Chapter 3.5 History of Life and Biology

Introduction

In contrast to our study of the creation of the Internet, we are now interested in two distinct processes: the development of life on earth and our attempt to understand that life. Life has evolved over several billion years while our understanding of life at the genetic and chemical level started a little over a hundred years ago.

History of life

Fossils represent an extensive history of the development of life on earth. Of course there are many gaps, and we don't know how much of the history is missing. The spectacular fossils are remains of the larger plants and animals, the ones that had fairly durable frameworks: bones, shells, stems, and trunks. But the fundamental molecular machines of life, the ones that enable information storage and transfer, had already evolved in the bacteria. Unfortunately bacteria make poor fossils, and more importantly the molecular features that constitute the information system are not preserved. Thus, we have to make many guesses. Working backward in time we think we know some of the important events but need to guess about many others.

It is certainly true that early life evolved in an atmosphere almost devoid of oxygen. In fact, it is the metabolism of living organisms that has produced the oxygen in our atmosphere. As mentioned before, the precursors of the hemoglobin in our blood were actually proteins that protected early bacteria from toxic effects of oxygen.

The RNA world

It has only been a few decades now since we realized that early life was most probably based on RNA; protein and DNA were subsequent improvements. We know that RNA is quite capable of carrying information, it's just not as stable as DNA and the replication system is more error prone. However, that is not a big problem for small genomes. RNA can catalyze chemical reactions, so a purely RNA organism should in principle be able to replicate and carry out metabolic reactions to generate energy and produce the substrates needed for that replication. We haven't found an organism that is entirely based on RNA, except for some viruses, and they use the host machinery to carry out most of their needed processes. Entirely RNA based organisms are unfortunately extinct (unfortunate for the biologist), having been replaced by those using protein and DNA.

In addition to developing an RNA based chemistry a critical step must have been the development of a cell membrane. A real organism has to be able to contain the chemicals it makes. Without a membrane there is no individually, it's just a soup, and evolution must have proceeded very slowly indeed.

Making organic molecules

The initial hurdle might seem to be generation of the complicated molecules needed to run even a simple organism. Even if early life needed only RNA to store genetic information and catalyze reactions, how did it get started, since it takes many enzymes

to make even the four nucleotide precursors of RNA from simple molecules. However, in specialized environments an amazing assortment of complicated organic molecules are made, including amino acids and nucleoside bases. The Russian biochemist Oparin suggested in the 1920s that complex organic compounds were formed spontaneously (at least without life) in the reducing atmosphere of early earth. In 1953 Miller passed an electric arc through a mixture of methane, ammonia, hydrogen, and water and found that a complex mixture of amino acids and organic acids was made. Later experiments by other workers demonstrated that molecules as complex as nucleosides could be formed in similar systems. It was assumed that the synthesis of the molecules that became primitive organism occurred on the surface of the primitive earth in liquid pools exposed to sunlight and electrical discharges from the atmosphere.

However, spectrographic analysis of comets and other cool areas of space (comets are a mixture of frozen water and minerals) shows that fairly complex organic molecules, including amino acids, are present. The energy to form these molecular species comes from the intense ultraviolet radiation in space, while the low temperatures of these environments presumably prevent their degradation. It thus seem at least possible that the complex molecules that formed early life arrived on the surface of the earth on meteors or similar cosmic debris.

A more likely environment for the creation of life seems to be the immediate region around thermal vents on the surface of the ocean floor. These vents, only discovered in 1977, are concentrated along the boundaries between the huge crustal plates of the earth's surface that float on the molten interior. One of the largest and most studied boundary is the Mid-Oceanic Ridge, in the North Atlantic half way between America and Europe and in the South Atlantic half way between South America and Africa. Water rich in minerals and hydrogen sulfide gushes from these vents at a temperatures several hundred degrees (Centigrade), but since the vents are at a depth of a mile or more, the high pressure prevents the water from boiling. There is a high temperature gradient around the vent, so there is an environment for everyone, and a rich community of specialized organisms lives there. Many of these organism reduce the hydrogen sulfide to sulfur to obtain the energy to form organic molecules. Heat, minerals, water, and a compound that can be reduced; what more do you need?

In summary, there are several scenarios for the production of complex organic molecules, and a general consensus that once you could make RNA you were alive. The steps in between are still rather fuzzy.

Chronology of life

The earth is about 4.5 billion years old, but for the first few hundred million years it must have been fluid, since the oldest rocks are 3.8 billion years old. The first fossils believed to represent bacteria are 3.5 billion years old, so life seems to have taken "only" 300 million years to accomplish the steps proposed above. Oxygen, produced by the original organisms, was a significant fraction of our atmosphere by 2.2. billion years ago and life began to be complicated with the evolution of the eukaryotes 400 million years later. Comparison of the sequences of genes allows us to make guesses about evolution back to this time, before significant fossils were created. Figure LifeTimeLine in chapter 3.3 presents a logarithmic plot of the history of life on earth.

There is a great deal known about the evolution of life once organisms were of substantial size, 0.1 mm or larger, and contained silicon or calcium shells and bones. However, by that time the basic molecular machinery we have discussed previously was operational, and thus information storage and transfer occurred mainly as it is today. Thus, while paleontology is a fascinating science, and does tell us something about how species are created and change, it is unfortunately not very relevant to our main interests here.

History of biology

A central theme in modern biology, the study of life, is that there is nothing special about life. By that I mean that life is no more special than a sunset, a hurricane, the sun, a galaxy, or the hydrogen atom. Nothing special means that the laws of physics and chemistry are sufficient to explain life. There are of course many people today that have philosophies or religions that posit a special position to life; a belief in vitalism. Only a few decades ago many respected scientists, while believing that life could be rationally studied and understood, still suspected that new laws of physics, or at least new interpretations of the established laws, would be needed to understand life. However, one by one, the activities of living organisms have been explained using "ordinary" chemistry and physics. We don't know everything by a long shot, but it looks like we could get very close with enough work. Perhaps the last really deep mystery is consciousness; how do we know we are we? It is less mysterious that some of the best molecular biologists have migrated into this field of study.

Biology can be divided into disciplines, of which three major ones are: the observation and classification of organisms into groups (systematics), the chemistry of life (biochemistry), and the inheritance of traits (genetics). Of course there are many other fields which represent the study of living organisms, each from its own perspective, e.g. anatomy, biophysics, embryology, immunology, molecular biology, physiology. The discovery of the structure of DNA in 1953 caused a collapse in the separation between genetics and biochemistry. Academia still requires the study of biology to be divided into departments, but the names have changed and the separation is much less absolute. Biochemistry has always been a branch of chemistry, despite academic rivalry. Physics started to encroach into chemistry with the study of the behavior of electrons in crystals, solid state physics, so important to the commercial rise of semi-conductors. There were always physical chemists in departments of chemistry. As biology became more molecular it became more interesting to physicists. Today almost anyone can work on and talk about biological problems. Of course, that is one of my justifications for this book.

In the following description of the evolution of biology I emphasize the fields of biochemistry and genetics, because they provide the foundations for our understanding of the digital information that defines most of our structure when we are born. While one chapter of this book describes the analog information used to keep an animal operational, I feel that a history of all the work that underpins that area is far too ambitious for this modest chapter.

Systematics and evolution

The scientific, or at least systematic, beginning of biology has been associated with the Swedish naturalist Carolus Linnaeus, who became a professor in Uppsala in 1741. Linnaeus realized that all would be (or already was) chaos unless everyone used the same name for an organism. Since Latin was the language understood by all learned scholars, scientific names for living organisms would naturally be in Latin. Linnaeus also realized the usefulness of grouping related organisms together, and proposed that the name would contain both the group name or Genus (capitalized) as well as the name for the individual, the species. Thus, Mytilus edulis is the bay mussel described by Linnaeus in 1758, while Mytilus californianus is a closely related mussel found in California and described by Conrad in 1837. In the time of Linnaeus evolution was certainly not the organizing principal of biology, but yet, it was clear that living organism could be identified as members of a set of larger and larger related groups.

The 19th century was a period of intense interest in biology by professionals and amateurs alike. Many voyages of exploration and survey collected and described new animals and plants, and brought specimens back to Europe. This is how Charles Darwin started his scientific career, as a biologist on a two year cruise aboard the English ship HMS Beagle. He collected many species of animals and plants which he sent back to England. While on the voyage he repeatedly noticed relations between groups of animals that suggested the evolution of some species from others. As but one example he observed that species found on islands were very similar to those found on the nearby mainland. He concluded that the islands were formed relatively recently and were then populated by a few animals that swam, flew, or floated from the mainland. Over thousands or millions of years he supposed that the island dwellers gradually changed to better complete and survive in their new environment. And if island animals were formed by a process of evolution, Darwin reasoned that all animals and plants could have been generated by evolution. Note that the evolution of life on earth is often intertwined with the evolution of the earth itself, i.e. geology. Indeed, a model for the evolution of the surface of the earth is necessary to relate the location of fossils to the evolution of the life these fossils represent. Darwin certainly wasn't the first to propose that living organisms evolved, however, his book Origin of Species, published in 1859, contained such a methodical exposition of arguments for evolution and careful refutation of possible objections to it, that he is considered to be the father of evolution.

Biochemistry- small molecules

Biochemistry is essentially a bottom up approach. It proposes that if you want to understand life you need to know the chemistry, the molecular structure, of living organisms. It doesn't matter much where you start, particularly in the beginning, since not much is known and you are going to have to study all molecules eventually. Thus there doesn't really need to be any grand strategy or hope of understanding the fundamental questions of life. There are a large number of interesting and medically important insights that can be obtained before you come close to a fundamental understanding of life, if that is even possible, or if a majority of academics could agree what would constitute understanding.

A scientist in any field is limited by the available instruments and methods. Science is also fundamentally incremental, each research project builds on the present knowledge, so the early biochemists studied small, simple molecules.

At the beginning of the 19th century it was thought that molecules unique to living organisms, containing carbon, could only be made by living organisms (and not by chemists in the laboratory). In this view there is a fundamental distinction between organic molecules made by a hypothetical vital force of life, and other molecules. However, in 1828 Wohler synthesized urea, then considered uniquely organic, from non-organic cyanic acid and ammonia. After making urea there was no reason to believe the other organic molecules couldn't be synthesized and manipulated in the laboratory. Synthesis of other molecules was important because the structure of a compound was typically inferred from the individual steps required to synthesize it. In the ensuing years more and more chemical species found in living organisms were synthesized in the laboratory. Thus we know there is nothing special about the molecules in living cells, except perhaps the fact that all the larger ones contain carbon[1]. Even today the study and synthesis of molecules containing carbon is called organic chemistry.

Over the next hundred years the structures of most of the smaller molecules found in living cells were determined. DNA was isolated and partially characterized by Miescher in 1869. Of course he didn't know how large the molecules were or the helical 3-D structure. But he did know they were found in the nucleus, had sugars, nitrogen rings, and acid phosphate groups; they were nucleic acids. Emil Fischer was one of the giants of this phase of biochemistry, receiving the Nobel in 1907 for determination of the structures of sugars and amino acids. The sugars are a real bear; so many similar structures with slightly different numbers of hydroxyl groups pointing in different directions.

Biochemistry- enzymes

Once you have seen the zoo of small molecules in living organisms you ask how they are created and destroyed. The environment in the living cell is aqueous and the temperature low, while the chemist uses organic solvents and high temperatures to synthesize molecules. The original answer was that only the vital force of life can perform biochemical reactions. Thus it was important for Bucher to show that sugar could be converted, or fermented, to alcohol by a cell free extract of yeast, and he was awarded the Nobel Prize for this work in 1907. But, as we have learned, enzymes are

[1] Thus it is not rational to purchase expensive tablets of vitamin C purified from rose hips if "non–organic" vitamin C tablets are less expensive. Unless perhaps you believe the "non-organic" vitamin C has not been sufficiently purified and thus might contain toxic byproducts of the synthesis. However, it can be a different logic to for you to want to get vitamin C by eating an orange. The vitamins have been defined by demonstration of an absolute requirement. If you get absolutely no vitamin C you will eventually die. However, there are no doubt many vitamin-like molecular species that are beneficial to health, but are not absolutely essential. These are difficult, if not impossible to identify experimentally. It thus does not seem irrational to think that a varied diet containing the foods that humans, and similar animals, probably ate during their evolution, would be a healthy diet.

very large, thus determining the structure of enzymes is at least an order of magnitude more difficult than the determination of structures of the amino acids and sugars.

The first step in characterizing an enzyme was to purify it so you can do some real chemistry. The next important accomplishment was obtaining crystals of the enzyme. The basic idea is to produce a very concentrated solution of the pure protein, typically in a salt solution which reduces the solubility. Some enzymes are easy to crystallize, others are difficult, and it's not easy to predict which it's going to be before the crystals form. In 1926 Summer crystallized urease, an enzyme that hydrolyzes urea to ammonia and CO_2 , and in 1929 Northrop crystallized pepsin, an enzyme that hydrolyzes proteins to produce smaller peptides. Summer in particular was instrumental in proving beyond any doubt that the urease enzyme was a protein. In 1946 they received the Nobel Prize for this work. Crystallization was of great pedagogical importance because it was the gold standard for purity in the discipline of organic chemistry. The fact that these enzymes could be crystallized again put them in the category of "ordinary" chemicals and dealt still another blow to vitalism.

After an enzyme has been purified and shown to be a protein, i.e. a polypeptide, the chemist has to determine the amino acid sequence to be able to write down a chemical structure. This was, and still is, difficult, for proteins of even modest size. There were many biochemists that made contributions to amino acid sequence analysis. However, the final, most powerful, and perhaps most elegant method was developed by Sanger. His technique consisted of a series of chemical reactions that sequentially removed and identified amino acids from the amino terminal end of the protein chain. The method was first applied to insulin, with his laboratory obtaining the complete sequence in 1955. Now all biochemists were forced to admit that proteins had a definite chemical structure, just like other molecules; many had previously thought the amino acid sequences were random! Sanger was awarded the Nobel Prize in 1958. Soon biochemists all over the world were applying Sanger's method to other proteins. The genetic information database was under construction!

To know how enzymes work we need to know the complete 3-D structure. However, Pauling accomplished a key intermediate goal by predicting a common structural motif that represents a significant portion of the structure of most proteins. Pauling had been working for many years on the precise structure and electronic nature of covalent and hydrogen bonds. His theory of resonance proposed that bonds were often hybrids of several states, e.g. a bond could be the time average of a single and double bond. As part of this program of research he had collected precise distances and angles of the bonds in amino acids and peptides, using the technique of x-ray diffraction. This data allowed him to predict, using wooden models of the bonded atoms, that chains of amino acids would form a helix in which the peptide amino group formed a hydrogen bond with the carboxyl group in the next twist of the helix. Pauling and Corey described this alpha helix in 1950, and were awarded the Nobel Prize in 1954.

To determine the structure of a large protein molecule the x-ray diffraction method needed to be augmented. The basic problem is that the pattern and intensity of the diffracted spots on the film does not uniquely determine the 3-D structure of the molecule. No amount of fancy mathematics can be used to find the structure, there just isn't enough information in the spots. The mathematical way to describe this problem is that the spots only give the intensities of the reflections, but both intensities and

phases are required to calculate the structure (phases are the amounts the waves are retarded or advanced when they bounce off the crystal). In previous studies of small molecules, the known covalent structure and the symmetry of the crystal restricted the number and nature of possible 3-D structures. With a great deal of work and a little luck, the x-ray diffraction spots could be uniquely associated with distances in the crystal and the molecular structure solved. However, even small proteins have too many possible configurations for this method to succeed.

In 1953 Perutz invented the isomorphic replacement method. Protein crystals were prepared in solutions of heavy metals, and with luck sometimes the metals bound only to a few specific positions in the protein, but the overall structure of the protein was not changed. The x-ray diffraction patterns of the heavy metal derivatives, compared with the original patterns, allowed the phases to be determined and thus the structure to be obtained. In 1957, using this method Kendrew was able to determine the structure of myoglobin, and in 1960 the structure of hemoglobin was determined by Perutz. They were given the Nobel Prize for this work in 1962.

Genetics

Genetics is the study of the patterns of inheritance of parental traits by progeny. This field of biology yielded a great deal of information many decades before the molecular basis for inheritance was known. The existence of inheritance was of course realized before the Greeks and Romans. For thousands of years animal and plant farmers selected and bred parents to produce progeny that possessed desirable traits. However, three important advances in our understanding of and expectations for genetics were made in the 19^{th} century.

As mentioned previously, Darwin published the book “Origin of Species” in 1859. The first thesis of the book was that the characteristics of an animal or plant species did not necessarily remain constant over long periods of time, the species could evolve. While the individuals in a species had a great deal of similarity, there was some diversity. If the environment changed, a subgroup in the species with some common characteristics would have a selective advantage, i.e. they would produce more progeny. Thus there would be a gradual enrichment for these traits in the species, and the species would be observed to change. Darwin’s second proposal could be seen as an extension of the first. Species are not created independently of other species; rather all species evolved from other species. A proposed scenario for creation of two species was the physical separation of two groups of the same species followed by independent evolution of each group until they would be recognized as separate species. In Darwin’s model of life on earth, every living organism has a parent, and every living organism is related. If a supernatural force created life, it created the first living organism. After that event new species were generated by evolution. Thus inheritance, or genetics, has a central importance. All forms of life, present and past, have their place in a giant family tree.

In 1864 Pasteur demonstrated that the organisms that ferment sugar to produce alcohol and the organisms that cause meat to spoil, do not appear “spontaneously” in the media. Rather they are the progeny of organisms transported by air currents or on contaminated containers. If the media and container is sterilized by heating and kept from the air, it does not spoil. Again, living organisms are always the progeny of other

living organisms, even in the case of yeast and bacteria. Inheritance thus also links all microorganisms together.

While inheritance was central in the theories of both Darwin and Pasteur, neither had any real ideas or insights for a mechanism of this central process. Mendel, an obscure Austrian priest and high school teacher, discovered the discrete, or digital nature of inheritance. In 1865 he published the results of crosses between stains of peas, i.e. the characteristics of progeny obtained after the pollen of one strain was used to fertilize another strain. One key to his success was to follow simple traits; seeds were either smooth or wrinkled, green or yellow. He found these traits to be inherited in an all-or-none manner, the seeds of the progeny were either smooth or wrinkled. In many cases the ratios of seeds that had the two characteristics was typically a simple one, e.g. 1/4 wrinkled and 3/4 smooth. The explanation for this ratio (which he didn't know) is that peas are diploid, i.e. each has two copies of all genes. Each parent has one wrinkled and one smooth gene, and thus only 1/4 of the progeny will have two wrinkled genes. If the smooth gene is dominant, only one copy is needed to make the seed smooth, only progeny with no smooth genes, or two wrinkled genes will have smooth genes. Mendel's paper was probably not read by many influential scientists, and it seems to have made no impact on any that might have seen it. It wasn't until 1900 that three scientists independently "rediscovered" Mendel's paper and the field of genetics was launched.

In the 20th century the pace of discovery accelerated. Muller described the production of mutants by x-rays in 1927. This work had a practical impact; it enabled many geneticists to induce mutations at much greater rate than occurred by natural causes. More mutants meant that they could do more genetic crosses. The ability to induce mutations reinforces the idea that the genetic material, what ever it was, could be manipulated. Since it could be altered by x-rays, as other material was, the genetic material was a little less mysterious. Muller had to wait until 1946 for the Nobel Prize for this work.

Thomas Hunt Morgan and Alfred Sturtevant gave genetics a geometric structure. Morgan began to work with the fruit fly, Drosophlia melogaster, in 1909 while at Columbia. In 1928 he moved to Caltech along with his student Sturtevant. The fruit flies could be maintained in small milk bottles, and they had a short generation time which allowed many crosses to be done in a few months. Using this system they were able to study the frequency at which mutations on the same chromosome would recombine to produce the original genotype. Using these recombination frequencies they were able to create a genetic map. In 1913 they published studies of double recombination, where three mutants were separated in the progeny. The fundamental result of this work was to show that mutations could be arranged as a linear sequence, or map. on a chromosome.

Jumping ahead four decades, Semore Benzer constructed a genetic map with the ultimate resolution. He studied recombinants between mutants in the rII gene of the bacterial virus T4. The protein coded by the rII gene had not be isolated or characterized, and its biochemical function was unknown. However, rII mutants had a useful property; while mutants and non-mutants grew well on one stain of bacteria (permissive), only non-mutants (wild type) grew on another strain of bacteria (non-permissive). The experiment was basically simple. First infect a culture of permissive bacteria with equal amounts of two rII mutants so that most bacteria were infected

with both mutants, allow the virus to grow, recombine, and produce progeny. Secondly, determine the number of progeny that had lost both rII mutants through recombination by plating onto non-permissive bacteria. The magic of this experiment is that you could plate 10^8 viruses on one plate and detect the presence of only one wild-type recombinant. Using this system Benzer was able to quickly construct a linear genetic map of the rII mutants. The map of hundreds of mutants contained many locations that contained several independent mutants, e.g. when any pair of these mutants were "crossed" no wild-type viruses were obtained. Thus these mutants were either at the same site or the experimental method was not able to resolve them. However, all mutants that could be resolved in this map were several factors of ten farther apart than the resolution of the experiment. The closest distance observed between any two mutants was calculated to be about one base pair. Thus, this completely genetic experiment determined the fundamental grain or resolution of the genetic material: one base pair.

Various lines of evidence had suggested that the nucleus of the cell, and more specifically the chromosomes, contained the genetic information. However, visualization of genetic material reached a new level after the chance discovery in 1933 by Heitz and Bauer that DNA replicates many times but does not separate in some cells of the salivary glands of Drosophlia; the chromosomes just become thicker with each round of DNA replication. Stains reveal a complex pattern of thick and thin bands along the chromosomes corresponding to the distribution of histones. Examination of the chromosomes from flies known by genetic mapping to have large deletions revealed deletions in the banding pattern of the expected size. This started the field of cytogenetics in which genes, or at least a pattern associated with genes, can be seen and studied.

Genetics and biochemistry merge

Now that genes could be seen, or at least bands associated with genes could be seen, there was more speculation about their chemical structure. In 1944 the physicist Erwin Schrödinger published the thin book "What is life?" which had a profound effect on many young scientists who would later create the field of Molecular Biology. Schrödinger considered the general properties that genes must have in order to function and concluded "to reconcile the high durability of the hereditary substance with its minute size, we had to evade the tendency to disorder by 'inventing the molecule', in fact an unusually large molecule". Schrödinger believed that no new physical forces would be needed to describe the gene, but that the actual molecular structure would likely be novel compared to those molecules familiar to the chemist.

The biochemists working from the bottom up were not particularly useful at this stage because they believed proteins were the only class of molecules that came close to having the specificity required by the gene. Since the chromosomes were mostly protein, that seemed to be an appropriate assumption. However, in the year that "What is Life?" was published, 1944, Avery, MacLeod and McCarty showed that a preparation of DNA could transform one strain of bacteria to another. The actual process of bacterial transformation had been described by Griffin in 1928, but in his experiments crude extracts one strain were injected into infected mice and the transformed stains recovered from the animals. The Avery, MacLeod and McCarty experiments used purified bacterial extracts to directly transform bacterial cultures.

DNAse, but not protease, inactivated the transforming activity. However, these results had an impact on few practicing scientists. As Gunther Stent commented, experimental results that are too far ahead of their time and thus can't be connected to the main body of knowledge can not be assimilated into the theoretical framework of a field.

However, biochemists intensified their study of nucleic acids, especially DNA. As an example, Chargaff published many studies of the nucleotide composition of DNA isolated from dozens of species and found that the ratio of the four bases varied widely between species, but was identical among individuals of a species. He also showed that all 16 dinucleotide pairs could be found, and again the frequencies were characteristic of species. Finally Chargaff observed that for all DNA samples the frequency of A and T were equal as was the frequency of G and C. Gradually DNA was becoming more interesting. In 1952 Hershey and Chase labeled the DNA of the T4 bacterial virus with P^{32} and the protein with C^{14}. They found that only the DNA entered the infected bacterial cell, the protein remained on the outside and could be removed by blending without decreasing the production of virus progeny. DNA was hot.

Genetics becomes biochemistry

The pivotal event in defining genes in chemical terms was a proposal for the structure for DNA published by Watson and Crick in 1953. The evidence and data they used to make their model came from many studies using different techniques, and neither Watson or Crick had any part in most of that experimental work. However, Watson and Crick were very smart and intensely motivated. Specifically they believed the chemical-physical nature of the gene was the most important problem in biology, they were convinced that DNA was the genetic material and hoped the structure of the DNA molecule would reveal the mechanism of replication. These convictions drove them to construct models of DNA using all the information at their disposal, with the hope that only one structure would satisfy all the constraints. This strategy had been used by Pauling to deduce the structure of the protein alpha helix, and indeed Pauling was a major competitor in the race to find the structure of DNA.

The data from x-ray diffraction studies of DNA threads was part of the mix they used, and here Crick could claim some expertise, since he had just previously written a paper on the characteristics of the diffraction pattern of helical molecules. However, DNA threads were not crystals like the ones used in diffraction studies of proteins. The DNA threads were random fragments of large DNA molecules which were drawn out from a supersaturated solution in a manner analogous to pulling a thread out from a ball of wool fibers. The DNA molecules were aligned along their axis, but not in any other way. The diffraction pattern was a few dozen fuzzy blurs arranged with a X shaped symmetry, not the hundreds or thousands of sharp spots which characterize the diffraction pattern of a crystallized protein. However, Crick immediately recognized the X pattern as the signature of a helical molecule. The spacing of the fuzzy blurs gave a few important values for distances of repetition along the helix.

Watson and Crick first proposed an incorrect structure for DNA, which when shown to the other scientists working on DNA structure caused them some embarrassment and resulted in instructions from their supervisor to leave DNA to

more competent people at nearby Kings College where most of the x-ray work was being done by Rosalind Franklin. However, while increasingly sharp diffraction patterns were obtained, and some progress was made studying the differences between wet and dry forms of DNA, no structures were produced, and there seemed to even be no consistent agreement that DNA was helical. When Franklin decided to move to Paris and pursue other projects Watson and Crick were allowed to resume their model building.

The fact that the bases were parallel had been deduced by optical studies, and as mentioned previously, the fact that the fractional amount of A equals T and G equals C had been discovered by Chargaff. The structures of the four nucleotides in DNA were known, although the form (keto versus enol) of the portions of the bases that form the base pairs was incorrectly presented in most text books. Fortunately Watson shared an office with Jerry Donohue, who informed Watson during an accidental conversation that the bases were actually in the keto form in an aqueous environment. Using simple flat cardboard models of the bases (remember they are planar) Watson now discovered the A:T and G:C paring and the rest of the helix model rapidly followed. This structure was published in the April 25th, 1953 edition of the Journal Nature. The DNA double helix was quickly accepted by essentially all the scientists working DNA structure because it was consistent with all the data and provided an immediate mechanism for DNA replication. It was just too good not to be correct!

However, it took years for many biochemists and geneticists to accept the fact that DNA was the genetic material and that the double helix revealed how it was replicated. As with many important discoveries much of the controversy was about who should get the credit. Chargaff insisted that he knew about base pairing all along, but was just too conservative to publish a structure without solid evidence. Rosalind Franklin seems to have graciously accepted omission from authorship of the paper proposing the structure. This is perhaps not surprising since she had been quite vocal in criticizing Watson and Crick for trying to determine the structure by building a model with incomplete data. She did publish a description of the diffraction pattern and its implications in the same issue of Nature, and so shared a little in the glory.

Watson, Crick, and Wilkins, the head of the group that produced the x–ray diffraction patterns, received the Nobel Prize for the structure of DNA in 1962. The rule for the Nobel Committee is that only three people can share a prize, which has caused bad feelings after the award of several prizes. Recently several authors have claimed that Franklin should have received more recognition for her contributions. She couldn't have won the Nobel Prize because she died several years before the prize was given, and it can not be awarded posthumously. The fact that woman scientists were treated rather badly at English colleges in Franklin's time, and the publication in 1968 of a very popular but brutally frank and personal account of the discovery of the structure of DNA by Watson has fueled the fire[2]. However, that controversy has little to do with biology and information, so we move on.

---

[2] The Norton Critical Edition of "The Double Helix" by James D. Watson (W. W. Norton & Company, New York, 1980) contains the full text of the original book plus several reviews of the book by famous scientists. An extensive historical introduction and commentaries on the reviews by the molecular biologist Gunther Stent give the reader valuable insight into the history of molecular biology. See "The Path to the

Messenger RNA

It was not until 1961 that Brenner, Jacob, and Meselson identified mRNA and showed that protein was made by the mRNA-ribosome complex. Attempts to identify mRNA during these times used bacterial systems, but it was not known that mRNA in bacteria has a half life of about 3 minutes, and thus only a few percent of the total RNA is messenger. To study such unstable mRNA it is necessary to add a radioactive precursor to the bacterial culture for no longer than a minute or two, so that a large fraction of the isotope is incorporated into RNA before it has been degraded. Using this protocol up to half of the isotope is in mRNA. It seems amazing that a bacterium would only use a mRNA molecule for 3 minutes, when it could make a hundred or so protein chains at most, and then degrade it back to nucleotides. What a waste of energy! However, a consequence of the rapid turnover of mRNA is that when transcription of a gene is turned off, the level of mRNA and thus the rate of synthesis of protein coded by that gene, rapidly decreases. The need for rapid control, i.e. information transfer, is stronger than the need for energy efficiency. This is an example of the central role of information in biology.

The nucleotide to amino acid code

Also in 1961, Nirenberg showed that poly U RNA made polyphenyalanine. Thus at least one code word for the amino acid phenylalanine must be UUU. These first experiments to determine the RNA-amino acid code were done using RNA made by the enzyme polynucleotide phosphorylase, which polymerized nucleotides independent of a template, and thus produced random polymers. The systematic determination of the code required the disciplined approach of an organic chemist, in this case Khorana and his huge group at MIT, to make RNA with defined sequences. With these RNA species the complete code was quickly determined.

DNA sequencing

In 1968 Holly, along with Khorana and Nirenberg, was given the Nobel Prize. We already know why Khorana and Nirenberg were given the prize. Holly earned the award by developing and using techniques to sequence RNA, specifically tRNA. In order to sequence nucleic acid you need a specific end to start from. Transfer RNA molecules are about 75 nucleotides long, and thus were a natural target for Holly, but most DNA molecules are orders of magnitude longer. Thus serious work on DNA sequencing could not be started until Smith and Wilcox found that many bacteria produce DNAses that only cut DNA at specific sites, typically 4-8 nucleotides long. Smith and others working on these DNAses were awarded the Nobel Prize for this work in 1970.

In 1977 methods to sequence DNA were developed independently by Walter Gilbert and Fred Sanger, and they won the Nobel Prize for this work in 1980 (note that this was the second Nobel for Sanger). The Sanger method soon became preferred, and

---

Double Helix; the Discovery of DNA" by Robert Olby (1994, Dover Publications, Inc., New York) for a comprehensive history. "The Eight Day of Creation; The Makers of the Revolution in Biology" by Horace Judson (1979, Simon and Schuster, New York) is another history of this time in biology.

was continually improved over the next few decades. No one now sequenced RNA, the RNA was copied into DNA and sequenced.

However, even after 1980 determining long DNA sequences was labor intensive, and thus only "interesting" segments of DNA could be sequenced. Most of the money was in medical research, specifically determining the cause of inherited disease. The process of finding the altered gene and then the protein was extremely tedious. The early steps required study of families in which the disease was common. The method was to look for association of the disease in these families with genetic markers having known positions. The last step was to sequence the putative gene responsible for the disease and show a difference in the nucleotide sequence between affected and normal individuals. Several heroic projects using this method, called positional cloning, were successful, e.g. in 1989 Tsui, Riordan, and Collins identified the gene responsible for cystic fibrosis. However, other projects were unsuccessful, and even the successful ones accumulated a large amount of information which was not published because the sequences weren't the one the team was looking for.

The process of actually determining a DNA sequence was all the time becoming faster and cheaper. In 1986 Leroy Hood and Lloyd Smith at Caltech developed the first automated machine to produce DNA sequence data, and the next year Applied Biosystems manufactured and sold the first of these machines. More and more scientists began to see that it was time to bite the bullet and just sequence the entire human genome. The Department of Energy (DOE) had a significant research effort in biology, inherited from the days when it was the Atomic Energy Agency and needed to study the biological effects of radiation. DOE had special expertise in instrumentation, and in 1986 redirected $5.3 million to a program on genomic sequencing and announced that it would lead a national effort to sequence the human genome.

After much talk, many meetings, and various estimates of the cost, the National Institutes of Health (NIH) announced it, not DOE, would lead the national program to sequence the human genome and created the Office of Human Genome Research in 1988 with Watson as its director. Thus political and social problems threatened to be as difficult as the technical ones. In addition to the NIH-DOE conflict, which was solved by a "memorandum of understanding" which ceded the lead role to NIH, scientists working at NIH and the much larger number of scientists at Universities supported by NIH grants had to be reassured that the Human Genome Project wouldn't suck all research funds into a black hole. Finally, molecular biologists had be acclimated to the large research groups that the physics community was already accustomed to. In 1990 NIH stated that the Human Genome Project (HGP) had actually started, and planned to produce a complete sequence in 15 years.

In the early years of the Human Genome Project it was suspected that the Sanger method would be too slow and expensive for the genome, and thus a sizable effort was directed toward developing new methods. It turned out that the polymerase chain reaction (PCR) was the only really new method developed or needed. PCR was invented in 1985 by Kerry Mullis who was working on DNA diagnostics at Cetus Corporation, a biotechnology company in California that was not funded by NIH or DOE grants. PCR enables you to make copies of a segment of a DNA chain if you know the sequence of a dozen or so nucleotides at the ends of the sequence. It

revolutionized molecular biology, allowing a specific sequence on as little as one molecule could be copied, or amplified, to produce more than 10^9 copies.

The corporate genome

In 1992, J. Craig Venter, a scientist at the NIH, became impatient with his failure to obtain funding for a large scale effort to sequence all the translated regions of the human genome. His plan was to purify mRNA from human cells, produce DNA with complementary sequences (cDNA) using reverse transcriptase *in vitro*, and then sequence all the cDNAs. Since only 1 or 2 percent of the human genome is transcribed, and the project didn't include determining the order of the cDNA sequences on the genome, this would be much less work than sequencing the entire genome. Venter claimed that the cDNA sequences were the only part of the genome that was of medical interest anyway. If the funding committee at NIH was unenthusiastic about the project, the venture capital community was not. A pair of companies was created. Human Genome Sciences (HGS), headed by William Haseltine, would determine the cDNA sequences, and sell the data to pharmaceutical companies and anyone else that would pay. The Institute for Genomic Research (TIGR), headed by Venter, would assist HGH, but would also pursue independent sequencing projects and publish the results. These two groups had impressive funds at their disposal and were free of the social and political restraints of the academic community. Perhaps in response to the threat of much of the human genome sequence becoming proprietary Britain's Wellcome Trust committed $95 million to establish a large sequencing center which would complement the efforts of the US HGP, which was beginning to be called the public consortium. Now we have real money, real egos, and a real fight in progress.

HGH started producing sequences and several large pharmaceutical corporations subscribed to the data. Since you had to pay to see the results, there was a certain sense of mystery around the actual progress being made, but the subscribers seemed satisfied. In 1995, Venter, Claire Fraser, and Hamilton Smith published the genome sequence, obtained by TIGR, of the bacterium *H. influenza*. The genomes of several viruses had been determined, but this sequence was far larger and the first genome of a free living organism. Venter and TIGR were thus validated as serious contenders for the human genome sequence quest.

In 1998 Venter and Applied Biosystems (which had renamed itself Applera) announced that he would leave TIGR to head Celara Genomics, a new division of the corporation. The new division would sequence the entire human genome in three years, which would thus be several years ahead of the goal set for itself by the Public Consortium. As with HGS, the data would be available only to subscribers, again mostly the large pharmaceutical corporations. Of course Celera Genomics would purchase hundred's of Applied Biosystems sequencers, and the facility would run around the clock. Aside from money and publicity, there was an important shift in methodology by Celara. The public consortium was using a strategy that made sense ten years earlier. The first step was to identify hundreds of marker sequences along the genome, and determine the linear sequence of the markers by genetic methods. The genome was then broken into large fragments and cloned. The actual sequencing was done on fragments of these fragments, and the sequences were the assembled into larger and larger contiguous segments (contigs) until the entire genome was supposed to emerge. The advantage of this incremental process was that the relative position of

a sequence in the genome was usually known far before all the actual sequencing was done. The disadvantage was that the preliminary stages were laborious and time consuming. Venter believed that the cost of determining the nucleotide sequence of a DNA fragment had now become so low that a better method was to just cut the genome into a very large number of random fragments and determine the sequences. The entire genomic sequence could be deduced by fitting the random sequences together using overlapping segments. This was called the shotgun method. It required a certain amount of faith (or arrogance) since previous efforts had shown that it could be difficult to generate random fragments. Some segments of genomic DNA seemed to be resistant to cutting, thus the sequence determined by the shotgun method could contain large gaps.

Detractors of Venter's shotgun method acknowledged that it worked on small genomes, but claimed it wouldn't scale up to larger ones. Venter announced that he would, in his spare time, and in collaboration with university researchers, determine the genome sequence of the fruit fly, *D. melanogaster*, and of course publish the results. The sequence was published in 2000, just two years after the start of construction of the Celara labs. While the genetic and sequence data that had been obtained before Celara started work on the fly sequence was of considerable help, the contribution by Celara in the determination of the final sequence was substantial, and served to mute, or at least intimidate, most critics of the shotgun method.

Since both Celara and the public consortium were approaching the goal of determining the human genome sequence, there were several problems to solve. While Celera was still selling data, it wanted to published a summary of its results. The public consortium had been publishing sequence data on the Internet continually. After several efforts to broker a joint publication, in 2001 the HGP published in the journal Nature, while Celera published in the journal Science. President Clinton presided over a White House ceremony in which Francis Collins (then the head of the HGP) and Venter shook hands a la Arafat-Barak. The competition was over. Both groups made important contributions. It was clear that the shotgun approach used by Venter would have been able to generate the entire human genome sequence. However the competition was not symmetrical in that throughout the sequencing effort the Celera group had complete access to the public consortium data, while the reverse was not true.

The public sequence available in 2001 had many gaps and mistakes, and presumably the Celera data did also. Both groups continued sequencing and the data became better and better. In 2002 Venter left Celera and Tony White, the CEO of the umbrella corporation announced that Celera would become a pharmaceutical company, i.e. would develop drugs. A change at Celera was probably inevitable since what ever superiority the Celera data might have had at one time was rapidly diminished as the public data improved. After all, there is only one sequence, and after you are very, very close to that sequence you might think there isn't much more to do. However, as we see below, the nucleotide sequence is only the beginning as far as medical research is concerned.

Figure BioChron

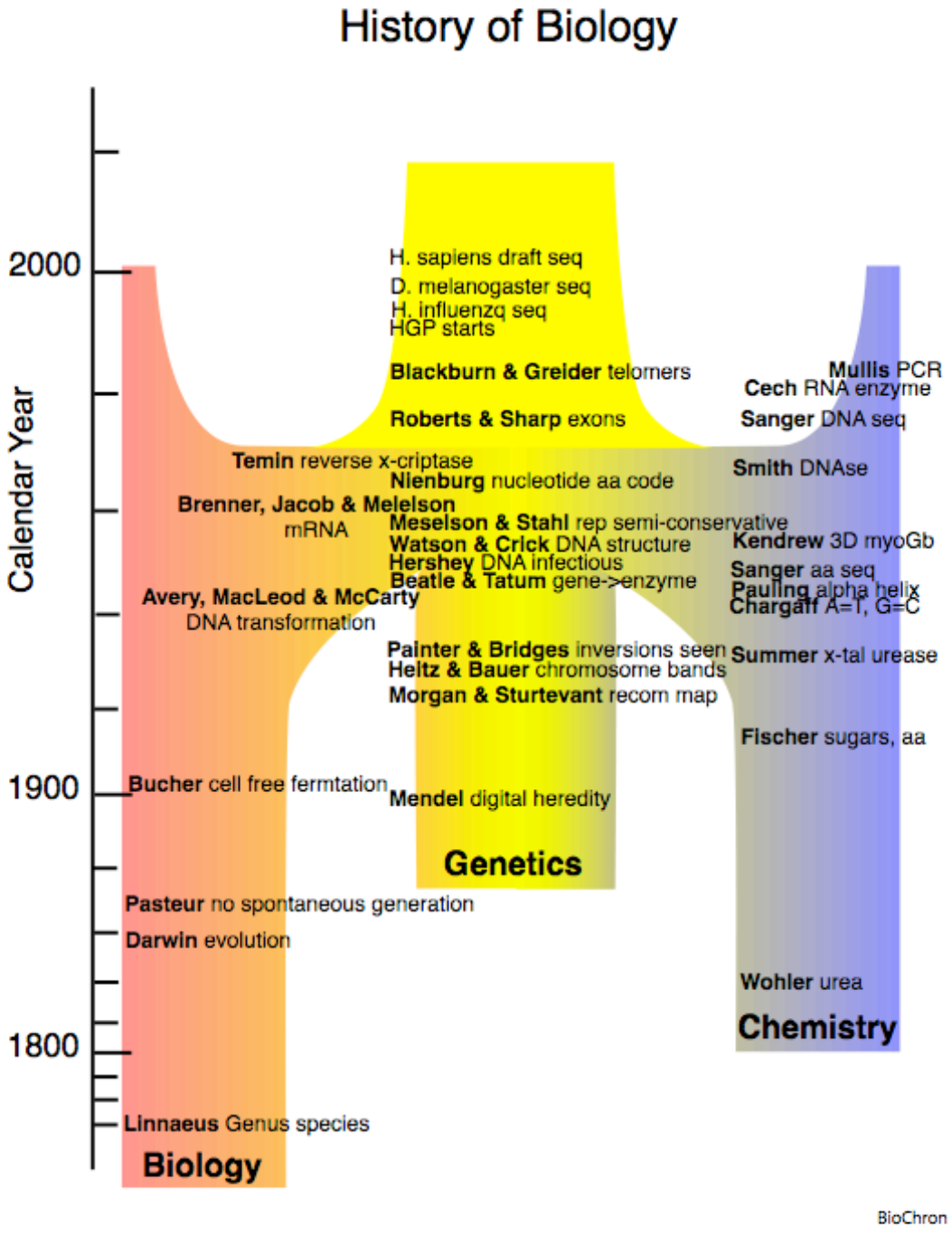

Figure BioChron. The study of biology originally meant observation and classification of living organisms, which Linnaeus systematized in 1741. The study of the chemistry of living organisms was a separate discipline, which Wohler started in 1828 by showing that urea, an organic compound found in organisms, could also be synthesized in the laboratory from inorganic chemicals. It had long been recognized that progeny inherit the characteristics of their parents, but Mendel was the first to document the transfer of discrete characteristics, which gives genetics its digital nature. While he published his work in 1865, it went unnoticed until 1900. Gradually the work of biologists, biochemists, and geneticists merged and with the chemical nature of the gene and its replication revealed in 1953 by Watson and Crick, the three areas merged. Today it is just possible to publish work in one of these areas without mention of the others, but any reader will integrate the results within a framework provided by the other two. The discoveries listed on this chart were selected primarily to illustrate the chronology of our understanding of the storage and transfer of information by life.

After the human genome

Of course obtaining the sequence of one human genome is only the beginning. To investigate the nature of inherited disease you need to have the sequences of at least several people with the disease and sequences of normal people, or rather sequences of people that don't have that particular disease, since no one is normal. There is some hope that mutations that cause disease are associated with marker mutations that have been given the name single nucleotide polymorphisms (SNP). If this is true, it may only be necessary to determine the profile of SNPs in a patient to infer the existence of the disease producing mutations.

The Holy Grail of human sequencing in the year 2005 (the year I am writing this) is the $1,000 genome, i.e. determining the sequence of the genome of one individual for $1,000. There is no reason to believe that this is possible anytime soon, but it has a nice ring to it, and there is no fundamental reason why it can not be done.

An more easily achievable goal is the sequencing of the genomes of many of the organisms that have been studied extensively, and use that data to increase our basic knowledge of biology. The genomic sequences of the model organisms described in the previous chapter, are an obvious first choice. But, in order to know what part of the genomic sequence of one of these model organisms is just "random" and what part is truly characteristic of the class, it is necessary to sequence the genomes of several closely related organisms.

Thus, both for medical studies and more fundamental research, there will be an ever increasing demand for DNA sequence information.

Chapter summary

The earth was formed about 4.5 billion years ago but the hard crust that now makes up the surface of the earth is only about 3.8 billion years old. The first fossils, which represent bacterial-like organisms are 3.5 billion years old, so life took "only" 300 million years to arise. The very first organisms are thought to have used RNA to store genetic information and to catalyze chemical reactions.

When organisms developed the ability to use sunlight as a source of energy they produced oxygen which converted the reducing atmosphere of the earth into the

oxygen rich one of the present 2.2 billion years ago. The first eukaryotes appeared 1.8 billion years ago.

The early studies of plants and animals were simple observations of existence and behavior; without microscopes and chemistry what else can you do. However, it became clear that living organisms can be organized into a hierarchical system of groups based on similarity of structure.

The publication of "On the Origin of Species" by Charles Darwin in 1859 revolutionized biology, as it asserted that the groups of animals and plants were the result of evolution from common ancestors. Evolution is the organizing principal of modern biology, in fact the ability to reproduce and evolve can be considered to be the defining characteristic of life.

The first chemists that studied living organisms believed that the "organic" compounds were the unique products of life and could not be made in any other way. However, first with urea and finally with DNA it has become apparent that biochemistry is chemistry.

The fields of biology, biochemistry, and genetics were merged when Watson and Crick proposed the structure of DNA. Each of these fields has a different perspective, but there is also now extensive overlap. Modern methods for obtaining the sequence of large DNA molecules have produced the entire nucleotide sequence of the human genome and the sequences of many other animal and plant genomes are accumulating in public data bases. Biology has entered the information age.