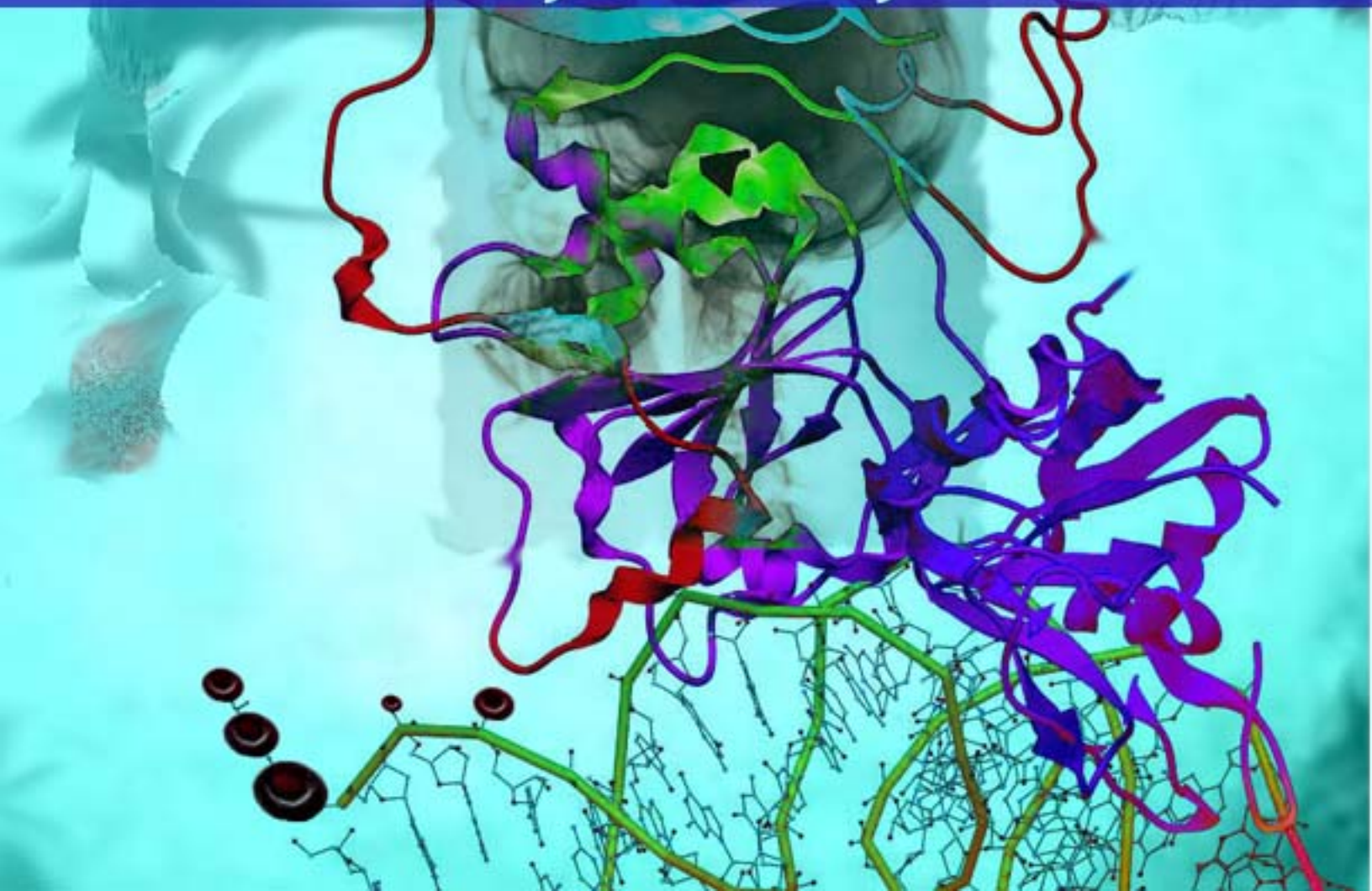
History and Research of Genetics

Mayola Motley



First Edition, 2012

ISBN 978-81-323-3303-6

Published by:
Research World
4735/22 Prakashdeep Bldg,
Ansari Road, Darya Ganj,
Delhi - 110002
Email: info@wtbooks.com

Table of Contents

Chapter- 1

History of Genetics

The **history of genetics** started with the work of the Augustinian friar Gregor Johann Mendel. His work on pea plants, published in 1866, described what came to be known as Mendelian Inheritance. In the centuries before—and for several decades after—Mendel's work, a wide variety of theories of heredity proliferated.

1900 marked the "rediscovery of Mendel" by Hugo de Vries, Carl Correns and Erich von Tschermak, and by 1915 the basic principles of Mendelian genetics had been applied to a wide variety of organisms—most notably the fruit fly *Drosophila melanogaster*. Led by Thomas Hunt Morgan and his fellow "drosophilists", geneticists developed the Mendelian, which was widely accepted by 1925. Alongside experimental work, mathematicians developed the statistical framework of population genetics, bring genetical explanations into the study of evolution.

With the basic patterns of genetic inheritance established, many biologists turned to investigations of the physical nature of the gene. In the 1940s and early 1950s, experiments pointed to DNA as the portion of chromosomes (and perhaps other nucleoproteins) that held genes. A focus on new model organisms such as viruses and bacteria, along with the discovery of the double helical structure of DNA in 1953, marked the transition to the era of molecular genetics.

In the following years, chemists developed techniques for sequencing both nucleic acids and proteins, while others worked out the relationship between the two forms of biological molecules: the genetic code. The regulation of gene expression became a central issue in the 1960s; by the 1970s gene expression could be controlled and manipulated through genetic engineering. In the last decades of the 20th century, many biologists focused on large-scale genetics projects, sequencing entire genomes.

Pre-Mendelian ideas on heredity

Ancient theories

The most influential early theories of heredity were that of Hippocrates and Aristotle. Hippocrates' theory (possibly based on the teachings of Anaxagoras) was similar to

Darwin's later ideas on pangenesis, involving heredity material that collects from throughout the body. Aristotle suggested instead that the (nonphysical) form-giving principle of an organism was transmitted through semen (which he considered to be a purified form of blood) and the mother's menstrual blood, which interacted in the womb to direct an organism's early development. For both Hippocrates and Aristotle—and nearly all Western scholars through to the late 19th century—the inheritance of acquired characters was a supposedly well-established fact that any adequate theory of heredity had to explain. At the same time, individual species were taken to have a fixed essence; such inherited changes were merely superficial.

In the 9th century CE, the Afro-Arab writer Al-Jahiz considered the effects of the environment on the likelihood of an animal to survive, and first described the struggle for existence. His ideas on the struggle for existence in the *Book of Animals* have been summarized as follows:

"Animals engage in a struggle for existence; for resources, to avoid being eaten and to breed. Environmental factors influence organisms to develop new characteristics to ensure survival, thus transforming into new species. Animals that survive to breed can pass on their successful characteristics to offspring."

In 1000 CE, the Arab physician, Abu al-Qasim al-Zahrawi (known as Albucasis in the West), wrote the first clear description of haemophilia, a hereditary genetic disorder, in his *Al-Tasrif*. In this work, he wrote of an Andalusian family whose males died of bleeding after minor injuries.

Plant systematics and hybridization

In the 18th century, with increased knowledge of plant and animal diversity and the accompanying increased focus on taxonomy, new ideas about heredity began to appear. Linnaeus and others (among them Joseph Gottlieb Kölreuter, Carl Friedrich von Gärtner, and Charles Naudin) conducted extensive experiments with hybridization, especially species hybrids. Species hybridizers described a wide variety of inheritance phenomena, include hybrid sterility and the high variability of back-crosses.

Plant breeders were also developing an array of stable varieties in many important plant species. In the early 19th century, Augustin Sageret established the concept of dominance, recognizing that when some plant varieties are crossed, certain characters (present in one parent) usually appear in the offspring; he also found that some ancestral characters found in neither parent may appear in offspring. However, plant breeders made little attempt to establish a theoretical foundation for their work or to share their knowledge with current work of physiology, although Gartons Agricultural Plant Breeders in England explained their system in their seed catalogue of 1901.

Mendel

In breeding experiments between 1856 and 1865, Gregor Mendel first traced inheritance patterns of certain traits in pea plants and showed that they obeyed simple statistical rules. Although not all features show these patterns of Mendelian inheritance, his work acted as a proof that application of statistics to inheritance could be highly useful. Since that time many more complex forms of inheritance have been demonstrated.

From his statistical analysis Mendel defined a concept that he described as an *allele*, which was the fundamental unit of heredity. The term *allele* as Mendel used it is nearly synonymous with the term *gene*, and now means a specific variant of a particular gene.

Mendel's work was published in 1866 as *"Versuche über Pflanzen-Hybriden" (Experiments on Plant Hybridization)* in the *Verhandlungen des Naturforschenden Vereins zu Brünn (Proceedings of the Natural History Society of Brünn)*, following two lectures he gave on the work in early 1865.

Post-Mendel, pre-re-discovery

Mendel's work was published in a relatively obscure scientific journal, and it was not given any attention in the scientific community. Instead, discussions about modes of heredity were galvanized by Darwin's theory of evolution by natural selection, in which mechanisms of non-Lamarckian heredity seemed to be required. Darwin's own theory of heredity, pangenesis, did not meet with any large degree of acceptance. A more mathematical version of pangenesis, one which dropped much of Darwin's Lamarckian holdovers, was developed as the "biometrical" school of heredity by Darwin's cousin, Francis Galton. Under Galton and his successor Karl Pearson, the biometrical school attempted to build statistical models for heredity and evolution, with some limited but real success, though the exact methods of heredity were unknown and largely unquestioned.

Classical genetics

The significance of Mendel's work was not understood until early in the twentieth century, after his death, when his research was re-discovered by other scientists working on similar problems. Hugo de Vries, Carl Correns and Erich von Tschermak

There was then a feud between Bateson and Pearson over the hereditary mechanism. Fisher solved this in The Correlation Between Relatives on the Supposition of Mendelian Inheritance

1865 Gregor Mendel's paper, *Experiments on Plant Hybridization*
1869 Friedrich Miescher discovers a weak acid in the nuclei of white blood cells that today we call DNA
1880-1890 Walther Flemming, Eduard Strasburger, and Edouard van Beneden elucidate chromosome distribution during cell division

1889 Hugo de Vries postulates that "inheritance of specific traits in organisms comes in particles", naming such particles "(pan)genes"
1903 Walter Sutton hypothesizes that chromosomes, which segregate in a Mendelian fashion, are hereditary units
1905 William Bateson coins the term "genetics" in a letter to Adam Sedgwick and at a meeting in 1906
1908 Hardy-Weinberg law derived.
1910 Thomas Hunt Morgan shows that genes reside on chromosomes
1913 Alfred Sturtevant makes the first genetic map of a chromosome
1913 Gene maps show chromosomes containing linear arranged genes
1918 Ronald Fisher publishes "The Correlation Between Relatives on the Supposition of Mendelian Inheritance" the modern synthesis of genetics and evolutionary biology starts.
1928 Frederick Griffith discovers that hereditary material from dead bacteria can be incorporated into live bacteria
1931 Crossing over is identified as the cause of recombination
1933 Jean Brachet is able to show that DNA is found in chromosomes and that RNA is present in the cytoplasm of all cells.
1941 Edward Lawrie Tatum and George Wells Beadle show that genes code for proteins.

The DNA era

1944 The Avery–MacLeod–McCarty experiment isolates DNA as the genetic material (at that time called transforming principle)
1950 Erwin Chargaff shows that the four nucleotides are not present in nucleic acids in stable proportions, but that some general rules appear to hold (e.g., that the amount of adenine, A, tends to be equal to that of thymine, T).
Barbara McClintock discovers transposons in maize
1952 The Hershey-Chase experiment proves the genetic information of phages (and all other organisms) to be DNA
1953 DNA structure is resolved to be a double helix by James D. Watson and Francis Crick
1956 Joe Hin Tjio and Albert Levan established the correct chromosome number in humans to be 46
1958 The Meselson-Stahl experiment demonstrates that DNA is semiconservatively replicated
1961-1967 Combined efforts of scientists "crack" the genetic code, including Marshall Nirenberg, Har Gobind Khorana, Sydney Brenner & Francis Crick
1964 Howard Temin showed using RNA viruses that the direction of DNA to RNA transcription can be reversed
1970 Restriction enzymes were discovered in studies of a bacterium, *Haemophilus influenzae*, enabling scientists to cut and paste DNA

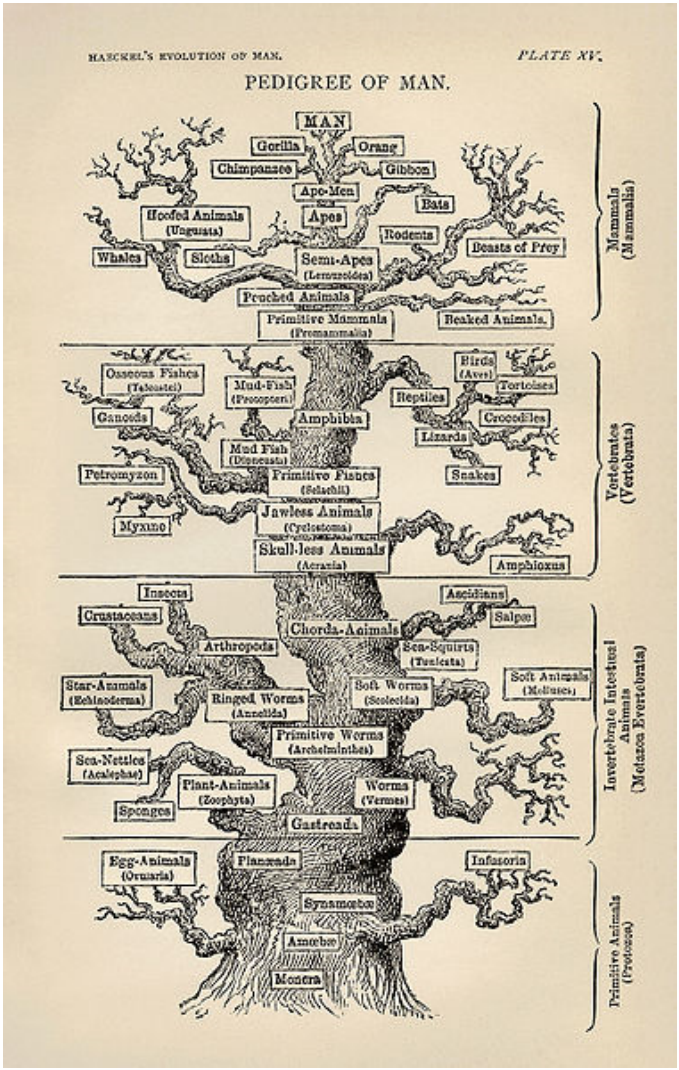
The genomics era

1972, Walter Fiers and his team at the Laboratory of Molecular Biology of the University of Ghent (Ghent, Belgium) were the first to determine the sequence of a gene: the gene for bacteriophage MS2 coat protein.

1976, Walter Fiers and his team determine the complete nucleotide-sequence of bacteriophage MS2-RNA
1977 DNA is sequenced for the first time by Fred Sanger, Walter Gilbert, and Allan Maxam working independently. Sanger's lab sequence the entire genome of bacteriophage Φ-X174.
1983 Kary Banks Mullis discovers the polymerase chain reaction enabling the easy amplification of DNA
1989 The human gene that encodes the CFTR protein was sequenced by Francis Collins and Lap-Chee Tsui. Defects in this gene cause cystic fibrosis.
1995 The genome of *Haemophilus influenzae* is the first genome of a free living organism to be sequenced
1996 *Saccharomyces cerevisiae* is the first eukaryote genome sequence to be released
1998 The first genome sequence for a multicellular eukaryote, *Caenorhabditis elegans*, is released
2001 First draft sequences of the human genome are released simultaneously by the Human Genome Project and Celera Genomics.
2003 (14 April) Successful completion of Human Genome Project with 99% of the genome sequenced to a 99.99% accuracy

Chapter- 2

History of Evolutionary Thought



The Tree of Life as depicted by Ernst Haeckel in *The Evolution of Man* (1879) illustrates the 19th-century view that evolution was a progressive process leading towards man.

Evolutionary thought, the conception that species change over time, has roots in antiquity, in the ideas of the ancient Greeks, Romans, and Chinese as well as in medieval Islamic science. However, until the 18th century, Western biological thinking was dominated by essentialism, the belief that every species has essential characteristics that are unalterable. This began to change during the Enlightenment when evolutionary cosmology and the mechanical philosophy spread from the physical sciences to natural history. Naturalists began to focus on the variability of species; the emergence of paleontology with the concept of extinction further undermined the static view of nature. In the early 19th century, Jean-Baptiste Lamarck proposed his theory of the transmutation of species, the first fully formed scientific theory of evolution.

In 1858, Charles Darwin and Alfred Russel Wallace published a new evolutionary theory that was explained in detail in Darwin's *On the Origin of Species* (1859). Unlike Lamarck, Darwin proposed common descent and a branching tree of life. The theory was based on the idea of natural selection, and it synthesized a broad range of evidence from animal husbandry, biogeography, geology, morphology, and embryology.

The debate over Darwin's work led to the rapid acceptance of the general concept of evolution, but the specific mechanism he proposed, natural selection, was not widely accepted until it was revived by developments in biology that occurred during 1920s through the 1940s. Before that time most biologists argued that other factors were responsible for evolution. Alternatives to natural selection suggested during the eclipse of Darwinism included inheritance of acquired characteristics (neo-Lamarckism), an innate drive for change (orthogenesis), and sudden large mutations (saltationism). The synthesis of natural selection with Mendelian genetics during the 1920s and 1930s founded the new discipline of population genetics. Throughout the 1930s and 1940s, population genetics became integrated with other biological fields, resulting in a widely applicable theory of evolution that encompassed much of biology—the modern evolutionary synthesis.

Following the establishment of evolutionary biology, studies of mutation and variation in natural populations, combined with biogeography and systematics, led to sophisticated mathematical and causal models of evolution. Paleontology and comparative anatomy allowed more detailed reconstructions of the history of life. After the rise of molecular genetics in the 1950s, the field of molecular evolution developed, based on protein sequences and immunological tests, and later incorporating RNA and DNA studies. The gene-centered view of evolution rose to prominence in the 1960s, followed by the neutral theory of molecular evolution, sparking debates over adaptationism, the units of selection, and the relative importance of genetic drift versus natural selection. In the late 20th century, DNA sequencing led to molecular phylogenetics and the reorganization of the tree of life into the three-domain system. In addition, the newly recognized factors of symbiogenesis and horizontal gene transfer introduced yet more complexity into evolutionary history.

Antiquity

Greeks

Several ancient Greek philosophers discussed ideas that involved change in living organisms over time. Anaximander (c.610–546 BC) proposed that the first animals lived in water and animals that live on land were generated from them. Empedocles (c. 490–430 BC) wrote of a non-supernatural origin for living things, suggesting that adaptation did not require an organizer or final cause. Aristotle summarized his idea: "Wherever then all the parts came about just what they would have been if they had come to be for an end, such things survived, being organized spontaneously in a fitting way; whereas those which grew otherwise perished and continue to perish ..." although Aristotle himself rejected this view.

Plato (left) and Aristotle (right), a detail of *The School of Athens*

Plato (c. 428–348 BC) was, in the words of biologist and historian Ernst Mayr, "the great antihero of evolutionism", as he established the philosophy of essentialism, which he called the Theory of Forms. This theory holds that objects observed in the real world are only *reflections* of a limited number of essences (*eide*). Variation merely results from an imperfect reflection of these constant essences. In his *Timaeus*, Plato set forth the idea that the Demiurge had created the cosmos and everything in it because, being good, and hence, "... free from jealousy, He desired that all things should be as like Himself as they could be". The creator created all conceivable forms of life, since "... without them the universe will be incomplete, for it will not contain every kind of animal which it ought to contain, if it is to be perfect". This "plenitude principle"—the idea that all potential forms of life are essential to a perfect creation—greatly influenced Christian thought.

Aristotle (384–322 BC), one of the most influential of the Greek philosophers, is the earliest natural historian whose work has been preserved in any real detail. His writings on biology resulted from his research into natural history on and around the isle of Lesbos, and have survived in the form of four books, usually known by their Latin names, *De anima* (on the essence of life), *Historia animalium* (inquiries about animals), *De generatione animalium* (reproduction), and *De partibus animalium* (anatomy). Aristotle's works contain some remarkably astute observations and interpretations—along with sundry myths and mistakes—reflecting the uneven state of knowledge during his time. However, for Charles Singer, "Nothing is more remarkable than [Aristotle's] efforts to [exhibit] the relationships of living things as a *scala naturæ*." This *scala naturæ*, described in *Historia animalium*, classified organisms in relation to a hierarchical "Ladder of Life" or "Chain of Being", placing them according to their complexity of structure and function, with organisms that showed greater vitality and ability to move described as "higher organisms".

Chinese

Ancient Chinese thinkers such as Zhuangzi (Chuang Tzu), a Taoist philosopher who lived around the 4th century BC, expressed ideas on changing biologic species. According to Joseph Needham, Taoism explicitly denies the fixity of biological species and Taoist philosophers speculated that species had developed differing attributes in response to differing environments. Taoism regards humans, nature and the heavens as existing in a state of "constant transformation" known as the *Tao*, in contrast with the more static view of nature typical of Western thought.

Romans

Titus Lucretius Carus (d. 50 BC), the Roman philosopher and atomist, wrote the poem *On the Nature of Things* (*De rerum natura*), which provides the best surviving explanation of the ideas of the Greek Epicurean philosophers. It describes the development of the cosmos, the Earth, living things, and human society through purely naturalistic mechanisms, without any reference to supernatural involvement. *On the Nature of things* would influence the cosmological and evolutionary speculations of philosophers and scientists during and after the Renaissance.

Augustine of Hippo

In line with earlier Greek thought, the 4th century bishop and theologian, St. Augustine of Hippo, wrote that the creation story in Genesis should not be read too literally. In his book *De Genesi ad litteram* ("On The Literal Interpretation of Genesis"), he stated that in some cases new creatures may have come about through the "decomposition" of earlier forms of life. For Augustine, "plant, fowl and animal life are not perfect ... but created in a state of potentiality", unlike what he considered the theologically perfect forms of angels, the firmament and the human soul. Augustine's idea 'that forms of life had been transformed "slowly over time"' prompted Father Giuseppe Tanzella-Nitti, Professor of Theology at the Pontifical Santa Croce University in Rome, to claim that Augustine had suggested a form of evolution.

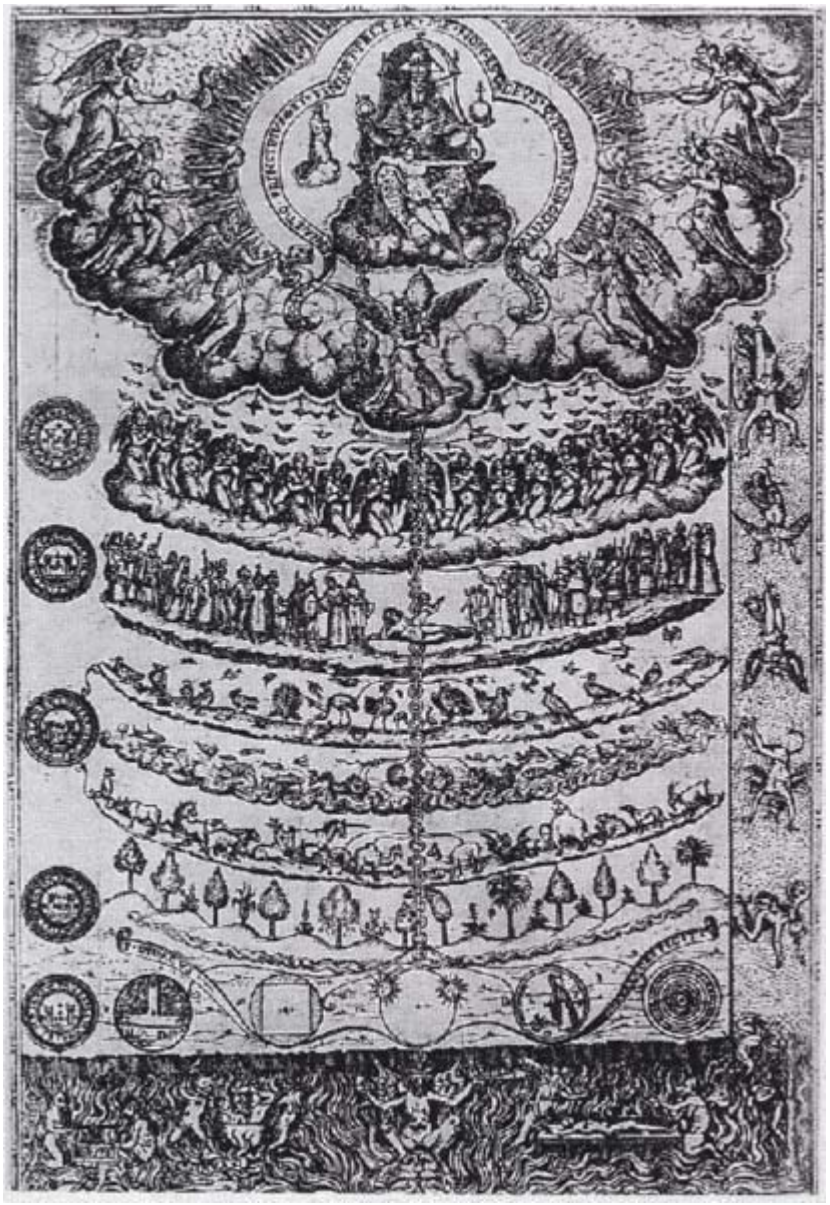
Middle Ages

Islamic philosophy and the struggle for existence

Although Greek and Roman evolutionary ideas died out in Europe after the fall of the Roman Empire, they were not lost to Islamic philosophers and scientists. In the Islamic Golden Age of the 8th to the 13th centuries, philosophers explored ideas about natural history. These ideas included transmutation from non-living to living: "from mineral to plant, from plant to animal, and from animal to man".

The first Muslim biologist and philosopher to publish detailed speculations about natural history, the Afro-Arab writer al-Jahiz, wrote in the 9th century. In the *Book of Animals*, he considered the effects of the environment on an animal's chances for survival, and described the struggle for existence. Al-Jahiz also wrote descriptions of food chains. Al-Jahiz speculated on the influence of the environment on animals and considered the effects of the environment on the likelihood of an animal to survive. For example, Al-Jahiz's wrote in his *Book of Animals*: "All animals, in short, can not exist without food, neither can the hunting animal escape being hunted in his turn. Every weak animal devours those weaker than itself. Strong animals cannot escape being devoured by other animals stronger than they".

Christian philosophy and the great chain of being



Drawing of the great chain of being from *Rhetorica Christiana* (1579) by Diego Valades

During the Early Middle Ages, Greek classical learning was all but lost to the West. However, contact with the Islamic world, where Greek manuscripts were preserved and expanded, soon led to a massive spate of Latin translations in the 12th century. Europeans were re-introduced to the works of Plato and Aristotle, as well as to Islamic thought. Christian thinkers of the scholastic school, in particular Abelard and Thomas Aquinas, combined Aristotelian classification with Plato's ideas of the goodness of God, and of all potential life forms being present in a perfect creation, to organize all inanimate, animate, and spiritual beings into a huge interconnected system: the *scala naturæ*, or great chain of being.

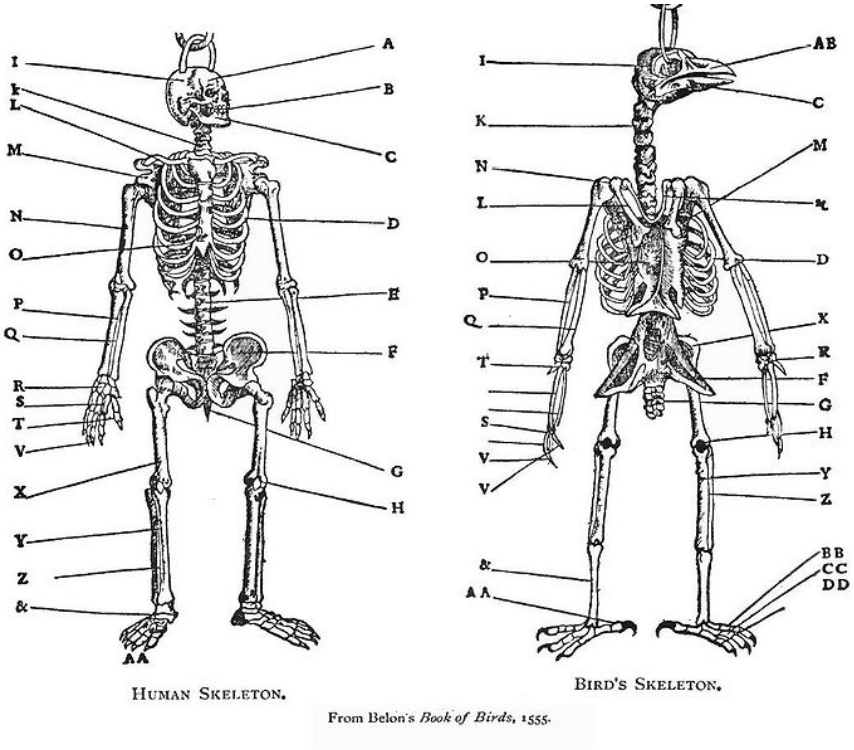
Within this system, everything that existed could be placed in order, from "lowest" to "highest", with Hell at the bottom and God at the top—below God, an angelic hierarchy marked by the orbits of the planets, mankind in an intermediate position, and worms the lowest of the animals. As the universe was ultimately perfect, the great chain was also perfect. There were no empty links in the chain, and no link was represented by more than one species. Therefore no species could ever move from one position to another. Thus, in this Christianized version of Plato's perfect universe, species could never change, but remained forever fixed, in accordance with the text of *Genesis*. For humans to forget their position was seen as sinful, whether they behaved like lower animals or aspired to a higher station than was given them by their Creator.

Creatures on adjacent steps were expected to closely resemble each other, an idea expressed in the saying: *natura non facit saltum* ("nature does not make leaps"). This basic concept of the great chain of being greatly influenced the thinking of Western civilization for centuries (and still has an influence today). It formed a part of the argument from design presented by natural theology. As a classification system, it became the major organizing principle and foundation of the emerging science of biology in the 17th and 18th centuries.

Thomas Aquinas on creation and natural processes

While the development of the great chain of being and the argument from design by Christian theologians contributed to the view that the natural world fit into an unchanging designed hierarchy, some theologians were more open to the possibility that the world might have developed through natural processes. Thomas Aquinas went even farther than Augustine of Hippo in arguing that scriptural texts like Genesis should not be interpreted in a literal way that conflicted with or constrained what natural philosophers learned about the workings of the natural world. He felt that the autonomy of nature was a sign of God's goodness and that there was no conflict between the concept of a divinely created universe, and the idea that the universe may have evolved over time through natural mechanisms. However, Aquinas disputed the views of those like the ancient Greek philosopher Empodocles who held that such natural processes showed that the universe could have developed without an underlying purpose. Rather holding that: "Hence, it is clear that nature is nothing but a certain kind of art, i.e., the divine art, impressed upon things, by which these things are moved to a determinate end. It is as if the shipbuilder were able to give to timbers that by which they would move themselves to take the form of a ship."

Renaissance and Enlightenment



Pierre Belon compared the skeletons of birds and humans in his *Book of Birds* (1555).

In the first half of the 17th century, René Descartes's mechanical philosophy encouraged the use of the metaphor of the universe as a machine, a concept that would come to characterise the scientific revolution. Between 1650 and 1800, some evolutionist theories supported the view that the universe, including life on Earth, had developed mechanically, entirely without divine guidance. In contrast, most contemporary theories of evolution, such of those of Gottfried Leibniz and J. G. Herder, held that evolution was a fundamentally *spiritual* process. In 1751, Pierre Louis Maupertuis veered toward more materialist ground. He wrote of natural modifications occurring during reproduction and accumulating over the course of many generations, producing races and even new species, a description that anticipated in general terms the concept of natural selection.

The word *evolution* (from the Latin *evolutio*, meaning "to unroll like a scroll") was initially used to refer to embryological development; its first use in relation to development of species came in 1762, when Charles Bonnet used it for his concept of "pre-formation", in which females carried a miniature form of all future generations. The term gradually gained a more general meaning of growth or progressive development.

Later in the 18th century, the French philosopher G. L. L. Buffon, one of the leading 18th century naturalists, suggested that what most people referred to as species were really just well-marked varieties, modified from an original form by environmental factors. For example, he believed that lions, tigers, leopards and house cats might all have a common ancestor. He further speculated that the 200 or so species of mammals then known might

have descended from as few as 38 original animal forms. Buffon's evolutionary ideas were limited; he believed each of the original forms had arisen through spontaneous generation and that each was shaped by "internal moulds" that limited the amount of change. Buffon's works, *Natural History* and *The Epochs of Nature*, containing well developed theories about a completely materialistic origin for the Earth and his ideas questioning the fixity of species, were extremely influential. Another French philosopher, Denis Diderot, also wrote that living things might have first arisen through spontaneous generation, and that species were always changing through a constant process of experiment where new forms arose and survived or not based on trial and error; an idea that can be considered a partial anticipation of natural selection. Between 1767 and 1792, James Burnett, Lord Monboddo included in his writings not only the concept that man had descended from primates, but also that, in response to the environment, creatures had found methods of transforming their characteristics over long time intervals. Charles Darwin's grandfather, Erasmus Darwin, published *Zoönomia* in 1796, which suggested that "all warm-blooded animals have arisen from one living filament". In his 1802 poem *Temple of Nature*, he described the rise of life from minute organisms living in mud to all of its modern diversity.

Early 19th century

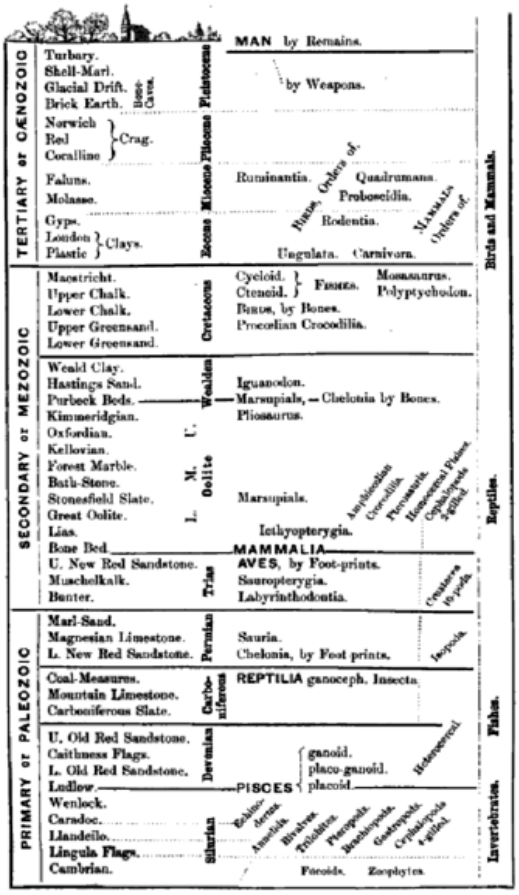


Diagram of the geologic timescale from an 1861 book by Richard Owen showing the appearance of major animal types

Paleontology and geology

In 1796, Georges Cuvier published his findings on the differences between living elephants and those found in the fossil record. His analysis demonstrated that mammoths and mastodons were distinct species, different from any living animal, effectively ending a long-running debate over whether the extinction of a species was possible. In 1788, James Hutton described gradual geological processes operating continuously over deep time. In the 1790s William Smith began the process of ordering rock strata by examining fossils in the layers while he worked on his geologic map of England. Independently, in 1811, Georges Cuvier and Alexandre Brongniart published an influential study of the geologic history of the region around Paris, based on the stratigraphic succession of rock layers. These works helped establish the antiquity of the Earth. Cuvier advocated catastrophism to explain the patterns of extinction and faunal succession revealed by the fossil record.

Knowledge of the fossil record continued to advance rapidly during the first few decades of the 19th century. By the 1840s, the outlines of the geologic timescale were becoming clear, and in 1841 John Phillips named three major eras, based on the predominant fauna of each: the Paleozoic, dominated by marine invertebrates and fish, the Mesozoic, the age of reptiles, and the current Cenozoic age of mammals. This progressive picture of the history of life was accepted even by conservative English geologists like Adam Sedgwick and William Buckland; however, like Cuvier, they attributed the progression to repeated catastrophic episodes of extinction followed by new episodes of creation. Unlike Cuvier, Buckland and some other advocates of natural theology among British geologists made efforts to explicitly link the last catastrophic episode proposed by Cuvier to the biblical flood.

From 1830 to 1833, Charles Lyell published his multi-volume work *Principles of Geology*, which, building on Hutton's ideas, advocated a uniformitarian alternative to the catastrophic theory of geology. Lyell claimed that, rather than being the products of cataclysmic (and possibly supernatural) events, the geologic features of the Earth are better explained as the result of the same gradual geologic forces observable in the present day—but acting over immensely long periods of time. Although Lyell opposed evolutionary ideas (even questioning the consensus that the fossil record demonstrates a true progression), his concept that the Earth was shaped by forces working gradually over an extended period, and the immense age of the Earth assumed by his theories, would strongly influence future evolutionary thinkers such as Charles Darwin.

Transmutation of species

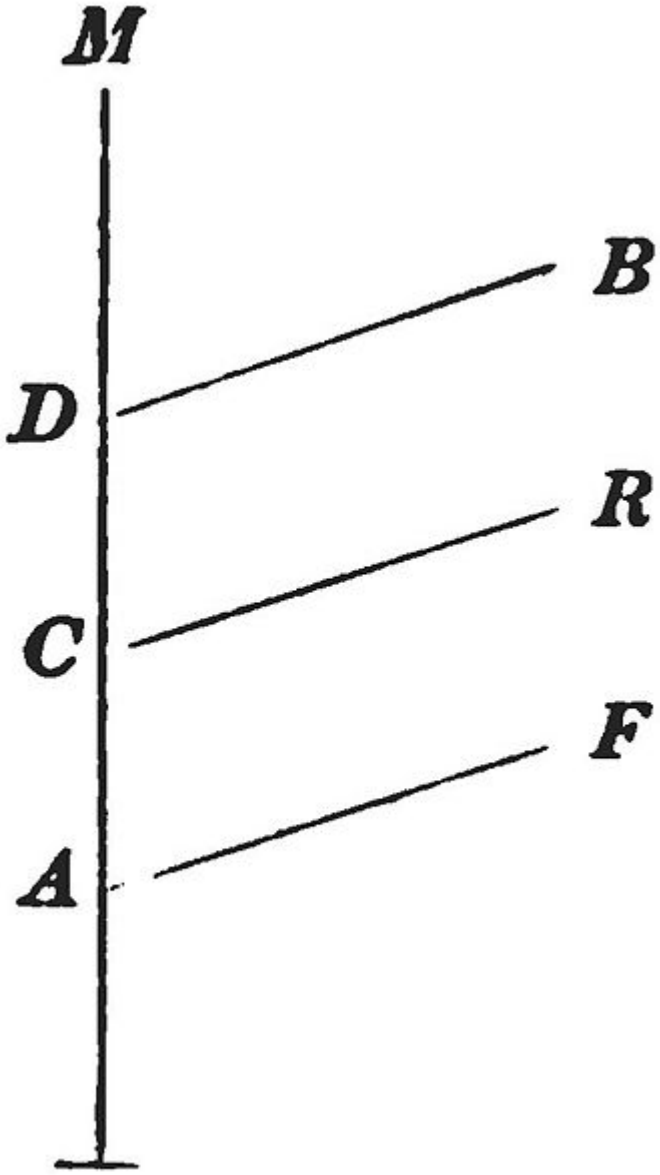


Diagram from *Vestiges of the Natural History of Creation* (1844) by Robert Chambers shows a model of development where fish (F), reptiles (R), and birds (B) represent branches from a path leading to mammals (M).

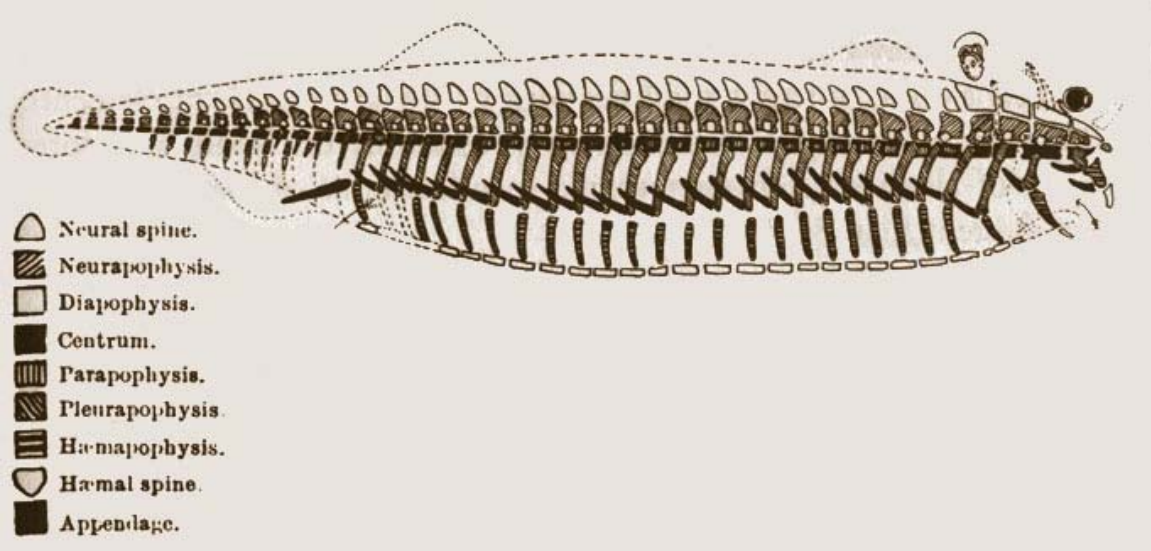
Jean-Baptiste Lamarck proposed, in his *Philosophie Zoologique* of 1809, a theory of the transmutation of species. Lamarck did not believe that all living things shared a common ancestor but rather that simple forms of life were created continuously by spontaneous generation. He also believed that an innate life force drove species to become more complex over time, advancing up a linear ladder of complexity that was related to the great chain of being. Lamarck recognized that species were adapted to their environment. He explained this by saying that the same innate force driving increasing complexity caused the organs of an animal (or a plant) to change based on the use or disuse of those

organs, just as muscles are affected by exercise. He argued that these changes would be inherited by the next generation and produce slow adaptation to the environment. It was this secondary mechanism of adaptation through the inheritance of acquired characteristics that would become known as Lamarckism and would influence discussions of evolution into the 20th century.

A radical British school of comparative anatomy that included the anatomist Robert Grant was closely in touch with Lamarck's French school of *Transformationism*. One of the French scientists who influenced Grant was the anatomist Étienne Geoffroy Saint-Hilaire, whose ideas on the unity of various animal body plans and the homology of certain anatomical structures would be widely influential and lead to intense debate with his colleague Georges Cuvier. Grant became an authority on the anatomy and reproduction of marine invertebrates. He developed Lamarck's and Erasmus Darwin's ideas of transmutation and evolutionism, and investigated homology, even proposing that plants and animals had a common evolutionary starting point. As a young student Charles Darwin joined Grant in investigations of the life cycle of marine animals. In 1826 an anonymous paper, probably written by Robert Jameson, praised Lamarck for explaining how higher animals had "evolved" from the simplest worms; this was the first use of the word "evolved" in a modern sense.

In 1844, the Scottish publisher Robert Chambers anonymously published an extremely controversial but widely read book entitled *Vestiges of the Natural History of Creation*. This book proposed an evolutionary scenario for the origins of the Solar System and life on Earth. It claimed that the fossil record showed a progressive ascent of animals with current animals being branches off a main line that leads progressively to humanity. It implied that the transmutations lead to the unfolding of a preordained plan that had been woven into the laws that governed the universe. In this sense it was less completely materialistic than the ideas of radicals like Robert Grant, but its implication that humans were only the last step in the ascent of animal life incensed many conservative thinkers. The high profile of the public debate over *Vestiges*, with its depiction of evolution as a progressive process, would greatly influence the perception of Darwin's theory a decade later.

Ideas about the transmutation of species were associated with the radical materialism of the Enlightenment and were attacked by more conservative thinkers. Georges Cuvier attacked the ideas of Lamarck and Geoffroy Saint-Hilaire, agreeing with Aristotle that species were immutable. Cuvier believed that the individual parts of an animal were too closely correlated with one another to allow for one part of the anatomy to change in isolation from the others, and argued that the fossil record showed patterns of catastrophic extinctions followed by re-population, rather than gradual change over time. He also noted that drawings of animals and animal mummies from Egypt, which were thousands of years old, showed no signs of change when compared with modern animals. The strength of Cuvier's arguments and his scientific reputation helped keep transmutational ideas out of the mainstream for decades.



This 1847 diagram by Richard Owen shows his conceptual archetype for all vertebrates.

In Britain the philosophy of natural theology remained influential. William Paley's 1802 book *Natural Theology* with its famous watchmaker analogy had been written at least in part as a response to the transmutational ideas of Erasmus Darwin. Geologists influenced by natural theology, such as Buckland and Sedgwick, made a regular practice of attacking the evolutionary ideas of Lamarck, Grant, and *The Vestiges of the Natural History of Creation*. Although the geologist Charles Lyell opposed scriptural geology, he also believed in the immutability of species, and in his *Principles of Geology* (1830–1833), he criticized Lamarck's theories of development. Idealists such as Louis Agassiz and Richard Owen believed that each species was fixed and unchangeable because it represented an idea in the mind of the creator. They believed that relationships between species could be discerned from developmental patterns in embryology, as well as in the fossil record, but that these relationships represented an underlying pattern of divine thought, with progressive creation leading to increasing complexity and culminating in humanity. Owen developed the idea of "archetypes" in the Divine mind that would produce a sequence of species related by anatomical homologies, such as vertebrate limbs. Owen led a public campaign that successfully marginalized Robert Grant in the scientific community. Darwin would make good use of the homologies analyzed by Owen in his own theory, but the harsh treatment of Grant, and the controversy surrounding *Vestiges*, showed him the need to ensure that his own ideas were scientifically sound.

Anticipations of natural selection

Several writers anticipated aspects of Darwin's theory, and in the third edition of *On the Origin of Species* published in 1861 Darwin named those he knew about in an introductory appendix, *An Historical Sketch of the Recent Progress of Opinion on the Origin of Species*, which he expanded in later editions.

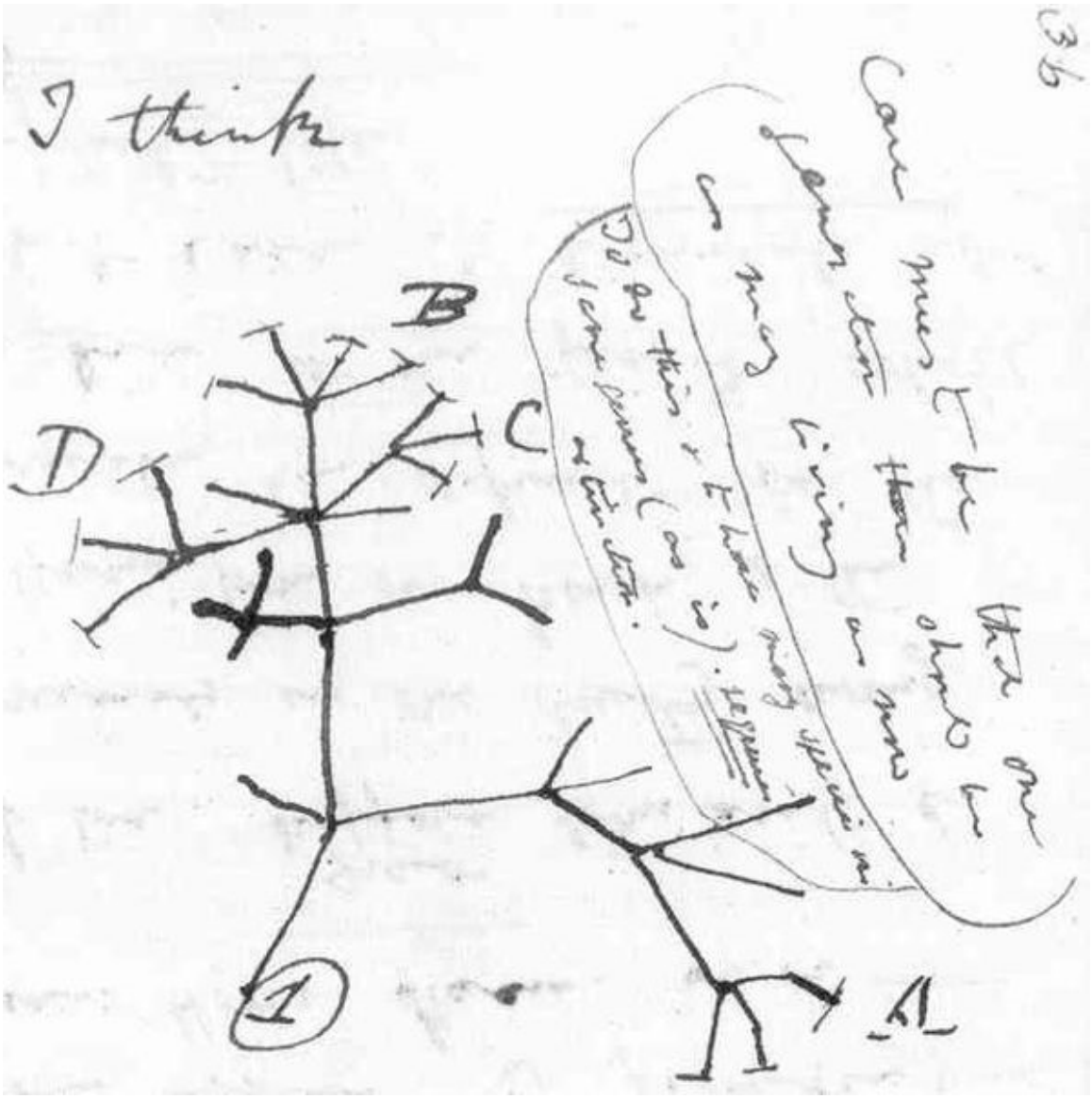
In 1813, William Charles Wells read before the Royal Society essays assuming that there had been evolution of humans, and recognising the principle of natural selection. Charles Darwin and Alfred Russel Wallace were unaware of this work when they jointly published the theory in 1858, but Darwin later acknowledged that Wells had recognised the principle before them, writing that the paper "An Account of a White Female, part of whose Skin resembles that of a Negro" was published in 1818, and "he distinctly recognises the principle of natural selection, and this is the first recognition which has been indicated; but he applies it only to the races of man, and to certain characters alone." When Darwin was developing his theory, he was influenced by Augustin de Candolle's *natural system* of classification, which laid emphasis on the war between competing species.

Patrick Matthew wrote in the obscure book *Naval Timber & Arboriculture* (1831) of "continual balancing of life to circumstance. ... [The] progeny of the same parents, under great differences of circumstance, might, in several generations, even become distinct species, incapable of co-reproduction." Charles Darwin discovered this work after the initial publication of the *Origin*. In the brief historical sketch that Darwin included in the 3rd edition he says "Unfortunately the view was given by Mr. Matthew very briefly in an Appendix to a work on a different subject ... He clearly saw, however, the full force of the principle of natural selection."

It is possible to look through the history of biology from the ancient Greeks onwards and discover anticipations of almost all of Darwin's key ideas. However, as historian of science Peter J. Bowler says, "Through a combination of bold theorizing and comprehensive evaluation, Darwin came up with a concept of evolution that was unique for the time." Bowler goes on to say that simple priority alone is not enough to secure a place in the history of science; someone has to develop an idea and convince others of its importance to have a real impact.

T. H. Huxley said in his essay on the reception of the *Origin of Species*:

The suggestion that new species may result from the selective action of external conditions upon the variations from their specific type which individuals present and which we call spontaneous because we are ignorant of their causation is as wholly unknown to the historian of scientific ideas as it was to biological specialists before 1858. But that suggestion is the central idea of the *Origin of Species*, and contains the quintessence of Darwinism.



Darwin's first sketch of an evolutionary tree from his *First Notebook on Transmutation of Species* (1837)

Natural selection

The biogeographical patterns Charles Darwin observed in places such as the Galapagos islands during the voyage of the Beagle caused him to doubt the fixity of species, and in 1837 Darwin started the first of a series of secret notebooks on transmutation. Darwin's observations led him to view transmutation as a process of divergence and branching, rather than the ladder-like progression envisioned by Lamarck and others. In 1838 he read the new 6th edition of *An Essay on the Principle of Population*, written in the late 18th century by Thomas Malthus. Malthus' idea of population growth leading to a struggle for survival combined with Darwin's knowledge on how breeders selected traits, led to the inception of Darwin's theory of natural selection. Darwin did not publish his

ideas on evolution for 20 years. However he did share them with certain other naturalists and friends, starting with Joseph Hooker, with whom he discussed his unpublished 1844 essay on natural selection. During this period he used the time he could spare from his other scientific work to slowly refine his ideas and, aware of the intense controversy around transmutation, amass evidence to support them. In September 1854 he began full time work on writing his book on natural selection.

Unlike Darwin, Alfred Russel Wallace, influenced by the book *Vestiges of the Natural History of Creation*, already suspected that transmutation of species occurred when he began his career as a naturalist. By 1855 his biogeographical observations during his field work in South America and the Malay Archipelago made him confident enough in a branching pattern of evolution to publish a paper stating that every species originated in close proximity to an already existing closely allied species. Like Darwin, it was Wallace's consideration of how the ideas of Malthus might apply to animal populations that led him to conclusions very similar to those reached by Darwin about the role of natural selection. In February 1858 Wallace, unaware of Darwin's unpublished ideas, composed his thoughts into an essay and mailed them to Darwin, asking for his opinion. The result was the joint publication in July of an extract from Darwin's 1844 essay along with Wallace's letter. Darwin also began work on a short abstract summarising his theory, which he would publish in 1859 as *On the Origin of Species*.

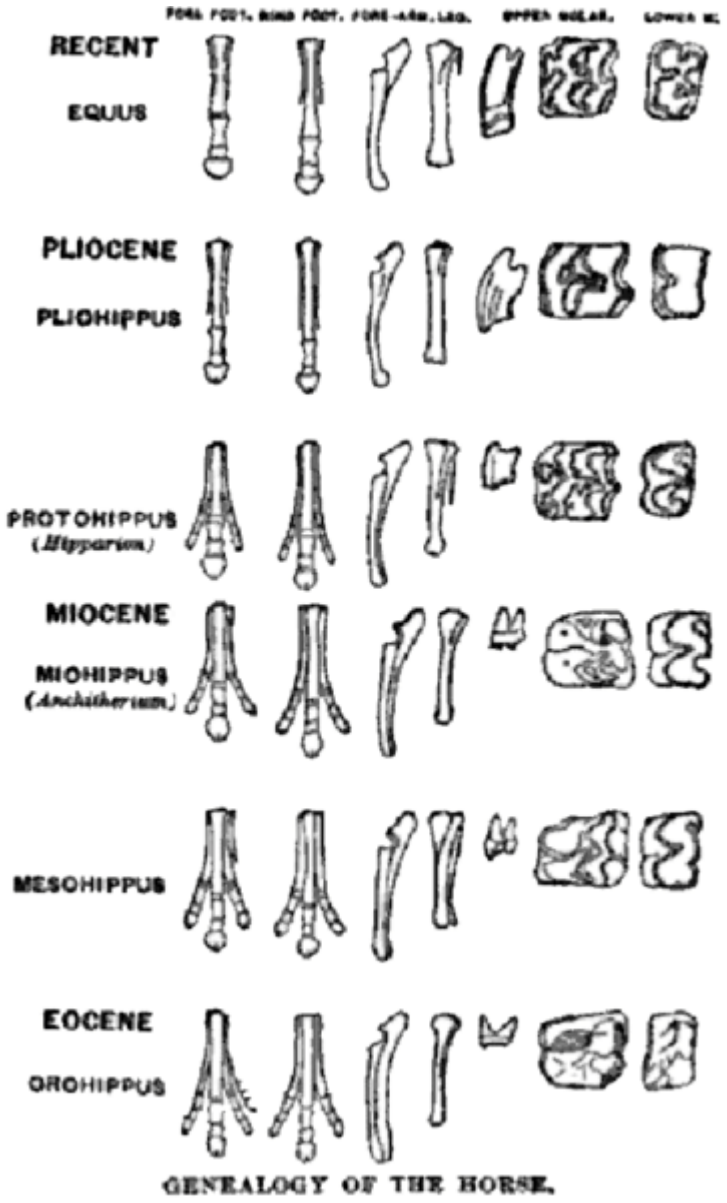


Diagram by O.C. Marsh of the evolution of horse feet and teeth over time as reproduced in T.H Huxley's 1876 book *Professor Huxley in America*

1859–1930s: Darwin and his legacy

By the 1850s, whether or not species evolved was a subject of intense debate, with prominent scientists arguing both sides of the issue. The publication of Charles Darwin's *On the Origin of Species* (1859) fundamentally transformed the discussion over biological origins. Darwin argued that his branching version of evolution explained a wealth of facts in biogeography, anatomy, embryology, and other fields of biology. He also provided the first cogent mechanism by which evolutionary change could persist: his theory of natural selection.

One of the first and most important naturalists to be convinced by *Origin* of the reality of evolution was the British anatomist Thomas Henry Huxley. Huxley recognized that unlike the earlier transmutational ideas of Lamarck and *Vestiges*, Darwin's theory provided a mechanism for evolution without supernatural involvement, even if Huxley himself was not completely convinced that natural selection was the key evolutionary mechanism. Huxley would make advocacy of evolution a cornerstone of the program of the X Club to reform and professionalise science by displacing natural theology with naturalism and to end the domination of British natural science by the clergy. By the early 1870s in English-speaking countries, thanks partly to these efforts, evolution had become the mainstream scientific explanation for the origin of species. In his campaign for public and scientific acceptance of Darwin's theory, Huxley made extensive use of new evidence for evolution from paleontology. This included evidence that birds had evolved from reptiles, including the discovery of *Archaeopteryx* in Europe, and a number of fossils of primitive birds with teeth found in North America. Another important line of evidence was the finding of fossils that helped trace the evolution of the horse from its small five-toed ancestors. However, acceptance of evolution among scientists in non-English speaking nations such as France, and the countries of southern Europe and Latin America was slower. An exception to this was Germany, where both August Weismann and Ernst Haeckel championed this idea: Haeckel used evolution to challenge the established tradition of metaphysical idealism in German biology, much as Huxley used it to challenge natural theology in Britain. Haeckel and other German scientists would take the lead in launching an ambitious programme to reconstruct the evolutionary history of life based on morphology and embryology.

Darwin's theory succeeded in profoundly altering scientific opinion regarding the development of life and in producing a small philosophical revolution. However, this theory could not explain several critical components of the evolutionary process. Specifically, Darwin was unable to explain the source of variation in traits within a species, and could not identify a mechanism that could pass traits faithfully from one generation to the next. Darwin's hypothesis of pangenesis, while relying in part on the inheritance of acquired characteristics, proved to be useful for statistical models of evolution that were developed by his cousin Francis Galton and the "biometric" school of evolutionary thought. However, this idea proved to be of little use to other biologists.

Application to humans

Photographically reduced from Diagrams of the natural size (except that of the Gibbon, which was twice as large as nature), drawn by Mr. Waterhouse Hawkins from specimens in the Museum of the Royal College of Surgeons.

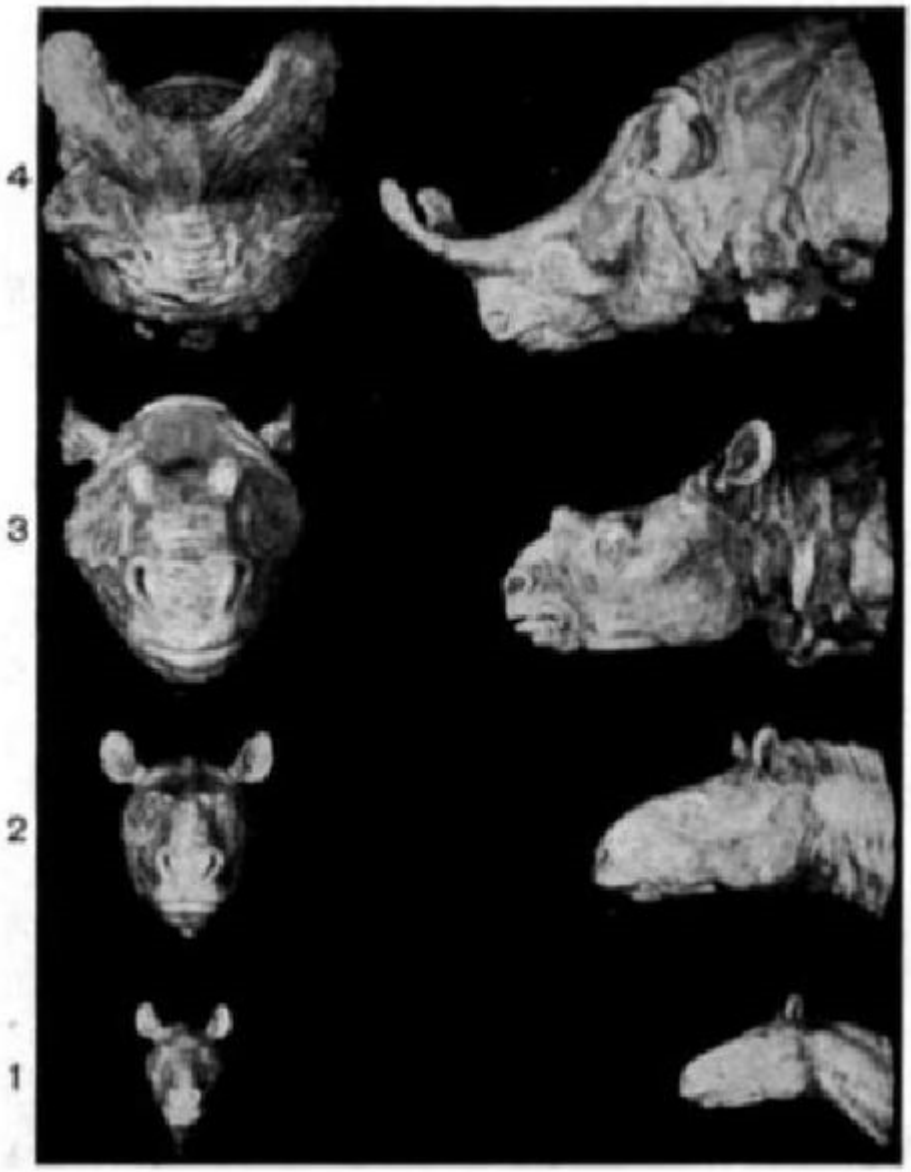
This illustration was the frontispiece of Thomas Henry Huxley's book *Evidence as to Man's Place in Nature* (1863). Huxley applied Darwins ideas to humans, using comparative anatomy to show that humans and apes had a common ancestor, which challenged the theologically important idea that humans held a unique place in the unniverse.

Charles Darwin was aware of the severe reaction in some parts of the scientific community against the suggestion made in *Vestiges of the Natural History of Creation* that humans had arisen from animals by a process of transmutation. Therefore he almost completely ignored the topic of human evolution in *The Origin of Species*. Despite this precaution, the issue featured prominently in the debate that followed the book's publication. For most of the first half of the 19th century, the scientific community believed that, although geology had shown that the Earth and life were very old, human beings had appeared suddenly just a few thousand years before the present. However, a series of archaeological discoveries in the 1840s and 1850s showed stone tools associated with the remains of extinct animals. By the early 1860s, as summarized in Charles Lyell's 1863 book *Geological Evidences of the Antiquity of Man*, it had become widely accepted that humans had existed during a prehistoric period – which stretched many thousands of years before the start of written history. This view of human history was more compatible with an evolutionary origin for humanity than was the older view. On the other hand, at that time there was no fossil evidence to demonstrate human evolution. The only human fossils found before the discovery of Java man in the 1890s were either of anatomically modern humans or of Neanderthals that were too close, especially in the critical characteristic of cranial capacity, to modern humans for them to be convincing intermediates between humans and other primates.

Therefore the debate that immediately followed the publication of *The Origin of Species* centered on the similarities and differences between humans and modern apes. Carolus Linnaeus had been criticised in the 18th century for grouping humans and apes together as primates in his ground breaking classification system. Richard Owen vigorously defended the classification suggested by Cuvier and Johann Friedrich Blumenbach that placed humans in a separate order from any of the other mammals, which by the early 19th century had become the orthodox view. On the other hand, Thomas Henry Huxley sought to demonstrate a close anatomical relationship between humans and apes. In one famous incident, Huxley showed that Owen was mistaken in claiming that the brains of gorillas lacked a structure present in human brains. Huxley summarized his argument in his highly influential 1863 book *Evidence as to Man's Place in Nature*. Another viewpoint was advocated by Charles Lyell and Alfred Russel Wallace. They agreed that humans shared a common ancestor with apes, but questioned whether any purely materialistic mechanism could account for all the differences between humans and apes, especially some aspects of the human mind.

In 1871, Darwin published *The Descent of Man, and Selection in Relation to Sex*, which contained his views on human evolution. Darwin argued that the differences between the human mind and the minds of the higher animals were a matter of degree rather than of kind. For example, he viewed morality as a natural outgrowth of instincts that were beneficial to animals living in social groups. He argued that all the differences between humans and apes were explained by a combination of the selective pressures that came from our ancestors moving from the trees to the plains, and sexual selection. The debate over human origins, and over the degree of human uniqueness continued well into the 20th century.

Alternatives to natural selection



This photo from Henry Fairfield Osborn's 1918 book *Origin and Evolution of Life* shows models depicting the evolution of Titanothere horns over time, which Osborn claimed was an example of an orthogenic trend in evolution.

The concept of evolution was widely accepted in scientific circles within a few years of the publication of *Origin*, but the acceptance of natural selection as its driving mechanism was much less widespread. The four major alternatives to natural selection in the late 19th century were theistic evolution, neo-Lamarckism, orthogenesis, and saltationism. Theistic evolution (a term promoted by Darwin's greatest American advocate Asa Gray) was the idea that God intervened in the process of evolution to guide it in such a way that the living world could still be considered to be designed. However, this idea gradually fell out of favor among scientists, as they became more and more committed to the idea

of methodological naturalism and came to believe that direct appeals to supernatural involvement were scientifically unproductive. By 1900, theistic evolution had largely disappeared from professional scientific discussions, although it retained a strong popular following.

In the late 19th century, the term neo-Lamarckism came to be associated with the position of naturalists who viewed the inheritance of acquired characteristics as the most important evolutionary mechanism. Advocates of this position included the British writer and Darwin critic Samuel Butler, the German biologist Ernst Haeckel, and the American paleontologist Edward Drinker Cope. They considered Lamarckism to be philosophically superior to Darwin's idea of selection acting on random variation. Cope looked for, and thought he found, patterns of linear progression in the fossil record. Inheritance of acquired characteristics was part of Haeckel's recapitulation theory of evolution, which held that the embryological development of an organism repeats its evolutionary history. Critics of neo-Lamarckism, such as the German biologist August Weismann and Alfred Russel Wallace, pointed out that no one had ever produced solid evidence for the inheritance of acquired characteristics. Despite these criticisms, neo-Lamarckism remained the most popular alternative to natural selection at the end of the 19th century, and would remain the position of some naturalists well into the 20th century.

Orthogenesis was the hypothesis that life has an innate tendency to change, in a unilinear fashion, towards ever-greater perfection. It had a significant following in the 19th century, and its proponents included the Russian biologist Leo Berg and the American paleontologist Henry Fairfield Osborn. Orthogenesis was popular among some paleontologists, who believed that the fossil record showed a gradual and constant unidirectional change. Saltationism was the idea that new species arise as a result of large mutations. It was seen as a much faster alternative to the Darwinian concept of a gradual process of small random variations being acted on by natural selection, and was popular with early geneticists such as Hugo de Vries, William Bateson, and early in his career, T. H. Morgan. It became the basis of the mutation theory of evolution.

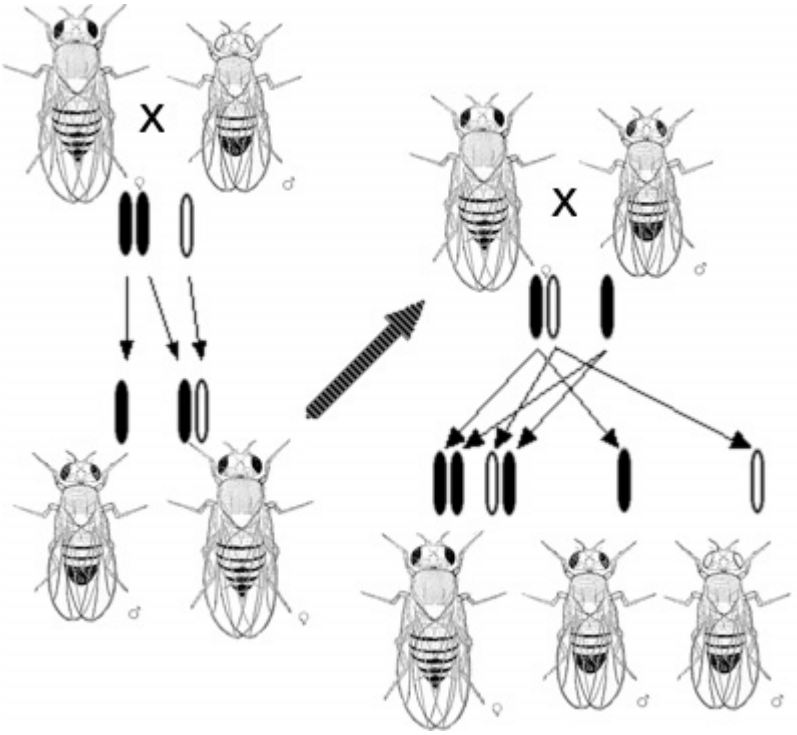


Diagram from T.H. Morgan's 1919 book *The Physical Basis of Heredity*, showing the sex-linked inheritance of the white-eyed mutation in *Drosophila melanogaster*

Mendelian genetics, biometrics, and mutation

The so-called rediscovery of Gregor Mendel's laws of inheritance in 1900 ignited a fierce debate between two camps of biologists. In one camp were the Mendelians, who were focused on discrete variations and the laws of inheritance. They were led by William Bateson (who coined the word *genetics*) and Hugo de Vries (who coined the word *mutation*). Their opponents were the biometricians, who were interested in the continuous variation of characteristics within populations. Their leaders, Karl Pearson and Walter Frank Raphael Weldon, followed in the tradition of Francis Galton, who had focused on measurement and statistical analysis of variation within a population. The biometricians rejected Mendelian genetics on the basis that discrete units of heredity, such as genes, could not explain the continuous range of variation seen in real populations. Weldon's work with crabs and snails provided evidence that selection pressure from the environment could shift the range of variation in wild populations, but the Mendelians maintained that the variations measured by biometricians were too insignificant to account for the evolution of new species.

When T. H. Morgan began experimenting with breeding the fruit fly *Drosophila melanogaster*, he was a saltationist who hoped to demonstrate that a new species could be created in the lab by mutation alone. Instead, the work at his lab between 1910 and 1915 reconfirmed Mendelian genetics and provided solid experimental evidence linking it to chromosomal inheritance. His work also demonstrated that most mutations had relatively

small effects, such as a change in eye color, and that rather than creating a new species in a single step, mutations served to increase variation within the existing population.

1920s–1940s

Biston betularia f. typica is the white-bodied form of the peppered moth.

Biston betularia f. carbonaria is the black-bodied form of the peppered moth.

Population genetics

The Mendelian and biometrician models were eventually reconciled with the development of population genetics. A key step was the work of the British biologist and statistician R.A. Fisher. In a series of papers starting in 1918 and culminating in his 1930 book *The Genetical Theory of Natural Selection*, Fisher showed that the continuous variation measured by the biometricians could be produced by the combined action of many discrete genes, and that natural selection could change gene frequencies in a population, resulting in evolution. In a series of papers beginning in 1924, another British geneticist, J.B.S. Haldane, applied statistical analysis to real-world examples of natural selection, such as the evolution of industrial melanism in peppered moths, and showed that natural selection worked at an even faster rate than Fisher assumed.

The American biologist Sewall Wright, who had a background in animal breeding experiments, focused on combinations of interacting genes, and the effects of inbreeding on small, relatively isolated populations that exhibited genetic drift. In 1932, Wright introduced the concept of an adaptive landscape and argued that genetic drift and inbreeding could drive a small, isolated sub-population away from an adaptive peak, allowing natural selection to drive it towards different adaptive peaks. The work of Fisher, Haldane and Wright founded the discipline of population genetics. This integrated natural selection with Mendelian genetics, which was the critical first step in developing a unified theory of how evolution worked.

Modern evolutionary synthesis

In the first few decades of the 20th century, most field naturalists continued to believe that Lamarckian and orthogenic mechanisms of evolution provided the best explanation for the complexity they observed in the living world. But as the field of genetics continued to develop, those views became less tenable. Theodosius Dobzhansky, a postdoctoral worker in T. H. Morgan's lab, had been influenced by the work on genetic diversity by Russian geneticists such as Sergei Chetverikov. He helped to bridge the divide between the foundations of microevolution developed by the population geneticists and the patterns of macroevolution observed by field biologists, with his 1937 book *Genetics and the Origin of Species*. Dobzhansky examined the genetic diversity of wild populations and showed that, contrary to the assumptions of the population geneticists, these populations had large amounts of genetic diversity, with marked differences between sub-populations. The book also took the highly mathematical work of the population geneticists and put it into a more accessible form. In Great Britain E.B. Ford, the pioneer of ecological genetics, continued throughout the 1930s and 1940s to demonstrate the power of selection due to ecological factors including the ability to maintain genetic diversity through genetic polymorphisms such as human blood types. Ford's work would contribute to a shift in emphasis during the course of the modern synthesis towards natural selection over genetic drift.

Evolutionary biologist Ernst Mayr was influenced by the work of the German biologist Bernhard Rensch showing the influence of local environmental factors on the geographic distribution of sub-species and closely related species. Mayr followed up on Dobzhansky's work with the 1942 book *Systematics and the Origin of Species*, which emphasized the importance of allopatric speciation in the formation of new species. This form of speciation occurs when the geographical isolation of a sub-population is followed by the development of mechanisms for reproductive isolation. Mayr also formulated the biological species concept that defined a species as a group of interbreeding or potentially interbreeding populations that were reproductively isolated from all other populations.

In the 1944 book *Tempo and Mode in Evolution*, George Gaylord Simpson showed that the fossil record was consistent with the irregular non-directional pattern predicted by the developing evolutionary synthesis, and that the linear trends that earlier paleontologists had claimed supported orthogenesis and neo-Lamarckism did not hold up to closer examination. In 1950, G. Ledyard Stebbins published *Variation and Evolution in Plants*, which helped to integrate botany into the synthesis. The emerging cross-disciplinary consensus on the workings of evolution would be known as the modern evolutionary synthesis. It received its name from the book *Evolution: The Modern Synthesis* by Julian Huxley.

The evolutionary synthesis provided a conceptual core—in particular, natural selection and Mendelian population genetics—that tied together many, but not all, biological disciplines. It helped establish the legitimacy of evolutionary biology, a primarily historical science, in a scientific climate that favored experimental methods over historical ones. The synthesis also resulted in a considerable narrowing of the range of

mainstream evolutionary thought (what Stephen Jay Gould called the "hardening of the synthesis"): by the 1950s, natural selection acting on genetic variation was virtually the only acceptable mechanism of evolutionary change (panselectionism), and macroevolution was simply considered the result of extensive microevolution.

1940s–1960s: Molecular biology and evolution

The middle decades of the 20th century saw the rise of molecular biology, and with it an understanding of the chemical nature of genes as sequences of DNA and their relationship, through the genetic code, to protein sequences. At the same time, increasingly powerful techniques for analyzing proteins, such as protein electrophoresis and sequencing, brought biochemical phenomena into realm of the synthetic theory of evolution. In the early 1960s, biochemists Linus Pauling and Emile Zuckerkandl proposed the molecular clock hypothesis: that sequence differences between homologous proteins could be used to calculate the time since two species diverged. By 1969, Motoo Kimura and others provided a theoretical basis for the molecular clock, arguing that—at the molecular level at least—most genetic mutations are neither harmful nor helpful and that genetic drift (rather than natural selection) causes a large portion of genetic change: the neutral theory of molecular evolution. Studies of protein differences *within* species also brought molecular data to bear on population genetics by providing estimates of the level of heterozygosity in natural populations.

From the early 1960s, molecular biology was increasingly seen as a threat to the traditional core of evolutionary biology. Established evolutionary biologists—particularly Ernst Mayr, Theodosius Dobzhansky and G. G. Simpson, three of the architects of the modern synthesis—were extremely skeptical of molecular approaches, especially when it came to the connection (or lack thereof) to natural selection. The molecular-clock hypothesis and the neutral theory were particularly controversial, spawning the neutralist-selectionist debate over the relative importance of drift and selection, which continued into the 1980s without a clear resolution.

Late 20th century

Gene-centered view

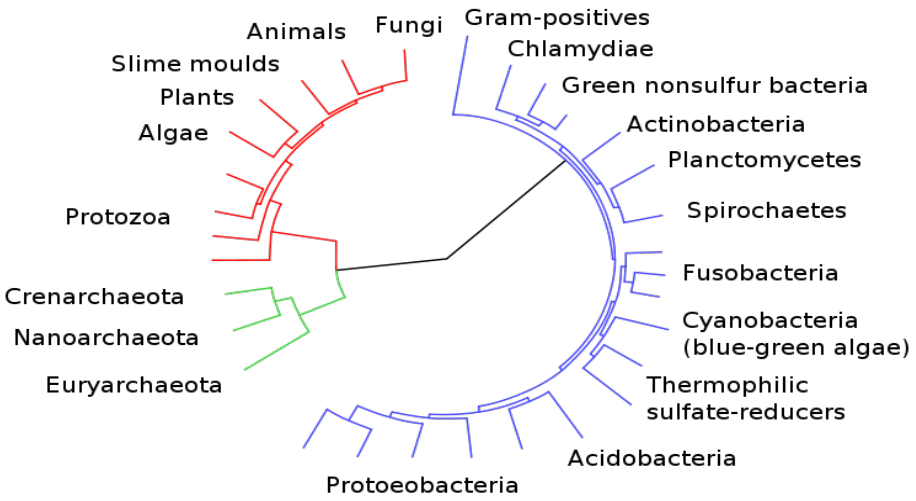
In the mid-1960s, George C. Williams strongly critiqued explanations of adaptations worded in terms of "survival of the species" (group selection arguments). Such explanations were largely replaced by a gene-centered view of evolution, epitomized by the kin selection arguments of W. D. Hamilton, George R. Price and John Maynard Smith. This viewpoint would be summarized and popularized in the influential 1976 book *The Selfish Gene* by Richard Dawkins. Models of the period showed that group selection was severely limited in its strength; though newer models do admit the possibility of significant multi-level selection.

In 1973, Leigh Van Valen proposed the term "Red Queen", which he took from *Through the Looking-Glass* by Lewis Carroll, to describe a scenario where a species involved in

one or more evolutionary arms races would have to constantly change just to keep pace with the species with which it was co-evolving. Hamilton, Williams and others suggested that this idea might explain the evolution of sexual reproduction: the increased genetic diversity caused by sexual reproduction would help maintain resistance against rapidly evolving parasites, thus making sexual reproduction common, despite the tremendous cost from the gene-centric point of view of a system where only half of an organism's genome is passed on during reproduction. The gene-centric view has also led to an increased interest in Darwin's old idea of sexual selection, and more recently in topics such as sexual conflict and intragenomic conflict.

Sociobiology

W. D. Hamilton's work on kin selection contributed to the emergence of the discipline of sociobiology. The existence of altruistic behaviors has been a difficult problem for evolutionary theorists from the beginning. Significant progress was made in 1964 when Hamilton formulated the inequality in kin selection known as Hamilton's rule, which showed how eusociality in insects (the existence of sterile worker classes) and many other examples of altruistic behavior could have evolved through kin selection. Other theories followed, some derived from game theory, such as reciprocal altruism. In 1975, E.O. Wilson published the influential and highly controversial book *Sociobiology: The New Synthesis* which claimed evolutionary theory could help explain many aspects of animal, including human, behavior. Critics of sociobiology, including Stephen Jay Gould and Richard Lewontin, claimed that sociobiology greatly overstated the degree to which complex human behaviors could be determined by genetic factors. They also claimed that the theories of sociobiologists often reflected their own ideological biases. Despite these criticisms, work has continued in sociobiology and the related discipline of evolutionary psychology, including work on other aspects of the altruism problem.



A phylogenetic tree showing the three-domain system. Eukaryotes are colored red, Archaea green, and Bacteria blue.

Evolutionary paths and processes

One of the most prominent debates arising during the 1970s was over the theory of punctuated equilibrium. Niles Eldredge and Stephen Jay Gould proposed that there was a pattern of fossil species that remained largely unchanged for long periods (what they termed *stasis*), interspersed with relatively brief periods of rapid change during speciation. Improvements in sequencing methods resulted in a large increase of sequenced genomes, allowing the testing and refining of evolutionary theories using this huge amount of genome data. Comparisons between these genomes provide insights into the molecular mechanisms of speciation and adaptation. These genomic analyses have produced fundamental changes in the understanding of the evolutionary history of life, such as the proposal of the three-domain system by Carl Woese. Advances in computational hardware and software allow the testing and extrapolation of increasingly advanced evolutionary models and the development of the field of systems biology. One of the results has been an exchange of ideas between theories of biological evolution and the field of computer science known as evolutionary computation, which attempts to mimic biological evolution for the purpose of developing new computer algorithms. Discoveries in biotechnology now allow the modification of entire genomes, advancing evolutionary studies to the level where future experiments may involve the creation of entirely synthetic organisms.

Microbiology, horizontal gene transfer, and endosymbiosis

Microbiology was largely ignored by early evolutionary theory. This was due to the paucity of morphological traits and the lack of a species concept in microbiology, particularly amongst prokaryotes. Now, evolutionary researchers are taking advantage of their improved understanding of microbial physiology and ecology, produced by the comparative ease of microbial genomics, to explore the taxonomy and evolution of these organisms. These studies are revealing unanticipated levels of diversity amongst microbes.

One particularly important outcome from studies on microbial evolution was the discovery in Japan of horizontal gene transfer in 1959. This transfer of genetic material between different species of bacteria was first recognized because it played a major role in the spread of antibiotic resistance. More recently, as knowledge of genomes has continued to expand, it has been suggested that lateral transfer of genetic material has played an important role in the evolution of all organisms. These high levels of horizontal gene transfer have led to suggestions that the family tree of today's organisms, the so-called "tree of life", is more similar to an interconnected web or net.

Indeed, as part of the endosymbiotic theory for the origin of organelles, a form of horizontal gene transfer has been a critical step in the evolution of eukaryotes such as fungi, plants, and animals. The endosymbiotic theory holds that organelles within the cells of eukorytes such as mitochondria and chloroplasts, had descended from independent bacteria that came to live symbiotically within other cells. It had been suggested in the late 19th century when similarities between mitochondria and bacteria

were noted, but largely dismissed until it was revived and championed by Lynn Margulis in the 1960s and 70s; Margulis was able to make use of new evidence that such organelles had their own DNA that was inherited independently from that in the cell's nucleus.

Evolutionary developmental biology

In the 1980s and 1990s the tenets of the modern evolutionary synthesis came under increasing scrutiny. There was a renewal of structuralist themes in evolutionary biology in the work of biologists such as Brian Goodwin and Stuart Kauffman, which incorporated ideas from cybernetics and systems theory, and emphasized the self-organizing processes of development as factors directing the course of evolution. The evolutionary biologist Stephen Jay Gould revived earlier ideas of heterochrony, alterations in the relative rates of developmental processes over the course of evolution, to account for the generation of novel forms, and, with the evolutionary biologist Richard Lewontin, wrote an influential paper in 1979 suggesting that a change in one biological structure, or even a structural novelty, could arise incidentally as an accidental result of selection on another structure, rather than through direct selection for that particular adaptation. They called such incidental structural changes "spandrels" after an architectural feature. Later, Gould and Vrba discussed the acquisition of new functions by novel structures arising in this fashion, calling them "exaptations".

Molecular data regarding the mechanisms underlying development accumulated rapidly during the 1980s and '90s. It became clear that the diversity of animal morphology was not the result of different sets of proteins regulating the development of different animals, but from changes in the deployment of a small set of proteins that were common to all animals. These proteins became known as the "developmental toolkit". Such perspectives influenced the disciplines of phylogenetics, paleontology and comparative developmental biology, and spawned the new discipline of evolutionary developmental biology.

More recent work in this field by Mary Jane West-Eberhard has emphasized phenotypic and developmental plasticity. It has been suggested, for example, that the rapid emergence of basic animal body plans in the Cambrian explosion was due in part to changes in the environment acting on inherent material properties of cell aggregates, such as differential cell adhesion and biochemical oscillation. The resulting forms were later stabilized by natural selection. Experimental and theoretical research on these and related ideas have been presented in the multi-authored volume *Origination of Organismal Form*.

21st century

Epigenetic inheritance

Yet another area where developmental biology has led to the questioning of some tenets of the modern evolutionary synthesis is in the field of epigenetics, the study of how environmental factors affect the way genes express themselves during development. By

the first decade of the 21st century it had become accepted that in some cases such environmental factors could affect the expression of genes in subsequent generations even though the offspring were not exposed to the same environmental factors, and there had been no genetic changes. This shows that in some cases non genetic changes to an organism can be inherited and it has been suggested that such inheritance can help with adaptation to local conditions and affect evolution. Some have suggested that in some cases a form of Lamarckian evolution may occur.

Unconventional evolutionary theory

Omega Point

Pierre Teilhard de Chardin's metaphysical Omega Point Theory describes the gradual development of the universe from subatomic particles to human society, which he viewed as its final stage and goal.

Gaia hypothesis

Teilhard de Chardin's ideas have been seen as being connected to the more specific Gaia theory by James Lovelock, who proposed that the living and nonliving parts of Earth can be viewed as a complex interacting system with similarities to a single organism. The Gaia hypothesis has also been viewed by Lynn Margulis and others as an extension of endosymbiosis and exosymbiosis. This modified hypothesis postulates that all living things have a regulatory effect on the Earth's environment that promotes life overall.

Transhumanism

Futurists have often viewed scientific and technological progress as a continuation of biological evolution. Among these, transhumanists often view such technological evolution itself as a goal in their philosophy, possibly in the form of a technological singularity.

Chapter- 3

History of Molecular Biology

The **history of molecular biology** begins in the 1930s with the convergence of various, previously distinct biological disciplines: biochemistry, genetics, microbiology, and virology. With the hope of understanding life at its most fundamental level, numerous physicists and chemists also took an interest in what would become molecular biology.

In its modern sense, molecular biology attempts to explain the phenomena of life starting from the macromolecular properties that generate them. Two categories of macromolecules in particular are the focus of the molecular biologist: 1) nucleic acids, among which the most famous is deoxyribonucleic acid (or DNA), the constituent of genes, and 2) proteins, which are the active agents of living organisms. One definition of the scope of molecular biology therefore is to characterize the structure, function and relationships between these two types of macromolecules. This relatively limited definition will suffice to allow us to establish a date for the so-called "molecular revolution", or at least to establish a chronology of its most fundamental developments.

General overview

In its earliest manifestations, molecular biology—the name was coined by Warren Weaver of the Rockefeller Foundation in 1938—was an ideal of physical and chemical explanations of life, rather than a coherent discipline. Following the advent of the Mendelian-chromosome theory of heredity in the 1910s and the maturation of atomic theory and quantum mechanics in the 1920s, such explanations seemed within reach. Weaver and others encouraged (and funded) research at the intersection of biology, chemistry and physics, while prominent physicists such as Niels Bohr and Erwin Schrödinger turned their attention to biological speculation. However, in the 1930s and 1940s it was by no means clear which—if any—cross-disciplinary research would bear fruit; work in colloid chemistry, biophysics and radiation biology, crystallography, and other emerging fields all seemed promising.

In 1940, George Beadle and Edward Tatum demonstrated the existence of a precise relationship between genes and proteins. In the course of their experiments connecting genetics with biochemistry, they switched from the genetics mainstay *Drosophila* to a more appropriate model organism, the fungus *Neurospora*; the construction and exploitation of new model organisms would become a recurring theme in the development of molecular biology. In 1944, Oswald Avery, working at the Rockefeller

Institute of New York, demonstrated that genes are made up of DNA. In 1952, Alfred Hershey and Martha Chase confirmed that the genetic material of the bacteriophage, the virus which infects bacteria, is made up of DNA. In 1953, James Watson and Francis Crick discovered the double helical structure of the DNA molecule. In 1961, Francois Jacob and Jacques Monod hypothesized the existence of an intermediary between DNA and its protein products, which they called messenger RNA. Between 1961 and 1965, the relationship between the information contained in DNA and the structure of proteins was determined: there is a code, the genetic code, which creates a correspondence between the succession of nucleotides in the DNA sequence and a series of amino acids in proteins. At the beginning of the 1960s, Monod and Jacob also demonstrated how certain specific proteins, called regulative proteins, latch onto DNA at the edges of the genes and control the transcription of these genes into messenger RNA; they direct the "expression" of the genes.

The chief discoveries of molecular biology took place in a period of only about twenty-five years. Another fifteen years were required before new and more sophisticated technologies, united today under the name of genetic engineering, would permit the isolation and characterization of genes, in particular those of highly complex organisms.

The exploration of the molecular dominion

If we evaluate the molecular revolution within the context of biological history, it is easy to note that it is the culmination of a long process which began with the first observations through a microscope. The aim of these early researchers was to understand the functioning of living organisms by describing their organization at the microscopic level. From the end of the 18th century, the characterization of the chemical molecules which make up living beings gained increasingly greater attention, along with the birth of physiological chemistry in the 19th century, developed by the German chemist Justus von Liebig and following the birth of biochemistry at the beginning of the 20th, thanks to another German chemist Eduard Buchner. Between the molecules studied by chemists and the tiny structures visible under the optical microscope, such as the cellular nucleus or the chromosomes, there was an obscure zone, "the world of the ignored dimensions," as it was called by the chemical-physicist Wolfgang Ostwald. This world is populated by colloids, chemical compounds whose structure and properties were not well defined.

The successes of molecular biology derived from the exploration of that unknown world by means of the new technologies developed by chemists and physicists: X-ray diffraction, electron microscopy, ultracentrifugization, and electrophoresis. These studies revealed the structure and function of the macromolecules.

A milestone in that process was the work of Dr. Linus Pauling in 1949, which for the first time linked the specific genetic mutation in patients with sickle cell disease to a demonstrated change in an individual protein, the hemoglobin in the erythrocytes of heterozygous or homozygous individuals.

The encounter between biochemistry and genetics

The development of molecular biology is also the encounter of two disciplines which made considerable progress in the course of the first thirty years of the twentieth century: biochemistry and genetics. The first studies the structure and function of the molecules which make up living things. Between 1900 and 1940, the central processes of metabolism were described: the process of digestion and the absorption of the nutritive elements derived from alimentation, such as the sugars. Every one of these processes is catalyzed by a particular enzyme. Enzymes are proteins, like the antibodies present in blood or the proteins responsible for muscular contraction. As a consequence, the study of proteins, of their structure and synthesis, became one of the principal objectives of biochemists.

The second discipline of biology which developed at the beginning of the 20th century is genetics. After the rediscovery of the laws of Mendel through the studies of Hugo de Vries, Carl Correns and Erich von Tschermak in 1900, this science began to take shape thanks to the adoption by Thomas Hunt Morgan, in 1910, of a model organism for genetic studies, the famous fruit fly (Drosophila melanogaster). Shortly after, Morgan showed that the genes are localized on chromosomes. Following this discovery, he continued working with Drosophila and, along with numerous other research groups, confirmed the importance of the gene in the life and development of organisms. Nevertheless, the chemical nature of genes and their mechanisms of action remained a mystery. Molecular biologists committed themselves to the determination of the structure, and the description of the complex relations between, genes and proteins.

The development of molecular biology was not just the fruit of some sort of intrinsic "necessity" in the history of ideas, but was a characteristically historical phenomenon, with all of its unknowns, imponderables and contingencies: the remarkable developments in physics at the beginning of the 20th century highlighted the relative lateness in development in biology, which became the "new frontier" in the search for knowledge about the empirical world. Moreover, the developments of the theory of information and cybernetics in the 1940s, in response to military exigencies, brought to the new biology a significant number of fertile ideas and, especially, metaphors.

The choice of bacteria and of its virus, the bacteriophage, as models for the study of the fundamental mechanisms of life was almost natural - they are the smallest living organisms known to exist - and at the same time the fruit of individual choices. This model owes its success, above all, to the fame and the sense of organization of Max Delbrück, a German physicist, who was able to create a dynamic research group, based in the United States, whose exclusive scope was the study of the bacteriophage: the *School of the Phage*.

The geographic panorama of the developments of the new biology was conditioned above all by preceding work. The US, where genetics had developed the most rapidly, and the UK, where there was a coexistence of both genetics and biochemical research of highly advanced levels, were in the avant-garde. Germany, the cradle of the revolutions in

physics, with the best minds and the most advanced laboratories of genetics in the world, should have had a primary role in the development of molecular biology. But history decided differently: the arrival of the Nazis in 1933 - and, to a less extreme degree, the rigidification of totalitarian measures in fascist Italy - caused the emigration of a large number of Jewish and non-Jewish scientists. The majority of them fled to the US or the UK, providing an extra impulse to the scientific dynamism of those nations. These movements ultimately made molecular biology a truly international science from the very beginnings.

History of DNA biochemistry

The study of DNA is a central part of molecular biology.

First isolation of DNA

Working in the 19th century, biochemists initially isolated DNA and RNA (mixed together) from cell nuclei. They were relatively quick to appreciate the polymeric nature of their "nucleic acid" isolates, but realized only later that nucleotides were of two types—one containing ribose and the other deoxyribose. It was this subsequent discovery that led to the identification and naming of DNA as a substance distinct from RNA.

Friedrich Miescher (1844–1895) discovered a substance he called "nuclein" in 1869. Somewhat later, he isolated a pure sample of the material now known as DNA from the sperm of salmon, and in 1889 his pupil, Richard Altmann, named it "nucleic acid". This substance was found to exist only in the chromosomes.

In 1919 Phoebus Levene at the Rockefeller Institute identified the components (the four bases, the sugar and the phosphate chain) and he showed that the components of DNA were linked in the order phosphate-sugar-base. He called each of these units a nucleotide and suggested the DNA molecule consisted of a string of nucleotide units linked together through the phosphate groups, which are the 'backbone' of the molecule. However Levene thought the chain was short and that the bases repeated in the same fixed order. Torbjorn Caspersson and Einar Hammersten showed that DNA was a polymer.

Chromosomes and inherited traits

Max Delbrück, Nikolai V. Timofeeff-Ressovsky, and Karl G. Zimmer published results in 1935 suggesting that chromosomes are very large molecules the structure of which can be changed by treatment with X-rays, and that by so changing their structure it was possible to change the heritable characteristics governed by those chromosomes. In 1937 William Astbury produced the first X-ray diffraction patterns from DNA. He was not able to propose the correct structure but the patterns showed that DNA had a regular structure and therefore it might be possible to deduce what this structure was.

In 1943, Oswald Theodore Avery and a team of scientists discovered that traits proper to the "smooth" form of the *Pneumococcus* could be transferred to the "rough" form of the

same bacteria merely by making the killed "smooth" (S) form available to the live "rough" (R) form. Quite unexpectedly, the living R *Pneumococcus* bacteria were transformed into a new strain of the S form, and the transferred S characteristics turned out to be heritable. Avery called the medium of transfer of traits the transforming principle; he identified DNA as the transforming principle, and not protein as previously thought. He essentially redid Frederick Griffith's experiment. In 1953, Alfred Hershey and Martha Chase did an experiment (Hershey-Chase experiment) that showed, in T2 phage, that DNA is the genetic material (Hershey shared the Nobel prize with Luria).

Discovery of the structure of DNA

In the 1950s, three groups made it their goal to determine the structure of DNA. The first group to start was at King's College London and was led by Maurice Wilkins and was later joined by Rosalind Franklin. Another group consisting of Francis Crick and James D. Watson was at Cambridge. A third group was at Caltech and was led by Linus Pauling. Crick and Watson built physical models using metal rods and balls, in which they incorporated the known chemical structures of the nucleotides, as well as the known position of the linkages joining one nucleotide to the next along the polymer. At King's College Maurice Wilkins and Rosalind Franklin examined X-ray diffraction patterns of DNA fibers. Of the three groups, only the London group was able to produce good quality diffraction patterns and thus produce sufficient quantitative data about the structure.

Helix structure

In 1948 Pauling discovered that many proteins included helical shapes. Pauling had deduced this structure from X-ray patterns and from attempts to physically model the structures. (Pauling was also later to suggest an incorrect three chain helical DNA structure based on Astbury's data.) Even in the initial diffraction data from DNA by Maurice Wilkins, it was evident that the structure involved helices. But this insight was only a beginning. There remained the questions of how many strands came together, whether this number was the same for every helix, whether the bases pointed toward the helical axis or away, and ultimately what were the explicit angles and coordinates of all the bonds and atoms. Such questions motivated the modeling efforts of Watson and Crick.

Complementary nucleotides

In their modeling, Watson and Crick restricted themselves to what they saw as chemically and biologically reasonable. Still, the breadth of possibilities was very wide. A breakthrough occurred in 1952, when Erwin Chargaff visited Cambridge and inspired Crick with a description of experiments Chargaff had published in 1947. Chargaff had observed that the proportions of the four nucleotides vary between one DNA sample and the next, but that for particular pairs of nucleotides — adenine and thymine, guanine and cytosine — the two nucleotides are always present in equal proportions.

Crick and Watson DNA model built in 1953, was reconstructed largely from its original pieces in 1973 and donated to the National Science Museum in London.

Using X-ray diffraction, as well as other data from Rosalind Franklin and her information that the bases were paired, James D. Watson and Francis Crick arrived at the first accurate model of DNA's molecular structure in 1953, which was accepted through inspection by Rosalind Franklin. The discovery was made on February 28, 1953; the first Watson/Crick paper appeared in Nature on April 25, 1953. Sir Lawrence Bragg, the director of the Cavendish Laboratory, where Watson and Crick worked, gave a talk at Guys Hospital Medical School in London on Thursday, May 14, 1953 which resulted in an article by Ritchie Calder in The News Chronicle of London, on Friday, May 15, 1953, entitled "Why You Are You. Nearer Secret of Life." The news reached readers of The New York Times the next day; Victor K. McElheny, in researching his biography,

"Watson and DNA: Making a Scientific Revolution", found a clipping of a six-paragraph New York Times article written from London and dated May 16, 1953 with the headline "Form of `Life Unit' in Cell Is Scanned." The article ran in an early edition and was then pulled to make space for news deemed more important.(The New York Times subsequently ran a longer article on June 12, 1953). The Cambridge University undergraduate newspaper also ran its own short article on the discovery on Saturday, May 30, 1953. Bragg's original announcement at a Solvay conference on proteins in Belgium on 8 April 1953 went unreported by the press. In 1962 Watson, Crick, and Maurice Wilkins jointly received the Nobel Prize for Physiology or Medicine for their determination of the structure of DNA.

"Central Dogma"

Watson and Crick's model attracted great interest immediately upon its presentation. Arriving at their conclusion on February 21, 1953, Watson and Crick made their first announcement on February 28. In an influential presentation in 1957, Crick laid out the "Central Dogma", which foretold the relationship between DNA, RNA, and proteins, and articulated the "sequence hypothesis." A critical confirmation of the replication mechanism that was implied by the double-helical structure followed in 1958 in the form of the Meselson-Stahl experiment. Work by Crick and coworkers showed that the genetic code was based on non-overlapping triplets of bases, called codons, and Har Gobind Khorana and others deciphered the genetic code not long afterward (1966). These findings represent the birth of molecular biology.

History of RNA tertiary structure

Pre-history: the helical structure of RNA

The earliest work in RNA structural biology coincided, more or less, with the work being done on DNA in the early 1950s. In their seminal 1953 paper, Watson and Crick suggested that van der Waals crowding by the 2`OH group of ribose would preclude RNA from adopting a double helical structure identical to the model they proposed - what we now know as B-form DNA. This provoked questions about the three dimensional structure of RNA: could this molecule form some type of helical structure, and if so, how? As with DNA, early structural work on RNA centered around isolation of native RNA polymers for fiber diffraction analysis. In part because of heterogeneity of the samples tested, early fiber diffraction patterns were usually ambiguous and not readily interpretable. In 1955, Grunberg-Manago *et al.* published a paper describing the enzyme polynucleotide phosphorylase, which cleaved a phosphate group from nucleotide diphosphates to catalyze their polymerization. This discovery allowed researchers to synthesize homogenous nucleotide polymers, which they then combined to produce double stranded molecules. These samples yielded the most readily interpretable fiber diffraction patterns yet obtained, suggesting an ordered, helical structure for cognate, double stranded RNA that differed from that observed in DNA. These results paved the way for a series of investigations into the various properties and propensities of RNA. Through the late 1950s and early 1960s, numerous papers were publihed on various

topics in RNA structure, including RNA-DNA hybridization, triple stranded RNA, and even small-scale crystallography of RNA di-nucleotides - G-C, and A-U - in primitive helix-like arrangements.

The beginning: crystal structure of $tRNA^{PHE}$

In the mid-1960s, the role of tRNA in protein synthesis was being intensively studied. At this point, ribosomes had been implicated in protein synthesis, and it had been shown that an mRNA strand was necessary for the formation of these structures. In a 1964 publication, Warner and Rich showed that ribosomes active in protein synthesis contained tRNA molecules bound at the A and P sites, and discussed the notion that these molecules aided in the peptidyl transferase reaction. However, despite considerable biochemical characterization, the structural basis of tRNA function remained a mystery. In 1965, Holley *et al.* purified and sequenced the first tRNA molecule, initially proposing that it adopted a cloverleaf structure, based largely on the ability of certain regions of the molecule to form stem loop structures. The isolation of tRNA proved to be the first major windfall in RNA structural biology. Following Holley's publication, numerous investigators began work on isolation tRNA for crystallographic study, developing improved methods for isolating the molecule as they worked. By 1968 several groups had produced tRNA crystals, but these proved to be of limited quality and did not yield data at the resolutions necessary to determine structure. In 1971, Kim *et al.* achieved another breakthrough, producing crystals of yeast $tRNA^{PHE}$ that diffracted to 2-3 Ångström resolutions by using spermine, a naturally occurring polyamine, which bound to and stabilized the tRNA. Despite having suitable crystals, however, the structure of $tRNA^{PHE}$ was not immediately solved at high resolution; rather it took pioneering work in the use of heavy metal derivatives and a good deal more time to produce a high-quality density map of the entire molecule. In 1973, Kim *et al.* produced a 4 Ångström map of the tRNA molecule in which they could unambiguously trace the entire backbone. This solution would be followed by many more, as various investigators worked to refine the structure and thereby more thoroughly elucidate the details of base pairing and stacking interactions, and validate the published architecture of the molecule.

The $tRNA^{PHE}$ structure is notable in the field of nucleic acid structure in general, as it represented the first solution of a long-chain nucleic acid structure of any kind - RNA or DNA - preceding Dickerson's solution of a B-form dodecamer by nearly a decade. Also, $tRNA^{PHE}$ demonstrated many of the tertiary interactions observed in RNA architecture which would not be categorized and more thoroughly understood for years to come, providing a foundation for all future RNA structural research.

The renaissance: the hammerhead ribozyme and the group I intron: P_{4-6}

For a considerable time following the first tRNA structures, the field of RNA structure did not dramatically advance. The ability to study an RNA strcuture depended upon the potential to isolate the RNA target. This proved limiting to the field for many years, in part owing to the fact that other know targets - i.e. the ribosome - were significantly more

difficult to isolate and crystallize. Further, because other interesting RNA targets had simply not been identified, or were not sufficiently understood to be deemed interesting, there was simply a lack of things to study structurally. As such, for some twenty years following the original publication of the $tRNA^{PHE}$ structure, the structures of only a handful of other RNA targets were solved, with almost all of these belonging to the transfer RNA family. This unfortunate lack of scope would eventually be overcome largely because of two major advancements in nucleic acid research: the identification of ribozymes, and the ability to produce them via *in vitro* transcription.

Subsequent to Tom Cech's publication implicating the *Tetrahymena* group I intron as an autocatalytic ribozyme, and Sidney Altman's report of catalysis by ribonuclease P RNA, several other catalytic RNAs were identified in the late 1980s, including the hammerhead ribozyme. In 1994, McKay *et al.* published the structure of a 'hammerhead RNA-DNA ribozyme-inhibitor complex' at 2.6 Ångström resolution, in which the autocatalytic activity of the ribozyme was disrupted via binding to a DNA substrate. The conformation of the ribozyme published in this paper was eventually shown to be one of several possible states, and although this particular sample was catalytically inactive, subsequent structures have revealed its active-state architecture. This structure was followed by Doudna's publication of the structure of the P4-P6 domains of the *Tetrahymena* group I intron, a fragment of the ribozyme originally made famous by Cech. The second clause in the title of this publication - *Principles of RNA Packing* - concisely evinces the value of these two structures: for the first time, comparisons could be made between well described tRNA structures and those of globular RNAs outside the transfer family. This allowed the framework of categorization to be built for RNA tertiary structure. It was now possible to propose the conservation of motifs, folds, and various local stabilizing interactions.

In addition to the advances being made in global structure determination via crystallography, the early 1990s also saw the implementation of NMR as a powerful technique in RNA structural biology. Coincident with the large-scale ribozyme structures being solved crystallographically, a number of structures of small RNAs and RNAs complexed with drugs and peptides were solved using NMR. In addition, NMR was now being used to investigate and supplement crystal structures, as exemplified by the determination of an isolated tetraloop-receptor motif structure published in 1997. Investigations such as this enabled a more precise characterization of the base pairing and base stacking interactions which stabilized the global folds of large RNA molecules. The importance of understanding RNA tertiary structural motifs was prophetically well described by Michel and Costa in their publication identifying the tetraloop motif: "..it should not come as a surprise if self-folding RNA molecules were to make intensive use of only a relatively small set of tertiary motifs. Identifying these motifs would greatly aid modeling enterprises, which will remain essential as long as the crystallization of large RNAs remains a difficult task".

The modern era: the age of RNA structural biology

The resurgence of RNA structural biology in the mid-1990s has caused a veritable explosion in the field of nucleic acid structural research. Since the publication of the hammerhead and P_{4-6} structures, numerous major contributions to the field have been made. Some of the most noteworthy examples include the structures of the Group I and Group II introns, and the Ribosome. It should be noted that the first three structures were produced using *in vitro* transcription, and that NMR has played a role in investigating partial components of all four structures - testaments to the indispensability of both techniques for RNA research. Most recently, the 2009 Nobel Prize in Chemistry was awarded to Ada Yonath, Venkatraman Ramakrishnan, and Thomas Steitz for their structural work on the ribosome, demonstrating the prominent role RNA structural biology has taken in modern molecular biology.

History of protein biochemistry

First isolation and classification

Proteins were recognized as a distinct class of biological molecules in the eighteenth century by Antoine Fourcroy and others. Members of this class (called the "albuminoids", *Eiweisskörper*, or *matières albuminoides*) were recognized by their ability to coagulate or flocculate under various treatments such as heat or acid; well-known examples at the start of the nineteenth century included albumen from egg whites, blood serum albumin, fibrin, and wheat gluten. The similarity between the cooking of egg whites and the curdling of milk was recognized even in ancient times; for example, the name *albumen* for the egg-white protein was coined by Pliny the Elder from the Latin *albus ovi* (egg white).

With the advice of Jöns Jakob Berzelius, the Dutch chemist Gerhardus Johannes Mulder carried out elemental analyses of common animal and plant proteins. To everyone's surprise, all proteins had nearly the same empirical formula, roughly $C_{400}H_{620}N_{100}O_{120}$ with individual sulfur and phosphorus atoms. Mulder published his findings in two papers (1837,1838) and hypothesized that there was one basic substance (*Grundstoff*) of proteins, and that it was synthesized by plants and absorbed from them by animals in digestion. Berzelius was an early proponent of this theory and proposed the name "protein" for this substance in a letter dated 10 July 1838

The name protein that I propose for the organic oxide of fibrin and albumin, I wanted to derive from [the Greek word] πρωτειος, because it appears to be the primitive or principal substance of animal nutrition.

Mulder went on to identify the products of protein degradation such as the amino acid, leucine, for which he found a (nearly correct) molecular weight of 131 Da.

Purifications and measurements of mass

The minimum molecular weight suggested by Mulder's analyses was roughly 9 kDa, hundreds of times larger than other molecules being studied. Hence, the chemical structure of proteins (their primary structure) was an active area of research until 1949, when Fred Sanger sequenced insulin. The (correct) theory that proteins were linear polymers of amino acids linked by peptide bonds was proposed independently and simultaneously by Franz Hofmeister and Emil Fischer at the same conference in 1902. However, some scientists were sceptical that such long macromolecules could be stable in solution. Consequently, numerous alternative theories of the protein primary structure were proposed, e.g., the colloidal hypothesis that proteins were assemblies of small molecules, the cyclol hypothesis of Dorothy Wrinch, the diketopiperazine hypothesis of Emil Abderhalden and the pyrrol/piperidine hypothesis of Troensgard (1942). Most of these theories had difficulties in accounting for the fact that the digestion of proteins yielded peptides and amino acids. Proteins were finally shown to be macromolecules of well-defined composition (and not colloidal mixtures) by Theodor Svedberg using analytical ultracentrifugation. The possibility that some proteins are non-covalent associations of such macromolecules was shown by Gilbert Smithson Adair (by measuring the osmotic pressure of hemoglobin) and, later, by Frederic M. Richards in his studies of ribonuclease S. The mass spectrometry of proteins has long been a useful technique for identifying posttranslational modifications and, more recently, for probing protein structure.

Most proteins are difficult to purify in more than milligram quantities, even using the most modern methods. Hence, early studies focused on proteins that could be purified in large quantities, e.g., those of blood, egg white, various toxins, and digestive/metabolic enzymes obtained from slaughterhouses. Many techniques of protein purification were developed during World War II in a project led by Edwin Joseph Cohn to purify blood proteins to help keep soldiers alive. In the late 1950s, the Armour Hot Dog Co. purified 1 kg (= one million milligrams) of pure bovine pancreatic ribonuclease A and made it freely available to scientists around the world. This generous act made RNase A the main protein for basic research for the next few decades, resulting in several Nobel Prizes.

Protein folding and first structural models

The study of protein folding began in 1910 with a famous paper by Henrietta Chick and C. J. Martin, in which they showed that the flocculation of a protein was composed of two distinct processes: the precipitation of a protein from solution was *preceded* by another process called denaturation, in which the protein became much less soluble, lost its enzymatic activity and became more chemically reactive. In the mid-1920s, Tim Anson and Alfred Mirsky proposed that denaturation was a reversible process, a correct hypothesis that was initially lampooned by some scientists as "unboiling the egg". Anson also suggested that denaturation was a two-state ("all-or-none") process, in which one fundamental molecular transition resulted in the drastic changes in solubility, enzymatic activity and chemical reactivity; he further noted that the free energy changes upon denaturation were much smaller than those typically involved in chemical reactions. In

1929, Hsien Wu hypothesized that denaturation was protein unfolding, a purely conformational change that resulted in the exposure of amino acid side chains to the solvent. According to this (correct) hypothesis, exposure of aliphatic and reactive side chains to solvent rendered the protein less soluble and more reactive, whereas the loss of a specific conformation caused the loss of enzymatic activity. Although considered plausible, Wu's hypothesis was not immediately accepted, since so little was known of protein structure and enzymology and other factors could account for the changes in solubility, enzymatic activity and chemical reactivity. In the early 1960s, Chris Anfinsen showed that the folding of ribonuclease A was fully reversible with no external cofactors needed, verifying the "thermodynamic hypothesis" of protein folding that the folded state represents the global minimum of free energy for the protein.

The hypothesis of protein folding was followed by research into the physical interactions that stabilize folded protein structures. The crucial role of hydrophobic interactions was hypothesized by Dorothy Wrinch and Irving Langmuir, as a mechanism that might stabilize her cyclol structures. Although supported by J. D. Bernal and others, this (correct) hypothesis was rejected along with the cyclol hypothesis, which was disproven in the 1930s by Linus Pauling (among others). Instead, Pauling championed the idea that protein structure was stabilized mainly by hydrogen bonds, an idea advanced initially by William Astbury (1933). Remarkably, Pauling's incorrect theory about H-bonds resulted in his *correct* models for the secondary structure elements of proteins, the alpha helix and the beta sheet. The hydrophobic interaction was restored to its correct prominence by a famous article in 1959 by Walter Kauzmann on denaturation, based partly on work by Kaj Linderstrøm-Lang. The ionic nature of proteins was demonstrated by Bjerrum, Weber and Arne Tiselius, but Linderstrom-Lang showed that the charges were generally accessible to solvent and not bound to each other (1949).

The secondary and low-resolution tertiary structure of globular proteins was investigated initially by hydrodynamic methods, such as analytical ultracentrifugation and flow birefringence. Spectroscopic methods to probe protein structure (such as circular dichroism, fluorescence, near-ultraviolet and infrared absorbance) were developed in the 1950s. The first atomic-resolution structures of proteins were solved by X-ray crystallography in the 1960s and by NMR in the 1980s. As of 2006, the Protein Data Bank has nearly 40,000 atomic-resolution structures of proteins. In more recent times, cryo-electron microscopy of large macromolecular assemblies and computational protein structure prediction of small protein domains are two methods approaching atomic resolution.

Chapter- 4

History of RNA Biology

Numerous key discoveries in biology have emerged from studies of RNA (ribonucleic acid), including seminal work in the fields of biochemistry, genetics, microbiology, molecular biology, molecular evolution and structural biology. As of 2010, 30 scientists have been awarded Nobel Prizes for experimental work that includes studies of RNA.

1930 - 1950

RNA and DNA have distinct chemical properties

When first studied in the early 1900s, the chemical and biological differences between RNA and DNA were not apparent, and they were named after the materials from which they were isolated; RNA was initially known as "yeast nucleic acid" and DNA was "pancreas nucleic acid". Using diagnostic chemical tests, carbohydrate chemists showed that the two nucleic acids contained different sugars, whereupon the common name for RNA became "ribose nucleic acid". Other early biochemical studies showed that RNA was readily broken down at high pH, while DNA was stable (although denatured) in alkali. Nucleoside composition analysis showed first that RNA contained similar nucleobases to DNA, with uracil instead of thymine, and that RNA contained a number of minor nucleobase components, e.g. small amounts of pseudouridine and dimethylguanine.

1951 - 1965

Messenger RNA (mRNA) carries genetic information that directs protein synthesis

The concept of messenger RNA emerged during the late 1950s, and is associated with Crick's description of his "Central Dogma of Molecular Biology", which asserted that DNA led to the formation of RNA, which in turn led to the synthesis of proteins. During the early 1960s, sophisticated genetic analysis of mutations in the lac operon of E. coli and in the rII locus of bacteriophage T4 were instrumental in defining the nature of both messenger RNA and the genetic code. The short-lived nature of bacterial RNAs, together

with the highly complex nature of the cellular mRNA population, made the biochemical isolation of mRNA very challenging. This problem was overcome in the 1960s by the use of reticulocytes in vertebrates, which produce large quantities of mRNA that are highly enriched in RNA encoding alpha- and beta-globin (the two major protein chains of hemoglobin).

Ribosomes make proteins

In the 1950s, results of labeling experiments in rat liver showed that radioactive amino acids were found to be associated with "microsomes" (later redefined as ribosomes) very rapidly after administration, and before they became widely incorporated into cellular proteins. Ribosomes were first visualized using electron microscopy, and their ribonucleoprotein components were identified by biophysical methods, chiefly sedimentation analysis within ultracentrifuges capable of generating very high accelerations (equivalent to hundreds of thousands times gravity). Polysomes (multiple ribosomes moving along a single mRNA molecule) were identified in the early 1960s, and their study led to an understanding of how ribosomes proceed to read the mRNA in a 5' to 3' direction, processively generating proteins as they do so.

Transfer RNA (tRNA) is the physical link between RNA and protein

Biochemical fractionation experiments showed that radioactive amino acids were rapidly incorporated into small RNA molecules that remained soluble under conditions where larger RNA-containing particles would precipitate. These molecules were termed soluble (sRNA) and were later renamed transfer RNA (tRNA). Subsequent studies showed that (i) every cell has multiple species of tRNA, each of which is associated with a single specific amino acid, (ii) that there are a matching set of enzymes responsible for linking tRNAs with the correct amino acids, and (iii) that tRNA anticodon sequences form a specific decoding interaction with mRNA codons.

The genetic code is solved

The genetic code consists of the translation of particular nucleotide sequences in mRNA to specific amino acid sequences in proteins (polypeptides). The ability to work out out the genetic code emerged from the convergence of three different areas of study--(i) new methods to generate synthetic RNA molecules of defined composition to serve as artificial mRNAs, (ii) development of *in vitro* translation systems that could used to translate the synthetic mRNAs into protein, and (iii) experimental and theoretical genetic work which established that the code was written in three letter "words" (codons). Today, our understanding of the genetic code permits the prediction of the amino sequence of the protein products of the tens of thousands of genes whose sequences are being determined in genome studies.

RNA polymerase is purified

The biochemical purification and characterization of RNA polymerase from the bacterium Escherichia coli enabled the understanding of the mechanisms through which RNA polymerase initiates and terminates transcription, and how those processes are regulated to regulate gene expression (i.e. turn genes on and off). Following the isolation of E. coli RNA polymerase, the three RNA polymerases of the eukaryotic nucleus were identified, as well as those associated with viruses and organelles. Studies of transcription also led to the identification of many protein factors that influence transcription, including repressors, activators and enhancers. The availability of purified preparations of RNA polymerase permitted investigators to develop a wide range of novel methods for studying RNA in the test tube, and led directly to many of the subsequent key discoveries in RNA biology.

1966 - 1975

First complete nucleotide sequence of a biological nucleic acid molecule

Although determining the sequence of proteins was becoming somewhat routine, methods for sequencing of nucleic acids were not available until the mid-1960s. In this seminal work, a specific tRNA was purified in substantial quantities, and then sliced into overlapping fragments using a variety of ribonucleases. Analysis of the detailed nucleotide composition of each fragment provided the information necessary to deduce the sequence of the tRNA. Today, the sequence analysis of much larger nucleic acid molecules is highly-automated and enormously faster.

Evolutionary variation of homologous RNA sequences reveals folding patterns

Additional tRNA molecules were purified and sequenced. The first comparative sequence analysis was done and revealed that the sequences varied through evolution in such a way that all of the tRNAs could fold into very similar secondary structures (two-dimensional structures) and had identical sequences at numerous positions (e.g. CCA at the 3' end). The radial four-arm structure of tRNA molecules is termed the 'cloverleaf structure', and results from the evolution of sequences with common ancestry and common biological function. Since the discovery of the tRNA cloverleaf, comparative analysis of numerous other homologous RNA molecules has led to the identification of common sequences and folding patterns.

First complete genomic nucleotide sequence

The 3569 nucleotide sequence of all of the genes of the RNA bacteriophage MS2 was determined by a large team of researchers over several years, and was reported in a series of scientific papers. These results enabled the analysis of the first complete genome, albeit an extremely tiny one by modern standards. Several surprising features were

identified, including genes that partially overlap one another and the first clues that different organisms might have slightly different codon usage patterns.

Reverse transcriptase can copy RNA into DNA

Retroviruses were shown to have a single-stranded RNA genome and to replicate via a DNA intermediate, the reverse of the usual DNA-to-RNA transcription pathway. They encode a RNA-dependent DNA polymerase (reverse transcriptase) that is essential for this process. Some retroviruses can cause diseases, including several that are associated with cancer, and HIV-1 which causes AIDS. Reverse transcriptase has been widely used as an experimental tool for the analysis of RNA molecules in the laboratory, in particular the conversion of RNA molecules into DNA prior to molecular cloning and/or polymerase chain reaction (PCR).

RNA replicons evolve rapidly

Biochemical and genetic analyses showed that the enzyme systems that replicate viral RNA molecules (reverse transcriptases and RNA replicases) lack molecular proofreading (3' to 5' exonuclease) activity, and that RNA sequences do not benefit from extensive repair systems analogous to those that exist for maintaining and repairing DNA sequences. Consequently, RNA genomes appear to be subject to significantly higher mutation rates than DNA genomes. For example, mutations in HIV-1 that lead to the emergence of viral mutants that are insensitive to antiviral drugs are common, and constitute a major clinical challenge.

Ribosomal RNA (rRNA) sequences provide a record of the evolutionary history of all life forms

Analysis of ribosomal RNA sequences from a large number of organisms demonstrated that all extant forms of life on Earth share common structural and sequence features of the ribosomal RNA, reflecting a common ancestry. Mapping the similarities and differences among rRNA molecules from different sources provides clear and quantitative information about the phylogenetic (i.e. evolutionary) relationships among organisms. Analysis of rRNA molecules led to the identification of a third major kingdom of organisms, the archaea, in addition to the prokaryotes and eukaryotes.

Non-encoded nucleotides are added to the ends of RNA molecules

Molecular analysis of mRNA molecules showed that, following transcription, mRNAs have non-DNA-encoded nucleotides added to both their 5' and 3' ends (guanosine caps and poly-A, respectively). Enzymes were also identified that add and maintain the universal CCA sequence on the 3' end of tRNA molecules. These events are among the first discovered examples of RNA processing, a complex series of reactions that are needed to convert RNA primary transcripts into biologically active RNA molecules.

1976 - 1985

Small RNA molecules are abundant in the eukaryotic nucleus

Small nuclear RNA molecules (snRNAs) were identified in the eukaryotic nucleus using immunological studies with autoimmune antibodies, which bind to small nuclear ribonucleoprotein complexes (snRNPs; complexes of the snRNA and protein). Subsequent biochemical, genetic, and phylogenetic studies established that many of these molecules play key roles in essential RNA processing reactions within the nucleus and nucleolus, including RNA splicing, polyadenylation, and the maturation of ribosomal RNAs.

RNA molecules require a specific, complex three-dimensional structure for activity

The detailed three-dimensional structure of tRNA molecules was determined using X-ray crystallography, and revealed highly complex, compact three dimensional structures consisting of tertiary interactions laid upon the basic cloverleaf secondary structure. Key features of tRNA tertiary structure include the coaxial stacking of adjacent helices and non-Watson-Crick interactions among nucleotides within the apical loops. Additional crystallographic studies showed that a wide range of RNA molecules (including ribozymes, riboswitches and ribosomal RNA) also fold into specific structures containing a variety of 3D structural motifs. The ability of RNA molecules to adopt specific tertiary structures is essential for their biological activity, and results from the single-stranded nature of RNA. In many ways, RNA folding is more highly analogous to the folding of proteins rather than to the highly repetitive folded structure of the DNA double helix.

Genes are commonly interrupted by introns that must be removed by RNA splicing

Analysis of mature eukaryotic messenger RNA molecules showed that they are often much smaller than the DNA sequences that encode them. The genes were shown to be discontinuous, composed of sequences that are not present in the final mature RNA (introns), located between sequences that are retained in the mature RNA (exons). Introns were shown to be removed after transcription through a process termed RNA splicing. Splicing of RNA transcripts requires a highly precise and coordinated sequence of molecular events, consisting of (a) definition of boundaries between exons and introns, (b) RNA strand cleavage at exactly those sites, and (c) covalent linking (ligation) of the RNA exons in the correct order. The discovery of discontinuous genes and RNA splicing was entirely unexpected by the community of RNA biologists, and stands as one of the most shocking findings in molecular biology research.

Alternative pre-mRNA splicing generates multiple proteins from a single gene

The great majority of protein-coding genes encoded within the nucleus of metazoan cells contain multiple introns. In many cases, these introns were shown to be processed in more than one pattern, thus generating a family of related mRNAs that differ, for example, by the inclusion or exclusion of particular exons. The end result of alternative splicing is that a single gene can encode a number of different protein isoforms that can exhibit a variety of (usually related) biological functions. Indeed, most of the proteins encoded by the human genome are generated by alternative splicing.

Discovery of catalytic RNA (ribozymes)

An experimental system was developed in which an intron-containing rRNA precursor from the nucleus of the ciliated protozoan Tetrahymena could be spliced *in vitro*. Subsequent biochemical analysis shows that this group I intron was self-splicing; that is, the precursor RNA is capable of carrying out the complete splicing reaction in the absence of proteins. In separate work, the RNA component of the bacterial enzyme ribonuclease P (a ribonucleoprotein complex) was shown to catalyze its tRNA-processing reaction in the absence of proteins. These experiments represented landmarks in RNA biology, since they revealed that RNA could play an active role in cellular processes, by catalyzing specific biochemical reactions. Before these discoveries, it was believed that biological catalysis was solely the realm of protein enzymes.

RNA was likely critical for prebiotic evolution

The discovery of catalytic RNA (ribozymes) showed that RNA could both encode genetic information (like DNA) and catalyze specific biochemical reactions (like protein enzymes). This realization led to the RNA World Hypothesis, a proposal that RNA may have played a critical role in prebiotic evolution at a time before the molecules with more specialized functions (DNA and proteins) came to dominate biological information coding and catalysis. Although it is not possible for us to know the course of prebiotic evolution with any certainty, the presence of functional RNA molecules with common ancestry in all modern-day life forms is a strong argument that RNA was widely present at the time of the last common ancestor.

Introns can be mobile genetic elements

Some self-splicing introns can spread through a population of organisms by "homing", inserting copies of themselves into genes at sites that previously lacked an intron. Because they are self-splicing (that is, they remove themselves at the RNA level from genes into which they have inserted), these sequences represent transposons that are genetically silent, i.e. they do not interfere with the expression of the gene into which they become inserted. These introns can be regarded as examples of selfish DNA. Some mobile introns encode homing endonucleases, enzymes that initiate the homing process by specifically cleaving double-stranded DNA at or near the intron-insertion site of

alleles lacking an intron. Mobile introns are frequently members of either the group I or group II families of self-splicing introns.

Spliceosomes mediate nuclear pre-mRNA splicing

Introns are removed from nuclear pre-mRNAs by splicesomes, large ribonucleoprotein complexes made up of snRNA and protein molecules whose composition and molecular interactions change during the course of the RNA splicing reactions. Spliceosomes assemble on and around splice sites (the boundaries between introns and exons in the unspliced pre-mRNA) in mRNA precursors and use RNA-RNA interactions to identify critical nucleotide sequences and, probably, to catalyze the splicing reactions. Nuclear pre-mRNA introns and spliceosome-associated snRNAs show similar structural features to self-splicing group II introns. In addition, the splicing pathway of nuclear pre-mRNA introns and group II introns shares a similar reaction pathway. These similarities have led to the hypothesis that these molecules may share a common ancestor.

1986 - 2000

RNA sequences can be edited within cells

Messenger RNA precursors from a wide range of organisms can be edited before being translated into protein. In this process, non-encoded nucleotides may be inserted into specific sites in the RNA, and encoded nucleotides may be removed or replaced. RNA editing was first discovered within the mitochondria of kinetoplastid protozoans, where it has been shown to be extensive. For example, some protein-coding genes encode fewer than 50% of the nucleotides found within the mature, translated mRNA. Other RNA editing events are found in mammals, plants, bacteria and viruses. These latter editing events involve fewer nucleotide modifications, insertions and deletions than the events within kinetoplast DNA, but still have high biological significance for gene expression and its regulation.

Telomerase uses a built-in RNA template to maintain chromosome ends

Telomerase is an enzyme that is present in all eukaryotic nuclei which serves to maintain the ends of the linear DNA in the linear chromosomes of the eukaryotic nucleus, through the addition of terminal sequences that are lost in each round of DNA replication. Before telomerase was identified, its activity was predicted on the basis of a molecular understanding of DNA replication, which indicated that the DNA polymerases known at that time could not replicate the 3' end of a linear chromosome, due to the absence of a template strand. Telomerase was shown to be a ribonucleoprotein enzyme that contains an RNA component that serves as a template strand, and a protein component that has reverse transcriptase activity and adds nucleotides to the chromosome ends using the internal RNA template.

Ribosomal RNA catalyzes peptide bond formation

For years, scientists had worked to identify which protein(s) within the ribosome were responsible for peptidyl transferase function during translation, because the covalent linking of amino acids represents one of the most central chemical reactions in all of biology. Careful biochemical studies showed that extensively-deproteinized large ribosomal subunits could still catalyze peptide bond formation, thereby implying that the sought-after activity might lie within ribosomal RNA rather than ribosomal proteins. Structural biologists, using X-ray crystallography, localized the peptidyl transferase center of the ribosome to a highly-conserved region of the large subunit ribosomal RNA (rRNA) that is located at the place within the ribosome where the amino-acid-bearing ends of tRNA bind, and where no proteins are present. These studies led to the conclusion that the ribosome is a ribozyme. The rRNA sequences that make up the ribosomal active site represent some of the most highly conserved sequences in the biological world. Together, these observations indicate that peptide bond formation catalyzed by RNA was a feature of the last common ancestor of all known forms of life.

Combinatorial selection of RNA molecules enables in vitro evolution

Experimental methods were invented that allowed investigators to use large, diverse populations of RNA molecules to carry out in vitro molecular experiments that utilized powerful selective replication strategies used by geneticists, and which amount to evolution in the test tube. These experiments have been described using different names, the most common of which are "combinatorial selection", "in vitro selection", and SELEX (for Systematic Evolution of Ligands by Exponential Enrichment). These experiments have been used for isolating RNA molecules with a wide range of properties, from binding to particular proteins, to catalyzing particular reactions, to binding low molecular weight organic ligands. They have equal applicability to elucidating interactions and mechanisms that are known properties of naturally-occurring RNA molecules to isolating RNA molecules with biochemical properties that are not known in nature. In developing in vitro selection technology for RNA, laboratory systems for synthesizing complex populations of RNA molecules were established, and used in conjunction with the selection of molecules with user-specified biochemical activities, and in vitro schemes for RNA replication. These steps can be viewed as (a) mutation, (b) selection, and (c) replication. Together, then, these three processes enable in vitro molecular evolution.

2001 - present

Many mobile DNA elements use an RNA intermediate

Transposable genetic elements (transposons) are found which can replicate via transcription into an RNA intermediate which is subsequently converted to DNA by reverse transcriptase. These sequences, many of which are likely related to retroviruses, constitute much of the DNA of the eukaryotic nucleus, especially so in plants. Genomic

sequencing shows that retrotransposons make up 36% of the human genome and over half of the genome of major cereal crops (wheat and maize).

Riboswitches bind cellular metabolites and control gene expression

Segments of RNA, typically embedded within the 5'-untranslated region of a vast number of bacterial mRNA molecules, have a profound effect on gene expression through a previously-undiscovered mechanism that does not involve the participation of proteins. In many cases, riboswitches change their folded structure in response to environmental conditions (e.g. ambient temperature or concentrations of specific metabolites), and the structural change controls the translation or stability of the mRNA in which the riboswitch is embedded. In this way, gene expression can be dramatically regulated at the post-transcriptional level.

Small RNA molecules regulate gene expression by post-transcriptional gene silencing

Another previously unknown mechanism by which RNA molecules are involved in genetic regulation was discovered in the 1990s. Small RNA molecules termed microRNA (miRNA) and small interfering RNA (siRNA) are abundant in eukaryotic cells and exert post-transcriptional control over mRNA expression. They function by binding to specific sites within the mRNA and inducing cleavage of the mRNA via a specific silencing-associated RNA degradation pathway.

Noncoding RNA controls epigenetic phenomena

In addition to their well-established roles in translation and splicing, members of noncoding RNA (ncRNA) families have recently been found to function in genome defense and chromosome inactivation. For example, piwi-interacting RNAs (piRNAs) prevent genome instability in germ line cells, while Xist (X-inactive-specific-transcript) is essential for X-chromosome inactivation in mammals.

Nobel Laureates in RNA biology

Name	Dates	Institution	Awards
Altman, Sidney	1939-	Yale University	1989 Nobel Prize in Chemistry
Baltimore, David	1938-	California Institute of Technology	1975 Nobel Prize in Physiology or Medicine
Barré-Sinoussi, Françoise	1947-	Pasteur Institute	2008 Nobel Prize in Physiology or Medicine
Blackburn, Elizabeth	1948-	University of California, San Francisco	2009 Nobel Prize in Physiology or Medicine

Brenner, Sydney	1927-	Salk Institute	2002 Nobel Prize in Physiology or Medicine
Cech, Thomas	1947-	University of Colorado, Boulder	1989 Nobel Prize in Chemistry
Crick, Francis	1916-2004	Salk Institute	1962 Nobel Prize in Physiology or Medicine
Dulbecco, Renato	1914-	CNR Institute of Biomedical Technologies (Italy)	1975 Nobel Prize in Physiology or Medicine
Fire, Andrew	1959-	Stanford University	2006 Nobel Prize in Physiology or Medicine
Gilbert, Walter	1932-	Harvard University	1980 Nobel Prize in Chemistry
Greider, Carol	1961-	Johns Hopkins University	2009 Nobel Prize in Physiology or Medicine
Holley, Robert	1922-1993	Cornell University	1968 Nobel Prize in Physiology or Medicine
Jacob, François	1920	Pasteur Institute	1965 Nobel Prize in Physiology or Medicine
Khorana, H. Gobind	1922-	Massachusetts Institute of Technology	1968 Nobel Prize in Physiology or Medicine
Klug, Aaron	1926-	Medical Research Council (UK)	1982 Nobel Prize in Chemistry
Kornberg, Roger	1947-	Stanford University	2006 Nobel Prize in Chemistry
Mello, Craig	1960-	University of Massachusetts Medical School	2006 Nobel Prize in Physiology or Medicine
Monod, Jacques	1910-1976	Pasteur Institute	1965 Nobel Prize in Physiology or Medicine
Montagnier, Luc	1932-	Pasteur Institute	2008 Nobel Prize in Physiology or Medicine
Nirenberg, Marshall	1927-2010	National Institutes of Health (USA)	1968 Nobel Prize in Physiology or Medicine
Ochoa, Severo	1905-1993	New York University	1959 Nobel Prize in Physiology or Medicine
Temin, Howard	1934-	University of Wisconsin,	1975 Nobel Prize in

	1994	Madison	Physiology or Medicine
Ramakrishnan, Venkatraman	1952-	Medical Research Council (UK)	2009 Nobel Prize in Chemistry
Roberts, Richard	1943-	New England Biolabs	1993 Nobel Prize in Physiology or Medicine
Sharp, Philip	1944-	Massachusetts Institute of Technology	1993 Nobel Prize in Physiology or Medicine
Steitz, Thomas	1940-	Yale University	2009 Nobel Prize in Chemistry
Szostak, Jack	1952-	Harvard University	2009 Nobel Prize in Physiology or Medicine
Todd, Alexander	1907-1997	University of Cambridge	1957 Nobel Prize in Chemistry
Watson, James	1928-	Cold Spring Harbor Laboratory	1962 Nobel Prize in Physiology or Medicine
Yonath, Ada	1939-	Weizmann Institute of Science	2009 Nobel Prize in Chemistry

Chapter- 5

DNA Sequencing

The term **DNA sequencing** refers to sequencing methods for determining the order of the nucleotide bases—adenine, guanine, cytosine, and thymine—in a molecule of DNA.

Knowledge of DNA sequences has become indispensable for basic biological research, other research branches utilizing DNA sequencing, and in numerous applied fields such as diagnostic, biotechnology, forensic biology and biological systematics. The advent of DNA sequencing has significantly accelerated biological research and discovery. The rapid speed of sequencing attained with modern DNA sequencing technology has been instrumental in the sequencing of the human genome, in the Human Genome Project. Related projects, often by scientific collaboration across continents, have generated the complete DNA sequences of many animal, plant, and microbial genomes.

The first DNA sequences were obtained in the early 1970s by academic researchers using laborious methods based on two-dimensional chromatography. Following the development of dye-based sequencing methods with automated analysis, DNA sequencing has become easier and orders of magnitude faster.

History

RNA sequencing was one of the earliest forms of nucleotide sequencing. The major landmark of RNA sequencing is the sequence of the first complete gene and the complete genome of Bacteriophage MS2, identified and published by Walter Fiers and his coworkers at the University of Ghent (Ghent, Belgium), between 1972 and 1976.

Prior to the development of rapid DNA sequencing methods in the early 1970s by Frederick Sanger at the University of Cambridge, in England and Walter Gilbert and Allan Maxam at Harvard, a number of laborious methods were used. For instance, in 1973, Gilbert and Maxam reported the sequence of 24 basepairs using a method known as wandering-spot analysis.

The chain-termination method developed by Sanger and coworkers in 1975 soon became the method of choice, owing to its relative ease and reliability.

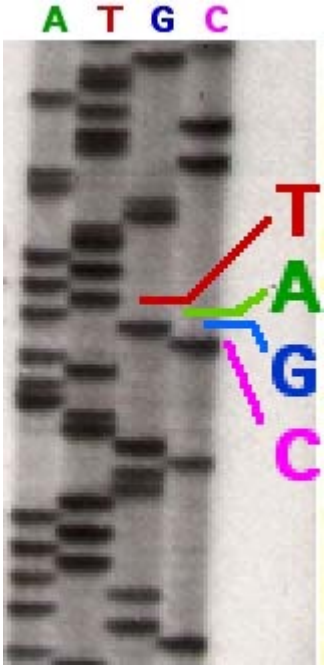
Maxam–Gilbert sequencing

In 1976–1977, Allan Maxam and Walter Gilbert developed a DNA sequencing method based on chemical modification of DNA and subsequent cleavage at specific bases. Although Maxam and Gilbert published their chemical sequencing method two years after the ground-breaking paper of Sanger and Coulson on plus-minus sequencing, Maxam–Gilbert sequencing rapidly became more popular, since purified DNA could be used directly, while the initial Sanger method required that each read start be cloned for production of single-stranded DNA. However, with the improvement of the chain-termination method (see below), Maxam-Gilbert sequencing has fallen out of favour due to its technical complexity prohibiting its use in standard molecular biology kits, extensive use of hazardous chemicals, and difficulties with scale-up.

The method requires radioactive labeling at one 5' end of the DNA (typically by a kinase reaction using gamma-^{32}P ATP) and purification of the DNA fragment to be sequenced. Chemical treatment generates breaks at a small proportion of one or two of the four nucleotide bases in each of four reactions (G, A+G, C, C+T). For example, the purines (A+G) are depurinated using formic acid, the guanines (and to some extent the adenines) are methylated by dimethyl sulfate, and the pyrimidines (C+T) are methylated using hydrazine. The addition of salt (sodium chloride) to the hydrazine reaction inhibits the methylation of thymine for the C-only reaction. The modified DNAs are then cleaved by hot piperidine at the position of the modified base. The concentration of the modifying chemicals is controlled to introduce on average one modification per DNA molecule. Thus a series of labeled fragments is generated, from the radiolabeled end to the first "cut" site in each molecule. The fragments in the four reactions are electrophoresed side by side in denaturing acrylamide gels for size separation. To visualize the fragments, the gel is exposed to X-ray film for autoradiography, yielding a series of dark bands each corresponding to a radiolabeled DNA fragment, from which the sequence may be inferred.

Also sometimes known as "chemical sequencing", this method originated in the study of DNA-protein interactions (DNase I footprinting) and nucleic acid structure, and within these it still has important applications.

Chain-termination methods



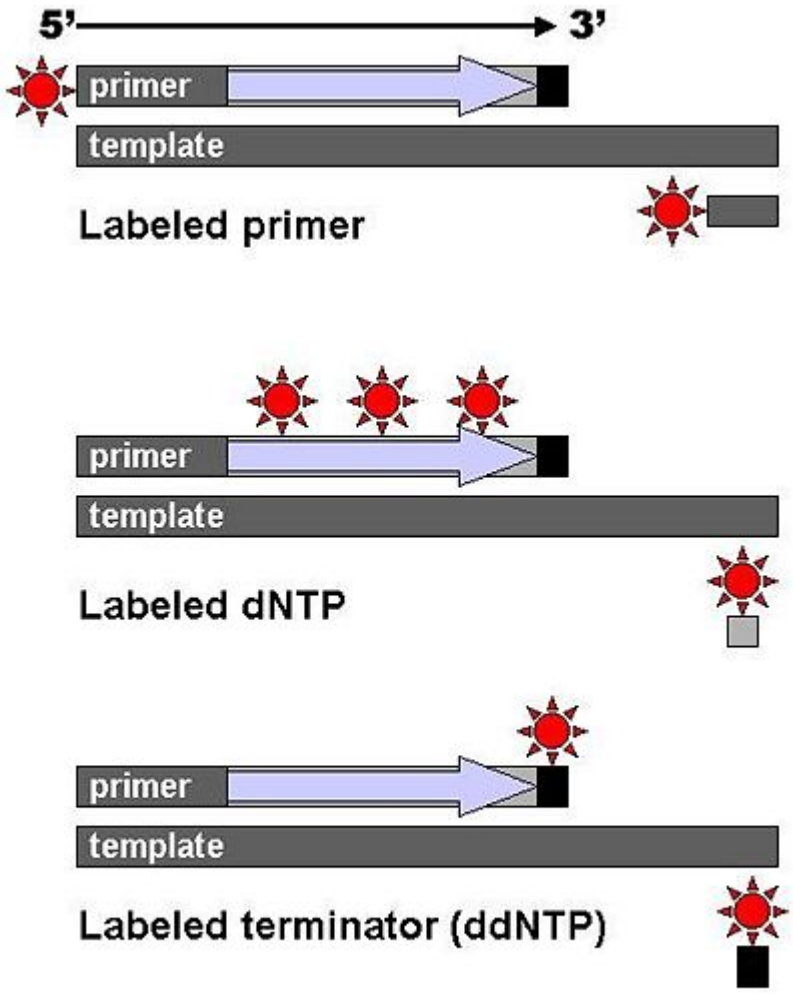
Part of a radioactively labelled sequencing gel

Because the chain-terminator method (or Sanger method after its developer Frederick Sanger) is more efficient and uses fewer toxic chemicals and lower amounts of radioactivity than the method of Maxam and Gilbert, it rapidly became the method of choice. The key principle of the Sanger method was the use of dideoxynucleotide triphosphates (ddNTPs) as DNA chain terminators.

The classical chain-termination method requires a single-stranded DNA template, a DNA primer, a DNA polymerase, normal deoxynucleotidephosphates (dNTPs), and modified nucleotides (dideoxyNTPs) that terminate DNA strand elongation. These ddNTPs will also be radioactively or fluorescently labelled for detection in automated sequencing machines. The DNA sample is divided into four separate sequencing reactions, containing all four of the standard deoxynucleotides (dATP, dGTP, dCTP and dTTP) and the DNA polymerase. To each reaction is added only one of the four dideoxynucleotides (ddATP, ddGTP, ddCTP, or ddTTP) which are the chain-terminating nucleotides, lacking a 3'-OH group required for the formation of a phosphodiester bond between two nucleotides, thus terminating DNA strand extension and resulting in DNA fragments of varying length.

The newly synthesized and labelled DNA fragments are heat denatured, and separated by size (with a resolution of just one nucleotide) by gel electrophoresis on a denaturing polyacrylamide-urea gel with each of the four reactions run in one of four individual lanes (lanes A, T, G, C); the DNA bands are then visualized by autoradiography or UV light, and the DNA sequence can be directly read off the X-ray film or gel image. In the image on the right, X-ray film was exposed to the gel, and the dark bands correspond to

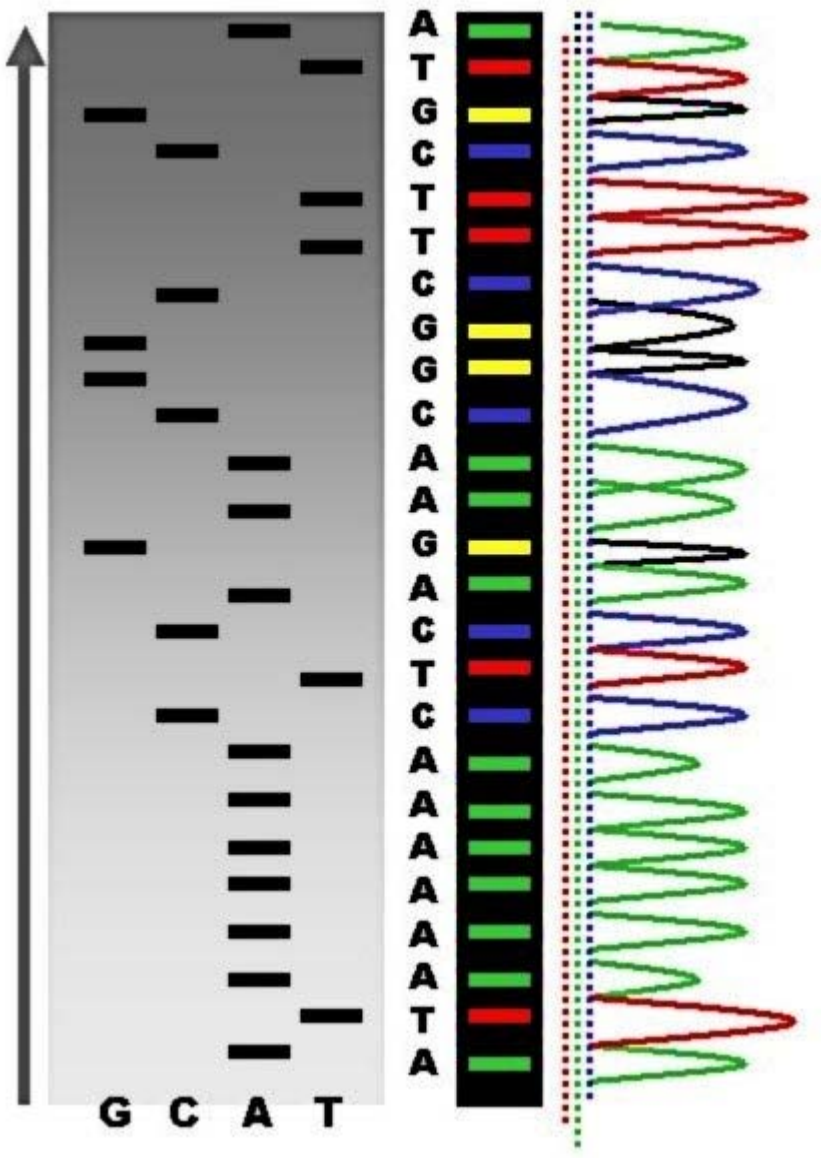
DNA fragments of different lengths. A dark band in a lane indicates a DNA fragment that is the result of chain termination after incorporation of a dideoxynucleotide (ddATP, ddGTP, ddCTP, or ddTTP). The relative positions of the different bands among the four lanes are then used to read (from bottom to top) the DNA sequence.



DNA fragments are labelled with a radioactive or fluorescent tag on the primer (1), in the new DNA strand with a labeled dNTP, or with a labeled ddNTP.

Technical variations of chain-termination sequencing include tagging with nucleotides containing radioactive phosphorus for radiolabelling, or using a primer labeled at the 5' end with a fluorescent dye. Dye-primer sequencing facilitates reading in an optical system for faster and more economical analysis and automation. The later development

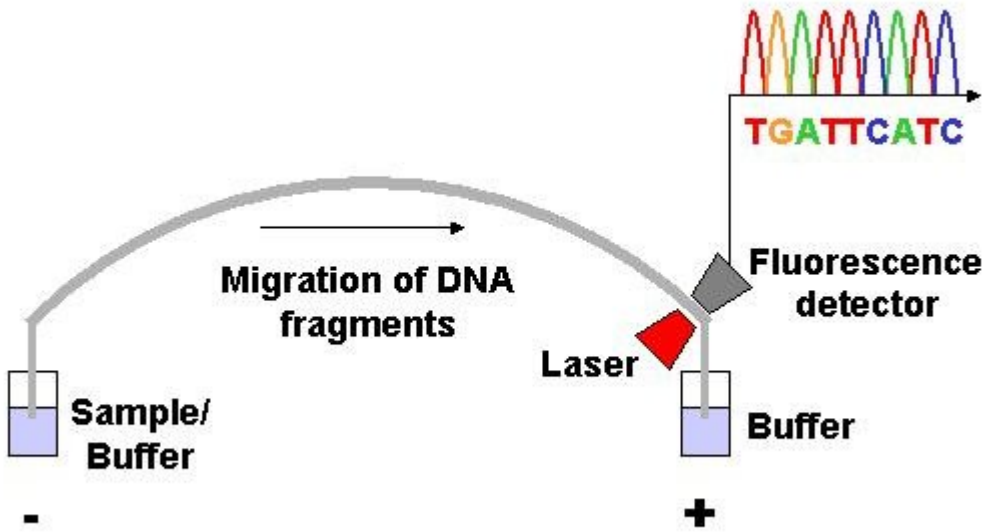
by Leroy Hood and coworkers of fluorescently labeled ddNTPs and primers set the stage for automated, high-throughput DNA sequencing.



Sequence ladder by radioactive sequencing compared to fluorescent peaks

Chain-termination methods have greatly simplified DNA sequencing. For example, chain-termination-based kits are commercially available that contain the reagents needed for sequencing, pre-aliquoted and ready to use. Limitations include non-specific binding of the primer to the DNA, affecting accurate read-out of the DNA sequence, and DNA secondary structures affecting the fidelity of the sequence.

Dye-terminator sequencing



Capillary electrophoresis

Dye-terminator sequencing utilizes labelling of the chain terminator ddNTPs, which permits sequencing in a single reaction, rather than four reactions as in the labelled-primer method. In dye-terminator sequencing, each of the four dideoxynucleotide chain terminators is labelled with fluorescent dyes, each of which emit light at different wavelengths.

Owing to its greater expediency and speed, dye-terminator sequencing is now the mainstay in automated sequencing. Its limitations include dye effects due to differences in the incorporation of the dye-labelled chain terminators into the DNA fragment, resulting in unequal peak heights and shapes in the electronic DNA sequence trace chromatogram after capillary electrophoresis.

This problem has been addressed with the use of modified DNA polymerase enzyme systems and dyes that minimize incorporation variability, as well as methods for eliminating "dye blobs". The dye-terminator sequencing method, along with automated high-throughput DNA sequence analyzers, is now being used for the vast majority of sequencing projects.

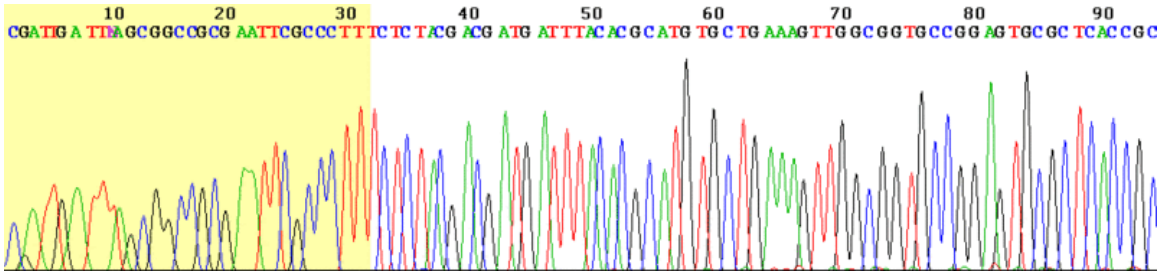
Challenges

Common challenges of DNA sequencing include poor quality in the first 15–40 bases of the sequence and deteriorating quality of sequencing traces after 700–900 bases. Base calling software typically gives an estimate of quality to aid in quality trimming.

In cases where DNA fragments are cloned before sequencing, the resulting sequence may contain parts of the cloning vector. In contrast, PCR-based cloning and emerging sequencing technologies based on pyrosequencing often avoid using cloning vectors. Recently, one-step Sanger sequencing (combined amplification and sequencing) methods such as Ampliseq and SeqSharp have been developed that allow rapid sequencing of target genes without cloning or prior amplification.

Current methods can directly sequence only relatively short (300–1000 nucleotides long) DNA fragments in a single reaction. The main obstacle to sequencing DNA fragments above this size limit is insufficient power of separation for resolving large DNA fragments that differ in length by only one nucleotide. In all cases the use of a primer with a free 5' end is essential.

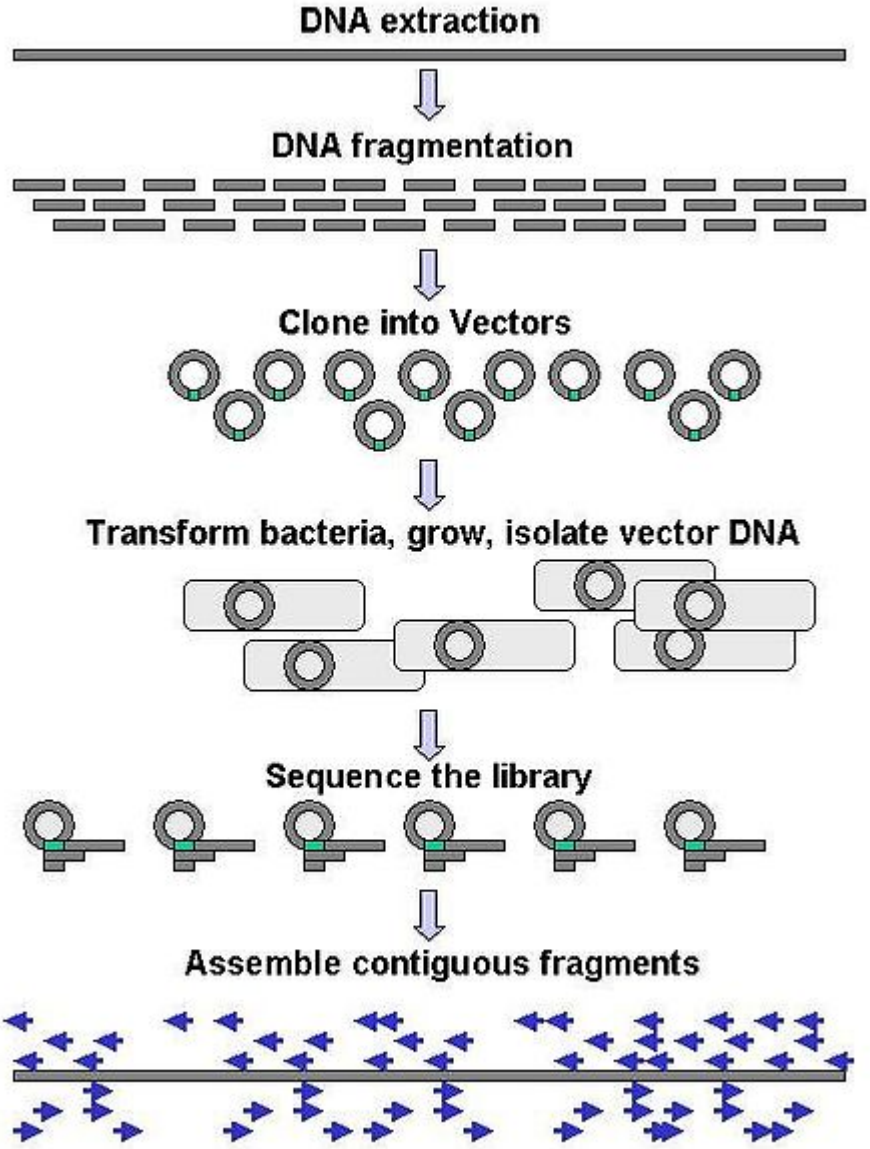
Automation and sample preparation



View of the start of an example dye-terminator read

Automated DNA-sequencing instruments (DNA sequencers) can sequence up to 384 DNA samples in a single batch (run) in up to 24 runs a day. DNA sequencers carry out capillary electrophoresis for size separation, detection and recording of dye fluorescence, and data output as fluorescent peak trace chromatograms. Sequencing reactions by thermocycling, cleanup and re-suspension in a buffer solution before loading onto the sequencer are performed separately. A number of commercial and non-commercial software packages can trim low-quality DNA traces automatically. These programs score the quality of each peak and remove low-quality base peaks (generally located at the ends of the sequence). The accuracy of such algorithms is below visual examination by a human operator, but sufficient for automated processing of large sequence data sets.

Amplification and clonal selection



Genomic DNA is fragmented into random pieces and cloned as a bacterial library. DNA from individual bacterial clones is sequenced and the sequence is assembled by using overlapping DNA regions.

Large-scale sequencing aims at sequencing very long DNA pieces, such as whole chromosomes. Common approaches consist of cutting (with restriction enzymes) or shearing (with mechanical forces) large DNA fragments into shorter DNA fragments. The fragmented DNA is cloned into a DNA vector, and amplified in *Escherichia coli*. Short DNA fragments purified from individual bacterial colonies are individually sequenced and assembled electronically into one long, contiguous sequence.

This method does not require any pre-existing information about the sequence of the DNA and is referred to as *de novo* sequencing. Gaps in the assembled sequence may be filled by primer walking. The different strategies have different tradeoffs in speed and accuracy; *shotgun methods* are often used for sequencing large genomes, but its assembly is complex and difficult, particularly with sequence repeats often causing gaps in genome assembly.

Most sequencing approaches use an *in vitro* cloning step to amplify individual DNA molecules, because their molecular detection methods are not sensitive enough for single molecule sequencing. Emulsion PCR isolates individual DNA molecules along with primer-coated beads in aqueous droplets within an oil phase. Polymerase chain reaction (PCR) then coats each bead with clonal copies of the DNA molecule followed by immobilization for later sequencing. Emulsion PCR is used in the methods by Marguilis et al. (commercialized by 454 Life Sciences), Shendure and Porreca et al. (also known as "Polony sequencing") and SOLiD sequencing, (developed by Agencourt, now Applied Biosystems).

Another method for *in vitro* clonal amplification is *bridge PCR*, where fragments are amplified upon primers attached to a solid surface, used in the Illumina Genome Analyzer. The single-molecule method developed by Stephen Quake's laboratory (later commercialized by Helicos) is an exception: it uses bright fluorophores and laser excitation to detect pyrosequencing events from individual DNA molecules fixed to a surface, eliminating the need for molecular amplification.

High-throughput sequencing

The high demand for low-cost sequencing has driven the development of high-throughput sequencing technologies that parallelize the sequencing process, producing thousands or millions of sequences at once. High-throughput sequencing technologies are intended to lower the cost of DNA sequencing beyond what is possible with standard dye-terminator methods.

Lynx Therapeutics' Massively Parallel Signature Sequencing (MPSS)

The first of the "next-generation" sequencing technologies, MPSS was developed in 1990s at Lynx Therapeutics, a company founded in 1992 by Sidney Brenner and Sam Eletr. MPSS was a bead-based method that used a complex approach of adapter ligation followed by adapter decoding, reading the sequence in increments of four nucleotides; this method made it susceptible to sequence-specific bias or loss of specific sequences. Because the technology was so complex, MPSS was only performed 'in-house' by Lynx Therapeutics and no machines were sold; when the merger with Solexa later lead to the development of sequencing-by-synthesis, a more simple approach with numerous advantages, MPSS became obsolete. However, the essential properties of the MPSS output were typical of later "next-gen" data types, including hundreds of thousands of short DNA sequences. In the case of MPSS, these were typically used for sequencing

cDNA for measurements of gene expression levels. Lynx Therapeutics merged with Solexa in 2004, and this company was later purchased by Illumina.

Polony Sequencing

Polony sequencing, developed in George Church's lab at Harvard, was among the first next-generation sequencing systems used to sequence a full genome in 2005. It combined an in vitro paired-tag library with emulsion PCR, an automated microscope, and ligation-based sequencing chemistry to sequence an E. coli genome at an accuracy of > 99.9999% and a cost approximately 1/10th that of Sanger sequencing. The technology was licensed to Agencourt Biosciences, subsequently spun out into Agencourt Personal Genomics, and ultimately incorporated into the Applied Biosystems SOLiD platform.

454 pyrosequencing

A parallelized version of pyrosequencing was developed by 454 Life Sciences. The method amplifies DNA inside water droplets in an oil solution (emulsion PCR), with each droplet containing a single DNA template attached to a single primer-coated bead that then forms a clonal colony. The sequencing machine contains many picolitre-volume wells each containing a single bead and sequencing enzymes. Pyrosequencing uses luciferase to generate light for detection of the individual nucleotides added to the nascent DNA, and the combined data are used to generate sequence read-outs. This technology provides intermediate read length and price per base compared to Sanger sequencing on one end and Solexa and SOLiD on the other. 454 Life Sciences has since been acquired by Roche Diagnostics.

Illumina (Solexa) sequencing

Solexa, now part of Illumina developed a sequencing technology based on reversible dye-terminators. DNA molecules are first attached to primers on a slide and amplified so that local clonal colonies are formed (bridge amplification). Four types of ddNTPs are added, and non-incorporated nucleotides are washed away. Unlike pyrosequencing, the DNA can only be extended one nucleotide at a time. A camera takes images of the fluorescently labeled nucleotides then the dye along with the terminal 3' blocker is chemically removed from the DNA, allowing a next cycle.

SOLiD sequencing

Applied Biosystems' SOLiD technology employs sequencing by ligation. Here, a pool of all possible oligonucleotides of a fixed length are labeled according to the sequenced position. Oligonucleotides are annealed and ligated; the preferential ligation by DNA ligase for matching sequences results in a signal informative of the nucleotide at that position. Before sequencing, the DNA is amplified by emulsion PCR. The resulting bead, each containing only copies of the same DNA molecule, are deposited on a glass slide. The result is sequences of quantities and lengths comparable to Illumina sequencing.

Future methods

Sequencing by hybridization is a non-enzymatic method that uses a DNA microarray. A single pool of DNA whose sequence is to be determined is fluorescently labeled and hybridized to an array containing known sequences. Strong hybridization signals from a given spot on the array identifies its sequence in the DNA being sequenced. Mass spectrometry may be used to determine mass differences between DNA fragments produced in chain-termination reactions.

DNA sequencing methods currently under development include labeling the DNA polymerase, reading the sequence as a DNA strand transits through nanopores, and microscopy-based techniques, such as AFM or electron microscopy that are used to identify the positions of individual nucleotides within long DNA fragments (>5,000 bp) by nucleotide labeling with heavier elements (e.g., halogens) for visual detection and recording.

In microfluidic Sanger sequencing the entire thermocycling amplification of DNA fragments as well as their separation by electrophoresis is done on a single glass wafer (approximately 10 cm in diameter) thus reducing the reagent usage as well as cost. In some instances researchers have shown that they can increase the throughput of conventional sequencing through the use of microchips. Research will still need to be done in order to make this use of technology effective.

In October 2006, the X Prize Foundation established an initiative to promote the development of full genome sequencing technologies, called the Archon X Prize, intending to award $10 million to "the first Team that can build a device and use it to sequence 100 human genomes within 10 days or less, with an accuracy of no more than one error in every 100,000 bases sequenced, with sequences accurately covering at least 98% of the genome, and at a recurring cost of no more than $10,000 (US) per genome."

Each year NHGRI promotes grants for new research and developments in genomics. 2010 grants and 2011 candidates include continuing work in microfluidic, polony and base-heavy sequencing methodologies

Major landmarks in DNA sequencing

- 1953 Discovery of the structure of the DNA double helix.
- 1972 Development of recombinant DNA technology, which permits isolation of defined fragments of DNA; prior to this, the only accessible samples for sequencing were from bacteriophage or virus DNA.
- 1977 The first complete DNA genome to be sequenced is that of bacteriophage φX174.

- 1977 Allan Maxam and Walter Gilbert publish "DNA sequencing by chemical degradation". Frederick Sanger, independently, publishes "DNA sequencing with chain-terminating inhibitors".

- 1984 Medical Research Council scientists decipher the complete DNA sequence of the Epstein-Barr virus, 170 kb.

- 1986 Leroy E. Hood's laboratory at the California Institute of Technology and Smith announce the first semi-automated DNA sequencing machine.

- 1987 Applied Biosystems markets first automated sequencing machine, the model ABI 370.

- 1990 The U.S. National Institutes of Health (NIH) begins large-scale sequencing trials on *Mycoplasma capricolum*, *Escherichia coli*, *Caenorhabditis elegans*, and *Saccharomyces cerevisiae* (at US$0.75/base).

- 1991 Sequencing of human expressed sequence tags begins in Craig Venter's lab, an attempt to capture the coding fraction of the human genome.

- 1995 Craig Venter, Hamilton Smith, and colleagues at The Institute for Genomic Research (TIGR) publish the first complete genome of a free-living organism, the bacterium Haemophilus influenzae. The circular chromosome contains 1,830,137 bases and its publication in the journal Science marks the first use of whole-genome shotgun sequencing, eliminating the need for initial mapping efforts.

- 1996 Pål Nyrén and his student Mostafa Ronaghi at the Royal Institute of Technology in Stockholm publish their method of pyrosequencing

- 1998 Phil Green and Brent Ewing of the University of Washington publish "phred" for sequencer data analysis.

- 2000 Lynx Therapeutics publishes and markets "MPSS" - a parallelized, adapter/ligation-mediated, bead-based sequencing technology, launching "next-generation" sequencing.

- 2001 A draft sequence of the human genome is published.

- 2004 454 Life Sciences markets a parallelized version of pyrosequencing. The first version of their machine reduced sequencing costs 6-fold compared to automated Sanger sequencing, and was the second of a new generation of sequencing technologies, after MPSS.

Chapter- 6

Neanderthal Genome Project

Max Planck Institute for Evolutionary Anthropology, in Leipzig, Germany

The Neanderthal genome project is a collaboration of scientists coordinated by the Max Planck Institute for Evolutionary Anthropology in Germany and 454 Life Sciences in the United States to sequence the Neanderthal genome.

Founded in July 2006, the project published their results in the May 2010 journal *Science* detailing an initial draft of the Neanderthal genome based on the analysis of four billion base pairs of Neanderthal DNA. The study determined that some mixture of genes occurred between Neanderthals and anatomically modern humans and presented evidence that elements of their genome remain in that of non-African modern humans.

Findings

At roughly 3.2 billion base pairs, the Neanderthal genome is about the size of the modern human genome. According to preliminary sequences, 99.7% of the base pairs of the modern human and Neanderthal genomes are identical, compared to humans sharing around 98.8% of base pairs with the chimpanzee. (Other studies concerning the commonality between chimps and humans have modified the commonality of 98% to a commonality of only 94%, showing that the genetic gap between humans and chimps is bigger than originally thought.) The researchers recovered ancient DNA of Neanderthals by extracting the DNA from the femur bones of three 38,000-year-old female Neanderthal specimens from Vindija Cave, Croatia, and other bones found in Spain, Russia, and Germany. Only about half a gram of the bone samples was required for the sequencing, but the project faced many difficulties, including the contamination of the samples by the bacteria that had colonized the Neanderthal's body and humans who handled the bones at the excavation site and at the laboratory.

Additionally, in 2010, the announcement of the discovery and analysis of Mitochondrial DNA (mtDNA) from the Denisova hominin in Siberia revealed that this specimen differs from that of modern humans by 385 bases (nucleotides) in the mtDNA strand out of approximately 16,500, whereas the difference between modern humans and Neanderthals is around 202 bases. In contrast, the difference between chimpanzees and modern humans is approximately 1,462 mtDNA base pairs. Analysis of the specimen's nuclear DNA is under way and is expected to clarify whether the find is a distinct species. Even though the Denisova hominin's mtDNA lineage predates the divergence of modern humans and Neanderthals, coalescent theory does not preclude a more recent divergence date for her nuclear DNA.

History

In 2006, two research teams working on the same Neanderthal sample published their results, Richard Green and his team in *Nature*, and Noonan et al. in *Science*. The results were received with some criticism, mainly surrounding the issue of a possible admixture of Neanderthals into the modern human genome. The speech-related gene FOXP2 with the same mutations as in modern humans was discovered in ancient DNA in the El Sidrón 1253 and 1351c specimens, suggesting Neanderthals might have shared some basic language capabilities with modern humans.

Svante Pääbo, director of the Department of Genetics at the Max Planck Institute for Evolutionary Anthropology and head of its Neanderthal genome project

In 2006, Richard Green's team had used a then new sequencing technique developed by 454 Life Sciences that amplifies single molecules for characterization and obtained over a quarter-million unique short sequences ("reads"). The technique delivers randomly located reads, so that sequences of interest, e.g. genes that differ between modern humans and Neanderthals, show up at random as well. However, this form of direct sequencing destroys the original sample so to obtain new reads more sample must be destructively sequenced.

Noonan et al., led by Edward Rubin, used a different technique, one in which the Neanderthal DNA is inserted into bacteria, which make multiple copies of a single

fragment. They demonstrated that Neanderthal genomic sequences can be recovered using a metagenomic library-based approach. All of the DNA in the sample is "immortalized" into metagenomic libraries. A DNA fragment is selected, then propagated in microbes. The Neanderthal DNA can be sequenced or specific sequences can be studied.

Overall, their results were remarkably similar. One group suggested there was a hint of mixing between human and Neanderthal genomes, while the other found none, but both teams recognized that the data set was not large enough to give a definitive answer.

The publication by Noonan et al. revealed Neanderthal DNA sequences matching chimpanzee DNA, but not modern human DNA, at multiple locations, thus enabling the first accurate calculation of the date of the most recent common ancestor of *H. sapiens* and *H. neanderthalensis*. The research team estimates the most recent common ancestor of their *H. neanderthalensis* samples and their *H. sapiens* reference sequence lived 706,000 years ago (divergence time), estimating the separation of the human and Neanderthal ancestral populations to 370,000 years ago (split time).

Earlier mitochondrial DNA research led by Pääbo in 1997 had indicated present day *Homo sapiens* and Neanderthals mtDNA split into separate lineages approximately 500,000 years ago.

Green et al. calculated a divergence time of 516,000 years ago and do not indicate a split, while they claim the average divergence time between alleles within humans is thus 459,000 years with a 95% confidence interval between 419,000 and 498,000 years. These two dates (~500k) were calculated with assumption on non-selective pressure. If positive selection forced mtDNA changes then the split time may be shorter. In this study, the team stated:

"Neanderthal genetic differences to humans must therefore be interpreted within the context of human diversity."

On the other hand, Noonan et al. found no evidence of Neanderthal admixture to the modern human genome, but they did not preclude admixture of up to 20% with a certainty better than 95%, and hence did not claim to present a definite answer to the question.

In February 2009, the Max Planck Institute's team, led by geneticist Svante Pääbo, announced that they had completed the first draft of the Neanderthal genome An early analysis of the data suggested in "the genome of Neanderthals, a human species driven to extinction" "no significant trace of Neanderthal genes in modern humans". New results suggested that some adult Neanderthals were lactose intolerant. On the question of potentially cloning a Neanderthal, Pääbo commented, "Starting from the DNA extracted from a fossil, it is and will remain impossible."

In May 2010, the project released a draft of their report on the sequenced Neanderthal Genome. Contradicting the results discovered while examining mitochondrial DNA, they demonstrated a range of genetic contribution to non-African modern humans ranging from 1% to 4%. From their Homo sapiens samples in Eurasia (French, Han Chinese & Papuan) the authors state that it is likely that interbreeding occurred in the Levant before Homo sapiens migrated into Europe. However, this finding is disputed because of the lack of archeological evidence supporting their statement. The fossil evidence does not place Neanderthals and modern humans in close proximity at this time and place.

Criticism

A 2007 review of the data by Wall and Kim reanalyses the data obtained from the published papers of Noonan et al. and Green et al., and it holds that the results are inconsistent with each other. The review proposes serious problems with the data quality in one of the studies, possibly due to modern human DNA contaminants and/or a high rate of sequencing errors.

The reanalyses confirmed both results to the Human-Neanderthal DNA Sequence Divergence Time (common ancestor), that is 706 kya (thousands of years ago) to the Noonan et al. analysis and 516 kya to the Green et al. analysis. The modern European-Neanderthal population split time was estimated at 35 kya for the Green et al. data, and 325 kya for the Noonan et al. data. Before, no split time was estimated by the Green et al. study, and according to Wall and Kim the split time originally estimated by Noonan et al. was even higher: 440 kya (the Noonan et al. paper mentions 370 kya).

While Noonan et al. were unable to definitively conclude that interbreeding between the two species of humans did not occur, they proclaim little likelihood of it having occurred at any appreciable level. The study opts for a 0% contribution of Neanderthal DNA to the modern European gene pool, based on the 95% confidence interval that indicates a margin between 0% and 20% contribution. The reanalyses of Wall and Kim yielded interbreeding margins between 0% and 39% to the data of Noonan et al., and margins between 81% and 100% to the data of Green et al. These vastly inconsistent results could only be reconciled by assuming a very recent split time between the two populations of 60 kya or less. However, such a recent split time would not be consistent with the estimated modern European-Neanderthal population split time from the Noonan et al. data.

The key assumption of Noonan et al. is that the 38,000 years of fossilisation suffered by the Neanderthal DNA should have the genome analysis focus on ancient DNA fragments of about 50 to 70 base pairs in length. Green et al. do not make such an assumption; they generalized towards the exclusion of modern human nuclear DNA contamination by finding little evidence of modern human DNA contamination. Such mitochondrial DNA tends to remain preserved longer than nuclear DNA. However, Wall and Kim noted a length dependence of the results, having the small fragments pointing to a divergence time similar to the results of Noonan et al. and the large fragments much more similar on average to modern human DNA - even to the extent of indicating an estimated human-

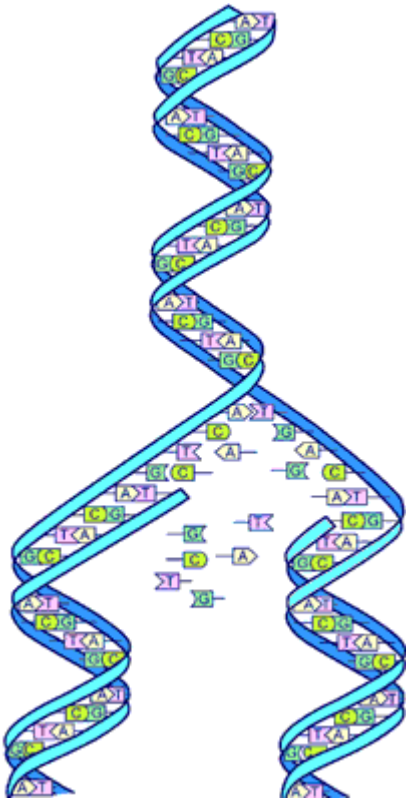
Neanderthal sequence divergence time that is less than the estimated divergence time of two extant members of one referenced population in West Africa. Although Wall and Kim hold modern human contamination to be size-biased, since Neanderthal DNA would be expected to have a tendency to be degraded into short fragments, they noted that length dependence of the results means that alignment issues alone are unlikely to be a sufficient explanation, since longer fragments would be easier to align and thus the data from longer fragments should be more accurate. Still they mark this as a signal of *potential* contamination in the data of Green et al. No similar signal of potential contamination was found in the data of Noonan et al.

Contamination in the data of Green et al. should have decreased the Neanderthal-specific sequence divergence in this study. Since this is not the case, the assumption of contamination also would indicate a higher sequencing error rate in the Green et al. data, since sequence errors would look the same as Neanderthal-specific mutations. These Neanderthal-specific mutations already were considered prone to error due to post-mortem DNA damage in both studies, and were excluded from the results.

In summary, Wall and Kim consider a model with 78% contamination more likely than a model with no contamination and 94% admixture.

Chapter- 7

Human Genome Project



DNA Replication

The **Human Genome Project (HGP)** is an international scientific research project with a primary goal of determining the sequence of chemical base pairs which make up DNA and to identify and map the approximately 20,000–25,000 genes of the human genome from both a physical and functional standpoint.

The project began in 1990 and was initially headed by Ari Patrinos, head of the Office of Biological and Environmental Research in the U.S. Department of Energy's Office of Science. Francis Collins directed the National Institutes of Health National Human Genome Research Institute efforts. A working draft of the genome was announced in 2000 and a complete one in 2003, with further, more detailed analysis still being published. A parallel project was conducted outside of government by the Celera

Corporation, which was formally launched in 1998. Most of the government-sponsored sequencing was performed in universities and research centers from the United States, the United Kingdom, Japan, France, Germany, and China. The mapping of human genes is an important step in the development of medicines and other aspects of health care.

While the objective of the Human Genome Project is to understand the genetic makeup of the human species, the project has also focused on several other nonhuman organisms such as *E. coli*, the fruit fly, and the laboratory mouse. It remains one of the largest single investigative projects in modern science.

The Human Genome Project originally aimed to map the nucleotides contained in a human haploid reference genome (more than three billion). Several groups have announced efforts to extend this to diploid human genomes including the International HapMap Project, Applied Biosystems, Perlegen, Illumina, JCVI, Personal Genome Project, and Roche-454.

The "genome" of any given individual (except for identical twins and cloned organisms) is unique; mapping "the human genome" involves sequencing multiple variations of each gene. The project did not study the entire DNA found in human cells; some heterochromatic areas (about 8% of the total genome) remain un-sequenced.

Project

Background

The project began with the culmination of several years of work supported by the United States Department of Energy, in particular workshops in 1984 and 1986 and a subsequent initiative of the US Department of Energy. This 1987 report stated boldly, "The ultimate goal of this initiative is to understand the human genome" and "knowledge of the human as necessary to the continuing progress of medicine and other health sciences as knowledge of human anatomy has been for the present state of medicine." Candidate technologies were already being considered for the proposed undertaking at least as early as 1985.

James D. Watson was head of the National Center for Human Genome Research at the National Institutes of Health (NIH) in the United States starting from 1988. Largely due to his disagreement with his boss, Bernadine Healy, over the issue of patenting genes, Watson was forced to resign in 1992. He was replaced by Francis Collins in April 1993, and the name of the Center was changed to the National Human Genome Research Institute (NHGRI) in 1997.

The $3-billion project was formally founded in 1990 by the United States Department of Energy and the U.S. National Institutes of Health, and was expected to take 15 years. In addition to the United States, the international consortium comprised geneticists in the United Kingdom, France, Germany, Japan, China, and India.

Due to widespread international cooperation and advances in the field of genomics (especially in sequence analysis), as well as major advances in computing technology, a 'rough draft' of the genome was finished in 2000 (announced jointly by then US president Bill Clinton and the British Prime Minister Tony Blair on June 26, 2000). This first available rough draft assembly of the genome was completed by the UCSC Genome Bioinformatics Group, primarily led by then graduate student Jim Kent. Ongoing sequencing led to the announcement of the essentially complete genome in April 2003, 2 years earlier than planned. In May 2006, another milestone was passed on the way to completion of the project, when the sequence of the last chromosome was published in the journal Nature.

State of completion

There are multiple definitions of the "complete sequence of the human genome". According to some of these definitions, the genome has already been completely sequenced, and according to other definitions, the genome has yet to be completely sequenced. There have been multiple popular press articles reporting that the genome was "complete." The genome has been completely sequenced using the definition employed by the International Human Genome Project. A graphical history of the human genome project shows that most of the human genome was complete by the end of 2003. However, there are a number of regions of the human genome that can be considered unfinished:

- First, the central regions of each chromosome, known as centromeres, are highly repetitive DNA sequences that are difficult to sequence using current technology. The centromeres are millions (possibly tens of millions) of base pairs long, and for the most part these are entirely un-sequenced.
- Second, the ends of the chromosomes, called telomeres, are also highly repetitive, and for most of the 46 chromosome ends these too are incomplete. It is not known precisely how much sequence remains before the telomeres of each chromosome are reached, but as with the centromeres, current technological restraints are prohibitive.
- Third, there are several loci in each individual's genome that contain members of multigene families that are difficult to disentangle with shotgun sequencing methods – these multigene families often encode proteins important for immune functions.
- Other than these regions, there remain a few dozen gaps scattered around the genome, some of them rather large, but there is hope that all these will be closed in the next couple of years.

In summary: the best estimates of total genome size indicate that about 92.3% of the genome has been completed and it is likely that the centromeres and telomeres will remain un-sequenced until new technology is developed that facilitates their sequencing. Most of the remaining DNA is highly repetitive and unlikely to contain genes, but it cannot be truly known until it is entirely sequenced. Understanding the functions of all the genes and their regulation is far from complete. The roles of junk DNA, the evolution

of the genome, the differences between individuals, and many other questions are still the subject of intense interest by laboratories all over the world.

Goals

The sequence of the human DNA is stored in databases available to anyone on the Internet. The U.S. National Center for Biotechnology Information (and sister organizations in Europe and Japan) house the gene sequence in a database known as GenBank, along with sequences of known and hypothetical genes and proteins. Other organizations such as the University of California, Santa Cruz, and Ensembl present additional data and annotation and powerful tools for visualizing and searching it. Computer programs have been developed to analyze the data, because the data itself is difficult to interpret without such programs.

The process of identifying the boundaries between genes and other features in a raw DNA sequence is called genome annotation and is the domain of bioinformatics. While expert biologists make the best annotators, their work proceeds slowly, and computer programs are increasingly used to meet the high-throughput demands of genome sequencing projects. The best current technologies for annotation make use of statistical models that take advantage of parallels between DNA sequences and human language, using concepts from computer science such as formal grammars.

Another, often overlooked, goal of the HGP is the study of its ethical, legal, and social implications. It is important to research these issues and find the most appropriate solutions before they become large dilemmas whose effect will manifest in the form of major political concerns.

All humans have unique gene sequences. Therefore the data published by the HGP does not represent the exact sequence of each and every individual's genome. It is the combined "reference genome" of a small number of anonymous donors. The HGP genome is a scaffold for future work in identifying differences among individuals. Most of the current effort in identifying differences among individuals involves single-nucleotide polymorphisms and the HapMap.

Findings

Key findings of the draft (2001) and complete (2004) genome sequences include

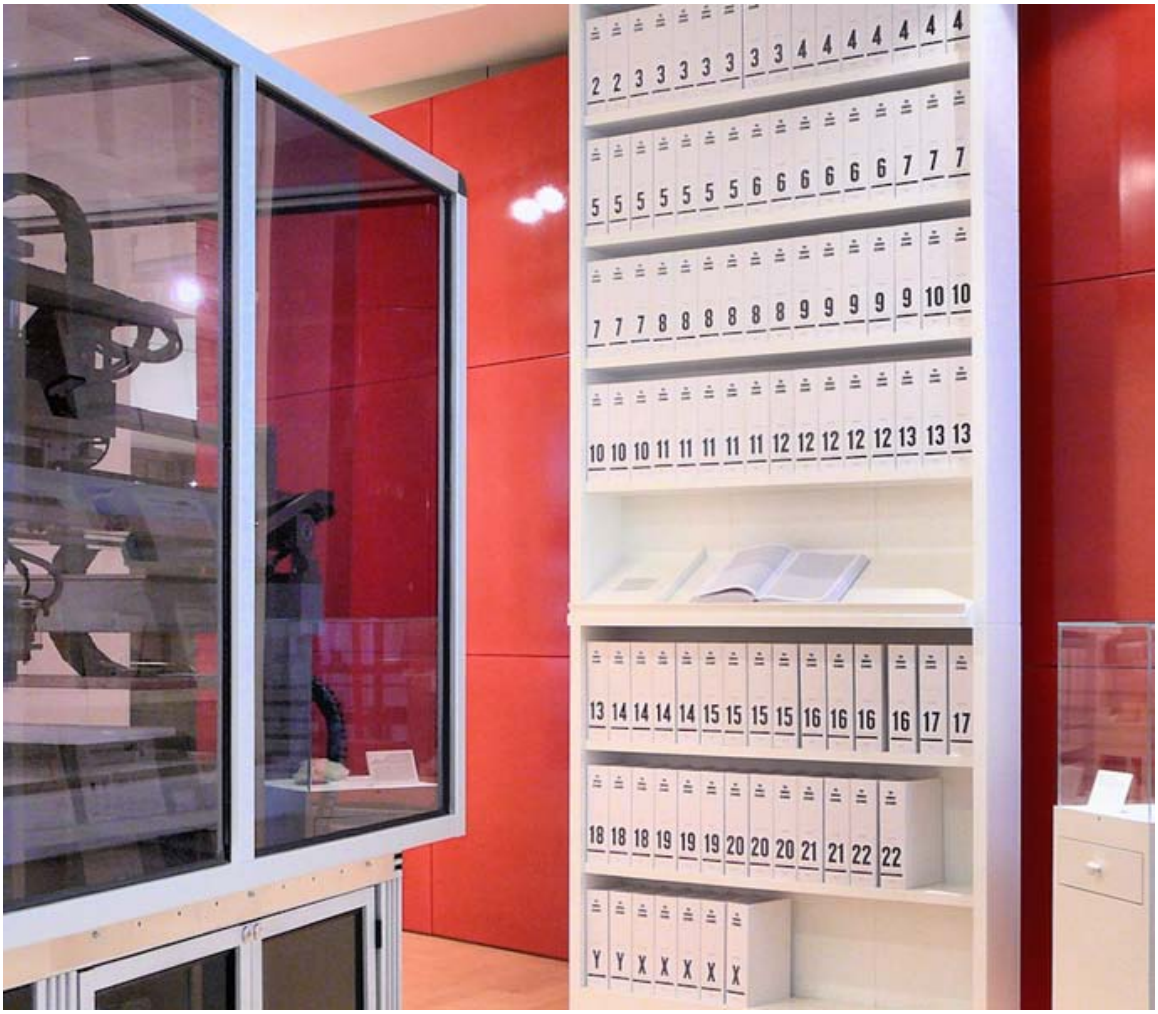
1. There are approximately 20,500 genes in human beings, the same range as in mice and twice that of roundworms. Understanding how these genes express themselves will provide clues to how diseases are caused.

2. Between 1.1% to 1.4% of the genome's sequence codes for proteins

3. The human genome has significantly more segmental duplications (nearly identical, repeated sections of DNA) than other mammalian genomes. These sections may underlie the creation of new primate-specific genes

4. At the time when the draft sequence was published less than 7% of protein families appeared to be vertebrate specific

How it was accomplished



The first printout of the human genome to be presented as a series of books, displayed at the Wellcome Collection, London

The Human Genome Project was started in 1989 with the goal of sequencing and identifying all three billion chemical units in the human genetic instruction set, finding the genetic roots of disease and then developing treatments. With the sequence in hand, the next step was to identify the genetic variants that increase the risk for common diseases like cancer and diabetes.

It was far too expensive at that time to think of sequencing patients' whole genomes. So the National Institutes of Health embraced the idea for a "shortcut", which was to look just at sites on the genome where many people have a variant DNA unit. The theory behind the shortcut was that since the major diseases are common, so too would be the genetic variants that caused them. Natural selection keeps the human genome free of variants that damage health before children are grown, the theory held, but fails against variants that strike later in life, allowing them to become quite common. (In 2002 the National Institutes of Health started a $138 million project called the HapMap to catalog the common variants in European, East Asian and African genomes.)

The genome was broken into smaller pieces; approximately 150,000 base pairs in length. These pieces were then ligated into a type of vector known as "bacterial artificial chromosomes", or BACs, which are derived from bacterial chromosomes which have been genetically engineered. The vectors containing the genes can be inserted into bacteria where they are copied by the bacterial DNA replication machinery. Each of these pieces was then sequenced separately as a small "shotgun" project and then assembled. The larger, 150,000 base pairs go together to create chromosomes. This is known as the "hierarchical shotgun" approach, because the genome is first broken into relatively large chunks, which are then mapped to chromosomes before being selected for sequencing.

Funding came from the US government through the National Institutes of Health in the United States, and a UK charity organization, the Wellcome Trust, as well as numerous other groups from around the world. The funding supported a number of large sequencing centers including those at Whitehead Institute, the Sanger Centre, Washington University in St. Louis, and Baylor College of Medicine.

The Human Genome Project is considered a Mega Project because the human genome has approximately 3.3 billion base-pairs.

If the sequence obtained was to be stored in book form, and if each page contained 1000 base-pairs recorded and each book contained 1000 pages, then 3300 such books would be needed in order to store the complete genome. However, if expressed in units of computer data storage, 3.3 billion base-pairs recorded at 2 bits per pair would equal 786 megabytes of raw data. This is comparable to a fully data loaded CD.

Public versus private approaches

In 1998, a similar, privately funded quest was launched by the American researcher Craig Venter, and his firm Celera Genomics. Venter was a scientist at the NIH during the early 1990s when the project was initiated. The $300,000,000 Celera effort was intended to proceed at a faster pace and at a fraction of the cost of the roughly $3 billion publicly funded project.

Celera used a technique called whole genome shotgun sequencing, employing pairwise end sequencing, which had been used to sequence bacterial genomes of up to six million

base pairs in length, but not for anything nearly as large as the three billion base pair human genome.

Celera initially announced that it would seek patent protection on "only 200–300" genes, but later amended this to seeking "intellectual property protection" on "fully-characterized important structures" amounting to 100–300 targets. The firm eventually filed preliminary ("place-holder") patent applications on 6,500 whole or partial genes. Celera also promised to publish their findings in accordance with the terms of the 1996 "Bermuda Statement," by releasing new data annually (the HGP released its new data daily), although, unlike the publicly funded project, they would not permit free redistribution or scientific use of the data. The publicly funded competitor UC Santa Cruz was compelled to publish the first draft of the human genome before Celera for this reason. On July 7, 2000, the UCSC Genome Bioinformatics Group released a first working draft on the web. The scientific community downloaded one-half trillion bytes of information from the UCSC genome server in the first 24 hours of free and unrestricted access to the first ever assembled blueprint of our human species.

In March 2000, President Clinton announced that the genome sequence could not be patented, and should be made freely available to all researchers. The statement sent Celera's stock plummeting and dragged down the biotechnology-heavy Nasdaq. The biotechnology sector lost about $50 billion in market capitalization in two days.

Although the working draft was announced in June 2000, it was not until February 2001 that Celera and the HGP scientists published details of their drafts. Special issues of *Nature* (which published the publicly funded project's scientific paper) and *Science* (which published Celera's paper) described the methods used to produce the draft sequence and offered analysis of the sequence. These drafts covered about 83% of the genome (90% of the euchromatic regions with 150,000 gaps and the order and orientation of many segments not yet established). In February 2001, at the time of the joint publications, press releases announced that the project had been completed by both groups. Improved drafts were announced in 2003 and 2005, filling in to ≈92% of the sequence currently.

The competition proved to be very good for the project, spurring the public groups to modify their strategy in order to accelerate progress. The rivals at UC Santa Cruz initially agreed to pool their data, but the agreement fell apart when Celera refused to deposit its data in the unrestricted public database GenBank. Celera had incorporated the public data into their genome, but forbade the public effort to use Celera data.

HGP is the most well known of many international genome projects aimed at sequencing the DNA of a specific organism. While the human DNA sequence offers the most tangible benefits, important developments in biology and medicine are predicted as a result of the sequencing of model organisms, including mice, fruit flies, zebrafish, yeast, nematodes, plants, and many microbial organisms and parasites.

In 2004, researchers from the International Human Genome Sequencing Consortium (IHGSC) of the HGP announced a new estimate of 20,000 to 25,000 genes in the human genome. Previously 30,000 to 40,000 had been predicted, while estimates at the start of the project reached up to as high as 2,000,000. The number continues to fluctuate and it is now expected that it will take many years to agree on a precise value for the number of genes in the human genome.

History

In 1976, the genome of the RNA virus Bacteriophage MS2 was the first complete genome to be determined, by Walter Fiers and his team at the University of Ghent (Ghent, Belgium). The idea for the shotgun technique came from the use of an algorithm that combined sequence information from many small fragments of DNA to reconstruct a genome. This technique was pioneered by Frederick Sanger to sequence the genome of the Phage Φ-X174, a virus (bacteriophage) that primarily infects bacteria that was the first fully sequenced genome (DNA-sequence) in 1977. The technique was called shotgun sequencing because the genome was broken into millions of pieces as if it had been blasted with a shotgun. In order to scale up the method, both the sequencing and genome assembly had to be automated, as they were in the 1980s.

Those techniques were shown applicable to sequencing of the first free-living bacterial genome (1.8 million base pairs) of *Haemophilus influenzae* in 1995 and the first animal genome (~100 Mbp) It involved the use of automated sequencers, longer individual sequences using approximately 500 base pairs at that time. Paired sequences separated by a fixed distance of around 2000 base pairs which were critical elements enabling the development of the first genome assembly programs for reconstruction of large regions of genomes (aka 'contigs').

Three years later, in 1998, the announcement by the newly-formed Celera Genomics that it would scale up the pairwise end sequencing method to the human genome was greeted with skepticism in some circles. The shotgun technique breaks the DNA into fragments of various sizes, ranging from 2,000 to 300,000 base pairs in length, forming what is called a DNA "library". Using an automated DNA sequencer the DNA is read in 800bp lengths from both ends of each fragment. Using a complex genome assembly algorithm and a supercomputer, the pieces are combined and the genome can be reconstructed from the millions of short, 800 base pair fragments. The success of both the public and privately funded effort hinged upon a new, more highly automated capillary DNA sequencing machine, called the Applied Biosystems 3700, that ran the DNA sequences through an extremely fine capillary tube rather than a flat gel. Even more critical was the development of a new, larger-scale genome assembly program, which could handle the 30–50 million sequences that would be required to sequence the entire human genome with this method. At the time, such a program did not exist. One of the first major projects at Celera Genomics was the development of this assembler, which was written in parallel with the construction of a large, highly automated genome sequencing factory. Development of the assembler was led by Brian Ramos. The first version of this assembler was demonstrated in 2000, when the Celera team joined forces with Professor

Gerald Rubin to sequence the fruit fly Drosophila melanogaster using the whole-genome shotgun method. At 130 million base pairs, it was at least 10 times larger than any genome previously shotgun assembled. One year later, the Celera team published their assembly of the three billion base pair human genome.

The Human Genome Project was a 13 year old mega project, that was launched in the year 1990 and completed in 2003. This project is closely associated to the branch of biology called Bio-informatics. The human genome project international consortium announced the publication of a draft sequence and analysis of the human genome—the genetic blueprint for the human being. An American company—Celera, led by Craig Venter and the other huge international collaboration of distinguished scientists led by Francis Collins, director, National Human Genome Research Institute, U.S., both published their findings.

This Mega Project is co-ordinated by the U.S. Department of Energy and the National Institute of Health. During the early years of the project, the Wellcome Trust (U.K.) became a major partner, other countries like Japan, Germany, China and France contributed significantly. Already the atlas has revealed some starting facts. The two factors that made this project a success are:

1. Genetic Engineering Techniques, with which it is possible to isolate and clone any segment of DNA.
2. Availability of simple and fast technologies, to determining the DNA sequences.

Being the most complex organisms, human beings were expected to have more than 100,000 genes or combination of DNA that provides commands for every characteristics of the body. Instead their studies show that humans have only 30,000 genes – around the same as mice, three times as many as flies, and only five times more than bacteria. Scientist told that not only are the numbers similar, the genes themselves, baring a few, are alike in mice and men. In a companion volume to the Book of Life, scientists have created a catalogue of 1.4 million single-letter differences, or single-nucleotide polymorphisms (SNPs) – and specified their exact locations in the human genome. This SNP map, the world's largest publicly available catalogue of SNP's, promises to revolutionize both mapping diseases and tracing human history. The sequence information from the consortium has been immediately and freely released to the world, with no restrictions on its use or redistribution. The information is scanned daily by scientists in academia and industry, as well as commercial database companies, providing key information services to bio-technologists. Already, many genes have been identified from the genome sequence, including more than 30 that play a direct role in human diseases. By dating the three millions repeat elements and examining the pattern of interspersed repeats on the Y-chromosome, scientists estimated the relative mutation rates in the X and the Y chromosomes and in the male and the female germ lines. They found that the ratio of mutations in male Vs female is 2:1. Scientists point to several possible reasons for the higher mutation rate in the male germ line, including the fact that there are a greater number of cell divisions involved in the formation of sperm than in the formation of eggs.

Methods

The IHGSC used pair-end sequencing plus whole-genome shotgun mapping of large (≈100 Kbp) plasmid clones and shotgun sequencing of smaller plasmid sub-clones plus a variety of other mapping data to orient and check the assembly of each human chromosome.

The Celera group emphasized the importance of the "whole-genome shotgun" sequencing method, relying on sequence information to orient and locate their fragments within the chromosome. However they used the publicly available data from HGP to assist in the assembly and orientation process, raising concerns that the Celera sequence was not independently derived.

Genome donors

In the IHGSC international public-sector Human Genome Project (HGP), researchers collected blood (female) or sperm (male) samples from a large number of donors. Only a few of many collected samples were processed as DNA resources. Thus the donor identities were protected so neither donors nor scientists could know whose DNA was sequenced. DNA clones from many different libraries were used in the overall project, with most of those libraries being created by Dr. Pieter J. de Jong. It has been informally reported, and is well known in the genomics community, that much of the DNA for the public HGP came from a single anonymous male donor from Buffalo, New York (code name RP11).

HGP scientists used white blood cells from the blood of two male and two female donors (randomly selected from 20 of each) -- each donor yielding a separate DNA library. One of these libraries (RP11) was used considerably more than others, due to quality considerations. One minor technical issue is that male samples contain just over half as much DNA from the sex chromosomes (one X chromosome and one Y chromosome) compared to female samples (which contain two X chromosomes). The other 22 chromosomes (the autosomes) are the same for both genders.

Although the main sequencing phase of the HGP has been completed, studies of DNA variation continue in the International HapMap Project, whose goal is to identify patterns of single-nucleotide polymorphism (SNP) groups (called haplotypes, or "haps"). The DNA samples for the HapMap came from a total of 270 individuals: Yoruba people in Ibadan, Nigeria; Japanese people in Tokyo; Han Chinese in Beijing; and the French Centre d'Etude du Polymorphisms Humain (CEf) resource, which consisted of residents of the United States having ancestry from Western and Northern Europe.

In the Celera Genomics private-sector project, DNA from five different individuals were used for sequencing. The lead scientist of Celera Genomics at that time, Craig Venter, later acknowledged (in a public letter to the journal Science) that his DNA was one of 21 samples in the pool, five of which were selected for use.

On September 4, 2007, a team led by Craig Venter published his complete DNA sequence, unveiling the six-billion-nucleotide genome of a single individual for the first time.

Benefits

The work on interpretation of genome data is still in its initial stages. It is anticipated that detailed knowledge of the human genome will provide new avenues for advances in medicine and biotechnology. Clear practical results of the project emerged even before the work was finished. For example, a number of companies, such as Myriad Genetics started offering easy ways to administer genetic tests that can show predisposition to a variety of illnesses, including breast cancer, disorders of hemostasis, cystic fibrosis, liver diseases and many others. Also, the etiologies for cancers, Alzheimer's disease and other areas of clinical interest are considered likely to benefit from genome information and possibly may lead in the long term to significant advances in their management.

There are also many tangible benefits for biological scientists. For example, a researcher investigating a certain form of cancer may have narrowed down his/her search to a particular gene. By visiting the human genome database on the World Wide Web, this researcher can examine what other scientists have written about this gene, including (potentially) the three-dimensional structure of its product, its function(s), its evolutionary relationships to other human genes, or to genes in mice or yeast or fruit flies, possible detrimental mutations, interactions with other genes, body tissues in which this gene is activated, diseases associated with this gene or other datatypes.

Further, deeper understanding of the disease processes at the level of molecular biology may determine new therapeutic procedures. Given the established importance of DNA in molecular biology and its central role in determining the fundamental operation of cellular processes, it is likely that expanded knowledge in this area will facilitate medical advances in numerous areas of clinical interest that may not have been possible without them.

The analysis of similarities between DNA sequences from different organisms is also opening new avenues in the study of evolution. In many cases, evolutionary questions can now be framed in terms of molecular biology; indeed, many major evolutionary milestones (the emergence of the ribosome and organelles, the development of embryos with body plans, the vertebrate immune system) can be related to the molecular level. Many questions about the similarities and differences between humans and our closest relatives (the primates, and indeed the other mammals) are expected to be illuminated by the data from this project.

The Human Genome Diversity Project (HGDP), spinoff research aimed at mapping the DNA that varies between human ethnic groups, which was rumored to have been halted, actually did continue and to date has yielded new conclusions. In the future, HGDP could possibly expose new data in disease surveillance, human development and anthropology. HGDP could unlock secrets behind and create new strategies for managing the

vulnerability of ethnic groups to certain diseases. It could also show how human populations have adapted to these vulnerabilities.

Advantages of Human Genome Project:

1. Knowledge of the effects of variation of DNA among individuals can revolutionize the ways to diagnose, treat and even prevent a number of diseases that affects the human beings.
2. It provides clues to the understanding of human biology.

Ethical, legal and social issues

The project's goals included not only identifying all of the approximately 24,000 genes in the human genome, but also to address the ethical, legal, and social issues (ELSI) that might arise from the availability of genetic information. Five percent of the annual budget was allocated to address the ELSI arising from the project.

Debra Harry, Executive Director of the U.S group Indigenous Peoples Council on Biocolonialism (IPCB), says that despite a decade of ELSI funding, the burden of genetics education has fallen on the tribes themselves to understand the motives of Human genome project and its potential impacts on their lives. Meanwhile, the government has been busily funding projects studying indigenous groups without any meaningful consultation with the groups.

The main criticism of ELSI is the failure to address the conditions raised by population-based research, especially with regard to unique processes for group decision-making and cultural worldviews. Genetic variation research such as HGP is group population research, but most ethical guidelines, according to Harry, focus on individual rights instead of group rights. She says the research represents a clash of culture: indigenous people's life revolves around collectivity and group decision making whereas the Western culture promotes individuality. Harry suggests that one of the challenges of ethical research is to include respect for collective review and decision making, while also upholding the Western model of individual rights.

Chapter- 8

Genetic Engineering

Genetic engineering, also called **genetic modification**, is the direct human manipulation of an organism's genetic material in a way that does not occur under natural conditions. It involves the use of recombinant DNA techniques, but does not include traditional animal and plant breeding or mutagenesis. Any organism that is generated using these techniques is considered to be a genetically modified organism. The first organisms genetically engineered were bacteria in 1973 and then mice in 1974. Insulin producing bacteria were commercialized in 1982 and genetically modified food has been sold since 1994.

The most common form of genetic engineering involves the insertion of new genetic material at an unspecified location in the host genome. This is accomplished by isolating and copying the genetic material of interest, generating a construct containing all the genetic elements for correct expression, and then inserting this construct into the host organism. Other forms of genetic engineering include gene targeting and knocking out specific genes via engineered nucleases such as zinc finger nucleases or engineered homing endonucleases.

Genetic engineering techniques have been applied in numerous fields including research, biotechnology, and medicine. Medicines such as insulin and human growth hormone are now produced in bacteria, experimental mice such as the oncomouse and the knockout mouse are being used for research purposes and insect resistant and/or herbicide tolerant crops have been commercialized. Genetically engineered plants and animals capable of producing biotechnology drugs more cheaply than current methods (called pharming) are also being developed and in 2009 the FDA approved the sale of the pharmaceutical protein antithrombin produced in the milk of genetically engineered goats.

Definition

Genetic engineering alters the genetic makeup of an organism using techniques that introduce heritable material prepared outside the organism either directly into the host or into a cell that is then fused or hybridized with the host. This involves using recombinant nucleic acid (DNA or RNA) techniques to form new combinations of heritable genetic material followed by the incorporation of that material either indirectly through a vector system or directly through micro-injection, macro-injection and micro-encapsulation

techniques. Genetic engineering does not include traditional animal and plant breeding, in vitro fertilisation, induction of polyploidy, mutagenesis and cell fusion techniques that do not use recombinant nucleic acids or a genetically modified organism in the process. Cloning and stem cell research, although not considered genetic engineering, are closely related and genetic engineering can be used within them. Synthetic biology is an emerging discipline that takes genetic engineering a step further by introducing artificially synthesized genetic material from raw materials into an organism.

If genetic material from another species is added to the host, the resulting organism is called transgenic. If genetic material from the same species or a species that can naturally breed with the host is used the resulting organism is called cisgenic. Genetic engineering can also be used to remove genetic material from the target organism, creating a knock out organism. In Europe genetic modification is synonymous with genetic engineering while within the United States of America it can also refer to conventional breeding methods.

History

Humans have altered the genomes of species for thousands of years through artificial selection and more recently mutagenesis. Genetic engineering as the direct manipulation of DNA by humans outside breeding and mutations has only existed since the 1970s. The term "genetic engineering" was first coined by Jack Williamson in his science fiction novel *Dragon's Island*, published in 1951, one year before DNA's role in heredity was confirmed by Alfred Hershey and Martha Chase, and two years before James Watson and Francis Crick showed that the DNA molecule has a double-helix structure.

In 1972 Paul Berg created the first recombinant DNA molecules by combined DNA from the monkey virus SV40 with that of the lambda virus. In 1973 Herbert Boyer and Stanley Cohen created the first transgenic organism by inserting antibiotic resistance genes into the plasmid of an *E. coli* bacterium. A year later Rudolf Jaenisch created a transgenic mouse by introducing foreign DNA into its embryo, making it the world’s first transgenic animal. In 1976 Genentech, the first genetic engineering company was founded by Herbert Boyer and Robert Swanson and a year later and the company produced a human protein (somatostatin) in *E.coli*. Genentech announced the production of genetically engineered human insulin in 1978. In 1980, the U.S. Supreme Court in the Diamond v. Chakrabarty case ruled that genetically altered life could be patented. The insulin produced by bacteria, branded humulin, was approved for release by the Food and Drug Administration in 1982.

The first field trials of genetically engineered plants occurred in France and the USA in 1986, tobacco plants were engineered to be resistant to herbicides. The People’s Republic of China was the first country to commercialize transgenic plants, introducing a virus-resistant tobacco in 1992. In 1994 Calgene attained approval to commercially release the Flavr Savr tomato, a tomato engineered to have a longer shelf life. In 1994, the European Union approved tobacco engineered to be resistant to the herbicide bromoxynil, making it the first genetically engineered crop commercialized in Europe. In 1995, Bt Potato was

approved safe by the Environmental Protection Agency, making it the first pesticide producing crop to be approved in the USA. In 2009 11 transgenic crops were grown commercially in 25 countries, the largest of which by area grown were the USA, Brazil, Argentina, India, Canada, China, Paraguay and South Africa.

In 2010, scientists at the J. Craig Venter Institute, announced that they had created the first synthetic bacterial genome, and added it to a cell containing no DNA. The resulting bacterium, named Synthia, was the world's first synthetic life form.

Process

Isolating the Gene

Elements of genetic engineering

First, the gene to be inserted into the genetically modified organism must be chosen and isolated. Presently, most genes transferred into plants provide protection against insects or tolerance to herbicides. In animals the majority of genes used are growth hormone genes. Once chosen the genes must be isolated. This typically involves multiplying the gene using polymerase chain reaction (PCR). If the chosen gene or the donor organism's genome has been well studied it may be present in a genetic library. If the DNA sequence is known, but no copies of the gene are available, it can be artificially synthesized. Once isolated, the gene is inserted into a bacterial plasmid.

Constructs

The gene to be inserted into the genetically modified organism must be combined with other genetic elements in order for it to work properly. The gene can also be modified at this stage for better expression or effectiveness. As well as the gene to be inserted most constructs contain a promoter and terminator region as well as a selectable marker gene. The promoter region initiates transcription of the gene and can be used to control the location and level of gene expression, while the terminator region ends transcription. The selectable marker, which in most cases confers antibiotic resistance to the organism it is expressed in, is needed to determine which cells are transformed with the new gene. The constructs are made using recombinant DNA techniques, such as restriction digests, ligations and molecular cloning.

Gene Targeting

The most common form of genetic engineering involves inserting new genetic material randomly within the host genome. Other techniques allow new genetic material to be inserted at a specific location in the host genome or generate mutations at desired genomic loci capable of knocking out endogenous genes. The technique of gene targeting uses homologous recombination to target desired changes to a specific endogenous gene. This tends to occur at a relatively low frequency in plants and animals and generally requires the use of selectable markers. The frequency of gene targeting can be greatly enhanced with the use of engineered nucleases such as zinc finger nucleases, engineered homing endonucleases, or nucleases created from TAL effectors. In addition to enhancing gene targeting, engineered nucleases can also be used to introduce mutations at endogenous genes that generate a gene knockout.

Transformation

A. tumefaciens attaching itself to a carrot cell

About 1% of bacteria are naturally able to take up foreign DNA but it can also be induced in other bacteria. Stressing the bacteria for example, with a heat shock or an electric shock, can make the cell membrane permeable to DNA that may then incorporate into their genome or exist as extrachromosomal DNA. DNA is generally inserted into animal cells using microinjection, where it can be injected through the cells nuclear envelope directly into the nucleus or through the use of viral vectors. In plants the DNA is generally inserted using *Agrobacterium*-mediated recombination or biolistics.

In *Agrobacterium*-mediated recombination the plasmid construct must also contain T-DNA. *Agrobacterium* naturally inserts DNA from a tumor inducing plasmid into any susceptible plant's genome it infects, causing crown gall disease. The T-DNA region of this plasmid is responsible for insertion of the DNA. The genes to be inserted are cloned into a binary vector, which contains T-DNA and can be grown in both *E. Coli* and *Agrobacterium*. Once the binary vector is constructed the plasmid is transformed into *Agrobacterium* containing no plasmids and plant cells are infected. The *Agrobacterium* will then naturally insert the genetic material into the plant cells.

In biolistics particles of gold or tungsten are coated with DNA and then shot into young plant cells or plant embryos. Some genetic material will enter the cells and transform them. This method can be used on plants that are not susceptible to *Agrobacterium* infection and also allows transformation of plant plastids. Another transformation method for plant and animal cells is electroporation. Electroporation involves subjecting the plant or animal cell to an electric shock, which can make the cell membrane permeable to plasmid DNA. In some cases the electroporated cells will incorporate the DNA into their genome. Due to the damage caused to the cells and DNA the transformation efficiency of biolistics and electroporation is lower than agrobacterial mediated transformation and microinjection.

Selection

Not all the organism's cells will be transformed with the new genetic material; in most cases a selectable marker is used to differentiate transformed from untransformed cells. If a cell has been successfully transformed with the DNA it will also contain the marker gene. By growing the cells in the presence of an antibiotic or chemical that selects or marks the cells expressing that gene it is possible to separate the transgenic events from the non-transgenic. Another method of screening involves using a DNA probe that will only stick to the inserted gene. A number of strategies have been developed that can remove the selectable marker from the mature transgenic plant.

Regeneration

As often only a single cell is transformed with genetic material the organism must be regrown from that single cell. As bacteria consist of a single cell and reproduce clonally regeneration is not necessary. In plants this is accomplished through the use of tissue culture. Each plant species has different requirements for successful regeneration through tissue culture. If successful an adult plant is produced that contains the transgene in every cell. In animals it is necessary to ensure that the inserted DNA is present in the embryonic stem cells. When the offspring is produced they can be screened for the presence of the gene. All offspring from the first generation will be heterozygous for the inserted gene and must be mated together to produce a homozygous animal.

Confirmation

Further tests using PCR, Southern Blots and Bioassays are needed to confirm that the gene is expressed and functions correctly. The organism's offspring are also tested to ensure that the trait can be inherited and that it follows a Mendelian inheritance pattern.

Applications

Genetic engineering has applications in medicine, research, industry and agriculture and can be used on a wide range of plants, animals and micro organism.

Medicine

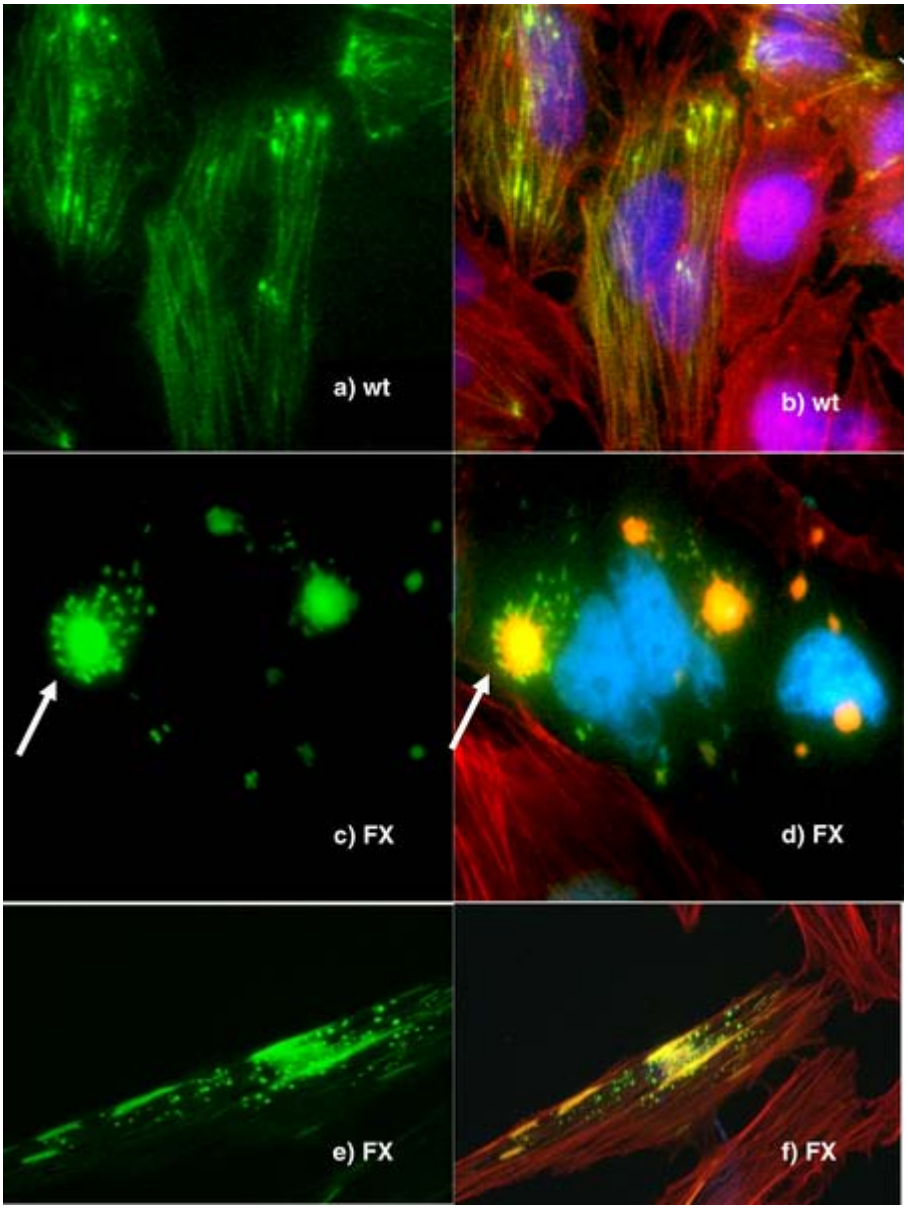
In medicine genetic engineering has been used to mass-produce insulin, human growth hormones, follistim (for treating infertility), human albumin, monoclonal antibodies, antihemophilic factors, vaccines and many other drugs. Vaccination generally involves injecting weak live, killed or inactivated forms of viruses or their toxins into the person being immunized. Genetically engineered viruses are being developed that can still confer immunity, but lack the infectious sequences. Mouse hybridomas, cells fused together to create monoclonal antibodies, have been humanised through genetic engineering to create human monoclonal antibodies.

Genetic engineering is used to create animal models of human diseases. Genetically modified mice are the most common genetically engineered animal model. They have been used to study and model cancer (the oncomouse), obesity, heart disease, diabetes, arthritis, substance abuse, anxiety, aging and Parkinson disease. Potential cures can be tested against these mouse models. Also genetically modified pigs have been bred with the aim of increasing the success of pig to human organ transplantation.

Gene therapy is the genetic engineering of humans by replacing defective human genes with functional copies. This can occur in somatic tissue or germline tissue. If the gene is inserted into the germline tissue it can be passed down to that person's descendants. Gene therapy has been used to treat patients suffering from immune deficiencies (notably Severe combined immunodeficiency) and trials have been carried out on other genetic disorders. The success of gene therapy so far has been limited and a patient (Jesse Gelsinger) has died during a clinical trial testing a new treatment. There are also ethical concerns should the technology be used not just for treatment, but for enhancement, modification or alteration of a human beings' appearance, adaptability, intelligence, character or behavior. The distinction between cure and enhancement can also be difficult to establish. Transhumanists consider the enhancement of humans desirable.

Research

Knockout mice



Human cells in which some proteins are fused with green fluorescent protein to allow them to be visualised

Genetic engineering is an important tool for natural scientists. Genes and other genetic information from a wide range of organisms are transformed into bacteria for storage and modification, creating genetically modified bacteria in the process. Bacteria are cheap, easy to grow, clonal, multiply quickly, relatively easy to transform and can be stored at -80°C almost indefinitely. Once a gene is isolated it can be stored inside the bacteria providing an unlimited supply for research.

Organisms are genetically engineered to discover the functions of certain genes. This could be the effect on the phenotype of the organism, where the gene is expressed or

what other genes it interacts with. These experiments generally involve loss of function, gain of function, tracking and expression.

- **Loss of function experiments**, such as in a gene knockout experiment, in which an organism is engineered to lack the activity of one or more genes. A knockout experiment involves the creation and manipulation of a DNA construct *in vitro*, which, in a simple knockout, consists of a copy of the desired gene, which has been altered such that it is non-functional. Embryonic stem cells incorporate the altered gene, which replaces the already present functional copy. These stem cells are injected into blastocysts, which are implanted into surrogate mothers. This allows the experimenter to analyze the defects caused by this mutation and thereby determine the role of particular genes. It is used especially frequently in developmental biology. Another method, useful in organisms such as Drosophila (fruit fly), is to induce mutations in a large population and then screen the progeny for the desired mutation. A similar process can be used in both plants and prokaryotes.
- **Gain of function experiments**, the logical counterpart of knockouts. These are sometimes performed in conjunction with knockout experiments to more finely establish the function of the desired gene. The process is much the same as that in knockout engineering, except that the construct is designed to increase the function of the gene, usually by providing extra copies of the gene or inducing synthesis of the protein more frequently.
- **Tracking experiments**, which seek to gain information about the localization and interaction of the desired protein. One way to do this is to replace the wild-type gene with a 'fusion' gene, which is a juxtaposition of the wild-type gene with a reporting element such as green fluorescent protein (GFP) that will allow easy visualization of the products of the genetic modification. While this is a useful technique, the manipulation can destroy the function of the gene, creating secondary effects and possibly calling into question the results of the experiment. More sophisticated techniques are now in development that can track protein products without mitigating their function, such as the addition of small sequences that will serve as binding motifs to monoclonal antibodies.
- **Expression studies** aim to discover where and when specific proteins are produced. In these experiments, the DNA sequence before the DNA that codes for a protein, known as a gene's promoter, is reintroduced into an organism with the protein coding region replaced by a reporter gene such as GFP or an enzyme that catalyzes the production of a dye. Thus the time and place where a particular protein is produced can be observed. Expression studies can be taken a step further by altering the promoter to find which pieces are crucial for the proper expression of the gene and are actually bound by transcription factor proteins; this process is known as promoter bashing.

Industrial

By engineering genes into bacterial plasmids it is possible to create a biological factory that can produce proteins and enzymes. Some genes do not work well in bacteria, so

yeast, a eukaryote, can also be used. Bacteria and yeast factories have been used to produce medicines such as insulin, human growth hormone, and vaccines, supplements such as tryptophan, aid in the production of food (chymosin in cheese making) and fuels. Other applications involving genetically engineered bacteria being investigated involve making the bacteria perform tasks outside their natural cycle, such as cleaning up oil spills, carbon and other toxic waste.

Agriculture

Bt-toxins present in peanut leaves (bottom image) protect it from extensive damage caused by European corn borer larvae (top image).

One of the best-known and controversial applications of genetic engineering is the creation of genetically modified food. There are three generations of genetically modified

crops. First generation crops have been commercialized and most provide protection from insects and/or resistance to herbicides. There are also fungal and virus resistant crops developed or in development. They have been developed to make the insect and weed management of crops easier and can indirectly increase crop yield.

The second generation of genetically modified crops being developed aim to directly improve yield by improving salt, cold or drought tolerance and to increase the nutritional value of the crops. The third generation consists of pharmaceutical crops, crops that contain edible vaccines and other drugs. Some agriculturally important animals have been genetically modified with growth hormones to increase their size while others have been engineered to express drugs and other proteins in their milk.

The genetic engineering of agricultural crops can increase the growth rates and resistance to different diseases caused by pathogens and parasites. This is beneficial as it can greatly increase the production of food sources with the usage of fewer resources that would be required to host the world's growing populations. These modified crops would also reduce the usage of chemicals, such as fertilizers and pesticides, and therefore decrease the severity and frequency of the damages produced by these chemical pollution.

Ethical and safety concerns have been raised around the use of genetically modified food. A major safety concern relates to the human health implications of eating genetically modified food, in particular whether toxic or allergic reactions could occur. Gene flow into related non-transgenic crops, off target effects on beneficial organisms and the impact on biodiversity are important environmental issues. Ethical concerns involve religious issues, corporate control of the food supply, intellectual property rights and the level of labeling needed on genetically modified products.

Other uses

In materials science, a genetically modified virus has been used to construct a more environmentally friendly lithium-ion battery. Some bacteria have been genetically engineered to create black and white photographs while others have potential to be used as sensors by expressing a fluorescent protein under certain environmental conditions. Genetic engineering is also being used to create BioArt and novelty items such as blue roses, and glowing fish.