Evolution: the best possible search algorithm?

"Molecular evolution has a practically limitless field for exploration and experiment, enabling it to elaborate the huge network of cybernetic interconnections, … whose performances appear to transcend, if not escape, the laws of chemistry."

Jacques Monod[1]

Abstract

What is it that makes natural evolution so successful at finding solutions to real-world problems? If this question can be understood and answered we will be in possession of a powerful new tool for overcoming many of the difficulties now being faced by the more mathematical approaches to communication problems. Evolution will not tolerate Lamarckian control. Indeed unfettered diversity is the key to the success of Darwinian evolution and nature has devised a remarkable molecular mechanism to guarantee that change is protected from any form of 'intelligent' guidance. The origin of life can be traced along a path of increasing diversity arriving at today's creatures with their 'junk' DNA. We see the fruits of evolutionary search in our own immunity to disease and in the developing field of combinatorial chemistry. The implications for future computer search technology are discussed.

Lamarckian and Darwinian Theories

Jean Lamarck first postulated the idea of the inheritance of acquired characteristics. Acquired characteristics are those benefits which parents acquire during their efforts to cope with their environment. According to Lamarck giraffes have long necks because generations of giraffes had been stretching for food and each parent had passed on the advantages of a stretched neck to the fortunate offspring through the mechanisms of heredity. Sons of blacksmiths are very likely to grow up with galls on their hands because their fathers and grandfathers spent their entire lives building up such protection on their palms. Evolution proceeds in a continuous manner accumulating the experience of successive generations and giving a certain purposefulness to life itself. The idea appeals to common sense.

Darwin presented his theory of natural selection in 1858 to the Linnean Society and in his publication of the Origin of Species in 1859. Darwin held that creatures evolve through a process of random mutation followed by natural selection. Each generation of offspring is subjected to changes and the pressures of the environment select against individuals. Those mutants, which are able to survive, will live to reproduce and pass their advantage down the generations. Their weaker cousins will be unable to cope, perhaps will be unable to find food, or lack the immunity to some disease and die.

It is fundamental to Darwinian evolution that mutations are not influenced by the needs of the individual; mutations are utterly impervious to any 'hints' from the outside world. The doctrine of the "continuity and inalterability of the germ-tract" was propounded by the German zoologist, August Weismann, in 1885 and still holds sway today[2]. The genetic data, which is passed to offspring, is held sacrosanct in the germ cells (the ova and sperm cells in animals) which are kept separate from the somatic cells, the body. Somatic cells do generate replacements and can be affected

by the experiences of the owner but they never stand any chance of promoting their own lineage through the generations, that is, until cloning becomes commonplace. Our skin cells, blood cells and thankfully cancer cells all have no influence on our children. Lamarckism on the other hand would expect a mechanism, which *does* affect the germ-tract accumulating parental alterations with each generation.

Today Lamarckism has been outlawed and Darwinian evolutionary principles are accepted by most scientists, although there is still considerable debate and the spectre of Lamarck is again raised from time to time[3].

It is quite surprising that Darwin produced his theory without having a clear idea of how heredity works. He could not see how offspring maintained their differences if they merely received the "average" of their parents' features. In time, the members of each species should all end up much the same and all diversity would be eliminated which was clearly contrary to all experience. This was a major challenge to his thinking and could have turned out to be a fatal flaw, but he was not discouraged and fortunately an Augustinian monk called Gregor Mendel produced the required answer. His experiments with garden peas revealed that there existed physical structures within cells that were responsible for two forms of each plant characteristic he studied. These structures we now call genes and the two forms are called alleles. He deduced that there must be two alleles within the cell, one dominant and the other recessive. Mendel was able to deduce fundamental laws of genetics, laws that we still recognise today as encapsulating the underlying principles of inheritance.

Nature abhors the constraints of Lamarckian guidance and unfettered diversity is key to the success of Darwinian evolution. Nature herself has endorsed Darwin's ideas by devising a remarkable molecular mechanism to guarantee that change is protected from any form of 'intelligent' guidance.

The Molecular Mechanism

In 1953 Watson and Crick discovered the famous double helix structure of DNA (Deoxyribonucleic Acid). They showed how this molecule held the genetic information, which was passed from parents to children and determined how living things developed from an embryo into an adult. Watson and Crick shared the 1962 Nobel Prize for Physiology and Medicine with Maurice Wilkins for their pioneering work, but sadly, Rosalind Franklin, whose X-ray crystallographic data greatly contributed to the key discovery, died before this date, and was not honoured with the prize.

Watson and Crick revealed a beautiful structure and a mechanism, which provided a deeper understanding of the work of Darwin and Mendel. It became obvious that DNA behaved like the characteristics that Mendel described and that hereditary factors were actually located within the DNA molecules. The first trait to be identified with DNA material or the chromosomes was sex. Observation under the microscope discovered the presence of the X and Y-chromosomes, a Y chromosome indicating masculinity.

The chemistry of life has since been explored in very great detail and is now culminating in a world-wide co-ordinated programme to enunciate the entire nucleotide sequence of the human genome by the year 2005[4] or sooner. An important part of the Human Genome Project is to identify and locate genes using a process

known as transcript mapping. Human DNA contains as many as 100,000 genes[5] in 23 pairs of chromosomes, so the task is a pretty daunting one.

DNA is the repository of all the information needed to build every component in a new living organism; it defines the way in which that organism functions, reproduces and indeed survives in whatever environment it finds itself. A full copy of the DNA is found in most cells in the human body (not blood cells or germ cells, for example) and so in theory every cell contains the information to clone a copy of the body of which it is already a part. DNA is a long polymeric molecule consisting of long chains of just four different individual units or nucleotides, called Adenine, Cytosine, Guanine and Thymine, A, C, G and T, for short. When we say DNA is long we mean it! There could be as many as 3,000,000,000 nucleotides chemically attached to each other in a human cell and if all the DNA in our cells was unravelled and strung out it would reach to the sun[6]! There are no chemical restrictions on the order in which the nucleotides can join together so at any point in the chain the nucleotide can be an A, C, G or T. This means that the number of possible sequences of length 3,000,000,000 is the unbelievably huge number $4^{3,000,000,000}$. This enormous variability enables the DNA to represent a practically infinite number of states that may or may not correspond to a viable life form.

Molecules of DNA are relatively stable and not surprisingly achieve lifetimes comparable with that of the organism within which they reside. Much of the stability is derived from a double helix structure in which one strand of DNA unites with a complementary strand in a well-defined way. Guanine always links with Cytosine and Adenine is always opposite Thymine; no other pairings are possible. The DNA double helix therefore only consists of a sequence of the two base pairs A-T and C-G).

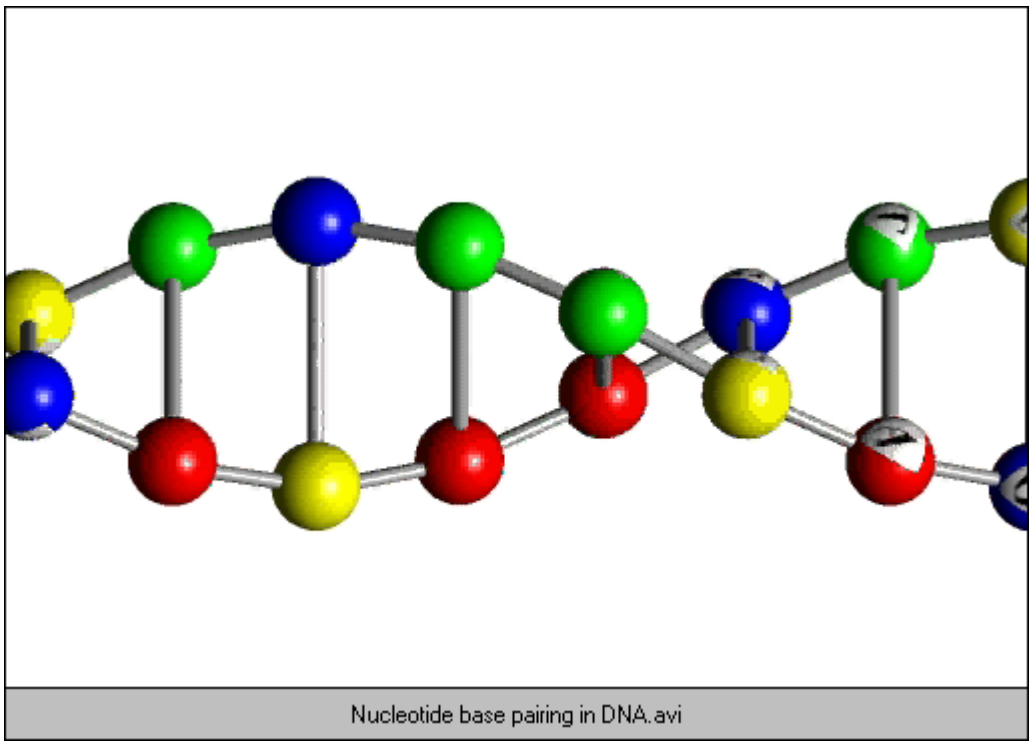

Figure 1 Nucleotide base pairing in DNA

DNA replication takes place through the action of a host of enzymes. First some of the turns[7] in the double helix are unwound, then some of the base pairs are broken exposing complementary single strands. Daughter strands are then built up using both exposed strands as templates until the entire molecule is duplicated. Although DNA replication is virtually error free (perhaps one change in10^9 base pairs[8]) occasionally mistakes do occur and a random mismatching base pair will be created (eg G-T instead of A-T). When this molecule is replicated one of the daughters will have an incorrect nucleotide base pair (G-C in this example) which might lead to a change in the functionality of the cell and constitutes a mutation. Mutation rates can be raised by the presence of chemicals or physical agents such as ultra violet radiation which interacts directly with the DNA.

A gene is simply a segment of a DNA molecule which may contain as many as 2,000,000 base pairs or as few as 75. The information contained in the sequence in effect defines a series of operations that leads to the synthesis of an enzyme or protein that expresses the gene's function.

In the first stage of gene expression one of the DNA strands unwinds from the other and behaves as a template in the production of messenger ribonucleic acid (mRNA). The base pairings are the same as in DNA except that Uracil replaces Thymine. The resulting mRNA is complementary to one of the strands of DNA which it copies and is identical to the other (with U substituted for T) (Figure 2). Again there are no chemical restrictions on the sequence of A, C, G and U units within a mRNA molecule and sideways links between the mRNA units are identical in nature. In addition the RNA molecule usually exists as a single strand unlike DNA that wraps around a second strand to form a double helix.

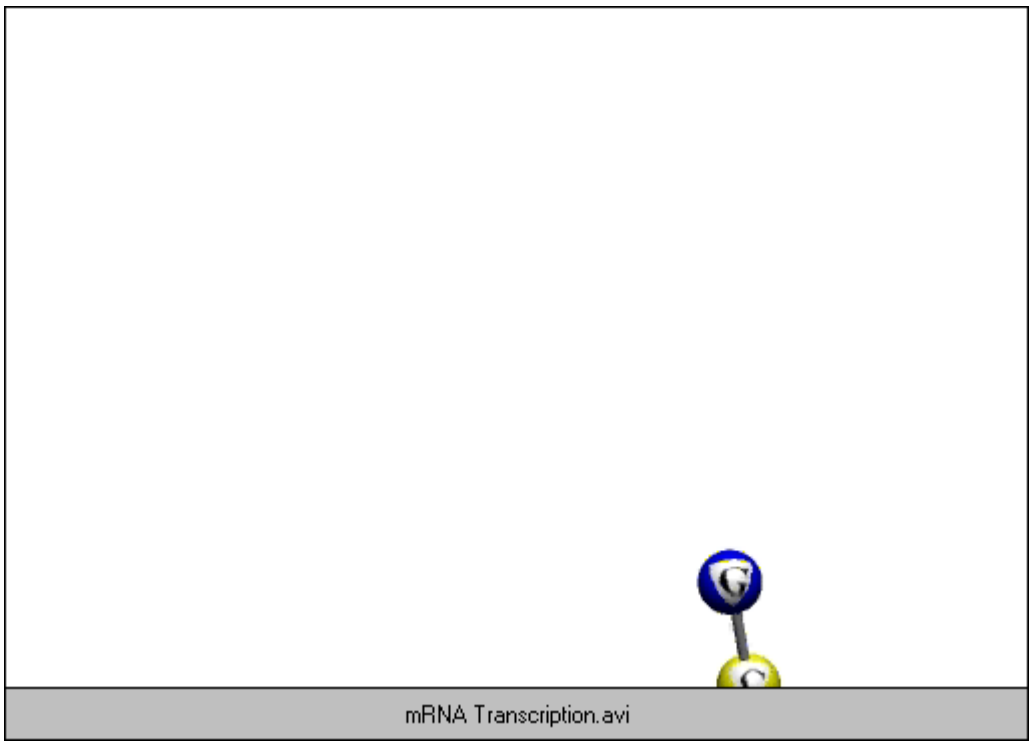

Figure 2 mRNA transcription

Frequently the transcribed RNA molecule is the end product of gene expression[9] but mRNA moves on to a second translation stage and the synthesis of specific proteins.

An intriguing coding scheme, which was only finally elucidated in 1966, is now brought into operation. It turns out that triplets of base units in mRNA each correspond to one of 20 amino acids and the sequence of triplets or codons determines how the amino acids are chained together at the end of the translation process. There are 4 different base units and therefore there are 4x4x4=64 different codons which represent the 20 amino acids[10].

The code is degenerate in the sense that all amino acids except methionine and tryptophan have more than one codon, and three codons do not code for an amino acid at all but instead act as punctuation marks in the translation process[11].

How are the proteins deciphered from the mRNA using this coding scheme? Enter the Ribosome. Ribosomes are large molecules which are the factories for protein synthesis. They attach themselves to the mRNA and travel along its length decoding and synthesising a chain of proteins (polypeptide) as they go (Figure 3). Molecules called transfer RNA's (tRNA) bring each of the 20 proteins to the ribosome and snap them into place if the tRNAs fit the codons on the mRNA. A series of tRNAs arrive and unload their amino acids in the right order to build the polypeptide. The translation process stops when the ribosome hits a punctuation mark or 'stop codon'.

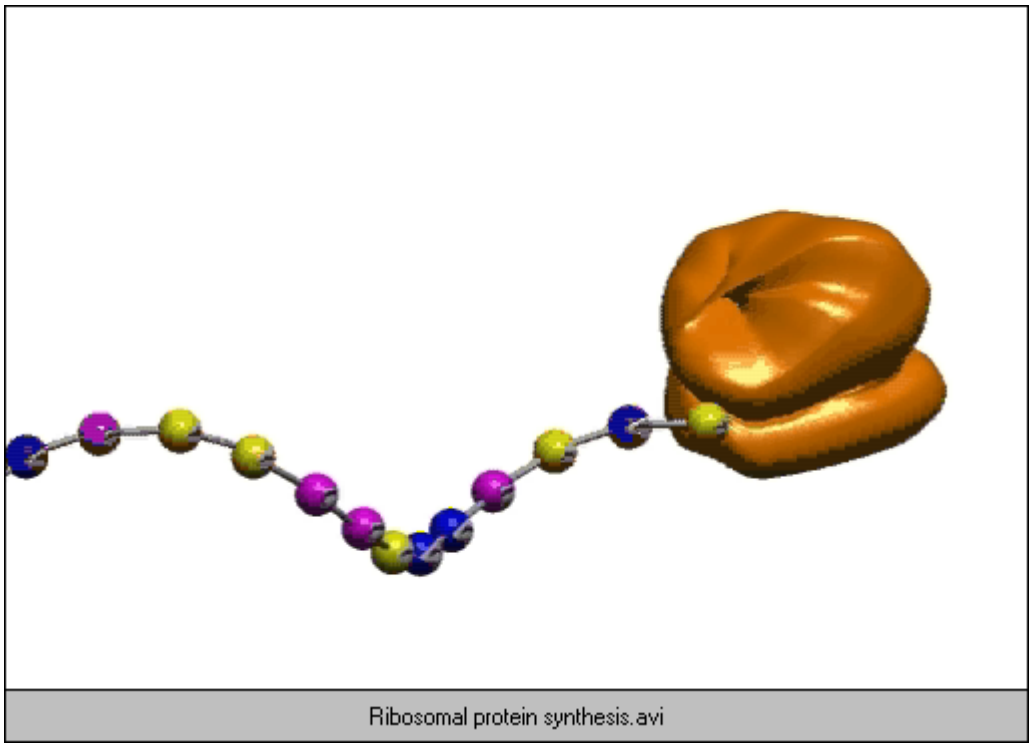

Figure 3 Ribosomal protein synthesis

tRNA is a fascinating object. On one end of the molecule is an anticodon that complements the codon triplet on the mRNA which codes for the amino acid it carries. This is the bit which must match the mRNA triplet if the ribosome is to accept the amino acid into the synthesised polypeptide. On the other end is the acceptor site to which the amino acid is attached. One might imagine that empty tRNAs go out into the cell and attach themselves to one of the 20 proteins which

match some distinctive part of their structure designed to correspond to the meaning of their anticodons. Surprisingly this is not the case. In fact the acceptor sites for all tRNAs are *identical* and there is no special template into which an amino acid will fit. The tRNA cannot do the job of fetching an amino acid by itself.

The charging of tRNA with a protein is catalysed by yet another molecule having the name of aminoacyl-tRNA synthetase. There is a different aminoacyl-tRNA synthetase for each tRNA and is able to recognise both the correct amino acid and the appropriate tRNA (Figure 4). How this is done is not yet understood in detail and there may be even more steps in what appears already to be a complex process. It is at first sight astonishing that nature has gone to the trouble of creating a long and protracted process for producing proteins when there could be much simpler alternatives that require less energy and much less cellular infrastructure.

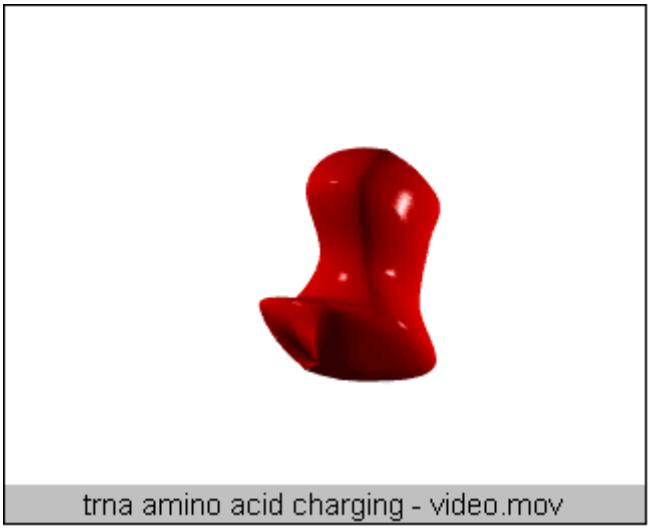

Figure 4 tRNA amino acid charging – *aminoacyl-tRNA synthase (grey) catalyses the attachment of a protein (red) to a tRNA (green)*

The reader should appreciate that much detail has been omitted from the molecular events described above[12]; the initiation of transcription and translation and the detailed workings of the ribosome are just two examples, but these omissions should not affect the discussion.

Diversity vs Constraint

How has the genetic code come about? Why has evolution produced a complex multi-stage synthesis process to generate the proteins of life? Surely something simpler and more direct would have greater survival value; indeed it would make sense if all the products of protein synthesis were derived purely by transcription from DNA without all the bother and energy for translation. But transcription alone presents barriers to the options available for the progress of evolution.

RNA is chemically very closely related to the DNA template from which it is derived. This means that any agent which might cause an evolutionary change at a particular point in the hereditary material will also have to function in the presence of other

proteins which are very likely to involve themselves in the same reaction which gives rise to the change. Such changes are perfectly possible, but they will be constrained because of chemical restrictions imposed by related proteins that surround the reaction. This will always be the case in any cascade or cycle of reactions in which all the intermediate products are present in varying proportions; the sequence of chemical events can be controlled to a certain extent by external influences, but the cycle cannot literally take any course simply because of the constraints imposed by chemistry.

So translation comes to the rescue. Translation enables the chemicals representing the store of hereditary information to be divorced from the products it produces via the genetic code. It prevents the majority of proteins and enzymes from having any significant chemical influence on what evolutionary changes are available and frees the process of evolution to trial a much larger universe of mutants. The triplets of amino acids in mRNA only have a connection with the proteins they represent through the tRNA's and the coding scheme that they implement. It is remarkable that the obstructive effects of chemistry are diminished still further by the tRNA's using identical acceptor sites for all 20 proteins which they carry. Amazingly there is no distinctive chemical link between each tRNA and any of the basic protein building blocks and there is therefore no connection between processes of change to DNA and the products which emerge from translation. A common sense template strategy in which tRNA's simply match an mRNA triplet on one part of their structure and a corresponding protein on another would suffer from the very chemical associations defined by the templating mechanism which would limit the evolutionary options available to DNA. However, it could be argued that the catalysing enzyme, aminoacyl-tRNA synthetase, actually does define a weak similarity between the protein and the corresponding tRNA causing some restrictive influence, and it might be interesting to speculate the participation of a cascade of further enzymes which facilitate the creation of the amino acid-aminoacyl-tRNA synthetase complex and which isolate the protein products from their origin still further.

We can conclude that the fundamental processes of transcription and translation have evolved to isolate molecular evolution itself from the constraints of basic chemistry and for it to enjoy as much freedom and diversity as it is possible to give.

Life versus Chemistry

It goes without saying that for evolution to take place, organisms must possess a structure that can accept and implement change. Organisms that are able to evolve quicker than others have a tremendous selective advantage over those who are sluggish to adapt and who perhaps have reached an evolutionary plateau. New species can be spawned into a changing environment effectively wiping out potential competitors who are unable to respond. This means that *evolvability* itself must be evolving in a Darwinian sense as new mechanisms have been introduced to extend the universe of possible organisms and increase the options available for change.

Let us look back at the origin of life for a moment and conjecture how evolution might have progressed through a series of stages which steadily removes the restrictions imposed by the environment to allow greater and greater numbers of alternatives to be trialed by natural selection. Stanley Miller in his classic experiment in the early 1950's showed that amino acids are formed by electric discharges in an

atmosphere of water, nitrogen, ammonia, hydrogen cyanide and methane[13]. He demonstrated that it was perfectly possible for the protein building blocks of life to be created in just the sort of conditions which might be expected on primitive earth[14]. But how could these simple proteins come together to produce the complex molecular structures we know today?

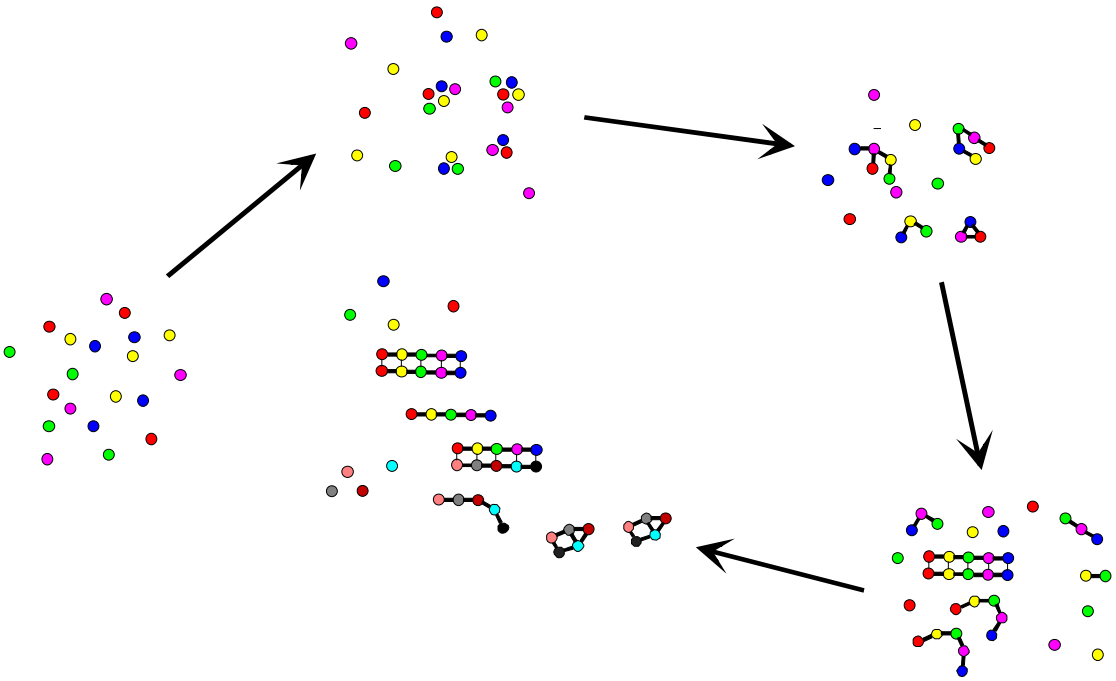

Figure 5 Stages of increasing diversity

Evolution probably began with localised concentrations of various mixtures of simple proteins. Those concentrations that promoted template replication would have a selective advantage over others that relied upon chance alone. The first obvious restriction to be faced by the evolving prebiotic soup was that of the simple geographical separation of reacting components. Molecular components which attached themselves to each other whilst maintaining template replication would speed evolution by reducing the effects of separation. One might expect relatively short chains of nucleotides, the precursors of RNA, to appear at around this time. Several varieties of proto-RNA could have appeared, but all would have been constrained from further evolution by their own three dimensional structure which in most cases would have prevented template replication from taking place. It would have been the RNA that was templated from a separate DNA precursor which overcame this the second barrier to evolution and evolvability. Much larger folded RNA molecules could now appear with complex functionality dependent upon their three-dimensional shape without the restrictive requirement of self-replication. Meanwhile the proto-DNA retained the ability to replicate itself and increase the related populations of proto-RNA's.

At this point we would imagine that the prebiotic soup consisted of a multitude of proteins interacting and following a variety of chemical pathways. Evolutionary changes would have a lasting effect if they occurred in the proto-DNA and were translated and transcribed into proteins.

However, as has been argued earlier, such changes would have to take place in the presence of a host of related templated products which will tend to encourage some changes and block many others. This constraint on variation must have been a tough one to overcome. Protein products needed to be totally unrelated to the chemistry of the process that generated them to increase diversity. No series of simple templates would achieve that, although countless such schemes must have been trialed. The breakthrough might have taken place as soon as enzymes appeared which carried proteins to an RNA template without requiring the protein itself to fit the template - the enzymes, or proto-tRNA, did the matching. This divorced the chemistry of the transported proteins from the synthesising process thereby boosting the potential for diversity and hence the speed of evolution. The proto-tRNAs in fact encapsulated the first primitive genetic code and gave that particular part of the soup another selective advantage.

Clearly the speed of evolution must still have been severely affected by geographical separation and those groups of molecules which could manage to stay close and maintain a continuous interaction would benefit. Enter the cellular structure in which some form of boundary prevented essential components from drifting off and allowed the mechanisms for improving the accuracy of replication and translation to be refined without interference from other contaminating products. The success of the cellular strategy has led to the rich variety of multicellular life on earth today and its ability to adapt to the extremes of almost any environment.

Sex and Junk DNA

Physical separation still imposes a restriction on genetic variety even today with our finely tuned cellular mechanisms, this time within the cell. Mutations certainly introduce variation into the genetic material, but there is another important mechanism that rearranges the sequences in DNA and yields further diversity. Sexual reproduction has accelerated evolution by creating new arrangements of genes which would have been highly unlikely if reliance was placed solely on a series of low level mutations. It gives a species access to the diversity of the entire population by allowing the genes to mix in all manner of combinations; it allows a highly beneficial mutation in a gene in one individual to pass in combination from one generation to the next and to spread throughout the population.

Most of our cells contain 23 pairs of chromosomes, with one of each pair inherited from each parent. The sex cells (sperm and ova) are produced through a multi-stage process called *meiosis* in which a full 46-chromosome cell divides twice to generate four sex cells with just a half complement of 23 single chromosomes. Before the first cell division, genetic material may be exchanged between the chromosomes in each pair in a process called recombination or “crossing-over”(Figure 6). This results in the merging and mixing of parental genomes which are passed on to the next generation through the sex cells. It allows the genes from each parent to be shuffled in new ways so that successive generations of individuals will each hold unique combinations of genes and hence unique combinations of advantages and disadvantages to be tested by natural selection[15].

However, the shuffling of genes does not necessarily take place as freely as Mendel’s original experiments indicated. Mendel was ‘lucky’ that the occurrence of the characteristics of the pea plants he studied were independent of each other. An

experiment conducted by William Bateson and others in 1905 on different characteristics of the sweet pea showed that Mendel's results do not always hold. The ratios of numbers of plant types were at variance with those predicted by Mendel. The explanation for these effects lies in the location of the genes on the chromosomes that give rise to the traits. If the genes are located on *different* chromosomes then they segregate independently; if they are located on the *same* chromosome, they are partly linked and will not disassociate as often during crossing over. In fact the closer a pair of genes resides on a chromosome, the less likely they are to be separated into different sex cells. As the frequency of occurrence of pairs of linked traits is a direct measure of their separation on the chromosome, it is used as a technique for mapping genes in organisms where the generation times are not too great.

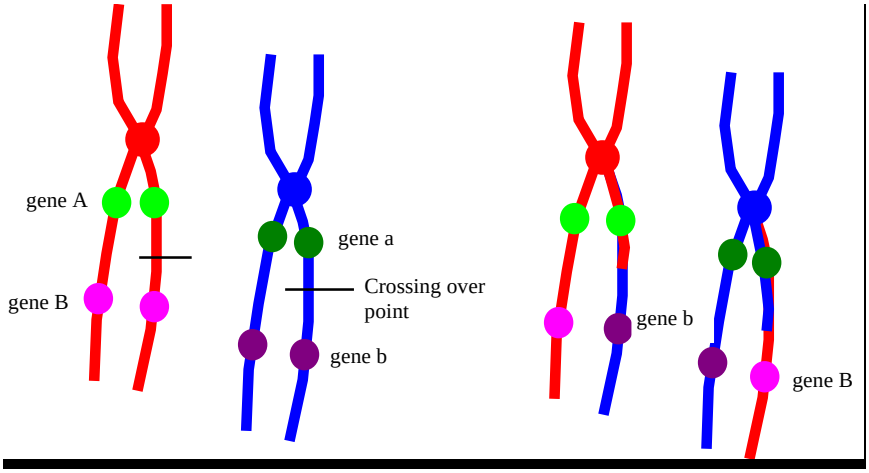

Figure 6 Crossing over

Rather surprisingly 90% of the DNA in the genome of most organisms does not code for any protein[16]. Much of it consists of highly repetitive sequences of nucleotides either dispersed or clustered in the genome; the patterns of repetition vary from person to person and are the data of genetic fingerprints[17]. Duplicating lots of redundant DNA takes time so more primitive bacterial organisms that rely on fast cell division would be heavily penalised for having redundant DNA.

Why did so much useless DNA creep into our cells in the first place? Linkage is a restriction on evolutionary diversity. In whatever way genes are arranged on the chromosomes, some have to be close to each other and are very unlikely to be uncoupled and recombined in new ways in later generations. This limitation would be mitigated to some extent by spreading the genes out on the chromosomes and incorporating plenty of unused real estate between them. Crossing over would then stand a chance of swapping more genes and presenting nature with more alternatives to select against[18]. Although it does not have an obvious purpose during the lifetime of the individual, spare DNA is there surely to oil the wheels of evolution.

It allows the process of crossing over to be much more thorough and generates many more combinations during considerably shorter time periods[19]. Such a mechanism confers a great advantage to an organism which is able to exchange genes rapidly throughout a population and adapt to changing environments within very few generations.

Immunity by Evolution

Backboned animals possess a staggeringly effective system for telling the difference between self and nonself. It can respond to the invasion of bacterial and viral organisms that would otherwise overwhelm the individual and yet leave friendly cells unharmed. Specialised cells called lymphocytes produce molecules of antibodies that fit specific regions of the foreign material, or antigen, and render it harmless. The system possesses memory for the antibodies and becomes more effective following a second attack by the same intruder. It is extraordinary that a targeted response takes place even for foreign molecules which have been synthesised for the very first time and have never existed before in history[20].

It is natural to believe that the immune response is based on a simple templating mechanism in which the invading molecule transfers a complement of its shape and structure to a deformable antibody molecule. The antibody then duplicates itself and binds to all foreign molecules having the same properties. How else can the structure of the antigen be communicated to the immune system for such a precise and specific response? This common sense explanation turns out to be false and has been replaced by a theory which at first sight is complex and unnecessary.

It is now understood that the body is able to manufacture a huge repertoire of antibody molecules, all having different binding sites. When a bacterium enters the body it faces a population of antibodies a few of which may bind approximately to its surface. This stimulates the corresponding mother lymphocytes to divide repeatedly and produce more and more of the same sort of antibody which soon kills the invader and its companions. So through a process of clonal selection the population of lymphocytes changes in composition to fabricate magic bullets which are lethal only to the aggressor. After the battle is won some lymphocytes that played a part in the process of antibody production remain in the body and will produce an accelerated response if ever the troublesome antigen returned.

The lymphocytes can generate huge universes of antibodies by virtue of a mechanism which randomises the genetic code specifying certain variable regions of the polypeptide chain making up the antibody. This in effect means that we are observing Darwinian evolution in miniature in which the diverse population of antibodies is selected against a criterion that measures success in terms of binding to antigens. Why does evolution itself choose evolution to search for solutions? Why does nature not use the common sense template approach based on a series of chemical events derived from the antigen? The reason is again because evolutionary search is more effective, and it is more effective because it is not restricted by any specific set of chemical rules for generating antibodies that require information about the shape of the antigen. An organism with a non-evolutionary immune mechanism would always be in danger from an attack from an antigen that could not be ‘read’ by the specific immune system in question. It would never be able to anticipate the effects of all antigens and would succumb in a world full of hostility and trickery.

Implications for Technology

What clues to our thinking does nature give us? We see nature struggling to break the shackles of the environment and especially the internal limitations of the body. The baggage of earlier evolution is a real impediment to progress and avenues of viable evolutionary change are restricted by the cellular process in which they take place. Evidence shows that mechanisms that lessen this burden and increase innovation are favoured. Wherever we look we see examples of nature seizing opportunities to gain more freedom for change. Darwinian natural selection wins out over Lamarck's acquired characteristics; the sequences of A, C, G, T bases in DNA are not influenced by the chemical bonds between the bases themselves; transcription allows proteins of all shapes and sizes to be synthesised which are not bound by the need to self reproduce; translation frees protein synthesis almost completely from the immutable rules of chemistry, and junk DNA permits genes to be mixed much more efficiently. Even evolution itself judges evolution to be the best search process for an immune response. Nature has been tremendously successful in producing life forms to fit every environmental niche on earth and possesses the evolutionary mechanism to do it. Life today is just a single moment in an unending process of change, change not only to individuals and species, but also change to the evolutionary process itself. Unfortunately we cannot expect to speed up evolution simply by a detailed emulation of one of nature's organisms because it will be subject to the very same limitations suffered by the real-life counterpart. It will duplicate all the complexities of molecular biology set down to free itself from the laws of chemistry and other constraints, it will hit the same evolutionary plateaux and possess the same evolutionary blind spots. Surely a search for higher performance using computers should press for the removal of as many constraints as possible and certainly leave behind all those suffered by biological evolution.

Combinatorial Chemistry

It is extremely difficult to predict the properties of chemicals without actually trying them out in experiment. If we cannot even say in advance how a protein chain will fold upon itself, we can hardly be confident about its biological activity. This presents a serious problem for the pharmacological companies who are in the business of discovering new wonder drugs. A medicinal chemist using traditional synthesis methods might synthesise only one compound in a week and then set about testing it against a target disease, clearly a slow and expensive process which might take years to find anything useful. In fact such a process demands a great deal of theory simply to justify the choice of lead compound structures for lengthy and costly investigation. And this is where the problem lies. How many powerful drugs have been proposed and rejected in the past because the theory at the time dictated that they would 'probably' be ineffective?

Combinatorial chemistry is an approach to chemical synthesis which represents a shift from a reliance upon theory towards trial and error and a dramatic improvement in the success rate for the discovery of new drugs. Dozens of pharmaceutical companies devoted to combinatorial chemistry are now in operation. We are seeing the application of Darwinian evolution to the selection of highly performing compounds from amongst an artificially created universe of diverse substances, rather than a directed Lamarckian process wholly founded upon past experience. The modern chemist is in fact imitating the body's immune system which screens huge numbers of

randomly generated antibodies and identifies those that work best against an invading bacteria.

A method invented by Árpád Furka for generating a combinatorial library of chemicals is known as the pool-and-split or split-and-mix synthesis. First one from a set A of perhaps 50 different chemical building block molecules is added to a solution containing inert, microscopic beads made of polystyrene. Some of the molecules become attached to the beads and any unreacted material is washed away leaving behind molecules that are only attached to beads.

This is again repeated in separate containers for the other 49 compounds. The contents of the 50 containers are then pooled and split out into 50 containers again, but this time each container contains *all* of the 50 sorts of molecule in set A. Now 50 compounds from another set B are each added separately to the 50 vessels containing A type molecules thereby providing all possible 2500 combinations. The same pool-and-split process can be repeated for set C and then D yielding 125,000 and 6,250,000 potentially different reactions.

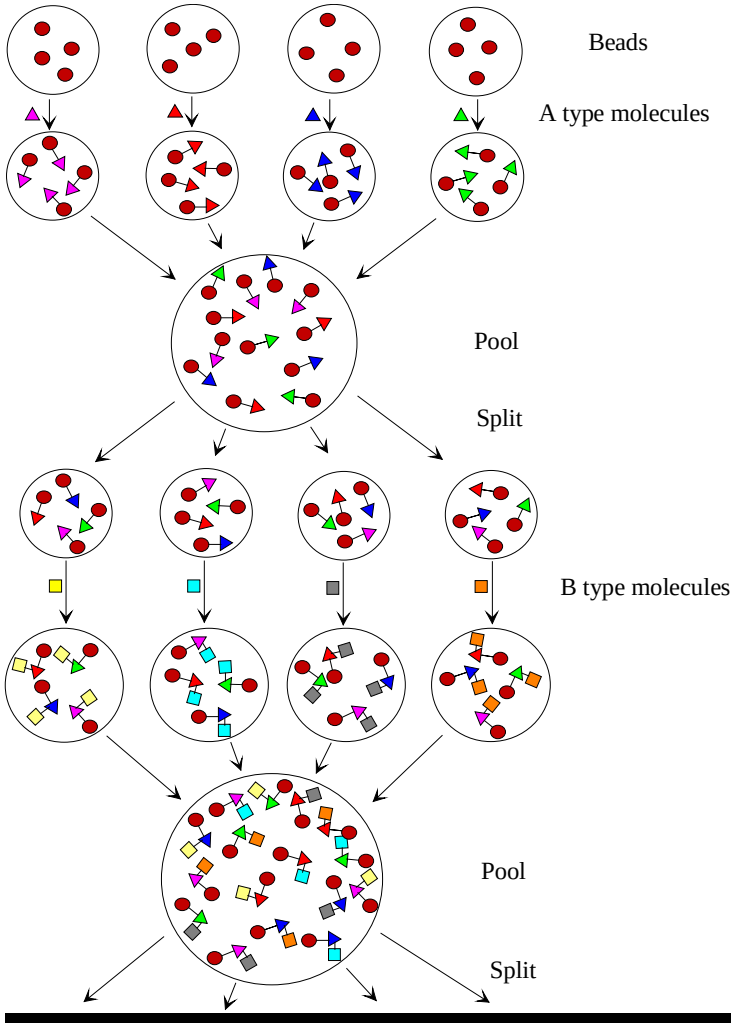

Figure 7 Two stages of pool and split synthesis

The problem now remains of how to pick out the useful products. One method is to screen the final mixtures of compounds to determine the average activity of each batch and to identify the most promising building block in set D. The best performing D molecule can the be added to all of the 50 combinations of C with AB mixtures to identify the best performing C molecule, and so on to get the best combination from all four sets of molecules. This technique does not always work satisfactorily, and methods are now being devised for labelling the beads to indicate the order in which building blocks have been added to the structure and hence the identity of the compound on that bead[21].

This approach to combinatorial chemistry is reminiscent of the prebiotic soup in which multitudes of assorted amino acid building blocks mingled together eventually producing concentrations of compounds which were judged successful simply on the basis of their ability to produce those concentrations. Today the chemists are mixing their own building blocks together and are taking success to mean how well a drug hits a target and yields medicinal value. There is also a striking parallel between nature's use of the genetic code sequences to produce life's proteins and the chemist's use of molecular tags to label beads holding particular compounds. The bead and its attachments are the chemist's equivalent of charged tRNA.

Genetic Algorithms and Crossing Over

Evolution thrives in a universe of diversity and will go to any length to increase that diversity. Some of evolution's greatest battles have been against the laws of chemistry and the principles of mechanics and has achieved amazing victories by deploying the genetic code and sexual reproduction. These are thankfully not our problems any more and evolution's brilliant solutions are not generally relevant to the questions we face today. Occasionally they do work, for example, the task of cutting up material for the manufacture of garments with minimum wastage has challenged scientists for many years and a whole battery of cutting algorithms are now in place driving automated machinery with unhesitating confidence and supreme efficiency.

One approach to this problem has applied the heuristic of 'crossing-over' in chromosomes in order to search for better cutting patterns. Just as good and bad genes are shuffled during meiosis to yield weak and strong offspring, so areas of dense packing with little wasted material in 'parent' cutting patterns can be swapped with less efficient arrangements of the same shapes in other parents. For example optimal packing in the left-hand part of a cutting pattern can be swapped with the loose arrangement in the left-hand part of another pattern to produce a design which incorporates the best ideas from both. The crossing-over heuristic is a good one in this example, but only because there are direct parallels with nature which uses sexual reproduction to overcome the effects of linkage along the chromosomes and to generate countless unique combinations of genes for testing by natural selection. In our example the 'genes' are the groups of shapes strung out along each strip of cloth and 'crossing over' allows well packed groups to spread throughout the evolving population of best cutting patterns. However, it would be wrong now to believe that chromosomes and the crossing over heuristic have to be modelled in the solution of *any* search problem. To deliberately select a representation which forces the use of crossing over, is to use a solution which was originally designed to solve another and more difficult problem, and imposes the dreadful limitations of linkage on the search

process. Why copy an approach that has run out of steam and encapsulates an artificial and arbitrary constraint?

The best search strategies

Nature is telling us not to impede our search algorithms with Lamarckian heuristics and thereby to avoid the exclusion of large tracts of solution space. The molecular mechanisms Nature has put in place maximise diversity and evolvability by eliminating as many constraints as possible in her search for perfection. By the same token computer search techniques should strive to employ representations that do not preclude solutions or impose a bias upon the results however tempted we are to interfere. This conclusion is endorsed by the potential consequences of Gödel's Incompleteness theorem which effectively states that there are truths which cannot be reached by any 'informed' rule-based framework for search.

The only search method that does not suffer from any form of constraint is a purely random search, often the first technique to be dismissed. In a random search all possible solutions are accessible, but at great expense if the search space is large. Even nature does not resort to this extreme, but instead has relied upon the separate incremental evolution of smaller component elements or genes. This strategy still brings with it the inevitable constraint of the evolutionary path itself; only a limited number of changes are possible once an organism has assumed significant structure, and the effects of linkage on chromosomes is also present. Fortunately computers do not need chromosomes to hold information and can eliminate all effects of linkage, but once an evolutionary search has begun there are still dangers that the process will stagnate because the best result achieved so far simply might not lie on any path leading to the optimal solution. One answer is to invoke multiple searches from different initial conditions to increase the chance of hitting a more fruitful part of the search space; again this is unlikely to yield satisfactory results in a large space of possibilities.

The success of nature's genes tells us that practical evolutionary search principles will succeed on problems that can be partitioned into component sub-problems each of which may be solved separately but using a criterion for success that requires good performance in concert with other sub-problem solutions. This strategy frees the search to select from a huge number of possible alternative paths and imposes minimal restrictions upon the searchable solution space - there are no vestiges of Lamarckian guidelines here. Typical problems[22] that are amenable to evolutionary search tend to be found in the domain of pattern recognition where the task frequently reduces to the determination of a set of independent features capable of classifying an unlimited number of noisy and distorted real-life patterns[23]. Whether such an evolutionary process plays an important part in the recognition and anticipation taking place in the brain is a subject currently being hotly debated[24].

---

[1] J Monod, "Chance and Necessity", Collins, 1972.

[2] The so called Central Dogma of molecular biology propounded by Francis Crick states that transfers do not take place from protein to DNA and only rarely from RNA to DNA in certain virus infected cells. - F Crick, "Central Dogma of Molecular Biology", Nature, Vol 227, Aug 8 1970.

[3] Steele E J, "Somatic selection and adaptive evolution", Toronto, Williams and Wallace, 1979.

[4] Approximately half of all human genes has been sampled as of 15 June, 1996.
[5] Curiously the salamander possesses 30 times as much genetic material in the chromosomes as humans.
[6] Each cell has about 6 feet of DNA and a human body has about 50 billion cells in it.
[7] There may be as many as 400,000 turns in the helix so there is plenty of scope for getting in a tangle!
[8] Alberts B et al, "The Molecular Biology of the Cell", Garland Publishing Inc, New York & London, 1994
[9] Products other than mRNA directly transcribed from DNA include Transfer RNA (tRNA) and Ribosomal RNA (rRNA) both of which figure in the mechanism of gene expression.
[10] The Genetic Code:

Codon	Amino acid	Codon	Amino acid	Codon	Amino acid	Codon	Amino acid
UUU	phenylalanine	UCU	serine	UAU	tyrosine	UGU	cysteine
UUC	phenylalanine	UCC	serine	UAC	tyrosine	UGC	cysteine
UUA	leucine	UCA	serine	UAA	Stop	UGA	Stop
UUG	leucine	UCG	serine	UAG	Stop	UGG	tryptophan
CUU	leucine	CCU	proline	CAU	histidine	CGU	arginine
CCU	leucine	CCC	proline	CAC	hisidine	CGC	arginine
CUA	leucine	CCA	proline	CAA	glutamine	CGA	arginine
CUG	leucine	CCG	proline	CAG	glutamine	CGG	arginine
AUU	isoleucine	ACU	threonine	AAU	asparagine	AGU	serine
ACU	isoleucine	ACC	threonine	AAC	asparagine	AGC	serine
AUA	isoleucine	ACA	threonine	AAA	lysine	AGA	arginine
AUG	methionine	ACG	threonine	AAG	lysine	AGG	arginine
GUU	valine	GCU	alanine	GAU	aspartic acid	GGU	glycine
GCU	valine	GCC	alanine	GAC	aspartic acid	GGC	glycine
GUA	valine	GCA	alanine	GAA	glutamic acid	GGA	glycine
GUG	valine	GCG	alanine	GAG	glutamic acid	GGG	glycine

[11] The genetic code was assumed to be universal for many years, but the community was surprised when it was discovered that the human mitochondrial genes use a slightly different genetic code. Non-standard codes were discovered in a few other organisms, and there is at least one example of a codon having a different meaning in different genes within the same organism.
[12] See T A Brown, "Genetics a molecular approach", Chapman & Hall, 1993.
[13] S Miller & H Urey, "A production of amino acids under possible primitive conditions", Science, 117, pages 528 - 529, 1953.
[14] Opinions differ on the composition of the primordal atmosphere.
[15] One of the 23 pairs of chromosomes present in males consists of an X and a Y chromosome. The Y chromosome determines maleness, but only exchanges genetic material with its fellow X at the tip, and appears to have lost much of its functionality as a result of being unable to shed and acquire new genes.
[16] In the case of humans only about 3% of DNA is thought to specify the portions of our genes that encode proteins.
[17] Enzymes are used to break up the DNA at points corresponding to specific nucleotide patterns. This produces a large number of strands of various lengths which reflect aspects of the original sequence. The distribution of strand lengths yields the unique fingerprint.
[18] J Cohen, "Reproduction", London, Butterworths, 1977.
[19] Assuming a linear distribution of genes across the unused DNA one might expect adjacent genes to

be at least ten times more likely to be segregated during crossing over.
[20] G Edelman, “Bright air, brilliant fire”, Penguin, 1994.
[21] M J Plunkett et al, “Combinatorial chemistry and new drugs”, Scientific American, April 1997.
[22] Shackleton M et al, “Nature Inspired Computation: toward novel and radical computing”, BTTJ, this issue.
[23] F W M Stentiford, “Automatic feature design for OCR using an evolutionary search procedure”, IEEE Trans PAMI, Vol 7, No 3, May 1985.
[24] D C Dennett, “Darwin’s Dangerous Idea”, Penguin, 1995.