Chapter 4.1 The Internet versus Life

The premise of this book is that while the Internet and life are different in many respects, there are sufficient similarities in the way information is stored, transmitted, and used in the two systems that we can learn about both by comparison. We now reflect on the two systems and try to make that comparison.

Information formats

Internet

Data sent over the Internet and the networks feeding it are coded using formats that are applied sequentially, thus each format can be considered a layer. The highest layer is constructed by the application and the lowest layer is built by the device actually transmitting the digital data. Thus the lowest layer is picked to be efficient for the media being used to carry the data, i.e. phone line, coax cable, optical fiber. One of the functions of routers is to translate an incoming format to another needed for the next leg of the transmission.

The lowest layer may continually transmit pulses or a carrier signal, with data represented by variation in the timing of pulses, or intensity and frequency of a carrier signal. However, at the higher layers data is transmitted in relatively small packets, with no packets transmitted if there is no data to transmit. This frugality allows many users with modest average needs to share a high speed transmission line. Thus we have only a few optical fibers coming into a router instead of the thousands of copper wires which feed a telephone switching station.

When on the Internet each data packet must have an Internet layer which, among other parameters contains the addresses of the source and destination. It is of course the presence of the addresses that enable packets from many users to share one link. Higher layers either facilitate packet delivery, e.g. TCP, or are specific to the applications (programs) producing and using the data.

Of course all formats are described explicitly in manuals. It may be difficult to get software engineers to write readable descriptions, but documentation is so important that technical writers are often hired to do a proper job. The format parameters for a layer and error correction code are in either a header or footer or both, which are just before and after the data (where else would they be). There are occasionally small unused segments in the header or footer, saved for parameters that are not needed in a particular situation. However, there are not many bits which have no obvious function, i.e. junk bits.

Since the Internet is a packet system, the routers connecting the links are as important as the links themselves. It's thus of interest to consider the format of the internal data used to operate the router, to borrow a term used by biologists, data to carry out the housekeeping functions. Most of these data are now digital, or at least have a digital component. The instructions routers implement to process Internet data are of course digital. Monitors used to be cathode ray tubes, where analog values of voltages and currents determine where an electron beam hits the screen to produce light. However, most monitors are now solid state devices with discrete pixels, even

though the intensity of the light produced is basically analog. In many cases a router doesn't have a monitor, it is accessed as a web site over the Internet. Cooling fans are often under the control of a digital logic chip with input from a temperature sensor. Even the power supply has a digital component. The voltage across a capacitor is compared to a reference voltage; if it is lower than the reference and the line voltage is higher transistor is turned on to charge the capacitor, otherwise the transistor is off. Thus the household chores use digital and hybrid digital-analog based systems.

However, is the actual, intrinsic data that enters and exits the Internet digital or analog? In the situation where one instrument is sending digital data over the Internet to another piece of equipment you would say the real data is digital. But much of the traffic on the Internet is both directly generated and used by humans. To a human music is analog, still and video images are analog, and even text is analog because it is "read" as an image. However, on the Internet all data is digital.

Life

Genetic information, the data needed to build the organism, is stored as a digital four valued code in the DNA. In contrast to the Internet the values are not levels or intensities, but rather specific molecular structures, i.e. tokens. As with the Internet there are layers of formats. Some formats coded in DNA have nothing to do with genetic information, but specify the ends of the DNA molecules, or the locations along the long DNA chains where DNA replication starts, or the location where fibers attach to the DNA to pull copies into two daughters when the cell divides.

Copying the information into RNA requires many formats. On the first layer for each RNA there is a binding site for the enzyme complex that makes the RNA and the site where the enzyme stops and disassociates from the DNA. Once the RNA is made, there are sequences that direct enzymes to remove segments and make a shorter RNA. Then there are sites that direct the ribosome to bind and start making protein.

There are sites, usually near the N-terminal end of the protein, that specify that the protein is to be cut by a specific enzyme, with the small peptide discarded. Other sites near the end of the protein bind to other proteins that transport it to specific locations in the cell, e.g. the outer membrane.

The code used to translate sequential nucleotide triplets into the amino acids of proteins is almost, but not quite, universal. However, many of the codes for other steps in the process of translating DNA into protein and processing that protein vary significantly from one group of organisms to another.

Thus, genetic information is encoded in DNA as a series of layers of formats that can be seen as analogous to formats used on the Internet. However, these formats and codes are not described in manuals. In addition, the equivalents to headers and footers are often not immediately adjacent to the segments they refer to, e.g. transcription enhancers may be short segments thousands of base pairs upstream of the actual start of transcription. Finally, a code may closer to an algorithm than a literal nucleotide sequence, e.g. two sequences 20-24 nucleotides long, complementary to each other, separated by 8-16 nucleotides that contain at least 70 percent A or T. While the three nucleotide code that specifies amino acids is quite similar in different organisms, the other formats and codes are much more dissimilar.

Information transmission

Internet

The whole purpose of the Internet is to transmit information rapidly. The information is encoded as variations in an electromagnetic field which may represented by pulses of current confined to a wire, pulses of light confined to a glass fiber[1], or shifts in amplitude, frequency, or phase of a radio frequency transmission. However, in all cases the transmission has a defined speed; the distance traveled is proportional to time. The speed is within a factor of two the speed of light in a vacuum, 300,000 km/sec. Thus data could travel from San Francisco to New York, a distance of about 5000 km, in about 15 ms over a single link. However, in practice the transfer would use several links connected by routers and the trip would more likely take 150 ms or longer (this time is called the latency). The effective speed, is thus about 30,000 km/sec.

The amount of information that can be transmitted per unit time is dependent on the physical nature of the communication link and its support. Twisted pair telephone lines have a capacity of about 1 Mb/sec (SDL over a distance of a few miles), Ethernet a capacity of 10, 100, or 1000 Mb/sec, and optical fiber 50 Gb/sec (capacity of optical fiber is increasing so rapidly that this is just an estimate).

Information moving through the Internet is generated and received by computers and in between it passes through routers, which are just special purpose computers. The speed of a computer which is relaying data depends on what it has to do with the data. Relaying is perhaps a misleading term because the router has to read and decode each header just to know what to do with the data, and in many cases other manipulations are required. Computers can execute several billion instructions per second, and each instruction can process 64 bits. The high capacity optical fibers carry data on 10-100 different wavelengths, and a different router feeds each wavelength. Thus there is not a big mismatch between the fiber and the router. Routers have huge amounts of RAM, so most of the data transfers are fast. Information can be copied from one disk drive to another, i.e. replicated, at a rate of greater than 1 Gb/sec.

However, high data capacity is not the fundamental characteristic of the Internet. The credit for high capacity goes to the hardware engineers that designed the communication infrastructure used by all networks. The use of packets and the Internet protocol actually increases the latency compared to a dedicated circuit. However, the Internet allows a large number of users with modest information transfer requirements to efficiently share high capacity links. The resulting increase in latency is a small price for the efficiencies in scale that result from use of a packet network.

Life

Living organisms have no need to transfer genetic data rapidly over long distances. Living organisms do need to move, and as they move they carry their genetic information. Animals need to move to keep away from predators, to find mates, to migrate seasonally to find food, and to fill all available ecological niches that can support them and their children. The adult barnacles on a rock produce children that go

[1] Transmission using laser beams in free space between satellites is used in some scientific and military applications.

through a free swimming (or drifting) immature stage before metamorphosing into small adult forms that settle down on rocks to grow and produce more barnacles. But this kind of information transfer is like shipping new computers (loaded with operating systems) to new users. It's not the kind of information transfer that the Internet accomplishes as its main purpose.

Operational (as opposed to genetic) information is represented in the cell as concentrations of specific chemicals, and is thus analog in nature. Each chemical species can be considered to be analogous to a channel in a communication system; the chemical is produced at one location and detected at another. The transfer from one location to another is by diffusion, and thus the transit time increases as the square of the distance; there is not a constant speed. The reliance on diffusion for transport of all materials, oxygen and nutrients as well as chemical messengers, effectively limits the size of cells to a diameter of about 10 μ.

An animal needs to transmit information fairly rapidly in order to control movement of its limbs. The nervous system is the fast analog network for a human. Information can travel at a speed of 0.1 km/sec on nerves. This is more than 10^5 times slower than the speed on the Internet, however, a typical distance for communication in a human is 1 m, and the transit time is thus only 10 ms. A very fast runner can travel 100 m in 10 sec, thus the feet must move at a velocity of about 10 m/s, or 0.1 m in 10 ms. The speed of impulses on the neurons thus barely keep up with the very rapid movements of athletes (or people running from bears).

Information storage

Internet

An information transfer system doesn't necessarily need to store information. The original telephone system didn't; once the operator set up the circuit between callers they could talk and the information just flowed in both directions. However, since the Internet is a packets system they must be stored by the routers so the addresses can be read and decisions made on the best routers to send them on to. The "time to live" counter must be decremented, or if zero the packet is discarded. The format may need to be changed if the packet is being sent out on a different media. Finally, the packet often has to wait its turn when the output channel is occupied. In addition to storage for the actual Internet data, the router has an operating system and uses many applications to carry out its job. It generates routing tables for other routers and diagnostics for the technicians that keep the Internet working. All these tasks require memory, lots of memory.

Random access memory (RAM) stores information as electric charges in large arrays of capacitors; each capacitor contains one bit. One module of a modern chip can hold 2 GB (16 Gb), and has a volume of about 1000 mm^3 giving a density of 16 Mb/mm^3. However, a RAM chip requires supporting hardware to actually store and maintain the information. The volume of this hardware is a little difficult to pin down (no pun intended), but let's say it lowers the effective density of information by a factor of 2 to 4.

Long term storage of information is accomplished by creating patterns of magnetic domains on the surface of rotating disks. The actual volume of the magnetic domains is small because they are created in thin films only 30 nm thick at a surface density of

about 2 Gb/mm^2. However, the film must be held by a much thicker metal, glass, or ceramic disk. If this disk is 1 mm thick the volume density is 2 Gb/mm^3 much higher than a RAM chip. However, a motor must spin the disk at 5,000 to 10,000 rpm and the magnetic domains are created and read by a head that is positioned over specific locations on the disk. When all these supporting components are included, a disk drive that can contain 120 GB occupies a space of about 500 cm^3 for a density of 2 Mb/mm^3. Thus, the effective capacity per unit volume of a RAM and a hard disk is quite similar if you include the supporting machinery.

Life

Genetic information is stored as a nucleotide sequence in double stranded helical DNA molecules. However, the shape of these helixes is to a good approximation independent of nucleotide sequence. This feature is shared by the machines used by the Internet. On a microscopic scale the surface of a disk is a flat, smooth surface, only on an atomic scale are magnetic domains evident. For both DNA and the magnetic disk the information is encoded in a way that has a small effect on the gross physical structure of the media. This is important, because the basic operations done on material containing information need to be independent of the actual information encoded in the material. Machines that copy, delete, rearrange, or read information must be able to function independently of the information encoded in the media. The enzyme machine that replicates DNA depends on a double stranded helix with constant diameter and a repetitive phosphodiester backbone. The disk drive head that reads and writes the information on a magnetic disk requires a flat, smooth surface.

DNA molecules store 2 bits/base pair. However, let's consider the DNA plus histones plus assorted supporting enzymes, generously the volume of the entire nucleus, as the volume of genetic information. A generic human cell has a diameter of about 10 μ and the nucleus is at most about 1/3 rd the diameter, so its volume is about 20 μ^3. Much of the information stored in the DNA may not actually be used, but the machinery that processes the information treats all the nucleotides equally, so the total information in the nucleus is about 6 X 10^9 bits. Thus the storage density is about 300 X 10^9 Mb/mm^9 , or more than 10^9 more dense than computer memory. Computer hardware engineers drool at the thought of molecular memory. However, the actual memory is just the first step, engineers would need also need to develop molecular machines to read, write, and duplicate the molecular memory. That's not going to be an easy task, but a system analogous to DNA would seem to be a reasonable prototype.

The memory stored in a human brain might represent a high information density. However, since we can't quantify memories it's difficult to estimate the amount of information in them. We know so little about the format and representation of this information in the brain it is difficult to even estimate the density. It's possible that the storage density is modest, but it is organized in a very clever way.

Error correction

Internet

No real world communication system could operate without error correction, because even the best have error rates that would be intolerable for many applications. However, different types of data have different degrees of error tolerance.

audio: An occasional error may not even be noticed. Errors that will be noticed are sudden discontinuities in volume or frequency, and they can be recognized and removed by software at the receiving end of the transmission.

text: An occasional substitution of one letter for another might be annoying, but in fact most languages, including English, are quite redundant. Thus, it is unlikely that a slightly garbled text would not be read correctly. Of course, some mistakes could be extremely significant.

computer code: A single error will almost certainly cause the program to crash if the segment containing the error is executed. Reversal of a single bit will change an instruction or a memory address, and then the program goes off in a completely different direction, doing silly things like attempting to address memory that doesn't exist. It's a disaster.

Information transmitted on the Internet can pass through many types of links, some with high error rates and some with very low. Thus very robust software error correction must be available. The Internet is built with layers, and most of them have there own error correction mechanisms. On top of the IP layer, you have the option of more error correction. Some applications just transmit data and hope, while others use TCP, which ensures that the data is uncorrupted.

The more elaborate the error correction system is, the greater the overhead. Overhead at the sender is the time to transmit the additional data correction data, at the receiver it is the time required to execute the error correction algorithms. Thus it may be wise to only use the error correction that you need.

Life

Single stranded RNA viruses use little error correction when they replicate, and thus the frequency of errors is high compared to a DNA organism. However, since they are small, the total number of errors per generation is small. In addition, mutations that do occur can give them an advantage in an adverse environment, e.g. HIV replication in a human taking anti-viral drugs.

Organisms with double stranded DNA have much greater stability because DNA is chemically more stable than RNA and the two strands provide redundant information storage. In addition, they have more enzyme systems that repair damage. But again, complete accuracy would eliminate the creation of diversity which is essential to evolution.

Errors that do occur during replication can have an extremely variable effect.

non-coding regions: Errors here are not likely to have an effect.

promoters and enhancers: Errors here can decrease or increase expression levels, and thus can have a great effect.

deletion or insertion in coding region: This type of error will cause the following nucleotides to be read out of phase, producing non-sense. Within a few dozen codons a non-sense codon is likely to be formed causing termination of the protein chain.

substitution in coding region: Since the genetic code is redundant (particularly in the third position), it may have no effect. Even if a different amino acid is put into the

protein, it may not change the property of the protein significantly if it does not change the charge or size.

Important differences

If we think carefully about error correction techniques several rather fundamental differences between the Internet (or digital computer technology) and biology become apparent. For life all error correction is local, because there is no mechanism for implementing a more global strategy, such as a check sum. Living cells don't have a CPU and a scratch memory that can compute the sum of the nucleotides on a long segment (using some arbitrary numerical value for each nucleotide) and then store the sum as a small nucleotide sequence at the end of the segment. Instead, error correction relies on many small molecular machines that move along the DNA sensing errors.

All biological error correction relies on DNA containing two strands with the same information. During or immediately after DNA replication the error correction machines can again check to see that the new strand is a copy of the old. At a later date machines check to see if the nucleotides are indeed one of the set of four allowable ones. If the machines find junk they get rid of it and fill in the gap using the other strand as a template.

But looking for junk in information on the Internet would mean analyzing the wave form of each pulse and rejecting it if the profile didn't match the expected curve. Of course that would be completely contrary to the whole philosophy of a digital system: if the signal is above the threshold at the clock tick the data is 1, else it's 0. The simplicity of this rule is its power, it makes digital systems fast but dumb. Speed and simplicity outweigh lack of sophistication, or the designer should use an analog system.

However, since the DNA strand must contain a series of discrete nucleotides (or molecules that resemble nucleotides) you could argue that it is still a digital system but just not binary; each molecule represents a different signal level. The trouble with this argument is that there are a large number of molecular structures that are recognized as incorrect, and it is not really possible to say exactly how many incorrect structures there are. There may be only a small set of incorrect "nucleotides" that are commonly generated naturally, and that we have cataloged, but many more could exist. In addition, it is difficult to predict in advance which strange nucleotides will be recognized and removed and which will not. To complicate matters more if a strange nucleotide is incorporated and not removed it can alter the performance of the DNA as an information storage device even it normally forms "correct" base pairs. The thymine analog 5-bromo-uridine can be incorporated into the DNA of organisms grown in the laboratory. It escapes the error correction machines and indeed forms the "correct" hydrogen bonds with the adenine on the other strand; most of the time. However, the error rate of this nucleotide is higher than rate of the correct nucleotide.

This doesn't sound like a pure digital system to me but rather some kind of a pseudo-digital system.

Organization

Internet

Arguably the most innovative characteristic of the Internet is its decentralization. The sender only needs to know the first node to send packets to. The remainder of the trip is determined one router at a time. Each router builds its address tables from the routers it is connected to, its neighbors. This process allows new links to be added to the Internet without direction from any central authority. The domain name servers are also decentralized, except for the top level. However, even in principle the Internet does not promise to deliver a packet, it just makes its best effort. If you want to be sure the packet arrived you have to do that yourself, or use a standard application like TCP.

This informal approach doesn't appeal to everyone. If you want to deliver real time, high quality audio it might be better to pay for a dedicated circuit from you to your client. However, if the Internet infrastructure between sender and receiver is not loaded down, a small buffer that stores a 100 ms or so of data can smooth out the variable times for packet transmission. The cost of Internet transmission is so much lower than dedicated circuits that you can get used to a small time delay between transmissions, even for person to person conversations.

Life

The activities of a human are also mostly decentralization. Of course our conscious thoughts and perception of the external world appear to us to be centralized, although it is difficult to know what that really means. When we execute a complex physical task, e.g. soccer, the brain controls the aspects of movement we are conscious of, by definition. But much of the muscular coordination is not conscious, not centralized, and thus not at all obvious to the human footballer. That's what training is all about.

A large part of our brain handles "unconscious" information processing. These regions of the brain send general signals to many organs, controlling, for example, the average rate of heart activity and breathing. But there are neural loops between many muscles and the spinal cord that generate reflex activities that are the components of activity. The neural network within the heart produces the complex sequence of contractions that produces efficient pumping by coordinated action of the four chambers. The muscle contraction that force food down our intestines are not initiated by the brain. Some the hormone systems are triggered by the brain, while others are controlled by other organs.

The insulin system that controls blood glucose concentration is independent of the brain, although it is centralized, since the insulin concentration in the blood coordinates metabolism over the entire body.

If forces on skeletal bone are decreased, the calcium content decreases daily and thus the strength of the bone gradually decreases. This effect has been carefully documented in astronauts that orbit the earth in an environment without gravity. The demineralization of bone has also been observed in many disease situations. It is clear that bone structure is in equilibrium with the environment, and is a decentralized system.

If you increase the activity of muscles in one part of your body over many days, the muscle mass in that part of the body will increase. If activity is decreased, muscle mass decreases. This is a decentralized response.

The red blood cells have an effective life span of about 120 days. After this time they are destroyed in the spleen. Thus red blood cells must be constantly replaced by the daughters of stem cells in bone marrow. The division of these stem cells must be kept in balance with the concentration of red blood cells in the blood. The leukocyte cells that make up the immune system also have a finite life span, and thus must be constantly replaced.

When you cut your skin, the genes in the cells along the edges of the wound are activated to start cell division to make cells to fill in the gap. In addition, synthesis of collagen is increased. This fibrous protein acts as a glue to hold the wound together and provides a scaffold for the migration of new cells. This is a decentralized response.

If half of your liver is removed the remaining cells of the liver begin to grow and multiply and the lost liver tissue is replaced in a few months. This is a decentralized response. Unfortunately, if you loose one kidney, another kidney is not produced.

The constant turnover of cells in our body is normally not apparent, unless you have the experimental tools to detect it. Cancer, the inappropriate growth of cells in our body, is an unpleasant example of failure in the balance of cell growth and death. The more molecular and cellular biology you learn the more amazed you are that we don't all die of cancer at a young age.

It is during embryogenesis, the development of the fertilized egg into a complete human, that the most complex pattern of gene expression is seen. Most of this process is decentralized.

Turning off expression of some genes and turning on others is an analog process. However, the proteins and nucleic acids that make up the machine which turns genes on and off are specified by the digital information in DNA.

Evolution

Internet

The Internet was able to evolve rapidly because of the decision, made when its precursor ARPANet was built, to use a general purpose computer and software to format data and implement network protocols. In this way protocols could be quickly developed and tested until they worked. After the first system was functional, many improvements and additions were added to the software before it became the Internet. Improvements and extensions to the Internet continue. This evolution would have been slower and more restricted (if it had even succeeded) if it had been necessary to implement the changes in hardware.

Life

The evolution of life is also achieved by changing software, the sequence of nucleotides in DNA molecules that specifies the structure, and thus function, of RNA and protein molecules. The genetic information in DNA is changed by duplication, deletion and rearrangement of segments, as well as mutations of a single base. In order

for these changes to endure they must be made during the process of creating a new individual. Of course the DNA itself is hardware, just as computer memory itself is hardware. However, the pattern of nucleotides along a DNA chain is functionally software just as the pattern of magnetic domains on the surface of a magnetic disk is functionally software (here I don't distinguish software code from other data).

However, the DNA of an organism isn't just the program, it also specifies the instruction set; it's the blueprint for the computer. However, the potential for diversity has considerable limits at fundamental biochemical levels. All organisms have the same 20 amino acids and four basic nucleotides, with a few exceptions. The basic structure of the protein globin, which in humans holds the heme-iron-oxygen complex, is found in ancient bacteria. The fundamental machines of the cell, e.g. the ribosomes, depend on the cooperation of a large number of different molecules. Changing such a complicated, interdependent system is not trivial, or so we might infer from the fact that it changes only slightly and very slowly. Some systems that were thought to represent organization at a rather high level, and thus unique to more complex organisms, e.g. the homeobox system, are now realized to be more primitive. As we obtain the nucleotide sequences of more organisms we seem to find a greater universal framework within which evolution occurs. However, there are certainly many enzymes that can change rapidly, or be eliminated, or be created by modification of other enzymes.

The segments of DNA that we understand the most, the coding sequences, produce proteins and RNA molecules with structures that are very dependent on amino acid or nucleotide sequence. Like DNA, the sequence of amino acids or nucleotides represent information, however it is the three dimensional structure of these chemical machines that enable them to actually do something, e.g. convert A to B. These proteins and RNA molecules are thus analogous to computer instructions. However, in most cases they are conditional instructions, i.e. they execute dependent on the presence (concentration) of other molecules.

Life is immensely more diverse than computers because evolution occurs in parallel over time throughout the genome of the organism and is expressed as they reproduce. Imagine if every computer that was manufactured was given a slightly different operating system which contained a few random changes to a hybrid of two operating systems that had been used by older computers for at least one year, with the donor computers picked in proportion to speed and reliability. Of course any computer manufacturer that did this would not last very long because many of the operating systems would either not work at all, or would be very inferior to the original system. The fact that the average performance of all the working operating systems gradually improved would not balance the failure of several percent. However, it is known that about half of all fertilized human eggs and ten percent of humans which reach the embryo stage die, while the rate of obvious birth defects is about one percent. In the biological world it is cheap to make more organisms and it is essential to produce diversity.

Staying with the computer analogy for awhile, there is another problem with loading each new computer with a slightly different operating system. How do you know the computers will be able to run the collection of applications they are expected to run, e.g. MS Word, MS Excel, Adobe Photoshop? Will they be able to run the standard Internet applications that are necessary to communicate over the Internet, e.g. the TCP/IP stack? If this were a complete analogy to the biological world the hardware of each computer

would also be slightly different. How could you be sure that a "standard" hard drive could be used to replace a hard drive that had failed (we are assuming that the original components of the computer worked properly together). Would a standard CD disc be read correctly by all the new computers? Obviously our imagined computer world with constant evolution of each computer would be complete chaos.

The biological world is simplified by the fact that every organism doesn't need to communicate with all other organisms. The biological world isn't the Internet. In the biological world only members of the same species need to communicate with each other; that's the definition of a species. The species can generate variability while ensuring that most of the genetic information common to the species is preserved even when occasional progeny are produced that are defective. The interchange of information between members of a species is the mechanism that maintains similarity between members of the species.

In fact it is useful for members of a species to maintain informational isolation from members of other species, otherwise the genetic information of the species would be constantly diluted. There are many mechanisms for limiting exchange of genetic information between species. They may live in different environments, have physical or behavioral barriers to mating, have different chromosomal organization, etc.

Epistemology

Internet

Since the Internet was designed by humans, almost by definition it was designed for a purpose. This purpose was to transmit information from one user to one or more of a very large group of other users. The strategy of the Internet is to transmit the information in packets over a network, using a decentralized, locally defined routing system. The goals of Internet design and the procedures, methods, and protocols intended to achieve those goals are defined and freely available.

The actual performance of the Internet is a completely different matter. The network is complex enough that while the effectiveness of various routing and scheduling algorithms can be predicted, only experimentation tells us how the Internet actually performs. Thus, collection of data on actual performance is essential to design the changes that will make the Internet more effective. The design of the Internet is an experimental process; it evolves by selecting the changes that work the best.

The actual physical structure of the Internet at any given time, the network of links and routers, is unknown. In principle the structure could be determined, but since it is not specified by any central router or other device, structure, like performance, can only be determined by experimentation. However, the Internet is so large and is changing so rapidly that it is effectively impossible to determine its structure. In the time that it would take to send messages to all addresses the Internet would change significantly. Even Google, with it's thousands of powerful computers searching the World Wide Web, which is only part of the Internet, takes days to look at every www site.

Life

Life was not designed by humans and thus there are no manuals. To understand life requires observation and inductive reasoning.

The study of naturally occurring mutants and variants has revealed function of many genes. We now have the ability to delete specific genes from genomes. However, since there are about 30,000 genes in a mammal, such as a mouse or human, it will take a lot of time to sift through them all. In addition, knowing that a gene product is essential for a function does not mean that you know what it does.

Many proteins interact with other proteins, and the protein complexes have properties that the individual subunits lack. Thus eliminating or altering one protein can cause effects that are entangled with the functions of other proteins. A function of a protein can be important only at a specific age or under a specific challenge. How can you possibly know all possible situations to study in order to observe all possible responses of the organism to that situation. You can't.

Thus, it is perhaps amazing that we know as much about the operation of living organisms as we do. The rate of accumulation of knowledge is increasing, but there is so much more to learn. The more philosophically conservative scientists (of which I am one) believe that there is some sort of objective reality in scientific models of the world. But the process of obtaining useful models is certainly subjective. All fields of science are at least partially an art, but I believe biology is the most artistic, or individualistic. In the end though, it doesn't seem reasonable, or perhaps even meaningful, to believe we can understand life, or even one organism completely.

Biology is different from other sciences

The study of life, biology, is quite different from physics and chemistry[2]. The animals and plants biologists study today have a history, they have ancestors. The current menagerie is just a snapshot in time, and its composition only makes real sense as an extension to yesterdays life. The same principal applies to the affairs of humans, which is why the study of history is so important to anyone who tries to understand why humans are doing the things they are.

History isn't as important to physics and chemistry. These fields are defined by basic principals and laws, e.g. conservation of energy and momentum. Basic laws don't evolve, they existed in the past, exist now, and will exist forever. Of course the laws of physics define the history of physical systems. They allow you to predict the past and future position and velocity of a clock pendulum given the present position and velocity. But the history or structure of a clock pendulum is not a very significant part of physics, it is more an example of something that can be better understood by use of physics. The objects that are important in physics are the fundamental particles that matter and energy are composed of, e.g. electrons, protons, quarks, photons. But there are only dozens of these fundamental objects, not the millions of animals, plants, and bacteria that biologists study. Two electrons are the same, except for their position and energy. No two living organisms are really the same, even identical twins.

Of course the objects biologists study also obey the laws of chemistry and physics. In this book, for example, we have described the physics of diffusion because it gives us some insight into the constraints living systems must obey. However, these laws are

[2] Many of these ideas are derived from essays by Ernst Mayr, published as the collection: "What Makes Biology Unique", Cambridge University Press, 2004. Similar ideas are presented by Francis Crick in his autobiography "What Mad Pursuit", Basic Books, Inc., 1988.

merely constraints, they don't allow you to predict what an animal looks like, how many species of insects there are, or why copepods have one eye.

However, certainly at the biochemical level there is a suspicion that some characteristics of life have a physiochemical rational, even if others are just historical, i.e. random. Perhaps there a reason amino acids have the L configuration about the alpha carbon? Maybe, but there is no theory or experimental evidence that would predict this configuration. Amino acids with the D configuration are made by bacteria, and incorporated into peptides, even though the mechanism for the synthesis of these peptides is completely different from the t-RNA and ribosome machine that makes the L amino acid containing proteins. Why are most proteins made from 20 amino acids? We can guess that if life started all over again, and again it evolved to a protein based community, there would be at least several different amino acids and some would be hydrophilic while others were hydrophobic. But would there be 20 amino acids, and would leucine be one of the amino acids in this version of life? Probably not, our life is partially a historical accident. The nucleic acids, especially RNA, are thought to have been made by life before proteins. There is research that strongly suggests that only a pentose (a five carbon sugar) can be used to make a nucleic acid, and that of the many pentose isomers, ribose produces the most stable RNA chains. So the basic chemical structure of the units that make up living organisms are certainly not completely random, but which parameters are predetermined and which are historical accidents is certainly an area of research and debate.

To be inclusive there are at least two other sciences that have an historical component, geology and astronomy. While chemistry and physics play an important role in geology, the shape of the coast of California can only be understood by combining physics with history. The size of our sun is an historical fact, however the laws physics give an upper limit to its size (larger objects collapse to become neutron stars). The evolution of planets, stars, and galaxies from an initially more amorphous compact universe is an essential part of the story of astronomy, but while the general path of the evolution of the universe might be predicted the details can not.

Living organisms and computers have history

Living organisms are different from most inanimate objects in that they contain structures, large molecules, that are similar only because of history. As an example, all humans (several billion) are different, but more than 99 percent of humans contain about 10^{20} identical molecules of an identical hemoglobin. This identity is due to the identity of the amino acid sequence; it is not the result of a physio-chemical law. The amino acid sequence is identical because these humans have the same nucleotide sequence in the globin genes because it was copied from the nucleotide sequence of ancestors.

The only collection of objects on the earth that have a similar property are the millions of computers which have one of a few variants of a pattern of magnetic spots on disc drives: the code for versions of the Microsoft operating system. These patterns are also not identical because of a physio-chemical law, but because they were copied from the pattern of magnetic spots on other discs, their ancestors.

History, the Internet, and life.

To understand either the Internet or Life it is essential to know a little information theory, chemistry, and physics. Why else would I write this book and why would you read it. However, the laws of chemistry and physics don't dictate the existence of either the Internet or life. There must also be a big dose of arbitrary historical "accident". The Internet might never have been created, or its structure could have been quite different. If a thousand earths were created in parallel who knows what life forms would have developed on those earths and when.

What is a living organism?

The following section is not organized as a comparison between the Internet and life, but examines special questions and problems that humans have defining life. Answers to some of these questions affect political choices, legal decisions and define our moral framework. Even though the subject is biological, exposure to the concepts of information and its application to the Internet can provide a background to the discussion.

Historical definitions of life

Before we examine serious real world questions, let's just ask what is life: how do we know an object is alive? Living things move, they grow, and they can reproduce. But is there some more basic, structural characteristic of life that enables living organisms to move, grow, and reproduce? A seminal characteristic which can be studied to gain insight into the process of living.

It was once thought that living organisms contained a life, or vital force, which distinguished them from non-living objects. The trouble with this concept was the lack of any functional definition of the vital force, at least one which didn't rely on the definition of life itself. Thus the term vital force was eventually seen to be just a synonym for life, without adding anything to our understanding of life.

It was proposed that living organisms were built from unique, organic, super-molecules. These molecules had special properties that the ordinary chemicals made by chemists in their laboratories did not possess. However, it became progressively apparent that chemists could make any molecular species found in living organism. If the individual chemical species are not special, the collection of different molecules that make up a living thing must be special. However, was found that some viruses, complex objects that had been assumed to be living, could be crystallized. Since chemists had long used crystallization as the criteria for a defined chemical species, they were forced to conclude that an entire living organism was a defined chemical.

Even in the 1950s some molecular biologists thought the study of life might reveal new laws of physics or chemistry. However, as the reductionist program of molecular biology relentlessly exposed one process after the other as resulting from application of the laws of "ordinary chemistry and physics" mysticism became a scarce commodity in biological labs.

The modern definition

Today the accepted definition of a living organism is functional. A living organism is an entity that can reproduce itself, with the caveat that this reproduction must be associated with the possibility of evolution. However, we now also believe that replication is possible only if the object contains DNA (or RNA if we agree that a virus is alive). Thus an object can only be alive if it contains DNA which encodes its genetic information. Information is the essence of life.

If a virus then what?

The uncertainty in this definition is the lack of a specification of the environment in which the replication can occur. If replication can take place inside another organism, in particular inside a cell, a virus is alive. Even though a virus depends on a host cell, it goes through a stage in which it is physically distinct from that host cell. Perhaps this extra-cellular stage makes a virus seem separate from its host, and thus encourages the assignment of the virus as a distinct, living, organism.

However, we can move further along a slippery slope by considering genetic objects that do not have an extra-cellular existence. A transposon has a certain amount of independent existence in the genome. It replicates along with the host genome, but in addition copies can integrate into other sites in the genome. Thus it has some independence in its pattern of replication. Is a transposon alive? If it is, what about a gene?

What is a human?

More specifically, what is a live human? The answer to this question has great practical, legal, and moral consequences because in most of the world killing a human is prohibited unless it is done in immediate self defense (with the exception of some countries which still execute criminals).

Application of the logic of the previous section defines a living human as a being that can replicate. But of course no one would say that a woman that was past menopause or a man that had a vasectomy was not alive. Nor would anyone claim that someone who, because of a congenital defect was sterile, was not alive. If all the DNA in a human was destroyed (this is an approximation of the effect of a lethal dose of radiation) they would still live for a few days, and if during this time they were killed no one would say they were not murdered.

A human can think

No, in human society life is defined by the ability to think like a human. A severely injured or ill person is considered dead when they are “brain dead”, which means that they can not respond to stimuli and electrical activity in parts of the brain that are associated with cognitive activity have ceased. Thus life is still associated with information, but neural information processing not genetic information.

This definition for human life is important in considering the morality of abortion. Those who believe abortion is murder assert that an embryo even before the first trimester of development is a living human. But if we identify a human as a being with the ability to think, an embryo becomes a human when it can think. While development

from fertilized egg to the infant at birth is a continuous process, most scientists and physicians believe complex neural activity starts between 12 to 20 weeks after fertilization[3]. The morality of abortion is an extremely emotional and contentious issue and volumes have been written on both sides. I have only attempted to suggest the major issues in these paragraphs.

Can other animals think?

The genome of a human and a chimpanzee is very similar. Thus, the proteins made by human and chimpanzee cells are similar. Chimpanzees can think, but how close to human thought is chimp thought? Since chimpanzees don't talk or write we have to access their intellectual ability by observation (if they did publish I'm sure there would be at least as much disagreement on the value of their literature is there is controversy in reviews of human literature). Chimpanzees have been taught many complex skills while in captivity. In the wild they also do many clever things to "earn a living". As one example, they use special sticks to probe ant hills and then remove the sticks and lick off the ants stuck to the sticks. They make different sticks for different situations (sharp ends, fuzzy ends) and use them in appropriate ways.

I claim that observations which suggest that chimpanzees can think in ways similar to humans affect the similarity humans see between chimpanzees and themselves far more that any statistical comparison between the human and chimpanzee genome.

The potential to be a human

One argument used by those who believe abortion is murder is the assertion that while an early embryo, or even a fertilized egg, is not a functional human, it is a potential human. If left to it's own devices it will become a fully functional human, and thus deserves all the respect and protection a fully functional human commands. The fertilized egg has this property, we now know, because it contains the DNA genome which specifies all the protein and RNA molecules that will enable it to become an adult human. However, a weakness in this argument is that almost every cell in our body has a complete genome and could, in principal, develop into an adult human. As we have learned from the earlier chapters, there is no fundamental reason a somatic cell, say a cell from the epithelium of our mouth, could not develop into an individual. This individual would have the identical genetic information as we did, and would thus be a clone.

At the time of writing this paragraph no human had been cloned. However, sheep have been cloned. Nuclei of cells taken from the mammary gland of a sheep were injected into enucleated sheep egg cells, and then implanted in the uterus of a sheep. After the normal gestation time live sheep, clones of the parent were born. Thus, all the cells in the mammary gland of a sheep are potentially a new sheep.

[3] See for example "Developmental Biology" 7th ed., by Scott F. Gilbert, 2003, Sinauer

Chapter summary

Information formats: Information moving over the Internet and contained in living organisms is organized using formats, and in both systems the formats form layers.

Information transmission: Information carried over the Internet travels close to the speed of light, 3×10^8 m/sec, on each link, but the routers that connect the links introduce considerable delays.

Living organisms move genetic information very slowly. The most rapid information transmission within an animal is analog data carried by nerves; even then the velocity is less than 10^3 m/sec.

Information storage: The Internet requires computers that store information as patterns of electric charges or magnetic patterns. These devices have an information density of several Mb/mm^3.

Living organisms store genetic information as molecular patterns with a density of greater than 10^8 Mb/mm^3.

Error correction: The prototype for the Internet is a global check sum which is corrected by a retransmission. The error rate after all error corrections are implemented is essentially zero.

Living organisms also have many error correction systems but they all work locally by either detecting non-complementarity between the two strands of DNA or non-allowed nucleotides. The latter method reveals that the information is encoded in an analog-digital hybrid system. Error correction never reaches the low level seen on the Internet, and a moderate error rate is required to enable living systems to evolve.

Organization: On the Internet packet routing using IP numbers is completely decentralized. Translation of URLs into IP numbers is also decentralized except for the top level domain name servers.

The selective expression of genetic information stored in the genome is determined locally on a cell by cell basis. Cells in a multicellular organism are coordinated by hormones and the nervous system.

Evolution: The Internet was novel in its time for the extent it was implemented in software. This allowed formats and protocols to be more easily developed in an evolutionary way as the properties of the ARPAnet and then the Internet were discovered. The Internet evolved over a few decades.

There is no way for life to form than by evolution. New species are constantly created by evolution of existing species, sometimes in an obvious response to selection by changing environments and in other cases to fill an empty ecological niche. Mutation, recombination and gene duplication are the major mechanisms of evolution. Life

evolved over several billion years but the evolution of some organisms can be studied in the laboratory in a period of months or years.

Epistemology: The Internet was designed by humans and their goals are documented. All the formats and protocols are explicit and open, i.e. publicly available. That doesn't mean the design exactly satisfies the goals; engineering is ultimately an experimental science.

The goal of life is to survive, or more accurately the life we see is the life that survives. The structure and behavior of living organisms has to be determined experimentally. We have to use inductive reasoning to link the structures and functions we find to the needs that the organism appears to have. While our understanding of life has increased immensely in the last few decades there are many basic questions that remain. In addition, in a fairly fundamental way, it will be impossible to be sure that any aspect of life is fully understood. Evolution is not perfect in the sense that any given organism does not necessarily respond to all challenges in the best possible way. Evolution is a mixture of partially random responses to selective pressures and constraints due to responses made in the past to other pressures.