1 Introduction

Historically, communications engineers have dealt with electromagnetic forms of communication: in wireline communication, electric fields move currents down a wire; in wireless communication, electromagnetic waves in the radio-frequency spectrum propagate through free space; in fiber-optic communication, electromagnetic radiation in the visible spectrum passes through glass fibers.

However, this book is concerned with an entirely different form of communication: *molecular communication*, in which messages are carried in patterns of molecules. As we shall see in this book, molecular communication systems come in many forms. For example, message-bearing molecules may propagate through a liquid medium via simple Brownian motion, or they may be carried by molecular motors; the message may be conveyed in the number and timing of indistinct molecules, or the message may be inscribed directly on the molecule (like DNA); the nanoscale properties of individual molecules may be important, or only their macroscale properties (like concentration).

Molecular communication is literally all around us: it is the primary method of communication among microorganisms, including the cells in the human body. In spite of its importance, only in the past decade has molecular communication been studied in the engineering literature. In writing this book, our goal is to introduce molecular communication to the wider community of communications engineers, and collect all the current knowledge in the field into a single reference for the sake of researchers who want to break into this exciting field.

1.1 Molecular communication: Why, what, and how?

1.1.1 Why molecular communication?

Why would engineers want to design a system involving molecular communication? To motivate this question, suppose you are given the following design problem. Your goal is to perform *targeted drug delivery*: to deliver drugs within the human body exactly where they are needed (for example, directly to malignant tumors within the body, as chemotherapy). To accomplish this goal, you have decided to use *thousands of tiny, blood-cell-sized robots* that must cooperate with each other to autonomously navigate through the body, identify tumors, and release their drugs to destroy the tumor. To

cooperate, the robots must be able to communicate – so how would you design the communication system?

This is a challenging question: as a result of their size, the devices have very small energy reserves, and must glean whatever energy they can from the environment. The devices must also operate in the body without disrupting healthy tissues, or being destroyed by the immune system prior to completing the mission. These features are consistent with the communication challenge faced by microorganisms, and these organisms have solved the problem by exchanging signals composed of molecules – that is, *molecular communication.*

As a result, for engineered systems, molecular communication is a *biologically inspired* solution to the communication problem. This communication could be engineered in two ways: first, an entirely artificial device could be designed to communicate using signaling molecules; and second, the existing molecular communication capabilities of an engineered microorganism (e.g., a bacteria with custom DNA) could be used. Remarkably, both nanoscale robots [1] and artificial bacteria [2] are within the capabilities of contemporary technology. However, nanoscale communication techniques, such as molecular communication, are needed to permit cooperation and unlock the disruptive potential of these systems.

1.1.2 What uses molecular communication?

In the previous section, we gave an example of tiny robots swimming through the human bloodstream. This example opens us up to *biological nanomachines*, or *bio-nanomachines*, one of the primary application areas of molecular communication.

For our purposes, bio-nanomachines may be defined as follows:

- *Materials.* A bio-nanomachine is made of biological materials (e.g., protein, nucleic acid, liposome, biological cell), or a hybrid of biological and non-biological materials.
- *Size.* A bio-nanomachine's size ranges from the size of a macromolecule to the typical size of a biological cell ($\sim$100 μm).[1]
- *Functionality.* A bio-nanomachine's functionality is limited to simple computation (e.g., integrating two types of input signals to produce one output signal), simple sensing (e.g., sensing only one or two types of molecule), and simple actuation (e.g., producing simple mechanical motion).

Figure 1.1 gives an overview of a molecular communication system involving bio-nanomachines [3]. In molecular communication, information is encoded onto (and decoded from) molecules, rather than electrons or electromagnetic waves. First, an information source generates information to encode onto molecules and triggers a group of sender bio-nanomachines to start propagation of information-encoded molecules.

[1] The term "nano" sometimes refers to dimensions of 1–100 nm, which is included in this definition; however, biological cells are typically much larger. Recently, the term *mesoscopic* has been used to describe dimensions that span from atomic to microbiological scales.

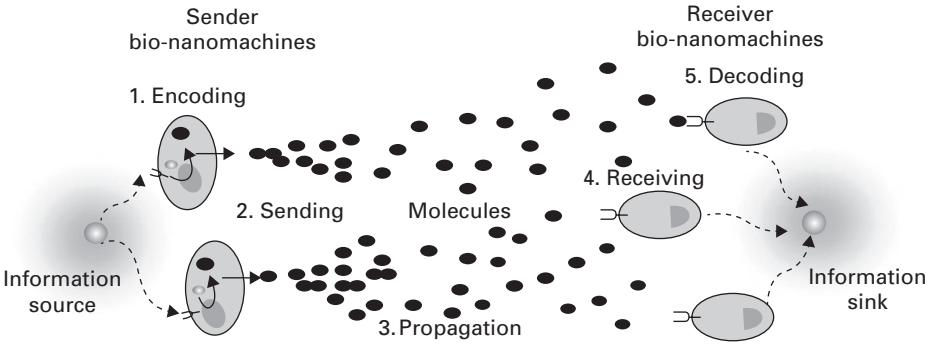

Figure 1.1 An outline of a molecular communication system incorporating bio-nanomachines [3].

Information-encoded molecules then propagate in the environment, and are detected by a group of receiver bio-nanomachines. Receiver bio-nanomachines may forward incoming molecules to next-hop bio-nanomachines or may pass them to an information sink for decoding information. We discuss this process in greater detail in Chapter 4.

Bio-nanomachines are not the only application for molecular communication – however, they are in many ways the primary motivating application, and the one that informs most of the analysis throughout this book. We give an introduction to applications in Section 1.3, and some detailed examples in Chapter 8.

1.1.3 How does it work? A quick introduction

How does molecular communication work? We spend the rest of this book answering this question, but here we give the reader a quick overview, and introduce the basic issues related to designing a molecular communication system.

First, we should be clear what we mean by "communication." We focus on artificial communication, where a manmade *message* needs to be conveyed from one point to another. A message can be discrete (like a sequence of bits, as in an IP packet), or continuous (like an analog waveform, as in AM radio), but for now, we will assume that the message is discrete. In the simplest form of communication, there are two terminals: a *transmitter*, which sends the message; and a *receiver*, which receives the message. (So far, this is general enough to include any point-to-point communication system, not just molecular communication. This setup can be generalized: in a network setting, there may be many senders and receivers, and a terminal can be both a sender and a receiver for different messages.)

To communicate, the transmitter makes a physical change to its environment, and that change must be measurable at the receiver. Again, this is true of any communication system: for instance, a wireless transmitter induces a changing EM field along an antenna, which can be detected in an antenna at the receiver. However, in molecular communication, the change must be molecular: the transmitter releases molecules into a shared medium, which propagate to (and are detected by) the receiver.

In order to convey distinct messages, each possible message is associated with a molecular *signal*: a unique pattern of molecules for each possible message, which can be distinguished at the receiver. Further, there must be a way for the receiver to *decide* which message was sent, based on the signal that it measures. For instance, say we want to send a message consisting of a single bit, 0 or 1. We can do this in many ways, but here are three possibilities:

- *Signaling with quantity.* Say we have $n > 0$ molecules available at the transmitter. We could send a 0 by releasing zero molecules, or a 1 by releasing n molecules. If the receiver observes 0 molecules, it can conclude that a 0 was sent; if it observes at least one molecule, it can conclude that a 1 was sent.
- *Signaling with identity.* Say we have two types of molecule available at the transmitter, A and B (where the receiver can distinguish A from B). We could send a 0 by releasing molecule A, or a 1 by releasing molecule B. The receiver would decide 0 or 1 if it observed A or B, respectively.
- *Signaling with timing.* Say we have a single molecule available at the transmitter. We could send a 0 by releasing that molecule right now, or we could send a 1 by waiting $t > 0$ seconds before releasing the molecule. The receiver would then decide whether 0 or 1 was sent by measuring the arrival time of the molecule.

This simple example, illustrated in Figure 1.2, encapsulates many of the general techniques that we will describe throughout the book. For example, generalizing a quantity signal, we can manipulate the concentrations of molecules in the medium. We may also wonder how to generate molecular signals; we will see throughout this book that many

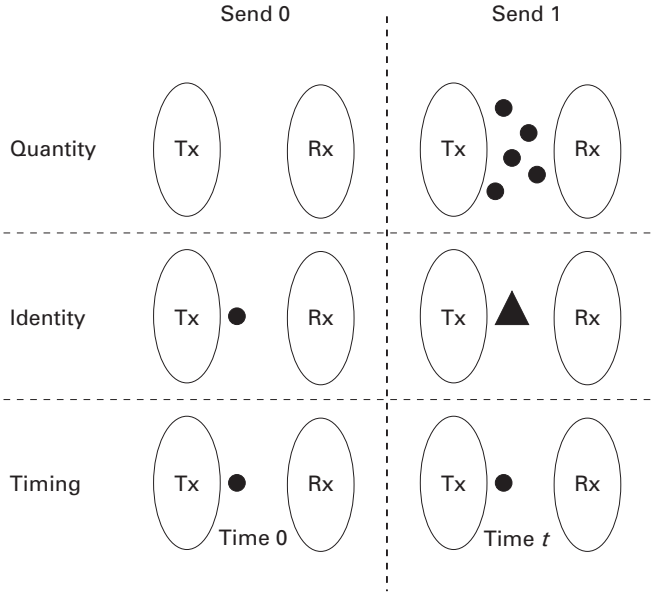

Figure 1.2 Illustration of three simple ways of generating a binary molecular signal.

biological "components" exist to emit and receive message-bearing molecules. As a result, molecular communication systems are often biologically based.

We also see that the propagation of molecules from transmitter to receiver must take place via diffusion: this could be viewed as either discrete Brownian motion, for small numbers of molecules; or continuous diffusion, for large numbers of molecules. Later, we will see that diffusion is a significant source of distortion and constraint on molecular communication systems: for instance, discrete Brownian motion might mean that message-bearing molecules are lost, or that they take an arbitrarily long time to arrive; further, continuous diffusion is a very slow process, which limits the possible rate of information transfer.

Figure 1.3 shows an example of molecular communication in the laboratory. A sender cell is stimulated at time $t = 0$, and encodes a molecular signal, using inositol trisphosphate (IP_3) and adenosine triphosphate (ATP). Here, information about the stimuli is encoded in the *type* and *number* (i.e., the concentration) of molecules. The sender cell broadcasts the molecular signal into the environment, through external pathways (in the extracellular space) or internal pathways (gap junction channels). The molecular signals diffuse through the two pathways, and receiver cells in the environment detect the molecular signals using receptors. The receiver cells then increase the concentration of intracellular molecules (e.g., calcium), and decode the signals using molecular mechanisms inside the cells. More detail of this form of molecular communication is provided in Chapter 3, as well as many other forms of molecular communication found in biological systems.

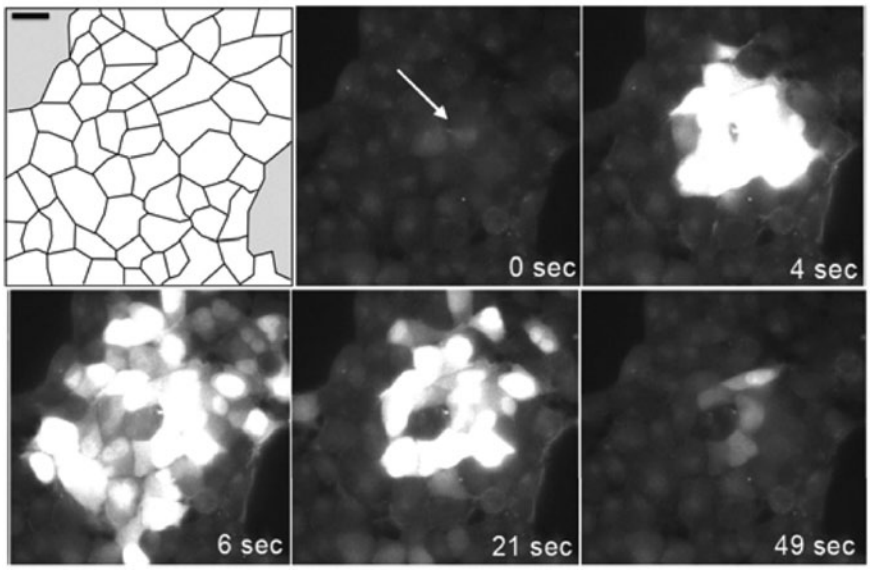

Figure 1.3 Series of images from a lab experiment. A sender cell, upon stimulation, broadcasts molecular signals and the receiver cells in the environment respond to the molecular signals [4].

1.2 A history of molecular communication

The field of nanotechnology is commonly traced back to the Nobel-laureate physicist Richard Feynman, and his famous 1959 lecture to the American Physical Society, entitled *There's Plenty of Room at the Bottom* (transcribed in [5]). Feynman argued that the laws of physics permit very small devices, far smaller than contemporary technology had managed to produce. Since 1959, Feynman's vision of extreme miniaturization has been realized in many fields, such as integrated circuitry and microscopy. Moreover, new fields of research, such as micro- and nano-electro-mechanical systems (MEMS and NEMS), were spawned to extend this miniaturization into robotics.

Meanwhile, it has long been recognized that microorganisms, including cells and bacteria, gain information from their environment by gathering chemical messengers sent by their neighbors. A simple example is quorum sensing [6], in which bacteria send molecular messages to one another in order to estimate the local population of their species; the bacteria can take action based on this estimate, such as forming a colony or seeking out larger numbers of their species. Further, the means by which cells send messages to one another and control each others' behavior is a well-studied area of biology known as *cell signaling* (see, e.g., [7]).

The engineering aspects of molecular communication have a research background that stretches back decades. In this section, we give a brief review of this field's history. We begin with a review of the (mostly theoretical) work done by early communications researchers. We then discuss more recent theoretical and implementational work, and conclude with a short review of contemporary research in this field.

1.2.1 Early history and theoretical research

Work by early researchers, such as Shannon [8] and Nyquist [9], established information theory and communication theory as mathematical disciplines. The focus was on telegraphic communication, so these theories developed (and remain) largely as subfields of electrical engineering. As abstract models, these techniques can be used in more general studies of communication, such as molecular or biological communication. However, this direction of research has remained on the fringes of information theory until recently, perhaps because Shannon himself discouraged it [10].[2]

Nonetheless, there has long been interest in information theory as a tool for explaining biological behavior, especially in terms of biomolecular interactions. To the knowledge of the authors, the first discussion of information theory in the context of biomolecular interactions occurred in [11], which analyzed the efficiency of the kidney by recognizing its operation in terms of information processing: the kidney examines molecules and makes decisions on them, either keeping them in the bloodstream or

[2] Shannon's point was not that chemistry or biology are inherently inappropriate applications for information theory, it was that the reputation of a rapidly growing field depends on scientific rigor and high-quality work. At the time, such work was found in electrical applications. Reference [10] is certainly worth re-reading, and its lessons worth remembering, as our field of molecular communication appears poised for rapid growth.

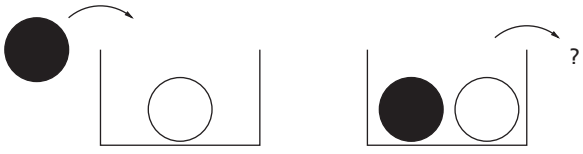

Figure 1.4 Illustration of Blackwell's "chemical channel." In the first figure, a black ball is introduced to the bag, which already contains a white ball. In the second figure, one of the two balls in the bag is selected at random and removed to form the channel output.

rejecting them as waste, in an operation reminiscent of Maxwell's demon. The key observation was that gathering molecular information has a minimum energy cost, so information processing explained the kidney's energy consumption. This work was extended by Berger [12], who showed that molecular energy efficiency could be explicitly described in terms of rate distortion theory. This result was cited as "visionary" in a book review [13].

Meanwhile, Blackwell [14] described a highly abstract channel model, where successive channel outputs are statistically dependent. This model was called the *chemical channel* by subsequent authors,[3] making it possibly the first molecular communication channel model to appear in the literature. In this model, colored balls are used to communicate: there are two colors, white and black, and the balls are otherwise identical. At the beginning of the communication session, a bag is filled with a given number of balls of unknown colors. Communication then proceeds as follows: first, the transmitter chooses a color and drops a new ball of that color into the bag, then the receiver selects a ball from the bag at random, removes it, and notes its color; this process is repeated as often as necessary to send a message.

EXAMPLE 1.1 Consider the chemical channel in Figure 1.4, where the bag initially contains one ball. You can send one bit of information for every three balls by using a *repetition code*: inserting three white balls in a row, or three black balls in a row. The receiver can tell what color the transmitter sent by picking the majority of colors out of every group of three: at most one ball will be the wrong color. This is not as good as you can do, however; capacity of the trapdoor channel is an open problem.

If the bag in this example contains many balls, then the random selection is a coarse analog to random diffusion. For instance, say we have molecules instead of balls: the transmitter inserts molecules into the channel, which diffuse randomly in the medium, and are ultimately removed by the receiver. If the molecules are perfectly mixed after each insertion, then we have something like this channel. Berger elaborated on these ideas, showing how they can be used to describe biological molecular communication in his Shannon prize lecture [16].

In its standard form, the trapdoor channel is a poor approximation to diffusion: the assumption of perfect mixing between insertions is not practical. The model can

[3] Early drafts of [15], available on arXiv, credit Thomas Cover with coining the term "chemical channel," though the claim is missing from the published version. The term "trapdoor channel" is also used.

be refined; for example, each ball can have a different probability of being selected. However, it is worth remembering that the trapdoor channel was not originally intended to model diffusion; the diffusion application came later. More recently, researchers have examined *diffusion-mediated* models that explicitly view molecular diffusion as a communication system.

Diffusion can be viewed microscopically, as a process involving individual molecules, or macroscopically, as a process involving continuous concentrations. The latter approach has the advantage of being linear: the (continuous) diffusion equation is a linear partial differential equation, so the considerable body of linear system theory for communication systems can be applied. Early work in this direction emerged from the biological literature: in [17], information theory was used to present chemical signal transduction in the retina as a communication system (to the authors' knowledge, the first explicit use of information theory in chemical signaling).[4] Building on these results, [20] simulated and analyzed a detailed linear model of a diffusion-mediated cellular transduction system, evaluating its frequency response and its information-theoretic capacity.

1.2.2 More recent theoretical research

The past five years have seen a rapid increase in information-theoretic analysis of molecular communication. The general information-theoretic model of communication is broad enough to include new methods of information transfer, including molecular communication (and we describe this general model in Chapter 6). For molecular communication, the challenge is to develop information-theoretic equivalents for the components of the model, such as the transmitter, the receiver, and the channel.

Discrete Brownian motion, modeled as a communication system, focuses on idealized models and the ultimate limits of molecular communication. This is because continuous diffusion is merely the limiting process of discrete Brownian motion, as the number of molecules becomes large. Thus, if we can find the limits of discrete Brownian motion, we have the best that can be done with molecular communication. The first work on discrete diffusion was [21], in which some "ideal" modeling assumptions were made, and the primary source of distortion in the channel was assumed to be the random propagation time of message-bearing molecules from transmitter to receiver.

It is important to note that discrete diffusion systems require processing that is far beyond contemporary technology: for one thing, these systems require sensing and manipulation of individual molecules; for another, they often assume synchronization between transmitter and receiver. However, as research into the ultimate limits of molecular communication, it is natural to consider these systems in terms of information theory.

Theoretical work has been done in other directions as well: continuous diffusion, considering the propagation of concentrations of molecules, is less efficient than discrete

[4] Information theory has been used to analyze neural coding for over fifty years, e.g. [18, 19], but not explicitly to analyze a chemical communication system.

diffusion, but feasible to implement in practice: components exist that can detect and respond to changes in concentration of a given molecular species. The capacity of such systems was considered in [22]. Biomorphic systems, as the focus of implementational work on molecular communication, are natural to analyze with information and communication theory. The aforementioned work of [20] is an example of this type of after analysis; another early work is [23], which analyzed a ligand–receptor system in discrete time.

1.2.3 Implementational aspects

The term "molecular communication," meaning an engineered communication system where messages are conveyed in patterns of molecules, was coined in the title of a 2005 paper [24]. That paper, focusing on the possible designs and uses of diffusion-based communication systems, launched a body of research on the implementation of molecular communication. These works described a variety of biological or chemical components that could be used to assemble practical systems to conduct molecular communication: in other words, this work explores the "hardware" that would form the communication system. In many cases, laboratory experiments have been performed to show the feasibility of molecular communication, or to describe potential applications.

Various subsystems for communication have been identified. As one example, the gap junction, used by cells to exchange ions, could be used by collections of cells to pass concentrations of ions. If this were done under external control, a message could be passed from one side of the collection to the other [25, 26]. As another example, liposomes (i.e., spherical vesicles that act as "packages" of molecules) can be used to exchange messages: information-bearing molecules can be encapsulated into a liposome, and passed to communication partners. This possibility was explored by [27, 28], and its feasibility was demonstrated in lab experiments in [29].

The practical problem of transporting molecules from transmitter to receiver has also been explored. Though random diffusion is one solution to this problem, there are alternatives: for example, molecular motors are used in living cells to transport molecules from one place to another. In molecular communication, motors may be used to collect message-bearing molecules (or packages of molecules, e.g., in liposomes) and transport them from transmitter to receiver [30]. Experiments validating this approach were presented in [31].

We describe some of these components and implementations in detail later in this book. An excellent review of contemporary research in implementational aspects of molecular communication is found in [32].

1.2.4 Contemporary research

Work on molecular communication has accelerated in the last five years, thanks in part to a new focus on nanoscale communication networks, or *nanonetworks* [33, 34]. Nanonetworks involve collections of very small devices that communicate and cooperate with each other, and in which essential features of the network have nanoscale dimensions. For example, swarms of nanorobots, which may be used in some of the

applications described earlier in this chapter, may form a nanonetwork to accomplish their task. Molecular communication has recently been recognized as an enabling technology for nanonetworking [35].

At the time of writing, molecular communication is increasingly popular among traditional communication engineers. As the background of these researchers is primarily theoretical and simulation-based, there has been a rapid increase in theoretical and simulation-based analysis of molecular communication. Without attempting to be comprehensive, we give four major themes of contemporary research:

- Channel modeling and noise analysis are key directions of research. Traditional communication and information theory are based on a set of mathematically precise channel models, such as the additive white Gaussian noise channel. Moreover, within each such channel, there exists a source of distortion, or "noise." However, no widely accepted general channel or noise model exists for molecular communication; depending on the scenario, it is likely that several different channel models are required. Adding to historical work on channel modeling, recent results include [36], which developed a complete end-to-end model of molecular communication based on continuous diffusion, and [37], which modeled the noise of an active-transport molecular communication system.
- The information-theoretic capacity of molecular communication, or the maximum rate at which data can be reliably transmitted, is an important open problem. The fully general problem of finding capacity is known to be difficult, but many recent papers have sought either bounds on capacity or the capacity in simplified scenarios, such as: [38], which considered continuous diffusion, simplifying the concentration to a binary variable (taking values of "high" and "low" concentration); [39], which found bounds on capacity for a general model of discrete diffusion; and [40], which considered a similar setup with generalized transmission schemes and possible molecular losses. Another direction is described in [41], which examines the symmetries in possible capacity-achieving input strategies, and bounds the general channel capacity.
- From the simulation side, system design for molecular communication is an important research topic. Compared to traditional communication, molecular communication seems less amenable to closed-form analysis and optimization; as a result, simulations are a key tool for determining the performance of the system. (Obviously, laboratory experiments are the most accurate way of determining performance, but these are significantly more expensive and difficult to perform than simulations.) A wide variety of design work has been done in simulation, such as optimization of distance-estimation techniques [42], design of channel shapes for microfluidic molecular communication [43], design of routing schemes in networks [44], and design of signaling techniques [45]. The design and analysis of simulation techniques themselves are an important open problem, and some papers are devoted entirely to that topic (e.g., [46]).
- System-level research has also attracted much recent attention. The problem described at the beginning of this chapter – operation of bio-nanomachines in the human body – was reviewed in [47], and major challenges identified. Molecular