DNA-based Molecular Communications

Bilgesu A. Bilgin, Ergin Dinc, *Member, IEEE*, Ozgur B. Akan, *Fellow, IEEE*
Electrical Engineering Division, Department of Engineering
University of Cambridge, UK
Email: bab46@cam.ac.uk, ed502@cam.ac.uk, oba21@cam.ac.uk

***Abstract*—In this paper, we propose a novel DNA-based molecular communication (MC) protocol towards high capacity communication between nanomachines for the first time in the literature. In the proposed protocol, transmitter is capable of emitting DNA strands having different lengths as information carrying molecules. Receiver contains receptor nanopores through which these negatively charged DNA strands pass, and duration of translocation event is utilized for selective sensing. We develop an analytical model for the proposed protocol to model diffusion, capturing, detection and reception processes. In MC literature, the processing times at the receiver is mostly neglected, but our protocol is the first MC protocol which considers the effect of processing times that are dependent on the DNA lengths. In addition, the number of detected DNA strands show significant dependence on diffusion constant, which changes according to DNA length. Therefore, we introduced a novel technique to minimize the effects of inter-symbol interference by adjusting the threshold level of each DNA strand according to its diffusion dynamics and detection rates. Furthermore, the proposed analytical model is exploited to derive information and communication theory metrics, i.e., capacity and bit error rate, for different communication metrics such as DNA lengths, the number of symbols, molecule thresholds and communication range by using realistic system parameters that are taken from experimental studies in the literature. At the end, the presented results show that the proposed DNA-based MC protocol is able to achieve capacity levels close to 6bps.**

***Index Terms*—Molecular Communication, Channel Modeling, Capacity, Internet of Nano Things**

I. Introduction

Molecular communication (MC) is a bio-inspired communication technique for transmitting and receiving information by means of molecules, in the same way the biological entities communicate between each other, i.e., ion channels between cells up to a few micrometers, hormonal communications up to several meters and pheromones up to several kilometers [1], [2]. Since the fundamentals of MC has been created by the nature itself through 3.5 billion years of evolution, MC is inherently biocompatible and energy-efficient. For these reasons, MC is the most promising method of communication for nanonetworks, i.e., set of nanomachines performing various tasks such as sensing, actuating and computing, towards the concept of Internet of Nano Things (IoNT) [1], [3]. Therefore, various aspects of MC has been studied in the literature such as MC in synaptic channel [4], multi-user MC [5], but the capacity of MC is quite limited due to the slow nature of communication via diffusion as analyzed in [6], [7]. In order to tackle this issue, the number of bits, i.e. the number of molecule types that can be selectively distinguished, is required to be increased to achieve higher data rates. Towards this purpose, we propose a DNA-based MC protocol, where the information is encoded with DNA strands having different properties.

In the nature, DNA is carrying information one generation to another. In a similar manner, we can exploit the properties of DNA for bit-wise MC to carry information from one place to another. Recent advancements in DNA sequencing and synthesis techniques have enabled DNA-encoded MC [8], [9]. For information transmission, communication symbols can be realized with DNA strands having different properties, i.e., length [10], dumbbell hairpins [8], [11], and short sequence motifs/labels [9]. For information detection, solid-state [9] and DNA-origami [12] based nanopores can be utilized to distinguish information symbols, i.e., the properties of DNA strands by examining the current characteristics while DNA strands pass through the nanopores. The utilization of nanopores for DNA symbol detection also enables the miniaturization of MC capable devices towards the realization of IoNT. According to [8], 3-bit barcode coded DNA strands with dumbbell hairpins can be detected through nanopores with 94% accuracy. The time of this process depends on the voltage, concentration and length of the DNA symbols, and the translocation of symbols can take up to a few ms up to hundred ms time frames [8], [11]. Considering the slow diffusion channel in MC, transmission/detection of DNA-encoded symbols do not introduce a bottleneck and multiple detections can be performed during each symbol transmission. Therefore, the utilization of DNA strands is promising for high capacity communication between nanomachines as the number of symbols can be increased by exploiting multiple properties of DNA: length, dumbbell hairpins, and short sequence motifs/labels.

In MC, several modulation schemes have been proposed to encode information into concentration, type or composition of molecules: concentration shift-keying (CSK), molecule shift-keying (MoSK), isomer shift-keying (ISK), Nucleotide Shift-Keying (NSK) [13]. CSK is akin to amplitude shift keying (ASK) such that the information is embedded to the concentration of the information molecule. At the simplest case, i.e., on-off keying or 1-bit ASK, the molecular receiver (Rx) decodes high logic when the concentration of the information molecule exceeds a certain threshold and low logic otherwise. Although it is possible to increase the number of symbols by increasing the number of available concentration levels, interference prone nature of the communication via diffusion significantly limits the number of bits. That's why, most studies consider the simplest version of CSK, on-off keying. In MoSK, the information is encoded by using multiple molecules, and the utilization of k molecules provide 2^k symbols for achieving

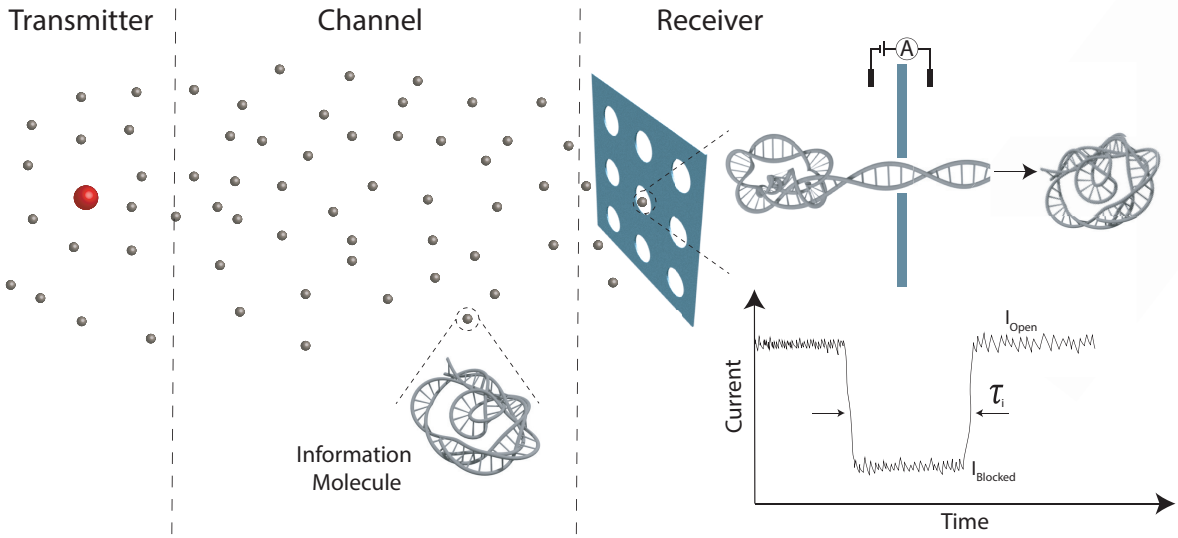


Fig. 1: System model for the DNA-based molecular communication protocol.

higher capacity. MoSK requires more complex Rx that can distinguish k molecules for successful communication, and the modulation of individual molecules are based on on-off keying as in CSK case. MoSK can achieve higher capacity, but the main limiting factor for this modulation type is the number of molecules that can be selectively received.

Furthermore, the authors of [14] proposed the utilization of isomer, i.e., the molecules having the same atoms in a different orientation, under different modulation schemes CSK, MoSK and molecule ratio-keying, where the information is encoded into ratio of received molecule types. In nucleotide shift-keying (NSK), information is encoded into the base sequences of deoxyribonucleic acids (DNAs), i.e., Nucleotide Shift-Keying (NSK). The proposed MC-TxRx designs are capable of transceiving DNA molecules. Using DNA, with its huge information storage capacity and robustness against environmental conditions, will enable high-rate and reliable MC. However, the DNA reading/writing speed and cost at the moment limits the utilization of NSK in a practical system. However, the utilization of DNA as information carrying molecule paves the way for high capacity links between nanomachienes by enabling higher number of molecules that can be selectively received at the Rx.

In this work, we propose a novel DNA-based MC protocol towards high capacity communication between nanomachines for the first time in the literature. In the proposed protocol, transmitter (Tx) is capable of emitting DNA strands of k different lengths as information carrying molecules. As presented in Fig. 1, Rx contains receptor nanopores through which these negatively charged DNA strands pass thanks to the applied potential, and as they do, they obstruct ionic currents that normally flow through. The duration of the current obstruction is proportional to the length of the DNA strand that passes through, which is utilized for selective sensing. We develop an analytical model for the proposed protocol to model diffusion, capturing, detection and reception processes. During the detection process, the nanopore is blocked for a certain time, which depends on the size of the DNA strand, and no other molecule can be captured during this processing time. In the molecular communication literature, the processing times at the Rx is mostly neglected, but our protocol is the first molecular communication protocol which considers the effect of processing times that are dependent on the DNA lengths. Due to the processing time, there are finite number of possible sequences that can be received during a sampling period. In our proposed analytical model, we calculate the probability of detecting all possible detection sequences. In addition to the processing times, diffusion constants of DNA strands also significantly depend on their length. Therefore, the expected numbers of detecting different DNA strands show significant variations according to their diffusion constants. Hence, decision thresholds for different DNA strands are required to be adjusted accordingly, and we calculate the optimal threshold values to minimize the effect of inter-symbol interference (ISI). Furthermore, the proposed analytical model is exploited to derive ICT metrics, i.e., capacity, for different communication metrics such as DNA lengths, the number of symbols, molecule thresholds and communication range by using realistic system parameters that are taken from experimental studies in the literature. At the end, the presented results show that the proposed DNA-based MC protocol is able to achieve capacity levels close to 6bps.

The remainder of the manuscript is organized as follows. Section II introduces the analytical model for diffusion, capturing, detection and reception processes of the proposed DNA-based molecular communication protocol. In Section III the main parameters for the DNA-based MC protocol are explained. Section IV includes the simulation results for the proposed protocol. Lastly, the conclusions and future work of our research are discussed in Section V.

II. Analytical Model

A. Model Setup and Communications Protocol

We assume that, Tx and Rx are d distance apart from each other situated inside a fluidic 3-dimensional medium, and that, Tx transmits an r-bit symbol $s_T \in \mathcal{S}_r$, where $\mathcal{S}_r$ is the space of r-bit words with size $|\mathcal{S}_r| = 2^r$, by controlling the release of r types of DNA strands of different lengths l_i measured in kilo base pairs (kbp), $1 \leq i \leq r$. Tx releases N_i many of the i^{th} type DNA strands, if and only if $s_T(i)$, the i^{th} bit of the symbol s_T, is logic one. Tx transmits these symbols regularly with an inter-transmission period of length T, more precisely, Tx transmits $s_T^{(k)}$ at $t = kT$, $k \in \mathbb{N}$. We further assume that, Rx is synchronized with Tx, that is, to detect s_k Rx samples its environment during the sampling period $[kT, (k+1)T)$, where it receives the r-bit symbol $s_R^{(k)}$.

Rx is assumed to have a densely packed array of M nanopores, or detection sites, where at each site DNA strands are captured successively and for each type the number of strands captured during a sampling period is recorded. Here, the process of detection of a single strand includes the capturing of the strand, which is characterized by the capturing rate constant K_i in units $mol^{-1}s^{-1}$, as well as, the process of determination of its type after capture, which is derived from the time τ_i the captured strand takes to go through the nanopore and is related proportionally to its length by the translocation speed v, i.e., $v = l_i/\tau_i$ for all $1 \leq i \leq r$. After summing up all the detections for each type from its M detection sites during the given sampling period, Rx decides that the i^{th} bit of the transmitted word is received as high logic if and only if the total number of i^th type detections, n_i, is not less than a prefixed threshold number n_i^{th}, i.e., $n_i \geq n_i^{th}$. n_i^{th} is an optimisation parameter, which, if selected high, decreases ISI in expense of decreasing reception probability.

B. Diffusion, Capturing, Detection and Reception Processes

In the following analysis we assume that, a site, that is processing a captured strand to determine its type, can not capture another strand until the determination process is over. In the derivation of the following analytical model, we will also assume that, the dimensions of the nanopore array at Rx are negligibly small compared to d, the movements of individual DNA strands are mutually independent, and that the effects of DNA strand capturing at Rx on strand concentrations near Rx is negligible.

Upon a transmission of a symbol s_T from Tx at time $t = 0$, the released DNA strands diffuse across the encompassing fluidic medium to reach Rx. Thanks to the assumptions given above, and also assuming that the medium is devoid of other sources of DNA strands, the expected concentration, $c_i(t)$, of i^{th} type of strand near Rx at a given time $t > 0$ can be approximated by the solution of the diffusion equation

$$\frac{d}{dt}c_i(x,t) - D_i \Delta c_i(x,t) = 0,$$

in $\mathbb{R}^3$, namely

$$c_i(t) = \frac{s_T(i) N_i}{(4\pi D_i t)^{3/2}} \exp\left\{\frac{d^2}{4D_i t}\right\}, \tag{1}$$

where D_i is the diffusion coefficient of i'^{th} type of strand.

On Rx, for a single detection site S, we define the function

$$S(t) = \begin{cases} 0 & \text{, if } S \text{ is free at time } t, \\ i & \text{, if } S \text{ is occupied by } i^{th} \text{ type of strand at time } t. \end{cases}$$

Now, the capturing rate of i^{th} type of DNA strand is given by

$$\lambda_i(t) = K_i c_i(t), \quad t \geq 0.$$

In absence of other types of strands, i.e., $c_j = 0$ for all $j \neq i$, $f_{C,i}(t)$, the probability distribution function (PDF) for the capturing process of i^{th} type of strand, provided that $S(s) = 0$, is given by the exponential distribution with time-varying rate $\lambda_i(t)$

$$f_{C,i}(t) \Big|_{\substack{c_j=0, j\neq i; \\ S(s)=0}} = \chi_{[s,\infty)}(t)\lambda_i(t) e^{-\int_s^t \lambda_i(\tau)\, d\tau},$$

where $\chi_I(t)$ is the characteristic function of the interval I, i.e., it is one when t belongs to I and zero otherwise. Thus, the probability that the first capture of an i^{th} type of strand falls into the time interval (s,t) is given by

$$P_{C,i}(s,t) \Big|_{S(s)=0} = \int_s^t f_{C,i}(\tau) d\tau.$$

It follows that, the probability of having no capture of i^{th} type in (s,t), provided that $S(s) = 0$, can be found as

$$P_{C,i}^{(0)}(s,t) \Big|_{S(s)=0} = 1 - P_{C,i}(s,t) \Big|_{S(s)=0} = e^{-\int_s^t \lambda_i(\tau)\, d\tau}.$$

Now, when multiple types of strands are present, by the assumption that the motions of individual DNA strands are mutually independent from each other, given that $S(s) = 0$, the probability of S not capturing any strand during (s,t) can be written as the product of non-capturing probabilities of each type, i.e.,

$$P_C^{(0)}(s,t) \Big|_{S(s)=0} = \prod_{i=1}^{r} P_{C,i}^{(0)}(s,t) \Big|_{S(s)=0} = e^{-\int_s^t \sum_{i=1}^{r} \lambda_i(\tau)\, d\tau}.$$

It follows that, the PDF describing the capturing process by S in presence of all types strands, provided that $S(s) = 0$, reads as

$$f_C(t) \Big|_{S(s)=0} = \sum_{i=1}^{r} f_{C,i}(t) \Big|_{S(s)=0},$$

with

$$f_{C,i}(t) \Big|_{S(s)=0} = \lambda_i(t) e^{-\int_s^t \sum_{j=1}^{r} \lambda_j(\tau)\, d\tau},$$

which can be identified as the PDFs corresponding to the event that the first captured strand will be of i^{th} type.

Next, since the detection process at S is combination of capturing and type determination processes, where the latter is just a delay by τ_i with certainty, given $S(s) = 0$, the PDF

describing the event of first detecting i^{th} type of strand is given by

$$\begin{aligned} f_{D,i}(t) \Big|_{S(s)=0} &= f_{C,i}(t-\tau_i) \Big|_{S(s)=0} \\ &= \chi_{[s+\tau_i,\infty)}(t)\lambda_i(t-\tau_i)e^{-\int_s^{t-\tau_i}\sum_{j=1}^{r}\lambda_j(\tau)\,d\tau}. \end{aligned} \tag{2}$$

Hence, the PDF for detecting the first DNA strand of any type can be found as the sum of the PDFs for first detection events of each type, that is,

$$f_D(t) \Big|_{S(s)=0} = \sum_{i=1}^{r} f_{D,i}(t) \Big|_{S(s)=0}.$$

Thus, the probability of first detection by S not happening in a given time interval (s,t), provided that $S(s) = 0$, is given by

$$P_D^{(0)}(s,t) \Big|_{S(s)=0} = 1 - \int_s^t f_D(\tau) \Big|_{S(s)=0} d\tau. \tag{3}$$

In what follows, it will be important to calculate the PDFs and probabilities given by (2) and (3), respectively, for all $(s,t) \subset (0,T)$, which, at first glance, hints that the calculation time for these parameters will scale with T^2. The following structural identities, however, which formulate these parameters in terms of PDFs and probabilities given $S(0) = 0$, allow us calculate them in linear time. For the PDFs in (2) it is easy to see that, we have for each $1 \le i \le r$

$$P_C^{(0)}(0,s) \Big|_{S(0)=0} f_{D,i}(t) \Big|_{S(s)=0} = \chi_{[s+\tau_i,\infty)}(t) f_{D,i}(t) \Big|_{S(0)=0}, \tag{4}$$

which simply says that, given S is initially free, having a first detection of i^{th} type after $s + \tau_i$ corresponds to having no capture until s and having i^{th} type detection given $S(s) = 0$. For probabilities given in (3), one can derive the relation

$$\begin{aligned} P_C^{(0)}(0,s) \Big|_{S(0)=0} &\left(1 - P_D^{(0)}(s,t) \Big|_{S(s)=0}\right) \\ &= \sum_{i=1}^{r} \chi_{[s+\tau_i,\infty)}(t) \int_{s+\tau_i}^{t} f_{D,i}(\tau) \Big|_{S(0)=0} d\tau, \end{aligned} \tag{5}$$

that is, given S is initially free, probability of capturing nothing until s and having a first detection in (s,t) given $S(s) = 0$ corresponds to the sum of probabilities of first detecting i^{th} type in $(s+\tau_i, t)$.

Now, S may detect more than one DNA strand during a sampling period, and because for each strand type the capturing rates $\lambda_i(t)$ are time-dependent and possibly distinct, detection probability of multiple types of strands by S during the sampling period depends on the order of the detected types, e.g., the event of detecting i^{th} type first and j^{th} type second has, in general, different probability of happening compared to the detection event in reverse order. Hence, a description of detection behaviour of S during a sampling period requires the consideration of all possible detection sequences $\mathcal{D}_T$ during a sampling period of length T. Thanks to the fact that the detection process for i^{th} type of strand involves τ_i delay, $\mathcal{D}_T$ contains finitely many finite detection sequences of the form $x = (d_1, \cdots, d_n)$, where d_j, $1 \le j \le n$, stands for the type of j^{th} detected strand, i.e., $1 \le d_j \le r$, and it can be characterized as

$$\mathcal{D}_T = \left\{ x = (d_1, \cdots, d_n) \,\middle|\, \sum_{j=1}^{n} \tau_{d_j} \le T \right\}.$$

In what follows, for any $x \in \mathcal{D}_T$, we will give an iterative description for, $P_D^{(x)}(0,T) \Big|_{S(0)=0}$, finding the probability that S will exactly detect the sequence x during the time interval $(0,T)$ given $S(0) = 0$. The iteration is done on $|x|$, the length of the sequence x. In particular, when $|x| = 0$, i.e., x is the empty sequence, the desired probability is given by (3). For $|x| \ge 1$, one can find this probability as

$$P_D^{(x)}(0,T) \Big|_{S(0)=0} = \int_0^T f_{D,x}(t) \Big|_{S(0)=0} P_D^{(0)}(t,T) \Big|_{S(t)=0} ds. \tag{6}$$

Here, $f_{D,x}(t) \Big|_{S(0)=0}$ is the PDF for detecting the sequence x first in $(0,t)$, and is given by

$$f_{D,x}(t) \Big|_{S(0)=0} = \int_0^t f_{D,\hat{x}}(\tau) \Big|_{S(0)=0} f_{D,x(|x|)}(t) \Big|_{S(\tau)=0} d\tau, \tag{7}$$

where $\hat{x} \in \mathcal{D}_T$ is the sequence obtained from x by removing its last entry $x(|x|)$.

The iteration relation (7) describes the detection of x first as a sum of all possible concatenations of detecting $\hat{x}$ first, and then, detecting $x(|x|)$ first. Note that, by our definition of detection, a detection by S at τ implies that S is free at τ, i.e., $S(\tau) = 0$. On the other hand, the relation (6) characterizes the probability of detecting exactly x in $(0,T)$ as the sum of all possible ways of detecting x first, and then, detecting nothing. Finally, utilizing the auxiliary relations (4) and (5) in (7) and (6), respectively, one obtains the more coding friendly relations.

Now, we are interested in the probability distribution of how many of each type of strand Rx will detect during the sampling period $(0,T)$, for which we define the probability mass function (PMF) for number of detections of i^{th} type by a single site S given $S(0) = 0$ as

$$P_{D,i}(n) \Big|_{S(0)=0} = \sum_{\substack{x \in \mathcal{D}_T, \\ n_i(x)=n}} P_D^{(x)}(0,T) \Big|_{S(0)=0}, \quad n \ge 0. \tag{8}$$

where $n_i(x)$ is the number of i^{th} type detections in x. For each $1 \le i \le r$ this PMF has finite support, indeed, it is zero for all $n \ge \frac{T}{\tau_i}$. However, it is not analytically trivial to verify that these PMFs have unit total mass, that is,

$$\sum_{n \ge 0} P_{D,i}(n) \Big|_{S(0)=0} = \sum_{n < \frac{T}{\tau_i}} P_{D,i}(n) \Big|_{S(0)=0} = 1, \quad 1 \le i \le r.$$

This equality, which is attained by the code developed following the analysis presented here, has served as a verification of our probabilistic derivations.

The PMF for the total number of detections of i^{th} type by Rx during the sampling period given that all sites on Rx are free at $t = 0$, $P_{D_{tot},i}(n) \Big|_{Rx(0)=0}$, can be derived from the PMF

for a single site by means of convolution under the assumption that, the detection processes of all sites on Rx are mutually independent, which reduces to the assumption that the effect of capturing process on DNA strand concentrations near Rx is negligible. More precisely, if there are M sites on Rx, then

$$P_{D_{tot},i}(n)\Big|_{Rx(0)=0} = \left(P_{D,i}(n)\Big|_{S(0)=0}\right)^{*M},$$

where for a discrete function f defined on $\mathbb{Z}$ f^{*M} denotes the convolution of f with itself M times.

Finally, by our protocol, in order to receive a high logic in i^{th} bit, Rx needs to detect i^{th} type of strand at least n_i^{th} many times, the probability of which is found as

$$P_{R,i}\Big|_{Rx(0)=0} = \sum_{n\geq n_i^{th}} P_{D_{tot},i}(n)\Big|_{Rx(0)=0}, \tag{9}$$

which we will denote as just $P_{R,i}$ for sake of ease of notation in the next discussion. In fact, we further change the notation to $P_{R,i}\Big|_{s_T}$ to emphasize the fact that the reception probabilities given by (9) depend on the symbol transmitted by Tx.

C. ICT Analysis

Next, we turn our attention to the ICT performance metrics of the communication link between Tx and Rx, namely, the bit error rate (BER) and capacity. In calculating BER and capacity we will assume that the input has maximum entropy, that is, Tx transmits any of the possible 2^r symbols with the same probability, i.e., with probability 2^{-r}.

To find BER, let us denote by $P_{R_{Err},i}\Big|_{s_T}$ the probability that Rx will receive the i^{th} bit of the transmitted symbol s_T erroneously, which is given as

$$P_{R_{Err},i}\Big|_{s_T} = \begin{cases} P_{R,i}\Big|_{s_T} & \text{, if } s_T(i)=0, \\ 1-P_{R,i}\Big|_{s_T} & \text{, if } s_T(i)=1. \end{cases}$$

Then, BER corresponding to i^{th} type of strand can be found as

$$BER_i = 2^{-r}\sum_{s_T\in\mathcal{S}_r} P_{R_{Err},i}\Big|_{s_T}. \tag{10}$$

Next, the capacity per transmission between Tx and Rx is given by the mutual information between the two nodes

$$I(Tx,Rx) = H(Rx) - H(Rx|Tx), \tag{11}$$

and the capacity between Tx and Rx can be found as

$$C(Tx,Rx) = \frac{I(Tx,Rx)}{T}.$$

In (11), $H(Rx)$ is the entropy at Rx under the assumption that Tx has maximum entropy, and $H(Rx|Tx)$ is the average over all possible s_T of the entropies at Rx for fixed s_T from Tx and is a measure of noise in the transmission. The entropy at Rx is given by

$$H(R_x) = -\sum_{s_R\in\mathcal{S}_r} P_{s_R}\log_2 P_{s_R},$$

where P_{s_R} is the probability that Rx will receive the signal s_R during a transmission, and is given as the average over all possible s_T of the conditional probability that s_R is received provided that s_T is transmitted

$$P_{s_R} = 2^{-r}\sum_{s_T\in\mathcal{S}_r} P_{s_R}\Big|_{s_T}.$$

In turn, these conditional probabilities can be calculated according to the formula

$$P_{s_R}\Big|_{s_T} = \prod_{i=1}^{r} P_{s_R,i}\Big|_{s_T},$$

where $P_{s_R,i}\Big|_{s_T}$ is the probability that the i^{th} bit received coincides with that of s_R given s_T is transmitted, and satisfies

$$P_{s_R,i}\Big|_{s_T} = \begin{cases} P_{R,i}\Big|_{s_T} & \text{, if } s_R(i)=1, \\ 1-P_{R,i}\Big|_{s_T} & \text{, if } s_R(i)=0. \end{cases}$$

Finally, $H(Rx|Tx)$ can be calculated by the formula

$$H(Rx|Tx) = 2^{-r}\sum_{s_T\in\mathcal{S}_r}\left(-\sum_{s_R\in\mathcal{S}_r} P_{s_R}\Big|_{s_T}\log_2 P_{s_R}\Big|_{s_T}\right).$$

The derivations so far do not take into account the effects of FD and ISI. In what follows we incorporate these into the model. For false detection (FD), we introduce the $r\times r$ FD matrix, A_{FD}, where its entry at i^{th} row and j^{th} column, $(A_{FD})_{ij}$, corresponds to the probability of i^{th} type detection given the captured strand was really of j^{th} type. We also define the r-vector first detection PDF, $\vec{f}_D(t)\Big|_{S(s)=0}$, with its i^{th} entry

$$\left(\vec{f}_D(t)\Big|_{S(s)=0}\right)_i = f_{D,i}(t)\Big|_{S(s)=0}, \quad 1\leq i\leq r,$$

given by (2). Then, FD can be incorporated into our model by defining the modified first detection PDFs accounting for FD as the entries of the modified vector PDF

$$\vec{f}_{D_{FD}}(t)\Big|_{S(s)=0} = A_{FD}\,\vec{f}_D(t)\Big|_{S(s)=0},$$

and use these PDFs in calculations instead.

In incorporating ISI into our model, we only consider interactions between two consecutive transmissions. In the derivations so far we had the assumption that, the initial DNA-strand concentrations are zero, which is not valid in the case of existence of a prior signal. We resolve this by considering, instead of the concentration in (1), the modified concentrations given by

$$c_i^{s_{T_p}}(t) = \frac{s_T(i)N_i}{(4D_it)^{3/2}}\exp\left\{\frac{d^2}{4D_it}\right\} + \frac{s_{T_p}(i)N_i}{(4D_i(t+T))^{3/2}}\exp\left\{\frac{d^2}{4D_i(t+T)}\right\},$$

for all possible previously transmitted symbols $s_{T_p}\in\mathcal{S}_r$. Then, the average reception probabilities for each bit can be found as

$$P_{R,i}^{ISI}\Big|_{Rx(0)=0} = 2^{-r}\sum_{s_{T_p}\in\mathcal{S}_r} P_{R,i}^{s_{T_p}}\Big|_{Rx(0)=0}.$$

We note that this simple approach of modelling ISI does not take into account for the effect of initial blocking of capturing sites at the beginning of a given sampling period caused by strands captured towards the end of the previous period. However, even though the detection of these strands would fall into the given period, according to our communication protocol we assume that, Rx does not take these into account, and hence, the initial partial occupancy is the only effect neglected by this ISI model.

III. Simulation Environment

This section introduces the parameters that are required for the capacity and BER calculations.

A. Diffusion Constant and Capturing Rate

Diffusion constant of DNA molecule depends on the size of the molecule, and the diffusion constant scales with the number of base pairs (bps) as [15], [16]

$$D = D_0 N^{-0.6}, \tag{12}$$

where D_0 is 5.9×10^{-10} m^2/s, and N is the number of bps.

The capturing event at the nanopores depends on the pore size, applied voltage, salt concentration of the medium, DNA concentration and DNA size. At low voltage values, the dependence on the DNA size becomes negligible [17], [10]. Since high voltage difference values (> 300mV) results in sub-ms translocation times due to the increased electrical forces, we assume that the DNA Rx will work under low voltage regime around 120mV, and capturing rate is independent of the DNA size. In the literature, wide range of translocation times from $51.26\mathrm{M}^{-1}\mathrm{ms}^{-1}$ [17] up to 1×10^6 $\mathrm{M}^{-1}\mathrm{ms}^{-1}$ [10] has been reported for different conditions. The low capturing rates are measured in 2° with sub-5nm nanopores, whereas the high values are reported for 15nm nanopores at 20° by using high voltage differences. For DNA-based molecular communication, extremely high capturing rates are not desirable as the concentration near Rx changes considerably and the effects of ISI would be amplified. By altering the pore size, salt concentration and voltage difference capturing rates between $51.26\mathrm{M}^{-1}\mathrm{ms}^{-1}$ and 1×10^6 $\mathrm{M}^{-1}\mathrm{ms}^{-1}$ can be achieved. Hence, we utilize $K_i = 5000\mathrm{M}^{-1}\mathrm{ms}^{-1} \forall i = 1, \cdots, r$.

B. Translocation Times

The average translocation speed can be assumed constant for any DNA size having more than 12 base pairs [18]. However, the distribution of the translocation times is not constant due to the randomness in the folding/unfolding of DNA while passing through the nanopore as illustrated in Fig. 1. In order to calculate the successful and FD of information molecules, we need the distribution of the translocation times. The translocation times of DNA strands show a complex distribution, which can be modeled as half gaussian and half falling exponential as suggested in [19], [20]. Therefore, we assume that the translocation time show a Gaussian distribution up to τ and than decrease with a falling distribution function, and the distribution can be expressed as

$$f_\tau(t) = \begin{cases} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\tau_p)^2}{2\sigma^2}} & t \leq \tau_p \\ B e^{-t/\psi} & t > \tau_p \end{cases}, \tag{13}$$

where σ is the standard deviation of the Gaussian distribution, ψ is the falling rate of the exponential part, B is a constant to make sure that the continuity of the pdf and is calculated as $B = e^{\tau_p/\psi}/\sqrt{2\pi\sigma^2}$, τ_p is the peak translocation time which is associated with the highest value in the translocation time PDF. The integral of $f_\tau(t)$ over all t needs to be 1, and this can be satisfied only if $\psi = \sqrt{\pi\sigma^2/2}$. The relationship between the peak time τ_p and the width of the distribution τ_w ($e^{-0.5}$ of the peak of the distribution) can be utilizes to calculate σ. The width of the distribution can be calculated as $\tau_w = \sigma + \psi/2 = \tau_p \chi$, where is the scale factor (τ_w/τ_p). In this work, we use a scale factor of 0.55 in order to get the translocation widths demonstrated in [21].

To calculate τ_p, constant peak translocation speed is utilized as $v_p = 0.015$nm/μs [18]. Therefore, the peak translocation times can be found with

$$\tau^p = L_0/v_p, \tag{14}$$

where L_0 is the length of the DNA in bp (0.34nm/base) [18]. Figure 2 shows the distribution of translocation times for DNA strands with 1kbp, 3kbp and 5kbp.

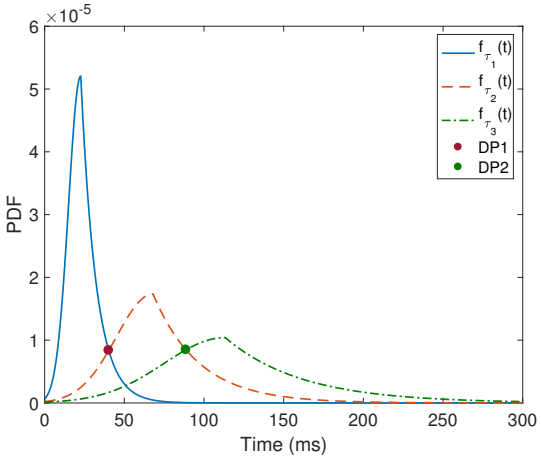


Fig. 2: Distribution of translocation times for 1kbp, 3kbp and 5kbp.

C. False Detection (FD)

The overlapping regions in the translocation time PDFs as seen in Figure 2 result in FD The DNA Rx adapts the maximum likelihood estimation technique, in which the symbol estimation is based on maximizing the likelihood function. That is, the decision point at Rx is selected as the intersection of neighboring PDFs as illustrated in Figure 2. Hence, FD probabilities when there are r number of distinct DNA strands,

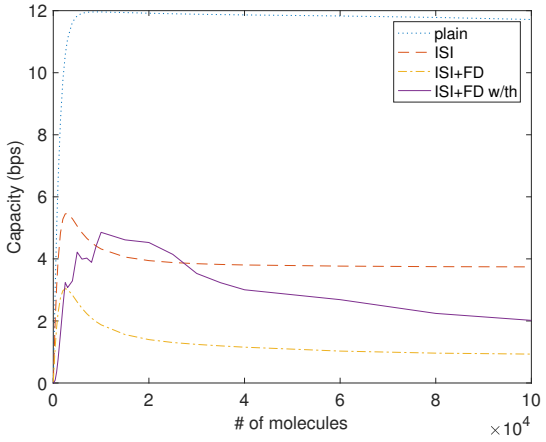


Fig. 3: Capacity vs. # of transmitted DNAs for different scenarios in case of three types of strands, 1kbp, 3kbp and 5kbp.

e.g. detecting i^{th} DNA strand while j^{th} strand captured, can be calculated as

$$\begin{aligned} A_{FD}^{(i,j)} &= \int_{DP_j}^{DP_{j+1}} f_{\tau_i}(t)dt \text{ for } i > j, \\ A_{FD}^{(i,j)} &= \int_{DP_j-1}^{DP_j} f_{\tau_i}(t)dt \text{ for } j > i, \\ A_{FD}^{(i,i)} &= 1 - A_{FD}^{(1:r\setminus i,j)}. \end{aligned} \tag{15}$$

where $DP_{0,\cdots,j+1}$ are the decision thresholds and $DP_0 = 10$ ms and $DP_{j+1} = \infty$. in this work, we assumed that the detection of the DNA strands only depends on the translocation times, i.e. the blockage time of the current through the nanopore. The detection process can be further improved by considering event charge deficit, i.e., the multiplication of mean current and the translocation times. As shown in [10], the event charge deficit has lower variance than the translocation times. However, in this work we assumed that the detection of the DNA strands only depends on the translocation times, i.e. the blockage time of the current through the nanopore, due to the unavailability of mean currents as a function of DNA strands.

IV. Simulation Results

The simulations are performed in MATLAB. In the simulations, the nominal values of the communication parameters can be found in Table I, and these parameters are utilized in the simulations if otherwise not stated.

TABLE I: Nominal Simulation Parameters.

Parameter	***Value***	***Unit***
Sampling Period T	*250*	*ms*
Number of Released Molecules N	10^4	-
Range d	*1*	µm
Minimum DNA length	*1*	*kbp*
# of nanopores	*25*	-
Capturing rate K	*5000*	$M^{-1}ms^{-1}$

A. Number of Transmitted DNAs and Nanopores

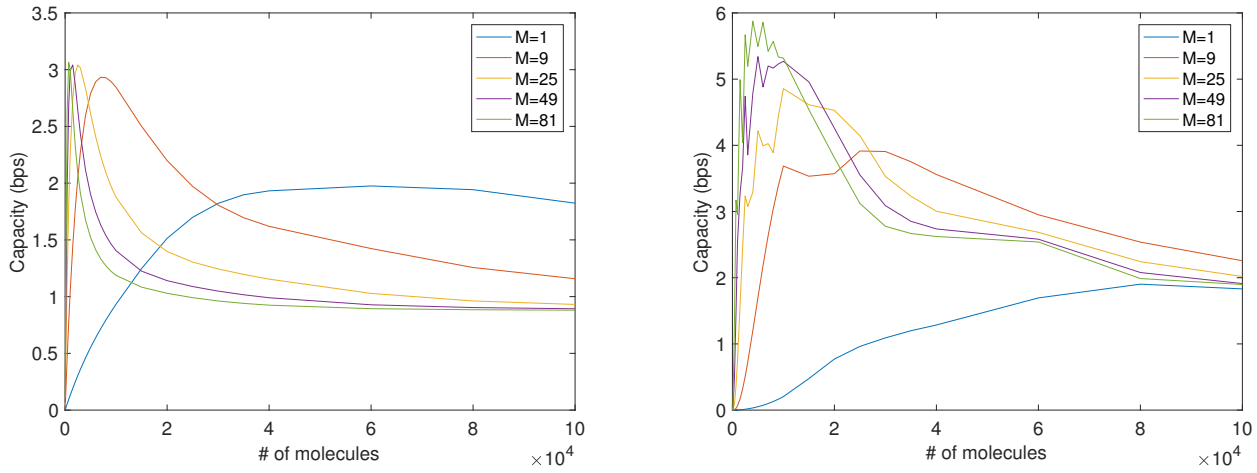
The capacity of communication of proposed protocol is highly dependant on the number of transmitted DNAs from Tx and the number of nanopores situated on Rx. Figure 3 illustrates the dependence of capacity on the number of transmitted DNAs in the case of three symbols $(1, 3, 5\text{kbp})$ for various scenarios, namely, when both ISI and FD are ignored (plain), with ISI but no FD (ISI), and with both ISI and FD (ISI+FD). All these scenarios assume minimal detection threshold values $(1, 1, 1)$ for reception, i.e., detection of a single molecule suffices for reception, except, for the case of both ISI and FD existing, we also present results for higher threshold values (ISI+FD w/th) obtained from an analytical guess described in next subsection. Results show that, one can achieve full rate, i.e., 12bps equivalent to three bits per transmission with T = 250ms, when ISI and FD are not accounted for. In case of ISI, as expected, capacity initially increases with increasing number of released DNAs, bot eventually decreases due to ISI. We observe that, with our guessed threshold values it is possible to nearly double the achievable rates, where the maximum is achieved at 10^4 transmitted molecules, which sets the basis for our choice of the nominal value in Table I.

The dependence of capacity on number of nanopores M on Rx is illustrated in Figure 4 for minimal and guessed threshold values. We observe that, with increasing M the capacity peaks shift to lower number of transmitted DNAs. This can be explained by increased sensitivity of Rx enabling reliable communication with low number of transmitter molecules, a regime where degrading effects of ISI is diminished. Also note that, with minimal thresholds capacity peaks saturates with increasing M to a limit (~ 3bps), and $M = 25$ almost provides this rate, which constitutes the basis of the nominal choice for M in Table I. Finally, one can observe that, with guessed thresholds, it is possible to improve channel capacity, and the extent of improvement is positively correlated with M.

B. Determination of Reception Thresholds

As mentioned before, in our proposed molecular communication protocol, Rx decides on the reception of a certain bit, if the corresponding type of DNA strand is captured more than a prefixed threshold number of times during a sampling period. This feature of our protocol allows us to partially alleviate the effects of ISI, from which diffusion based molecular communication is known to suffer particularly. The data rates that can be achieved are highly sensitive to these threshold numbers, and the proper choice of them is a highly non-trivial problem. Figure 5 shows, in the cases of two DNA strand types of lengths 1kbp and 3kbp both with and without FD, an exhaustive search of rates over the first 25 threshold values for both types. Expectedly, data rates decrease with the introduction of FD, where in both cases the maximum is achieved at thresholds $(5, 8)$ (first coordinate belongs to 1kbp strands) with ~ 6bps and ~ 7bps. The comparatively higher value of optimum threshold for longer DNA strands arises from their increased contribution to ISI due to their slower diffusion rates.

Thus, an educated choice of threshold values is of critical importance. As a first approximation, we consider the average



(a) Minimal detection thresholds $(1, 1, 1)$ **(b)** Guessed detection thresholds

Fig. 4: Capacity vs. # of transmitted DNAs for varying # of nanopores M in case of three types of strands, 1kbp, 3kbp and 5kbp.

capturing rates contributing to detection $\overline{\lambda}$ of each type of strand in the absence of others

$$\overline{\lambda} = \frac{1}{T} \int_0^{T-\tau} \lambda(s) ds.$$

The average time for detection $\overline{t}_D$ is the sum of average times of capture and processing, from which we define average detection rate of a single site as

$$\overline{\lambda}_D = \frac{1}{\overline{t}_D} = \frac{\overline{\lambda}}{1 + \overline{\lambda}\tau}.$$

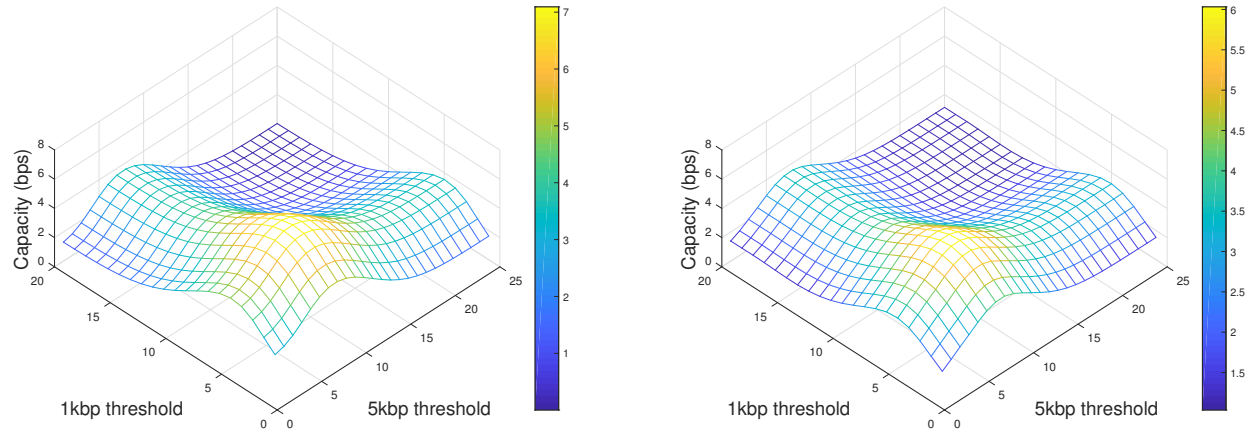
Thus, the expected number of detections by Rx can roughly be approximated by

$$[N_{D,Rx}] \approx M \overline{\lambda}_D T,$$

where M is the number detection sites on Rx. Similar calculations can be carried out with and without ISI mediated concentrations taken into account, and their difference ΔN_D can be thought of as the amount of contribution to detection by ISI. Thus, in case of no FD, we take our threshold values to be $C\Delta N_D$ for each type to diminish the effects of ISI on reception, where C is a positive constant to be chosen. To account for FD, we use the thresholds obtained by multiplying the FD matrix A_{FD} with the vector $\Delta\vec{N}_D$ containing the differences for each type and scaling obtained values by C. For both cases, numerical studies revealed that $C = 3$ yields improved rates. For instance, this guess threshold is $(6, 9)$, very close to the optimum value of $(5, 8)$, for the 2 symbol case depicted in Figure 5. Throughout the rest, unless otherwise stated, we assume these threshold values.

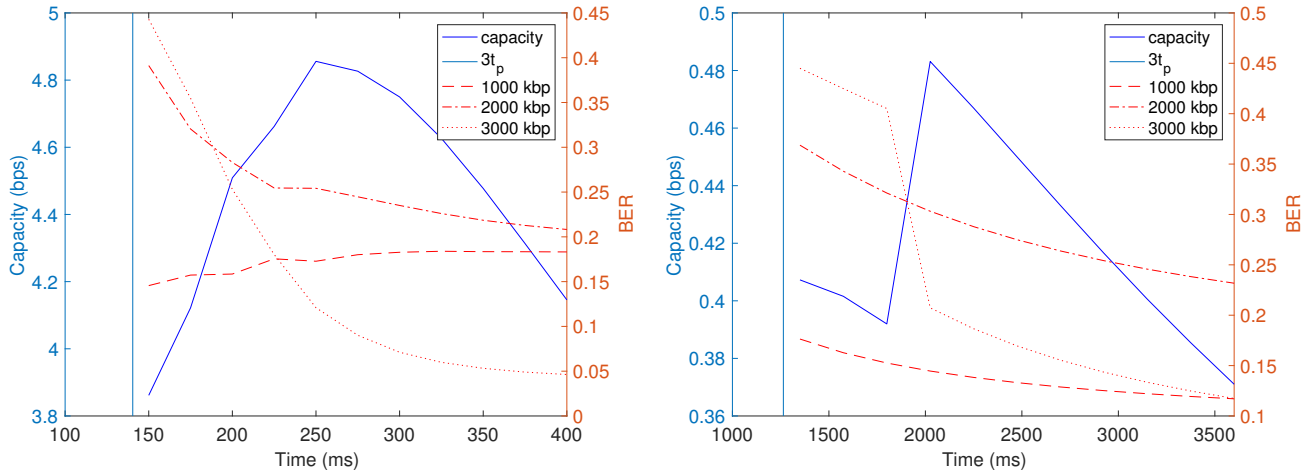
C. Sampling Period and Distance

Upon release of molecules by Tx, dictated by the diffusion process, the peak concentration times at Rx scale with the



(a) No FD **(b)** With FD

Fig. 5: Communication capacities with varying threshold values in case of two types of strands, 1kbp and 3kbp.



(a) d = 1µm **(b)** d = 3µm

Fig. 6: Capacity and BERs as a function of sampling periods at two distances. The vertical line sits at $T = 3t_p$.

square of the distance d as $t_p = d^2/6D$. Thus, accordingly the sampling period T has to be adjusted with distance, and to justify the fact that we are neglecting the effects of ISI from signals prior to previous one, we consider sampling periods with $T \geq 3t_p$ corresponding to the largest t_p, i.e., smallest diffusion coefficient D, which belongs to the longest DNA strand. Figure 6 depicts the dependence of communication capacities and BERs to sampling period for the case of three types of strands at distances 1µm and 3µm.

D. DNA Sizes

Figure 7 shows the capacity of the channel for different number of DNA strands r, where r number of symbols are generated with linear and equal spacing between 1kbp and 5kbp. As noticed, increasing the number of DNA types improved the channel capacity up to a point, and the peak capacity is observed at three DNA types. After this point, spacing between symbols get closer other such that the effect of FD becomes a dominant factor. Hence, the capacity of the channel decreases as the number of DNA types becomes more than two.

Figure 8 presents the capacity results for different DNA size increments N_{inc} and number of DNA strands r, where the lowest DNA strand size is 1kbp. The highest capacity value of ≈ 5.7bps is reached for 2000bp increment with 2 distinct DNA molecules such that 1 and 3kbp DNA lengths are utilized in the MC channel. For lower increment values, the capacity values are decreased because low increment values have higher FD probabilities. In addition, increasing the number of DNA strands also causes a decrease in the capacity as additional DNA strands having higher number of bp suffer from lower diffusion constants, which increase the level of ISI in the channel compared to lower level increments. At the end, we can conclude that DNA-based MC protocol is able to achieve high capacity levels close to 5.7bps as shown in Figure 8.

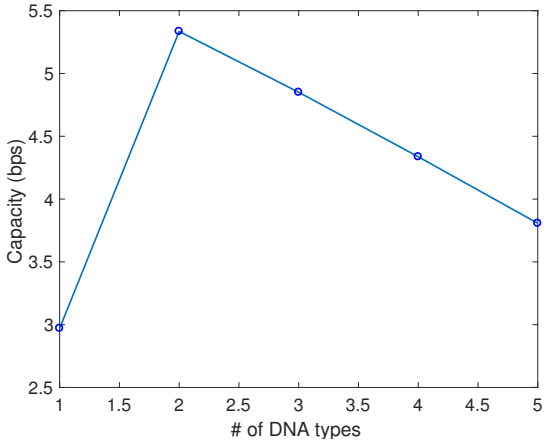


Fig. 7: Capacity vs. # of DNA types for fixed DNA range (1-5kbp).

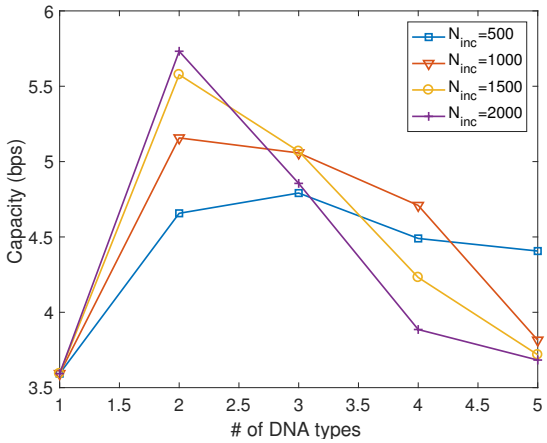


Fig. 8: Capacity vs. # of DNA types for fixed increments N_{inc}.

V. Conclusions and Future Work

In this paper, we propose a novel DNA-based MC protocol for achieving high capacity communication between nanoma-

chines up to 6bps at 1μm communication range. A powerful analytical model is introduced to analyze diffusion, capturing, detection and reception processes in DNA-based communication system, where Tx is capable of emitting DNA strands having different lengths as information carrying molecules, and Rx contains receptor nanopores to selectively detect DNA strands. The proposed model is capable of including processing times of the information molecules and adaptive threshold values are calculated based on the diffusion dynamics of different DNA lengths. In the proposed model, there are multiple parameters, e.g., the number of molecules released, the number of DNA lengths, the DNA length increments, the number of pores, and sampling period, that can be optimized according to desired communication range. The optimization of these parameters stands as an important open research problem in this field. In addition, the proposed analytical model is based on information carrying DNA molecules with different lengths, but these analysis can be easily extended to MC channels using DNA strands having dumbbell hairpins, and short sequence motifs/labels. In this way, the number of DNA strands can be further increased in order to further improve the capacity of DNA-based MC.

ACKNOWLEDGMENT

This work was supported in part by ERC project MINERVA (ERC-2013-CoG #616922).

REFERENCES

[1] O. B. Akan, H. Ramezani, T. Khan, N. A. Abbasi, and M. Kuscu, "Fundamentals of molecular information and communication science," *Proceedings of the IEEE*, vol. 105, no. 2, pp. 306–318, 2017.

[2] N. Farsad, W. Guo, and A. W. Eckford, "Tabletop molecular communication: Text messages through chemical signals," *PLoS ONE*, vol. 8, no. 12, 2013.

[3] I. F. Akyildiz, M. Pierobon, S. Balasubramaniam, and Y. Koucheryavy, "The internet of Bio-Nano things," *IEEE Communications Magazine*, vol. 53, no. 3, pp. 32–40, 2015.

[4] B. A. Bilgin and O. B. Akan, "A Fast Algorithm for Analysis of Molecular Communication in Artificial Synapse," *IEEE Transactions on Nanobioscience*, vol. 16, no. 6, pp. 408–417, 2017.

[5] E. Dinc and O. B. Akan, "Theoretical limits on multiuser molecular communication in internet of nano-bio things," *IEEE transactions on nanobioscience*, vol. 16, no. 4, pp. 266–270, 2017.

[6] B. Atakan and O. B. Akan, "On channel capacity and error compensation in molecular communication," in *Transactions on computational systems biology X*. Springer, 2008, pp. 59–80.

[7] M. Pierobon and I. F. Akyildiz, "Capacity of a diffusion-based molecular communication system with channel memory and molecular noise," *IEEE Transactions on Information Theory*, vol. 59, no. 2, pp. 942–954, 2013.

[8] N. A. Bell and U. F. Keyser, "Digitally encoded DNA nanostructures for multiplexed, single-molecule protein sensing with nanopores," *Nature Nanotechnology*, vol. 11, no. 7, pp. 645–651, 2016.

[9] K. Chen, M. Juhasz, F. Gularek, E. Weinhold, Y. Tian, U. F. Keyser, and N. A. Bell, "Ionic Current-Based Mapping of Short Sequence Motifs in Single DNA Molecules Using Solid-State Nanopores," *Nano Letters*, vol. 17, no. 9, pp. 5199–5205, 2017.

[10] N. A. Bell, M. Muthukumar, and U. F. Keyser, "Translocation frequency of double-stranded DNA through a solid-state nanopore," *Physical Review E*, vol. 93, no. 2, 2016.

[11] N. A. Bell and U. F. Keyser, "Direct measurements reveal non-markovian fluctuations of dna threading through a solid-state nanopore," *arXiv preprint arXiv:1607.04612*, 2016.

[12] S. Hernández-Ainsa and U. F. Keyser, "DNA origami nanopores: Developments, challenges and perspectives," *Nanoscale*, vol. 6, no. 23, pp. 14 121–14 132, 2014.

[13] M. S. Kuran, H. B. Yilmaz, T. Tugcu, and I. F. Akyildiz, "Modulation Techniques for Communication via Diffusion in Nanonetworks," in *2011 IEEE International Conference on Communications (ICC)*, 2011, pp. 1–5. [Online]. Available: http://ieeexplore.ieee.org/document/5962989/

[14] N. R. Kim and C. B. Chae, "Novel modulation techniques using isomers as messenger molecules for nano communication networks via diffusion," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 12, pp. 847–856, 2013.

[15] S. S. Sorlie and R. Pecora, "A Dynamic Light Scattering Study of Four DNA Restriction Fragments," *Macromolecules*, vol. 23, no. 2, pp. 487–497, 1990.

[16] R. M. Robertson, S. Laib, and D. E. Smith, "Diffusion of isolated DNA molecules: Dependence on length and topology," *Proceedings of the National Academy of Sciences*, vol. 103, no. 19, pp. 7310–7314, 2006. [Online]. Available: http://www.pnas.org/cgi/doi/10.1073/pnas.0601903103

[17] A. Meller, "Dynamics of polynucleotide transport through nanometre-scale pores," *Journal of physics: condensed matter*, vol. 15, no. 17, p. R581, 2003.

[18] A. Meller and D. Branton, "Single molecule measurements of DNA transport through a nanopore," *Electrophoresis*, vol. 23, no. 16, pp. 2583–2591, 2002.

[19] A. Meller, L. Nivon, and D. Branton, "Voltage-driven DNA translocations through a nanopore," *Physical Review Letters*, vol. 86, no. 15, pp. 3435–3438, 2001.

[20] C. Y. Kong and M. Muthukumar, "Modeling of polynucleotide translocation through protein pores and nanotubes," *Electrophoresis*, vol. 23, no. 16, pp. 2697–2703, 2002.

[21] A. J. Storm, C. Storm, J. Chen, H. Zandbergen, J. F. Joanny, and C. Dekker, "Fast DNA translocation through a solid-state nanopore," *Nano Letters*, vol. 5, no. 7, pp. 1193–1197, 2005.