The Information Challenge
By Richard Dawkins
Article in The Skeptic Vol 18, No 4 Dec 1998

In September 1997, I allowed an Australian film crew into my house in Oxford without realising that their purpose was creationist propaganda. In the course of a suspiciously amateurish interview, they issued a truculent challenge to me to "give an example of a genetic mutation or an evolutionary process which can be seen to increase the information in the genome." It is the kind of question only a creationist would ask in that way, and it was at this point I tumbled to the fact that I had been duped into granting an interview to creationists - a thing I normally don't do, for good reasons. In my anger I refused to discuss the question further, and told them to stop the camera. However, I eventually withdrew my peremptory termination of the interview as a whole. This was solely because they pleaded with me that they had come all the way from Australia specifically in order to interview me. Even if this was a considerable exaggeration, it seemed, on reflection, ungenerous to tear up the legal release form and throw them out. I therefore relented.

My generosity was rewarded in a fashion that anyone familiar with fundamentalist tactics might have predicted. When I eventually saw the film a year later 1, I found that it had been edited to give the false impression that I was incapable of answering the question about information content 2. In fairness, this may not have been quite as intentionally deceitful as it sounds. You have to understand that these people really believe that their question cannot be answered! Pathetic as it sounds, their entire journey from Australia seems to have been a quest to film an evolutionist failing to answer it.

With hindsight - given that I had been suckered into admitting them into my house in the first place - it might have been wiser simply to answer the question. But I like to be understood whenever I open my mouth - I have a horror of blinding people with science - and this was not a question that could be answered in a soundbite. First you first have to explain the technical meaning of "information". Then the relevance to evolution, too, is complicated - not really difficult but it takes time. Rather than engage now in further recriminations and disputes about exactly what happened at the time of the interview (for, to be fair, I should say that the Australian producer's memory of events seems to differ from mine), I shall try to redress the matter now in constructive fashion by answering the original question, the "Information Challenge", at adequate length - the sort of length you can achieve in a proper article.

Information

The technical definition of "information" was introduced by the American engineer Claude Shannon in 1948. An employee of the Bell Telephone Company, Shannon was concerned to measure information as an economic commodity. It is costly to send messages along a telephone line. Much of what passes in a message is not information: it is redundant. You could save money by recoding the message to remove the redundancy. Redundancy was a second technical term introduced by Shannon, as the inverse of information. Both definitions were mathematical, but we can convey Shannon's intuitive meaning in words.

Redundancy is any part of a message that is not informative, either because the recipient already knows it (is not surprised by it) or because it duplicates other parts of the message. In the sentence "Rover is a poodle dog", the word "dog" is redundant because "poodle" already tells us that Rover is a dog. An economical telegram would omit it, thereby increasing the informative proportion of the message. "Arr JFK Fri pm pls mt BA Cncrd flt" carries the same information as the much longer, but more redundant, "I'll be arriving at John F Kennedy airport on Friday evening; please meet the British Airways Concorde flight". Obviously the brief, telegraphic message is cheaper to send (although the recipient may have to work harder to decipher it - redundancy has its virtues if we forget economics). Shannon wanted to find a mathematical way to capture the idea that any message could be broken into the information (which is worth paying for), the redundancy (which can, with economic advantage, be deleted from the message because, in effect, it can be

reconstructed by the recipient) and the noise (which is just random rubbish).

"It rained in Oxford every day this week" carries relatively little information, because the receiver is not surprised by it. On the other hand, "It rained in the Sahara desert every day this week" would be a message with high information content, well worth paying extra to send. Shannon wanted to capture this sense of information content as "surprise value". It is related to the other sense - "that which is not duplicated in other parts of the message" - because repetitions lose their power to surprise. Note that Shannon's definition of the quantity of information is independent of whether it is true. The measure he came up with was ingenious and intuitively satisfying. Let's estimate, he suggested, the receiver's ignorance or uncertainty before receiving the message, and then compare it with the receiver's remaining ignorance after receiving the message. The quantity of ignorance-reduction is the information content. Shannon's unit of information is the bit, short for "binary digit". One bit is defined as the amount of information needed to halve the receiver's prior uncertainty, however great that prior uncertainty was (mathematical readers will notice that the bit is, therefore, a logarithmic measure).

In practice, you first have to find a way of measuring the prior uncertainty - that which is reduced by the information when it comes. For particular kinds of simple message, this is easily done in terms of probabilities. An expectant father watches the Caesarian birth of his child through a window into the operating theatre. He can't see any details, so a nurse has agreed to hold up a pink card if it is a girl, blue for a boy. How much information is conveyed when, say, the nurse flourishes the pink card to the delighted father? The answer is one bit - the prior uncertainty is halved. The father knows that a baby of some kind has been born, so his uncertainty amounts to just two possibilities - boy and girl - and they are (for purposes of this discussion) equal. The pink card halves the father's prior uncertainty from two possibilities to one (girl). If there'd been no pink card but a doctor had walked out of the operating theatre, shook the father's hand and said "Congratulations old chap, I'm delighted to be the first to tell you that you have a daughter", the information conveyed by the 17 word message would still be only one bit.

Computer information

Computer information is held in a sequence of noughts and ones. There are only two possibilities, so each 0 or 1 can hold one bit. The memory capacity of a computer, or the storage capacity of a disc or tape, is often measured in bits, and this is the total number of 0s or 1s that it can hold. For some purposes, more convenient units of measurement are the byte (8 bits), the kilobyte (1000 bytes or 8000 bits), the megabyte (a million bytes or 8 million bits) or the gigabyte (1000 million bytes or 8000 million bits). Notice that these figures refer to the total available capacity. This is the maximum quantity of information that the device is capable of storing. The actual amount of information stored is something else. The capacity of my hard disc happens to be 4.2 gigabytes. Of this, about 1.4 gigabytes are actually being used to store data at present. But even this is not the true information content of the disc in Shannon's sense. The true information content is smaller, because the information could be more economically stored. You can get some idea of the true information content by using one of those ingenious compression programs like "Stuffit". Stuffit looks for redundancy in the sequence of 0s and 1s, and removes a hefty proportion of it by recoding - stripping out internal predictability. Maximum information content would be achieved (probably never in practice) only if every 1 or 0 surprised us equally. Before data is transmitted in bulk around the Internet, it is routinely compressed to reduce redundancy.

That's good economics. But on the other hand it is also a good idea to keep some redundancy in messages, to help correct errors. In a message that is totally free of redundancy, after there's been an error there is no means of reconstructing what was intended. Computer codes often incorporate deliberately redundant "parity bits" to aid in error detection. DNA, too, has various error-correcting procedures which depend upon redundancy. When I come on to talk of genomes, I'll return to the three-way distinction between total information capacity, information capacity actually used, and true information content.

It was Shannon's insight that information of any kind, no matter what it means, no matter whether it

is true or false, and no matter by what physical medium it is carried, can be measured in bits, and is translatable into any other medium of information. The great biologist J B S Haldane used Shannon's theory to compute the number of bits of information conveyed by a worker bee to her hivemates when she "dances" the location of a food source (about 3 bits to tell about the direction of the food and another 3 bits for the distance of the food). In the same units, I recently calculated that I'd need to set aside 120 megabits of laptop computer memory to store the triumphal opening chords of Richard Strauss's "Also Sprach Zarathustra" (the "2001" theme) which I wanted to play in the middle of a lecture about evolution. Shannon's economics enable you to calculate how much modem time it'll cost you to e-mail the complete text of a book to a publisher in another land. Fifty years after Shannon, the idea of information as a commodity, as measurable and interconvertible as money or energy, has come into its own.

DNA information

DNA carries information in a very computer-like way, and we can measure the genome's capacity in bits too, if we wish. DNA doesn't use a binary code, but a quaternary one. Whereas the unit of information in the computer is a 1 or a 0, the unit in DNA can be T, A, C or G. If I tell you that a particular location in a DNA sequence is a T, how much information is conveyed from me to you? Begin by measuring the prior uncertainty. How many possibilities are open before the message "T" arrives? Four. How many possibilities remain after it has arrived? One. So you might think the information transferred is four bits, but actually it is two. Here's why (assuming that the four letters are equally probable, like the four suits in a pack of cards). Remember that Shannon's metric is concerned with the most economical way of conveying the message. Think of it as the number of yes/no questions that you'd have to ask in order to narrow down to certainty, from an initial uncertainty of four possibilities, assuming that you planned your questions in the most economical way. "Is the mystery letter before D in the alphabet?" No. That narrows it down to T or G, and now we need only one more question to clinch it. So, by this method of measuring, each "letter" of the DNA has an information capacity of 2 bits.

Whenever prior uncertainty of recipient can be expressed as a number of equiprobable alternatives N, the information content of a message which narrows those alternatives down to one is $\log_2 N$ (the power to which 2 must be raised in order to yield the number of alternatives N). If you pick a card, any card, from a normal pack, a statement of the identity of the card carries $\log_2 52$, or 5.7 bits of information. In other words, given a large number of guessing games, it would take 5.7 yes/no questions on average to guess the card, provided the questions are asked in the most economical way. The first two questions might establish the suit. (Is it red? Is it a diamond?) the remaining three or four questions would successively divide and conquer the suit (is it a 7 or higher? etc.), finally homing in on the chosen card. When the prior uncertainty is some mixture of alternatives that are not equiprobable, Shannon's formula becomes a slightly more elaborate weighted average, but it is essentially similar. By the way, Shannon's weighted average is the same formula as physicists have used, since the nineteenth century, for entropy. The point has interesting implications but I shall not pursue them here.

Information and evolution

That's enough background on information theory. It is a theory which has long held a fascination for me, and I have used it in several of my research papers over the years. Let's now think how we might use it to ask whether the information content of genomes increases in evolution. First, recall the three way distinction between total information capacity, the capacity that is actually used, and the true information content when stored in the most economical way possible. The total information capacity of the human genome is measured in gigabits. That of the common gut bacterium Escherichia coli is measured in megabits. We, like all other animals, are descended from an ancestor which, were it available for our study today, we'd classify as a bacterium. So perhaps, during the billions of years of evolution since that ancestor lived, the information capacity of our genome has gone up about three orders of magnitude (powers of ten) - about a thousandfold. This is satisfyingly plausible and comforting to human dignity. Should human dignity feel wounded, then, by the fact that the crested newt, Triturus cristatus, has a genome capacity estimated at 40 gigabits, an order of magnitude larger than the human genome? No, because, in any case, most of the

capacity of the genome of any animal is not used to store useful information. There are many nonfunctional pseudogenes (see below) and lots of repetitive nonsense, useful for forensic detectives but not translated into protein in the living cells. The crested newt has a bigger "hard disc" than we have, but since the great bulk of both our hard discs is unused, we needn't feel insulted. Related species of newt have much smaller genomes. Why the Creator should have played fast and loose with the genome sizes of newts in such a capricious way is a problem that creationists might like to ponder. From an evolutionary point of view the explanation is simple (see The Selfish Gene pp 44-45 and p 275 in the Second Edition).

Gene duplication

Evidently the total information capacity of genomes is very variable across the living kingdoms, and it must have changed greatly in evolution, presumably in both directions. Losses of genetic material are called deletions. New genes arise through various kinds of duplication. This is well illustrated by haemoglobin, the complex protein molecule that transports oxygen in the blood.

Human adult haemoglobin is actually a composite of four protein chains called globins, knotted around each other. Their detailed sequences show that the four globin chains are closely related to each other, but they are not identical. Two of them are called alpha globins (each a chain of 141 amino acids), and two are beta globins (each a chain of 146 amino acids). The genes coding for the alpha globins are on chromosome 11; those coding for the beta globins are on chromosome 16. On each of these chromosomes, there is a cluster of globin genes in a row, interspersed with some junk DNA. The alpha cluster, on Chromosome 11, contains seven globin genes. Four of these are pseudogenes, versions of alpha disabled by faults in their sequence and not translated into proteins. Two are true alpha globins, used in the adult. The final one is called zeta and is used only in embryos. Similarly the beta cluster, on chromosome 16, has six genes, some of which are disabled, and one of which is used only in the embryo. Adult haemoglobin, as we've seen contains two alpha and two beta chains.

Never mind all this complexity. Here's the fascinating point. Careful letter-by-letter analysis shows that these different kinds of globin genes are literally cousins of each other, literally members of a family. But these distant cousins still coexist inside our own genome, and that of all vertebrates. On a the scale of whole organism, the vertebrates are our cousins too. The tree of vertebrate evolution is the family tree we are all familiar with, its branch-points representing speciation events - the splitting of species into pairs of daughter species. But there is another family tree occupying the same timescale, whose branches represent not speciation events but gene duplication events within genomes.

The dozen or so different globins inside you are descended from an ancient globin gene which, in a remote ancestor who lived about half a billion years ago, duplicated, after which both copies stayed in the genome. There were then two copies of it, in different parts of the genome of all descendant animals. One copy was destined to give rise to the alpha cluster (on what would eventually become Chromosome 11 in our genome), the other to the beta cluster (on Chromosome 16). As the aeons passed, there were further duplications (and doubtless some deletions as well). Around 400 million years ago the ancestral alpha gene duplicated again, but this time the two copies remained near neighbours of each other, in a cluster on the same chromosome. One of them was destined to become the zeta of our embryos, the other became the alpha globin genes of adult humans (other branches gave rise to the nonfunctional pseudogenes I mentioned). It was a similar story along the beta branch of the family, but with duplications at other moments in geological history.

Now here's an equally fascinating point. Given that the split between the alpha cluster and the beta cluster took place 500 million years ago, it will of course not be just our human genomes that show the split - possess alpha genes in a different part of the genome from beta genes. We should see the same within-genome split if we look at any other mammals, at birds, reptiles, amphibians and bony fish, for our common ancestor with all of them lived less than 500 million years ago. Wherever it has been investigated, this expectation has proved correct. Our greatest hope of finding a vertebrate that does not share with us the ancient alpha/beta split would be a jawless fish like a

lamprey, for they are our most remote cousins among surviving vertebrates; they are the only surviving vertebrates whose common ancestor with the rest of the vertebrates is sufficiently ancient that it could have predated the alpha/beta split. Sure enough, these jawless fishes are the only known vertebrates that lack the alpha/beta divide.

Gene duplication, within the genome, has a similar historic impact to species duplication ("speciation") in phylogeny. It is responsible for gene diversity, in the same way as speciation is responsible for phyletic diversity. Beginning with a single universal ancestor, the magnificent diversity of life has come about through a series of branchings of new species, which eventually gave rise to the major branches of the living kingdoms and the hundreds of millions of separate species that have graced the earth. A similar series of branchings, but this time within genomes - gene duplications - has spawned the large and diverse population of clusters of genes that constitutes the modern genome.

The story of the globins is just one among many. Gene duplications and deletions have occurred from time to time throughout genomes. It is by these, and similar means, that genome sizes can increase in evolution. But remember the distinction between the total capacity of the whole genome, and the capacity of the portion that is actually used. Recall that not all the globin genes are actually used. Some of them, like theta in the alpha cluster of globin genes, are pseudogenes, recognizably kin to functional genes in the same genomes, but never actually translated into the action language of protein. What is true of globins is true of most other genes. Genomes are littered with nonfunctional pseudogenes, faulty duplicates of functional genes that do nothing, while their functional cousins (the word doesn't even need scare quotes) get on with their business in a different part of the same genome. And there's lots more DNA that doesn't even deserve the name pseudogene. It, too, is derived by duplication, but not duplication of functional genes. It consists of multiple copies of junk, "tandem repeats", and other nonsense which may be useful for forensic detectives but which doesn't seem to be used in the body itself.

Once again, creationists might spend some earnest time speculating on why the Creator should bother to litter genomes with untranslated pseudogenes and junk tandem repeat DNA.
Information in the genome

Can we measure the information capacity of that portion of the genome which is actually used? We can at least estimate it. In the case of the human genome it is about 2% - considerably less than the proportion of my hard disc that I have ever used since I bought it. Presumably the equivalent figure for the crested newt is even smaller, but I don't know if it has been measured. In any case, we mustn't run away with a chaunvinistic idea that the human genome somehow ought to have the largest DNA database because we are so wonderful. The great evolutionary biologist George C Williams has pointed out that animals with complicated life cycles need to code for the development of all stages in the life cycle, but they only have one genome with which to do so. A butterfly's genome has to hold the complete information needed for building a caterpillar as well as a butterfly. A sheep liver fluke has six distinct stages in its life cycle, each specialised for a different way of life. We shouldn't feel too insulted if liver flukes turned out to have bigger genomes than we have (actually they don't).

Remember, too, that even the total capacity of genome that is actually used is still not the same thing as the true information content in Shannon's sense. The true information content is what's left when the redundancy has been compressed out of the message, by the theoretical equivalent of Stuffit. There are even some viruses which seem to use a kind of Stuffit-like compression. They make use of the fact that the RNA (not DNA in these viruses, as it happens, but the principle is the same) code is read in triplets. There is a "frame" which moves along the RNA sequence, reading off three letters at a time. Obviously, under normal conditions, if the frame starts reading in the wrong place (as in a so-called frame-shift mutation), it makes total nonsense: the "triplets" that it reads are out of step with the meaningful ones. But these splendid viruses actually exploit frame-shifted reading. They get two messages for the price of one, by having a completely different message embedded in the very same series of letters when read frame-shifted. In principle you could even

get three messages for the price of one, but I don't know whether there are any examples.

Information in the body

It is one thing to estimate the total information capacity of a genome, and the amount of the genome that is actually used, but it's harder to estimate its true information content in the Shannon sense. The best we can do is probably to forget about the genome itself and look at its product, the "phenotype", the working body of the animal or plant itself. In 1951, J W S Pringle, who later became my Professor at Oxford, suggested using a Shannon-type information measure to estimate "complexity'. Pringle wanted to express complexity mathematically in bits, but I have long found the following verbal form helpful in explaining his idea to students.

We have an intuitive sense that a lobster, say, is more complex (more "advanced", some might even say more "highly evolved") than another animal, perhaps a millipede. Can we measure something in order to confirm or deny our intuition? Without literally turning it into bits, we can make an approximate estimation of the information contents of the two bodies as follows. Imagine writing a book describing the lobster. Now write another book describing the millipede down to the same level of detail. Divide the word-count in one book by the word-count in the other, and you have an approximate estimate of the relative information content of lobster and millipede. It is important to specify that both books describe their respective animals "down to the same level of detail". Obviously if we describe the millipede down to cellular detail, but stick to gross anatomical features in the case of the lobster, the millipede would come out ahead.

But if we do the test fairly, I'll bet the lobster book would come out longer than the millipede book. It's a simple plausibility argument, as follows. Both animals are made up of segments - modules of bodily architecture that are fundamentally similar to each other, arranged fore-and-aft like the trucks of a train. The millipede's segments are mostly identical to each other. The lobster's segments, though following the same basic plan (each with a nervous ganglion, a pair of appendages, and so on) are mostly different from each other. The millipede book would consist of one chapter describing a typical segment, followed by the phrase "Repeat N times" where N is the number of segments. The lobster book would need a different chapter for each segment. This isn't quite fair on the millipede, whose front and rear end segments are a bit different from the rest. But I'd still bet that, if anyone bothered to do the experiment, the estimate of lobster information content would come out substantially greater than the estimate of millipede information content.

It's not of direct evolutionary interest to compare a lobster with a millipede in this way, because nobody thinks lobsters evolved from millipedes. Obviously no modern animal evolved from any other modern animal. Instead, any pair of modern animals had a last common ancestor which lived at some (in principle) discoverable moment in geological history. Almost all of evolution happened way back in the past, which makes it hard to study details. But we can use the "length of book" thought-experiment to agree upon what it would mean to ask the question whether information content increases over evolution, if only we had ancestral animals to look at.

The answer in practice is complicated and controversial, all bound up with a vigorous debate over whether evolution is, in general, progressive. I am one of those associated with a limited form of yes answer. My colleague Stephen Jay Gould tends towards a no answer. I don't think anybody would deny that, by any method of measuring - whether bodily information content, total information capacity of genome, capacity of genome actually used, or true ("Stuffit compressed") information content of genome - there has been a broad overall trend towards increased information content during the course of human evolution from our remote bacterial ancestors. People might disagree, however, over two important questions: first, whether such a trend is to be found in all, or a majority of evolutionary lineages (for example parasite evolution often shows a trend towards decreasing bodily complexity, because parasites are better off being simple); second, whether, even in lineages where there is a clear overall trend over the very long term, it is bucked by so many reversals and re-reversals in the shorter term as to undermine the very idea of progress. This is not the place to resolve this interesting controversy. There are distinguished biologists with good arguments on both sides.

Supporters of "intelligent design" guiding evolution, by the way, should be deeply committed to the view that information content increases during evolution. Even if the information comes from God, perhaps especially if it does, it should surely increase, and the increase should presumably show itself in the genome. Unless, of course - for anything goes in such addle-brained theorising - God works his evolutionary miracles by nongenetic means.

Perhaps the main lesson we should learn from Pringle is that the information content of a biological system is another name for its complexity. Therefore the creationist challenge with which we began is tantamount to the standard challenge to explain how biological complexity can evolve from simpler antecedents, one that I have devoted three books to answering (The Blind Watchmaker, River Out of Eden, Climbing Mount Improbable) and I do not propose to repeat their contents here. The "information challenge" turns out to be none other than our old friend: "How could something as complex as an eye evolve?" It is just dressed up in fancy mathematical language - perhaps in an attempt to bamboozle. Or perhaps those who ask it have already bamboozled themselves, and don't realise that it is the same old - and thoroughly answered - question.

The Genetic Book of the Dead

Let me turn, finally, to another way of looking at whether the information content of genomes increases in evolution. We now switch from the broad sweep of evolutionary history to the minutiae of natural selection. Natural selection itself, when you think about it, is a narrowing down from a wide initial field of possible alternatives, to the narrower field of the alternatives actually chosen. Random genetic error (mutation), sexual recombination and migratory mixing, all provide a wide field of genetic variation: the available alternatives. Mutation is not an increase in true information content, rather the reverse, for mutation, in the Shannon analogy, contributes to increasing the prior uncertainty. But now we come to natural selection, which reduces the "prior uncertainty" and therefore, in Shannon's sense, contributes information to the gene pool. In every generation, natural selection removes the less successful genes from the gene pool, so the remaining gene pool is a narrower subset. The narrowing is nonrandom, in the direction of improvement, where improvement is defined, in the Darwinian way, as improvement in fitness to survive and reproduce. Of course the total range of variation is topped up again in every generation by new mutation and other kinds of variation. But it still remains true that natural selection is a narrowing down from an initially wider field of possibilities, including mostly unsuccessful ones, to a narrower field of successful ones. This is analogous to the definition of information with which we began: information is what enables the narrowing down from prior uncertainty (the initial range of possibilities) to later certainty (the "successful" choice among the prior probabilities). According to this analogy, natural selection is by definition a process whereby information is fed into the gene pool of the next generation.

If natural selection feeds information into gene pools, what is the information about? It is about how to survive. Strictly it is about how to survive and reproduce, in the conditions that prevailed when previous generations were alive. To the extent that present day conditions are different from ancestral conditions, the ancestral genetic advice will be wrong. In extreme cases, the species may then go extinct. To the extent that conditions for the present generation are not too different from conditions for past generations, the information fed into present-day genomes from past generations is helpful information. Information from the ancestral past can be seen as a manual for surviving in the present: a family bible of ancestral "advice" on how to survive today. We need only a little poetic licence to say that the information fed into modern genomes by natural selection is actually information about ancient environments in which ancestors survived.

This idea of information fed from ancestral generations into descendant gene pools is one of the themes of my new book, Unweaving the Rainbow. It takes a whole chapter, "The Genetic Book of the Dead", to develop the notion, so I won't repeat it here except to say two things. First, it is the whole gene pool of the species as a whole, not the genome of any particular individual, which is best seen as the recipient of the ancestral information about how to survive. The genomes of particular individuals are random samples of the current gene pool, randomised by sexual

recombination. Second, we are privileged to "intercept" the information if we wish, and "read" an animal's body, or even its genes, as a coded description of ancestral worlds. To quote from Unweaving the Rainbow: "And isn't it an arresting thought? We are digital archives of the African Pliocene, even of Devonian seas; walking repositories of wisdom out of the old days. You could spend a lifetime reading in this ancient library and die unsated by the wonder of it."

1 The producers never deigned to send me a copy: I completely forgot about it until an American colleague called it to my attention.

2 See Barry Williams (1998): "Creationist Deception Exposed", The Skeptic 18, 3, pp 7-10, for an account of how my long pause (trying to decide whether to throw them out) was made to look like hesitant inability to answer the question, followed by an apparently evasive answer to a completely different question.