



NSF Workshop on Molecular Communication/ Biological Communications Technology

February 20-21, 2008

Hilton Arlington, Virginia

Organizers: Tadashi Nakano¹ and Tatsuya Suda^{1,2}

¹University of California, Irvine

²National Science Foundation

<http://netresearch.ics.uci.edu/mc/nsfws08>

NSF Workshop on Molecular Communication/ Biological Communications Technology

Tadashi Nakano¹ and Tatsuya Suda^{1,2}

¹University of California, Irvine

²National Science Foundation

<http://netresearch.ics.uci.edu/mc/nsfws08>

Abstract: The workshop on Molecular Communication/Biological Communications Technology (February 20-21, 2008) has provided a forum to discuss the emerging interdisciplinary field of Biological Communications Technology. The workshop has focused on various key themes at the intersection of biology and computing/communications technology. The workshop has in particular focused on the following five topics; (1) component technology, (2) system design methodology, (3) coding theory, (4) novel computing, and (5) applications, and identified grand research challenges for further advancement of this field. This report describes in detail grand research challenges identified during the two day workshop. This report also provides a summary of suggested actions for NSF in order to promote and support interdisciplinary education and research.

Table of Contents

1. Executive Summary	1
2. Grand Research Challenges	4
2.1 Communication Components and Systems	4
2.2 System Design Methodologies	8
2.3 Coding Theory and Channel Capacity	13
2.4 Novel Computing Machines and Paradigms	24
2.5 Biological Communications and its Potential Applications	31
3. Suggestions to NSF	34
3.1 Promoting Interdisciplinary Education	34
3.2 Supporting Interdisciplinary Research	35
4. References	36
Appendix A: Workshop Organization	45
Appendix B: Workshop Agenda	46
Appendix C: Participant List	48

1. Executive Summary

The workshop on Molecular Communication/Biological Communications Technology was held at the Hilton Arlington, Arlington, VA (February 20-21, 2008). The workshop provided a forum to discuss an emerging interdisciplinary field of Biological Communications Technology. The workshop brought together 36 leading researchers from Biology, Chemistry, Physics, Mathematics, Nanotechnology, Computer Science, and Engineering and discussed key research issues, the current state of the art, and future directions of this important field and its related areas, and identified grand research challenges through discussions and brainstorming. This

workshop also discussed to provide suggestions to NSF on what directions to pursue to establish a new integrated and transformative science in the above mentioned areas; Biology, Chemistry, Physics, Mathematics, Nanotechnology, Computer Science, and Engineering.

The central theme of the workshop is biology and computing/communications technology, ultimately to create and establish a new integrated and transformative science. The specific goals of this area are to understand biological computing and communication processes and to apply the obtained insights to create biological-material-based or biologically-inspired computing and communication systems that are innovative and move beyond incremental and evolutionary technological advances.

In order to achieve the goals of this area, researchers must address fundamental research issues related to computing and communication as follows.

Objective 1: Understand biological computing and communication processes through experimentation, simulation, modeling, analysis, synthesis and verification.

There are a number of questions that need to be answered to understand biological systems – e.g., how do biological entities represent or code information? What materials are used and what are the physical properties that encode information? How do we detect and measure the information encoded in biological systems? How do biological entities deal with environmental noise and errors in information representation? What makes biological information robust? How do biological systems compute? What is the architecture of biological computing? What are the design principles of biological computing? How do we measure the computing capacity of biological entities? How are biological computing systems controlled, reset and fine tuned? How do biological systems communicate? What is the architecture of biological communication? How do biological communication systems scale? Are biological communication systems modular? How does redundancy in biological systems help achieve robustness of biological communication? How is information transmitted and directed in a biological system? How does global behavior of a biological system emerge from local communication and interaction of biological entities?

To answer these questions, we need to perform experimentation, simulation, and theoretical modeling and analysis of biological computing and communication processes. Key disciplines include information and communication theory (to understand how biological entities encode, receive, digest, process, decode information, and how they utilize environmental noise in such processes), computational/synthetic/systems biology and computational neuroscience (to understand, through experiments, simulation and theoretical modeling and analysis, how biological entities encode information, and how they compute and communicate), complex systems (to understand how biological entities exhibits useful emergent behaviors), and network systems (to understand biological computing and communication processes as a network system). This also involves developing techniques and tools, as well as creating infrastructure to share biological data, to help understand biological computing and communication processes. Related declines include bioinformatics (to develop techniques to search, integrate and model on the basis of large-scale, heterogeneous biological databases), computational biology (to develop multi-scale simulators to simulate biological computing and communication processes), network systems theory (to create large scale simulation models of and tools for dynamic networks within biological systems).

Objective 2: Develop innovative techniques and useful tools to understand and apply the basic principles underlying biological computing and biological communication.

This involves creating computer scientific representation of biological computing and communication processes; Creating new models, languages and common representation (algorithmic, algebraic, and formal logic representation) to describe biological entities and

systems; developing descriptions of biological systems using concepts from networks such as protocols, layers, and architectures; using the computer scientific representation of biological systems to understand biological systems and predict their behaviors; developing techniques to model and analyze large-scale and heterogeneous biological data; developing innovative applications to validate the basic principles of biological computing and biological communication.

Existing disciplines may be applicable including algorithmic, algebraic and formal logic methods (to describe biological computing and communication processes, as well as process of biological emergent behavior), computing theory (to describe biological computing and communication processes using, for instance, process algebras (pi-calculus) and state-machines, to use model checking techniques to verify invariants of such processes), and network systems theory (to describe biological computing and communication processes using concepts from networks such as protocols, layers, and architectures).

Objective 3: Apply biological insights, and design and construct new computing and network systems.

Specific goals in this category are recreating and duplicating biological systems to construct new computing and network systems from biological materials; designing biological components and interfaces, and integrate them into a system; Designing and constructing new bio-inspired computing and network systems; developing techniques to control highly dynamic and large scale bio-inspired computing and network systems; developing methodology to evaluate bio-inspired computing and network systems.

Applicable disciplines include coding theory (to design novel coding schemes for next generation communications), complex system design (to design large complex systems from biological inspirations), system evaluation (to evaluate desired features of bio inspired systems such as scalability, robustness, adaptability, self-organization, through empirical study, simulation and theoretical modeling and analysis), and control theory (to control highly dynamic and large scale bio-inspired-networks). Important disciplines also include nano-scale biological component technology and synthetic biology [to design and create components (e.g., computing gates using DNA molecules, signal amplifier using cells) of biological computing and communication systems, to design and create biological nanomachines and nano robots], computing, communication and network systems theory (to integrate various biological components into a coherent computing, communication and network system architecture). Example disciplines also include synthetic biology (to develop biocompatible systems for medical applications) and unconventional computing (to use biological computing and communication systems and solve unconventional computing problems such as solving maze using “mold” like computing/communication elements).

The workshop has formed five sessions, (1) **system component technology session** to discuss necessary components for biological computing and communication systems (what are necessary system components and their functions? What are the characteristics and capacities that systems components may possess? How are components designed and engineered?); (2) **system design methodology session** to discuss system level design issues (how are independent components are interfaced to fully function as an integrated system? how are robustness and stability achieved in the presence of noise? How does a generic architecture look like?); (3) **coding theory and channel capacity session** to discuss information theory for biological computing and communication systems (is Shannon Information Theory applicable to biological systems? how are encoding and decoding done in biological systems? how are information processing and channel capacities quantified? how does noise affect the capacities?); (4) **novel computing session** to discuss unconventional computing and communication (how are problems solved in nature? what biological materials are applicable to

problem-solving? what are applicable domains of new ways of computing and why not silicon-technology?); (5) **potential applications session** to discuss novel applications enabled by biological computing and communications technology (what new applications may potentially emerge from this area? what common and standard framework are necessary for facilitating applications design and development? what engineering technology is in need to develop applications?). With the specific focuses, all the five sessions together discussed the above-mentioned research issues to a great extent.

During the workshop, each session has discussed and identified specific grand research challenges. Here follows a summary of grand research challenges identified at each session.

- The grand research challenge from the system component technology session is to develop and deploy networks of large number of components that are capable of operating in multiple modalities and at multiple spatial and time scales, and that can be interfaced with biological systems at the cellular and molecular levels. Networking requires effective communications and new technology (e.g., biochemical communication) may need to be devised to satisfy constraints such as biocompatibility. Key communication components that need to be developed may include encoders, decoders, transmitters, receivers, amplifiers, and power sources.
- The grand research challenges from the system design methodologies session are to address system level design issues of biological systems as a complex system and to develop design paradigms that lead to the rational organization of elements into highly interactive functional collectives. This involves identifying, understanding and applying biological architectures for networking at the extremes of density, scale, and stochasticity.
- The grand research challenges from the coding and channel capacity session lie in the following three subareas; (1) understanding how to use information theory to learn about biology, in particular as it relates to the different scales of the system, from molecules to behavior, (2) how to use biology to learn more information theory and (3) how to apply these ideas to bioinformatics, engineering systems and technologies at the macro-, micro- and nanoscales.
- The grand research challenge from the novel computing session is to develop novel paradigms and design principles through experimental exploration to use living matter as a computing and communication media. A particularly important application of living matter as a computing and communication media is artificial morphogenesis for computing and communication, which grows into a complex three dimensional system.
- The grand research challenge from the applications session is to understand dynamics and regulation of biological communications, and based on such understanding, to create systems of practical applications.

2. Grand Research Challenges

2.1 Communication Components and Systems¹

This session was charged with discussing communication system components. We felt that we needed a system and application context, and enlarged the scope of the session accordingly. We identified a grand challenge, dubbed “Communication Components and Systems for Biology”, which raises most of the component-related issues we discussed. It involves the design and

¹ Subsection 2.1 authors: Ram Datar (Oak Ridge National Laboratory), Jun Li (Kansas State University), Jun Ni, (University of Iowa), Kazuhiro Oiwa (National Institute of Communications and Technology, Japan), Ari Requicha (Chair) (University of Southern California), and Louis Rossi (University of Delaware).

implementation of networks of sensing, computing and acting components at a scale appropriate for interaction with biological cells and molecules. We believe that such networked systems could bring about revolutionary advances in our understanding of biology.

2.1.1 Objectives

The ultimate goal of this work is to develop and deploy networks of large numbers of components (including sensors and, in the long run, actuators) that are capable of operating in multiple modalities and at multiple spatial and time scales, and that can be interfaced with the cellular and molecular systems which are pervasive in biology. Sensing in real time the values of the many variables of potential importance for a biological system would be a breakthrough in our data acquisition capabilities, and would undoubtedly lead to new understanding of biological function through experimentation, modeling, analysis and model refinement. This understanding, in turn, would enable new biomedical applications such as early disease detection and intervention.

The potential and importance of networks for biological systems, including the human body, have been recognized since the early times of sensor network research. However, it is only recently that the confluence of several technologies is making it possible to launch a systematic attack on the difficult problem of building such networks. Ideally, components should have dimensions comparable to those of the cells and molecules being sensed, to be able to interact intimately with them, and should also disturb the systems as little as possible. This implies that label-free sensing with very small sensors is desirable. Recent advances in nanotechnology, sensor networks, and distributed robotics are providing us with the necessary tools, and the time is now ripe to tackle the problem. But many hurdles remain, as we shall discuss below. These hurdles, together with the huge potential rewards, make this truly a grand challenge.

2.1.2 Technical Challenges

Nanosensing technology is progressing rapidly, with hundreds of new papers on it being published annually. For example, nanowire and nanotube sensors have been demonstrated, which are capable of label-free sensing of a great many quantities of biological interest—see e.g. [CURRELI05, ZHENG05]. These sensors directly produce electric output signals, which can be processed by electronic circuitry. Nevertheless, significant challenges remain:

- Development and networking of multiple nanosensors, for multiple variables (such as biological markers). Here an important issue is the large range of different sensitivities required for the different sensing modalities. Exquisite specificity is needed for some applications in which a very low concentration analyte (e.g., a metastatic tumor cell) exists in an environment of high concentration materials that can generate non-specific noise (e.g., serum).
- Sensor regeneration for re-usability. For example, sensors based on molecular recognition will become unusable when all the recognition sites are full; techniques are needed for resetting the sensors to their original conditions, otherwise they will have short life spans.
- Sensor calibration and stability, especially for applications that require long time spans.
- Neural interfaces. Here the entities being sensed are primarily electrical, such as voltages or ionic currents. Neural sensing raises specialized issues such as stability of interfaces with soft materials, ionic/electric current conversion, electrical or chemical feedback, and access to neurotransmitters.
- Power sources. This is a major challenge, not only for sensors per se, but also for any system component that is to be embedded in a biological environment. Research on nanosensing has thus far finessed the problem by simply using micro or macro electrodes, but clearly the problem must eventually be faced and solved.

- Biocompatibility and biotic/abiotic interfaces. Like power, this is a general problem, and will be crucial for embedded, in vivo systems. Toxicity to the biological system is a major consideration, but there are many other sources of potential problems, from phagocytosis (in essence, being eaten by a immune system cell) to encapsulation by tissue (which precludes contact between the sensor and the sensed entity). Biocompatibility is currently being addressed in many bionanotechnology research projects, e.g., those related to drug delivery.

Communication among sensors and between sensors and other nodes is a fundamental requirement for network operation. Regardless of communication medium, a communication system requires a set of components such as encoders/modulators, transmitters, receivers/detectors, decoders/demodulators, amplifiers and power supplies. All of these components need to be developed for the networks we are discussing. The design of these components, however, will depend strongly on the chosen signal transmission approach.

Chemical communication is well suited for biological systems and it is the method normally used by nature. In a chemical communication system, natural or artificial, the information is encoded into molecules or particles, and these move from sender to receiver, where decoding takes place. Building a human-made, artificial system raises a host of challenging problems:

- Choice of molecules or particles. Note that two types of molecules or particles might be used, one as the information-bearer and the other as a physical carrier of the information-bearing particles. Proteins might be a good choice of information carriers—see below.
- How do we encode and decode information into and out of the molecules or particles?
- Where do these molecules or particles come from? Having a reservoir of molecules or particles in each node of the network is not an attractive solution. Making them on the fly by using available stock materials seems a better approach, but requires non-trivial machinery to synthesize the desired molecules or particles.
- How long do these molecules or particles survive in the medium? This may have significant implications on the communication strategies, and depends strongly on the chosen information carriers.
- Are they eventually disposed of? How? Can they be recycled? These questions can be satisfactorily answered for proteins, which were suggested above as possible information carriers. First off, proteins can be made by using the standard cell machinery, by expressing suitable genes. They can be degraded after they do their job, and the resulting amino acids can be recycled to build new proteins. Note also that proteins can serve as information carriers, but they also are machines. Therefore, a communication system that uses proteins not only sends information to receivers, but also sends them machinery they may be able to use.
- How are information carriers transmitted? The simplest approach here is to use diffusion, which is analogous to omnidirectional wireless transmission. A small molecule in water at room temperature diffuses a distance of a micrometer in milliseconds, but takes several hours to move over a few centimeters. Therefore diffusion is only effective over very small distances. Diffusion also may have difficulties in dealing with cluttered environments such as the interiors of cells. Therefore, alternative approaches that are analogous to wired communication and employ “tracks” may be more attractive. These are used in nature by biomotors that move along microtubules or actin fibers, and cell-mimetic tracked transport systems may prove useful.
- What is the performance of chemical communication channels? How is noise characterized? What errors can be expected? What are the achievable transmission rates?
- How can we model and simulate biotransport from the molecular to the macroscopic level, and at multiple time scales? Modeling and simulation are important tools for answering performance questions such as those posed above.

- How is transport achieved across biological barriers? For example, how does information get across cell membranes?

Chemical communication is not the only possible approach. Even in nature, some of the signals are transmitted electrically, in neurons. This is analogous to wired communication. For artificial systems, electromagnetic transmission might be advantageous because of its inherent speed, and because of our familiarity with electromagnetic technology. Non-chemical nanoscale capabilities have been lacking, but recently a few developments show promise. Optical and plasmonic antennas using nanowires or nanotubes, as well as electromechanical resonators have been reported—see e.g. [CUBUKCU06, JENSEN07]. Operation in the near-infrared part of the spectrum may be advantageous since the attenuation in tissue is reasonable and there are few sources of biological noise in that frequency range. Therefore, alternative approaches to signal transmission are worth exploring and evaluating; integration of several different approaches may prove fruitful.

Thus far we have only discussed sensor networks. In the long run, we need to develop actuators and integrate them into the network. Initially, actuation may amount simply to coordinated motion of the sensor nodes. This would provide capabilities such as adaptive sampling, which would generate data at a high spatial resolution where it is needed, by using mobile sensors. Techniques for dealing with mobile sensors have been under study by the sensor network and distributed robotics communities for some time, and are reasonably well understood—see e.g. [BATALIN04, ZHANG07]. Understanding the role of communication in biological swarms may prove useful to impart a degree of self-organization to the network.

Later on, more sophisticated forms of actuation should be investigated. For example, techniques for inducing cell death or for directly killing cells are of obvious importance. To be able to exploit actuation in a sensor network requires the development of decision-making logic, to interpret the sensory data and drive the actuation. Ideally, this logic should be embeddable, so that the entire system can be embedded in *in vitro* cell systems or, eventually, in animal bodies.

Finally, research on sensor/actuator networks should not take place in an application vacuum. Choice of a compelling biological application to drive the research is important, and the collaboration of biological scientists should be secured from the outset.

2.1.3 Impact and Applications

The development of the networks we have been discussing will have a near-term impact both on network technology and on basic biology. The biological environment and the spatial scale of the nodes are likely to pose new constraints, and imply new trade-offs for network development. As far as we know, network research has not explored this application area until now.

The networks proposed here will be able to provide real time data in multiple modalities and at very small spatial scales, with resolution comparable to the sizes of the cells or macromolecules of interest in biology. They will do this with minimum interference with the workings of the biological system because of the small size of the components, the lack of labels, and the lack of connecting wires.

For concreteness, we consider here a potential generic application, which we call an Instrumented Cell System. Initially these systems will be *in vitro*, possibly using microfluidic facilities for maintaining cell cultures. Eventually, they will evolve towards dealing with explants such as slices of tissue, and ultimately towards *in vivo* studies in what we call an Instrumented Body. In the long run, the Instrumented Body will be human, but this is not expected to happen for at least a decade.

Real-time monitoring of many variables and for a long period of time in an Instrumented Cell System is an unprecedented capability. We can envisage major applications to study cell signaling and behavior, to determine the effects of chemicals on cell systems, and to assess the evolution of the system, to name a few. The new understanding of the biology gained through experimentation, modeling and analysis with Instrumented Cell Systems will eventually be translated into medicine, perhaps under NIH support, or joint support of the NSF and NIH. Applications can be foreseen in early disease detection, treatment evaluation and therapeutics. Medical applications, however, are long term.

An Instrumented Cell System will likely have nodes within and outside cells, and also “mother ships”, i.e., larger and more capable nodes which will be able to communicate with outside the system. Some of the nodes may be mobile, others fixed. This application raises most of the issues we listed earlier in this report, such as: how to communicate between nodes inside and outside cells, and with the outside world; what and how to sense (chemical concentrations, electric potentials, physical variables such as pressure and temperature, ...); where to get energy; and so forth.

The current state of the art in Instrumented Cell Systems may be summarized briefly as follows. Microfluidic systems for building in vitro structures for cell systems exist today. Inorganic nanowire and nanotube sensors, as well as functionalization techniques suitable for detecting a great variety of chemicals have been demonstrated, but additional experiments in physiological conditions are needed. Current nanosensors must be wired. This is probably the major hurdle facing us right now, but, as noted above, promising new technologies are appearing. Algorithms and software for network processing exist and continue to be developed rapidly; however, an Instrumented Cell System may require a new approach—we will have to attack the problem to understand what are the requirements it poses on the associated software. Finally, a concrete and compelling biological application needs to be selected and pursued, to ensure that the research is firmly grounded.

2.2 System Design Methodologies²

This session was charged to address a set of issues that naturally divided into questions of understanding and questions of design:

Understanding

- How do biological entities communicate?
- How is biological information encoded onto and decoded from molecules?
- What are the recurring architectures of biological communications?
- What are the robustness-fragility trade-offs in biological communications systems?

Design

- How can different biological systems be interfaced to enhance communication?
- How may communication mechanisms employed by biological entities be applied to create an artificial communication system using biological materials?

As a result, our attention was drawn to the coupling between *understanding* and *design* in complex systems, a thought that is particularly well-captured in the Feynman quote: “What I cannot create, I do not understand.” Furthermore, we noted that the inverse of this statement (what I do not understand, I cannot create) is equally valid. Design, the reiterative process of

² Subsection 2.2 authors: Michael Simpson (Chair) (Oak Ridge National Laboratory/University of Tennessee), John Doyle (California Institute of Technology), Leor Weinberger (University of California, San Diego), Jeff Hasty (University of California, San Diego), Ido Golding (University of Illinois at Urbana-Champaign), Eric Bachelor (Harvard University), Eric Stabb (University of Georgia), and Sasitharan Balasubramaniam (Waterford Institute of Technology)

creating function through the intentional interconnection of components, is an indispensable tool for unraveling complexity, and unraveling complexity is the key to continued technological progress. Accordingly, we focused on this intersection of *understanding* and *design* in biological systems and the promise that this intersection holds for technology.

Scientific Context

Perhaps the semiconductor revolution represents the most successful program of applied research yet seen. Indeed, one of the great hopes of the National Nanotechnology Initiative is that rapid advances in new materials and fabrication can lead to the continuation or even acceleration of the Moore's law increase in computational density and speed that has occurred over the last few decades. However, the scale and density of the new systems enabled by advances in both conventional semiconductor and nanoscale technologies present both significant new challenges and unprecedented – for manmade systems – opportunities.

As Anderson observed “at each level of complexity entirely new properties appear, and the understanding of the new behaviors require research.....as fundamental in its nature as any other” [ANDERSON72]. This observation highlights a *Grand Challenge* that confronts the continuing advancement of technology: moving beyond just the synthesis of individual or homogeneous arrays to complex arrangements of elements that lead to new classes of functionality. This must be understood as primarily an *information transport* (i.e. communication) problem as the complexity of function scales with the networking capabilities of the individual components. For example, individual cells or self-organized groups of cells perform extremely complex functions that include sensing, communication, navigation, cooperation, and fabrication and organization of synthetic nanoscale materials into highly functional ensembles. These high levels of functionality emerge from the networks of interacting components. Can the deliberate design and assembly of synthetic components lead to systems with cell-like functionality? What lessons must we learn from biology to enable such technological pathways?

The large discrepancy between the functional density (i.e. the number of components or interconnection of components per unit volume) of cells and engineered systems highlights the inherent challenges we face at this intersection of understanding and design. A simple example compares *Escherichia coli* (~2- μm^2 cross-sectional area) with an equivalent area on a silicon integrated circuit [SIMPSON01]. The *E. coli* cell has an ~4.6 million base-pair chromosome (the equivalent of a 9.2 megabit memory) that codes for as many as 4,300 different polypeptides under the inducible control of several hundred different promoters, while the same space on a silicon chip could provide only a very small fraction of this memory or a few simple logic gates. Clearly, the operational scale of biological systems is significantly smaller than that of conventionally engineered systems, but if the differences were just scale, the functionality gap could be closed with new nanomaterials organized in much the same way as today's integrated circuits (Figure 1). However, both scale and *complexity* differentiate the natural from synthetic systems. In the biological substrate, dynamic systems exploit weak interactions arranged to provide desired specificity, and take place in a non-stationary environment. These features lead from simply high spatial density to high *functional* density and the realization of robust, adaptable systems.

As nanoscale science and technology advance, the emulation of biological design principles using synthetic components becomes feasible. Potentially, as systems of such elements approach biological-scale functional density, they can begin to assume cell-like characteristics including: (1) construction from an inhomogeneous mixture of materials with different properties, modes and strengths of interactions, and relative abundances; (2) the encoding of information within small populations (e.g. biomolecules or electrons); (3) function emerging from an environment with large stochastic fluctuations (a consequence of (2)); and (4) the efficient transduction of information, energy, and materials that emanates from the molecular

scale. It is an intriguing possibility that as our ability to control the synthesis and direct the assembly of synthetic nanoscale elements increases, we may attempt the bottom-up design and construction of systems with cell-like complexity and capabilities. Below we present objectives for the scientific exploration of the intersection of design and understanding in cell-scale systems.

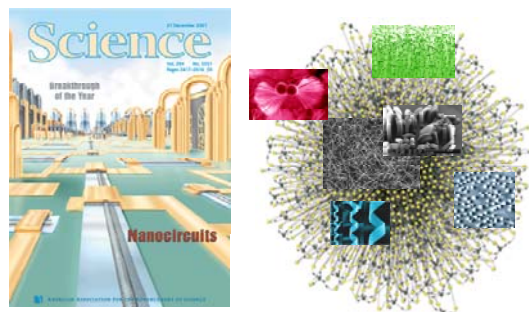


Figure 1. (Left) Electronic nanocircuits were the *Science* magazine breakthrough of the year in 2001. As envisioned then by *Science*, these new materials would be organized in old (i.e. tradition integrated circuit) ways. (Right) In contrast, the lessons of nature suggest that reaching new levels of functionality requires new architectures.

2.2.1 Objectives

Rational design paradigms involve the synergistic coupling of theory, synthesis, and characterization. Unfortunately, although there are numerous *ad hoc* approaches with limited applicability, *there is no general theory to guide the rational organization of elements into highly interactive functional collectives*. Accordingly, the central objective here is to integrate a fourth component into the design paradigm: the observation and mimicking of the organization of natural complex systems through the pursuit of the following research directions:

- *Discover conserved architectures that manage the flow of information in genetic network*
 Engineers can learn from biological systems, which are robust and evolvable in the face of even large changes in environment and system components (although they can be extremely fragile to rare small perturbations [DOYLE07]). One approach flows from chaos, fractals, random graphs and power laws, which inspire a view of complexity in which typically unpredictable behaviors emerge from simple interconnections among like components [ALON03, BARABASI99]. For example, metabolic networks are found to be ‘small-world’ and ‘scale-free’ [BARABASI99, JEONG00, RAVASZ02]. Small-world implies that any element in the network is connected to any other element through a small number of interconnections. Scale-free implies that even though the average number of connections to elements may be relatively small, a significant subset of elements will have many more than the average number of connections, and these highly connected hubs are important to the network function. A different approach focuses on organization, protocols and architecture [DOYLE07], and views genetic networks as ‘plug-and-play’ modules that communicate through a highly conserved set of common protocols [DOYLE07, CSETE04] (e.g. bacterial quorum sensing [DUNN07]). The goal in this objective is to develop the understanding of these biological organizational principles to the point that they can be applied in synthetic complex systems.
- *Discover biological strategies for the conservation and distribution of noise across information carriers in dense networks*

Stochasticity is an inherent feature at the nanoscale as information is represented by small populations of objects (e.g. electrons or protein molecules). Numerous studies have described the origins, processing and consequences of noise in genetic systems [ELOWITZ02, ARKIN98, VOLFSON06, AUSTIN06, STERN07]. However, a systems-level question that remains is understanding how noise distribution across system components may be optimized. A typical cell may contain 50M molecules of a few thousand different types, and the coefficients of variation (CV; standard deviation/average) for these populations would range from 1-100%. Fluctuations (noise) in any one molecular population could vary from negligible to dominant, and can be made negligible just by the selection of a large population. However, because of the bounded (limited size and power), crowded (many different functional elements competing for space and power) environment, such a selection made in favor of one comes at the cost of moving this stochasticity to other species. That is, in bounded complex systems total stochasticity is ‘conserved’, which may be stated analytically as follows:

$$\sigma_n^2 \propto \langle P_n \rangle,$$

where σ_n^2 is the variance in the population of element type n and $\langle P_n \rangle$ is the average population of element type n . However, in a bounded environment with a mixed population of N different element types,

$$\langle P_T \rangle = \sum_N \langle P_n \rangle = C_p,$$

where $\langle P_T \rangle$ is the total population of all elements, and C_p is approximately a constant that is set by the space/power constraints of the system. As a consequence,

$$\sum_N \langle \sigma_n^2 \rangle \propto C_p.$$

Therefore, the reduction of the variance in one element type can only be accomplished by distributing this stochasticity to the other element types, and little is known to guide this distribution in synthetic systems. Two recent studies of system-wide stochasticity shed some light on the pertinent issues [NEWMAN06, BAR-EVEN06], suggesting that there is some (at present unknown) strategies for the optimal distribution of stochasticity across the elements of complex systems. The key question for this objective is “*what can biology tell us about how stochasticity should be distributed to the different functions or classes of function in synthetic information transport systems?*”

- *Create synthetic dense networks that apply these strategies*

This objective ‘completes the loop’ in the coupling between understanding and design, and leads both to advances in technology and in science. Whereas modeling, simulation and experimental analyses have a tendency to focus attention on the details of individual elements, design requires grappling with the trade-offs and compromises needed to enable system function. Along these lines, synthetic biology efforts follow a strategy of constructing deliberately simplified systems to comprehend molecular and cellular regulatory processes from the bottom up [GUIDO06, SPRINZAK05, HASTY02, ANDRIAN06]. Similarly, efforts towards constructing minimal cells either add, subtract or manipulate components to realize simple systems with desired capabilities [FORSTER06, LUISI06]. In both cases, iterative design is a fundamental aspect of the approach and represents a major step towards true bottom-up construction of biological complexity. However, these approaches still depend on a platform of an existing cellular environment or the use of biomolecules (nucleic acids, proteins, lipids) to jump start cellular function. Thus, the questions remain – what new technologies could be developed and what new biological understanding could be learned from true bottom-up efforts to reconstitute cell-like complexity in synthetic systems?

2.2.2 Technical Challenges

The technical challenges presented by the research envisioned here breakdown into the broad categories of ‘seeing’, ‘conceptualizing’, and ‘design methodologies’.

“Seeing is the beginning of understanding.” [LAUGHLIN00]. The power of science to describe and explain the universe around us begins with observation, leading through hypothesis and experiment to physical laws and theoretical models. Feynman [FEYNMAN60] identified improvements in electron microscopy as a key advance in making “room at the bottom;” the beginnings of fundamental nanoscience and nanotechnology has been ascribed by Toumey [TOUMEY05] to the unforeseen invention of scanning probe microscopy by a small group at IBM Labs. Controlling and designing properties that lead to a desired outcome require the ability to see and manipulate the functionality at the length, time, and energy scales of the problem. Obviously, there are numerous microscopy tools and techniques for seeing biology, but these often fall short as *system* characterization tools due to a lack of specificity and/or bandwidth. On the one hand many of the ‘omics’ tools (gene arrays, PCR, etc.) provide a great deal of static (snapshot) and bulk (averaged over many cells) information about biological systems, but are limited in their ability to probe dynamics, particularly at the single cell level. Conversely, reporter proteins – primarily fluorescent proteins derived from GFP – have been the workhorse for the study of dynamics in single cells, but are limited to probing only a portion of the expression process for (at most) a few genes simultaneously. Compared to the powerful tools of electronics system characterization (e.g. oscilloscopes, logic analyzers, network analyzers), these biological tools are extremely limited in their ability to provide systems observability.

Conceptualization as used here means developing an understanding of biological systems in systems terms. As described earlier, this means an understanding of biology with a focus on organization, protocols, and architecture. Most often this means going beyond first approximations in the classical view based on rates [GOLDING05], concentrations [GOLDING06], or function driven by steady-state conditions [WEINBERGER08]. Furthermore, this requires analysis, modeling, simulation, and experimentation that fully appreciate and integrate stochasticity. Indeed, this may be a challenge that the NSF is uniquely suited to address as it requires the intellectual and technical integration of two very different scientific cultures.

Finally, design methodologies that enable genetic network scale integration are the practical tools that couple understanding and creation. In contrast to the very powerful design tools used for electronic system design, these future tools must embrace a great diversity of technologies from inorganic semiconductors to living systems, and from nanoscale to macroscale.

2.2.3 Impact and Applications

Central to the research described here is the creation of organizational schemes, protocols, and architecture that enable the design of function within complex arrangements of individual components. As a result, the far-reaching impacts and applications would extend into any area where function is a consequence of network structure. Although the impacts would extend much more broadly, they are perhaps easiest to see in electronic systems where both conventional top-down and newer bottom-up fabrication are pushing the boundaries on scale, density, and complexity. Indeed, these systems are on the edge of facing the challenges of power management, interconnectivity, and stochasticity that to this point have only been met by biological networks. Accordingly, the research envisioned here has the promise of

- *Solving the density-power challenge of networks-on-chips* by providing strategies that reduce power requirements. For example, one of the limits on the reduction of chip power supply

voltages is the need to provide noise margins. Architectures that use noise to functional advantage would remove this key barrier to on-chip power reduction.

- *Realizing the functional promise of networks of synthetic nanoscale elements* by providing organizational schemes, protocols, and architectures that can take advantage of the scale and network complexity made possible by these new materials.
- *Creating synthetic systems with a level of functionality rivaling that of living systems.* The research in other areas is providing the means to synthesize and assemble systems that approach the complexity of cells. Unfortunately, it is often assumed that these new levels of functionality can be obtained by arranging the materials in old ways (Figure 1). The research envisioned here offers a new approach grounded in the architectures of nature.

2.3 Coding Theory and Channel Capacity^{3 4 5}

Hartley in 1928 (Hartley 1928), in an attempt to quantify the flow of information for the signals in wired communication systems, proposed the logarithmic measure of information. Probabilistic information theory, as we know it today was fully developed by Claude Shannon in 1948 and 1949 [SHANNON48, SHANNON49], and has become the basis of modern communication systems. An intriguing clue that Shannon's theory may apply to biology came when he wrote the uncertainty equation:

$$H = -\sum_i p_i \log_2 p_i \quad (\text{bits per symbol [or second]})$$

as it closely resembles the Maxwell-Boltzmann form for statistical entropy [SCHNEIDER91b]. This formula was used in biology as early as 1949 [DANCOFF53, KAY00] by Dancoff and Quastler and extensively in 1956 [YOCKEY58a, YOCKEY58b]. Since entropic changes drive life processes, this formula can be used to measure biological phenomena by computing the decrease in uncertainty as a measure of information ($R = -\Delta H$).

A practical example is the sequence logo, which displays patterns in the binding sites of proteins on nucleic acid or protein structures [SCHNEIDER96]. Logos display sequence information computed as the decrease of uncertainty from before sites are bound to after they are bound (or aligned).

In addition to developing a fundamental and practical measure for information, Shannon also found a way to measure the capability of a communications line to handle the information, the channel capacity:

$$C = W \log_2(P/N + 1) \quad (\text{bits per second})$$

³ Subsection 2.3 authors: Thomas D. Schneider (National Cancer Institute at Frederick), Hubert P. Yockey, Dónall A. Mac Dónaill, (University of Dublin, Ireland), Elebeoba E. May (Sandia National Labs), Christopher Rose (Rutgers WINLAB), John S. Garavelli (EBI, Cambridge, UK), Andreas G. Andreou (Johns Hopkins), and Andrew W. Eckford (York University, Canada)

⁴ This report in subsection 2.3 is biased by the perspectives of the members of the workshop section who, by necessity, only work on a few of the many possible applications of information theory to biology. We apologize to those who might have contributed their insights but who were not able to attend the meeting. Pierce [PIERCE80] discussed how information theory is so general as to apply to many fields. In contrast, at the height of the initial excitement after the publication of his papers, Shannon wrote a small paper "The Bandwagon" [SHANNON56] warning that a careful application of information theory would be necessary for fields outside communications. We feel that after 60 years of development, the available data and the conceptual advances exemplified in this report warrant optimism that many additional advances in understanding biology can be obtained by using information theory.

⁵ The authors of subsection 2.3 would like to thank Cynthia Yockey for useful suggestions and editing. The authors were greatly saddened to learn of the loss of Sergio Servetto who perhaps could have contributed greatly to this meeting and report.

where W is the bandwidth, P the signal power and N the noise power at the receiver. Shannon showed if the rate R is less than or equal to C , communication may have as few errors as desired.

This theorem has been ‘transported’ to molecular biology terms [QUASTLER58, YOCKEY58c, YOCKEY74, SCHNEIDER91a, SCHNEIDER91b, SCHNEIDER94] in which messages correspond to distinct states of molecules. A surprising result is that the channel capacity theorem only applies to biology and not non-living physical or thermodynamic systems [SCHNEIDER06].

Attneave [ATTNEAVE54] and Barlow [BARLOW61], pioneered the use of information theory for rigorously quantifying stimulus-response function perception in the nervous system. This has led to a flourishing field aimed at understanding neural coding in terms of the rigorous tools and methods of communication and information theory at the level of single cell components, i.e. molecules, single cells [PAMELA01], as well as network of neural cells [ATICK92, LEVY96, ABSHIRE01, GOLDBERG04].

Thus research challenges lie in three areas: (1) understanding how to use information theory to learn about biology, in particular as it relates to the different scales of the system, from molecules to behavior, (2) how to use biology to learn more information theory and (3) how to apply these ideas to bioinformatics, engineering systems and technologies at the macro-, micro- and nanoscales.

2.3.1 Understanding Biology Using Information and Coding Theory

2.3.1.1 Objectives

The objective of this area is to understand how information and coding theories apply to biological systems with the objective of further developing biology as a rigorous theoretical science. In reaching this objective we will establish probability theory, information theory and coding theory as one of the mathematical foundations of theoretical biology.

2.3.1.2 Technical Challenges

- *Identify biological systems that can be investigated experimentally using information theory.*

The Symposium on Information Theory in Biology was held in 1956, only a few years after Shannon’s famous papers had been published [QUASTLER58]. At that time authors considered a variety of applications in biology, with topics including membrane phenomena, antigenic specificity, protein structure, morphogenesis, effects of radiation and aging. However, even today information theory is neither widely understood nor used to advantage in modern biology. We suggest that the reason for this lack of application is that sometimes information theory has been used only for description and not for prediction as a guide to experimentation. Furthermore, there is a chasm between the methodologies and approaches that drive modern biology and the rigorous foundation of tools that the mathematics of information theory offers. The challenge, then, is to find biological systems where the application of the theory can advance our understanding of fundamental biology.

There have been, however, fundamental breakthroughs where information theory has provided the key link between in the iterative process of theory, prediction, experimentation. An example at the molecular level is the discovery that the second step of DNA replication and transcription initiation, after proteins bind to the DNA, is likely to be one or a few bases flipping out of the DNA. This was observed clearly and predicted using information theory [PAPP93, PAPP94, T. D. SCHNEIDER01]. Experiments to date have confirmed the proposal

[LYAKHOV01]. What other biological systems can be dissected using the clarity provided by information theory?

- *Given currently available or experimentally generated data, test the application of information and coding theory to biology.*

At the molecular level, we now have enormous amounts of biological information scattered across a vast literature. A serious challenge is to gather the relevant information for analysis. Collecting a set of proven natural binding sites for proteins on DNA or RNA is the major factor limiting the application of sequence logos [SCHNEIDER90] and sequence walkers [SCHNEIDER98] to the thousands of genetic control systems that are now known. One challenge we must meet is to support databases that produce annotations for experimentally proven natural sites with original literature citations. In the field of neuroscience, at the cellular/network level, a wealth of data has been accumulated over decades of experimental psychophysical and behaviour studies which are awaiting rigorous methods of analysis to elucidate the principles of neural codes.

- *What is the structural/physical basis of codes in molecular biology?*

Digital binary codes are employed in modern electronic and computational technologies to detect and correct errors in transmitted and recorded data [HAMMING50, CIT86]. Offering a powerful mechanism for error control, it is not surprising that codes have also been found in molecular biology, where the particular composition of the nucleotide alphabet of A, C, G and T can be formally interpreted in familiar error coding terms. In this case information is expressed through hydrogen donor/acceptor (D/A) patterns [DOONAIL02, DOONAIL03, SZATHMARY03]. Another study [MAY06b], suggests that error-coding is also used in translation initiation. Recent experimental results on the DNA binding protein Fis also suggest a coding mechanism [SHULTZABERGER07].

Information transmission permeates molecular biology, and it is highly likely that codes are widely employed. What are they, and how can we identify them? What is their physico-chemical basis? Over the past two decades the field of supramolecular chemistry has provided a vast array of molecular systems capable of molecular recognition and assembly [LEHN02, CORBIN02]. Drawing on this, what physico-chemical properties are necessary and most useful for the construction of molecular codes? In what ways may chemical constraints cause molecular and biological codes to differ from electronic codes? Why, and when, would some molecular codes prove superior to others?

- *At a given level of description, how closely do individual biological systems approach the channel capacity?*

Given that we can precisely measure the information in at least one particular biological system, namely DNA and RNA binding sites of proteins and other macromolecules, the important question is how close the biological systems come to the channel capacity limit determined by Shannon.

The equivalent of channel capacity for molecules has been determined as the number of distinct states that the molecule can reach for a given energy dissipation [SCHNEIDER91a, SCHNEIDER91b, SCHNEIDER94]. It has been demonstrated that several DNA binding proteins such as EcoRI do operate at the channel capacity (T. D. Schneider, in preparation). Our challenge is determining how closely biological systems do come to the channel capacity. Do they use coding tricks that are unknown in modern coding theory?

- *What current assumptions in modern information/coding theory must be relaxed or modified to apply information theory to molecular systems?*

Modern coding methods used in the communications industry have adopted a number of assumptions which are less general than Shannon's original vision of coding in which codes are represented by the packing of spheres in high dimensional spaces [SHANNON49]. These assumptions include two state (binary) systems, rigid timing and non-changing coding schemes. It is likely that biology uses much more general and fluid systems. Can we extend the modern communications methods into analysis of biological systems, or will we require an entirely new paradigm in order to understand them?

- *Develop a consensus on terminology to eliminate fundamental confusions.*

Applications of information theory to biology frequently falter on misunderstandings caused by imprecise use of technical terms. For example, when someone uses the word 'entropy' they must make it clear whether they are talking about the thermodynamic entropy, $dS \geq dQ/T$, the Gibbs-Boltzmann formula of statistical mechanics, $S = -k \sum p_i \ln p_i$, or the Shannon 'uncertainty', which some people call the Shannon entropy, $H = -\sum p_i \log_2 p_i$ [JAYNES03]. Likewise within biology and the scientific press, some basic terms such as 'alphabet,' 'code' and 'coding' are so widely misunderstood and misapplied that not only can basic concepts like the recording and transfer of information remain murky, but also descriptions of biological communication processes are not framed correctly ([ITE05] pp. 10-15). What needs to be emphasized is that the fundamental terms of information theory must be defined so that they are used in the field of biology correctly and consistently.

- *Understand how noise affects biological systems.*

A challenge is to apply information theory to measure the noise in the transfer of genetic information in each living entity from the genome to the proteome to determine how closely each one approaches the channel capacity. Understanding noise will help in the study of aging, diseases, evolution and the effects of exposure to radiation, which is of interest both in medicine and the effects of radiation on humans in space exploration.

2.3.1.3 Impact and Applications

- *Paradigm shift of biology from a descriptive science to one with a theoretical foundation capable of prediction and extrapolation.*

Formal efforts to integrate theoretical biology into a unified discipline with a mathematical foundation date from the founding of the Committee on Mathematical Biophysics in 1934 at the University of Chicago by Dr. Nicolas Rashevsky. It became the Committee on Mathematical Biology in 1947 [MAINI04] (p. 596). In 2008, theoretical biology now has a clear mathematical foundation in information theory, coding theory and probability theory—yet 74 years after the first seminal attempt, mathematical biology still does not exist as an integrated discipline with its own university departments and full-time faculties. The National Science Foundation is in a unique position to facilitate the paradigm shift of biology from a descriptive science to one with a mathematical foundation by providing a unified vision, holding multi-disciplinary meetings (such as the workshop that led to this report) and providing grants aimed at achieving that goal. Without such a unified vision and action coordinated at the national level, the benefits to the various disciplines—and university departments—of staying separate from one another may keep them separate indefinitely.

- *Information theory can provide the underlying explanations for molecule-to-molecule biological interaction parameters which can then be used in the construction of higher level systems biology.*

To understand biology from a theoretical perspective is to be able to predict the response of the system at the different levels from molecules to behavior. While this may seem to be a daunting task, the engineering tools of information theory, communication theory as well as control and signal processing provide a solid foundation in this endeavor. These theoretical tools provide guidance as to the plausible explanations that can be given to certain experimental observations. More important however, the theoretical tools of information and information theory provide the means to explore theoretical hypotheses that pertain to the system at different levels of description. For example while seemingly radically different, coding at the cellular level in a network of cells and coding at the molecular level within a single cell share fundamental principles rooted in the fields of information and communication theory [PIERCE80]. An analysis by synthesis approach offers through the rich landscape of information and communication theory the theoretical biologist the insights necessary to make progress.

The inputs and outputs of cells are continuous signals, expressed as currents, voltages, and chemical concentrations. These signals are propagated throughout tissue, for example neural tissue in the brain, by a massively connected communication network. Because cells have structures that are distributed in nature they are electrically leaky, so communication of signals beyond the immediate vicinity in a network of cells must be encoded as spikes which are amenable to active restoration and transmission over long distances.

How spike encoding and decoding distorts neural signals and what hypothetical strategies are that neuron cells could use to minimize this distortion remains an open problem. A comparison of spike codes can tell us which codes are the most effective and, in turn, are most likely to be employed by neuronal cells in the nervous systems. The notion that neural systems are organized to realize efficient representation of information dates back to the seminal work of Attneave [ATTNEAVE54] and Barlow [BARLOW61]. The organism that can process information the fastest with a fixed amount of resources can react more quickly to a constantly changing world, and in turn is more fit to survive and procreate. Similarly, the organism that can process a fixed amount of information with fewer resources is also more fit. In these ways, evolution should have optimized biological structures for efficiency. The optimization criteria may be determined by the function of the biological system or the constraints that guided its development.

One way to quantify the transfer of information over a “cellular link” is to compute its channel capacity—the maximum amount of information (per unit time) that can be transmitted over the link. A channel is characterized by an input alphabet, an output alphabet, and a probabilistic mapping between the symbols in the two alphabets. The channel capacity for a spiking link, a specific instance of inter-cellular communication that uses alphabets that consist of the number of spikes in a window, has been derived by MacKay and McCulloch [MACKAY52]. An analysis based on channel capacity assumes that the encoder exploits the full capacity of the channel and that the decoder is capable of recovering all of the information present in the spike train. However, a spiking link will not achieve the channel capacity if a great deal of resources are required to achieve capacity. For this reason, it is important to quantify the transfer of information over a spiking link by accounting for the noise incurred by biological encoders and decoders [GOLDBERG03].

This is because we can hypothesize that a neural system will use the most efficient code—the encoder/decoder pair that has the most favorable trade-off between performance and cost. Ideally, we would like to systematically search through all possible inter-cellular communication codes to determine which are the most efficient in order to test this hypothesis. Such a search is at the moment intractable, so a fruitful research program must focus on specific encoders and decoders that have been discussed in the literature or new methods must be developed. To be tractable, such encoders/decoders pairs must be

characterized by few parameters, and therefore facilitate a parametric exploration of the inter-cell communication performance when in a network. This is in contrast to decoders that can be found via optimization techniques, such as the optimal linear decoder [GOLDBERG07].

Furthermore, biophysically detailed models of encoders and decoders are required to estimate the cost of inter-cell communication. Such models will enable us to further examine how energy consumption and other constraints influence coding.

Traditionally, the engineering of communication systems has followed the “separation” principle. On one hand “channel coding” transforms a “noisy” channel into one that allows for reliable information transmission at capacity C . In a complementary way, “source coding” reduces the source rate to match the capacity of the reliable channel. In the latter case, source coding can introduce distortion. While much progress has been made in applying information theory to biological problems at different levels of abstractions, in recent work by Gastpar et. al. it is argued that other than channel capacity, the study of rate-distortion tradeoffs may be more appropriate in the study of biological systems [GASTPAR03].

2.3.2 Understanding Information and Coding Theory Using Biology

2.3.2.1 Objectives

The objective of this area is to advance coding and information theory by learning from biological information systems at all scales. Molecular biological systems have combined the information for spatial compression, error control and methods to cryptographically protect chromosomal information within the DNA sequence. As such biological sequences appear to have a type of nested encoding structure [BATTAIL06], where one layer of information is ‘encoded’ and combined with additional information. We can then think of transcription and translation like peeling off layers of an onion where we need to expose the information (so spatially uncompress) by ‘unwinding’ the DNA in the chromatin; then we unpeel the next layer through transcription and the final layer through translation.

2.3.2.2 Technical Challenges

- Determine whether the kinds of codes used by biological systems have been studied previously.

In order for biological information and coding theory to impact the mathematical and engineering sciences, we must first identify coding principles used by biological systems and determine whether our current coding models can adequately describe them. One approach to inferring the coding theory principles used by molecular biology systems involves taking encoding/decoding methods used in the field of communication theory and testing if various genetic sequences “fit” the encoder model and then providing some assessment of the degree of “fit” [MAY07, MAY04a, ROSEN06, LIEBOVITCH96]. Additional methods based on physical and chemical properties of nucleic acid bases and the relationship between hybridization and phenotypic response need to be explored [DOONAILL02, MAY06a].

An objective of “living” systems is to persist/survive and transmit/reproduce. Therefore, as in engineered systems, biological systems must contain mechanisms and methods for representation, storage/maintenance, protection and reliable transmission/replication of biological information. The question is what are they, how similar are they to our approach to communication (in an engineering sense), and how can we best quantify them?

Representation of biological information: In engineered systems the fundamental representation of information is in the form of a bit. Engineering codes are composed by

combining bits according to rules that obey finite field mathematics. In molecular biology, the fundamental unit of genetic information can be defined as the base. Sequence logos and walkers [BINDEWALD06, SCHNEIDER90, SCHNEIDER98] represent patterns in DNA, RNA and proteins but it is not clear that they depict the actual codes used. Can we define a mathematical “field” of sorts that captures the physico-chemical properties of nucleic acid bases and the interaction between bases? Mac Dónaill and colleagues have begun exploring this question [DOONAILL02].

Biological Storage Codes: In source coding, the purpose is to remove the redundancy in the information, usually by compression. Spatial compression of DNA into chromosomes is empirically known to involve the interaction of histone proteins with chromosomal DNA. This interaction results in a three-dimensionally folded structure. This physical compaction is different from sequence information compression which is not dependent on the physical path of the molecule. Previous work on nucleic acid sequence compression [LOEWENSTERN99, LOEWENSTERN98, POWELL98, DELGRANGE99] as well as studies of codes with structure, such as Turbo Codes [BERROU93, GUIZZO04], may offer plausible starting points for providing ideas on how to mathematically capture possible DNA spatial compression codes.

Biological Protection Codes: Encryption and subversion of cryptographic codes is a critical area in protection of information. While public key encryption and similar types of codes are common within human information transmission systems, identifying biological cryptographic mechanisms used by viral systems to thwart the host defense at the molecular level and obtain “write access” to host genome remains relatively unexplored.

Biological Transmission Codes: Channel coding (error-control coding) adds redundancy in order to protect the transmitted information from errors introduced by the channel. Several parallels between molecular biology and information transmission systems have been recognized and various authors have made these comparisons qualitatively and quantitatively [EIGEN93, GATLIN72, MAY06b, MAY04b, SCHNEIDER91] [ITMB92] (pp. 110-111, Figures 5.1 and 5.2), [ITE05] (pp. 34-35). Existence of redundant genomic information is accepted. Since in engineered systems redundancy is the hallmark of error-control encoded information, it is also plausible to accept the argument that nature employs some method of coding to transmit and protect its information. The question is what is the nature of this code? Is it linear or non-linear? Does it incorporate memory like convolutional codes or are the encoded segments disjoint and more akin to block codes? Does the structure of the biological code tend towards nested, concatenated codes? What are the coding parameters such as the coding rate, the memory length if dealing with convolutional code models and the error detection/correction limits. Some initial work on the use of entropy measures may provide plausible methods to investigate coding parameters [MAY03]. Researchers have explored the parallel between the ribosome during protein translation and a channel decoder [MAY07, MAY04b, ROSEN06, LIEBOVITCH96]. In addition to the ribosome other biological mechanisms and machines “decode” and transmit, such as the RNA polymerase and the mechanism of post-transcriptional modification [FORSDYKE81]. The role of these mechanisms in the biological information transmission network need to be explored and mathematically described.

- *Understand how multidimensional coding is generated and used in proteins and other biological structures.*

In his 1949 paper [SHANNON49], Claude Shannon presented an elegant model for communications: “Essentially, we have replaced a complex entity (say, a television signal) in a simple environment [the signal requires only a plane for its representation as $f(t)$] by a simple entity (a point) in a complex environment ($2TW$ dimensional space).” He then went on

to show that because of this mapping from a low dimension to a high dimension, there will be discontinuities. Apparently the equivalent discontinuity has been observed for DNA binding proteins [SHULTZABERGER07], which implies that they use sophisticated coding. Exactly what are the codes that DNA binding proteins use? Will these codes be ones we know about already or will they be new?

- *Learn how three dimensional coding, such as found between the surfaces of interacting proteins, operates.*

The binding site of a protein on DNA stretches along the molecule and, for the most part, has no correlations between one base and the next [STEPHENS92] so representations such as sequence logos [SCHNEIDER90] generally do not require an additional dimension [BINDEWALD06]. Therefore, even though the protein recognizes a three-dimensional surface on the DNA, it is easily represented by single letters (A,C,G,T). In the case when two proteins bind to each other, there is also an interaction in three-dimensions, but the pattern is over a surface. It is not at all clear how to deal with this theoretically because the surfaces of proteins contain many chemical moieties that interact with many other molecules such as DNA, small molecules and other proteins at all possible displacements and angles. If molecular interactions are coded as is suggested by recent experimental results [SHULTZABERGER07], then the code is likely to be expressed in these interacting two dimensional surfaces. How can we think about two-dimensional surface codes and their evolution?

2.3.2.3 Impact and Applications

- *Creation of novel coding techniques that approach the channel capacity for application at both the molecular nanotechnology and macroscopic levels.*

Due to their low energy and complexity costs, molecular communication systems have great potential in nanotechnological contexts. For example, molecular communication may be used as a means for controlling nanoscale machines, or for gathering information from nanoscale sensors. However, this mode of communication is poorly understood by communications engineers. Given accurate statistical models of these channels, it is possible to determine their Shannon capacity – that is, the maximum data rate at which reliable communication is possible [ECKFORD07]. These limits give an idea of the suitability of molecular communication for its intended applications in nanoscale control and sensing. However, since error-control coding is typically required to approach the Shannon capacity, it is necessary to consider how to perform such coding in biological and nanotechnological contexts. Inspiration can be gathered from the study of biological error-control mechanisms, for example in DNA translation [MAY06b], to find appropriate error-control codes and decoders.

2.3.3 Applying Information and Coding Theory to Bioinformatics and Bio/Nanotechnology

2.3.3.1 Objectives

The objective of this area is to extend the application of information theory to novel biotechnological and nanotechnological problems.

2.3.3.2 Technical Challenges

- *Understanding the limits of binding site predictions using individual information.*

As well as the binding of proteins to nucleic acids, information theory is used in studying the interaction of proteins, how they bind and react. The Sequence Logo [SCHNEIDER90] has been particularly useful for visualizing significant patterns in the binding or modification sites of proteins, and they are being incorporated into database projects for ProSite [HULO06] and the RESID Database [GARAVELLI04].

Sequence logos graphically represent the average information found in the ensemble of sequences for a binding site [SCHNEIDER96]. Information theory also allows the ensemble of sequences to be split into individual contributions by computing the information of each sequence relative to the entire set [SCHNEIDER97]. The individual sequence contributions can be displayed using a graphic called a sequence walker [SCHNEIDER98].

Simple binding sites bound by a single protein are well modeled this way but nature has many more complex sites such as ribosome binding sites and promoters that contain several components. Each component such as the -35 and -10 of the σ^{70} promoters in *E. coli* can be modeled as a rigid object and then the variable spacing can be determined. The probabilities of each spacing are known so the uncertainty added to the overall pattern can be computed. When the rigid parts are combined in this way by a flexible spacer, the resulting models for *E. coli* ribosome binding sites [SHULTZABERGER01], σ^{70} promoters [SHULTZABERGER07] and multi-part promoters in yeast [SHULTZABERGER99] are quite reasonable. These models are similar to models constructed by training hidden markov models (HMM) [KROGH94] except that the probabilities of various spacings are taken into account. Thus the challenge has been reasonably well met already but combining HMMs with information theory and accounting for variable spacing probabilities needs to be addressed.

An unsolved challenge appears in the Fur binding sites of *E. coli*. Fur binds to clusters of overlapping sites [CHEN07]. How to separate out the contributions of direct binding from those of neighboring sites has not been solved. One might think that SELEX and other cyclic evolutionary schemes may be used to identify single sites to construct models, but this failed for unknown reasons in the Lrp system [SHULTZABERGER99]. Thus it is not clear that experimental approaches, which use potentially unnatural *in vitro* conditions can be used to construct reasonable models.

An additional challenge is understanding the plethora of homodimers and heterodimeric transcriptional activation proteins found in eukaryotes. The first challenge is to identify the binding sites to which the set of proteins binds, but often several members of a family will bind with varying affinity to the same sites. Again, *in vitro* experiments may give anomalous results [SHULTZABERGER99].

A clear understanding of molecular interactions from a coding and information theory perspective will give us insights into disease processes such as those caused by mutations in splice junctions [ROGAN95, ROGAN98, JAYNES03]. With this understanding biotechnologies for handling the diseases can be developed rationally by engineering.

- *How can we apply our increasing knowledge of biological codes to technological applications?*

Construction of coded molecular systems such as the Medusa(TM) DNA Sequencer. During the process of adapting the channel capacity formula to molecular biology, it was observed that molecular systems are mostly coded in space, whereas the classical Shannon model is coded in time [SCHNEIDER91]. A combined space and time device would code in both of these dimensions. A concept for a DNA sequencing device may be the first application of these possibilities. The Medusa(TM) DNA Sequencer [SCHNEIDER05] consists of a single molecule that emits a series of four distinct spectra to indicate the DNA sequence being read. An arrangement of the emitting fluorophores similar to a Hamming code can be designed to

indicate when the device has been broken. What other devices can be built on these principles?

- *Development of multiple, gapped sequence alignment algorithms to avoid systematic bias in phylogenetic analysis.*

Sequence alignment algorithms are the basis of every evolutionary and comparative study, and errors in alignments can lead to significant errors in the interpretation of evolutionary information in genomic sequences. Sequence logos, whether nucleotide or protein, critically rely upon having robust methods for producing alignments and evaluating sequence homology.

Sequence logos are usually prepared from alignments of sequences considered to be orthologous. The judgement of whether selected sequences are orthologs, or even homologs, would ideally be based upon experimental biochemical evidence relating the function or structure of the sequences. The capability for comparing very large arrays of sequences has been enabled by the huge advances in genomic research and in information technology over the last 12 years. By comparison, proteomic methodologies to provide supporting biochemical evidence are still relatively young, and their capability for providing the large scale results required are still being explored.

Some very recent studies have suggested that the algorithms used to produce large-scale sequence alignments might not be as robust as had been expected from smaller scale comparisons. Wong, Suchard and Huelsenbeck [WONG08] hypothesized that, “Statistical methods that until recently would have been applied to a single alignment, carefully constructed, are now applied to a large number of alignments, many of which may be of uncertain quality and cause the underlying assumptions of the methods to fail.” They investigated seven different alignment methods on a large set of presumptive orthologous gene sequences from seven yeast species, and found that uncertainties in the alignments lead to the different alignment methods producing inconsistent conclusions. Other workers who had seen similar effects in other studies [LÖYTENOJA08] suspect that traditional multiple sequence alignment methods produce inconsistent conclusions because they disregard the phylogenetic implications of alignment gap patterns. In order to produce useful large-scale alignments for functional and phylogenetic research, it will be extremely important to have robust alignment algorithms, producing reliable results.

Previous work on using information theory to align DNA binding sites by maximizing the information [SCHNEIDER96] was successful [HENGGEN97] but the method did not account for gaps. A major challenge is to find fast and robust methods for sequence alignment with gaps based on information maximization.

2.3.3.3 Impact and Applications

- *Improved clinical diagnosis from information theoretic analysis of genetic sequences.*

Information theory can be used to investigate the information in individual binding sites [SCHNEIDER97, SCHNEIDER98]. For example, 15% of all point mutations that cause genetic diseases are in RNA splice junctions [KRAWKZAK92] and these can be analyzed to predict whether a sequence change is merely a polymorphism, a likely cause of the disease, or merely a mild disease [ROGAN95, ROGAN98, INUI08]. These same information theory based techniques can be applied to transcription factors to understand sequence alternations in promoters and to predict previously unidentified genes under a particular control [VYHLIDAL04]. These techniques provide a single unified theory for understanding all

binding sites. Although they have been successfully applied in many genetic systems it is not clear what their limits are.

- *Development of robust coded nanotechnologies.*

Understanding biological coding theory can help us design coded nanotechnologies such as the deoxyribozyme-based computational biosensor [MAY08] which uses coding theory to robustly classify 15-base DNA target sequences into strain-specific subtypes and identify mutations within the target sequence. This paradigm shifting technology enables simultaneous detection and classification *in vitro*. In another example, a single-molecule sequencing device being developed reports a DNA sequence by a series of light pulses [SCHNEIDER05]. If the device is broken, it can report this by a coding similar to a Hamming code. The general theory of such devices is known [SCHNEIDER91].

- *Understanding the quantitative medical aspects of how cells communicate with neighbors and diseases as communication disruptions.*

Shannon information theory [SCHANNON48, COVER91] enables specification of two inviolable bounds – the lowest information rate needed to faithfully represent a message source, and the highest rate of reliable message delivery through some medium under transmission energy constraints. These bounds are the *source entropy rate* and the *channel capacity*, respectively. This basic communications framework and associated bounds are applicable to almost any scenario comprised of a sender with messages, a channel through which messages can be conveyed, and a receiver that cares about the messages.

In the multicellular context, the message can be any information a sender cell (or group of cells) seeks to convey. Such messages could be cell fate instructions during morphogenesis, tissue maintenance signals which preclude genotypically malignant cells from becoming phenotypically malignant [WEABER97, BISSEL05] or any number of other messages which affect how cells express themselves. Energy constraints arise naturally from signaling compound manufacture costs and signaling apparatus upkeep. The channel is the physics which allows transport of signaling agents and the set of actuators between where the message resides and where actions are taken.

Identification of specific signaling agents and molecular mechanisms is a daunting experimental task and one which already consumes a significant part of the biological research community. However, though details are necessary to understand (and influence) specific biological systems, the beauty of a communications theory approach is that often the detailed methods by which information is conveyed do not affect the bounds on the possible amount of information or how rapidly it can be reliably delivered. For instance, in many telecommunications systems the figure of merit necessary to specify channel capacity is the ratio of available signaling energy to the underlying noise levels at the receiver. This simple “signal to noise ratio” captures all that need be known about message transmission in what might be called a *mechanism-blind* fashion. A variety of similar bounds also exist for networks of communicating elements [COVER91]. It is this generality and implicit reduction of complexity that constitutes the power of a communications-theoretic approach.

Therefore, by analyzing channel physics and using energy-efficiency as a barometer, we expect communications theory to help refine and extend our understanding of multicellular communications and its role in morphogenesis, tissue maintenance, aging and disease – all complex systems where cells communicate in spatially precise ways.

2.4 Novel Computing Machines and Paradigms⁶

For more than half a century, the *von Neumann computer architecture* and the abstract *Turing machine* have largely dominated computer science in many variants and refinements. In view of the upcoming physical limitations of further miniaturization of silicon-based devices [KISH02], we need to ask what the future of these two major paradigms will look like.

The 21st century promises to be the century of information-, bio-, and nanotechnology [ROCO02, BALL00]. Without disruptive new technologies, it is expected that the ever-increasing computing performance and storage capacity achieved with existing technologies will eventually reach a plateau. In order to understand, model, predict, and simulate tomorrow's natural and man-made complex systems, we need computer science to keep going at least at the same pace. Molecular and nanoscale technology and machinery are particularly promising technologies, however, massive investments are required to go beyond the standard models and machines of computation in order to address tomorrow's complex large-scale grand challenges. Whereas traditional silicon-based computers will not simply disappear, there is a growing number of challenges that need to be addressed if

- we want to continue the current pace of progress under Moore's law and
- if we want to open new and unseen application domains and environments for computers.

These broad, long-term, and very challenging quests, in both theoretical and practical dimensions, are further motivated by a certain number of observations and insights (in no particular order):

- there is a huge gap between information processing in nature and in artifacts [BROOKS01,HAREL03,TEUSCHER06], i.e., machines are good where nature isn't and vice versa;
- existing formalisms and computational paradigms are often not suitable to describe, predict, and control the complex information processing in biological, chemical, and physical systems;
- the downfall of Moore's law [KISH02] and sky-rocketing fab costs;
- a drastic increase in complexity and the number of individual components involved due to ongoing miniaturization and novel computing substrates (e.g., nano-scale and molecular electronics);
- a drastic widening of the "design gap," i.e., our ability to design and program computers is not keeping up with the number and the complexity of the individual components available [HENKEL03];
- the increasing need to deal with defect and unreliable components due to miniaturization, novel materials, novel manufacturing techniques, and larger number of components involved;
- the difficulty to program massive parallel and spatial computing systems;
- more dynamic, complex, and harsher environments due to an increasing pervasiveness of computing machines, i.e., the computer literally gets everywhere; and
- an increasing need for adaptive, intelligent, self-configuring, self-diagnosing, self-repairing, self-assembling, self-* machines in order to master the rapidly increasing

⁶ Subsection 2.4 authors: Kobi Benenson (Harvard University), Peter Dittrich (Friedrich-Schiller University Jena, Germany), Jerzy Gorecki (Institute of Physical Chemistry, Warsaw, Poland), Bruce MacLennan (University of Tennessee), Chien-Chung Shen (University of Delaware), Christof Teuscher (Chair) (Los Alamos National Laboratory)

system complexity and to deal with more dynamic, uncertain, and complex environments.

It's all about Communication

Traditional silicon interconnects

In recent years, the importance of interconnects on electronic chips has outrun the importance of transistors as a dominant factor of performance and it is widely conceded that technology alone cannot solve the on-chip global interconnect problem with current design methodologies. With increasing system complexity and the continuing miniaturization of silicon technology, radically new interconnect approaches will be necessary if we want to sustain the current pace of progress [MEINDL03]. Two main factors potentially limit performance [HO01,DAVIS01]:

- the miniaturization of wires, unlike transistors, does not enhance their performance, which is why wires are now more important than transistors [MEIDL03], and
- global wires that communicate signals across the whole chip increase delays and therefore limit the system scalability.

The 2005 ITRS roadmap [ITRS2005] (see also the more recent updates) lists a more detailed number of critical challenges for interconnects. In recent years, true 3D architectures and associated design methodologies have emerged, which offer an attractive option to address some of the current interconnect challenges.

Communication in biological systems

Communication in biological systems differs significantly from communication in traditional computers. While electrical signals provide in general a highly reliable and efficient means to communicate information from a source to a destination, biological systems not only use different transport media but are also much more unreliable and slower in general. For example, chemical communication, where the information carriers are molecules, is very frequently used in biological systems. This type of communication tends to be much slower than communication based on electrical signals, and lots of additional challenges need to be addressed if one wants to efficiently and reliably communicate by means of chemical signals. However, biological systems also make wide use of electrical signals, for example in the brain.

Complex networks and network topologies

Most real networks, such as brain networks [SPORNS04,EGUELUZ05], electronic circuits [CANCO01], the Internet, and social networks share the so-called *small-world* (SW) property [WATTS98]. Compared to purely locally and regularly interconnected networks (such as for example the cellular automata interconnect), small-world networks have a very short average distance between any pair of nodes, which makes them particularly interesting for efficient communication.

The classical Watts-Strogatz small-world network [WATTS98] is built from a regular lattice with only nearest neighbor connections. Every link is then rewired with a *rewiring probability* p to a randomly chosen node. Thus, by varying p , one can obtain a fully regular ($p=0$) and a fully random ($p=1$) network topology. The rewiring procedure establishes “shortcuts” in the network, which significantly lower the average distance (i.e., the number of edges to traverse) between any pair of nodes. In the original model, the length distribution of the shortcuts is uniform since a node is chosen randomly. If the rewiring of the connections is done proportional to a power law, $l^{-\alpha}$, where l is the wire length, then we obtain a *small-world power-law network*. The exponent α affects the network's communication characteristics [KOZMA05] and navigability [KLEINBERG00], which is better than in the uniformly generated small-world network. One can think of other distance-proportional distributions for the rewiring, such as for example a

Gaussian distribution, which has been found between certain layers of the rat's neocortical pyramidal neurons [HELLWIG00]. Studying the connection probabilities and the average number of connections in biological systems, especially in neural systems, can give us important insights on how nearly optimal systems evolved in Nature under limited resources and various other physical constraints.

In a real network, it is fair to assume that local connections have a lower cost (in terms of the associated wire-delay and the area required) than long-distance connections. Physically realizing small-world networks with uniformly distributed long-distance connections is thus not realistic and distance, i.e., the wiring cost, needs to be taken into account, a perspective that recently gained increasing attention [PETERMANN05,PETERANN0]. On the other hand, a network's topology also directly affects how efficient problems can be solved. For example, it has been shown that both small-world [TOMASSINI05] as well as random Erdos-Renyi topologies [MESOT05] offer better performance than regular lattices and are easier to evolve to solve the global synchronization and density classification task, two toy problems commonly used in the cellular automata community.

In summary: there is trade-off between (1) the physical realizability and (2) the communication characteristics for a network topology. A locally and regularly interconnected topology is in general easy to build and only involves minimal wire and area cost, but it offers poor global communication characteristics and scales-up poorly with system size. On the other hand, a random Erdos-Renyi topology scales-up well and has a very short-average path length, but it is not physically plausible because it involves costly long-distance connections established independently of the Euclidean distance between the nodes.

Novel computing paradigms and machines

Unconventional computation (also *non-classical*, *novel*, or *emerging computation*) [ADAMATZKY07,STEPNEY07,MUNAKATA07,COOPER08] is a broad and interdisciplinary research area with the main goal to go beyond the standard models and practical implementations of computers, such as the von Neumann computer architecture and the abstract Turing machine, which have dominated computer science for more than half a century. This undertaking is motivated by a number of trends. First, it is expected that, without disruptive new technologies, the ever-increasing computing performance and storage capacity achieved with existing technologies, will eventually reach a plateau. The main reason for this are fundamental physical limits on the miniaturization of today's silicon-based electronics (see e.g., [KISH02]). Second, novel ways to synthetically fabricate chemical and biological assemblies, for example through self-assembly, self-replication (e.g., [FELLERMANN07]), or bio-engineering (e.g., [TABOR07,BASU05]) allow one to create systems of unimagined complexity. However, we currently lack the methodologies and the tools to design and program such massively parallel and spatially extended unconventional "machines." Third, many of today's most important computational challenges, such as for example understanding complex biological and physical systems by simulations or identifying significant features in large, heterogeneous, and unstructured data sets, may not be well suited for classical computing machines. That is, while a classical Turing-universal computer, at least from a theoretical perspective, can in principle solve all of these challenging problems (as any other algorithmic problem), the general hope is that unconventional computers might solve them much more efficiently, i.e., orders of magnitude faster and using much less resources.

Not everything that looks like a computation in a physical system is useful and not everything that does some information processing can be used to solve problems. A common, if slightly abused, example is that of a falling ball, which can be interpreted as an "unconventional" computer that solves the second order differential equation of Newton's second law. As a matter of fact, a significant research effort has been spent on similar examples, with the goal to

characterize the types of computations, i.e., the laws governing the underlying dynamics behind various physical phenomena. However, a falling ball is a pretty useless computer that can only solve one particular equation with different initial conditions. Interpreting the solution, storing and recalling it, and interfacing the computing unit with other units to perform further computations, is virtually impossible. Thus, while most physical systems solve some equations and most biological organisms process information in some way, we should refrain from calling these systems “computers” until we can harness and interpret the underlying processes with a specific computation in mind. Given a physical, a biological, or a chemical system that is supposed to act as a computer, the question is not only *what*, if anything, this system computes, but also, and more importantly,

- *What are the characteristics of such a computation? (in terms of speed, size, integration density, or power consumption)?*
- *What are the limitations?*
- *What kind of problems can be solved and how efficiently?*
- *How can we “program” the system to perform a specific computation? and*
- *How can we interface the result of the computation with traditional computers to post-process, analyze, and store it?*

Example: non-classical networks-on-chips for future and emerging electronics

Traditionally, on-chip interconnects are designed in a very top-down way, either to form an *ad hoc* network or a regular topology. However, in future bottom-up fabricated circuits, achieving the same regularity and perfection as with top-down devices is generally not straightforward.

Teuscher et al. [TEUSCHER07,TEUSCHER08] have shown that a certain class of irregular and physically plausible 3D interconnect fabrics, which are considered easily and cheaply implementable by self-assembling nanowires or nanotubes, have major advantages over regular fabrics in terms of performance and robustness. The same principles and insights could essentially be applied to biological systems, such as colonies of bacteria or simple cells.

Computation in irregular and heterogeneous self-assemblies of bio and nano components and interconnections is a highly appealing paradigm, both from the perspective of fabrication as well as performance and robustness. This is obviously a radically new technological and unconventional approach with many open questions. For example, there are essentially no methodologies and tools that would allow (1) to map an arbitrary computing architecture or a logical system on an irregularly assembled physical substrate, (2) to do arbitrary computations with such an assembly, and (3) to systematically analyze performance and robustness within a rigorous mathematical framework. There are also many open questions regarding the self-assembling fabrication techniques, which will need to be further explored in the future.

In the following, we will describe in detail two research challenges we believe are highly relevant for the future of molecular communication and computation. The second research challenge described in subsection 2.4.2 can be considered a subset of the first research challenge described in subsection 2.4.1, however, we believe it is important enough to be mentioned separately in this report.

2.4.1 Living Matter as Computing and Communication Media – Paradigms, Design Principles, and Experimental Exploration

2.4.1.1 Objectives

A grand challenge in computer science consists in developing architectures, design methodologies, formal frameworks, and tools that allow to reliably compute and efficiently solve problems with future and emerging devices at the common interface of bio and nano-

technology that are build in a bottom-up instead of a top-down way. It is hypothesized that this would allow to build much more complex systems at a lower cost as the technology becomes more mature. Building a scalable computing architecture on top of a potentially very unreliable physical substrate, such as for example molecular electronics, is a challenging task, which is guided by a number of major trade-offs in the design space, such as the number and the characteristics of the resources available, the required performance, the energy consumption, and the reliability. The lack of systematic understanding of these issues and of clear design methodologies makes the process still more of an art than of a scientific endeavor. We believe the following objectives need to be addressed:

- Understanding and applying the principles of information processing and transmission in natural systems. One needs to understand what information is in the context of biological systems, how it is stored, manipulated, and transmitted. While some information processing is well understood, e.g., cell-cell signaling, or certain neural processing in the brain, biological systems offer many more information processing and communication mechanisms.
- New communication and computing paradigms that respect the media. While binary logic is a perfect fit for silicon electronics, it is not necessarily for biological systems. For example, multi-valued logic or analog computation may be more appropriate. The basic idea is to work *with* the media, and not against it. The more we can harness the dynamics and complexity of an underlying device, the more efficient a system will be on the next level. For example, one can question the approach of building logical gates with a reaction-diffusion media, instead, one should try to remove levels of abstractions to make the system more efficient. In case of a reaction-diffusion system, for example, one would be better off solving differential equations instead of building logical gates, and then trying to solve a problem in a classical algorithmic way.
- Combine theory and experiments to establish viable, application-oriented engineering discipline. Similar to a hardware-software co-design approach, we need to develop a simulation-experiment co-design framework for biological systems. Most computing systems that are based on bio-components are designed in a very *ad hoc* way, which makes it different to go beyond a certain complexity.
- Identify the risks and benefits of bio-computing devices, limit their autonomy, regulate waste and legal issues, educate. The more we use bio-engineering to modify or even build from scratch biological systems, the bigger the risk that something goes wrong and that things get out of control in the worst case. How do we make such systems safe for humans and the environment? How do we dispose waste safely? Should we allow engineered bio-components to be patented?

2.4.1.2 Technical Challenges

- Control and exploit noise, unreliability, and stochasticity. Instead of trying to build perfect systems and to tame noise and unreliability, we should encompass and harness it for our purposes. An example are stochastic optimization algorithms.
- Deal with massive numbers of components. It is generally relatively easy to produce massive numbers of simple components, e.g., by means of chemical self-assembly. However, it is not usually obvious how to harness this complexity for doing computations and communications. Oftentimes, the components are highly unreliable and simple.
- Design methodologies, including validation, simulation, and testing. As for digital logic and silicon-based electronics, we need sophisticated design tools that will enable us to

“engineer” biological systems. An example would be a C-to-molecules compiler, which provides the bio-engineer a high-level abstraction for all the underlying communications and computations.

- Self-* properties. Due to the increasing complexity, it is believed that decentralized and autonomous approaches would help to leverage some of the challenges.
- Develop tools for rapid prototyping and testing. As part of a design flow, we need to develop rapid prototyping tools to accelerate and simplify building bio-molecular systems.
- Actually building the system. Oftentimes, building the actual systems is highly challenging because we lack control. Automated discovery approaches, tools, and design methodologies are needed.

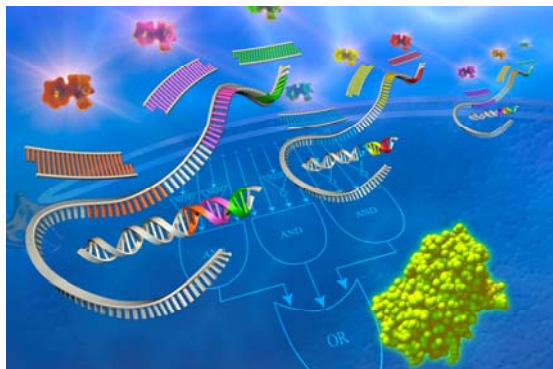


Figure 2: Illustration of biological components used for computations.

Source and copyright: K. Benenson.

2.4.1.3 Impact and Applications

- New hardware and software for hard and massive-scale problems. As opposed to the traditional top-down way of building machines, biology builds organisms in a bottom-up way. If we can draw inspiration from these principles, we will more likely be able to build (or self-assemble) massive-scale systems.
- Enabling technology for bio-molecular devices. Many of the design methodologies and (digital) computing paradigms cannot easily be applied to bio-molecular devices. We need new paradigms, tools, and architectures.
- Interfacing with biological systems. Interfacing silicon logic with biological systems is challenging. If we can perform most, if not all, computations *in vivo*, we could avoid significant interfacing issues.
- Go beyond rational design. More automated design approaches are generally believed to enable the creation of more complex systems.
- New generation of diagnostic and therapeutic tools for personalized medicine. Medical applications are one of the very hot topics for *in vivo* computers. For example, if we can guide medication in an adaptive and autonomous way to the place in the body where it needs to be applied, that would increase the efficiency and decrease the side effects. However, this requires tools and methodologies to build reliable computing machines from highly unreliable bio-components.

2.4.2 Artificial Morphogenesis – Understanding and Engineering a Growing Complex 3D System

2.4.2.1 Objectives

In developmental biology, *morphogenesis* is one of three fundamental aspects, along with cellular growth and differentiation. Here we use the term in a rather loose way to describe bio-inspired computing machines which have a developmental aspect. Biological systems grow, live, adapt, and reproduce, characteristics that are not truly encompassed by any existing computing system. The concept of “living” has a number of consequences in terms of adaptation, interaction with the environment, and the ability to deal with limited resources. Methodologies and technologies that enable the construction of artificial systems that live, grow, adapt, and reproduce in hardware would allow a quantum leap in performance for many computing systems known so far. While evolution, learning, and adaptation have been explored and applied in computer science for many years, developmental aspects have been neglected. An exception is represented by the *Embryonics* project (Embryonics stands for embryonic electronics) [MANGE99,MANGE00,TEUSCHER03], which is inspired by the development of multicellular organisms and by the cellular division and differentiation of living organisms and their stem cells. The final goal of this approach was to develop extremely robust integrated circuits able to self-repair and to self-replicate. The new hardware paradigm, called *autonomous reconfigurable tissue*, is based on a homogeneous, infinitely expandable tissue structured in three layers: an input, an output, and a logic layer. A molecule—the tissue's basic element—consists of a reconfigurable digital circuit. A finite set of molecules makes up a cell, a small processor with an associated memory. The cell is capable of *autonomous self-replication* and can thus create the finite set of cells making up an organism, an application-specific multiprocessor system. The final organism can itself replicate, giving rise to a population of identical organisms, i.e., clones.

While the Embryonics project used silicon components as its building blocks, the grand challenge we propose here is to use bio-molecular components:

- Understand and engineer complex 3D bio-molecular computing and communicating systems. We lack the understanding and engineering principles to gradually assemble complex 3D structures from simple basic components, such as bio-molecules.
- Deep hierarchical structure: We need to be able to assemble systems with detailed, specific structures from the nanoscale up to the macroscale. Animals are examples of such systems and morphogenesis provides examples of how to accomplish this. We face similar challenges in the assembly of artificial brains, robots, etc.
- Develop methods and technologies for controlling growth, shape and functionality of 3D macroscopic bio-systems using engineered or synthetic cells. As for digital logic and silicon-based electronics, we need sophisticated design tools that will enable us to “engineer” developing biological systems.
- Identify the conditions that allow for learning, adaptation, autonomy and self-repair. The level of complexity we would like to achieve in building complex 3D systems requires to have decentralized communications and computations. Learning, adaptation, autonomy, self-repair, etc. are mechanisms that could lead us to that goal.

2.4.2.2 Technical Challenges

- Develop decentralized, adaptive, and robust control mechanisms.
- Ensure far from equilibrium operation.

- Reach brain-/Avogadro-scale (6×10^{23}) complexity.
- Establish principles of molecular communication and computing in large 3D structures.
- Bio-technical micro factories.
- Methods and technologies that limit the autonomy of living technology, e.g., apoptosis.
- Setting regulations for waste, recycling, ethics, legal issues, intellectual property.

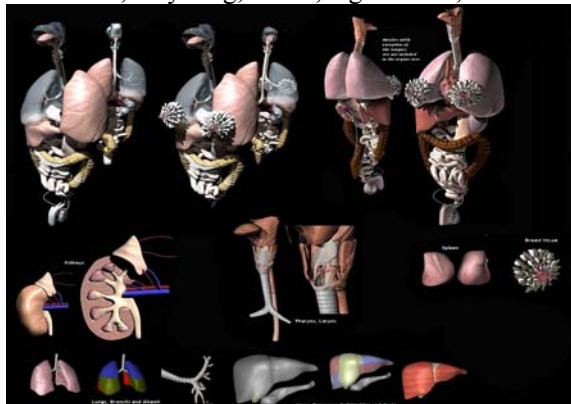


Figure 3: Complex 3D organs.

2.4.2.3 Impact and Applications

- Prosthetic devices and regeneration medicine. If we master the technology of controlling, for example, the growth of human tissue cells, we could regenerate tissue at the right place in a controlled way.
- Micro/nano/molecular robots that grow, self-repair, and self-heal. Medical, sensing, environmental, and many other applications could benefit from autonomous micro, nano, or molecular scale agents that operate autonomously, are able to grow into complex bodies from simple components, and can self-repair if necessary. Naturally, this raises ethical and safety concerns, which need to be addressed as part of this grand challenge.
- Bio-technological micro-factories. Wouldn't it be great if you could synthesize your own personal medication at home, which would be highly optimized for your own body and issues? This would require us to understand how to synthesize the basic molecular building blocks and then to (self-) assemble them into a macroscopic object.
- *In vivo* information processing devices. Instead of injecting read-made information processing devices into a body, we imagine that such devices, especially of a more significant size or for locations that are hard to reach, could gradually be assembled inside the body.

2.5 Biological Communications and its Potential Applications⁷

Cells constantly respond to and process diverse signals from their environment. In particular, they often communicate with each other to carry out certain functions in a coordinated manner.

⁷ Subsection 2.5 authors: Lingchong You (Chair) (Duke University), Anand Asthagiri (California Institute of Technology), William Bentley (University of Maryland, College Park), Cynthia Collins (Rensselaer Polytechnic Institute), Yuki Moritani (NTT DoCoMo), Christopher Rao (University of Illinois at Urbana-Champaign), and John Tyson (Virginia Tech).

The communication process is controlled by complex regulatory networks inside cells. Unlike electrical circuits, cellular networks have to process signals in the presence of “noise,” which is often due to environmental perturbations and low copy numbers of cellular components, among other factors. A better understanding of biological communication will impact broad areas spanning biology and engineering, including the development of cell-based biocomputing and innovative cellular systems for applications in biotechnology, environment, and medicine. This session brought together leading scientists with computational and/or experimental expertise in analyzing biological communication in diverse systems and in using such communication to create innovative artificial systems with broad applications. The key questions addressed by the panel included: What is biological communication? What types of communication are present in biological systems? Are there universal properties of various types of communication? How can improving our understanding of biological communication lead to creation of novel systems with practical applications?

The active discussion among panelists facilitated clarification of the key concepts related to biological communication. Dr. Tyson (Virginia Tech) discussed decomposition of complex cellular networks underlying cell cycle regulation into well-defined regulatory motifs and modules that communicate with each other. Dr. Rao (University of Illinois) described efforts to elucidate dynamics of regulatory motifs underlying virulence development in the bacterium *S. typhimurium*. Dr. Asthagiri (California Institute of Technology) described dynamics involved in cellular differentiation in response to environmental cues and inter-cell communication. Dr. Bentley (University of Maryland) described modeling and experimental analysis of quorum sensing (inter-cell communication based on small, diffusible chemicals) in *E. coli*, as well as ongoing efforts to create novel technologies to analyze bacterial communication. Dr. Collins (RPI) described engineering and evolution of a synthetic microbial community. Mr. Moritani (NTT DoCoMo) discussed efforts to engineer molecular communication between artificial cells based on liposomes, and the potential applications of such technology to medicine. Dr. You (Duke) pointed out that the limited understanding of cellular physiology represents a major challenge in engineering robust cellular behavior using synthetic gene circuits.

2.5.1 Objectives

Through the presentations and discussions, the panel recognized the tremendous diversity of biological communication in different forms, which include communication between molecules, that between network modules, that between cells, and that between populations.

The overall goal of this suggested direction is to understand dynamics and regulation of biological communication at multiple time and length scales and, based on such understanding, to create systems of practical applications. It is also important to create the platform technologies, such as microfluidics, imaging, and biomaterials, to enable the proposed investigations. Specifically, the panel identified several key objectives for next step research:

- *Elucidate design principles of biological communication and information processing.* This should be carried out in different systems from single-celled organisms (e.g. bacteria, yeast) to multicellular systems (mammalian cells, plant cells, stem cells). The key challenge is to identify unifying, portable dynamic properties of regulatory networks across multiple time and length scales [TYSON03][GIURUMESCU08][LI06].
- *Integrate biological substrates and systems with “classical” communications modalities* (cell phones, laptops, microfabricated devices). Engineering of molecular communication between artificial cells (based on liposomes) represents an excellent example. Also, integration of cellular systems with culturing and imaging system represents another promising direction.

- *Translate biological communication architectures to other existing networks and generate new architectures and modes for communication.* One immediate application is to apply design principles identified in natural systems to the engineering of artificial networks.
- *Develop novel applications by exploiting biological communications.* For instance, engineered synthetic microbial community may serve as a robust platform for metabolic engineering or bioremediation [BRENNER08].

2.5.2 Technical Challenges

Given the diversity of communication systems in terms of their implementation and scales, many challenges need to be overcome to make progress in the identified research directions.

- *Identify and characterize regulatory modules at multiple levels.* Often times, modules are in the eyes of beholder. In order to develop deeper understanding of inter-module communication/interaction, a clear, unambiguous definition of modules should be developed. This may require development of metrics of modules [DEL08].
- *Develop interfaces between electronic and biological communication.* This is critically important for developing hybrid systems that encompass biological and other modes of communication and information processing.
- *Controlling and exploiting biological uncertainty.* The uncertainty includes the stochastic fluctuations involved in biological processes [RAO02] and the limited understanding of cell physiology [MARGUET07]. Such uncertainty will pose a significant challenge for engineering robust behavior using synthetic gene circuits, in single cells or in populations. On the other hand, the uncertainty may be exploited by natural or engineered cellular networks for specific functions.
- *Integrating diverse disciplines to understand problem and cultivate solutions and applications.* Often researchers with different background approach a common set of problems with divergent perspectives. The “cultural differences” may be difficult to reconcile at times. For instance, experimentalists may fail to recognize the potential and limits of theory, whereas theorists may fail to appreciate the complexity intrinsic to an experimental system. Also, engineers and computer scientists may fail to appreciate the ‘messiness’ of living, evolved communication systems.
- *Foundational technologies for analyzing and manipulating biological communication networks experimentally and computationally.* These include methods to engineer biological components (e.g., directed evolution as described in [COLLINS06]), observe communication (e.g., photonics), “wire” biological communication onto “classical” electron-mediated devices.
- *Modeling formalisms and software support platforms.* There may be niche for developing modeling platforms specific for computational challenges associated with biological communication and its engineered counterpart, similar to community efforts to develop the Systems Biology Markup Language for generic model representation and formulation (<http://www.sbml.org>).

2.5.3 Impact and Applications

Investment in research along this direction has tremendous potential pay off, which includes innovative systems with applications in a wide variety of directions. These applications in turn will define a paradigm to inspire future generation of scientists and engineers, trained in a multidisciplinary setting. Examples of applications include:

- *Biotechnology* – biofilms, protein engineering, bioreactors, cellular engineering.
- *Energy* – microbial fuel cells, electron transport
- *Biosensing/actuation* – exploit natural recognition capabilities, new recognition elements, transduction modalities, and actuation (e.g. biohybrid devices).
- *Hybrid communications* devices that incorporate biological elements.
- *Cell-based biocomputation*.
- *Medicine* – new therapeutic strategies based on in-depth understanding of cellular regulation or engineering of innovative systems.
- *Environment* – Environmental testing, improving sustainability, new synthesis processes that are environmentally benign, efficient, and less costly
- *Education* – new paradigm to educate and inspire future, interdisciplinary engineers and scientists.

3. Suggestions to NSF⁸

We are experiencing a “composite revolution” [ROCO02, BALL00] where the convergence of various sciences, along with their own related inspirations, is more likely to lead us to the destination we seek than any single one of them can, the session chair, Dr. Teuscher of Los Alamos National Laboratory says. Non-classical computation described in section 2.4 is a good example, which resides at the interface of various research areas, such as biology, chemistry, computer science, physics, and material science.

Other grand research challenges described in section 2 are equally highly interdisciplinary, requiring educated personnel that cover a wide range of capabilities. The standard NSF funding grants are not likely to be very effective. Here the workshop attendees suggest how NSF should promote interdisciplinary education and support interdisciplinary research in order to advance this technology field. A summary of their suggestions on interdisciplinary education and research is provided in the following.

3.1 Promoting Interdisciplinary Education

- Promote interdisciplinary education, vertically integrated from undergraduate to junior faculty. Provide incentives for young faculty (and their universities) to work in interdisciplinary areas and be evaluated on the basis of their interdisciplinary research output.
- Educate a new generation of students, postdoctoral researchers, and junior faculty in (a) computer science as well as (b) biological science.
 - (a) The current computer science curriculum is not sufficient and appropriate anymore for the upcoming challenges in this area. The CS curriculum is too focused on silicon-based technology and not interdisciplinary enough to deal with novel computing devices and substrates. A new curriculum would likely also get some of the excitement back to computer science and would help to improve the alarmingly decreasing number students over the past years.
 - (b) Most current molecular biologists lack a basic understanding of information theory. Even as recently as January of 2008 a referee for a paper asked “what does ‘bits’

⁸ Section 3 has been summarized by the five session chairs; Ari Requicha (University of Southern California), Michael Simpson (Oak Ridge National Laboratory/University of Tennessee), Thomas D. Schneider (National Institute of Health), Christof Teuscher (Los Alamos National Laboratory), Lingchong You (Duke University), and edited by Tadashi Nakano (University of California, Irvine).

mean?” indicating that even the most basic concepts of information theory are not wide spread despite many papers showing the usefulness, clarity and universality of information theory applied to binding sites, the session chair, Dr. Schneider of National Institute of Health, says. These issues can be addressed by increasing the training at all levels in the application of mathematics to biology. Training in mathematical biology must begin at the undergraduate level in order to provide sufficient time for instruction in probability theory, information theory and coding theory as applied to biological problems, in addition to computer science, thermodynamics, chemistry, physics and biology. An undergraduate degree in mathematical biology may require the number of hours usually completed for a double major.

- Textbooks are needed that integrate multiple disciplines (e.g., molecular biology and evolution with information theory and coding theory, and all levels of biology with probability theory and statistics.)
- Promote to standardize nomenclatures and develop ontologies in order to bring together the various disciplines to the area. These efforts can eliminate the confusion when various disciplines use the same words differently.
- Beyond training, it will eventually be important to unify multiple disciplines as a field of study by establishing university departments of the unified disciplines, thus creating a clear career path from undergraduate and graduate programs to faculty fully devoted to teaching, publishing and obtaining grants in the unified disciplines.
- Initiate IGERT, REU site, EFRI, ERC, and other mechanisms that relate to and support research on biological communications technology.

3.2 Supporting Interdisciplinary Research

- Create multidisciplinary programs that include proposals *and* panelists from different directorates.
- Support interdisciplinary research groups at a substantial level and for a significant period to tackle grand challenges—for example, at ~ \$1 M/year, for 5 years. Interdisciplinary projects usually have a slow start because of communication and cultural problems across disciplines; when an effective group has been painstakingly assembled, the usual 2-3 year project duration is almost finished and the group is about to be dismantled. This is not a cost-effective approach. An increased level of funding over standard grants is necessary because of the many disciplines and associated personnel involved. For the grand challenges proposed in section 2, expertise is needed from biology, chemistry, computer science, electrical engineering, materials science and physics.
- Establish a mechanism for follow-up funding, which allows an interdisciplinary research group to work towards a challenging goal that is 10-20 years.
- Support postdoctoral researchers. These play a very important role in these projects by providing a continuity and know-how that is difficult or impossible to maintain only with graduate students.
- Support high-risk and high pay-off projects. The EU has maintained several highly programs in the area of Future and Emerging Technologies (FET). There is clearly a need to fund transformative science which takes more risk than usual for NSF projects, but which will lead to major breakthroughs if successful.
- Fund grand challenge oriented proposals/ideas, which allow to push an area and to solve challenging problems systematically over multiple years. Traditional peer-review might not be appropriate in that case. In the multidisciplinary field of biological communications technology, a real impact is more likely if the effort is concentrated.

- Provide access to infrastructure, either through new initiatives or by coordination with existing ones. The following are commonly needed facilities for the interdisciplinary research of biological communications technology: micro/nano fabrication and characterization facilities; high performance computing; biological model systems and experimental data.
- Promote and organize integrative conferences and workshops to make focused progress and to disseminate results.
- Establish mechanisms for world-wide collaborations, e.g., joint EU/NSF projects. That would likely result in more high-profile projects, a broader and deeper impact, and in more qualified teams.

4. References

- [ABSHIRE01] P. A. Abshire and A. G. Andreou, "Capacity and energy cost of information in biological and silicon photoreceptors," *Proceedings of the IEEE*, vol. 89, no. 7, pp. 1052–1064, July 2001, (Invited Paper).
- [ADAMATZKY07] A. Adamatzky, L. Bull, B. De Lacy Costello, S. Stepney, and C. Teuscher, editors. *Unconventional Computing 2007*. Luniver Press, Beckington, UK, 2007.
- [ALON03] Alon, U. *Biological Networks: The Tinkerer as an Engineer*. *Science* 301, 1866–1867 (2003).
- [ANDERSON72] Anderson, P. W. *More Is Different - Broken Symmetry and Nature of Hierarchical Structure of Science*. *Science* 177, 393-& (1972).
- [ANDRIAN06] Andrianantoandro, E., Basu, S., Karig, D. K. & Weiss, R. *Synthetic biology: new engineering rules for an emerging discipline*. *Mol Syst Biol* 2 (2006).
- [ATICK92] J. J. Atick, "Could information theory provide an ecological theory of sensory processing?" *Network*, vol. 3, pp. 213–251, 1992.
- [ARKIN98] Arkin, A., Ross, J. & McAdams, H. H. *Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected Escherichia coli cells*. *Genetics* 149, 1633–1648 (1998).
- [ATTNEAVE54] F. Attneave, "Some informational aspects of visual perception," *Psychological Review*, vol. 61, pp. 183–193, 1954.
- [AUSTIN06] Austin, D. W. et al. *Gene network shaping of inherent noise spectra*. *Nature* 439, 608–611 (2006).
- [BALL00] P. Ball. *Chemistry meets computing*. *Nature*, vol. 406, 118–120, 2000
- [BARABASI99] Barabasi, A. L. & Albert, R. *Emergence of Scaling in Random Networks*. *Science* 286, 509–512 (1999).
- [BAR-EVEN06] Bar-Even, A. et al. *Noise in protein expression scales with natural protein abundance*. *Nat Genet* 38, 636–643 (2006).
- [BATTAIL06] G. Battail, "Should genetics get an information-theoretic education?" *Engineering in Medicine and Biology Magazine*, vol. 25, pp. 34–45, Jan.-Feb. 2006.
- [BASU05] S. Basu, Y. Gerchman, C. H. Collins, F. H. Arnold, and R. Weiss. *A synthetic multicellular system for programmed pattern formation*. *Nature*, 434(7037):1130–1134, 2005.
- [BATALIN04] M. Batalin and G. S. Sukhatme, "Coverage, Exploration and Deployment by a Mobile Robot and Communication Network," *Telecommunication Systems Journal*, Special Issue on Wireless Sensor Networks, Vol. 26, No. 2, pp. 181–196, 2004.
- [BARLOW61] H. B. Barlow, "Possible principles underlying the transformations of sensory messages," in *Sensory Communication*, W. A. Rosenblith, Ed. Cambridge, MA: MIT Press, 1961, pp. 217–234.

- [BHZ06] Beyond the horizon: Anticipating future and emerging information society technologies. Initiative within the "Future and Emerging Technologies" (FET) program of the European Commission, 2006. <http://www.beyond-the-horizon.net>
- [BINDEWALD06] E. Bindewald, T. D. Schneider, and B. A. Shapiro, "CorreLogo: An online server for 3D sequence logos of RNA and DNA alignments," *Nucleic Acids Res.*, vol. 34, pp. w405–w411, 2006, <http://www.ccrnp.ncifcrf.gov/~toms/papers/correlogo/>.
- [BISSELL05] M. J. Bissell and M. A. Labarge, "Context, tissue plasticity, and cancer: are tumor stem cells also regulated by the microenvironment?" *Cancer Cell*, vol. 7, pp. 17–23, 2005.
- [BRROU93] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding and decoding: Turbo-codes," *Proc. of IEEE*, vol. 2, pp. 1064–1070, May 1993.
- [BRENNER08] Brenner, K., L. You, and F. H. Arnold (2008). Engineering microbial consortia: a new frontier for synthetic biology. Trends in Biotechnology. In press
- [BROOKS01] R. Brooks. The relationship between matter and life. *Nature*, 409(6818):409–411, January 18 2001.
- [CANCHO01] R. Ferrer i Cancho, C. Janssen, and R. V. Sole. Topology of technology graphs: small world patterns in electronic circuits. *Phys. Rev. E* 64, 046119, 2001
- [CHEN07] Z. Chen, K. A. Lewis, R. K. Shultzaberger, I. G. Lyakhov, M. Zheng, B. Doan, G. Storz, and T. D. Schneider, "Discovery of Fur binding site clusters in *Escherichia coli* by information theory models," *Nucleic Acids Res.*, vol. 35, pp. 6762–6777, 2007.
- [CIT86] *Coding and Information Theory*. Englewood Cliffs NJ: Prentice-Hall, Inc., 1986.
- [CNLS07] Unconventional computation: Quo vadis? Santa Fe, NM, USA, March 21-23 2007. <http://cnls.lanl.gov/uc07>
- [CORBIN02] P. S. Corbin, L. J. Lawless, Z. Li, Y. Ma, M. J. Witmer, and S. C. Zimmerman, "Discrete and polymeric self-assembled dendrimers: hydrogen bond-mediated assembly with high stability and high fidelity," *Proc Natl Acad Sci U S A*, vol. 99, pp. 5099–5104, 2002.
- [COVER91] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. N. Y.: John Wiley & Sons, Inc., 1991.
- [COMPUTING06] 2020 – Future of Computing. *Nature*, 440:398-419, March 2006. <http://www.nature.com/nature/focus/futurecomputing>
- [COLLINS06] Collins CH, Leadbetter JR, Arnold FH. Dual selection enhances the signaling specificity of a variant of the quorum-sensing transcriptional activator LuxR. *Nat Biotechnol.* 2006 Jun;24(6):708
- [COOPER08] S. B. Cooper, B. Lowe, and A. Sorbi, editors. *New Computational Paradigms*. Springer, New York, NY, USA, 2008.
- [CSETE04] Csete, M. & Doyle, J. Bow ties, metabolism and disease. *Trends in Biotechnology* 22, 446-450 (2004).
- [CUBUKCU06] E. Cubukcu, E. A. Cort, K. B. Crozier and F. Capasso, "Plasmonic laser antenna", *Applied Physics Letters*, Vol. 89, 093120, 2006.
- [CURRELI05] M. Curreli, C. Li, Y. Sun, B. Lei, M. A. Gundersen, M. E. Thompson and C. Zhou, "Selective functionalization of In₂O₃-nanowire mat devices for biosensing applications", *J. American Chemical Society*, Vol. 127, No. 19, pp. 6922-6923, 18 May 2005.
- [DANCOFF53] S. M. Dancoff and H. Quastler, "The information content and error rate of living things," in *Essays on the Use of Information Theory in Biology*, H. Quastler, Ed. Urbana: University of Illinois Press, 1953, pp. 263–273.
- [DAVIS01] J. A. Davis, R. Venkatesan, A. Kaloyeros, M. Beylansky, S. J. Souri, K. Banerjee, K. C. Saraswat, A. Rahman, R. Reif, and J. D. Meindl. Interconnect limits on gigascale integration (GSI) in the 21st century. *Proceedings of the IEEE*, 89(3):305–324, 2001.

- [DELGRANGE99] O. Delgrange, M. Dauchet, and E. Rivals, "Location of repetitive regions in sequences by optimizing a compression method," in *Pac. Symp. Biocomput.*, 1999, pp. 254–65.
- [DEL08] Del Vecchio D, Ninfa AJ, Sontag ED. Modular cell biology: retroactivity and insulation. *Mol Syst Biol.* 2008;4:161
- [DOONAIL02] D. A. Mac Dónaill, "A parity code interpretation of nucleotide alphabet composition," *Chem Commun (Camb)*, vol. 18, pp. 2062–2063, 2002.
- [DOONAIL03] D. A. Mac Dónaill, "Why nature chose A, C, G and U/T: an error-coding perspective of nucleotide alphabet composition," *Orig Life Evol Biosph*, vol. 33, pp. 433–455, 2003.
- [DOYLE07] Doyle, J. & Csete, M. Rules of engagement - Complex engineered and biological systems share protocol-based architectures that make them robust and evolvable, but with hidden fragilities to rare perturbations. *Nature* 446, 860-860 (2007).
- [DUNN07] Dunn, A. K. & Stabb, E. V. Beyond quorum sensing: the complexities of prokaryotic parliamentary procedures. *Analytical and Bioanalytical Chemistry* 387, 391-398 (2007).
- [ECKFORD07] A. W. Eckford, "Nanoscale communication with Brownian motion," *Proc. 41st Conference on Information Sciences and Systems*, pp. 160–165, 2007.
- [EGUEL05] V. M. Eguéluz, D. R. Chialvo, G. A. Cecchi, M. Baliki, and A. V. Apkarian. Scale-free brain functional networks. *Phys. Rev. Letters*, 94, 018102, 2005.
- [EIGEN93] M. Eigen, "The origin of genetic information: viruses as models," *Gene*, vol. 135, pp. 37–47, 1993.
- [ELOWITZ02] Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. *Science* 297, 1183-1186 (2002).
- [FEYNMAN60] Feynman, R. P. There's Plenty of Room at the Bottom. *Engineering and Science* 23, 22-36 (1960).
- [FELLERMANN07] H. Fellermann, S. Rasmussen, H.-J. Ziock, and R. V. Sole. Life cycle of a minimal protocell—a dissipative particle dynamics study. *Artificial Life*, 13:319–345, 2007.
- [FORSDYKE81] D. R. Forsdyke, "Are introns in-series error-detecting sequences?" *J. Theor. Biol.*, vol. 93, pp. 861–866, 1981.
- [FORSTER06] Forster, A. C. & Church, G. M. Towards synthesis of a minimal cell. *Mol Syst Biol* 2 (2006).
- [GARAVELLI04] J. S. Garavelli, "The RESID Database of Protein Modifications as a resource and annotation tool," *Proteomics*, vol. 4, pp. 1527–1533, 2004.
- [GATLIN72] L. L. Gatlin, *Information Theory and the Living System*. New York, NY: Columbia University Press, 1972.
- [GASTPAR03] M. Gastpar, B. Rimoldi, and M. Vetterli, "To Code, or Not to Code: Lossy Source—Channel Communication Revisited," *IEEE Transactions on Information Theory*, vol. 49, pp. 1147–1158, 2003.
- [GIURUMESCU08] Giurumescu, C.A. and A.R. Asthagiri. Signal processing during developmental multicellular patterning. *Biotech. Prog.*, (2008). 24:80.
- [GOLDBERG07] D. H. Goldberg and A. G. Andreou, "Distortion of neural signals by spike coding," *Natural Comput*, vol. 19, pp. 2797-2839, 2007.
- [GOLDBERG04] D. Goldberg and A. G. Andreou, "Spike communication of dynamic stimuli: rate decoding versus temporal decoding," *Neurocomputing*, vol. 58-60, pp. 101–107, June 2004.
- [GOLDBERG03] D. Goldberg, A. Shripati, and A. G. Andreou, "Energy efficiency in a channel model for the spiking axon," *Neurocomputing*, vol. 52-54, pp. 39–44, June 2003.
- [GUIZZO04] E. Guizzo, "Closing in on the perfect code," *IEEE Spectrum*, vol. 41, no. 3, pp. 36–42, March 2004.
- [GOLDING05] Golding, I., Paulsson, J., Zawilski, S. M. & Cox, E. C. Real-time kinetics of gene activity in individual bacteria. *Cell* 123, 1025-1036 (2005).

- [GOLDING06] Golding, I. & Cox, E. C. Physical nature of bacterial cytoplasm. *Physical Review Letters* 96 (2006).
- [GUIDO06] Guido, N. J. et al. A bottom-up approach to gene regulation. *Nature* 439, 856-860 (2006).
- [HAMMING50] R. W. Hamming, "Error detecting and error correcting codes," *Bell System Tech. J.*, vol. 26, pp. 147-160, 1950.
- [HARTLEY28] R. V. L. Hartley, "Transmission of information," *Bell System Tech. Journ.*, pp. 535-563, 1928.
- [HAREL03] D. Harel. *Computers Ltd.: What They Really Can't Do*. Oxford University Press, 2003.
- [HASTY02] Hasty, J., McMillen, D. & Collins, J. J. Engineered gene circuits. *Nature* 420, 224-230 (2002).
- [HELLWIG00] B. Hellwig. A quantitative analysis of the local connectivity between pyramidal neurons in layers 2/3 of the rat visual cortex. *Biological Cybernetics*, 82:111-121, 2000.
- [HENKEL03] J. Henkel. Closing the SoC design gap. *IEEE Computer*, 36(9):119-121, 2003.
- [HENGEN97] P. N. Hengen, S. L. Bartram, L. E. Stewart, and T. D. Schneider, "Information analysis of Fis binding sites," *Nucleic Acids Res.*, vol. 25, no. 24, pp. 4994-5002, 1997, <http://www.ccrnp.ncifcrf.gov/~toms/paper/fisinfo/>.
- [HO01] R. Ho, K. W. Mai, and M. A. Horowitz. The future of wires. *Proceedings of the IEEE*, 89(4):490-504, 2001.
- [HULO06] N. Hulo, A. Bairoch, V. Bulliard, L. Cerutti, E. De Castro, P. S. Langendijk-Genevaux, M. Pagni, and C. J. Sigrist, "The PROSITE database," *Nucleic Acids Res*, vol. 34, pp. D227-230, 2006.
- [INUI08] H. Inui, K. S. Oh, C. Nadem, T. Ueda, S. G. Khan, A. Metin, E. Gozukara, S. Emmert, H. Slor, D. B. Busch, C. C. Baker, J. J. Digiovanna, D. Tamura, C. S. Seitz, A. Gratchev, W. H. Wu, K. Y. Chung, H. J. Chung, E. Azizi, R. Woodgate, T. D. Schneider, and K. H. Kraemer, "Xeroderma Pigmentosum-Variant Patients from America, Europe, and Asia," *J Invest Dermatol*, 2008.
- [ITMB92] *Information Theory in Molecular Biology*. Cambridge: Cambridge University Press, 1992.
- [ITE05] *Information Theory, Evolution, and The Origin of Life*. Cambridge: Cambridge University Press, 2005.
- [ITRS2005] International technology roadmap for semiconductors (ITRS). Semiconductor Industry Association, 2005. <http://www.itrs.net/Common/2005ITRS/Home2005.htm>
- [JAYNES03] E. T. Jaynes, *Probability Theory: The Logic of Science*, G. L. Bretthorst, Ed. Cambridge University Press, 2003.
- [JENSEN07] K. Jensen, J. Weldon, H. Garcia and A. Zettl, "Nanotube radio", *Nanoletters*, Vol. 7, No. 11, pp. 3508-3511, July 2007.
- [JEONG00] Jeong, H., Tombor, B., Albert, R., Oltval, Z. N. & Barabasi, A. L. The large-scale organization of metabolic networks. *Nature* 407, 651-654 (2000).
- [KAY00] L. E. Kay, *Who Wrote the Book of Life?: A History of the Genetic Code*. Stanford University Press, 2000.
- [KISH02] L. B. Kish. End of Moore's law: Thermal (noise) death of integration in micro and nano electronics. *Physics Letters A*, 305:144-149, 2002.
- [KLEINBERG00] J. K. Kleinberg. Navigation in a small world. *Nature*, 406:845, 2000.
- [KOZMA05] B. Kozma, M. B. Hastings, and G. Korniss. Diffusion processes on power-law small-world networks. *Physical Review Letters*, 95:018701, 2005.
- [KRAWCZAK92] M. Krawczak, J. Reiss, and D. N. Cooper, "The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences," *Hum Genet*, vol. 90, pp. 41-54, 1992.

- [KROGH94] A. Krogh, M. Brown, I. S. Mian, K. Sjölönder, and D. Haussler, “Hidden Markov models in computational biology, applications to protein modeling,” *J. Mol. Biol.*, vol. 235, pp. 1501–1531, 1994.
- [LAUGHLIN00] Laughlin, R. B., Pines, D., Schmalian, J., Stojkovic, B. P. & Wolynes, P. The middle way. Proceedings of the National Academy of Sciences of the United States of America 97, 32-37 (2000).
- [LEHN02] J. M. Lehn, “Toward complex matter: supramolecular chemistry and self-organization,” *Proc Natl Acad Sci U S A*, vol. 99, pp. 4763–4768, 2002.
- [LEVY96] W. B. Levy and R. A. Baxter, “Energy efficient neural codes,” *Neural Computation*, vol. 8, pp. 531–543, 1996.
- [LIEBOVITCH96] L. S. Liebovitch, Y. Tao, A. Todorov, and L. Levine, “Is there an Error Correcting Code in DNA?” *Biophysical Journal*, vol. 71, pp. 1539–1544, 1996.
- [LI06] Li, J., Wang, L., Hashimoto, Y., Tsao, C-Y., Wood, T.K., Valdes, J.J., Zafiriou, E., and Bentley, WE. Stochastic Model of E. coli AI-2 Quorum Signal Circuit Reveals Alternative Synthesis Pathways, *Molecular Systems Biology*. (2006)
- [LLOYD00] Lloyd, S. Ultimate physical limits to computation. *Nature*, 406:1047–1054, 2000
- [LÖYTYNOJA08] A. Löytynoja and N. Goldman, “Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis,” *Science*, vol. 320, pp. 1632–1635, 2008.
- [LOEWENSTERN99] D. Loewenstern and P. N. Yianilos, “Significantly lower entropy estimates for natural DNA sequences,” *J Computational Biology*, vol. 6, no. 1, pp. 125–42, 1999.
- [LOEWENSTERN98] D. M. Loewenstern, H. M. Berman, and H. Hirsh, “Maximum a posteriori classification of DNA structure from sequence information,” in *Pac. Symp. Biocomput.*, 1998, pp. 669–80.
- [LUI06] Luisi, P. L., Ferri, F. & Stano, P. Approaches to semi-synthetic minimal cells: a review. *Naturwissenschaften* 93, 1-13 (2006).
- [LYAKHOV01] I. G. Lyakhov, P. N. Hengen, D. Rubens, and T. D. Schneider, “The P1 Phage Replication Protein RepA Contacts an Otherwise Inaccessible Thymine N3 Proton by DNA Distortion or Base Flipping,” *Nucleic Acids Res.*, vol. 29, no. 23, pp. 4892–4900, 2001, <http://www.ccrnp.ncifcrf.gov/~toms/paper/repan3/>.
- [MACKAY52] D. M. MacKay and W. S. McCulloch, “The limiting information capacity of a neuronal link,” *Bulletin of Mathematical Biophysics*, vol. 14, pp. 127–135, 1952.
- [MAINI04] P. K. Maini, S. Schnell, and S. Jolliffe, “Bulletin of Mathematical Biology—facts, figures and comparisons,” *Bull Math Biol*, vol. 66, pp. 595–603, 2004.
- [MARGUET07] Marguet, P., F. Balagadde, C. Tan, and L. You (2007). Biology by design: reduction and synthesis of cellular components and behavior. *J. Royal Society Interface*. 4:607-623.
- [MANGE99] D. Mange, M. Sipper, and P. Marchal. Embryonic electronics. *BioSystems*, 51(3):145–152, September 1999.
- [MANGE00] D. Mange, M. Sipper, A. Stauffer, and G. Tempesti. Toward robust integrated circuits: The embryonics approach. Proceedings of the IEEE, 88(4):516–540, April 2000.
- [MAY08] E. May, P. Dolan, P. Crozier, S. Brozik, and M. Manginell, “Towards De novo design of Deoxyribozyme Biosensors for GMO detection,” *IEEE Sensors Journal*, vol. June, in press, 2008.
- [MAY07] E. May, “Error Control Codes and the Genome,” in *Genomics and Proteomics Engineering in Medicine and Biology*, M. Akay, Ed., 2007, pp. 173–208.
- [MAY06a] E. May, P. Dolan, P. Crozier, and S. Brozik, “Syndrome-based discrimination of single nucleotide polymorphism,” in *IEEE Engineering in Medicine and Biology Society International Conference*, 2006, pp. 4548–4551.

- [MAY06b] E. E. May, M. A. Vouk, and D. L. Bitzer, "Classification of *Escherichia coli* K-12 ribosome binding sites. An error-control coding model," *IEEE Eng Med Biol Mag*, vol. 25, no. 1, pp. 90–97, 2006.
- [MAY04a] E. E. May, M. A. Vouk, D. L. Bitzer, and D. I. Rosnick, "Coding theory based models for protein translation initiation in prokaryotic organisms," *Biosystems*, vol. 76, pp. 249–260, 2004.
- [MAY04b] E. E. May, M. A. Vouk, D. L. Bitzer, and D. I. Rosnick, "An error-correcting code framework for genetic sequence analysis," *Journal of the Franklin Institute*, vol. 341, pp. 89–109, 2004.
- [MAY03] E. E. May, A. M. Johnston, W. E. Hart, J.-P. Watson, R. J. Pryor, and M. D. Rintoul, "Detection and reconstruction of error control codes for engineered and biological regulatory systems," *SAND REPORT*, pp. 1–42, October 2003, Sandia National Laboratories, Albuquerque, New Mexico, SAND2003-3963, <http://www.cs.sandia.gov/~wehart/Papers/2003/MayJohHarWatPryRin03-sand.pdf>.
- [MEINDL03] J. D. Meindl. Interconnect opportunities for gigascale integration. *IEEE Micro*, 23(3):28–35, 2003.
- [MESOT05] B. Mesot and C. Teuscher. Deducing local rules for solving global tasks with random Boolean networks. *Physica D*, 211(1–2):88–106, 2005.
- [MUNAKATA07] T. Munakata (guest editor). Beyond silicon: New computing paradigms. *Communications of the ACM*, 50(9):30–72, Sep 2007.
- [MUNAKATA07] T. Munakata (guest editor). Beyond silicon: New computing paradigms. *Communications of the ACM*, 50(9):30–72, Sep 2007.
- [NEWMAN06] Newman, J. R. S. et al. Single-cell proteomic analysis of *S-cerevisiae* reveals the architecture of biological noise. *Nature* 441, 840-846 (2006).
- [PAPP04] P. P. Papp and D. K. Chattoraj, "Missing-base and ethylation interference footprinting of P1 plasmid replication initiator," *Nucleic Acids Res.*, vol. 22, pp. 152–157, 1994.
- [PAPP93] P. P. Papp, D. K. Chattoraj, and T. D. Schneider, "Information analysis of sequences that bind the replication initiator RepA," *J. Mol. Biol.*, vol. 233, pp. 219–230, 1993.
- [PAMELA01] Pamela A. Abshire and Andreas G. Andreou, "A communication channel model for information transmission in the blowfly photoreceptor," *Biosystems Journal*, vol. 62, pp. 113–133, 2001.
- [PETERMANN05] T. Petermann and P. De Los Rios. Spatial small-world networks: A wiring-cost perspective. <http://arXiv:cond-mat/0501420>, 2005.
- [PETERMANN06] T. Petermann and P. De Los Rios. Physical realizability of small-world networks. *Physical Review E*, 73:026114, 2006.
- [PIERCE80] J. R. Pierce, *An Introduction to Information Theory: Symbols, Signals and Noise*, 2nd ed. NY: Dover Publications, Inc., 1980.
- [POWELL98] D. R. Powell, D. L. Dowe, L. A. L., and T. I. Dix, "Discovering simple DNA sequences by compression," in *Pac. Symp. Biocomput.*, 1998, pp. 597–608.
- [QUASTLER58] H. Quastler, "The domain of information theory in biology," in *Symposium on Information Theory in Biology*, H. P. Yockey, R. L. Platzman, and H. Quastler, Eds. New York: Pergamon Press, 1958, pp. 187–196.
- [RAVASZ02] Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. & Barabasi, A.-L. Hierarchical Organization of Modularity in Metabolic Networks. *Science* 297, 1551-1555 (2002).
- [RAO02] Rao, C., D. Wolf, and A. Arkin, Control, exploitation and tolerance of intracellular noise, *Nature*, 420, 231-237, (2002).
- [ROGAN98] P. K. Rogan, B. M. Faux, and T. D. Schneider, "Information analysis of human splice site mutations," *Human Mutation*, vol. 12, pp. 153–171, 1998, erratum in: *Hum Mutat* 1999;13(1):82. <http://www.ccrnp.ncifcrf.gov/~toms/paper/rfs/>.

- [ROGAN95] P. K. Rogan and T. D. Schneider, "Using information content and base frequencies to distinguish mutations from genetic polymorphisms in splice junction recognition sites," *Human Mutation*, vol. 6, pp. 74–76, 1995, <http://www.ccrnp.ncifcrf.gov/~toms/paper/colonsplice/>.
- [ROSEN06] G. Rosen, "Examining coding structure and redundancy in DNA," *IEEE Engineering in Medicine and Biology Magazine*, vol. 25, pp. 62–68, 2006.
- [ROCO02] M. C. Roco and W. S. Bainbridge, editors. *Converging Technologies for Improving Human Performance: Nanotechnology, Biotechnology, Information Technology and Cognitive Science*. World Technology Evaluation Center (WTEC), Arlington, Virginia, June 2002. NSF/DOC-sponsored report. <http://www.wtec.org/ConvergingTechnologies>
- [SCHNEIDER91a] T. D. Schneider, "Theory of molecular machines. I. Channel capacity of molecular machines," *J. Theor. Biol.*, vol. 148, pp. 83–123, 1991, <http://www.ccrnp.ncifcrf.gov/~toms/paper/ccmm/>.
- [SCHNEIDER91b] Thomas D. Schneider, "Theory of molecular machines. II. Energy dissipation from molecular machines," *J. Theor. Biol.*, vol. 148, pp. 125–137, 1991, <http://www.ccrnp.ncifcrf.gov/~toms/paper/edmm/>.
- [SCHNEIDER94] T. D. Schneider, "Sequence logos, machine/channel capacity, Maxwell's demon, and molecular computers: a review of the theory of molecular machines," *Nanotechnology*, vol. 5, pp. 1–18, 1994, <http://www.ccrnp.ncifcrf.gov/~toms/paper/nano2/>.
- [SCHNEIDER97] T. D. Schneider, "Information content of individual genetic sequences," *J. Theor. Biol.*, vol. 189, no. 4, pp. 427–441, 1997, <http://www.ccrnp.ncifcrf.gov/~toms/paper/ri/>.
- [SCHNEIDER98] T. D. Schneider, "Sequence walkers: a graphical method to display how binding proteins interact with DNA or RNA sequences," *Nucleic Acids Res.*, vol. 25, pp. 4408–4415, 1997, <http://www.ccrnp.ncifcrf.gov/~toms/paper/walker/>, erratum: *NAR* 26(4): 1135, 1998.
- [SCHNEIDER01] T. D. Schneider, "Strong minor groove base conservation in sequence logos implies DNA distortion or base flipping during replication and transcription initiation," *Nucleic Acids Res.*, vol. 29, no. 23, pp. 4881–4891, 2001, <http://www.ccrnp.ncifcrf.gov/~toms/paper/baseflip/>.
- [SCHNEIDER02] T. D. Schneider, "Consensus Sequence Zen," *Applied Bioinformatics*, vol. 1, no. 3, pp. 111–119, 2002, <http://www.ccrnp.ncifcrf.gov/~toms/papers/zen/>.
- [SCHNEIDER06] T. D. Schneider, "Claude Shannon: Biologist," *IEEE Engineering in Medicine and Biology Magazine*, vol. 25, no. 1, pp. 30–33, 2006, <http://www.ccrnp.ncifcrf.gov/~toms/papers/shannonbiologist/>.
- [SCHNEIDER96] T. D. Schneider and D. Mastrorarde, "Fast multiple alignment of ungapped DNA sequences using information theory and a relaxation method," *Discrete Applied Mathematics*, vol. 71, pp. 259–268, 1996, <http://www.ccrnp.ncifcrf.gov/~toms/paper/malign>.
- [SCHNEIDER90] T. D. Schneider and R. M. Stephens, "Sequence logos: A new way to display consensus sequences," *Nucleic Acids Res.*, vol. 18, pp. 6097–6100, 1990, <http://www.ccrnp.ncifcrf.gov/~toms/paper/logopaper/>.
- [SCHNEIDER05] T. D. Schneider, I. Lyakhov, and D. Needle, "PROBE FOR NUCLEIC ACID SEQUENCING AND METHODS OF USE," 2005, US Patent application, patent pending, WO/2007/070572, <http://www.ccrnp.ncifcrf.gov/~toms/patent/medusa/>.
- [SHANNON48] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Tech. J.*, vol. 27, pp. 379–423, 623–656, 1948, <http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html>.
- [SHANNON49] C. E. Shannon, "Communication in the Presence of Noise," *Proc. IRE*, vol. 37, pp. 10–21, 1949.
- [SHANNON56] C. E. Shannon, "The bandwagon," *IRE Transactions-Information Theory*, vol. 2, no. 1, pp. 2–3, 1956.

- [SHULTZABERGER01] R. K. Shultzaberger, R. E. Bucheimer, K. E. Rudd, and T. D. Schneider, "Anatomy of *Escherichia coli* Ribosome Binding Sites," *J. Mol. Biol.*, vol. 313, pp. 215–228, 2001, <http://www.ccrnp.ncifcrf.gov/~toms/paper/flexrbs/>.
- [SHULTZABERGER07a] R. K. Shultzaberger, Z. Chen, K. A. Lewis, and T. D. Schneider, "Anatomy of *Escherichia coli* σ 70 promoters," *Nucleic Acids Res.*, vol. 35, pp. 771–788, 2007, <http://www.ccrnp.ncifcrf.gov/~toms/paper/flexprom/>.
- [SHULTZABERGER07b] R. K. Shultzaberger, D. Y. Chiang, A. M. Moses, and M. B. Eisen, "Determining physical constraints in transcriptional initiation complexes using DNA sequence analysis," *PLoS ONE*, vol. 2, p. e1199, 2007.
- [SHULTZABERGER07c] R. K. Shultzaberger, L. R. Roberts, I. G. Lyakhov, I. A. Sidorov, A. G. Stephen, R. J. Fisher, and T. D. Schneider, "Correlation between binding rate constants and individual information of *E. coli* Fis binding sites," *Nucleic Acids Res.*, vol. 35, pp. 5275–5283, 2007, <http://www.ccrnp.ncifcrf.gov/~toms/paper/fisbc/> <http://dx.doi.org/10.1093/nar/gkm471>.
- [SHULTZABERGER99] R. K. Shultzaberger and T. D. Schneider, "Using sequence logos and information analysis of Lrp DNA binding sites to investigate discrepancies between natural selection and SELEX," *Nucleic Acids Res.*, vol. 27, no. 3, pp. 882–887, 1999, <http://www.ccrnp.ncifcrf.gov/~toms/paper/lrp/>.
- [SIMPSON01] Simpson, M. L., Sayler, G. S., Fleming, J. T. & Applegate, B. Whole-cell biocomputing. *Trends in Biotechnology* 19, 317-323 (2001).
- [SPRINZAK05] Sprinzak, D. & Elowitz, M. B. Reconstruction of genetic circuits. *Nature* 438, 443-448 (2005).
- [SPORNS04] O. Sporns, D. R. Chialvo, M. Kaiser, and C. C. Hilgtag. Organization, development, and function of complex brain networks. *Trends in Cognitive Science*. 8, 418–425, 2004.
- [STEPHEN92] R. M. Stephens and T. D. Schneider, "Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites," *J. Mol. Biol.*, vol. 228, pp. 1124–1136, 1992, <http://www.ccrnp.ncifcrf.gov/~toms/paper/splice/>.
- [STEPNEY07] S. Stepney and S. Emmott, editors. Special Issue: Grand Challenges in Non-Classical Computation, *International Journal of Unconventional Computing*, volume 3, 2007.
- [STERN07] Stern, S., Dror, T., Stolovicki, E., Brenner, N. & Braun, E. Genome-wide transcriptional plasticity underlies cellular adaptation to novel challenge. *Mol Syst Biol* 3 (2007).
- [SZATHMARY03] E. Szathmary, "Why are there four letters in the genetic alphabet?" *Nat Rev Genet*, vol. 4, pp. 995–1001, 2003.
- [TASWELL08] C. Taswell, "DOORS to the Semantic Web and Grid With a PORTAL for Biomedical Computing," *IEEE Trans. Infor. Technol. Biomed.*, vol. 12, pp. 191–204, 2008.
- [TABOR07] J. J. Tabor. Programming living cells to function as massively parallel computers. In *Proceedings of the 44th Design Automation Conference (DAC'07)*, pages 638–639, 2007.
- [TOUMEY05] Toumey, C. P. Apostolic Succession: Does nanotechnology descend from Richard Feynman's 1959 talk? *Engineering and Science*, 16-23 (2005).
- [TEUSCHER08] C. Teuscher, N. Gulbahce and T. Rohlfs. Assessing Random Dynamical Network Architectures for Nanoelectronics. *Proceedings of the IEEE/ACM Symposium on Nanoscale Architectures, NANOARCH 2008*, Anaheim, CA, USA, Jun 12-13, 2008. <http://arxiv.org/abs/0805.2684>
- [TEUSCHER07] C. Teuscher. Nature-Inspired Interconnects for Self-Assembled Large-Scale Network-on-Chip Designs. *Chaos*, 17(2):026106, 2007.
- [TEUSCHER03] C. Teuscher, D. Mange, A. Stauffer, and G. Tempesti. Bio-inspired computing tissues: Towards machines that evolve, grow, and learn. *BioSystems*, 68(2–3):235–244, February–March 2003.

- [TOMASSINI05] M. Tomassini, M. Giacobini, and C. Darabos. Evolution and dynamics of small-world cellular automata. *Complex Systems*, 15(4):261–284, 2005.
- [TYSON03] Tyson, J.J., Chen, K.C. and Novak, B. (2003). Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. *Curr. Opin. Cell Biol.* 15:221-231.
- [TEUSCHER06] C. Teuscher. Biologically un-inspired computer science. *Communications of the ACM*, 49(11):27–29, November 2006.
- [VOLFSON06] Volfson, D. et al. Origins of extrinsic variability in eukaryotic gene expression. *Nature* 439, 861-864 (2006).
- [VYHLIDAL04] C. A. Vyhldal, P. K. Rogan, and J. S. Leeder, “Development and refinement of pregnane X receptor (PXR) DNA binding site model using information theory: insights into PXR-mediated gene regulation,” *J. Biol. Chem.*, vol. 279, pp. 46 779–46 786, 2004.
- [WATTS98] D. J. Watts and S. H. Strogatz. Collective dynamics of ‘small-world’ networks. *Nature* 393, 440–442, 1998.
- [WEAVER97] V. M. Weaver, O. W. Petersen, F. Wang, C. A. Larabell, P. Briand, C. Damsky, and M. J. Bissell, “Reversion of the malignant phenotype of human breast cells in three-dimensional culture and *in vivo* by integrin blocking antibodies,” *J Cell Biol*, vol. 137, pp. 231–245, 1997.
- [WEINBERGER08] Weinberger, L. S., Dar, R. D. & Simpson, M. L. Transient-mediated fate determination in a transcriptional circuit of HIV. *Nature Genetics* 40, 466-470 (2008).
- [WONG08] K. M. Wong, M. A. Suchard, and J. P. Huelsenbeck, “Alignment uncertainty and genomic analysis,” *Science*, vol. 319, pp. 473–476, 2008.
- [YOCKEY74] H. P. Yockey, “An application of information theory to the Central Dogma and the Sequence Hypothesis,” *J Theor Biol*, vol. 46, pp. 369–406, 1974.
- [YOCKEY58a] H. P. Yockey, “Some introductory ideas concerning the application of information theory in biology,” in *Symposium on Information Theory in Biology*, H. P. Yockey, R. L. Platzman, and H. Quastler, Eds. New York: Pergamon Press, 1958, pp. 50–59.
- [YOCKEY58b] “A primer on information theory,” in *Symposium on Information Theory in Biology*, H. P. Yockey, R. L. Platzman, and H. Quastler, Eds. New York: Pergamon Press, 1958, pp. 3–49.
- [YOCKEY58c] “A study of aging, thermal killing, and radiation damage by information theory,” in *Symposium on Information Theory in Biology*, H. P. Yockey, R. L. Platzman, and H. Quastler, Eds. New York: Pergamon Press, 1958, pp. 297–316.
- [ZHANG07] B. Zhang and G. S. Sukhatme, "Adaptive sampling for estimating a scalar field using a robotic boat and a sensor network," *Proc. IEEE Int'l Conf. on Robotics and Automation*, pp. 3673-3680, 2007.
- [ZHENG05] G. Zheng, F. Patolsky, Y. Cui, W. U. Wang and C. M. Lieber, "Multiplexed electrical detection of cancer markers with nanowire sensor arrays", *Nature Biotechnology*, Vol. 23, No. 10, pp. 1294-1301, October 2005.

Appendix A: Workshop Organization

Day 1 started with the five presentations covering the state-of-the-art of the five focused areas of the workshop; (1) *Design and Engineering of System Components* by Dr. Aristides Requicha (University of Southern California), (2) *System Design Methodology* by Dr. Michael L. Simpson (Oak Ridge National Laboratory), (3) *Coding Theory and Channel Capacity* by Dr. Thomas Schneider (National Institutes of Health), (4) *Novel Computing Paradigms* by Dr. Christof Teuscher (Los Alamos National Laboratory), and (5) *Biological Communications and its Potential Applications* by Dr. Lingchong You (Duke University), all concerned with the central theme of biology and computing/communications technology. The workshop was then broken into five breakout sessions; the 36 invited professionals were, according to their interest and expertise, divided into five breakout groups to discuss key research challenges.

Day 2 in the morning continued the breakout sessions. Participants in each breakout session discussed grand research challenges while formulating into presentation slides that illustrate identified grand research challenges. In the afternoon of Day 2, all participants met and each breakout group presented grand research challenges identified over the 2 day breakout sessions. Day 2 ended with a group discussion where participants provided comments and feedbacks on the workshop and together discussed to identify necessary action items for NSF to advance this technology field.

After the workshop, the workshop organizer, breakout session chairs and session members worked together to document in detail the grand research challenges identified from the workshop. A final report was then submitted to NSF for consideration for future funding programs.

Appendix B: Workshop Agenda

NSF Workshop on Molecular Communication/ Biological Communications Technology

Dates: Wednesday & Thursday Feb. 20 & 21st 2008.

Location: Hilton Arlington, Arlington, VA

DAY 1 (Wednesday Feb. 20th)

- 7:20 – 8:20 Registration and continental breakfast, *Masters Ballroom Pre-function Area*
- 8:20 – 9:00 Welcome and introductions, *Gallery 3*
- 8:20 – 8:25 Welcome remarks
Dr. Tadashi Nakano (University of California, Irvine)
- 8:25 – 8:35 Opening Talk (1): Dr. Jeannette Wing (NSF CISE/OAD)
- 8:35 – 8:45 Opening Talk (2): Dr. Michael Foster (NSF CCF)
- 8:45 – 9:00 Workshop organization: Dr. Tadashi Nakano
- 9:00 – 12:00 State-of-the-art specialist presentations, *Gallery 3*
- 9:00 – 9:30 “System components”
Dr. Aristides Requicha (University of Southern California)
- 9:30 – 10:00 “What I cannot create I do not understand:
The parallel synergetic pathways of design and discovery”
Dr. Michael L. Simpson (Oak Ridge National Laboratory)
- 10:00 – 10:30 Coffee break, *Masters Ballroom Pre-function Area*
- 10:30 – 11:10 “Shannon’s channel capacity theorem: is it about Biology?”
Dr. Thomas Schneider (National Institutes of Health)
Dr. Christopher Rose (Rutgers University)
- 11:10 – 11:40 “Novel computing paradigms: challenges and opportunities”
Dr. Christof Teuscher (Los Alamos National Laboratory)
- 11:40 – 12:10 “Biological communication: molecules, networks, and populations”
Dr. Lingchong You (Duke University)
- 12:10 – 12:15 “Molecular communication: research activities”
Dr. Tadashi Nakano (University of California, Irvine)
- 12:15 – 13:15 Lunch buffet, *restaurant within the Hilton Arlington*
- 13:30 – 17:00 Breakout sessions to discuss grand research challenges (by topics)
- “System components” (Chair: Dr. Aristides Requicha), *Picasso*
- “System design methodology” (Chair: Dr. Michael L. Simpson), *Da Vinci*
- “Coding theory and channel capacity in biology”
(Chair: Dr. Thomas Schneider), *Matisse*
- “Novel computing machines and paradigms”
(Chair: Dr. Christof Teuscher), *Rembrandt*
- “Biological communication: molecules, networks, and populations”
(Chair: Dr. Lingchong You), *Renoir*

- (15:00 – 15:30 Coffee break, *Masters Ballroom Pre-function Area*)
- 17:00 – 18:00 Summary of Day 1 discussions, *Gallery 3*
Brief report from each session (10 minute presentation from each session)
- 18:30 – 20:00 Dinner, *Gallery 2*
- 20:00 – 22:00 Breakout sessions (Optional), same rooms are available ~ 22:00.

DAY 2 (Thursday Feb. 21st)

- 7:30 – 8:30 Continental breakfast, *Masters Ballroom Pre-function Area*
- 8:30 – 11:30 Discussions on grand research challenges by topics (cont'd from Day 1)
(9:45 – 10:15 Coffee break, *Masters Ballroom Pre-function Area*)
- 11:30 – 12:30 Lunch buffet, *restaurant within the Hilton Arlington*
(Including session chair meeting on grand research challenges)
- 12:30 – 17:00 Presentations and discussions on grand research challenges (by topics), *Masters Ballroom*
“System components”
“System design methodology”
“Coding theory and channel capacity in biology”
“Novel computing machines and paradigms”
“Biological communication: molecules, networks, and populations”
(14:30 – 15:00 Coffee break, *Masters Ballroom Pre-function Area*)
- 17:00 – 18:00 Session chair meeting (session chairs only to discuss workshop report), *Renoir*

Appendix C: Participant List

Workshop organizers (2)

- Tadashi Nakano (University of California, Irvine)
- Tatsuya Suda (University of California, Irvine/NSF)

Invited professionals (36)

- Session 1: Ram H. Datar (Oak Ridge National Laboratory), Jun Li (Kansas State University), Jun Ni (University of Iowa), Kazuhiro Oiwa (National Institute of Information and Communications Technology), Aristides A. G. Requicha (University of Southern California), Louis F. Rossi (University of Delaware)
- Session 2: Sasitharan Balasubramaniam (Waterford Institute of Technology), Eric Batchelor (Harvard Medical School), John Doyle (California Institute of Technology), Ido Golding (University of Illinois at Urbana-Champaign), Jeff Hasty (University of California San Diego), Michael L. Simpson (Oak Ridge National Laboratory), Eric V. Stabb (University of Georgia), Leor S. Weinberger (University of California San Diego)
- Session 3: Andreas G. Andreou (Johns Hopkins University), Donal A. Mac Donaill (Trinity College), Andrew W. Eckford (York University), John S. Garavelli (European Molecular Biology Laboratory Outstation), Elebeoba (Chi-Chi) May (Sandia National Laboratories/University of New Mexico), Christopher Rose (Rutgers University, WINLAB), Thomas D. Schneider (National Institutes of Health), Hubert P. Yockey, Cynthia Yockey
- Session 4: Yaakov (Kobi) Benenson (Harvard University), Peter Dittrich (Friedrich Schiller University Jena), Jerzy Górecki (Institute of Physical Chemistry of Polish Academy of Sciences), Bruce J. MacLennan (University of Tennessee), Chien-Chung Shen (University of Delaware), Christof Teuscher (Los Alamos National Laboratory)
- Session 5: Anand R. Asthagiri (California Institute of Technology), William Bentley (University of Maryland, College Park), Cynthia Collins (University of Calgary), Yuki Moritani (NTT DoCoMo, Inc.), Christopher Rao (University of Illinois at Urbana-Champaign), John J. Tyson (Virginia Polytechnic Inst. & State Univ.), Lingchong You (Duke University)

NSF program directors (3)

- Jeannette Wing (CISE/OAD), Michael J. Foster (CISE/CCF), Deborah Crawford (CISE/OAD), Almadena Y. Chtchelkanova (CISE/CCF), Sirin Tekinay (CISE/CCF), Pinaki Mazumder (CISE/CCF), Mary Ann Horn (MPS/DMS), Sylvia Spengler (CISE/IIS), David Du (CISE/CNS), Shih-Chi Liu (ENG/CMMI), Darleen L. Fisher (CISE/CNS), Allison Mankin (CISE/CNS), Jie Wu (CISE/CNS)

Administrative assistants (3)

- Shun Watanabe (University of California, Irvine), Michael Moore (University of California, Irvine), Akihiro Enomoto (University of California, Irvine)