Information **2011**, *2*, 117-139; doi:10.3390/info2010117

OPEN ACCESS

information

ISSN 2078-2489

www.mdpi.com/journal/information

Article

Biological Information—Definitions from a Biological Perspective

Jan Charles Biro

Homulus Foundation, 612 S Flower Str., Los Angeles, CA 90017, USA; E-Mail: jan.biro@att.net

Received: 13 January 2011 / Accepted: 18 January 2011 / Published: 21 January 2011

Abstract: The objective of this paper is to analyze the properties of information in general and to define biological information in particular.

Keywords: biological; information; data; signal; message; knowledge; Shannon; DNA; protein; communication

1. Introduction

A young man traveling around the country became short of money and sent the shortest possible telegram to his father explaining his situation and asking for help: "Father! Send some money! Your son". His father was old with poor eyesight and therefore asked his neighbor, a farmer, to read the message. The farmer did not concern himself with unnecessary speech. He read the message and shouted to the father: **"FATHER-SEND-SOME-MONEY-YOUR-SON!!!"** The father was upset about the short and demanding request, and decided to ignore it.

However, several days later he asked another neighbor, the village priest, to read the same message to confirm that he had understood his son correctly. The priest quietly read the message, lifted his eyes upward and started to recite the message very slowly, as though he was speaking directly to God: "Father!" he paused. "Send….. some…. Money." He finished the message whispering, like a shy little boy: "your son". The old man shed some tears, and became emotional on hearing the quiet, polite, respectful message… and he sent money to his son.

I have heard versions of this story several times from my own father and this taught me that a message not only has content but also packaging. Moreover, I learned that recipients can understand or misunderstand the same message.

How should we define biological information? The word "information" can be used in several ways and has a complex meaning. Therefore, it is necessary to provide the context of what we are speaking about before giving a succinct definition. According to the Oxford English Dictionary, the earliest

historical meaning of the word 'information' was the act of informing, or giving form or shape to the mind, as in education, instruction or training. A quotation from 1387: *"Five books came down from heaven for the information of mankind"*.

The English word was thought to be arrived at by adding the common 'noun of action' ending '-ation' to the earlier verb 'to inform', in the sense of 'to give form to the mind, to discipline, instruct and teach'. The word 'Inform' derives from the Latin verb *informare*, which means to give form to, to form an idea of. Furthermore, the Latin language already contained the word *informatio*, meaning concept or idea. The ancient Greek word for 'form' was εἶδος (eidos), famously used in a technical philosophical sense by Plato (and later Aristotle) to denote the ideal identity or essence of something.

Message is the materialized form of the information. It has an origin, the *sender*, and a destination, the *recipient*. However, not everything in a message is information, only the portion that has the same meaning for both the sender and recipient can be considered information. Simply stated, ***information is a message received and understood***.

The message can be transmitted between sender and recipient in different forms (letters, barcode, sound, atoms etc). It is not the carrier itself, but the order of carrier elements (the pattern), that is essential in a message. This can give rise to the idea that information is independent of its carrier, something immaterial or spiritual, or having at least one material and one immaterial manifestation (similar to the philosophical concept of *dualism*).

Information has a reproducible effect; that is to say, the same information in the same system causes some change and the change is reproducible. The presence of information cannot be confirmed in the absence of effect, and the information-containing (non-random) nature of the message is not obvious if the effect is not reproducible.

Information is bi-directional because the sender and recipient must have similar or identical properties; they must speak the same language to be able to formulate (send) or identify (receive) the information. This property can lead to the illusion that information exists *between* sender and recipient, simply oscillating between them. It can also lead to the concept of materialization of information. However, imagine two tape-recorders endlessly playing the same message to one another. Is this yoyo message "information"?

What makes information directional? One pole of the bidirectional information path has to be equipped with an executive arm, which executes some effect triggered by the information and defines that pole as the recipient, while the other is the sender. This indicates that having an effect is an essential property of information.

The requirement that sender and receiver must have some similar properties suggests that the same information exists in at least two locations: the sender and the recipient. It may appear logical to speak about 'information transfer' between sender and recipient, but this is not the case: it is not information, rather data transfer. It is not possible to decide whether or not a sequence of data is information without recognition (reception) of the message. In other words, information exists at its origin and destination, but not between the two. The inference is clear: non-random flow of unrecognized data is not information until it is recognized.

Non-random dataflow indicates the existence of information because it clearly manifests the existence of a sender, and where there is a sender there is a potential recipient. However, non-random flow of data cannot be identified as information without locating the recipient and the relevant effect.

The world is likely to contain non-random data that we simply do not recognize, so it has never become information for us. There is a human perspective on information. Humans, as intelligent subjects—often only external observers of non-random data exchange—can locate senders and recipients. Such external observation requires the construction of artificial receptors that can recognize information flow. However, humans as external artificial recipients occupy a unique position: they recognize non-random data flow as information only if it makes sense to them, that is to say, if it has some human meaning in addition to the meaning recognized by the primary (intentional) destination of the message. ***Information is data that make sense.***

I met *Michael Waterman* in Bangkok, Thailand, in 2002. We were visiting InCoBio. I became upset at the arrogant lecturing style of a young fellow and rebuked him with critical remarks. Michael stated "I didn't like that man either". That was the beginning of our friendship. Michael is often overlooked as one of the most important contributors to recent bioinformatics (Smith-Waterman algorithm, 1979) owing to his informal style; he lectures wearing jeans and a T shirt but he is a member of the American Academy of Sciences and my work concerning BlastNP benefits from his equation. I remember we were drinking coffee in a Thai hotel surrounded by several advertisements, many written in the Thai language, of which I have no understanding (although I like the round letters). I asked Michael what his definition of biological information is. I'll never forget his surprised, almost irritated look; I was supposed to know the definitions of information and biological information. He responded: "You can find it in any dictionary". I understood the point and replied: "OK. Tell me; is that information on the wall or just gibberish?" I pointed to one of the Thai advertisements. He did not speak or write the Thai language and therefore was unable to answer. However, he understood my point and we continued our discussion with increased respect for one another.

2. Measuring Information Entropy

The view of information as a message became prominent with the publications of Ralph Hartley in 1928 [1] and Claude Sahnnon in 1948 [2]. These papers established the foundations of information theory and endowed the word 'information' not only with a technical meaning but also with a measure. If the sending device is equally likely to send any one of a set of N messages, then the preferred measure of 'the information produced when one message is chosen from the set' is the base two logarithm of N. (This measure is called self-information.) In his paper, Shannon continues:

"The choice of a logarithmic base corresponds to the choice of a unit for measuring information. If the base 2 is used the resulting units may be called *binary digits*, or more briefly *bits*, a word suggested by J. W. Tukey. A device with two stable positions, such as a relay or a flip-flop circuit, can store one bit of information. N such devices can store N bits…" [1].

The method suggested by Hartley and Shannon for the quantitative measurement of information is the summative measurement of certain properties of a message (or signal) such as density, length and variations, and has nothing to do with information (at least not in the original, established meaning of this word). Information exists at the sender and at the receiver but not between them. I dislike expressions such as 'information flow' as information does not flow, it translocates. The Shannon equation is practical for characterizing a signal (or message) and estimating the physical space it may occupy; most random sequences give the highest possible entropy value (bits).

This remark concerning Shannon's work might sound bold, provocative and impolite, but this is not the intention. It is widely recognized and discussed in the literature that his terminology is not always fortunate.

Shannon's *information entropy (H)* is often confused with the *physical entropy (S)* because both concepts have a very similar mathematical formulation. However, they have very different meanings. Thermodynamic entropy characterizes a statistical ensemble of molecular states, while Shannon's entropy characterizes a statistical ensemble of messages. In terms of thermodynamics, entropy concerns all the ways molecules or particles might be arranged, and greater entropy means that less physical work can be extracted from the system. In Shannon's usage, entropy concerns all the ways messages could be transmitted by an information source, and greater entropy means that the messages are more equally probable. It is disturbing and some authors, such as Tom Schneider [3], argue for dropping the word entropy for the H function of Information Theory and using Shannon's other term, *uncertainty* (average surprisal), instead.

Some remarks have been attached to the use of the word *information*. An important point to consider is that both the information source and the noise source are stochastic processes; both could be treated as information sources. The main difference between them is that the receiver is interested in the information source and wishes to ignore the noise source. In some situations, a noise source is intentionally observed, in which case it becomes an information source. However, sometimes information is copied unintentionally from one channel to another. This is called cross-talk, and the result is that someone's uninteresting information source is considered noise. One person's information is another person's noise and vice-versa. It is the interest of the observer that changes a stochastic process into an information source, *i.e.*, one particular message is more interesting to the receiver than any other, as it has meaning for the receiver while others do not.

Shannon's generous use of the term *information* is understandable if we consider that he worked for the Bell Telephone Laboratories and his famous paper, "A Mathematical Theory of Communication" was published in the Bell System Technical Journal. The publication was meant to address a technical question, the transfer of meaningful messages (information *vs.* noise) to a narrow audience consisting of other communication engineers. The year was 1948, five years before Watson and Crick suggested the structure of DNA (the carrier of biological "information").

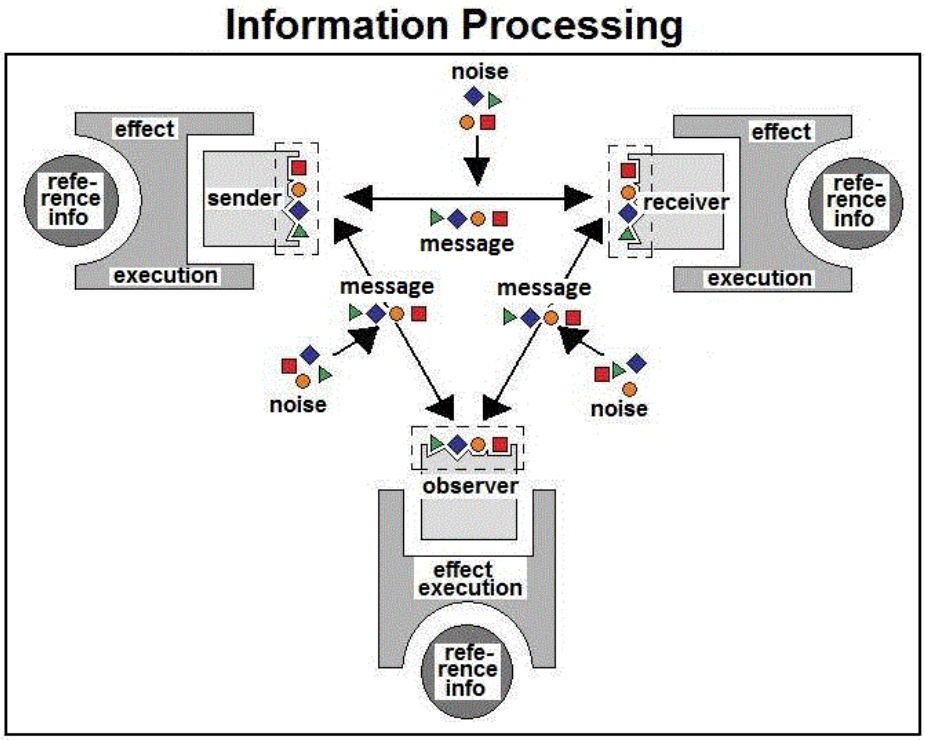
In information theory, signals are part of a process, not a substance; they do something. They do not contain any specific meaning, suggesting that the most random signals contain the most information, as they can be interpreted in any way and cannot be compressed. This is a nonsensical paradox as random sequences make no sense; their information content is zero. Information is related to non-randomness.

Speaking about *meaning* and *sense* in connection with the definition of information could have a negative consequence. *Meaning* and *sense* are most often associated with human consciousness and intelligence; for example, when an animal leaves a footprint, a signal is created with no consciousness or meaning. However, the footprint becomes information and contains meaning if somebody can identify the animal by 'reading' it. Generally, the term 'meaning' does not involve consciousness.

The bi-conditional definition of information (*i.e.*, simultaneous presence of a non-random message and a responsive/receiver) has additional consequences. The recipient has a very important role in defining and quantifying the information part of the message.

In optimal cases, the information content is equal to the signal entropy (as defined, for example, by Shannon). However, there are cases when signal recognition is suboptimal and the information is compromised. In other cases, a simple signal can trigger a cascade of reactions, where much more information is involved than the original signal can possibly contain (Figure 1).

Figure 1. Information Processing. Information (a given order of elements) exists at three places: the sender, receiver and observer. They are spatially separated but their construction is similar. They uniformly contain some reference information to distinguish signals from noise. There is an executive function in each that creates or stores the order in the signal, and is responsible for a response to the message. Noise is the un-ordered occurrence of elements.

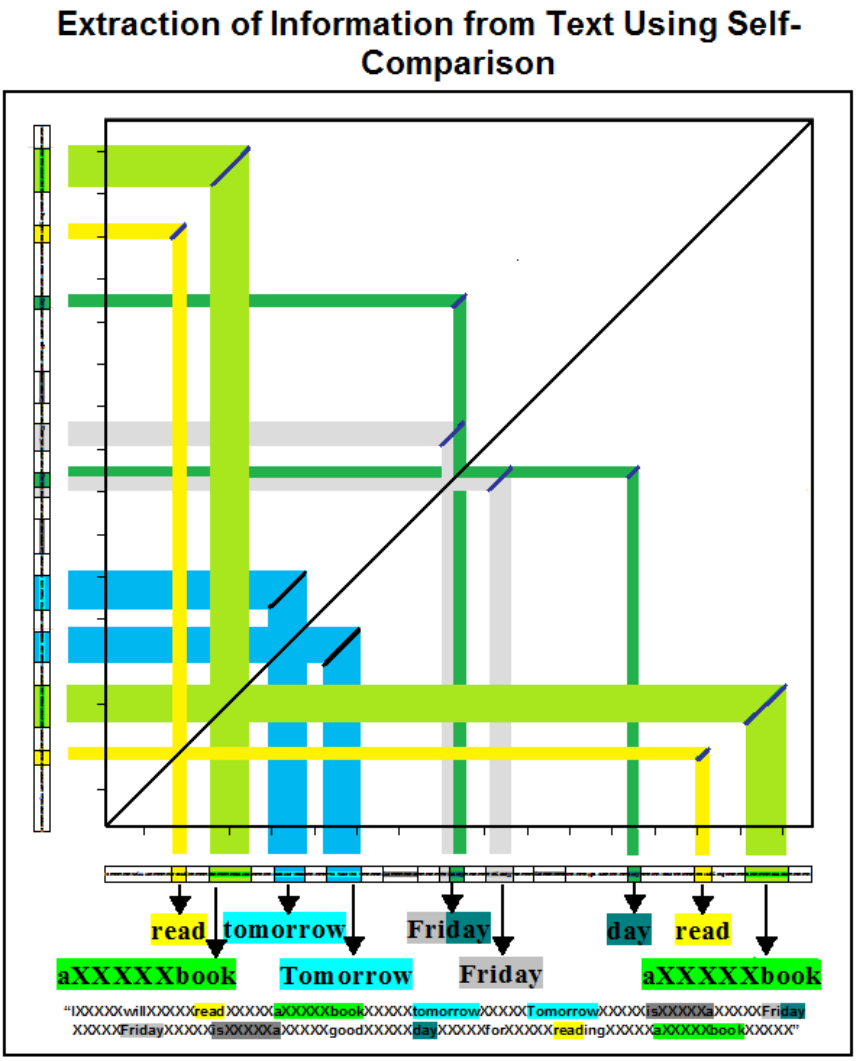


Signal, information, message and randomness have several definitions. The above description is chosen because it is easily understood without advanced mathematics (the words has an established meaning for the general public). In addition, it can logically be applied to biological systems, where signal and receptor are molecules that interact specifically with high affinity.

3. Information Gathering

The bi-conditional definition of information clearly states that information does not exist without recognition; a signal (data) is not itself information. How do you generate information when you have nothing other than the message? The answer is surprisingly simple: duplicate the message and use one copy as a tentative receptor. This method is effective for analyzing texts or text-like sequences (Figure 2).

Figure 2. Extraction of information from the text using self-comparison.



Suppose an extraterrestrial observer receives a message from the Earth:

"I will read a book tomorrow. Tomorrow is a Friday. Friday is a good day for reading a book."

The message appears like this on his receiver:

IXXXXXwillXXXXXreadXXXXXaXXXXXbookXXXXXtomorrow.XXXXXTomorrowXXXXXisXXXXXaXXXXXFriday.XXXXXFridayXXXXXisXXXXXaXXXXXgoodXXXXXdayXXXXXforXXXXXreadingXXXXXaXXXXXbook.XXXXX

A simple self-comparison of the message reveals that (1) it is 165 units (letters) long, (2) there are 19 different letters, (3) there are 19-times-5-unit-long uniform signals (X), (4) there are 19 longer units (words), (5) five of the longer units appear twice and it is possible to calculate that (6) the probability of repeated occurrence (by chance) of the longer unit "tomorrow" is $(1/17)^8 = 1.4E^{-10}$ under the given circumstances. The observer can conclude with certainty that this message is not randomly generated. This is a lot of information (data that make sense to an intelligent extraterrestrial observer) generated simply from only a few data. If you use the same type of self-comparison on the entire collection of Shakespeare's works, you can determine exactly the number of different letters and words in the

English language and its grammatical rules. However, you may still not understand a single word or concept in the English language.

This is how the first step of bioinformatics data analysis works. The protein and nucleic acid sequences are collected into large databases and analyzed for non-random regularities. The results are collected and sorted into secondary databases following predefined rules. These secondary databases are rather like dictionaries, listing returning signals or signal combinations (patterns) [4].

4. Data > Information > Knowledge: Knowledge Emerging from Information

The data > information transition is the first step of data processing. There is a second step, in which the information becomes integrated and transformed into knowledge. Some pieces of information complete each other, elucidating things and processes from different angles and perspectives. These complementary pieces of information attract one another in the human (intelligent) brain and have a tendency to fuse. This is similar to putting a puzzle together. We try the puzzle pieces two by two for compatibility and after a while a picture is revealed. The puzzle pieces acquire meaning together. Similarly, different pieces of compatible information in sufficient amounts can complete each other and provide a higher level of understanding. This higher level of understanding, or insight, is the birth of knowledge. The birth of knowledge is readily recognizable from its general properties:

1. A mass of information 'melts' together. The number of pieces of information necessary to describe the system is suddenly and dramatically reduced. Knowledge reduces the information necessary to describe a system accurately. Knowledge is **compact**.

2. A logical picture emerges where the elements are interconnected. You can remove any element from the picture; the missing element is predictable from those remaining and you can reproduce the picture. **Reproducibility** and **predictability** are key elements of knowledge.

3. The interconnectivity of information in knowledge is often unidirectional and a change in one element causes change in another, but not in the reverse direction. **Causality** is an important aspect of knowledge-based relationships.

There are many special properties of knowledge that should be considered when speaking more concretely about the subject:

4. Knowledge may be **partial and/or situated**, *i.e.*, valid only with limitations of time and place.

5. A distinction is often made in philosophy (epistemology) between ***a priori*** and ***a posteriori*** forms of knowledge. *A priori* knowledge is justified by reason or arguments (without experience) while *a posteriori* knowledge is justified by experience or experiments.

6. Knowledge can be gained by trial and error, intuition or learning. Trial and error is the slow, regular, 'difficult way'. Intuition bypasses the difficult way, if it occurs, but it is unpredictable. **Learning** is the 'easy way' but it only works after significant preparation for receiving knowledge (see later).

7. The validity (quality) of knowledge is highly variable. The existence of 'absolute' knowledge is highly questionable, but some natural laws are strong and generally valid with very few restrictions.

Others are weak with little or no predictive value. Validity distinguishes **science from religion**, knowledge from belief.

8. There is a human aspect to knowledge. Knowledge is a force creating order, and order is biologically highly rewarded: it feels good. Disorder causes discomfort (at least for those who are intelligent enough to be able to detect disorder). The transition between disorder and order gives the experience of '**eureka**' and satisfaction.

9. The border between knowledge and belief is often poorly defined. Belief, even religious, often has some experience/observation as its starting point. However, much science-based knowledge started as strong belief or **intuition**.

10. Knowledge is not static. It changes dynamically as its building blocks, interconnected information, can change with time. This change does not necessarily alter the content of knowledge, but it can alter its area of validity and application. Knowledge has to be continuously used (tested) to keep it alive. I would say that as **there is no religion without belief, there is no science without disbelief.**

11. Knowledge is usually '**know-how'**, but know-how is not necessarily knowledge. We have learned how to generate antibodies in rabbits by injecting them with antigens (know how); however, we do not know how to make such antibodies (the rabbit's biology knows that).

12. Knowledge often arises spontaneously when available information reaches a critical mass. However, this process can be artificially facilitated by science and by using scientific methods. Overall, information is the result of processing, manipulating and organizing data in such a way that they add to the knowledge of the person receiving them.

Knowledge can be obtained from several sources including books, the internet, and from speaking to other individuals. However, Aporoksha Gnyana (also spelled Aparoksha-Jnana) is the knowledge borne of direct experience, *i.e.*, knowledge that one discovers for oneself.

13. Statistically generated connections are not knowledge; neuronal networking and computer prediction generate no knowledge.

14. A distinction between knowledge (science) and belief (religion) should be made clear, but it is not easy. Philosophical debates generally start with Plato's formulation of knowledge as 'justified true belief'. Knowledge is a strict abstraction, while belief is a loose one—by others.

Information is integration of data used to make sense, and knowledge is integration of information to a compact meaning. Knowledge is understanding, a deep and compact dynamic reflection of stable or general connections among things and events.

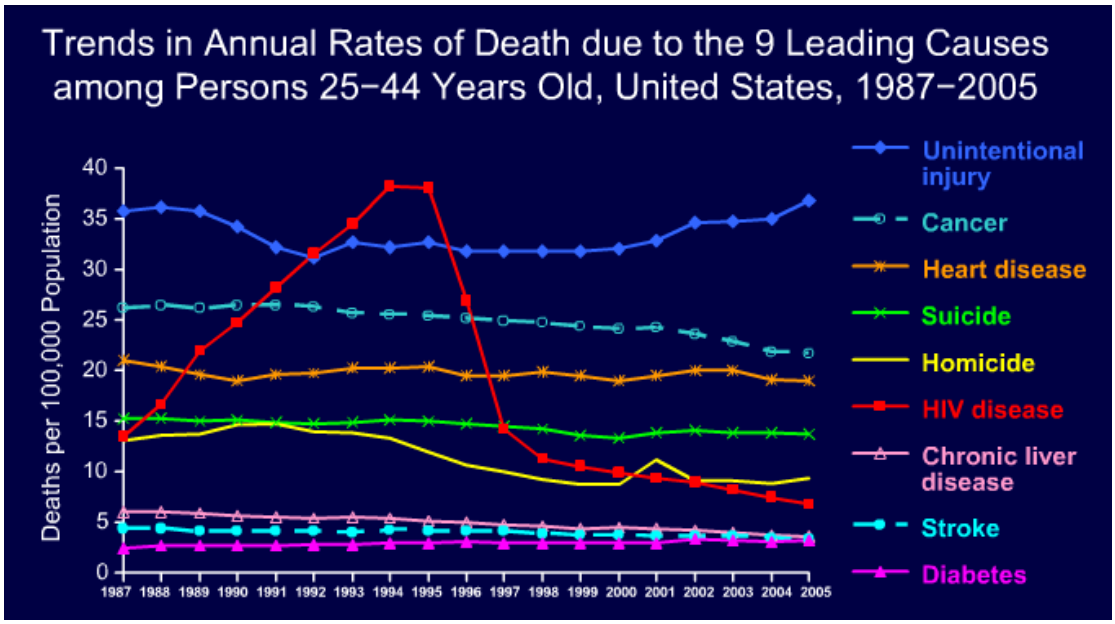
5. Growth of Biological Data and Information

It is usually possible to observe the absence or presence of knowledge, but to measure it quantitatively is challenging. The effects of knowledge might be measured in some way.

The number of HIV-related deaths has significantly and progressively decreased since 1997. The number of HIV infections has increased slightly (not shown). This indicates that effective treatment

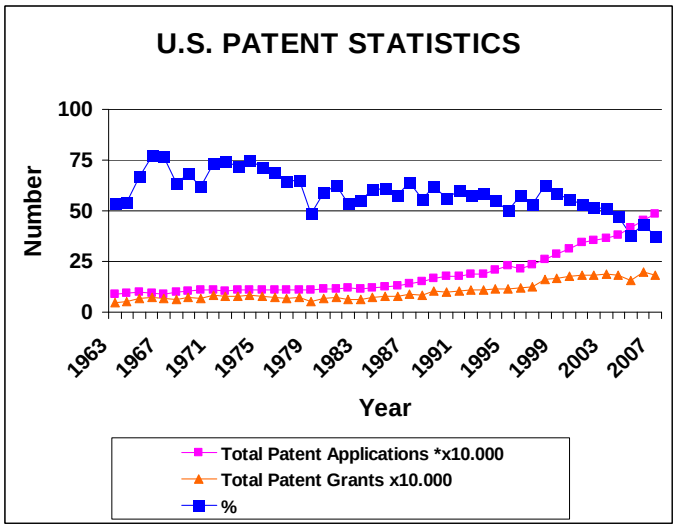
options have been introduced and successfully applied. This is new knowledge in action. At the same time, the death rate related to other conditions has remained unchanged, indicating the absence of new knowledge [2] (Figure 3).

Figure 3. Annual rates of death in USA [5].



The total number of patent applications increases exponentially, while the proportion of total, granted patents presents with a linear decrease, indicating that new knowledge grows much more slowly than information [6] (Figure 4).

Figure 4. U.S. Patent Statistics [6].



The discussion concerning patents (applications *vs.* grants) is not as simple as portrayed. The number of grants is always lower than the number of applications, and the assumption is that those rejected patents concerned inventions that were not new. There are a number of reasons why patent

applications are rejected including inventions that fall outside the statutory limits of the patenting process (this is changing, but not as fast as many new fields), and a variety of sophisticated strategic uses of the patenting system by commercial concerns to block competitors from erecting patent fences around areas of interest (which seems counterintuitive, but is a common practice), by abandoning applications during prosecution. Other applications are abandoned simply because the expense of further prosecution may not be warranted for a specific technology as the cost-benefit ratio may be too low.

This serves to illustrate the difficulties of quantifying knowledge.

Cancer-related deaths have been remarkably stable over the period 2000-2008, indicating the absence of new knowledge regarding cancer treatment during this period. However, detailed cancer death statistics for a seven-decade-long period demonstrate that fewer people are dying from some forms of cancer but more from others. This demonstrates that knowledge of how to prevent cancer deaths generally remains elusive [7] (Figure 5).

Figure 5. Estimated New Cancer Cases and Deaths in USA [7].

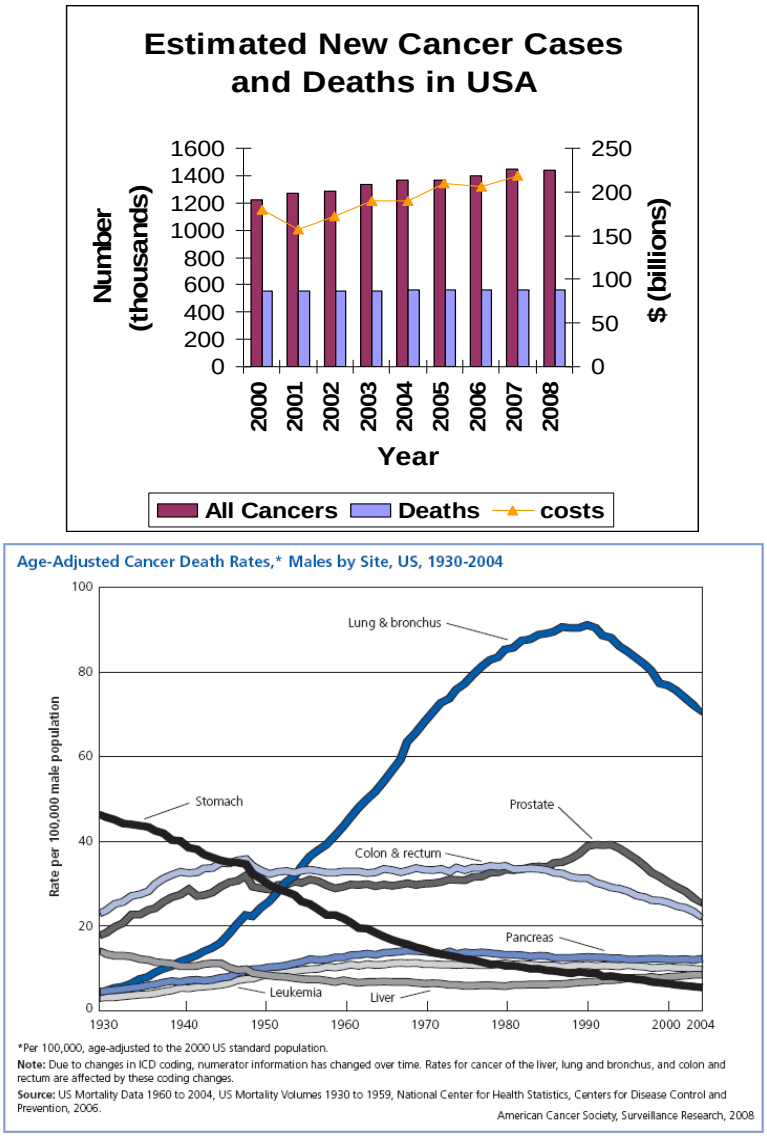
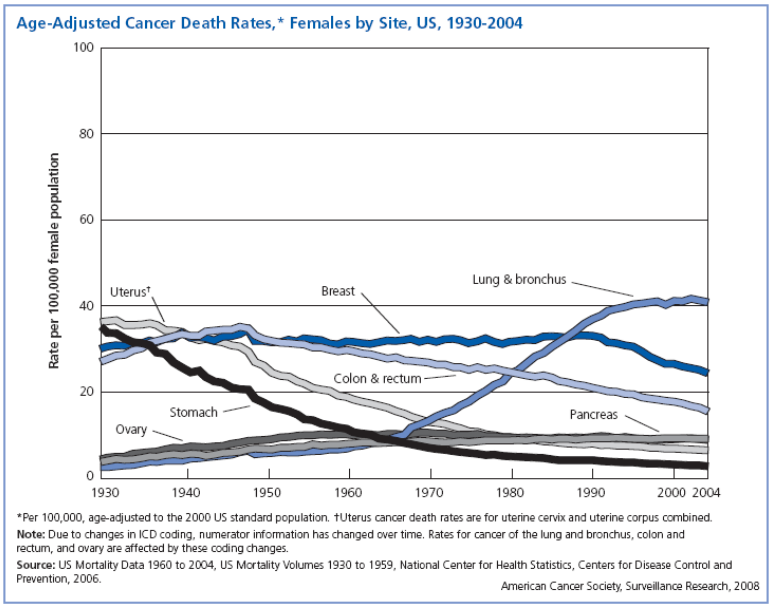
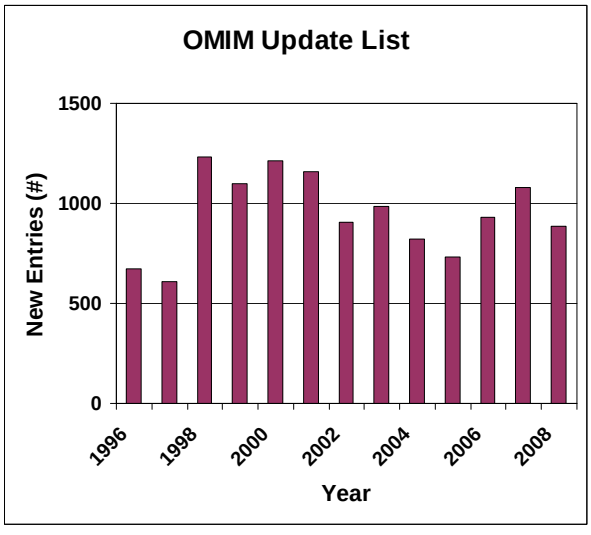


Figure 5. *cont.*



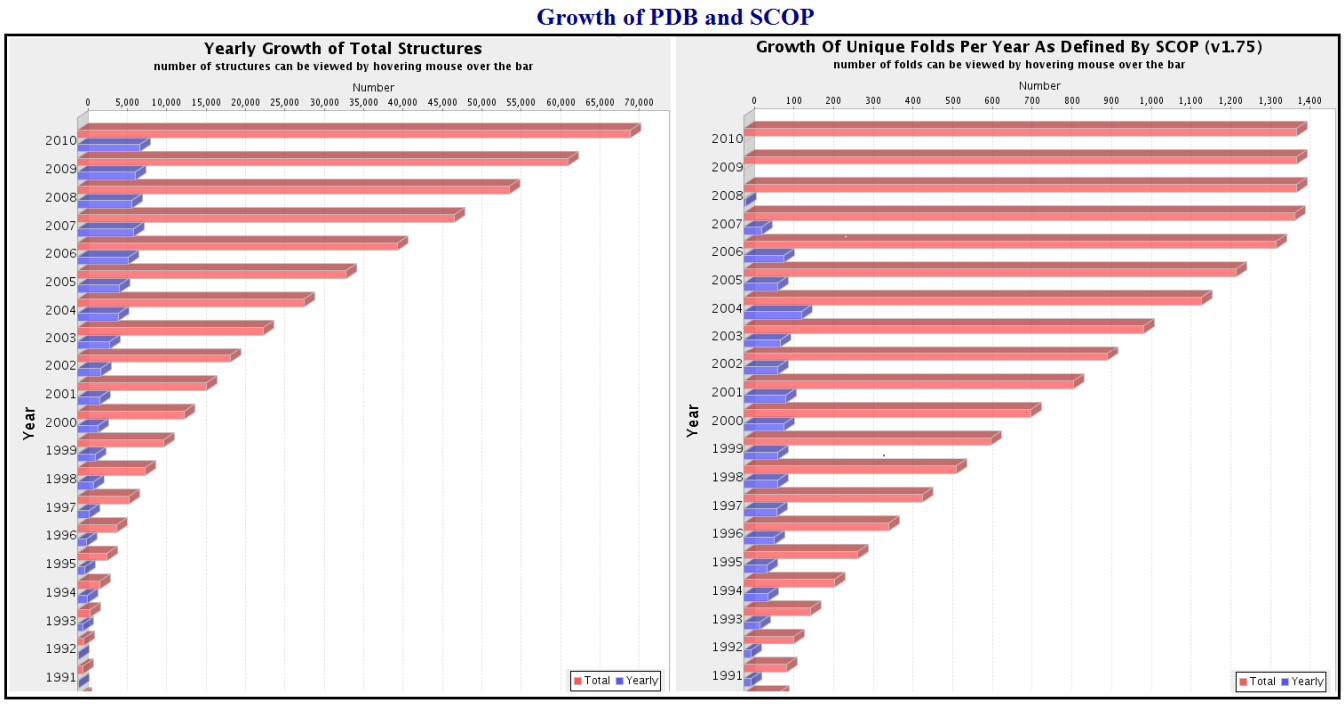
OMIM (Online Mendelian Inheritance of Man) is a well-known knowledge database connecting human genotypes to phenotypes (disease). The number of new entries appears not to be affected (yet) by the ongoing Human Genome Project (HGP, which was completed 2003). The HGP has provided data, but not information or knowledge, to this field [8] (Figure 6).

Figure 6. OMIM Update List [8].



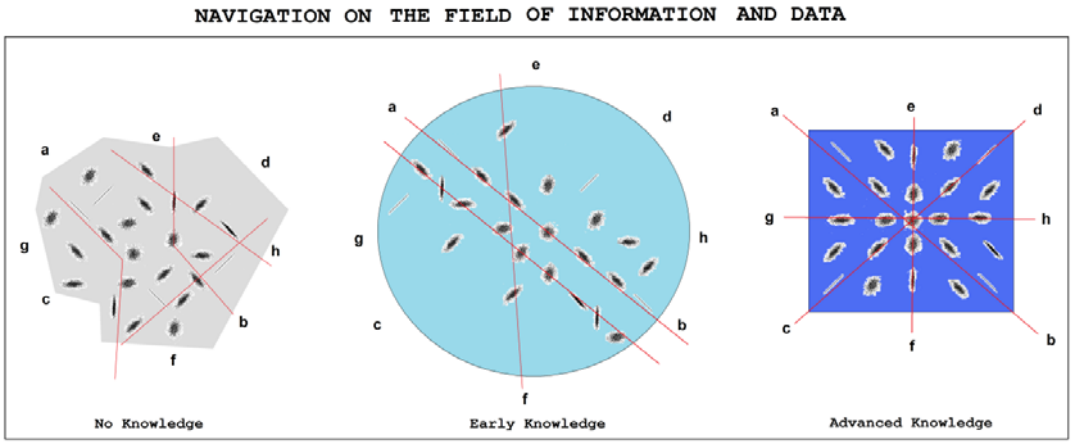
The number of protein structures in the PDB is growing exponentially. However, this steady growth is no longer accompanied by the addition of any unique folds to the SCOP database: this information has now been completely extracted from the structural data [9] (Figure 7).

Figure 7. Growth of PDB and SCOP [9].



The relationships of data to information, and of information to knowledge, are illustrated visually in Figure 8.

Figure 8. Navigation in the field of Information and Data. Data (dots) are sorted into information (elliptical forms). The information may or may not complete (connect) different fields of existing knowledge (a-h, lines).



Thousands of measurements (observations, *data points*, black dots) were analyzed and sorted into 21 groups (representing *information*, elliptical forms). Even if every single group of data (information) makes some sense (defining a direction in the space), the groups do not fit together; the connections among fields a-h remain unidentified (no *knowledge* is gained). Rearrangement of the information indicates the possibility of connecting a-b and e-f, but the picture is not clear; there are many exceptions and disturbing elements during this early stage of knowledge. After additional maturation

and rearrangements of the 21 available groups of information, a clear picture emerges and reveals a network of connections among all the available information. New knowledge has been gained and integrated into fields a-h, defined by different knowledge.

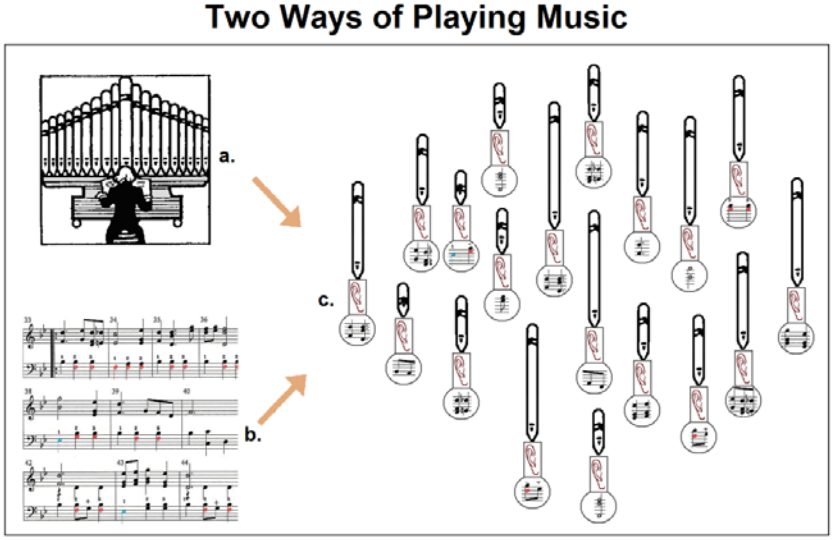
The situation described by this abstract example is familiar to most scientists. There are often several different, seemingly opposing ideas, suggestions and hypotheses in a newly-defined area of science. However, after some time (often decades of maturation), the conflicts are resolved, the questions are answered and new solid knowledge becomes accepted and consolidated. This consolidated situation will exist until it becomes necessary to incorporate additional knowledge (for example, i & j).

6. Human Ways of Thinking and Communicating

Data collection, information extraction and knowledge building are intimately related to the transfer of meaningful messages in a meaningful way: communication and signaling. Formally, the connection between sender and recipient may be linear (one to one), network-like (several to several) or somewhere between these two extremes.

The most widely recognized form of message formation, transfer and reception is illustrated in Figure 1. However, there is a fundamentally different way of generating and receiving signals, which we call **disseminated or fragmented**. I mention this as it is widely practiced *in vivo* by multicellular organisms, particularly those that have developed a nervous system. Imagine playing a musical instrument, reading a note and generating a melody. There are two different ways to do this (Figure 9). First, a trained musician (a) reads a note and operates the instrument (b); this is the centralized method of music-playing as it depends completely on the musician. An alternative way is to multiply and disseminate the notes to the units of the instrument (for example to the pipes of an organ). In this case, the 'pipes' listen to one another and they contribute their tones to the note at the places and in the sequence determined. The musician is removed and the placement of the pipes is no longer important. The organ is fragmented (disseminated) into its units, but the melody itself remains intact (c). This is the disseminated method of signal/message generation.

Figure 9. Two Ways of Playing Music.



A similarly fragmented system may work in the opposite direction and function as a receptor. Our sensory organs (vision and hearing) operate in this way. There are approximately 20,000 sensory units in the organ of Corti (cochlea, inner ear), each responding to a narrow sound frequency. Complex sounds (many different frequencies) excite several sensory units and form a neuronal pattern that is analyzed by the nervous system and becomes meaningful speech or melody. This disseminated method of signal generation and reception is well suited to multi-cellular biological systems.

7. Complexity (Simple—Complex), Chaos (Order—Disorder)

Something is 'simple' when it is composed of few or identical elements and the relationship between the elements is unidirectional or unilateral. The definition of 'complex' is the opposite: The thing consists of many different kinds of elements, the relationship between them is multilateral and multidirectional, and there is a high degree of interdependence. A simple system is easy to understand by the human observer (it is directly understandable) while complex systems need to be studied before their logic can be understood.

The question of complexity in science is old and recurrent. There are two main aspects to mention:

1. How to describe Nature and how to formulate observations regarding natural laws often present a dilemma: simply or complexly? Some scientists think that Nature is essentially complex and simple descriptions are reductionisms and not scientific approaches. Other scientists, including myself, strongly believe that fundamental natural laws are simple, and much so-called complexity is caused by misunderstanding of, or failure to understand, those fundamental laws.

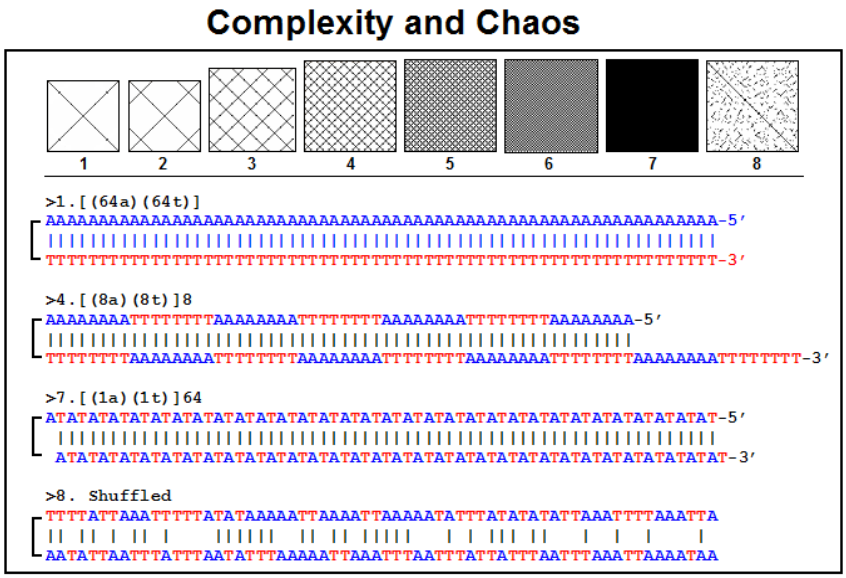
2. Another aspect of complexity is that it is the neighbor of chaos. Order and information content increase with increasing complexity and the time required to understand the system. There is a point on this scale at which the time required to understand the complexity is so great that it is practically impossible to provide. At this point, it is reasonable to regard the system as more chaotic than complex. This view is clearly compatible with the idea that chaotic events do not exist: Very large numbers of observations and a great deal of time spent analyzing these data can provide a logical explanation for any event. I do not reject this suggestion, and accept that chaos arises when you do not have time to understand.

The close relationship between complexity and chaos is easier to accept from everyday experiments. One compelling early experiment regarding chaos and chaos management comes from my medical practice. Every medicine has well-defined effects and side-effects, which are easy to distinguish from the symptoms of a disease. This picture becomes exponentially more complicated when increasing numbers of different medications are prescribed to the same person at the same time. There is a point, well known to practicing physicians, when the symptoms caused by the disease and the side-effects of the medications create an impenetrably complex picture: It is difficult to understand what is going on. In this situation a proficient doctor will reduce the number of treatments in order to clarify the situation; a poor doctor will increase the treatments…

Most people know the Ten Commandments of Christianity. They are easy to learn and remember when required. Statements and principles of law and ethics are more detailed in the 613 Jewish commandments (mitzvot), but can still be remembered. The human attitude to crime and punishment

was ingeniously regulated in the famous legal code written by Hammurabi, where the principle of exact reciprocity was introduced and explicitly used: the principle of 'an eye for an eye'. It is simple and clear. At least, it was simple and clear in ca. 1795–1750 BC in Babylon. The complexity of religious and human laws increased over time, making priests and attorneys practically unavoidable in modern times. Some societies developed a penchant for law-making and started to regulate every aspect of human life in detail. The network of laws became more and more complex in those countries, leading to the flourishing of bureaucracy. Finally, the law became impenetrable in real time, with the consequence that every citizen became potentially criminal no matter what their intention. In these countries (mostly socialist and communist societies), legal chaos (or day-by-day interpretation) is the law, where the 'dominant animal' wins the argument—exactly as in prehistoric times.

Figure 10. Complexity and chaos. A 128-base sequence contains 64 A and 64 T in regular, alternating arrangements. The altering periods are 64 (1), 32 (2), 16 (3), 8 (4), 4 (5), 2 (6) and 1 (7). Identical and complementary subsequences are visualized by the Dotler tool (diagonal lines in the upper quadrangular areas). The sequences form a hairpin-like structure (mFold method, lower part of the figure). No further shortening of the alternating periods in the sequences is possible (max. complexity in sequence and dotplot is near 7) and the system becomes chaotic (8).



We can learn a remarkable example of chaos management from Nature. It is now known that chromosomal damage accumulates in our genes over time. We know that acquired characteristics do not become inherited; genetic variation is the key to accommodating to a changing environment and such variation occurs during sexual reproduction and is manifest in new generations. We know that the human brain loses its flexibility during aging and becomes more and more saturated with old knowledge and anachronistic ideas. The brain loses its ability to see things from new perspectives. All these changes in the human body make it more and more complex, and rigid and chaotic with time. Nature's way of managing this chaos is to kill old individuals. The material and ideal things produced by each human remain intact for reuse by the species, but the single individual has to go. Some genes

survive, thoroughly reshuffled during sexual reproduction, but the original combination will never return. It is a simple and universal rule, no exceptions. (I am not happy about efforts at cloning; they seem to me a crude violation of one of Nature's most fundamental laws.)

The "simple", "complex" "chaos" transition is illustrated in Figure 10.

It is often impossible to decide whether a system is highly complex or chaotic. Any encrypted message is supposed to give the impression of chaos. A method of 'self-comparison' can provide some guidelines. If the pattern with self-comparison shows any regularity (order), the system may contain information. The human brain is excellent at *seeing* regularities in patterns, much better than a computer. Even the most devoted computer fan should learn to trust his brain to distinguish between chaotic and organized patterns.

8. Destruction of Information *vs.* not Receiving it: Translation, Junk DNA

The complexity *vs.* chaos dilemma raises the question of mixed signals *i.e.*, when the signal is not completely chaotic but contains no clear information. It is rather obvious that noise (chaos) destroys signal quality. It is also clear that information entropy (measured in bits as described by Claude Shannon) indicates the maximum possible quantity of information in a given message, without revealing anything about the real information content of that particular message.

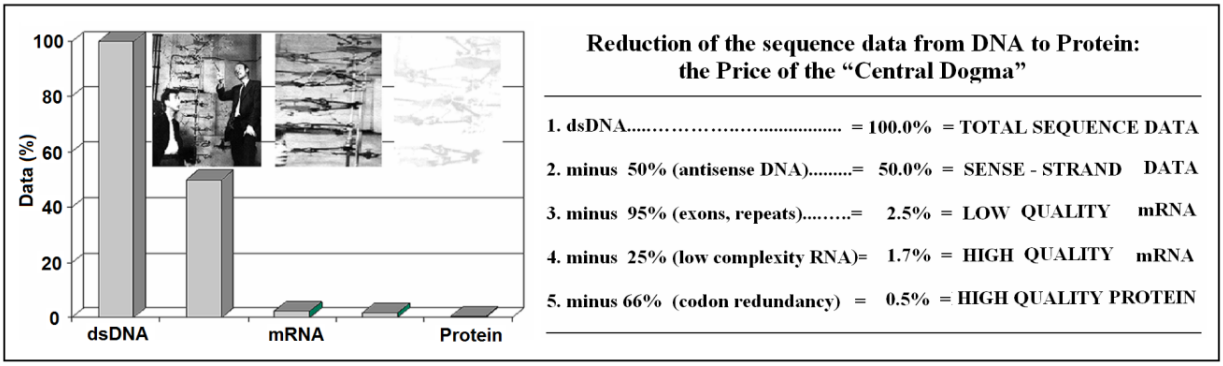
Molecular biology provides two famous examples of the dilemma concerning information and/or noise in the same message:

1. Translation: the process in which biological information in nucleic acids is transmitted into protein sequences. The exact rules of translation are described in Nirenberg's Genetic Code, which is simple and universal [10]. However, there is a hidden problem; every single amino acid is encoded by (on average) three different codons, *i.e.*, the Genetic Code is redundant. The information content of the codon is $\log_2 [4 \times 4 \times 4] = 6$ bits/codon, while the information content of an amino acid in a protein is $\log_2 [20] = 4.3$ [11]. What happens to the difference? Is it just noise? Is it a kind of redundancy necessary to protect against mutations? Is it further information, additional to the Genetic Code, that has a biological role but we do not have the connection yet? (I will return to this question.)

2. Only 2-3% of the genome comprises protein-coding sequences (exons). The remainder, introns, have poorly-defined functions, and were previously called "junk DNA" or "parasitic DNA" [12]. However, self-comparison reveals that this part of the genome is far from random; it contains highly ordered and conserved repeats. It is extremely arrogant to call highly-ordered structures "junk" just because we do not understand their function.

Estimates for the number of cells in the human body range between 10 trillion and 100 trillion [13]. The human genome is estimated to contain some three billion base pairs. This amounts to approximately 330 g DNA/adult human body; this is all genetic material which has been highly conserved for millions of years. Is it really a good idea to decide that 99% or more of this genetic material is 'junk'? (Figure 11)

Figure 11. Reduction of the Sequence Data from RNA to Protein: The Price of the "Central Dogma". There are 3×10^9 base pairs in the human genome (100% data), but only a tiny fraction (0.5%) are needed for protein synthesis. The three inserted pictures illustrate the catastrophic loss of meaning (information) when the number of pixels is reduced from 100 to 2.5% and finally to 0.5%.



It is clear that a huge number of data are lost during the transition from DNA to protein. It is also clear that the lost data are not random, but ordered and carefully preserved from one generation to the next. Therefore, they are not only data, but certainly biological information. However, the knowledge represented by this information is as yet unknown.

9. Signals as a Special kind of Information

A very important, distinctive characteristic of information is that it comprises data that make sense; it has meaning for the sender and the recipient. The 'meaningfulness' of information (quantity) depends on two factors: First, the complexity of the message, and second, the complexity of the sender and receiver. A complex sender with access to a large amount of information can create complex messages, and a complex receiver is prepared to make sense of complex messages.

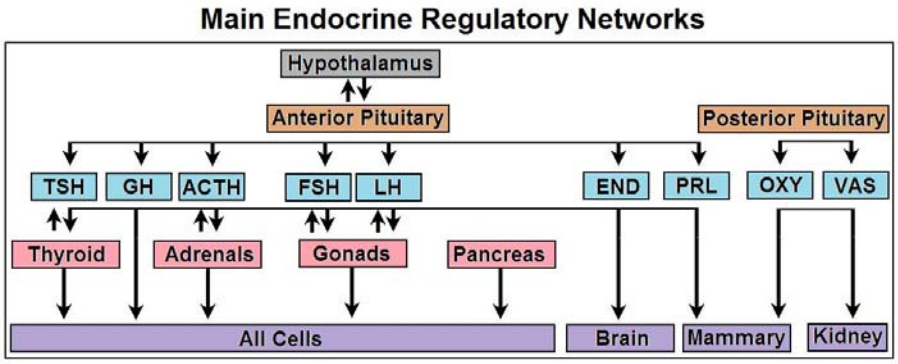
Information results in some kind of change in the recipient. It could be a very discrete change such as storage of the information in a databank for later use, or extensive modification of the receiver's behavior. The size and complexity of the message are not necessarily proportional to the size of the change it causes. For example, the description of an object can take up a lot of storage space in the receiver's database, while other messages are more instructive than descriptive.

Everyday human language is surprisingly poor in terms of information, approximately 1.3 bit/character [14]. There are 26 letters in the English alphabet and there are 60 letters in a single line of a printed book. This means that the number of maximal letter combinations in a line is 26^{60}, an astonishingly high number. Most letter combinations are never used to form words and most word combinations do not show up in meaningful texts. However, years of training are required to understand the words of a language, and to understand the meaning of some special word combinations may take many additional years of study. We can use language to give instructions. In this case, the message can be short but the sender and receiver must be properly trained. Alternatively, language can be used for detailed information transfer. In this case, the message will be long and complex, but the recipient does not need to be specially prepared.

The situation regarding chemical communication between cells is similar. Cells normally respond to very short molecular messages in very specific ways. This specificity and discrimination between signals develops during differentiation. Every cell in an organism contains the entire set of genetic information, carrying the characteristics of that species, at the beginning of its lifespan, and has the potential to perform a wide range of functions (pluripotent). However, as this wide potency is successively replaced by fewer very specific functions, the cell becomes oligo—or unipotent (differentiated and specialized). A specialized cell is able to respond to a few, short signals promptly and accurately.

Signals are short instructive messages that elicit pre-determined responses in the receiver. The endocrine system is a well-understood signaling network in organisms with developed circulatory systems; hormonal messages are transported by the circulatory system. It is well known that there is a local communication network between cells located adjacent to or very close to each other, called autocrine and paracrine regulation, respectively (Figure 12).

Figure 12. Main Endocrine Regulatory Networks.



The hypothalamus (part of the brain) controls the anterior pituitary gland, which is the 'conductor of the hormone orchestra' and has a vital but indirect influence on several aspects of metabolism through hormones produced by the thyroid, adrenal glands and gonads. Growth hormone (GH), Prolactin (PRL) and Endorphin (END) act directly on their target cells. Oxytocin (OXY) and Vasopressin (VAS) are two oligopeptides from the posterior pituitary gland that have direct effects on the mammary glands and kidneys, respectively. The endocrine pancreas (Islets of Langerhans) produces insulin, which controls and is controlled by the blood glucose level.

Biological signals may provide examples of the kind of dilemma a scientist might face when trying to measure the information content of a biological message using Hartley & Shannon's method without biological knowledge or consideration.

How much information (in bits) is necessary to describe the difference between man and woman? The answer depends on the person trying to calculate the answer. If you ask a librarian, you will end up with terabytes. It will be some similarly large number if you ask an artist, film-maker or psychologist. If you ask a young, well-educated bioinformatitian, the 'professional' answer might be 490 bits. His logic is the following: males have a Y chromosome, which is not present in females. Much of this Y chromosome is composed of sequences present in X chromosomes, except the SRY (Sex Related Y) gene. Ultimately, the genetic difference between genders is that males have the SRY

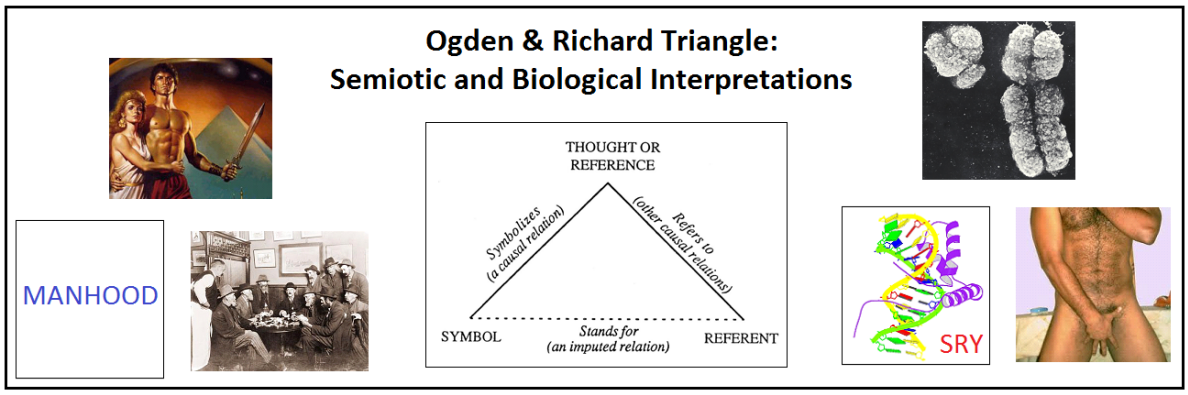
gene, females do not. This gene produces a 113 amino acid long protein which signals to the embryo to develop into a male. A 113 amino acid long protein (20 variables at 113 positions) is a maximum 490 bits large message. Some scholars call the SRY protein the most intriguing letter ever created… (Figure 13).

Figure 13. Human SRY Protein [15]—"the most intriguing letter ever created".

```
>HUMAN SRY PROTEIN
NVQDRVKRPM NAFIVWSRDQ RRKMALENPR MRNSEISKQL GYQWKMLTEA
EKWPFFQEAQ KLQAMHREKY PNYKYRPRRK AKMLPKNCCL LPADPASVLC
SEVQLDNRLY RDD
```

There is another way to perform a similar calculation. Testosterone is a powerful male hormone. A fetus exposed to testosterone *in utero* will develop a male phenotype even if she is genetically XX. What happens to a male embryo (XY) if the testosterone receptor is missing? He will not develop into a male phenotype, but remains female and develop a female body. The condition is usually discovered in puberty due to the absence of menstruation. They often marry and live a normal female life without biological descents (see testicular feminization for details [16]). The binding domain of the human androgen (testosterone) binding protein hormone is a 47 amino acid long sequence, representing approximately 203 bits. These kinds of calculations may be impressive but they do not help to understand the biological information of 'maleness'.

Figure 14. Ogden & Richard Triangle (middle) and its semiotic (left) and biological (right) interpretations. A *thought or reference* might be reflected by a *symbol* and a *referent* (continuous lines); the symbol stands for the referent but they are not in direct relationship with one another (dotted line). The world 'manhood' symbolizes an idea [say: "Be on the alert, stand firm in the faith, act like men, and be strong."—Corinthians 16:13 (NAS)] but it may also be used for a group of drinkers (referent). In biology, the concept of manhood is *written* in the XY genotype (chromosomes) and the referent is a masculine man. The SRY protein is the biological symbol (signal) of masculinity and responsible for the extra hair, muscle and testosterone.



Biological signaling resembles the linguistic phenomenon described by the Ogden & Richard semiotic triangle in 1923 [17]. It is a model of how linguistic symbols are related to the objects they represent. It has to be interpreted in the following way: there is no direct relationship between the word and 'symbol' denoted by it. This is symbolized by the broken line connecting the two, which Ogden and Richard characterize as 'an imputed relation', saying that the 'symbol' stands for the 'referent' (Figure 14).

10. Unique and Distinguishing Properties of Biological Information

1. Signals used in communication technology are always two dimensional, analog or digital changes in the carrier medium (which are normally electromagnetic waves in the space or electrons/photons in cables). Biological signals are always **3D molecules** transferred in wet medium. Biological communication is **wet** communication, in contrast to technical communications, which are usually **dry**.

2. Two very important categories of biological signals (molecules) are sequences. Nucleic acids (DNA and RNA) are composed of four variables (the bases) and proteins are built from 20 variables (the amino acids). These sequences are easily represented by binary units (bites) to estimate their signal density and to compress these binary signals as described by Shannon. However, the biological sequences **fold,** often have more than one 3D configuration and represent different kinds of biological information. It should be noted that these allosteric configurations have the same primary sequence and the same calculated signal density (or information content). It is possible to represent 3D structures as a series of binary signals and treat them like any other communication signal. However, it is an untested area and the 'fairness' of this treatment—from a biological point of vie—is not yet known. (The history of information science provided many examples of mistakes due to ambitious efforts to use Shannon's communication theory in areas where it does not belong).

3. The **receptor** (the recipient of the molecular signals) is an essential component for processing molecular (biological) information. The receptor is always (not surprisingly) a protein structure which forms a kind of mirror image (mold) of the molecule involved in communicating a biological signal (called a ligand). The receptor and ligand interact with one another and this interaction is highly specific and unique. Insulin binds to insulin receptors, other molecules cannot.

4. The receptor has high **affinity** for its specific ligand. The ligand concentration increases around its specific receptor. (Just imagine a radio which specifically and selectively increases the concentration of those radio-waves for which it is tuned, around its antenna).

5. Another distinctively biological phenomenon is the **plasticity** of the receptor (the signal recipient). Signals are often able to modify their specific receptors. Frequently returning patterns of biological signals improve the reception of that particular signal. This is one of the biological bases of learning (of any kind of learning at any stage of biological evolution). Biological learning does not require consciousness or cognition. A particularly fascinating type of biological learning is the maturation of the immune system. Newborns have an insufficient immune response. However, they are exposed to a strange biological environment and their immune system has to 'learn' how to react to the millions of biological signals in their environment. The immune response takes several weeks to occur after receiving a biological signal (called an antigen in this context) for the first time. However, the

second time an antigen is encountered the immune response will occur within hours, because the organism has developed specific receptors (antibodies) to that particular antigen.

[The induction of antibodies by antigens suggests that a biological signal is able to create a recipient (receptor) and transfers the message into information. It is not the case. The sender-signal-recipient system is necessary but not sufficient to assume the existence of biological information exchange. The presence of a sender-signal-recipient system is a strong indication of the existence of biological information. However, some biological meaning and response are necessary to confirm the presence of biological information].

The development of immune memory is not about collecting 'memories' in a database. It concerns developing a receptor (antibody) for a specific biological signal (the antigen).

6. Signal compression or **efficiency** of message transfer does not appear to be a primary concern in biology; there is lavish redundancy. The entire genetic information is present in every single cell of an organism. The genetic code has an approximate 3-fold redundancy; 64 codons code for the 20 amino acids.

7. The **'signal' and 'noise' concepts** are very different in biological and physical information systems. A signal for one organism is noise for another. Genes (genetic messages), used to organize and make one organism functional, are often present in another organism but never used. The genetic material of higher organisms (such as mammals, including humans) is particularly remarkable in this respect. More than 95% of the human genome consists of DNA sequences (introns) which are never used. We know that some of these introns were used during an earlier stage of human evolution, perhaps thousands or millions of years ago. These sequences are still present and carefully preserved from one generation to another. They were called nonsense or 'junk DNA' not too long ago.

8. Safety of information transfer is a vital issue in living systems. Redundancy is one way to achieve this. However, in biological systems **'feedback'** is used. Signals continue to flow from the sender until they are recognized by their target cells, and the target cell sends back a 'receipt' to the sender cells to confirm the reception of the signal. It is a wonderful regulatory mechanism, most evident in endocrinology.

9. Molecules used for messaging/signaling in biological systems are often not the smallest possible packages that could represent certain biological information. The pharmacological industry has provided several drugs that perform better than physiological molecules despite being smaller and less complex. The stress hormone cortisone has numerous artificial relatives that are several hundred times more powerful.

Conclusions

The science concerning biological information is very young. DNA was discovered in 1953 (Wilkins, Crick and Watson), the genetic code in 1961 (Nirenberg and Matthaei), protein sequencing in 1951 (Sanger), effective nucleic acid sequencing in 1977 (Sanger); the first sequence database was published in 1965 (Dayhoff), and human genome sequencing was completed in 2003 (led by Collins and Venter). Bioinformatics, a new interdisciplinary science field, probably emerged in 1981 when Smith & Waterman described their fundamental equation for sequence similarity searches [18]. This field has become fruitful during the past 10 years or so. The first generation of bio-informatitians—

mostly young scientists from a biological or computational background—will have to face the hard reality that bioinformatics, if it is carried out properly, is not the 'wet variant' of the well-developed field of physical informatics and communication sciences. This is an entirely new way of thinking about information.

Bioinformatics is somewhat disadvantaged compared with physical or mathematical informatics. Scientists working with the latter categories usually have an idea of what are they working with and often know that the message in question (even if coded) is meaningful, *i.e.*, it is information. Bioinformatitians do not have this luxury. The human genome contains 3×10^9 bases (letters of a four letters alphabet) which are approximately 60 gigabits of 'information' (uncompressed). We have reason to suppose that this is important biological information as it has allowed the species to exist for at least 35,000 years and it is carefully preserved from generation to generation. However, we have to *learn* the language of life (Francis Collins calls it the *"Language of God"* [19]) and the biological meaning of DNA and protein sequences to understand biological communication. Understanding produces information from data and knowledge from information.

References and Notes

1. Hartley, R.V.L. Transmission of information. *Bell Syst. Tech. J.* **1928**, *7*, 535–563.
2. Shannon, C.E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423, 623–656.
3. Schneider, T.D. Information Is Not Entropy, Information Is Not Uncertainty, 2009. http://www.lmmb.ncifcrf.gov/~toms/information.is.not.uncertainty.html (Accessed on January 18, 2011).
4. NCBI: Sequence Analyses. http://www.ncbi.nlm.nih.gov/guide/sequence-analysis/. (Accessed on January 18, 2011).
5. HIV/AIDS Statistics and Surveillance, Centers for Disease Control and Prevention. Department of Health and Human Services: http://www.cdc.gov/hiv/topics/surveillance/resources/slides/mortality/flash/index.htm (Accessed on January 18, 2011).
6. U.S. Patent Statistics Chart Calendar Years 1963–2007. U.S. Patent and Trademark Office: http://www.uspto.gov/go/taf/us_stat.htm (Accessed on January 18, 2011).
7. American Cancer Society. Statistics for 2008: http://www.cancer.org/docroot/STT/STT_0.asp (Accessed on January 18, 2011).
8. OMIM Update List. OMIM, Online Mendelian Inheritance in Man, 2009. http://www.ncbi.nlm.nih.gov/Omim/dispupdates.html (Accessed on January 18, 2011).
9. General Information, PDB Statistics. Protein Data Bank: http://www.pdb.org/pdb/static.do?p=general_information/pdb_statistics/index.html& (Accessed on January 18, 2011).
10. Nirenberg, M.W.; Matthaei, J.H. The Dependence of Cell-Free Protein Synthesis in E. coli upon Naturally Occurring or Synthetic Polyribonucleotides. *Proc. Natl. Acad Sci. USA* **1961**, *47*, 1588–1602.
11. Strait, B.J.; Dewey, T.G. The Shannon Information Entropy of Protein Sequences. *Biophys. J.* **1996**, *71*, 148–155.
12. Ohno, S. So much "junk" DNA in our genome. *Brookhaven Symp. Biol*. **1972**, *23*, 366–370.
13. Sears, C.L. A dynamic partnership: Celebrating our gut flora. *Anaerobe* **2005**, *11*, 247–251.

14. Shannon, C.E. Prediction and entropy of printed English. *Bell Syst. Tech. J*. **1951**, *30*, 50–64.
15. NCBI: Databases. http://www.ncbi.nlm.nih.gov/Database/ (Accessed on January 18, 2011).
16. MedicineNet.com: Definition of Testicular feminization syndrome. http://www.medterms.com/script/main/art.asp?articlekey=14430 (Accessed on January 18, 2011).
17. Ogden, C.K.; Richards, I.A. *The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism*, 10th ed; Routledge & Kegan Paul: London, UK, 1949.
18. Smith, T.F.; Waterman, M.S. Identification of Common Molecular Subsequences. *J. Mol. Biol.* **1981**, *147*, 195–197.
19. Collins, F.S. *The Language of God: A Scientist Presents Evidence for Belief*; Free Press: New York, NY, USA, 2006.

© 2011 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/3.0/).