Chapter 8

Chargaff's GC rule

"DNA is in its composition characteristic of the species from which it is derived. This can ... be demonstrated by determining the ratios in which the individual purines and pyrimidines occur ..... There appear to exist two main groups of DNA, namely the ‘AT type,’ in which adenine and thymine predominate, and the ‘GC type,’ in which guanine and cytosine are the major constituents."

Erwin Chargaff (1951) [1]

Evolutionary selective pressures sometimes act to preserve nucleic acid features at the expense of encoded proteins. That this might occur in the case of nucleic acid secondary structure was noted in Chapter 5. That this might also apply to the species-dependent component of the base composition, (G+C)%, was shown by Sueoka in 1961 [2]. The amino acid composition of the proteins of bacteria is influenced, not only by the demands of the environment on the proteins, but also by the (G+C)% of the genome encoding those proteins.

From the genetic code (Table 7-1) it can be inferred that AT-rich genomes tend to have more codons for the amino acids Phe, Tyr, Met, Ile, Asn, and Lys. These amino acids have single-letter abbreviations that collectively spell “FYMINK.” GC-rich genomes tend to have more codons for the amino acids Gly, Ala, Arg and Pro. These amino acids have single-letter abbreviations that collectively spell “GARP.” If you know the FYMINK/GARP ratio of an organism’s proteins, then you can make a good guess as to its (G+C)%.

Sueoka’s observation has since been shown to apply to a wide variety of animal and plant species. Could an organism’s (G+C)% have arisen randomly and yet be so powerful that it could force a gene to encode an amino acid “against its will” (i.e. to encode an amino acid that might not best serve the needs of the encoded protein, and hence, of the organism)? Can a selective force be recognized?
Uniformity of (G+C)%

Chargaff’s “GC rule” is that the ratio of (G+C) to the total bases (A+G+C+T) tends to be constant in a particular species, but varies between species. Sueoka further pointed out that for individual “strains” of Tetrahymena (ciliated protozoans) the (G+C)% (referred to as “GC”) tends to be uniform throughout the genome:

“If one compares the distribution of DNA molecules of Tetrahymena strains of different mean GC contents, it is clear that the difference in mean values is due to a rather uniform difference of GC content in individual molecules. In other words, assuming that strains of Tetrahymena have a common phylogenetic origin, when the GC content of DNA of a particular strain changes, all the molecules undergo increases or decreases of GC pairs in similar amounts. This result is consistent with the idea that the base composition is rather uniform not only among DNA molecules of an organism, but also with respect to different parts of a given molecule.”

Again, this observation has since been shown to apply to a wide variety of species, although many organisms have their genomes finely sectored into regions (“homostability regions” or “isochores”) of low or high (G+C)% (see later). Sueoka also noted a link between (G+C)% and reproductive isolation for strains of Tetrahymena:

“DNA base composition is a reflection of phylogenetic relationship. Furthermore, it is evident that those strains which mate with one another (i.e. strains within the same ‘variety’) have similar base compositions. Thus strains of variety 1 ..., which are freely intercrossed, have similar mean GC content.”

The Holy Grail

It seems that, in identifying (G+C)% as the component of the base composition that varies between species, Chargaff had uncovered what can now be recognized as the “holy grail” of speciation postulated by the Victorian physiologist George Romanes [3]. Romanes had drawn attention to the possibility of what we would now call non-genic variations (germ-line mutations that usually do not affect gene products). As manifest in the phenomenon of hybrid sterility, these would tend to isolate an individual reproductively from most members of the species to which its close ancestors had belonged, but not from individuals that had undergone the same non-genic variation. Romanes held that, in the general case, this isolation was an essential precondi-
tion for the preservation of the anatomical and physiological characteristics (genic characteristics) that were distinctive of a new species.

In the early decades of the twentieth century William Bateson also postulated non-genic inherited variations that tend to remain relatively constant (vary only within narrow limits) within a species, but would vary between species (i.e. a species member would not differ from its fellow species members, but would differ from members of allied species). The non-genic variations, in whatever was responsible for carrying hereditary information from generation to generation (not known at that time), would have the potential to lead to species differentiation, so that variant individuals (constituting a potential "not-self" incipient species) would end up not being able to reproduce with members of the main species ("self" species).

Reproduction being unsuccessful, the main species can be viewed as constituting a "reproductive environment" that moulds the genome phenotype ("reprotype") by negatively selecting (by denying reproductive success to) variant organisms that attempt (by mating and producing healthy, fertile offspring) to recross the emerging interspecies boundary. Thus, the main species positively selects itself by negatively selecting variants. Should these variants find compatible mates, then they might accumulate as a new species that, in turn, would positively select itself by negatively selecting further variants. This is "species selection," a form of group selection that many biologists have found hard to imagine. Indeed, Richard Dawkins, having scorned the "argument from personal incredulity," was obliged to resort to it when confronted with the possibility of species selection: "It is hard to think of reasons why species survivability should be decoupled from the sum of the survivabilities of the individual members of the species" [4].

When the latter sentence is parsed its logic seems impeccable. Hold tight, and we will see if we can work it out. "The species" is the established main species, members of which imperil themselves only marginally, if at all, by mating with (denying reproductive success to) members of a small potentially incipient species. Thus, in reproductive interactions between a main and an incipient species, survivability of the main species is coupled negatively to the sum of the survivabilities of individual members of the incipient species (i.e. it survives when they do not survive), much more than it is coupled positively to the sum of the survivabilities of its own individual members (i.e. it survives when they survive). In this sense, main species survivability is coupled to the sum of the survivabilities of individual members of the incipient species, and decoupled from the sum of the survivabilities of its own individual members.

Of course, by individual survivabilities is meant, not just mere survival, but survival permitting unimpeded production of fertile offspring. Survival of members of an incipient species occurs, not only when classical Darwinian
phenotypic interactions are favourable (e.g. escape from a tiger), but also when reprotoptic interactions are favourable (e.g. no attempted reproduction with members of the main species). Tigers are a phenotypic threat. Members of the main species are a reprotoptic threat [3].

Individual members of a main species that are involved (when there is attempted crossing) in the denial of reproductive success to individual members of an incipient species, are like individual stones in the walls of a species fortress against which the reproductive arrows of an incipient species become blunted and fall to the ground. Alternatively, the main species can be viewed as a Gulliver who barely notices the individual Lilliputian incipients brushed off or trampled in his evolutionary path. Just as individual cells acting in collective phenotypic harmony constitute a Gulliver, so individual members of a species acting in collective reprotoptic harmony constitute a species. That harmony is threatened, not by its own members, but by deviants that, by definition, are no longer members of the main species (since a species is defined as consisting of individuals between which there is no reproductive isolation). These deviants constitute a potential incipient species that might one day pose a phenotypic threat to the main species (i.e. they will become part of the environment of the latter).

It is true that a member of a main species that becomes irretrievably pair-bonded with a member of an incipient species (e.g. pigeons) will leave fewer offspring, so that both members will suffer the same fate (have decreased survivability in terms of number of fertile offspring). But, in the general case, one such infertile reproductive encounter with a member of an incipient species will be followed by many fertile reproductive encounters with fellow members of the main species. Members of the main species are most likely to encounter other members of the main species – hence, there will be fertile offspring. Members of an incipient species, being a minority, are also most likely to encounter members of the main species – hence, there will be infertile (sterile) offspring. Much more rarely, a member of an incipient species will encounter a fellow incipient species member with which it can successfully reproduce – an essential precondition for species divergence.

Once branching (reproductive isolation) is initiated (Fig. 7-4), the natural selection of Darwin should help the branches sprout (extend in length). Natural selection would favour linear species differentiation by allowing the survival of organisms with advantageous genic variations, and disallowing the survival of organisms with disadvantageous genic variations. These genic variations would affect an organism’s form and function (the classical phenotype).

Darwin thought that natural selection might itself suffice to bring about branching. Indeed, it appears to do so in certain circumstances, as when segments of a species have become geographically isolated from each other.
However, here the branching agency is whatever caused the geographical isolation, not natural selection. Speciation requires isolation in some shape or form. The problem of the origin of species is that of determining what form isolation takes in the general case. In his faith in the power of natural selection, Darwin was like the early chemists who were satisfied with atoms as the ultimate basis of matter. But for some chemists phenomena such as swinging compass needles (magnetism), falling apples (gravity), and (later) radioactivity, were manifestations of something more fundamental in chemistry than atoms. Likewise, for some biologists the phenomenon of hybrid sterility seemed to manifest something more fundamental in biology than natural selection [3].

Romanes referred to his holy grail (speciating factor) as an abstract “intrinsic peculiarity” of the reproductive system. Bateson described his as an abstract “residue” with which genes were independently associated. Goldschmidt’s was an abstract chromosomal “pattern” caused by “systemic mutations” that would not necessarily affect genic functions (see Chapter 7). These are just what we might expect of \((G+C)\%\). Indeed, in bacteria, which when so inclined intermittently transfer DNA in a sexual fashion [5], differences in \((G+C)\%\) appear early in the speciation process [6], in keeping with Sueoka’s above observations in ciliates.

As shown in Chapter 3, where different levels of genetic information were considered, a metaphor for the role \((G+C)\%\) might play in keeping individuals reproductively isolated from each other, is their accent [7]. A common language brings people together, and in this way is conducive to sexual reproduction. But languages can vary, first into dialects and then into independent sub-languages. Linguistic differences keep people apart, and this difference in the reproductive environment can militate against sexual reproduction.

At the molecular level, we see similar forces acting at the level of meiosis – the dance of the chromosomes. In the gonad similar paternal and maternal chromosomes (homologues) align. The early microscopists referred to this as “conjugation.” If there is sufficient sequence identity (i.e. the DNA “accents” match), then the band plays on. The chromosomes continue their minuet, progressing through various check-points [8], and gametes are formed. If there is insufficient identity (i.e. the DNA “accents” do not match) then the music stops. Meiosis fails, gametes are not formed, and the child is sterile – a “mule.” Thus, the parents of the child (their “hybrid”) are reproductively isolated from each other (i.e. unable to generate a line of descendants due to hybrid sterility), but not necessarily from other members of their species. At least one of the parents has the potential to be a founding member of a new species, provided it can find a mate with the same DNA “accent.” Differences in \((G+C)\%\) have the potential to initiate the speciation process creating
first “incipient species” with partial reproductive isolation, and then “species” that, by definition, are fully reproductively isolated. To see how this might work, we consider the chemistry of chromosome alignment at meiosis [9].

Crick’s Unpairing Postulate

In 1922 Muller suggested that the pairing of genes as parts of chromosomes undergoing meiotic synapsis in the gonad might provide clues to gene structure and replication [10]:

“It is evident that the very same forces which cause the genes to grow [duplicate] should also cause like genes to attract each other [pair] ... If the two phenomena are thus dependent on a common principle in the make-up of the gene, progress made in the study of one of them should help in the solution of the other.”

In 1954 he set his students an essay “How does the Watson-Crick model account for synapsis?” [11]. The model had the two DNA strands “inward-looking” (i.e. the bases on one strand were paired with the bases on the other strand). Crick took up the challenge in 1971 with his “unpairing postulate” by which the two strands of a DNA duplex would unpair to expose free bases in single-stranded regions [12]. This would allow a search for sequence similarity (homology) between two chromosomes (i.e. between two independent duplexes). Others later proposed that the single-stranded regions would be extruded as stem-loops. The “outward-looking” bases in the loops would be available to initiate the pairing process [13–15].

Thus, for meiotic alignment, maternal and paternal chromosomal homologues should mutually explore each other and test for “self” DNA complementarity, by the “kissing” mechanism noted in Chapter 6 [16–18]. Under this model (Fig. 8-1), the sequences do not commit themselves, by incurring strand-breakage, until a degree of complementary has been recognized. The mechanism is essentially the same as that by which tRNA anticodon loops recognize codons in mRNAs, except that the stem-loop structures first have to be extruded from DNA molecules that would normally be in classical duplex form. In all DNA molecules examined, base-order supports the formation of such secondary structures (see Chapter 5). If sufficient complementarity is found between the sequences of paternal and maternal chromosome homologues (i.e. the genomes are “reprotypically” compatible), then crossing over and recombination can occur (i.e. the “kissing” can be “consummated”).

The main adaptive values of this would be the proper assortment of chromosomes among gametes, and the correction of errors in chromosome sequences (see below and Chapter 14). “Kissing” turns out to be a powerful
metaphor, since it implies an exploratory interaction that may have reproductive consequences.

\[
\begin{align*}
\text{P} & \quad \text{M} \\
(G+C)\% &= X & X \\
\text{PAIRING} & \quad \text{NO PAIRING}
\end{align*}
\]

**Fig. 8-1.** The exquisite sensitivity of stem-loop extrusion from duplex DNA to differences in base composition can prevent the initiation of pairing between homologous DNA sequences. At the left, paternal (P) and maternal (M) duplexes have the same (G+C)\% value (X). As negative supercoiling progressively increases, the strands of each duplex synchronously open to allow formation of equivalent stem-loop secondary structures so that "kissing" interactions between loops can progress to pairing. At the right, paternal and maternal duplexes differ slightly in (G+C)\% (X, and X + 1). The maternal duplex of higher (G+C)\% opens less readily as negative supercoiling increases, so strand opening is not synchronous, "kissing" interactions fail, and there is no progress to pairing. In this model, chromosome pairing occurs before the strand breakage that accompanies recombination (not shown). Even if strand breakage were to occur first (as required by some models), unless inhibited by single-stranded DNA-binding proteins the free single strands so exposed would rapidly adopt stem-loop conformations. So the homology search could still involve kissing interactions between the tips of loops.
The model predicts that, for preventing recombination (i.e. creating reproductive isolation), a non-complementarity between the sequences of potentially pairing strands, in itself, might be less important than a non-complementarity associated with sequence differences that change the pattern of stem-loops. This implies differences in the quantities of members of the Watson-Crick base pairs in single strands (i.e. a parity difference). This is because parity between these bases would be needed for optimum stem formation. Parity differences should correlate with differences in stem formation, and hence, different stem-loop patterns, as will now be considered.

**(G+C)% Controls Pairing**

What role does the (G+C)% “accent” play in meiotic pairing? From calculated DNA secondary structures, it has been inferred that small fluctuations in (G+C)% have great potential to affect the extrusion of stem-loops from duplex DNA molecules and, hence, to affect the pattern of loops which would then appear (Fig. 5-2). A very small difference in (G+C)% (reprotypic difference) would mark as “not-self” a DNA molecule that was attempting to pair meiotically with another DNA (“self”). This would impair the kissing interaction with the DNA [19, 20], and so would disrupt meiosis and allow divergence between the two parental lines, thus initiating a potential speciation event.

The total stem-loop potential in a sequence window can be analysed quantitatively in terms of the relative contributions of base composition and base order, of which base composition plays a major role (see Chapter 5). Of the various factors likely to contribute to the base composition-dependent component of the folding energy of an extruded single stranded DNA sequence, the four simplest are the quantities of the four bases.

Two slightly more complex factors are the individual bases, from each potential Watson-Crick base pair, that are present in lowest amounts. For example, if the quantities of A, G, C and T in a 200 nucleotide sequence window are 60, 70, 30 and 40, respectively, then what may be referred to as “AT_{min}” would be 40, and the corresponding “GC_{min}” would be 30. These numbers would reflect the upper limit on the number of base pairs that could form stems, since the quantity of the Watson-Crick pairing partner that was least would place a limit on the possible number of base pairs. This value might be expected to correlate positively with folding stability.

Conversely, the excess of bases without a potential pairing partner (in the above example A-T = 20 and G-C = 40) might provide an indication of the maximum number of bases available to form loops. Since loops tend to destabilize stem-loop structures, these “Chargaff difference” values might be expected to correlate negatively with folding stability.
Although the bases are held in linear order, a vibrating single-stranded DNA molecule has the potential to adopt many structural conformations, with Watson-Crick interactions occurring between widely separated bases. Accordingly, pairing can also be viewed as if the result of random interactions between free bases in solution. This suggests that the two products of the quantities of pairing bases could be important (60 x 40, and 70 x 30, in the above example). The products would be maximal when pairing bases were in equal proportions in accordance with Chargaff’s second parity rule. Thus, [50 x 50] > [60 x 40] > [70 x 30]). The product values (2500 > 2400 > 2100) might provide an index both of the absolute quantities of the members of a Watson-Crick base pair, and of their relative proportions.

In an attempt to derive formulae permitting prediction of folding energy values directly from the proportions of the four bases, Jih-H. Chen [21] examined the relative importance of eight of the above ten factors in determining the base composition-dependent component of the folding energy (FORS-M; see Chapter 5). These factors were A, G, C, T, AT\text{min}, CG\text{min}, A x T, and G x C (where A, C, G, and T refer to the quantities of each particular base in a sequence window). The products of the quantities of the Watson-Crick pairing bases (A x T, and G x C) were found to be of major importance, with the coefficients of G x C (the strongly interacting S bases), greatly exceeding those of A x T (the weakly interacting W bases). Less important were AT\text{min} and CG\text{min}, and the quantities of the four bases.

All ten parameters were examined in an independent study, which confirmed the major role of the product of the quantities of the S bases in a segment (Table 8-1) [22].

| Number of predictors | Individual bases | Base products | Minimum bases | Chargaff differences | $r^2$ |
|----------------------|------------------|--------------|---------------|---------------------|------|
|                      | A    T    C    G | A x T    G x C | AT\text{min}  | CG\text{min}       | A - T | G - C | |
| 1                    | -     -     -     - | -          | -             | -                   | -    | -    | 0.92 |
| 2                    | -     -     +     - | -          | +             | -                   | -    | -    | 0.96 |
| 3                    | -     -     +     + | -          | +             | -                   | -    | -    | 0.98 |
| 4                    | -     -     +     + | -          | +             | -                   | -    | -    | 0.98 |

Table 8-1. Relative abilities of ten base composition-derived parameters to predict the base composition-dependent component of the folding energy of single-stranded DNA. Best predictors are marked as plus. With 1 predictor, the product of G and C has a correlation of 0.92, as assessed by the $r^2$ value (see Appendix 1). Adding further predictors (rows 2, 3 and 4) does not greatly improve on this value.
Of particular importance is that it is not just the absolute quantities of the S bases, but the product of the multiplication of these absolute quantities. This should amplify very small fluctuations in (G+C)%, and so should have a major impact on the folding energy of a segment and, hence, in the pattern of stem-loops extruded from the duplex DNA in a chromosome engaging in a “kissing” homology search for a homologous chromosome segment.

If stem-loops are of critical importance for the initiation of pairing between segments of nucleic acids at meiosis, then differences in (G+C)% could strongly influence the establishment of meiotic barriers, so leading to speciation. But barriers may be transient. Having served its purpose, an initial barrier may be superseded later in the course of evolution by a more substantial barrier (see Figure 7-4). In this circumstance evidence for the early transient barrier may be difficult to find. However, in the case of different, but related, virus species (allied species) that have the potential to coinfect a common host cell, there is circumstantial evidence that the original (G+C)% barrier has been retained.

**Mutational Meltdown**

Modern retroviruses, such as those causing AIDS (HIV-1) and human T cell leukemia (HTLV-1), probably evolved by divergence from a common ancestral retrovirus. Branching phylogenetic trees linking the sequences of modern retroviruses to such a primitive retroviral “Eve” are readily constructed, using either differences between entire sequences, or just (G+C)% differences [23]. The fewer the differences, the closer are two species on such trees.

Unlike most other virus groups, retroviruses are diploid. As indicated in Chapter 2, diploidy entails a considerable redundancy of information, a luxury that most viruses cannot afford. They need compact genomes that can be rapidly replicated, packaged and dispersed to new hosts. However, different virus groups have evolved different evolutionary strategies.

The strategy of retroviruses is literally to mutate themselves to the threshold of oblivion (“mutational meltdown”), so constituting a constantly moving target that the immune system of the host cannot readily adapt to. To generate mutants, retroviruses replicate their nucleic acids with self-encoded enzymes (polymerases) that do not have the error-correcting (“proof-reading”) function that is found in the corresponding enzymes of their hosts. Indeed, this is the basis of AIDS therapy with AZT (azidothymidine), which is an analogue of one of the nucleotide building blocks that are joined together (polymerized) to form linear nucleic acid molecules (“polymers;” see Chapter 2). AZT is recognized as foreign by host polymerases, which eject it. But retroviral polymerases cannot discriminate, and levels of mutation (in this case termination of the nucleic acid sequence) attain values above the obliv-
ion threshold ("hypermutation") from which it is impossible to recover ("error catastrophe"). Below the threshold, there is a most effective mechanism to counter mutational damage.

The retroviral counter-mutation strategy requires that two complete single-strand retroviral RNA genomes be packaged in each virus particle (i.e. diploidy). Each of these genomes will be severely mutated but, since mutations occur randomly, there is a chance that each genome will have mutations at different sites. Thus, in the next host cell there is the possibility of recombination (cutting and splicing) between the two genomes to generate a new genome with many less, or zero, mutations [24].

The copackaging of the two genomes requires a process analogous to meiotic pairing. On each genome a "dimer initiation" nucleotide sequence folds into a stem-loop structure. "Kissing" interactions between the loops precede the formation of a short length of duplex RNA (Fig. 6-8), so that the two genomes form a dimer. This allows packaging and, in the next host, recombination can occur.

**Coinfecting Viruses**

What if two diploid viruses both infected the same host cell, thus releasing four genomes into an environment conducive to recombination? In many cases this would be a most favorable circumstance, since there would now be four damaged genomes from which to regenerate, by repeated acts of recombination, an ideal genome. Thus, it would seem maladaptive for a virus with this particularly strategy to evolve mechanisms to prevent entry of another virus ("superinfection") into a cell that it was occupying, at least in the early stages of infection [25].

This presupposes that a co-infecting virus will be of the same species as the virus which first gained entry. However, HIV-1 and HTLV-1 are retroviruses of allied, but distinct, species. They have a common host (humans) and common host cell (known as the CD4 T-lymphocyte). When in the course of evolution these two virus species first began to diverge from a common ancestral retroviral species, a barrier to recombination had to develop as a condition of successful divergence. Yet, these two virus types needed to retain a common host cell in which they had to perform similar tasks. This meant that they had to retain similar genes. Many similar gene-encoded functions are indeed found. Similar genes implies similar sequences, and similar sequences implies the possibility of recombination between the two genomes. Thus, coexistence in the same host cell could result in the viruses destroying each other, as distinct species members, by mutually recombining (shuffling their genomes together).

Without a recombination barrier each virus was part of the selective environment of the other. This should have provided a pressure for genomic
changes that, while not interfering with conventional phenotypic functions, would protect against recombination with the other type. If \((G+C)\%\) differences could create such a recombination barrier (while maintaining, through choice of appropriate codons, the abilities to encode similar amino acid sequences), then such differences would be selected for. When we examine the \((G+C)\%\) values of each of these species there is a remarkable difference. HIV-1 is one of the lowest \((G+C)\%\) species known (i.e. it is AT-rich). HTLV-1 is one of the highest \((G+C)\%\) species known (i.e. it is GC-rich). This might be regarded as just a remarkable coincidence save for the fact that, in some other situations where two viruses from different but allied species occupy a common host cell, there are also wide differences in \((G+C)\%\) [3, 19]. As set out above, these \((G+C)\%\) differences alone should suffice to prevent recombination.

### Polyploidy

The plant which gives us tobacco, *Nicotiniana tabacum*, is a tetraploid which emerged some six million years ago when the two diploid genomes of *Nicotiniana sylvestris* and *Nicotiniana tomentosiformis* appeared to fuse. *Nicotiniana tabacum* is designated an allotetraploid (rather than an autotetraploid) since the two genomes were from different source species (Greek: *allos* = other; *autos* = same). The two species are estimated to have diverged from a common ancestral species 75 million years ago. As allied species they should have retained some sequence similarities; so within a common nucleus in the tetraploid there should have been ample opportunity for recombination between the two genomes. Yet, the genomes have retained their separate identities. This can be shown by backcrossing to the parental types. Half the chromosomes of the tetraploid pair at meiosis with chromosomes of one parent type. Thus, recombination of the other chromosomes of the tetraploid with chromosomes of that parent type is in some way prohibited. In 1940 Goldschmidt noted [26]:

"Clausen ... has come to the conclusion that *N. tabacum* is an allotetraploid hybrid, one of the genomes being derived from the species *sylvestris*, the other from *tomentosa*. By continuous backcrossing to *sylvestris* the chromosomes derived from *sylvestris* can be tested because they form tetrads with the *sylvestris*-chromosomes. They have been found to be completely different genetically [from *tomentosa*]. The idea of reaction system [pattern] thus becomes less generalized [less nebulous] and actually applies to the whole architecture of individual chromosomes."

The separateness of the genomes in allotetraploids is reflected in meiosis where, although there are four potentially pairing copies of each chromosome
(four homologous chromosomes), each parental type only pairs with its own type. The two sets of chromosomes differ in a key sequence characteristic, their (G+C)% [27].

**Gene Duplication**

The survival of a duplicate copy of a gene depends on a variety of factors, including (i) natural selection favouring organisms where a function encoded by the gene is either increased or changed (i.e. there is either concerted or divergent gene evolution), (ii) a recombination-dependent process known as gene conversion, and (iii) a recombination-dependent process that can lead to copy-loss (see Fig. 8-2). These intragenomic recombinations can occur when there is a successful search for similarity between DNA strands. This is likely to be greatly influenced by the (G+C)% environment of the original gene and the (G+C)% environment where the duplicate copy locates.

Once a (G+C)%-dependent speciation process has begun, factors other than (G+C)% are likely to replace the original difference in (G+C)% as an intergenomic barrier to reproduction (i.e. a barrier to recombination between diverged paternal and maternal genomes within their hybrid, if such a “mule” can be generated; Fig. 7-4). In this circumstance, (G+C)% becomes free to adopt other roles, such as the prevention of recombination within a genome (intragenomic recombination). This can involve the differentiation of regions of relatively uniform (G+C)%, that Japanese physicists Akiyoshi Wada and Akira Suyama referred to as having a “homostabilizing propensity” and Gior­gio Bernardi and his coworkers named “isochores” (Greek: iso = same; choros = group) [28, 29]. These have the potential to recombinationally isolate different parts of a genome.

Thus, the attempted duplication of an ancestral globin gene to generate the α-globin and β-globin genes of modern primates might have failed since sequence similarity would favour recombination between the two genes and incipient differences (early sequence divergence) could have been eliminated (“gene conversion;” Fig. 8-3). However, the duplication appears to have involved relocation to a different isochore with a different (G+C)%, so the two genes became recombinationally isolated to the extent that initially the sequences flanking the genes differed in (G+C)%. Later the new gene would have increased its recombinational isolation by mutating to acquire the (G+C)% of its host isochore. As a consequence of the differences in (G+C)% the corresponding mRNAs today utilize different codons for corresponding amino acids, even though both mRNAs are translated in the same cell using the same ribosomes and same tRNA populations. So it is most unlikely that the primary pressure to differentiate codons arose at the translational level.
Fig. 8-2 Model for possible outcomes of a gene duplication. The duplication from (a) can result in identical multicopy genes (b) that confer an ability to produce more of the gene product. If this is advantageous, then the multicopy state will tend to be favored by natural selection. If unmutated (white box in (b)) or only slightly mutated (light grey striped box in (c)), there are not sufficient differences between the duplicates to prevent a successful homology search (d). This allows the mutation (c) to be reversed to (b) by the process known as gene conversion (see Fig. 8-3). This maintains identical copies, so allowing concerted evolution of the multicopy genes to continue. However, the recombination necessary for gene conversion can also result in removal of a circular intermediate (e, f), and restoration of the single copy state (g). The risk of copy-loss due to recombination (d-g) can be decreased by further mutation (dark grey striped box in (h)). This will decrease the probability of a successful homology search. Being protected against recombination (i.e. preserved), the duplicate is then free to differentiate further by mutation (black box in (i)). If the product of the new gene confers an advantage, then the duplicate will be further preserved by natural selection (divergent gene evolution). In the general case, mutation facilitating recombinational isolation (h) precedes mutation facilitating functional differentiation (i) under positive Darwinian selection.
Fig. 8-3. Recombination repair and gene conversion. (a) Sequences 3.1 and 3.3 of Chapter 3 are taken as paternal (P) and maternal (M) versions (alleles) of a particular gene, with a single base-pair difference (underlined in the grey
maternal version). (b) Following a successful homology search, exchange of single-strands results in a P segment becoming part of the bottom duplex, and an M segment becoming part of the top duplex, without permanent strand breakage ("paranemic joint"). Because of the single base-pair difference, the recombination complex has an A mismatched with an A (top duplex), and a T mismatched with a T (bottom duplex). (c) Cross-over points can migrate without any permanent strand breakage (shown at left). (d) Strand breakage and resealing (ligation) occur, so that single-strand segments are exchanged between the duplexes (recombination). Without guidance, the "repair" of each mismatch has a 50% chance of restoring the status quo as in (a) (i.e. replacing the A-A non-Watson-Crick base pair with an A-T Watson-Crick base-pair in the top duplex, and replacing the T-T non-Watson-Crick base-pair with a T-A Watson-Crick base pair in the bottom duplex). In the alternative shown here, the status quo is restored to the top duplex (an A is mutated to T), but in the bottom duplex the T-T non-Watson-Crick base-pair is replaced with an A-T Watson-Crick base-pair (i.e. a T is mutated to an A). Thus, there has been conversion of the sequence of the original M allele to that of the P allele. There has been a loss of heterozygosity (as in (a)) and a gain of homozygosity (as in (d)). In this example, gene conversion involves copies of homologous genes (alleles) on different chromosomes. However, gene conversion can also involve homologous genes (non-allelic "paralogues") on the same chromosome (see Fig. 8-2). Note that, in Chapter 4, sequence 3.1 (P above) is shown to form a stem-loop with the central bases being located in the loop (sequence 4.4). Since the single base-pair difference between P and M versions is in this loop, then the M version has the potential to form a similar stem-loop. Because the loops differ slightly, during the initial homology search loop-loop "kissing" interactions might fail and prohibit subsequent steps. However, cross-over points can migrate (e.g. (b) to (c)), so that if crossing over is prohibited in one region there is some possibility of a migration from a neighboring region that would reveal mismatches. Thus, multiple incompatibilities (base differences) are most likely to inhibit the pairing of homologous chromosomes and the repairing of multiple mismatches.

Each isochore would have arisen as a random fluctuation in the base composition of a genomic region such that a copy of a duplicated gene that had transposed to that region was able to survive without recombination with the original gene for a sufficient number of generations to allow differentiation between the copy and its original to occur. This would have provided not only greater recombinational isolation, but also an opportunity for functional differentiation. If the latter differentiation were advantageous, organisms with the copy would be favoured by natural selection. The regional base compositional fluctuation would then have "hitch-hiked" through the generations by virtue of its linkage to the successful duplicate (i.e. the copy would have been positively selected). By preserving the duplicate copy from re-
combination with the original copy, the isochore would, in turn, have itself been preserved by virtue of its linkage to the duplicate copy.

When functional differentiation of a duplicate is necessary for it to be selected (divergent evolution), there is the danger that, before natural selection can operate, recombination-mediated gene conversion will reverse any incipient differentiation, or intragenic recombination between the copies (paralogues) will result in copy-loss. In the case of duplicate eukaryotic genes that have diverged in sequence, Koichi Matsuo and his colleagues noted that divergence was greatest at third codon positions, usually involving a change in (G+C)% [30–33]. Thus, there was a codon bias in favour of the positions of least importance for the functional differentiation that would be necessary for the operation of natural selection. Where amino acids had not changed, different gene copies used different synonymous codons. It was proposed that the (G+C)% change was an important “line of defence” against homologous recombination between the duplicates. Thus, recombinational isolation of the duplicate (largely involving third codon position differences in (G+C)%) would protect (preserve) the duplicate so allowing time for functional differentiation (largely involving first and second codon position differences), and hence, for natural selection to operate. In the general case, isolation would precede functional differentiation, not the converse. (G+C)% differentiation, largely involving third codon positions, would precede functional differentiation, largely involving first and second codon positions under positive Darwinian selection.

From all this it would be predicted that, if a gene from one isochore were transposed to an isochore of different (G+C)%, and its ability to recombine with its allele were advantageous, then the gene would preferentially accept mutations converting its (G+C)% to that of the new host isochore (i.e. organisms with those mutations would be genetically fitter and thus likely to leave more fertile offspring than organisms without the mutations). Indeed, there is evidence supporting this. The sex chromosomes (X and Y) tend not to recombine at meiosis except in a small region (the “pseudoautosomal” region; see Chapter 14). Transfer of a gene from a non-recombining part of a sex chromosome to the pseudoautosomal region forces the gene rapidly to change its (G+C)% value [34].

For various reasons (e.g. large demand for the gene product), certain genes are present in multiple identical copies. But, in the absence of some restraint, copies that are initially identical will inevitably diverge in sequence [3]. So how can multicopy genes (e.g. rRNA genes) preserve their similarity to each other? To prevent divergence through the generations (i.e. to allow “concerted evolution”), they should mutually correct each other to eliminate deviant copies. This is likely to occur by a recombination-dependent process – “gene conversion” (Figs. 8-2, 8-3; see Chapter 10). Thus, multicopy genes
should all be, either in the same isochore, or in isochores of very close \((G+C)\)%, so that recombination can occur.

**Isochores Early**

Before DNA sequencing methods became available, “isochores” were described as DNA segments that could be identified on the basis of their distinct densities in samples of duplex DNA obtained from organisms whose cells had nuclei (eukaryotes). The method involved physically disrupting DNA by hydrodynamic sheering to break it down to lengths of about 300 kilobases. The fragments were then separated as bands of distinct densities by centrifugation in a salt density gradient. The densities could be related to the average \((G+C)\)% values of the segments, since the greater these values, the greater the densities. This way of assessing the \((G+C)\)% of a duplex DNA segment distinguished one large segment from another, and largeness became a defining property of isochores.

Isochores, as so defined, were not identified in bacteria, which do not have distinct nuclear membranes (prokaryotes; see Chapter 10). Since prokaryotes (e.g. bacteria) and eukaryotes (e.g. primates) are considered to have evolved from a common ancestor, does this mean that the ancestor had isochores that were subsequently lost by prokaryotes during or after their divergence from the eukaryote lineage (isochores-early)? Or did the ancestor not have isochores, which were therefore freshly acquired by the eukaryotic lineage after its divergence from the prokaryotic lineage (isochores-late)? If prokaryotes could be shown to have isochores, then this would favour the isochores-early hypothesis.

Indeed, prior to modern sequencing technologies, physical methods demonstrated small segments of distinct \((G+C)\)% in the genomes of prokaryotes and their viruses. The 48 kb duplex genome of phage lambda (see Chapter 5) was extensively sheered to break it down to subgenome-sized fragments. These resolved into six distinct segments, each of relatively uniform \((G+C)\)%, by the density method [35], and into thirty four “gene sized” segments by another, more sensitive, method (thermal denaturation spectrophotometry) [36].

With the advent of sequencing technologies, in 1984 Mervyn Bibb and his colleagues were able to plot the average \((G+C)\)% values of every third base for small windows in the sequences of various bacteria (Fig. 8-4) [37]. Three plots were generated, the first beginning with the first base of the sequence (i.e. bases in frame 1, 4, 7, etc.), the second beginning with the second base of the sequence (i.e. bases in frame 2, 5, 8, etc.), and the third beginning with the third base of the sequence (i.e. bases in frame 3, 6, 9, etc.). In certain small regions \((G+C)\)% values were relatively constant within each frame. These regions of constant \((G+C)\)% corresponded to genes.
Fig. 8-4. The (G+C)% of every third base tends to be both distinct from neighboring bases, and constant, in the region of a gene. Bases were counted in windows of 126 bases (a) or 42 bases (b), in one of three frames, either beginning with base 1 (frame 1, 4, 7, etc.), or with base 2 (frame 2, 5, 8, etc.), or with base 3 (frame 3, 6, 9, etc.). Thus, for calculating average (G+C)% values, each frame takes into account either 42 bases (a) or 14 bases (b). The frames are here named according to the positions of bases in the codons of the gene (encoding the enzyme aminoglycoside phosphotransferase in the GC-rich bacterium *Streptomyces fradiae*). Vertical dashed lines and the grey rectangle indicate the limits of the rightward-transcribed gene. Note that the relative constancy of (G+C)% is most for the third codon position (mainly independent of the encoded amino acids), and least for the second codon position (most dependent on the encoded amino acids). The fluctuation in values at the second codon position is more apparent when a window size equivalent to 14 codons is used (b) than when a window size equivalent to 42 codons is used (a). This figure was redrawn from ref. [37]
Thus, individual genes have a relatively uniform \((G+C)\)% and each codon position makes a distinctive contribution to that uniformity. This is not confined to bacteria. Wada and Suyama noted that, whether prokaryotic or eukaryotic, “every base in a codon seems to work cooperatively towards realizing the gene’s characteristic value of \((G+C)\) content.” This was a “homostabilizing propensity” allowing a gene to maintain a distinct \((G+C)\)% relatively uniform along its length, which would differentiate it from other genes in the same genome [38]. Thus, each gene constitutes a homostabilizing region in DNA.

Stated another way, if large size is excluded as a defining property, many bacteria have isochores. When isochores are defined as DNA segments of relatively uniform \((G+C)\)% that are co-inherited with specific sequences of bases, then bacteria have isochores. To contrast with the classical isochores of Bernardi, these are termed “microisochores,” and their length is that of a gene, or small group of genes (see Chapter 9). Thus, classical eukaryotic isochores (“macroisochores”) can be viewed as constellations of microisochores of a particular \((G+C)\)%.

The proposed antirecombination role of \((G+C)\)% would required that, unless they represent multicopy genes, microisochores sharing a common macroisochores (i.e. they have a common \((G+C)\)% have other sequence differences that are sufficient to prevent recombination between themselves [39]. Within an organism, genes with similar \((G+C)\)% values may sometimes locate to similar tissues, so that there is a tissue-specific codon usage tendency [31].

Since both prokaryotic (e.g. bacterial) and eukaryotic (e.g. primate) lineages have some form of isochore, this appears most consistent with the isochores-early hypothesis. While not endorsing a particular role for \((G+C)\)%, this underlines the fundamental importance of \((G+C)\)% differences in biology. Let metaphors multiply! A given segment of DNA is co-inherited with a “coat” of a particular \((G+C)\)% “color.” A given segment of DNA “speaks” with a particular \((G+C)\)% “accent,” (and hence has a distinct potential vibrational frequency; see Fig. 5-2). A fundamental duality of information levels is again manifest.

**The Gene as a Unit of Recombination**

As will be further considered in Chapter 9, it is likely that differences in \((G+C)\)% serve to isolate recombinationally both genes within a genome, and genomes within a group of species (a taxonomic group). The power to recombine is fundamental to all life forms because, for a variety of reasons, it is advantageous (see Chapter 14). However, the same power threatens to homogenize (blend) genes within a genome, and to homogenize (blend) the genomes of members of allied species within a taxonomic group (i.e. genus). This would countermand evolution both within a species and between spe-
cies. Thus, functional differentiation, be it between genes in a genome, or between genomes in a taxonomic group (speciation), must, in the general case, be preceded (or closely accompanied) by the establishment of recombinational barriers. Species have long been defined in terms of recombinational barriers (see Chapter 7). In some contexts, genes are defined similarly. A species can be defined as a unit of recombination (or rather, of antirecombination with respect to other species). So can a gene.

Most definitions of the “gene” contain a loose or explicit reference to function. Thus, biologists talk of a gene encoding information for tallness in peas. Biochemists talk of the gene encoding information for growth hormone (a protein), and relate this to a segment of DNA (see legend to Fig. 10-1). However, before it can function, information must be preserved. Classical Darwinian theory proposes that function, through natural selection, is itself the preserving agent. Thus, function and preservation go hand-in-hand, but function is more fundamental than preservation. In 1966 biologist George Williams in the USA, an originator of the “selfish gene” concept, seemed to argue the converse when arriving at a new definition.

The function of any multipart entity, which needs more than one part for this function, is usually dependent on its parts not being separated. Preservation can be more fundamental than function. Williams proposed that a gene should be defined entirely by its property of remaining intact as it passes from generation to generation. He identified recombination as a major threat to that intactness. Thus, for Williams, “gene” meant any DNA segment that has the potential to persist for enough generations to serve as a unit for natural selection; this requires that it not be easily disruptable by recombination. The gene is a unit of recombination (or rather, of antirecombination with respect to other genes) [40].

“Socrates’ genes may be with us yet, but not his genotype, because meiosis and recombination destroy genotypes as surely as death. It is only the meiotically dissociated fragments of the genotype that are transmitted in sexual reproduction, and these fragments are further fragmented by meiosis in the next generation. If there is an ultimate indivisible fragment it is, by definition, ‘the gene’ that is treated in the abstract discussions of population genetics. Various kinds of suppression of recombination may cause a major chromosomal segment or even a whole chromosome to be transmitted entire for many generations in certain lines of descent. In such cases the segment, or chromosome, behaves in a way that approximates the population genetics of a single gene. ... I use the term gene to mean ‘that which segregates and recombines with appreciable frequency’. ... A gene is
one of a multitude of meiotically dissociable units that make up the genotypic message."

Despite this, Williams did not invoke any special chromosomal characteristic that might act to facilitate preservation. Pointing to "the now discredited theories of the nineteenth century," and lamenting an opposition that "arises ... not from what reason dictates, but from the limits of what the imagination can accept," his text *Adaptation and Natural Selection* made what seemed a compelling case for "natural selection as the primary or exclusive creative force." No other agency was required. This tendency, which can be infectious, to bolster the scientific with the *ad hominem* in otherwise rational discourse, will be considered in the Epilogue. In contrast, we have here considered intergenomic and intragenomic differences in \((G+C)\)% as an agency, essentially independent of natural selection, which preserves the integrity of species and genes, respectively.

Within a species individual genes differ in their \((G+C)\)% values. Relative positions of genes on the \((G+C)\)% scale are usually preserved through speciation events. If, in an ancestral species, gene \(A\) was of higher \((G+C)\)% than gene \(B\), this relationship has been sustained in the modern species that resulted from divergences within that ancestral species. Accordingly, when the \((G+C)\)% values of the genes of one of the modern species are plotted against the corresponding \((G+C)\)% values of similar (orthologous) genes in the other modern species, the points usually fit a close linear relationship (c.f. Fig. 2-5).

Species with intragenomic isochore differentiation can themselves further differentiate into new species. In this case, a further layer of intergenicomic \((G+C)\)% differentiation would be imposed upon the previous intragenomic differentiation. Again, when a sufficient degree of reproductive isolation had been achieved this initial barrier between species would usually be replaced by other barriers, thus leaving \((G+C)\)% free to continue differentiating in response to intragenomic demands. However, \((G+C)\)% is never entirely free. It can itself be constrained by demands on gene function (i.e. natural selection) that primarily affect first and second codon positions. Furthermore, as we shall see next, in extreme environments, natural selection can make direct demands on \((G+C)\)%, which might then conflict with its role as a recombinational isolator.

**Thermophiles**

There are few environments on this planet where living organisms are not found. Hot springs, oceanic thermal vents, and radioactive discharges of nuclear reactors, all contain living organisms ("extremophiles"). Fortunately, since heat and radiation are convenient ways of achieving sterilization in hospitals, none of these organisms has been found (or genetically engineered
to become) pathogenic (so far). Thermophiles are so-called because they thrive at high temperatures. Proteins purified from thermophiles may show high stability at normal temperatures, a feature that has attracted commercial interest (i.e. they have a long "shelf life"). Hence, the full genomic sequences of many prokaryotic thermophiles (bacteria and archaea) are now available.

Some thermophiles normally live at the temperature of boiling water. Nucleic acids in solution at this temperature soon degrade. So how do nucleic acids survive in thermophiles? The secondary structure of nucleic acids with a high (G+C)% is more stable than that of nucleic acids with a low (G+C)%. This is consistent with Watson-Crick G-C bonds being strong, and A-T or A-U bonds being weak (see Table 2-1). Do thermophiles have high (G+C)% DNA?

In the case of genes corresponding to RNAs whose structure is vital for RNA function, namely rRNAs and tRNAs, the answer is affirmative. Free of coding constraints (i.e. they are not mRNAs), yet required to form part of the precise structure of ribosomes where protein synthesis occurs, genes corresponding to rRNAs appear to have had the flexibility to accept mutations that increase G+C (i.e. organisms that did not accept such mutations perished by natural selection, presumably acting against organisms with less efficient protein synthesis at high temperatures). The G+C content of rRNAs is directly proportional to the normal growth temperature, so that rRNAs of thermophilic prokaryotes are highly enriched in G and C [41–43]. Yet, although optimum growth temperature correlates positively with the G+C content of rRNA (and hence of rRNA genes), optimum growth temperature does not correlate positively with the overall G+C content of genomic DNA, and hence with that of the numerous mRNA populations transcribed from the genes in that DNA (Fig. 8-5a). Instead, optimum growth temperature correlates positively with A+G content (Fig. 8-5b; see Chapter 12) [44].

The finding of no consistent trend towards a high genomic (G+C)% in thermophilic organisms has been interpreted as supporting the "neutralist" argument that variations in genomic (G+C)% are the consequences of mutational biases and are, in themselves, of no adaptive value, at least with respect to maintaining duplex stability [43, 45]. However, the finding is also consistent with the argument that genomic (G+C)% is too important merely to follow the dictates of temperature, since its primary role is related to other more fundamental adaptations.

The stability of duplex DNA at high temperatures can be achieved in ways other than by an increase in G+C content. These include association with small basic peptides (polyamines) and relaxation of torsional strain (supercoiling) [46, 47]. Thus, there is every reason to believe that, whatever their (G+C)% content, thermophiles are able, both to maintain their DNAs in classical duplex structures with Watson-Crick hydrogen-bonding between oppo-
site strands, and to adopt any necessary extruded secondary structures involving intrastrand hydrogen-bonding (i.e. stem-loops). This will be further considered in Chapter 9.

Fig. 8-5. Variation of average (a) (G+C)% and (b) (A+G)% of the genes of prokaryotic species (bacteria and archaea), with the temperature required for optimal growth of each species. The large number of data points corresponding to ambient temperatures (20-30°C) and human blood temperature (37°C) reflect the fact that relatively few thermophiles have been sequenced at this time. Note that, whereas in (a) only 5% of the variation between points can be explained by growth temperature ($r^2 = 0.05$), in (b) 21% can be explained on this basis ($r^2 = 0.21$; see Appendix 1)
Saltum?

Darwin held that biological evolution reflected the accumulation of frequent very small variations, rather than few intermittent large variations. That Nature did not work by means of large jumps was encapsulated in the Latin phrase “Natura non facit saltum.” However, Huxley, while supporting most of Darwin’s teachings, considered it more likely that evolution had proceeded in jumps (“Natura facit saltum”). According to the arguments of this chapter, both are correct. Within some members of a species small variations in the genome phenotype (i.e. in \((G+C)\%\)) accumulate, so that these members become progressively more reproductively isolated from most other members of the species, initially without major changes in the conventional phenotype. As it accrues, reproductive isolation increasingly favors rapid change in the conventional phenotype, often under the influence of natural selection. So, when their appearance is viewed on a geological time scale, new species can seem to “jump” into existence.

The rate increase reflects better preservation of frequent phenotypic micromutations rather than of infrequent phenotypic macromutations (i.e. of “hopeful monsters,” to use Goldschmidt’s unfortunately term). In other words, while there is continuity of variation at the genotype level, as far as speciation is concerned variants (mutant forms) seem to emerge discontinuously at the phenotype level. Being infrequent, and hence unlikely to find a member of the opposite sex with the same change, organisms with macromutations are not the stuff of evolution.

Summary

Single strands extruded from duplex DNA have the potential to form stem-loop structures that, through exploratory loop-loop “kissing” interactions, may be involved in the homology search preceding recombination. The total stem-loop potential in a sequence window can be analyzed quantitatively in terms of the relative contributions of base composition and base order, of which base composition, and particularly the product of the two \(S\) bases \((G \times C)\), plays a major role. Thus, very small differences in \((G+C)\%\) should impair meiotic pairing, resulting in hybrid sterility and the reproductive isolation that can initiate speciation (i.e. because their hybrid is sterile, the parents are, in an evolutionary sense, “reproductively isolated” from each other). In chemical terms, Chargaff’s species-dependent component of base composition, \((G+C)\%\), may be the “holy grail” responsible for reproductive isolation (non-genic) as postulated by Romanes, Bateson and Goldschmidt. Once a speciation process has initiated, other factors (often genic) may replace \((G+C)\%\) as a barrier to reproduction (preventing intergenomic recombination between species). This leaves \((G+C)\%\) free to assume other roles, such as...
preventing intragenomic (e.g. intergenic) recombination within a species. Thus, many organisms have “macroisochores,” defined as long segments of relatively uniform (G+C)% that are coinherited with specific sequences of bases. These may facilitate gene duplication. Indeed, each gene has a “homo-stabilizing propensity” to maintain itself as a “microisochore” of relatively uniform (G+C)%.

Protection against inadvertent recombination afforded by differences in (G+C)% facilitates the duplication both of genes, and of genomes (speciation). George Williams’ definition of a gene as a unit of recombination rather than of function is now seen to have a chemical basis, and many of the old disputes among evolutionists (e.g. saltum versus non-saltum) appear reconcilable.