

THREE-“LETTER” SEQUENCES, or codons, of DNA and RNA (*shown*) encode the individual amino acids that build and maintain all life on earth.

Evolution Encoded

New discoveries about the rules governing how genes encode proteins have revealed nature’s sophisticated “programming” for protecting life from catastrophic errors while accelerating evolution



On April 14, 2003, scientists announced to the world that they had finished sequencing the human genome—logging the three billion pairs of DNA nucleotides that describe how to make a human being. But finding all the working genes amid the junk in the sequence remains a further challenge, as does gaining a better understanding of how and when genes are activated and how their instructions affect the behavior of the protein molecules they describe. So it is no wonder that Human Genome Project leader Francis S. Collins has called the group’s accomplishment only “the end of the beginning.”

Collins was also alluding to an event commemorated that same week: the beginning of the beginning, 50 years earlier, when James D. Watson and Francis H. Crick revealed the structure of the DNA molecule itself. That, too, was an exciting time. Scientists knew that the molecule they were finally able to visualize contained nothing less than the secret of life, which permitted organisms to store themselves as a set of blueprints and convert this stored information back into live metabolism. In subsequent years, attempts to figure out how this conversion took place captivated the scientific world. DNA’s alphabet was known to consist of only four types of nucleotide. So the information encoded in the double helix had to be *decoded* according

By Stephen J. Freeland and Laurence D. Hurst

to some rules to tell cells which of 20 amino acids to string together to constitute the thousands of proteins that make up billions of life-forms. Indeed, the entire living world had to be perpetually engaged in frenetic decryption, as eggs hatched, seeds germinated, fungus spread and bacteria divided.

But so little was understood at the time about the cellular machinery translating DNA's message that attempts to crack this genetic code focused on the mathematics of the problem. Many ear-

ly proposals proved wrong, a few spectacularly, although their sheer ingenuity and creativity still provide fascinating reading. In fact, when the actual code was finally deciphered during the 1960s, it nearly disappointed. Nature's version looked less elegant than several of the theorists' hypotheses.



Certain codons were just redundant. MANY CAME TO VIEW NATURE'S REAL CODE as little more than a random accident.

ly proposals proved wrong, a few spectacularly, although their sheer ingenuity and creativity still provide fascinating reading. In fact, when the actual code was finally deciphered during the 1960s, it nearly disappointed. Nature's version looked less elegant than several of the theorists' hypotheses.

Only in recent years have new discoveries about the code revealed just how sophisticated a piece of programming it really is. Why nature chose these basic rules and why they have survived three billion or so years of natural selection have started to become clear. We can now show that the code's rules may actually speed evolution while protecting life from making disastrous errors in protein synthesis. Studying the code is also providing clues to solving some of those remaining challenges facing laboratories in the post-genome era. In going back to

the *very* beginning to understand the rules of life's underlying code, we are discovering tools for future research. When we speak of the "code" and "decoding," we are being quite literal. Genetic instructions are stored in DNA and RNA, both made of one type of biochemical molecule, nucleic acid. But organisms are mostly built from (and by) a very different type of molecule, protein. So although a gene is traditionally defined as the sequence of nucleotides that describes a single protein, the genetic sen-

tence containing that description must first be translated from one system of symbols into an entirely different kind of system, rather like converting from Morse code to English.

Cracking the Code

WHEN WATSON and Crick described DNA's structure in 1953, they and their contemporaries could see that genes are written in an alphabet of just four "letters"—the bases adenine, cytosine, guanine and thymine (A, C, G and T) that distinguish each nucleotide and form the rungs in DNA's now familiar twisted-ladder shape. The protein alphabet, in contrast, contained 20 different amino acids, so the need for a multinucleotide genetic "word" to specify any given amino acid was obvious. Two-letter combinations of the four bases would yield only 16 possible words, or "codons." But triplet com-

binations produce 64 possible codons, which would be plenty.

Little else was obvious at the time about how genes might be translated into proteins. Today we understand that gene sequences do use three-letter codons to specify individual amino acids and that several steps are needed for the gene's sequence of bases to be converted into a sequence of amino acids. The DNA gene is first copied and edited into a transcript made of RNA, employing similar nucleic acid bases, except that

DNA's thymine is replaced by uracil. This messenger RNA (mRNA) version of the gene is then read by cellular machinery, three letters at a time, while tiny cellular butlers known as transfer RNAs (tRNA) fetch the specified amino acids to be strung together.

But in the early 1950s this process was a black box, leaving only an intriguing mathematical puzzle. And the first proposed solution came not from a biologist but from physicist George Gamow, better known as an originator of the big bang theory. His "diamond code," published in 1954, elegantly combined the arithmetic of getting 20 amino acid meanings from a four-nucleotide alphabet with the physical structure of DNA itself.

Gamow theorized that at every turn in the double helix there was a diamond-shaped space bounded at its four corners by nucleotides. These gaps would allow DNA to act as a template against which amino acids would line up, determined by the nucleotide combinations present at each twist. His model eliminated one corner of each diamond, then sorted the 64 possible three-nucleotide codons into chemically related groups. It also allowed meaningful codons to overlap, depending on the "reading frame," or where one began reading the sequence of letters along the length of the DNA molecule. This kind of data compression was an efficien-

Overview/*Life's Code*

- Genetic instructions for the manufacture of proteins are written in three-letter "words" called codons, each specifying one of 20 amino acids or a "stop translating" sign. The arrangement of these codons and their amino acid meanings was once considered random, but recent discoveries indicate that natural selection has chosen and maintained this order.
- Computer simulations reveal why: compared with hypothetical alternatives, the standard code is exceptionally good at minimizing the harm caused by errors in genes themselves or in the process of translating genes into proteins.

NATURE'S CODE

IF A GENE SEQUENCE is a “sentence” describing a protein, then its basic units are three-letter “words,” or “codons,” each of which translates into one of 20 amino acids or a “stop translating” signal. Cellular machinery transcribes DNA genes into RNA versions—whose nucleotide building blocks are represented by

the letters A, C, G and U—and then translates the RNA genes, codon by codon, into a corresponding amino acid sequence. Nature’s exact amino acid definitions (*below*) were worked out during the early 1960s. But the significance of patterns in the code would not be fully appreciated for several decades.

SYNONYMS AND SIMILARITIES

Many of the 64 possible three-letter codons specify the same amino acid, providing alternative ways for genes to spell out most proteins. These synonymous codons tend to differ by just a single letter, usually the last, forming a pattern of blocks. Codons for amino acids with similar affinities for water also tend to differ by their last letter, and codons sharing the same first letter often code for amino acids that are products or precursors of one another. These features, as it turns out, are crucial to the survival of all organisms and may even help speed their evolution.

		Second Nucleotide Position			
		U	C	A	G
U	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	
C	CUU Leucine	CCU Proline	CAU Histidine	CAU Arginine	
	CUC Leucine	CCC Proline	CAC Histidine	CAC Arginine	
	CUA Leucine	CCA Proline	CAA Glutamine	CAA Arginine	
	CUG Leucine	CCG Proline	CAG Glutamine	CAG Arginine	
A	AUU Isoleucine	ACU Threonine	AAU Asparagine	AAU Serine	
	AUC Isoleucine	ACC Threonine	AAC Asparagine	AAC Serine	
	AUA Isoleucine	ACA Threonine	AAA Lysine	AAA Arginine	
	AUG Methionine	ACG Threonine	AAG Lysine	AAG Arginine	
G	GUU Valine	GCU Alanine	GAU Aspartate	GAU Glycine	
	GUC Valine	GCC Alanine	GAC Aspartate	GAC Glycine	
	GUA Valine	GCA Alanine	GAA Glutamate	GAA Glycine	
	GUG Valine	GCG Alanine	GAG Glutamate	GAG Glycine	

cy prized by coding theorists of the day. Unfortunately, amino acid chains were soon discovered that could not be accounted for by Gamow’s or any other overlapping codes.

At the same time, evidence was suggesting that DNA and amino acids were not interacting with one another directly. Crick developed a hypothesis that so-called adaptor molecules could be serving as intermediaries, and in 1957 he put forth a set of rules by which they might operate. Simply put, Crick’s adaptors recognized only 20 meaningful codons designating each of the 20 amino acids, making the remainder of the 64 possible triplets “nonsense.” Rather than overlapping, Crick’s code was “commaless”: meaningless codons were effectively invisible to the adaptors, so nature needed no figurative punctuation to designate the start of a reading frame. The commaless concept was so streamlined that it immediately won near universal acceptance—that is, until the data again proved an elegant theory wrong.

In the early 1960s experiments showed that even supposed nonsense codons could provoke protein synthesis in a

beaker, and by 1965 the actual amino acid meanings of all 64 possible triplet codons had been worked out in the lab. No tidy numerology was apparent: certain codons were just redundant, with some individual amino acids specified by two, four, even six different codons. After all the enthusiastic speculation, many came to view nature’s real code as little more than a random accident of history.

Frozen Accident?

INDEED, ONCE THE CODE was deciphered, scientists found that organisms as different as humans and bacteria employed the exact same coding rules. Seemingly no variations had occurred in the billions of years since the three basic domains of life—archaea, bacteria and eukaryotes—diverged from a single ancient common ancestor. Consequently, the simple and persuasive “frozen accident” argument, put forth by Crick himself in 1968, came to dominate scientific thinking until recently. “The allocation of codons to amino acids at this point was entirely a matter of chance,” he wrote, but once the code had appeared in any form, it was so fundamental to life that any further

changes would have been catastrophic.

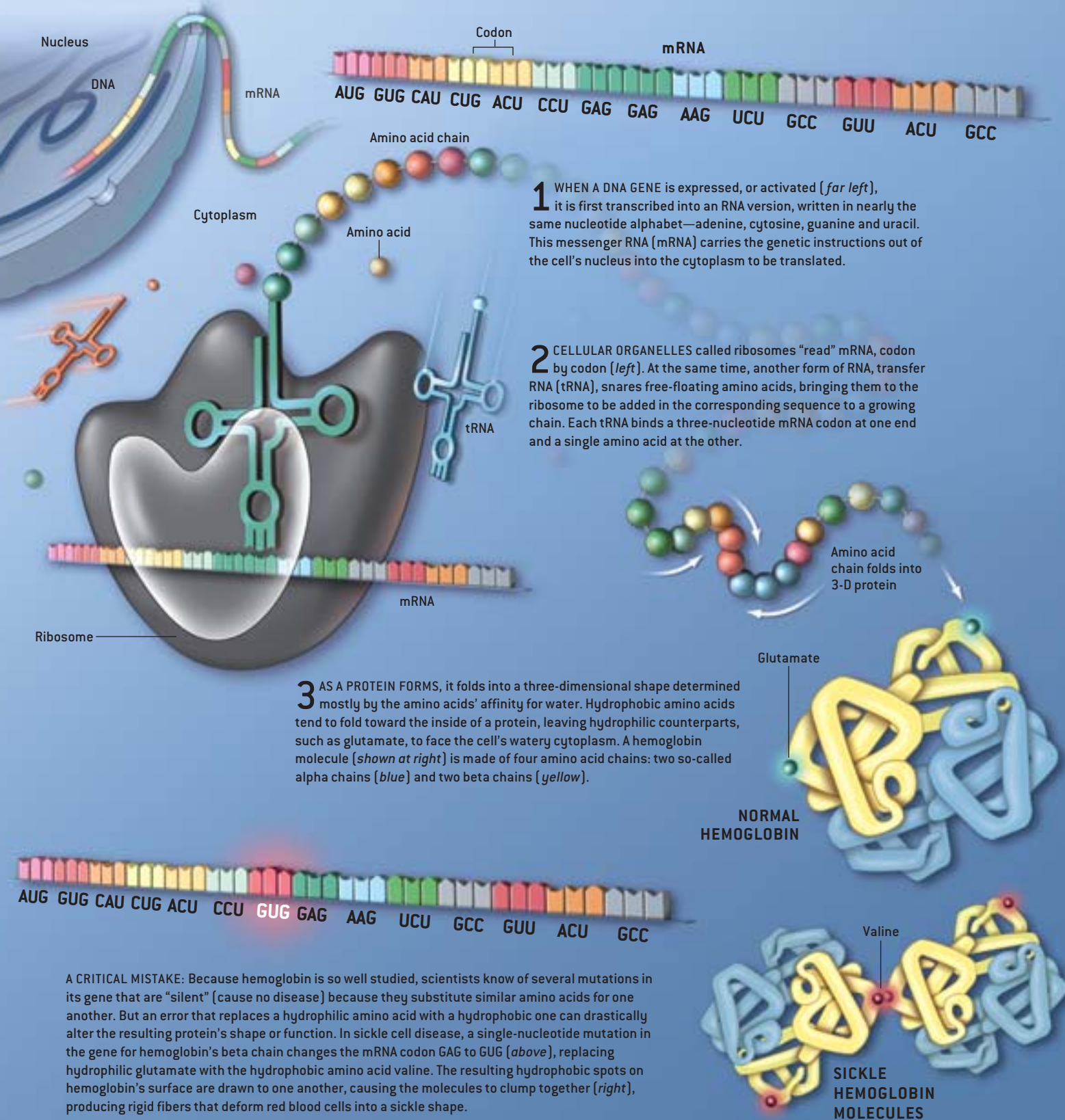
Darwinian natural selection rests on the premise that sometimes a small change in a single gene can prove beneficial if it allows organisms to fare better in their environment. But altering an organism’s decoding rules would be tantamount to simultaneously introducing changes at countless sites throughout its genetic material, producing an utterly dysfunctional metabolism. It would be the difference between introducing a single typo and rewiring the entire typewriter keyboard.

This attractively straightforward reasoning, however, has since proved simplistic. Although most living systems do employ the standard genetic code, scientists now know of at least 16 variants, distributed across a diverse array of evolutionary lineages, that assign different meanings to certain codons. The underlying system remains the same: triple-nucleotide codons are translated into amino acids. But where most organisms would read the RNA codon “CUG” to mean the amino acid leucine, many species of the fungus *Candida* translate CUG as serine. Mitochondria, the tiny power generators within all kinds of cells, have their own

Protecting Proteins

Nature's code minimizes the effects of genetic errors, whether resulting from mutations in genes themselves or mistakes in the translation process. A gene sequence is translated into a corresponding amino acid sequence, which dictates the final three-dimensional structure of the encoded protein [1, 2 and 3].

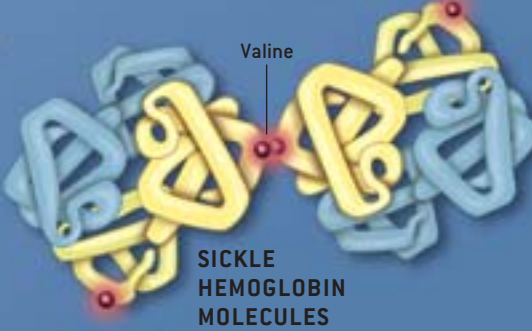
Even when an error inserts the wrong amino acid, the code's arrangement ensures that the substitute is often chemically similar to the intended amino acid, and the final protein is relatively unaltered. An exception that illustrates the kind of harm a single-nucleotide error can cause is sickle cell disease (*at bottom*).



1 WHEN A DNA GENE is expressed, or activated (*far left*), it is first transcribed into an RNA version, written in nearly the same nucleotide alphabet—adenine, cytosine, guanine and uracil. This messenger RNA (mRNA) carries the genetic instructions out of the cell's nucleus into the cytoplasm to be translated.

2 CELLULAR ORGANELLES called ribosomes “read” mRNA, codon by codon (*left*). At the same time, another form of RNA, transfer RNA (tRNA), snares free-floating amino acids, bringing them to the ribosome to be added in the corresponding sequence to a growing chain. Each tRNA binds a three-nucleotide mRNA codon at one end and a single amino acid at the other.

3 AS A PROTEIN FORMS, it folds into a three-dimensional shape determined mostly by the amino acids' affinity for water. Hydrophobic amino acids tend to fold toward the inside of a protein, leaving hydrophilic counterparts, such as glutamate, to face the cell's watery cytoplasm. A hemoglobin molecule (*shown at right*) is made of four amino acid chains: two so-called alpha chains (*blue*) and two beta chains (*yellow*).



AUG GUG CAU CUG ACU CCU **GUG** GAG AAG UCU GCC GUU ACU GCC

A CRITICAL MISTAKE: Because hemoglobin is so well studied, scientists know of several mutations in its gene that are “silent” [cause no disease] because they substitute similar amino acids for one another. But an error that replaces a hydrophilic amino acid with a hydrophobic one can drastically alter the resulting protein's shape or function. In sickle cell disease, a single-nucleotide mutation in the gene for hemoglobin's beta chain changes the mRNA codon GAG to GUG (*above*), replacing hydrophilic glutamate with the hydrophobic amino acid valine. The resulting hydrophobic spots on hemoglobin's surface are drawn to one another, causing the molecules to clump together (*right*), producing rigid fibers that deform red blood cells into a sickle shape.

genomes, and many have also developed their own codon assignments. For instance, in the mitochondrial genome of baker's yeast (*Saccharomyces cerevisiae*), four of the six codons that normally encode leucine instead encode threonine.

As discoveries of these variations proliferated during the 1990s, it became clear that the code is not frozen at all. It can evolve, which means that it probably *did* evolve. So nature's standard codon–amino acid assignments, refined and preserved by billions of years of natural selection, are no accident. In fact, their arrangement does an excellent job of minimizing the impact of accidents.

Damage Control

EVERY CODING SYSTEM has to contend with the possibility of mistakes, but not all errors are equally damaging. In

mistake here or there does not greatly affect the resulting protein.

Defining “similar” with regard to amino acids can be complex: the 20 amino acids differ from one another in all sorts of properties, from size to shape to electric charge. What Woese and others noted is that codons sharing two out of three bases tend to code for amino acids that are much alike in the extent to which they are repelled by or attracted to water. This property is crucial to the ultimate functioning of the protein. A newly made amino acid chain folds into a distinctive shape depending on the positioning of hydrophobic amino acids, which like to cluster together away from the cell's watery cytoplasm, leaving hydrophiles to form the protein's surface.

The remarkable feature of the genetic code is that when a single-nucleotide er-

the resulting amino acids' hydrophobicity caused by all possible single-letter changes to all 64 codons of the code. This value represents the genetic code's susceptibility to errors but is of little meaning on its own. We needed to know how nature's coding system stacks up against possible alternatives.

To generate these hypothetical alternative codes, we had to begin with certain assumptions about realistic restrictions under which a code would operate in a world made of DNA, RNAs and amino acids. One observation is that mistakes in translation of mRNA into a corresponding amino acid occur most frequently at the codon's third position. This spot is simply where the binding affinity between the mRNA and tRNA is weakest, which is why Crick dubbed the phenomenon “wobble.” But synonymous codons—

The code can evolve, which means that it probably did evolve. NATURE'S AMINO ACID ASSIGNMENTS are no accident.



English, vowels and consonants are very different, so that replacement of “s” by “a” makes this message significantly easier to understand. In contrast, the letters “s” and “z” have a similar sound, such that this phrase remains easily understandable. For an error-prone system, a good coding strategy would be one that reduces the effect of the inevitable occasional mistake.

In a living organism, errors come in many forms. Sometimes the original DNA version of a gene changes (a mutation). Sometimes the wrong adaptor (tRNA) binds to the mRNA transcript of a gene, misincorporating an amino acid into a nascent protein [see box on opposite page]. But even when scientists considered the code a product of chance, they noticed that it did seem to be arranged well in terms of ensuring that individual errors are of little consequence. As early as 1965 Carl R. Woese, then at the University of Illinois, observed that similar codons (those sharing two of three letters) usually specify similar amino acids, so a

ror occurs, the actual and intended amino acids are often similar in hydrophobicity, making the alteration in the final protein relatively harmless. But just *how* efficient is the code in this regard? This is where, in 1998, we stepped in to develop the observations of earlier scientists.

Testing the Code

FIRST WE TOOK a quantitative measure of the 20 amino acids' hydrophobicity. Next we used those values to calculate the genetic code's error value, which we define as the average change in

those coding for the same amino acid—usually differ by only their last letters, so such mistranslations often yield the same amino acid meaning.

Although this grouping of synonymous codons in itself reduces the error value of the code, the mechanics of wobble make the arrangement more likely to be a biochemical limitation rather than an evolutionary adaptation. Thus, to err on the side of caution when deriving our measure, we should consider only alternative codes that share this feature. Moreover, it is impossible to put a hydropho-

THE AUTHORS

STEPHEN J. FREELAND and LAURENCE D. HURST use bioinformatics to study evolutionary biology. Freeland is assistant professor of bioinformatics at the University of Maryland, Baltimore County, where he is working to convert insights about genetic code evolution into practical approaches to exploring genome data. His students are currently testing his theories by reengineering a human cancer gene to express in *Escherichia coli* according to that organism's codon preferences. Freeland earned his Ph.D. in evolutionary theory at the University of Cambridge, studying under Royal Society Research Fellow Laurence Hurst, who is now professor of evolutionary genetics at the University of Bath in England. Hurst's research concentrates on understanding the structure and evolution of genetic systems, particularly the evolutionary origin of phenomena such as sexes, the order of genes on chromosomes, genomic imprinting and the genetic code itself.

bicity value on the codons assigned to the “stop” signal, so we kept their number and codon assignments the same in all alternative codes.

Using these technical assumptions, we generated alternatives by randomizing the 20 meanings among the 20 codon blocks. This still defined some 2.5×10^{18} possible configurations (approximately equal to the number of seconds that have elapsed since the earth formed). So we took large random samples of these possibilities and found that from a sample of one million alternative codes only about 100 had a

peared orders of magnitude better still, outperforming all but one in a million of the alternatives.

A straightforward explanation for the genetic code’s remarkable resilience is that it results from natural selection. Perhaps there were once many codes, all with different degrees of error susceptibility. Organisms whose codes coped best with error were more likely to survive, and the standard genetic code simply won in the struggle for existence. We know that variant codes are possible, so this assumption is reasonable.

er predictions for an optimal code are limited to the criteria provided by the programmer, and most of the “better” codes that have been described so far are based on oversimplified assumptions about the types of errors that a code encounters in the real world. For example, they ignore the wobble phenomenon, which prevents their algorithms from perceiving the advantage of having synonymous codons differ only in their third letter.

This shortcoming emphasizes a second problem with designer-optimized codes. Natural selection is a “blind de-



By minimizing the effects of any mutation, THE CODE MAXIMIZES THE LIKELIHOOD that a gene mutation will improve the resulting protein.

lower error value than the natural code [see box below].

Still more striking was our finding when we incorporated additional restrictions to reflect observed patterns in the way DNA tends to mutate and the ways in which genes tend to be mistranscribed into RNA. Under these “real world” conditions, the natural code’s error value ap-

Evidence for error minimization as the driving evolutionary force behind the arrangement of the code has its critics, however. Sophisticated computer searches can certainly improve on nature’s choice, even when they accept the premise that a “good” code is one that minimizes the change in amino acid hydrophobicity caused by genetic errors. But comput-

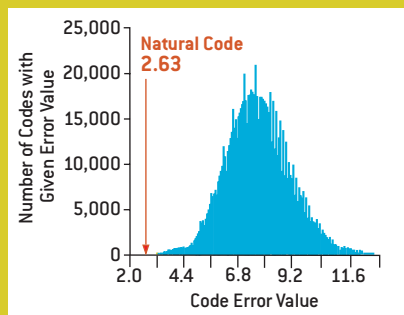
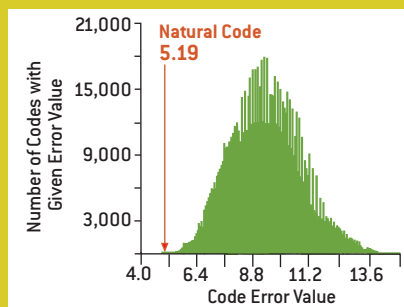
signer,” in that it can only grope toward an ideal by choosing the best alternative within a population of variants at each generation. When we simulate natural selection in this manner, we find that the degree of error minimization achieved by the standard genetic code is still rather impressive: typically less than 3 percent of random theoretical codes can evolve under selection to match its resilience.

In other words, the diamond and commaless codes once looked superior to nature’s own, and computers may generate yet more mathematically idealized codes. But merely demonstrating the possibility of better codes without taking into account the evolutionary process is of dubious relevance to understanding the strength of natural selection’s choice.

Indeed, the standard code is not only a product of natural selection; it may act as a search algorithm to speed evolution. The impact-minimizing properties of the code, with its blocks of both synonymous codons and those specifying biochemically similar amino acids, achieve more than damage control. “Smaller” mutations, in contrast with extreme alterations, are statistically more likely to be beneficial, so by minimizing the effects of any mutation, the code maximizes the likelihood that a gene mutation will lead to an improvement in the resulting protein.

Testing the Code

DAMAGE TO PROTEINS resulting from gene mutations or mistranslations is minimized when the errors substitute amino acids with similar affinities for water (hydrophobicity). If we define a code’s error value as the average change in amino acid hydrophobicity caused by all possible single-base changes within all codons of a code, then a high error value indicates that a code is very vulnerable to errors, and a low value means that a code minimizes their harm. We generated a large random sample of possible codes and found that only 100 of one million alternatives had a lower error value than nature’s code (*top*). When we factor in real-world patterns in the way genes mutate and are mistranslated, nature’s code outperforms all but one in a million of the alternatives (*bottom*).



Using the Code

UNDERSTANDING THE FORCES that shaped the code and how it in turn shapes evolution does more than provide an opportunity to admire nature's skill as a primordial software designer. These insights can also help solve some of the toughest problems facing laboratories in 2004.

Sifting through reams of raw genome sequence data to find the actual genes is a priority in molecular biology, but current searches are limited to matching the characteristics of genes that we already know about. Taking into account the way that the genetic code filters gene mutations can enhance these searches by allowing scientists to recognize highly diversified genes and perhaps infer the function of the proteins they encode. Researchers can even derive clues about the folded protein shape that an amino acid sequence dictates by looking at the error-minimizing properties of its codons and how substitutions might affect amino acid size, charge or hydrophobicity.

Biologists can also apply our awareness of organisms that deviate from the standard code to "disguise" genes for research. Because a single code is nearly universal to all life, it has become common practice to take a gene of interest, such as a human cancer gene, and insert it into an organism, such as *Escherichia coli*, that will churn out the protein the gene encodes. But occasionally the organism fails to express the gene at all, or it produces less of the protein than expected or a slightly different version of the protein found in humans.

This problem can play havoc with biology research, but we now realize that sometimes the failure arises because the organisms exhibit different preferences among synonymous codons. For example, the standard code contains six codons for the amino acid arginine, and human genes tend to favor using the codons AGA and AGG. *E. coli*, however, very rarely uses AGA and often mistranslates it. Knowing these variations and preferences enables us to design versions of the human gene that will work reliably when moved between different organisms.

One of our labs (Freeland's) is developing software applications to help mo-

The Evolving Code

At least 16 organisms from a diverse array of evolutionary lineages deviate from nature's standard code in the amino acid "meaning" they assign to specific codons. Many species of the green alga *Acetabularia*, for example, translate the standard "stop" codons UAG and UAA as the amino acid glycine. To *Candida* fungi, the RNA codon CUG, which normally means leucine, instead specifies serine.

The existence of such variations demonstrates that the code can evolve and may provide clues about how it did. In all three domains of life, a nonstandard 21st amino acid, selenocysteine, is sometimes fabricated in response to the standard stop codon UGA. Selenocysteine is created by chemical tweaking of serine while that amino acid is still attached to its tRNA in the ribosome. In two domains (archaea and bacteria), a 22nd amino acid, pyrrolysine, is produced in the same manner, in response to the standard stop codon UAG.

The code used by early life probably did not specify as many as 20 amino acids. Indeed, the more complex ones are produced solely as biologically modified derivatives of simpler ones.

In several bacterial species, for example, the amino acid glutamine is produced from its biochemical cousin glutamate while the latter is still attached to its tRNA. This phenomenon suggests that novel amino acids may have arisen as modifications to a smaller primordial set and that the newcomers "captured" a subset of the tRNAs and codons assigned to their chemically simpler relatives, just as certain codons appear to have been captured by standard amino acids in the organisms known to employ variant codes. These discoveries raise the question of how many more variant codes may be out there and whether the standard code will eventually expand to contain far more amino acids.

—S.J.F. and L.D.H.



MARINE ALGA *Acetabularia* can grow up to five centimeters tall, but each stalk is a single cell—the largest known to science.

lecular biologists turn such theoretical observations about the code into practical tools for genetic engineering, gene finding, and predicting protein shapes. And both of us, along with other researchers, are investigating how the code itself came to be—how RNA first started interacting with amino acids, how their association developed into a system of formal coding and how the amino acid

alphabet expanded during early evolution.

This approach may allow inroads into many additional unresolved questions: Why 20 and only 20 standard amino acids? Why are some amino acids assigned six codons, whereas others have just one or two? Could this pattern have anything to do with minimizing error? Cracking the code has proved merely the start to understanding its meaning. **SA**

MORE TO EXPLORE

A Quantitative Measure of Error Minimization in the Genetic Code. David Haig and Laurence D. Hurst in *Journal of Molecular Evolution*, Vol. 33, No. 5, pages 412–417; November 1991.

The Invention of the Genetic Code. Brian Hayes in *American Scientist*, Vol. 86, No. 1, pages 8–14; January-February 1998.

The Genetic Code Is One in a Million. Stephen J. Freeland and Laurence D. Hurst in *Journal of Molecular Evolution*, Vol. 47, No. 3, pages 238–248; September 1998.

The Case for an Error-Minimizing Genetic Code. Stephen J. Freeland, Nick Keulmann and Tao Wu in *Origins of Life and Evolution of the Biosphere*, Vol. 33, Nos. 4–5, pages 457–477; October 2003.

More information about the evolution of the genetic code can be found on Stephen J. Freeland's laboratory Web site at www.evoldingcode.net