises multiple challenges. As long strands of bases (i.e., nucleotides) cannot be read in one step, sequencing devices must read short fragments of contiguous bases and then combine the data together to retrieve the original sequence. This process has motivated recent research into reconstruction algorithms [7] [8] [9].

From the channel coding perspective, DNA data storage has relied so far on tools that correct deletions/insertions/substitutions (a single one or a burst), asymmetric coding techniques [9] [10], as well as codes in the Damerau distance that correct single or block transposition errors [1] combined with deletions [11]. All these kinds of codes offer firm error correction capability and the associated decoding algorithms use bounded-distance decoding, based on taking a hard decision at the input of the decoder. In this work, we specifically focus on DNA sequencing and the analysis of

[1] these errors relate to nucleotides changing their position in the DNA sequence

the output nanopore sequencers. This leads to a new DNA sequencer channel model that opens new directions in coding strategies for DNA data storage.

The contributions of this paper can be listed as: first, we present the principles of DNA storage in a practical way in order to lead to a simple model of the *store-and-retrieve-data* chain. This model uses some of the typical blocks of the classical digital communication chain. Secondly, we present the sources of errors, their characteristics and a survey on the different coding schemes that have been proposed so far to overcome them. The third contribution is related to *third-generation* sequencing technologies: we introduce an original channel model based on recent work from the nanotechnology domain. Based on this new channel model we propose to consider graph- or trellis-based codes and their associated soft-decoding algorithms to improve performance of nanopore sequencing. To the best of our knowledge, this kind of decoding algorithms have never been proposed before in the context of DNA sequencing.

The paper is organized as follows: Section II introduces the definitions related to DNA storage and presents a simple model for the DNA storage chain. Section III describes DNA synthesis, its associated errors and state-of-the-art coding methods to improve the reliability of this process. Section IV deals with DNA sequencing focusing on recent advances on solid-state nanopores. In Section V we introduce an original communication and channel model that describes the behaviour of modern nanopore sequencers. This model opens new research directions in the domain of advanced coding schemes for reliable DNA storing systems.

II. The DNA-storage chain

A. Some definitions and context

Nucleotides (or bases [2]) Adenine (A), Cytosine (C), Guanine (G) and Thymine (T) are the building units of DNA. For storage purposes a long DNA molecule can be considered as a string over a non-binary alphabet of size $q = 4$. All the recently proposed systems are limited to a few hundred bases due to cost constraints and hard technological challenges still to be solved [12].

The writing process is known as *DNA synthesis* and relies on biotechnologies that are able to create DNA fragments from a pool of oligonucleotide building blocks. These methods are based on chemical oligonucleotide synthesis and the most promising one is known as the *microarray-based synthesis* method [13] [14]. This method can synthesize sequences of lengths up to 200 nucleotides, with a cost of roughly 0.001$ per nucleotide [3]. However, its major drawback is its high error rate. From a general point of view, we can state that current methods either combine high-cost and high-accuracy or low-cost and low-accuracy, and ongoing research tries to reduce the gap between these two extremes. In our channel model, we consider the synthesis process as a source of substitution errors and single-deletion/single-insertion errors.

Editing DNA data is possible with classical techniques that allow deleting or inserting DNA substrings at specific controlled locations. This process consists in deleting and inserting DNA substrings at well-controlled locations or, also, editing can be performed by adding very specific point mutations [15] [16].

The reading process is known as *DNA sequencing* and its goal is to determine the exact nucleotides and their order in the DNA data sequence. There have been multiple generations of sequencing technologies. Sanger sequencing [17], originally from the seventies, is known as first-generation. Next-generation technologies ([18] [19] among others) have resulted in an impressive decrease of the sequencing costs over the last decade. In this work we specifically focus on the third-generation sequencers based on nanopores. This technology is based on the detection of changes in an ion current when a DNA sequence passes through a nanoscale hole. Each nucleotide base causes a different *amount* of *current drop*, this making the identification possible.

B. A simplified model for the DNA storage chain

Fig. 1 presents a simplified model for the DNA storage chain. In this model we do not consider compression (or source coding). The "binary to nucleotide mapping" block defines the correspondence between binary subsequences of length two to the DNA alphabet: $\{11 \rightarrow A, 10 \rightarrow T, 01 \rightarrow C, 00 \rightarrow G\}$.

The "Channel encoding" block includes error-correction methods at two levels, as presented in Fig. 2. The first level corrects deletions and insertions introduced by synthesis; the second level deals with substitution errors and the errors introduced by the sequencing techniques, which is the focus of this paper. We introduce graph and trellis-based error-correcting codes at this level in order to exploit the soft information at the output of the sequencer system. The "Channel decoding" block also performs at two levels: the first is to decode deletion/insertion codes, and the second level is a soft-decoder through a trellis or a bipartite graph. At the end of the chain, the "Nucleotide to binary mapping" block performs the inverse operation of the "binary to nucleotide mapping" block, providing a binary information sequence.

More details on the DNA synthesis, the kind of error it introduces in the chain and a brief state of the art of coding techniques are presented in Section IV. DNA editing is considered to be error-free in this model. The original contribution of this paper is at the "DNA sequencing" block. To be specific we introduce a modulation scheme and a channel model to characterize the output of nanopore sequencers, which are new solid state third-generation sequencing systems. Our contributions on DNA sequencers modelling are in Sections IV and V.

III. Error in DNA synthesis and state-of-the-art correction techniques

The dominant error events in the DNA synthesis are simple substitutions [2] [3] [6] and error rates mainly depend on the

[2] with a slight abuse of meaning, we use both terms.

[3] other synthesis methods report a cost of 0.15$ per nucleotide.

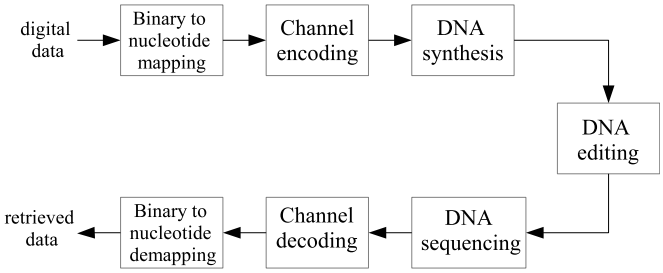

Fig. 1. DNA data storing and retrieving chain.

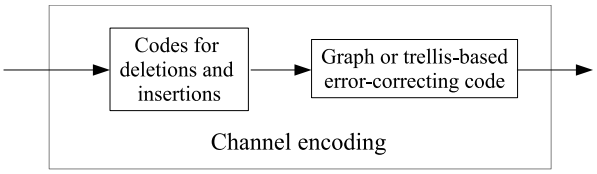

Fig. 2. Two-level channel encoding for the DNA storage chain.

cost of the technology [20] [21] [22]. Sequencing methods [4] using Polymerase Chain Reaction (PCR) *amplify* these substitutions errors by creating many copies of the synthesized sequence. Moreover, with high-throughput sequencing, synthesis errors propagate through a number of reads produced through sequencing. These issues have been addressed in [9] [24] [25] with the introduction of DNA profile codes.

Single insertion/deletion errors can also be introduced during synthesis. Tenengolts codes [26] are well adapted for this kind of errors and can be directly encoded into the DNA sequence. Also, the problem of reconstructing the DNA sequence from deletions/insertions followed by PCR techniques has been considered in [27] and [28].

Because we specifically focus on nanopore sequencing techniques, we assume that we are not concerned by sequence reconstruction. Note, however, that sequence reconstruction has motivated a large amount of recent research [28] [29] [30].

IV. Errors in DNA Sequencing

Sequencing technologies typically trade off accuracy for the length of the strands that can be read in a single shot. For example, modern third-generation sequencers offer longer reads, but also induce more errors compared to the first-generation methods [17]. As described in Section I, in this work we specifically focus on nanopore sequencers, which are based on the readout of current signals when the DNA sequence passes through a nanoscale hole in a membrane. Only one nucleotide base is read at a time, and its identity is determined by the current drop that it generates (i.e., each base leads to a different drop value because of their different atomic structures).

So far only two works have addressed the errors caused by the technological constraints of nanopore sequencers. The first [10] proposes a new family of codes, named Asymmetric Lee distance codes, to deal with substitution errors characterized by the impulse response distributions of the sequencer output signals. There errors are considered *asymmetric* because some substitutions are much more likely than others (e.g., base A is much more likely to be substituted by a T than by a G). The second work [31] addresses the problem of fast translocation speeds of DNA molecules across the nanopore in the ultrathin membranes, which leads to burst deletions [32]. To correct this kind of errors the authors created a non-binary burst-deletion correcting code based on their work in [33].

[4]including the Illumina platform, one of the most frequently used [23].

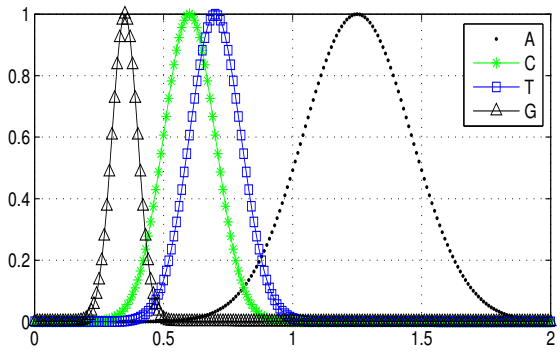

Fig. 3. Probability density functions of signal r_i.

In [34] the authors optimize the translocation speeds for single nucleotides by exploiting the high viscosity of room-temperature ionic liquids. This makes burst deletion errors much less common. The authors then use this approach to statistically detect the four kinds of nucleotides and provide histograms for the current drops. The overlapping of the current responses show that substitution errors become dominant (see Fig. 3). In the next Section, we introduce a new model that can be considered to characterize the sequencer output as a well-known modulation and channel from the digital communications domain.

V. Modulation and Channel Model for Nanopore Sequencers

We propose to model DNA nanopore sequencing approach of [34] with the digital modulation scheme depicted in Fig. 4. The "4-PAM modulation" block performs a memoryless Pulse-Amplitude Modulation where the $M = 4$ signal waveforms correspond to current drops in the nanopore sequencer. They are expressed as: $s_m = A_{mc} \cdot g(t)$ where $m = 1, \ldots, 4$, $g(t)$ is a pulse shape function and A_{mc} is a coefficient in a set of $M = 4$, each one corresponding to a nucleotide.

In the proposed model, after the modulation step, each signal goes through an Additive White Gausian Noise (AWGN) channel, such that the output signal (i.e., signal provided by the nanopore sequencer) is $r = s_i + n_i$. Note that n_i is a random variable that follows a Gaussian distribution with zero mean and a variance σ_i^2, $i = 1, \ldots, 4$, whose value depends on the nucleotide causing the current drop [34]. To be specific, the σ value is determined from the normalized histogram of current drops for each DNA homopolymer. Table I presents

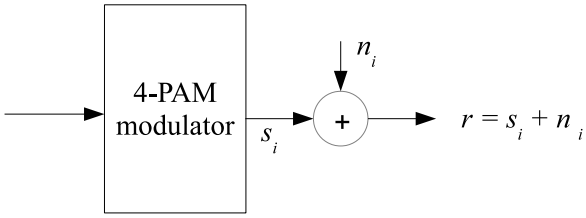

Fig. 4. 4-PAM modulation and AWGN channel to model the output of the nanopore sequencer in [29].

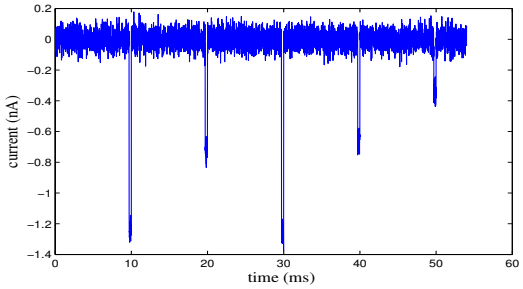

Fig. 5. Schematic waveform of sequencer output for nucleotide chain $\{A, T, A, C, G\}$.

the 4-PAM A_{mc} coefficients, as well as the current-drop and σ values for each nucleotide. These values correspond to the conditions reported in [34]. As an example, Fig. 5 shows the schematized shape of the sequencer output for a chain of nucleotides $\{A, T, A, C, G\}$.

TABLE I
PAM COEFFICIENTS, CURRENT-DROP AND σ VALUES FOR EACH NUCLEOTIDE

Nucleotide	A	C	T	G
A_{mc}	0.49	0.86	0.96	1.69
Current drop (nA)	1.25	0.64	0.71	0.36
σ value of n_i	0.21	0.1	0.1	0.05

A. Advanced error-correcting codes for speed-controlled nanopore sequencers

As an alternative to the models discussed in [10] and [31], we thus propose a model that uses graph or trellis-based codes over an alphabet of $q = 4$ elements. An example of such graph-based codes are LDPC codes defined on a Galois field of order $q = 4$ [35]. This kind of codes are defined by sparse matrices whose elements belong to the Galois field. Their representation with weighted-edge bipartite graphs allow for the use of message-passing decoding algorithms, which are considered as optimal in the digital communication/error-correcting coding domain [35] [36] and have shown performance very close to theoretical limits.

An example of a high-performance trellis-based code would be a duo-binary turbo-code [37]. This coding scheme has been massively studied and a number of decoding algorithms such the Soft Output Viterbi Algorithm [38] [39] or the Maximum A Posteriori algorithm [40] were proposed to obtain error-correcting performance that approaches information theory limits.

The channel model introduced in this Section opens new directions in the use of these high-performance channel coding schemes and their associated soft-decoding algorithms to the emerging field of DNA storage. So far, only algebraic codes and hard decoding approaches have been considered. However, with the proposed model, the use of soft-decoding algorithms has a potential to significantly reduce error rates in these devices.

VI. CONCLUSION

In this paper we first reviewed the principles of DNA data storage to introduce a simple model of its associated chain. We then provided an state of the art on error-correcting codes that help overcoming the technological constraints of this recent technology at each of its steps. The paper then focused on nanopore sequencers and proposed a modulation-and-channel model for a speed-controlled version of these. We finally proposed the usage of advanced error-correcting codes to highly exploit the soft information they provide. With this work, we open new paths in the area of channel coding for DNA-based storage devices. Future work will focus on the design of well-adapted coding schemes as well as analyzing/simulating their performance with the proposed model.

ACKNOWLEDGEMENTS

The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007-2013) under REA grant agreement n. PCOFUND-GA-2013-609102, through the PRESTIGE programme coordinated by Campus France. The authors would like to thank Siyi Yang for valuable help with references.

REFERENCES

[1] T. Hale, "How much data does the world generate every minute?" *IFL Science website*, July 2017. [Online]. Available: http://www.iflscience.com/technology/how-much-data-does-the-world-generate-every-minute/

[2] G. M. Church, Y. Gao, and S. Kosuri, "Next-generation digital information storage in DNA," *Science*, vol. 337, no. 6102, pp. 1628–1628, 2012.

[3] N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E. M. LeProust, B. Sipos, and E. Birney, "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA," *Nature*, vol. 494, no. 7435, pp. 77–80, February 2013. [Online]. Available: https://www.nature.com/articles/nature11875

[4] I. S. Reed and G. Solomon, "Polynomial codes over certain finite fields," *Journal of the Society for Industrial and Applied Mathematics*, vol. 8, no. 2, pp. 300–304, 1960.

[5] R. N. Grass, R. Heckel, M. Puddu, D. Paunescu, and W. J. Stark, "Robust chemical preservation of digital information on DNA in silica with error-correcting codes," *Angewandte Chemie International Edition*, vol. 54, no. 8, pp. 2552–2555.

[6] S. Yazdi, Y. Yuan, J. Ma, H. Zhao, and O. Milenkovic, "A rewritable, random-access dna-based storage system," *Sci. Rep*, vol. 5, no. 14138, pp. 397–409, 2015.

[7] F. Sala, R. Gabrys, C. Schoeny, and L. Dolecek, "Exact reconstruction from insertions in synchronization codes," *IEEE Transactions on Information Theory*, vol. 63, no. 4, pp. 2428–2445, April 2017.

[8] R. Heckel, I. Shomorony, K. Ramchandran, and D. N. C. Tse, "Fundamental limits of DNA storage systems," in *2017 IEEE International Symposium on Information Theory (ISIT)*, June 2017, pp. 3130–3134.

[9] H. M. Kiah, G. J. Puleo, and O. Milenkovic, "Codes for DNA sequence profiles," *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 3125–3146, June 2016.

[10] R. Gabrys, H. M. Kiah, and O. Milenkovic, "Asymmetric Lee distance codes for DNA-based storage," *IEEE Transactions on Information Theory*, vol. 63, no. 8, pp. 4982–4995, Aug 2017.

[11] R. Gabrys, E. Yaakobi, and O. Milenkovic, "Codes in the Damerau distance for DNA storage," *CoRR*, vol. abs/1601.06885, 2016. [Online]. Available: http://arxiv.org/abs/1601.06885

[12] S. M. H. T. Yazdi, H. M. Kiah, E. Garcia-Ruiz, J. Ma, H. Zhao, and O. Milenkovic, "DNA-based storage: Trends and methods," *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, vol. 1, no. 3, pp. 230–248, Sept 2015.

[13] X. Gao, E. Gulari, and X. Zhou, "In situ synthesis of oligonucleotide microarrays," *Biopolymers*, vol. 73(5), pp. 579–96, April 2004.

[14] S. Singh-Gasson *et al.*, "Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array," *Nature Biotechnology*, vol. 17(10), pp. 974–8, Oct 1999.

[15] I. Wataru, I. Hiroshi, and K. Yoshikazu, "A general method for introducing a series of mutations into cloned dna using the polymerase chain reaction," *Gene*, vol. 102, no. 1, pp. 67 – 70, 1991.

[16] R. Higuchi, B. Krummel, and R. Saiki, "A general method of in vitro preparation and specific mutagenesis of DNA fragments: study of protein and DNA interactions," *Nucleic Acids Res.*, no. 16, pp. 7351–67, August 1988.

[17] F. Sanger, S. Nicklen, and A. Coulson, "DNA sequencing with chain-terminating inhibitors," *Proc Natl Acad Sci USA*, no. 74(12), pp. 5463–7, Dec. 1977.

[18] D. A. Wheeler *et al.*, "The complete genome of an individual by massively parallel DNA sequencing," *Nature*, vol. 452, p. 872–876, April 2008.

[19] M. G. Ross, C. Russ, M. Costello, A. Hollinger, N. J. Lennon, R. Hegarty, C. Nusbaum, and D. B. Jaffe, "Characterizing and measuring bias in sequence data," *Genome Biology*, vol. 14, no. 5, p. R51, May 2013.

[20] W. Wan *et al.*, "Error removal in microchip-synthesized DNA using immobilized muts." *Nucleic Acids Res.*, vol. 42(12), Jul 2014.

[21] X. Gao, E. Gulari, and X. Zhou, "In situ synthesis of oligonucleotide microarrays," *Biopolymers*, vol. 73(5), pp. 579–96, April 2004.

[22] D. G. Gibson *et al*, "Chemical synthesis of the mouse mitochondrial genome," *Nature Methods*, vol. 7, p. 901–903, 2010.

[23] E. Schadt, S. Turner, and A. Kasarskis, "A window into third-generation sequencing." *Hum Mol Genet.*, vol. 19(R2), no. R227-40, Oct. 2010.

[31] C. Schoeny, F. Sala, and L. Dolecek, "Novel combinatorial coding results for dna sequencing and data storage," *IEEE Asilomar Conference on Signals, Systems, and Computers*, vol. abs/1801.04882, Oct. 2017.

[24] H. M. Kiah, G. J. Puleo, and O. Milenkovic, "Codes for DNA sequence profiles," in *2015 IEEE International Symposium on Information Theory (ISIT)*, June 2015, pp. 814–818.

[25] ——, "Codes for dna storage channels," in *2015 IEEE Information Theory Workshop (ITW)*, April 2015, pp. 1–5.

[26] G. Tenengolts, "Nonbinary codes, correcting single deletion or insertion (corresp.)," *IEEE Transactions on Information Theory*, vol. 30, no. 5, pp. 766–769, Sep 1984.

[27] F. Sala, R. Gabrys, C. Schoeny, and L. Dolecek, "Three novel combinatorial theorems for the insertion/deletion channel," in *2015 IEEE International Symposium on Information Theory (ISIT)*, June 2015, pp. 2702–2706.

[28] F. Sala, R. Gabrys, C. Schoeny, K. Mazooji, and L. Dolecek, "Exact sequence reconstruction for insertion-correcting codes," in *2016 IEEE International Symposium on Information Theory (ISIT)*, July 2016, pp. 615–619.

[29] A. Lenz, P. H. Siegel, A. Wachter-Zeh, and E. Yaakobi, "Coding over sets for DNA storage," *CoRR*, vol. abs/1801.04882, 2018.

[30] R. Heckel, I. Shomorony, K. Ramchandran, and D. N. C. Tse, "Fundamental limits of DNA storage systems," *CoRR*, vol. abs/1705.04732, 2017. [Online]. Available: http://arxiv.org/abs/1705.04732

[32] C. R. O'Donnel, H. Wang, and W. B. Dunbar, "Error analysis of idealized nanopore sequencing," *Electrophoresis*, vol. 34(15), pp. 2137 – 2144, Aug. 2013.

[33] C. Schoeny, A. Wachter-Zeh, R. Gabrys, and E. Yaakobi, "Codes correcting a burst of deletions or insertions," *IEEE Transactions on Information Theory*, vol. 63, no. 4, pp. 1971–1985, April 2017.

[34] J. Feng, K. Liu, R. D. Bulushev, S. Khlybov, D. Dumcenco, A. Kis, and A. Radenovic, "Identification of single nucleotides in MoS_2 nanopores," *Nature Nanotechnology*, vol. 10, pp. 1070–1078, Dec 2015.

[35] M. C. Davey and D. J. C. MacKay, "Low density parity check codes over GF(q)," *IEEE Communications Letters*, vol. 2, no. 6, pp. 159–166, June 1998.

[36] D. Declercq and M. Fossorier, "Decoding algorithms for nonbinary LDPC codes over GF(q)," *IEEE Trans. on Commun.*, vol. 55, pp. 633–643, April 2007.

[37] C. Berrou, M. Jézéequel, C. Douillard, and S. Kerouédan, "The advantages of non-binary turbo codes," in *Proceedings 2001 IEEE Information Theory Workshop (Cat. No.01EX494)*, 2001, pp. 61–63.

[38] J. Hagenauer and L. Papke, "Decoding turbo-codes with the soft output viterbi algorithm (SOVA)," in *Proceedings of 1994 IEEE International Symposium on Information Theory*, Jun 1994, pp. 164–.

[39] M. Fossorier, F. Burkert, S. Lin, and J. Hagenauer, "On the equivalence between SOVA and max-log-map decodings," *IEEE Communications Letters*, vol. 2, no. 5, pp. 137–139, May 1998.

[40] P. Robertson, E. Villebrun, and P. Hoeher, "A comparison of optimal and sub-optimal MAP decoding algorithms operating in the log domain," in *Communications, 1995. ICC '95 Seattle, 'Gateway to Globalization', 1995 IEEE International Conference on*, vol. 2, Jun 1995, pp. 1009–1013 vol.2.