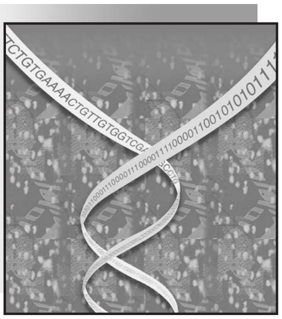

©EYEWIRE

# The Quest for Error Correction in Biology

## Recent Developments in Codes and Biology

BY MANISH K. GUPTA

The early excitement of application of coding and information theory to biology could not continue further due to several reasons like underdevelopment of both the fields. Now after 50 years, our understanding of biology is increasing day by day due to genomic research. The goal of this article is to give a brief account of the recent developments in codes and biology. In particular, we focus on the existence of error correction in biology.

Computing and communications are two broad areas where information theory has direct impact. The field of information theory begins with the work of Shannon in 1948, and since then, it has found applications and interactions in various fields including biology and chemistry [1]. The early excitement about its applications to biology could not continue further as the following observation was made in 1956 in a conference organized by Yockey:

> Information Theory is very strong on the negative side, i.e., in demonstrating what cannot be done; on the positive side its application to the study of living things has not produced many results so far; it has not yet led to the discovery of new facts, nor has its application to known facts been tested in critical experiments. To date, a definitive and valid judgment of the value of information theory in biology is not possible. [2]

This conclusion seems to be obvious because it was given only three years after the discovery of DNA structure and after nine years of the discovery of Shannon's information theory. Both the fields were quite young, and it was ahead of time when people were mixing the two. Yockey also wrote a book on this subject [3]. In another development on the counterpart field of information theory, that is, coding theory, there was work motivated by the discovery of the genetic code [4]. Several people including Golomb worked on comma-free codes [4].

At the threshold of a new century, our understanding of biology is now increasing day by day, and with the advancement of sequencing technology, we have the whole human genome stretched out before us. We now have with us the periodic table of biology. There is a need to look again with the new information in our hands for the applications of these communications technologies (coding and information theory) to biology. We are fortunate that we can safely divide the field into several subfields. Currently, the discipline can be divided roughly into four major directions:

1) applications of information theory to biology
2) existence of error correction in biological information processing
3) applications of coding theory to biomolecular computing (e.g., DNA computing)
4) applications of coding theory to computational molecular biology and bioinformatics.

Of course there is some overlap in these fields. In this article, we focus on second development and provide a brief account of the work. The first idea of possible connections between them (more precisely, the following question: Is there an error correction in the DNA sequence?) arose in the author's mind during his Ph.D. studies around 1998. After searching the literature, this author became aware of the work of Liebovitch et al. [5]. They could not find simple error-correcting codes in the base sequence of DNA. They were searching for simple error-correcting codes. Clearly, their approach was very basic, and it is unlikely to work. From their work, this author came to know the work of Forsdyke (1981) [6] and Rzeszowska-Wolny (1983) [7]. The first paper looked at error detection in intron sequences, and the second paper looked at the question of error correction in the genetic code. These papers are earlier attempts. There is a mixed response about the error correction in DNA sequences. On one side, there are biologists, like Forsdyke [8], who are optimistic about it and, on the other hand, biophysicists, like Patel [9], who are sure that there is no error correction. Finally in August 2002, a parity code (error-detecting code) was discovered in the nucleotide alphabet by Mac Dónaill [10]. This author also became aware that most of the coding theorists are unaware about this area partly because most of these papers are in biology or chemistry journals.  Schneider, together with his colleagues, has tried to develop a theory of molecular machines and molecular information theory (see [11]–[15] and the references therein) in recent years. In fact, more recently, Toby Berger of Cornell University gave a Shannon lecture at the IEEE International Symposium on Information Theory 2002

0739-5175/04/$20.00©2006IEEE

in Switzerland on “Living Information Theory.” He describes a mathematical model of neural coalition [16]. Finally, we should mention the following words of von Neumann:

> It is easy to note that the number of nerve actuation’s which occur in a normal lifetime must be of the order of $10^{20}$. Obviously, during this chain of events, there never occurs a malfunction which cannot be corrected by the organism itself, without any significant outside intervention. The system must, therefore, contain the necessary arrangements to diagnose errors as they occur, to readjust the organism so as to minimize the effects of errors. [17]

## Brief History and Preliminaries

To make the article self-contained, we include a brief summary of molecular biology, information theory, and coding theory. We start with some fundamental questions. Erwin Schrodinger [18] wrote a book, *What Is Life?,* in 1944. He was a physicist writing about life before the discovery of DNA structure. After 60 years, we have now another book written by a computer scientist Eric Baum [19] *What Is Thought?* He tried to explain thought process on the basis of computation. One can ask another question: What is the difference between life and matter? In other words, we can ask what the smallest unit of life is. Perhaps it is a cell or maybe a virus. Is it information? See the review of Casti [20]. Gitt tried to explain even the origin of life in his book *In the Beginning Was Information* [21]. It is worthwhile to mention that there is a spin-glass model of the origin of life [22], and, surprisingly, spin-glass has a naive connection with coding theory [23]. Ising spin-glasses are just a collection of $N$ particles with spin $\pm 1$. In the problem of magnetism, the Ising spin represents whether the microscopic magnetic moment is pointing up or down. The energy of a spin glass depends on the values of the spins and the strengths of the interactions among the particles. The energy of the whole system for a specific configuration is given by the Hamiltonian (see [24] for details). A lot of work has been done on spin glass and coding theory, and, remarkably, this gives Shannon capacity achieving codes. An analogy between the concepts in coding theory and spin glasses is shown below (see Table 1) [24].

For a connection between spin glass and biology see [25]. It appears that nature is already using optimal coding techniques, if any. One has to be very careful while trying to generalize Shannon’s idea of information theory to biology. Shannon’s information theory deals with point-to-point communication, and it is also like interorganism communication. To understand various biological communication systems and biological computing, we need to answer several questions. We have to answer first what information is [26]. Some related work between information theory and biology can be found in [27]–[35]. There are some comments about the role of information theory in biology by Shannon himself and Peter Elias (see [16]). In fact, von Neumann tried to develop the mathematical foundations of biology during last stages of his life [36]. The idea of von Neumann about genetic information is discussed by Chaitin at greater length in [37]. It is also worthwhile to mention the work of Bennett about the biosynthesis of messenger RNA as an example of reversible computation [38].

### Coding and Information Theory

In 1948, Shannon [39] set the ground work for today’s Information Technology. He gave *A Mathematical Theory of Communication,* now known as information theory. Information theory is about sending the information from here to there (transmission) and sending information from now to then (storage). Information theory sets bounds on what can be done or what cannot be done, but it never tells you how to do that. A constructive counterpart to Shannon’s theory is algebraic and combinatorial coding theory, which tells you how to do it. The father of coding theory is Richard W. Hamming, who was also at Bell Labs at the time of the birth of Shannon’s information theory. It was out of a frustration of errors made by that era’s computers that Hamming created his binary codes for correcting single errors in computers of that time. We should mention that the Hamming codes were known to Fisher [40], [41] in a different context. He discovered binary Simplex codes (dual of Hamming codes) in 1942 and later generalized them to prime powers in 1945. There is an old history for binary codes as well. It appears that the first five-letter binary code was discovered by Francis Bacon in 1605 in advancement of learning called *Omnia per Omnia*. This was the time when there were only 24 letters in English alphabet. Early in the 19th century in France, Joseph Marie Jacquard designed the first binary-coded punched cards for operating looms. George Boole gave the algebra of propositional calculus that forms the basis of the modern design of computer logic. A French engineer, Emile Baudot, discovered a binary cyclic-permuted code (now, often called Gray code because it was patented by Frank Gray on 17 March 1953 [42]). Gray code represented a major advancement in telegraphy and has various other applications. We will see later how Gray code is connected to molecular biology. Readers who are unfamiliar with the next 50 years of information theory and coding theory can see a commemorative issue of *IEEE Transactions on Information Theory* published in 1998 [1], [43]. It took almost 50 years to achieve Shannon capacity by iterative decoding procedures, low-density parity check codes, and turbo codes. Coding and information theory is now facing new challenges in wireless communications, multiple-input, multiple-output (MIMO) communication systems, and networking. Is it the right time to ask about biological coding theory, biological information theory, living coding theory, or living information theory?

### Molecular Biology

This section is based on a book written by Lander and Waterman [44]. Molecular biology grew out of two complementary experimental approaches to studying biological function: genetics and biochemistry. Genetics can be traced

**Table 1. Analogy between coding theory and spin glasses.**

| Coding Theory | Spin Glasses |
|---|---|
| Error-correcting code | Spin Hamiltonian |
| Signal to noise | $\frac{J0^2}{\Delta J^2}$ |
| Maximum likelihood decoding | Find a ground state |
| Error probability per bit | Ground state magnetization |
| Sequence of most probable symbols | Magnetization at temperature T = 1 |
| Convolutional codes | One-dimensional spin-glasses |
| Viterbi decoding | Transfer matrix algorithm |

back to Gregor Mendel, whose experiments on peas generated much interest in 1865 and showed the existence of genes in a mathematical way in living organisms. Fisher analyzed Mendel's data many years later and concluded that they fit statistical expectation very well [45]. Biochemistry deals with fractionating the molecules in a living organism with a goal of purifying and characterizing the chemical components responsible for carrying out a particular function. It was found that living organisms are composed of carbon, hydrogen, oxygen, and nitrogen. They also contain small amounts of other elements such as sodium, potassium, magnesium, sulfur, manganese, and selenium. These elements are combined in a vast array of complex macromolecules that can be classified into a number of major types: proteins, nucleic acids, lipids (fat), and carbohydrates (starch and sugar). Proteins are molecular miracles made of amino acids. Proteins have the most diverse range of functions (examples include enzymes, which catalyze chemical reactions such as the digestion of food: structure molecules, which make up hair, skin, and cell walls: transporters of substances, such as hemoglobin, which carries oxygen in blood: transporters of information, such as receptors in the surface of cells and insulin and other hormones). Most of the functions of the cell are done by proteins. There are 20 distinct amino acids, each with its own chemical properties. Each protein is defined by its unique sequence of amino acids. There are about 100,000 distinct proteins in the human body. The amino acid sequence of a protein causes it to fold into a particular three-dimensional shape having the lowest energy. This gives a protein its specific biochemical properties, i.e., its function. Predicting the structure of a protein is an extremely challenging problem in mathematical optimization. Readers who want to learn more about molecular biology and biochemistry are referred to the excellent books [44], [46], and [47]. Some other references about biological information processing are [48]–[50].

### Existence of Coding Theory in Molecular Biology

This is probably one of the most difficult areas to understand among the previous defined fields. Mojzsis et al. [51] have mentioned that coding and information theory has been in place in biology for at least 3.85 billions years. A very recent connection has been established by Yockey between the origin of life on earth and Shannon's theory of communication and evolution [52], [53]. One can look for the possible error-correction/detection schemes in living organisms at places where some sort of information processing is going on. Living organisms process the information at various places [54]–[56], although the processing of information and the corresponding encoding could be quite different [54], [16]. This information processing could be both classical and quantum. We need to find out at what place what type of information processing is going on. Our experience suggests that we should look for a Turing machine first. A good starting place is a series of papers written by Patel [57], [58], [54]. He speculates neatly about the biological information processing and the structure of information encoding [54]. He also gave an analogy between living organisms and computers (see Table 2). This could be a good starting point for us.

**Table 2. Biological, functional, computational analogy (from (58)).**

| Living organisms | Task | Computers |
|---|---|---|
| Signals from environment | Input | Data |
| Sense organs | High level | Preprocessor |
| Nervous system + brain | Translation | Operating system + compiler |
| Electrochemical signals | Low level | Machine code |
| Proteins | Execution | Electrical signals |
| DNA | Program | Programmer |

At some places, living systems process information without coding [59]. While searching for error correction in living organisms, we need to understand error correction very well. In [60], von Neumann proposes to view error "not as an extraneous and misdirected or misdirecting accident, but as an essential part of the process under consideration." We also need to understand how the information is encoded in living organisms (for example, as a concentration of chemicals etc).

In [59], Berger expresses his ideas about coding theory in biology as follows:

> Given that DNA is a long "block" code in a finite alphabet with some redundancy, there would seem to be the potential for coding there, and as I also point out such coding might be more suitable to a high-latency, interorganism communication such as takes place during mating and gestation. But in a brief discussion with the renowned Francis Crick of Crick & Watson fame about three years ago, he said he was unaware at that time of any solid evidence to that effect. Also in my Shannon lecture write up, I suggest in a long footnote that there may be some form of space-time coding taking place with the emphasis heavily on space. The reason why coding over time is of dubious value is that one must keep firmly in mind that latency is crucial in many perception problems; there is no a prior determined time at which a decision will be taken. Rather, the organism must have a (suboptimal) decision ready at all times based on what's been processed so far. This need to be "greedy" speaks against the employment of long block and convolutional codes in the case of sensory perception/pattern recognition, too. However, there are now iterative decoders of the turbo code type for certain simple convolutional code and other code families (MDPC families, e.g.) with relatively short constraint lengths which have the advantage that a turbo (i.e., iterative, effectively maximum-likelihood-seeking) decoder (if there can be said to be a decoder in the brain and anyone can locate where) would be capable of producing a suboptimum decision at a moment's notice. Moreover, this suboptimal decision would become increasingly close to optimum over time if it does turn out that the situation allows for sufficient time prior to decision making.

In the following sections, we will consider two of the important connections: one with genetic code and Gray code and the other with the Mac Dónaill code found in purine-pyrimidine and hydrogen donor-acceptor patterns in a nucleotide base. The section on codes and quantum biology considers the topological quantum error correction proposed by Porter [61] in microtubule and a suggestion of Patel about Grover's quantum search algorithm and genetic code [57].

### *The Central Dogma of Molecular Biology*

The central dogma of molecular biology was suggested by Crick [62], [63]. It suggests that information flows from DNA to mRNA and to protein (see Figure 1). It was observed by Yockey [53] that this is the property of genetic code. His arguments are based on Shannon's information theory and the fact that entropy of the DNA sequence is $\log_2(64) = 6$ and the entropy of the protein sequence is $\log_2(20)$. A model of central dogma as a communication system has been studied by Yockey (in particular, for a model of the channel see [3, p. 111]).

There have been some other attempts by Battail [64], Eigen [65], May [66], and Roman-Roldan et al. [67], viewing it as a communication channel. Some recent views are in [68]. Hopfield [69]–[71] in a series of papers writes about error correction. For more of his work, see the book by MacKay [72]. May, together with her colleagues, has written several papers recently on modeling the process of protein synthesis as a coding theory problem in prokaryotic organisms [73], [74]. The principle hypothesis in her model is: If mRNA is viewed as a noisy encoded signal to interpret the genetic translation, it is feasible to use principles of error-control coding theory in initiation mechanism. Ribosome is viewed as an error control decoder. She has tried even convolution code models [75] and iterative methods [76]. These models have been applied to the Escherichia coli K-12. Clearly these models are a good starting point, but a lot of work is needed [77]. Recently, Rosen and Moore have used finite field framework to find redundant coding structure in DNA [78], [79] and Battail examines the possible existence of soft codes, nested codes, and turbo codes [80], [81]. The work of Battail is related to the more general one of Barbieri regarding the existence of biological codes at different information levels [82].

### Gray Code and Genetic Code

Gray code is a binary linear code such that if you write all the code words as an array, then each codeword differs from the next codeword by a Hamming distance of one. For example, at length 2, the set $C = \{00, 01, 11, 10\}$ is a Gray code of length 2. This particular code was used as a mapping from the ring of integers modulo 4, i.e., $Z_4 \rightarrow (Z_2)^2$ and is a unique map that gives a new definition to binary nonlinear Kerdock and Preparata codes and solves the mystery about their duality [83]. Various other applications of Gray codes are known. There is a rich history of the genetic code [3], [84], [4]. In fact, an interesting and short story about its invention is given in [4]. Genetic code is at the heart of the famous central dogma of molecular biology. Genetic code contains instructions to make protein. DNA stores information in an alphabet of size 4 viz {A, C, G, T}. Each letter represents nucleotide bases adenine (A), cytosine (C), guanine (G) and thymine (T). When a cell decides to make a protein, it reads the portion of DNA that codes for that protein. The portion of the DNA that codes for a particular protein is called the *gene* for that protein. The reading of the gene is done by a special molecule called RNA polymerase (yet another protein), which then produces a transcript of the DNA sequence in the form of a strand of messenger RNA. RNA is ribonucleic acid, a substance that is like DNA in that it is made up of four nucleotides, which in turn are made of organic bases and ribose phosphate. But the ribose in RNA has a particular oxygen atom where the ribose in DNA does not. mRNA has an alphabet of size 4, {A, C, G, U}, similar to DNA but uracil (U) is in place of thymine (T).

A polymer is a molecule that is built up as a chain of smaller molecules, or monomers. Proteins are polymers with the monomers as amino acids. (See Figure 2 for a list of all 20 amino acids). Thus, proteins have an alphabet of size 20, {All 20 amino acids}. The genetic code is a mapping from mRNA to amino acids (see Table 3). Each amino acid is made of three nucleotides called a *triplet* or *codon*. The process of creating an mRNA from DNA is called *transcription*, and the process of making protein from mRNA is called *translation*.

By assigning attributes of 0 or 1, one can get a unique binary vector of length 6 from each amino acid as follows (see [85] and [86] and the references therein). According to chemical type and hydrogen binding, each base in DNA (mRNA) can be categorized. Thus we assign $A \rightarrow 00$, $G \rightarrow 01$, $C \rightarrow 11$, and $U \rightarrow 10$.

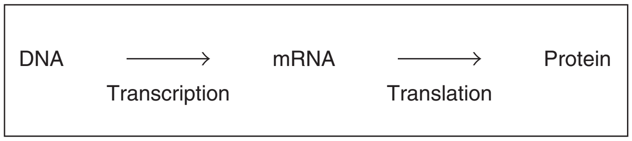

**Fig. 1.** Genetic information flow.

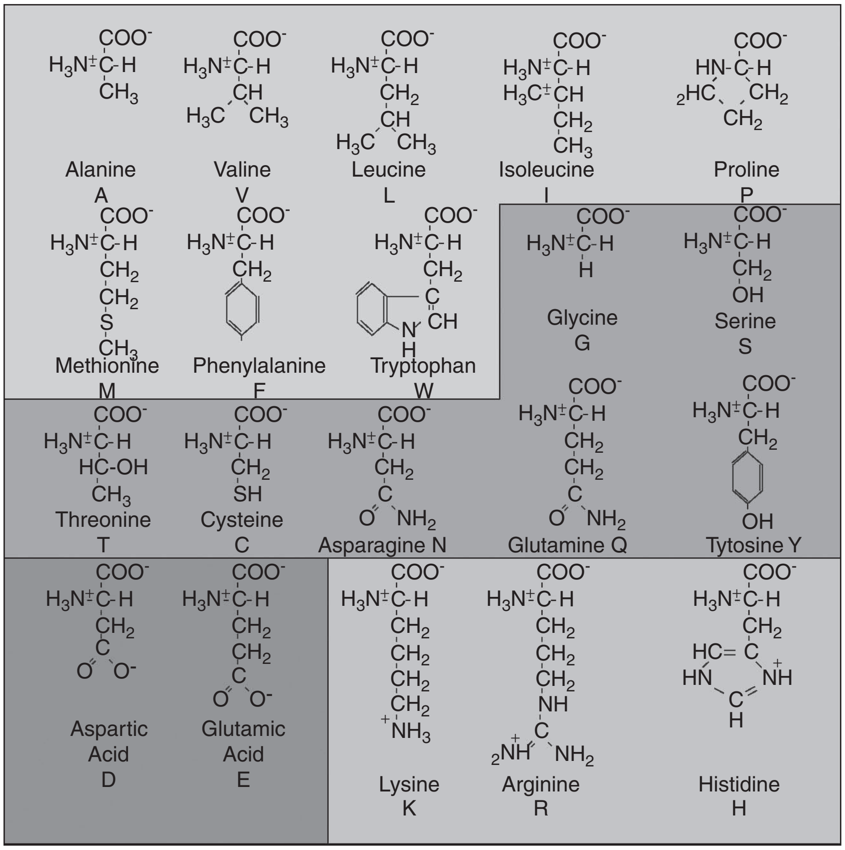

**Fig. 2.** A list of all 20 amino acids' molecular structures.

The first attribute is chemical character and the second attribute is the hydrogen-bond character. For example, A and G fall into same chemical type purines, whereas G and C fall into different chemical types (G falls into purines and C falls into pyrimidines). Similarly, A and U have weak hydrogen bonds, and G and C have strong hydrogen bonds. If we apply these maps to our genetic code codons, we will get a unique binary vector of length 6 for each amino acid, and eventually we get a Gray code representation of genetic code; see Figure 3. Note that all of the 64 codons are at the vertices in the figure, while some codons represent the same amino acid. One can associate a number of different Gray codes, depending upon the order of importance of the bits in a codeword [85]. Obviously, the next job is to check how good this representation of genetic code is from a coding theory point of view. This still has to be investigated in the sense of a Gray map. A lot has been said about the symmetry of the genetic code (for example, [3], [87]). Recently, Gonzalez [88] gave another mathematical description of the genetic code, describing the theoretical possibility of parity coding along the sequences of DNA.

### Mac Dónaill Code

Mac Dónaill of Dublin College, Ireland, discovered in August 2002 a binary (4, 4, 2) even parity code in the nucleotide alphabet by assigning a binary vector of length 4 to each base. This assignment was based on hydrogen donor-acceptor patterns found in nucleotides [10], [89]. This work was motivated by Yockey's arbitrary assignment of a binary vector of length 5 to each base and the work of Szathmary [90]. Here we describe it in detail.

We know that nucleotide bases can be classified according to their rings: purines, with two rings; R:(A, G), and pyrimidines, with one ring; Y:(C, T/U). In a DNA molecule, we only have C and T in pyrimidines. These bases form a complementary pair depending upon the hydrogen bonds between them, for example, G and C form a complementary pair. One can represent the donor-acceptor pattern of each nucleotide as a string of three bits. For example, if a donor is (arbitrarily) represented as 1 and an acceptor as 0, the pattern 100 would encode C and 011 would encode G. Further, if a purine is represented by 0 and a pyrimidine by 1, the full codeword for C would be 100, 1 and for G would be 011, 0. Thus, we can assign the following:

C 1001
G 0110
A 1010
T 0101.

In other words, nucleotides may be depicted as positions on a hypercube, represented by a cube within a cube. The position of a nucleotide is determined by its donor/acceptor pattern, while the purine/pyrimidine nature determines whether it belongs on the inner cube (pyrimidines) or outer cube (purines).

Table 3. Genetic code (from (44)).

| 5′ END | U | C | A | G | 3′END |
|---|---|---|---|---|---|
| U | Phe | Ser | Tyr | Cys | U |
|  | Phe | Ser | Tyr | Cys | C |
|  | Leu | Ser | Stop | Stop | A |
|  | Leu | Ser | Stop | Trp | G |
| C | Leu | Pro | His | Arg | U |
|  | Leu | Pro | His | Arg | C |
|  | Leu | Pro | Gln | Arg | A |
|  | Leu | Pro | Gln | Arg | G |
| A | Ile | Thr | Asn | Ser | U |
|  | Ile | Thr | Asn | Ser | C |
|  | Ile | Thr | Lys | Arg | A |
|  | Met | Thr | Lys | Arg | G |
| G | Val | Ala | Asp | Gly | U |
|  | Val | Ala | Asp | Gly | C |
|  | Val | Ala | Glu | Gly | A |
|  | Val | Ala | Glu | Gly | G |

## Codes and Quantum Biology

### *Microtubules and Codes*

Microtubules (MTs) [91], [92] are cylindrical polymers of the protein tublin and are 25 nanometers in diameter (see Figure 4). In fact, much of the cell cytoskeleton is made of microtubule (MT). They self assemble to determine cell shape and function. It is like a strong tunnel that links distant parts of the cell to each other. Tubulin by virtue of its tertiary structure likes to polymerize and form strong tunnels. Each tubulin is a peanut-shaped 8-nm dimer consisting of $\alpha$ and $\beta$ monomers. The tubulin dimer within MTs is arranged in a skew hexagonal lattice, which is slightly twisted. Each tubulin has an electric dipole moment due to an asymmetric charge distribution. Thus, MT is a lattice of oriented dipoles that can be in different phases including spin-glass phase. It has been proposed that MTs can store both classical and quantum information and actually do the computation. Tubulin work as a cellular automaton in the walls of MTs. Tubulin can exist in two conformations

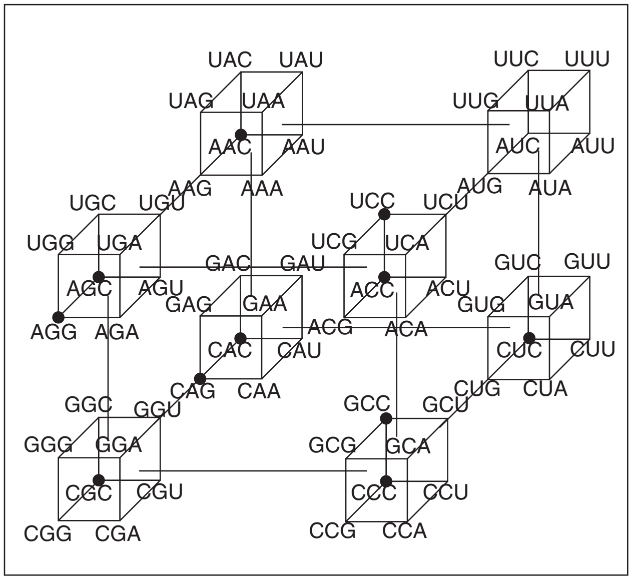

**Fig. 3.** A Gray code representation of the genetic code.

determined by quantum London forces in a hyperbolic pocket or superposition of both conformations (see Figure 5). The motion of an object between two conformational states of tubulin is equivalent to two curvatures in space-time geometry represented as a two-dimensional (2-D) space-time sheet.

Thus, MT appears to be most promising candidate for information processing. Several authors have looked at the possibility of classical information processing and quantum information processing in MTs [93], [94]. Therefore, a natural question is what kind of error-correction mechanism do they have? To answer this question, [93] is a good starting point for classical information processing. Penrose, Hameroff, Hagan, and Tuszynski make proposals about quantum computation in brain MTs (see [95] and references therein). In fact, Porter suggested a topological quantum error correction in MTs very recently [61]. Porter assumes special 2-D particles (he calls them *anyons*) that are involved in topological quantum computing by moving around each other on the wall of MT. His argument about anyon goes as follows. All known particles are "bosons" or "fermions." Bosons gather together while fermions stay apart. If you swap two bosons, their quantum state will not change; however, if we swap two fermions, their quantum state will be multiplied by $-1$. The spin-statistics theorem says that these are the only possibilities in three dimension. In two dimension, the phase factor modifying the quantum state can be any complex number of size 1; thus the name "any-on." None of the elementary particles are anyons, but there can be anyonic "quasiparticles" made of a group of electrons [61]. In Freedman's model of topological quantum computing, the anyons will be localized patterns of qubit flips created in pairs. Since the MT can be seen as an array of qubits, the creation of topological states requires the prior existence of a domain of coherently coupled dimer-qubits that rings the MT. Once such a "quantum ring" exists, anyonic motions can create a robust multiqubit entanglement. This speculation still must be investigated in more detail [61]. Penrose has suggested that Fibonacci patterns on microtubules may be optimal for error correction [96]. As these ideas are quite nascent, we need rigorous analysis and experimental verifications.

### Grover's Search Algorithm and Genetic Code

Patel studies DNA replication and protein synthesis from a computer science point of view [57]. According to his proposal, there is a quantum computer working behind four nucleotide bases and 20 amino acids. These numbers arise as a solution to an optimization problem. Grover gave an algorithm for searching an unordered database of $N$ objects on a quantum computer [97]. More precisely, this optimal quantum search algorithm relates the number of objects $N$ that can be distinguished by a number of yes/no queries $Q$ according to

$$(2Q+1)\sin^{-1}\left(\frac{1}{\sqrt{N}}\right)=\frac{\pi}{2}.$$

The solutions of this for small values of $Q$ have special significance for the number of building blocks involved in genetic information processing according to Patel

$$Q=1, N=4;\ Q=2, N=10.5;\ Q=3, N=20.2.$$

Identification of a binary quantum query with nucleotide base-pairing gives a natural explanation of why living organisms have four nucleotide bases and 20 amino acids. The second case shows that today's genetic code evolved from a simpler one with ten amino acids. So, if there is a quantum computer, a natural question is then what kind of error correction?

Recently, a quantum mechanical model of adaptive mutation has been experimented by McFadden and Khalili [98].

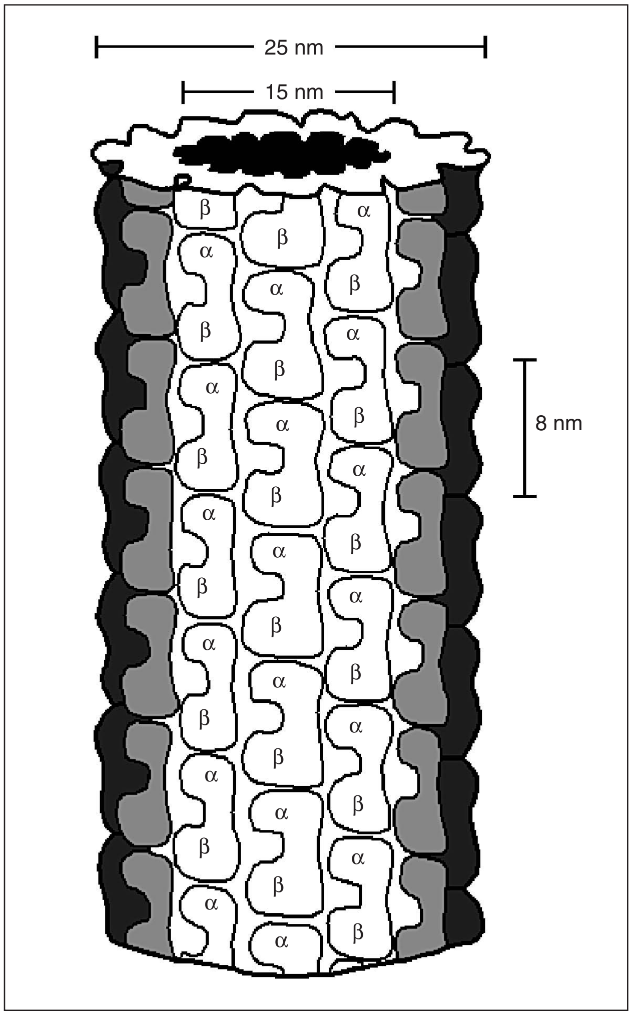

**Fig. 4.** A microtubule (94).

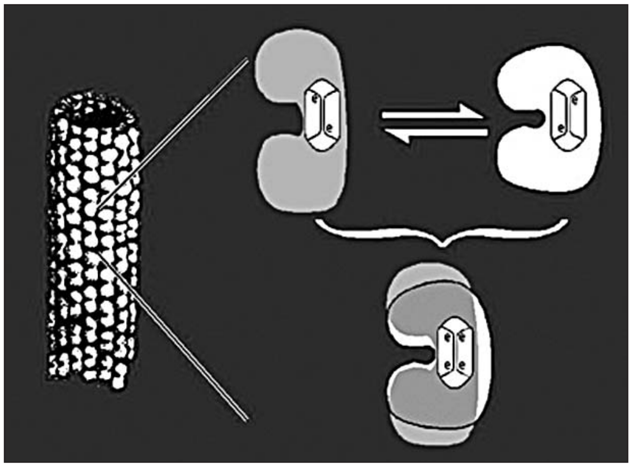

**Fig. 5.** Tubulin can exist in superposition of both conformations (96).

Many such hypotheses are emerging in quantum biology. Still, one has to investigate the error-correction mechanism in such systems.

## Conclusions

Information processing in biology is a fast-emerging field, and coding and information theory for biological systems is still in its infancy. We hope in the future that this new field will become very important to understand how biology is doing information technology at the molecular level. The quest for information processing principles in life sciences has begun seriously (for example, a main aim of the Howard Hughes Medical Institute's recently opened Janelia research farm is "the identification of the general principles that govern how information is processed by neuronal circuits" [100]. We have more questions than answers at this moment in this area. We need to search for classical and quantum codes and, in some cases, different types of coding. Perhaps a cellular automata model will be helpful in some cases. We need coding and information theorists to look at this. In particular, we need solid mathematical foundations. If we can solve the problem of protein folding with such techniques, it will change the future of medicine. We need the analogue of Shannon's theorems, and we need to classify the biological communication channels. Certainly, we need to extend the domain of coding and information theory in order to understand the biological communication systems [99]. Today, a revolution in biology is being led not only by biologists but by computer scientists, engineers, and mathematicians (one such example is Eric S. Lander, who was the key person involved in the human and mouse genome projects. He was trained as a coding theorist and not as a biologist).

## Acknowledgments

The author would like to thank number of people who had sent their pre-/reprints, and who had discussions with the author. In particular, the author wants to thank Cris Calude, D.A. Forskyde, Hesham El-gamal, T. Aaron Gulliver, Lila Kari, D. A. Mac Dónaill, D.J.C. MacKay, A. Patel, Patrick Sole, Jack Tuszynski, and H.P. Yockey. The author is thankful to Stuart Hameroff for allowing his figures to be included in the paper and to the anonymous referees for helpful suggestions. Instead of redrawing, some figures are taken from the public domain. The author is thankful to Navin Kashyap and to the Mathematics and Statistics Department at Queens University, Canada, for financial support.

**Manish K. Gupta** is currently serving as an adjunct assistant professor and postdoctoral fellow in the Mathematics Department, Queens University, Canada. Gupta received his B.S. degree in physics, chemistry, and mathematics and an M.S. degree in mathematics from the University of Lucknow, India. He earned his Ph.D. in mathematics in 2000 from the Indian Institute of Technology, Kanpur, India. He worked as a Marsden research fellow at the University of Canterbury, Christchurch, New Zealand, in a quantum error-correction project of the Royal Society of New Zealand from 2000–2002. From August 2002–May 2004, he was a faculty associate and postdoctoral fellow at the Department of Computer Science and Engineering at Arizona State University (ASU). and from May 2004–2005 he workedat the Information Processing Systems Laboratory, Ohio State University (OSU) as a postdoctoral fellow/lecturer. He taught several courses at the Department of Computer Science and Engineering at ASU and the Department of Electrical and Computer Engineering at OSU.

His research interests include information processing in biology, coding, and information theory, DNA, and quantum computing. He is a Member of the IEEE, and of the IEEE Information Theory, Communication and Computer Societies, the International Society for Computational Biology, and the International Society for Nanoscale Science, Computation and Engineering.

He is a coauthor of a forthcoming research monograph in quantum error correction and has written several research articles. He has served as a reviewer to several international journals in mathematics and communications and has also given several invited talks in the United States, Singapore, New Zealand, Korea, and India.

**Address for Correspondence:** Manish K. Gupta, Room 216, Department of Mathematics and Statistics, Queens University, Kingston, Ontario K7L3N6 Canada. Phone: +1 613 533 2409. Fax: +1 613 533 2964. E-mail: m.k.gupta@ieee.org.

## References

[1] S. Verdu, "Fifty years of Shannon theory," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2057–2078, 1998.

[2] H.P. Yockey, *Symposium on Information Theory in Biology,*. New York: Pergamon Press, 1956.

[3] H.P. Yockey, *Information Theory and Molecular Biology.* Cambridge, UK: Cambridge Univ. Press, 2005.

[4] B. Hayes, "The invention of the genetic code," *Amer. Scientist*, vol. 86, no. 1, pp. 8–14, 1998.

[5] L.S. Liebovitch, Y. Tao, A.T. Todorov, and L. Levine, "Is there an error correcting code in the base sequence in DNA?" *Biophys. J.*, vol. 71, no. 3, pp. 1539–1544, 1996.

[6] D.R. Forsdyke, "Are introns in-series error-detecting sequences?," *J. Theoretical Biol.*, vol. 93, no. 4, pp. 861–866, 1981.

[7] J. Rzeszowska-Wolny, "Is genetic code error-correcting?," *J. Theor. Biol.*, vol. 104, pp. 701–702, 1983.

[8] M.K. Gupta, private communication with Donald Forsdyke, 2001.

[9] M.K. Gupta, private communication with Apoorva Patel, 2001.

[10] D.A. Mac Dónaill, "A parity code interpretation of nucleotide alphabet composition," *Chem. Commun.*, no. 18, pp. 2062–2063, 2002.

[11] T.D. Schneider, G.D. Stormo, L. Gold, and A. Ehrenfeucht, "Information content of binding sites on nucleotide sequences," *J. Mol. Biol.*, vol. 188, no. 3, pp. 415–431, 1986.

[12] T.D. Schneider and R.M. Stephens, "Sequence logos: A new way to display consensus sequences," *Nucleic Acids Res.*, vol. 18, no. 20, pp. 6097–6100, 1990.

[13] 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, no. 1, pp. 1–18, 1994.

[14] T.D. Schneider, "Theory of molecular machines. I. Channel capacity of molecular machines," *J. Theor. Biol.*, vol. 148, no. 1, pp. 83–123, 1991.

[15] T.D. Schneider, "Theory of molecular machines. II. Energy dissipation from molecular machines," *J. Theor. Biol.*, vol. 148, no. 1, pp. 125–137, 1991.

[16] T. Berger, "Living information theory," *IEEE Information Theory Soc. Newslett.*, vol. 53, no. 1, pp. 1, 6–19, 2003.

[17] J. von Neumann, *The Computer and the Brain*. New Haven, CT: Yale Univ. Press, 1958.

[18] E. Schrodinger, *What Is Life*? Cambridge, UK: Cambridge University Press, 1944.

[19] E.B. Baum, *What Is Thought*? Cambridge, MA: MIT Press, 2004.

[20] J.L. Casti, "Steve Grand's creation: Life and how to make it (review)," *Nature*, vol. 409, pp. 17–18, 2001.

[21] W. Gitt, *In the Beginning Was Information*. Bielefeld, Germany: CLV, 2001.

[22] G. Rowe, *Theoretical Models in Biology: The Origin of Life, the Immune System, and the Brain*. New York: Oxford Univ. Press, 1994.

[23] H. Nishimori, *Statistical Physics of Spin Glasses and Information Processing: An Introduction*. New York: Oxford Univ. Press, 2001.

[24] N. Sourlas, "Statistical mechanics and error-correcting codes," 1998 [Online]. Available: http://arXiv.org/abs/cond-mat/9811406

[25] D.L. Stein, *Spin Glasses and Biology.* Singapore: World Scientific, 1992.

[26] R.L. Constable, “Notes on what is information? Workshop,” Cornell Univ., Ithaca, NY, Tech. Rep. 14853–7501, 2001.

[27] B. Hassenstein, *Information and Control in the Living Organism*. London: Chapman & Hall, 1971.

[28] L.L. Gatlin, *Information Theory and the Living System*. New York: Columbia Univ. Press, 1972.

[29] R. Baddeley, P. Hancock, and P. Foldiak, *Information Theory and the Brain*. Cambridge, UK: Cambridge Univ. Press, 2000.

[30] C. Adami, “Information theory in molecular biology,” *Physics Life Rev 1,* pp. 3–22, 2004.

[31] S. Ji, “Molecular information theory: Solving the mysteries of DNA,” in *Modeling in Moldecular Biology* (Natural Computing Series), G. Ciobanu and G. Rozenberg, Eds. Berlin: Springer, 2004, pp. 141–150.

[32] J. Avery, *Information Theory and Evolution*. Singapore: World Scientific, 2003.

[33] O. Milenkovic and B. Vasic, “Information theory and coding problems in genetics,” in *Proc. IEEE Information Theory Workshop*, Oct. 2004.

[34] S. Hussini, L. Kari, and S. Konstantinidis, “Coding properties of DNA languages,” *Theoretical Comput.Sci.*, vol. 290, no. 3, pp. 1557–1579, 2003.

[35] Stambuk, “On circular coding properties of gene and protein sequences,” *Croatia Chemica ACTA,* vol. 4, no. 4, pp. 999–1008, 1999.

[36] G. Chaitin, “Information-theoretic computation complexity,” *IEEE Trans. Inform. Theory*, vol. 20, no. 1, pp. 10–15, 1974.

[37] G.J. Chaitin, “To a mathematical definition of life,” *ACM SICACT News*, vol. 4, p. 12, 1970.

[38] C.H. Bennett, “Logical reversibility of computation,” *IBM J. Res. Develop*, vol. 17, no. 6, p. 525, 1973.

[39] C.E. Shannon, “A mathematical theory of communication,” *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656.

[40] R.A. Fisher, “The theory of confounding in factorial experiments in relation to the theory of groups,” *Ann. Eugenics*, vol. 11, pp. 341–353, 1942.

[41] R.A. Fisher, “A system of confounding for factors with more than two alternatives, giving completely orthogonal cubes and higher powers,” *Ann. Eugenics*, vol. 12, pp. 2238–2290, 1945.

[42] F. Gray, “Pulse code communication,” U.S. Patent 2 632 058,” Mar. 17, 1953.

[43] A.R. Calderbank, “The art of signaling: Fifty years of coding theory,” *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2561–2595, 1998.

[44] E.S. Lander and M.S. Waterman, *Calculating the Secrets of Life*. Washington, DC: National Research Council, 1995.

[45] R.A. Fisher, *The Genetical Theory of Natural Selection*. Oxford, UK: Oxford University Press, 1930.

[46] D. Voet and J.G. Voet, *Biochemistry*. New York: Wiley, 1995.

[47] J.D. Watson, T.A. Baker, S.P. Bell, A. Gann, M. Levine, R. Losick, *Molecular Biology of the Gene*. San Francisco, CA: Benjamin Cummings, 2005.

[48] G. Bock and J. Goode, *Complexity in Biological Information Processing*. New York: Wiley, 2001.

[49] S. Fraga, K.M.S. Saxena, and Manuel Torres, *Bio-molecular Information Theory*. New York: Elsevier, 1978.

[50] H.C. Luttgau and R. Necker, *Biological Signal Processing*. Weinheim, Germany: Sonderforschungsbereiche, 1989.

[51] S.J. Mojzsis and A.G. Kishnamurthy, “Before RNA and after: Geological and geochemical constraints on molecular evolution,” in *The RNA World, The Nature of Modern RNA Suggests a Prebiotic RNA*. Boca Raton, FL: Cold Spring Harbor Laboratory Press, 1998, pp. 1–47.

[52] H.P. Yockey, “Origin of life on earth and Shannon’s theory of communication,” *Computers and Chemistry*, vol. 24, no. 1, pp. 105–123, 2000.

[53] H.P. Yockey, *Information Theory, Evolution and the Origin of Life*: *Fundamentals of Life*. New York: Elsevier, 2002, pp. 335–348.

[54] A. Patel, “Carbon—The first frontier of information processing,” *J. Biosci*., vol. 27, no. 3, pp. 207–218, 2002.

[55] J.A. Tuszynski, “Entropy versus information: Is a living cell a machine or a computer?,” in *Proc. Int. Conf.Comput. Information Technol.*, 2001.

[56] J.A. Tuszynski, “Biomolecular quantum computers,” in *Molecular Computing,* T. Sienko, Ed. Cambridge, MA: MIT Press, 2001.

[57] A. Patel, “Quantum algorithms and the genetic code,” *Pramana-J. Physics*, vol. 56, no. 2–3, pp. 367–381, Feb./Mar. 2000.

[58] A. Patel, “Mathematical physics and life,” *Mathematical Sciences Series: Selected Topics in Computing and Information Science*, J.C. Misra, Ed., vol. 4. New Delhi, India: Narosa, 2003, pp. 270–293.

[59] M.K. Gupta, private communication with Toby Berger, Sep. 2003.

[60] J. von Neumann, “Probabilistic logics and the synthesis of reliable organisms from unreliable components,” in *Automata Studies*. Princeton, NJ: Princeton Univ. Press, 1956.

[61] M. Porter, “Topological quantum error correction: Applications to microtubules,” in *Proc. Quantum Mind Conf.,* Tucson*,* Arizona, 2003.

[62] F.H.C. Crick, “The origin of the genetic code,” *J. Mol. Biol.*, vol. 38, no. 3, pp. 367–379, 1968.

[63] F. Crick, “Central dogma of molecular biology,” *Nature*, vol. 227, pp. 561–563, 1970.

[64] G. Battail, “Does information theory explain biological evolution?,” *Europhysics Lett.*, vol. 40, no. 3, pp. 343–348, 1997.

[65] M. Eigen, “The origin of genetic information: viruses as models,” *Gene*, vol. 135, no. 1–2, pp. 37–47, 1993.

[66] E.E. May, “Comparative analysis of information based models for initiating protein translation in escherichia coli K-12,” M.S. thesis, NC State Univ., Raleigh, NC, 1998.

[67] R. Roman-Roldan, P. Bernaola-Galvan, and J.L. Oliver, “Applications of information theory to DNA sequence analysis: A review,” *Pattern Recognition*, vol. 29, no. 7, pp. 1187–1194, 1996.

[68] E.E. May, M.A. Vouk, D.L. Bitzer, and D.I. Rosnick, “A coding theory framework for genetic sequence analysis,” in *Proc. Workshop Genomic Signal Processing Statistics*, 2002.

[69] J.J. Hopfield, “Kinetic proofreading: A new mechanism for reducing errors in biosynthetic process requiring high specificity,”*Proc. Nat. Acad. Sci. USA*, vol. 71, no. 10, pp. 4135–4139, 1974.

[70] J.J. Hopfield, “The energy relay: A proofreading scheme based on dynamic cooperativity and lacking all characteristic symptoms of kinetic proofreading in DNA replication and protein synthesis,” *Proc. Nat. Acad. Sci. USA*, vol. 77, no. 9, pp. 5248–5252, 1980.

[71] J.J. Hopfield, “Origin of the genetic code: A testable hypothesis based on tRNA structure, sequence, and kinetic proofreading,” *Proc. Nat. Acad. Sci. USA*, vol. 75, no. 9, pp. 4334–4338, 1978.

[72] D.J.C. MacKay, *Information Theory, Inference, and Learning Algorithms*. Cambridge, UK: Cambridge Univ. Press, 2003.

[73] E.E. May, *Analysis of Coding Theory Based Models for Initiating Protein Translation in Prokaryotic Organisms*, Ph.D. dissertation, NC State Univ., Raleigh, NC, 2002.

[74] E.E. May, M.A. Vouk, D.L. Bitzer, and D.I. Rosnick, “Coding theory based models for protein translation initiation in prokaryotic organisms,” *J. BioSyst.*, vol. 76, no. 1–3, pp. 249–260, 2004.

[75] E.E. May, M.A. Vouk, D.L. Bitzer, and D.I. Rosnick, “Constructing optimal convolutional code models for prokaryotic translation initiation,” in *Proc. 2nd Joint EMBS/BMES Conf.,* 2002, vol. 3, pp. 2188–2189.

[76] E.E. May, “Use of iterative methods in biological coding theory: Applications,” in *Proc. IMACS 03*, 2003.

[77] E.E. May, “Towards a biological coding theory discipline,” *New Thesis*, vol. 1, no. 1, pp. 19–38, 2004.

[78] G.L. Rosen, “Finding near-periodic DNA regions using a finite-field framework,” in *Proc. IEEE Workshop Genomic Signal Processing (GENSIPS)*, May 2004.

[79] G.L. Rosen and J.D. Moore, “Investigation of coding structure in DNA,” in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Processing (ICASSP)*, Apr. 2003.

[80] G. Battail, “An engineer’s view on genetic information and biological evolution,” *Biosystems*, vol. 76, no. 1–3, pp. 279–90, 2004.

[81] G. Battail, “Can we explain the faithful communication of genetic information?,” in *Proc. DIMACS Working Group Theoretic. Advances Information Recording*, Mar. 2004.

[82] M. Barbieri, *The Organic Codes*. Cambridge, UK: Cambridge Univ. Press, 2003.

[83] A.R. Hammons, P.V. Kumar, A.R. Calderbank, N.J.A. Sloane, P. Sole, “Thelinearity of Kerdock, Preparata, Goethals, and related codes,” *IEEE Trans. Inform. Theory*, vol. 40, no. 2, pp. 301–319, 1994.

[84] L.E. Kay, *Who Wrote the Book of Life? A History of the Genetic Code*. Stanford, CA: Stanford Univ. Press, 2000.

[85] M.A. Jimenez-Montano, C.R. Mora-Basanez, and T. Poschel, “On the hypercube structure of the genetic code,” in *Proc. 3rd Int. Conf. Bioinformatics Genome Res.*,1994, p. 445.

[86] M.A. Jimenez-Montano, C.R. Mora-Basanez, and T. Poschel, “The hypercube structure of the genetic code explains conservative and non-conservative amino acid substitutions in vivo and in vitro,” *BioSyst.*, vol. 39, no. 2, pp. 117–125, 1996.

[87] I. Stewart, “Broken symmetry in the genetic code?,” *New Scientist*, vol. 1915, p. 16, Mar. 1994.

[88] D.L. Gonzalez, “Can the genetic code be mathematically described?,” *Med. Sci. Monit.*, vol. 10, no. 4, pp. HY11–17, 2004.

[89] D.A. Mac Donaill, “Why nature chose A,C, G and U/T, an error-coding perspective of nucleotide alphabet composition,” *Origins of Life and Evolution of the Biosphere*, vol. 33, pp. 433–455, 2003.

[90] E. Szathmary, “What is the optimum size for the genetic alphabet?,” in *Proc. Natl. Acad. Sci. USA*, vol. 89, no. 7, p. 2614, 1992.

[91] P. Cappuccinelli and N. Ronald Morris, *Microtubules in Microorganisms*. New York: Marcel Dekker, 1982.

[92] P. Dustin, *Microtubules*. Berlin: Springer, 1984.

[93] J. Faber, L.P. Rosa, and R. Portugal, “Information processing in brain microtubules,” in *Quantum Mind Conf.*, 2003.

[94] S. Hameroff and J. Tuszynski, *Search for Quantum and Classical Modes of Information Processing in Microtubules: Implications for the Living State*. Singapore: World Scientific, 2003.

[95] S. Hagan, S.R. Hameroff, and J.A. Tuszynski, “Quantum computation in brain microtubules: Decoherence and biological feasibility,” *Physical Rev. E*, vol. 65, no. 6, pt. 1, pp. 061901, 2002.

[96] S. Hameroff home page [Online]. Available: http://www.quantumconsciousness.org

[97] L.K. Grover, “A fast quantum mechanical algorithm for data base search,” in *Proc. Annual ACM Symp.Theory Computing (STOC)*, 1996, p. 212–219.

[98] J. McFadden and J. Al-Khalili, “A quantum mechanical model of adaptive mutation,” *BioSyst.*, vol. 50, no. 3, pp. 203–211, 1999..

[99] A. Patel, “Information processing beyond quantum computation” [Online]. Available: http://arxiv.org/abs/quant-ph/0306158

[100] Janelia Farm Research Campus [Online]. Available: http://www.hhmi.org/janelia