Chapter 2
Chargaff's First Parity Rule

> *The formation of different languages and of distinct species, and the proofs that both have been developed through a gradual process, are curiously the same. … Languages, like organic beings, can be classed in groups under groups. … A language, like a species, when once extinct, never … reappears. The same language never has two birthplaces … . The survival or preservation of certain favoured words in the struggle for existence is natural selection.*
>
> Charles Darwin (1871) [1]

Consider a creature, whose attributes you are probably familiar with, pausing in a posture similar to that which you are now perhaps adopting:

The cat sat on the mat.
The cat sat on the mat.

You might guess that, since the information is repeated and rhythms, it forms part of some artistic endeavor that we might refer to as poetry. Alternatively, the author may have had little faith in the typesetters and, to make assurance doubly sure, sent the message twice. It is with the latter explanation that we are most concerned.

Error-Detection

Since type-setting errors are usually random and rare, it is likely that, if an error were to occur, it would affect only one of the sentences. Instead of "mat" on the top line you might have seen "hat." Coming across the two parallel sentences for the first time, and knowing that the repetition was deliberate, how would you decide which was the correct sentence? You might read the top sentence first as the 'sense' text, and then read the bottom sentence to check that each of the letters of the alphabet is faithfully matched by its 'sense' equivalent (i.e. "a" is always matched by "a;" "t" is always matched by "t," etc.). You would check that there is *parity* between the two parallel lines.

D.R. Forsdyke, *Evolutionary Bioinformatics*, DOI 10.1007/978-3-319-28755-3_2

Finding that "h" in "hat" in the top line is mismatched with "m" in "mat" in the bottom line, you would know that an error had occurred. But, in the absence of other information you would not be able to decide whether "h" in the top line was correct, or that it should be corrected to "m" based on the information in the bottom line. All you would know was that there had been an error. If for some reason you were forced to decide, you might toss a coin. But, in the absence of other information, if you accepted the coin's guidance there would only be a 50:50 chance of your being correct.

In-Parallel Redundancy

To increase the chance of accurate error-*detection*, and hence of accurate error-*correction*, the author might have repeated the sentence, in-parallel, three times. If two lines contained the word "mat" and only one line the word "hat," then you would prudently choose the former. Your choice would be even more likely to be correct if the author repeated the sentence, in-parallel, four times, and only one line had the word "hat."

All this requires much redundant information, which both takes up space in the medium conveying the message (in this case, the printed page), and imposes extra labor on the reader. For some purposes it might suffice merely to detect that an error had occurred. Having been alerted, you might then be able to consult other sources of information should the need to distinguish between "hat" and "mat" be critical. For example, there are 25 possible alternatives to "h" as the first letter in "hat." Of the resulting words – aat, bat, cat, dat, eat, fat, gat, etc. – several can be excluded because there is no English dictionary equivalent, others can be excluded syntactically (e.g. "eat" is a verb not an object), and others can be excluded contextually (e.g. neighboring text might refer to mat, not to hat).

Thus, there is much to be gained by duplication, but with increasing levels of redundancy (triplication, quadruplication, etc.) the gains are less evident. At face value, this appears to be a strategy adopted by biological systems for accurately transferring information from generation to generation. Genetic messages are sent as duplexes, but with a 'twist' in more than one sense.

DNA Structure

Contrasting with the 26 letter English alphabet, the DNA alphabet has four letters – the bases **A** (adenine), **C** (cytosine), **G** (guanine), and **T** (thymine). Thus, a message in DNA might read:

TACGACGCCGATAGCGTCGTA (2.1)

With duplex in-parallel redundancy, the message could be sent from generation to generation as:

$$\begin{array}{c}\mathbf{TACGACGCCGATAGCGTCGTA}\\ \mathbf{TACGACGCCGATAGCGTCGTA}\end{array} \tag{2.2}$$

We can refer to this as "sense-sense" pairing since, like the cat sentences, both sentences read the same (i.e. **A** is matched with an **A**, and **T** is matched with a **T**, etc.). However, when arriving at their model for the duplex structure of DNA in 1953, James Watson and Francis Crick [2] took into account a 'rule' enunciated by Erwin Chargaff. He and his coworkers had found that bases did not match themselves. They matched other bases. In DNA, base **A** is quantitatively equivalent to base **T**, and base **G** is quantitatively equivalent to base **C**. Chargaff speculated in 1951 that this regularity might be important for DNA structure, noting [3]:

> It is almost impossible to decide at present whether these regularities are entirely fortuitous or whether they reflect the existence in all DNA preparations of certain common structural principles, irrespective of far-reaching differences in their individual composition and the absence of an easily recognizable periodicity.

In 1952 Canadian biochemist Gerard Wyatt went further, suggesting a spiral structure [4]:

> If you have a spiral structure … [it is quite possible to have the bases] sticking out free so that they don't interfere with each other. Then you could have a regular spacing down the backbone of the chain, in spite of the differences in sequence.

Later he added [5]:

> One is tempted to speculate that regular structural association of nucleotides of adenine with those of thymine and those of guanine with those of cytosine … in the DNA molecule requires that they be equal in number.

If the top message were 'sense,' the bottom message could be considered as 'antisense.' The above 'sense' message could then be sent in duplex form as:

$$\begin{array}{ll}\mathbf{TACGACGCCGATAGCGTCGTA} & \text{'}\textit{sense}\text{'}\\ \mathbf{ATGCTGCGGCTATCGCAGCAT} & \text{'}\textit{antisense}\text{'}\end{array} \tag{2.3}$$

Error-detection would still be possible. In this "sense-antisense" error-detection system, errors would be detected when an **A** was matched with **G**, **C** or another **A**, rather than with **T**. Similarly, if **G** was matched with **A**, **T** or another **G**, rather than with **C**, another error would have been detected.

That a base would not match itself was also right for chemical reasons. Just as the letters of the standard alphabet come as either vowels or consonants, so the bases of DNA are either purines (**A** and **G**) or pyrimidines (**C** and **T**; Table 2.1).

Vowels and consonants often match or 'complement' to the extent that vowels separate consonants giving words a structure, which facilitates their pronunciation. Purines are bigger than pyrimidines, and the chemical models that Watson and

Table 2.1 Symbols for groups of DNA bases. When picking symbols for collectivities of bases some logic is attempted. Since purines and pyrimidines both begin with the same letter, the second consonants **R** and **Y** are employed. Watson-Crick base-pairing involves interactions that are either 'weak' (**W**) in the case of **A**-**T** base-pairs, or 'strong' (**S**) in the case of **G**-**C** base-pairs

	R (Purines)	**Y** (Pyrimidines)
W (Weak)	**A** (Adenine)	**T** (Thymine)
S (Strong)	**G** (Guanine)	**C** (Cytosine)

Crick constructed required that a purine always match or 'complement' a pyrimidine. A molecular complex of two purines would be too big. A molecular complex of two pyrimidines would be too small. The solution is that the purine **A** pairs with the pyrimidine **T** and the purine **G** pairs with the pyrimidine **C**. By match we mean an actual structural (i.e. chemical) pairing. Although your eyes can detect that **A** on one line matches **T** on the other, inside our cells it is dark and there are no eyes to see. Matching is something molecules do for themselves by recognizing complementary shapes on their pairing partners, just as a key recognizes the lock with which it 'pairs.'

The key-lock analogy will serve us well here; however, pairing may also require subtleties such as similar molecular vibrations, or resonances [6]. To visualize this, in 1941 the geneticist Herman Muller likened molecular mixtures to imaginary mixtures of floating electromagnets each charged with an alternating current of a particular frequency. Since magnet polarity would depend on the direction of current flow, the polarity of each magnet would be constantly changing at a frequency determined by the frequency of the alternating current [7]:

> If we had a heterogenous mixture of artificial electromagnets, floating freely about and having different frequencies of reversal of sign, those of the same frequency would be found eventually to orient towards and attract one another, specifically seeking each other out to the exclusion of others.

Of course, the final twist of Watson and Crick was, literally, a twist. The two sequences in DNA are two molecular strands that are wound round each other to form a spatially compact helix (Fig. 2.1). Perhaps the most famous throwaway line ever written came at the end of Watson and Cricks' first paper [2]. Here, as an apparent afterthought, they casually noted: "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material." In other words, they were claiming not only to have discovered the fundamental structure of genetic information, but also to have discerned from that structure how the information would be faithfully replicated. When the underlying *chemistry* was understood, the *physiology* could be explained – a triumph for the 'reductionist' approach. They had shown how the chemical structure of DNA provided a basis for the continuity of inherited characteristics from organism

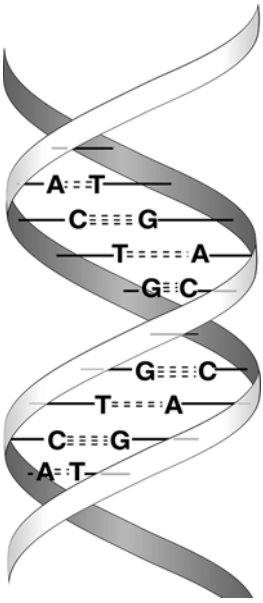

Fig. 2.1 Double helix model for DNA. The base 'message' is written on two helical strands, which are shown here as twisted ribbons – the 'medium.' Bases are arranged internally so that an **A** on one strand pairs with a **T** on the other strand (and vice versa), and a **G** on one strand pairs with a **C** on the other strand (and vice versa). The bases are attached to the strands by strong bonds, whereas the base-pairing interactions involve weak bonds (shown as dashed lines). Chemically, the bases are like flat discs that 'stack' on top of each other within the helical strands, like a pile of coins (rouleau). These stacking interactions stabilize the double-helical structure and, being largely entropy-driven (see Chapter 15), become greater as temperature increases. However, in solution at high temperatures (e.g. 80°C) this can be overcome, and the two strands separate (i.e. the duplex 'melts') to generate free single strands. This figure was kindly adapted by Richard Sinden from his book *DNA Structure and Function* [8]

to organism, and from cell to cell within an organism. In Bateson's words, they had discovered how "the allotment of characteristics among offspring is … accomplished." This was made explicit in a second paper [9]:

> Previous discussions of self-duplication [of genetic information] have usually involved the concept of a template or mould. Either the template was supposed to copy itself directly or it was to produce a 'negative,' which in its turn was to act as a template and produce the original 'positive' once again. … Now our model for deoxyribonucleic acid is, in effect, a *pair* of templates [Watson and Cricks' italics], each of which is complementary to the other. We imagine that prior to duplication … the two chains unwind and separate. Each chain then acts as a template for the formation on to itself of a new companion chain, so that eventually we shall have two pairs of chains, where we only had one before.

Armed with this powerful clue, within a decade biochemists such as Arthur Kornberg in the USA had shown Watson and Crick to be correct, and had identified key enzymes (e.g. DNA polymerase) that catalyze DNA replication [10]. The stunning novelty of the Watson-Crick model was not only that it was beautiful, but

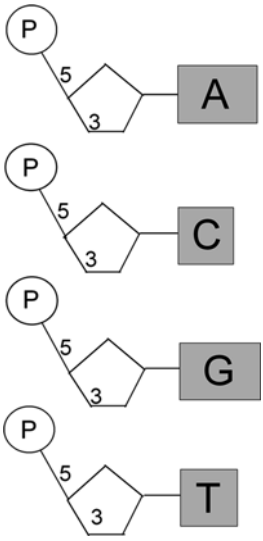

Fig. 2.2 The four nucleotide 'building blocks' of which DNA is composed. Each base is connected by a pentose sugar (pentagon) to a phosphate (circle). The purine bases (**A** and **G**) are shown as larger boxes than the pyrimidine bases (**T** and **C**). Nucleotides have in common a pentose sugar and a phosphate, and differ in their bases. The pentose sugar in DNA is deoxyribose (hence "deoxyribonucleic acid" = "DNA"). The sugar in RNA is ribose (hence "ribonucleic acid" = "RNA"). Pentose sugar carbon atoms are numbered to indicate the third and fifth. In RNA **T** is replaced by **U** (uracil) which, like **T**, pairs with **A**

that it also explained so much of the biology of heredity. One strand is the complement of the other, so that the text of one strand can be inferred from the text of the other. If there is an error in one strand, then there is the potential for its repair on the basis of the text of the opposite strand. When the cell divides the two strands separate. New 'child' strands are synthesized from nucleotide 'building blocks' corresponding to **A**, **C**, **G** and **T**. Each of these blocks, consisting of phosphate, ribose and a base (Fig. 2.2), replaces the former pairing partners of the separated strands, so that two new duplexes identical to the parental duplex are created. In each duplex one of the parental strands is conserved, being paired with a freshly synthesized child strand (Fig. 2.3).

All nucleotide 'building blocks' have in common phosphate and ribose, which continue the phosphate-ribose 'medium,' upon which the base 'message' or 'pattern' is 'written.' Thus, *any* nucleotide can serve to ensure continuity of the phosphate-ribose medium, and the message itself is determined only by which particular base-containing nucleotide is placed in a particular position. This, in turn, is determined by the complementary template provided by the parental DNA strands, which are recognized according to the specific base-pairing rules (Fig. 2.4).

The message you are now reading was imposed by the stepwise sequential addition of letters to a *pre-existing* medium (the paper). Each letter required a small local piece of the medium, but that medium was *already* in place when the letter

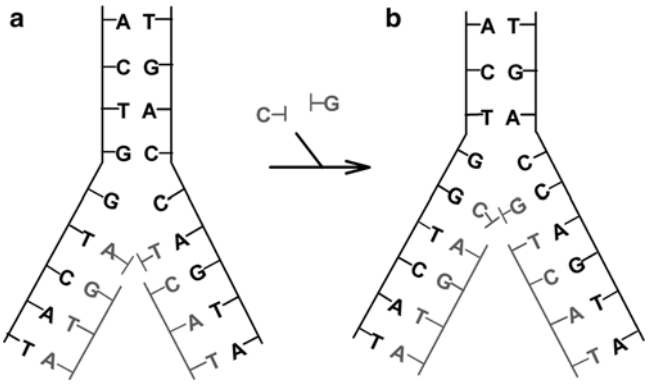

Fig. 2.3 DNA replication. Individual strands of a parental duplex partially separate, and fresh child strands are synthesized by sequential introduction and 'zipping' up (polymerization) of complementary base nucleotide 'building blocks' (shown in gray). Thus, DNA is a linear polymer (Greek: *poly* = many and *meros* = part) of nucleotide units (i.e. it is a polynucleotide). In (a), at the point of child strand growth in the left limb of the replication fork (inverted Y), an **A** (gray) is about to be joined to a **G** (gray). This join is complete in (b), where the two parental strands are further separated. The new duplexes each contain one parental strand (black), and one child strand (gray). Details of synthesis in the right limb are dealt with in Chapter 6 (Fig. 6.6)

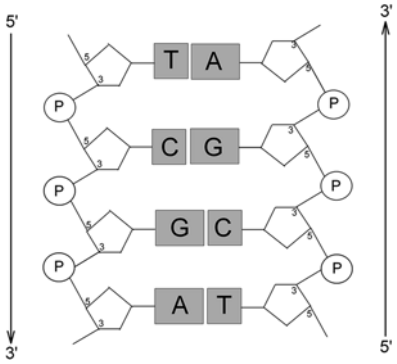

Fig. 2.4 Base-pairing between the two strands of a DNA duplex. The larger purines pair with the smaller pyrimidines, so the distance between the two strands remains relatively constant. Because of this size difference, the flat bases do not just 'stack' (form a 'pile of coins') above and below neighboring bases in the same strand (e.g. note that the two **G**'s on separate strands overlap each other, and thus partially stack together). Rather, base-pairs 'stack' with base-pairs, some better than others. Numbering associated with the pentose sugars indicates that strands have distinct directionality (polarity) that, by convention, is written from 5′ to 3′ (see vertical arrows). Thus, the left strand reads 5′**TCGA**3′ from top to bottom. The right strand reads 5′**TCGA**3′ from bottom to top. The two strands are described as "antiparallel." Since this short duplex *as a whole* has symmetry (i.e. putting the purines first, the order is **A**-**T** base pair, **G**-**C** base pair, **G**-**C** base pair, **A**-**T** base pair), then it can be said to show palindromic properties (see Chapter 4)

arrived. The medium had already been generated. When DNA is synthesized each base 'letter' arrives in a pre-existing association with a small piece of the medium (phosphate-ribose) that it locally requires. Thus, the message and the medium are generated *at the same time*. The message and the medium increase in length simultaneously. Remarkably, all this had been sensed by Muller, who had mentored Watson in the 1940s. In 1936, while attempting to defend Russian genetics against Lysenko (see Epilogue), he distinguished less and more variable parts of gene structure, the latter comprising its "specific pattern" [11]:

> The gene is, as it were, a modeller, and forms an image, a copy of itself, next to itself, and since all genes in the chain do likewise, a duplicate chain is produced next to each original chain, and no doubt lying in contact with a certain face of the latter. … There are thousands of different levels of genes, i.e. of genes having different patterns, … and … each of these genes has to reproduce its own specific pattern out of the surrounding materials common to them all. When, through some microchemical accident, or chance quantum absorption, a sudden change in the composition ('pattern') of the gene takes place, known to biologists as a 'mutation,' then the gene of the new type, so produced, reproduces itself according to this new type, i.e. it now produces precisely the new pattern.
>
> This shows that the copying property depends upon some more fundamental feature of gene structure [phosphate-ribose chain to the modern reader] than does the specific pattern which the gene has [base sequence to the modern reader], and that it is the effect of the former to cause a copying not only of itself but also of the latter, more variable, features. It is this fact which gives the possibility of biological evolution and which has allowed living matter ultimately to become so very much more highly organized than non-living. It is this which lies at the bottom of … growth, reproduction, and heredity.

As we shall see (Chapter 7), the "possibility of biological evolution" occurs because, although mutations are often repaired, sometimes they are not. A change in "specific pattern" is then passed on, by copying, to the next generation. When considering pairs of bases, care should be taken to distinguish between: (i) a Watson-Crick base-pair (i.e. two bases on separate strands, or separate parts of a strand, which are involved in the classical **A**-**T** and **G**-**C** pairings), (ii) a dinucleotide consisting of two ordered contiguous bases on the same strand (e.g. **CG** often written as **C**p**G**; see Chapter 18), and (iii) the base composition of a nucleic acid segment (e.g. (**G**+**C**)%; see Chapter 10).

Turnover and Channeling

The "microchemical accident," to which Muller referred might have a definite cause (e.g. Muller himself had noted increased mutations following X-irradiation), or might loosely be described as 'spontaneous.' The accident might result in one regular letter being substituted for another (e.g. "hat" rather than "mat"), or a regular letter might change to something else (e.g. "πat" rather than "mat"), or simply be eliminated (e.g. "at" rather than "mat"). As will be discussed later (under the heading "entropy;" Chapter 15), it seems to be a general property of the universe that the

elements that compose it, whatever their size, tend to become disordered and evenly distributed. This is true at the chemical level where macromolecules tend to break down to their micromolecular building blocks, and the building blocks themselves, either separate or when they are part of macromolecules, live under a constant threat of structural change and dismemberment into their constituent atoms.

Photographers sometimes want to photograph a busy city scene but without the people, traffic and parked cars. The solution is to use time-lapse photography. A fixed camera takes a picture once a day with a very short exposure time. The film is not wound on, so daily pictures are superimposed. The first picture, if developed, would show nothing (because of short exposure). However, over weeks and months static objects begin to appear, whereas the transient objects are never present long enough to register. Since macromolecules tend to be transient, a magic time-lapse camera that could see individual molecules in bodies would tend to register nothing – except for molecules of DNA (and a few structural proteins like collagen). From this crude metaphor one should not deduce that DNA molecules are static. Even buildings vibrate and move within their confines. So do DNA molecules.

Two cell strategies for dealing with the constant breakdown of its parts are *recycling* (so that macromolecules are degraded and then resynthesized from their component parts), and *repair*. The former strategy (turnover) applies mainly to four of the five major classes of macromolecules (lipids, carbohydrates, proteins and RNA). The latter strategy applies mainly to the fifth class, DNA. Thus, whereas a damaged amino acid in a protein (a polymer of amino acid units) leads to the protein being degraded by specific enzymes (proteases) to its constituent amino acids, a damaged nucleotide in a DNA molecule (a polymer of nucleotide units) often invokes a 'rapid response team' of repair enzymes that will do its best to effect on-site repair without necessarily interrupting macromolecular continuity.

Synthesis of macromolecules from component parts is something cells do well because the assembly lines (biochemical pathways) for making the components are well established. DNA polymerase does not have to stand idle, waiting for a suitable nucleotide to turn up. However, the nucleotides that supply the needs of the cell for DNA synthesis can also supply the needs of a foreign invader – a virus. Ideally, nucleotides would be carefully dispensed to match the cell's needs, but would be kept from the predators. Indeed, another cell strategy is to *channel* the nucleotides to the site of DNA synthesis [12]. Thus, the enzymes, both for the balanced synthesis of the four component parts, and for their incorporation into DNA, can exist as a large multi-enzyme aggregate close to the replication fork (Fig. 2.3).

Promiscuous DNA

Sometimes there is a break in a DNA duplex. The two ends may be reconnected by various enzymes ("ligases"). However, the tendency towards disorder sometimes means that a DNA segment is incorrectly reconnected. A random 'cut' followed by

a 'paste' may result in one segment of DNA recombining with a new segment of DNA so that the order of the information they contain is changed (transposed or inverted). To the extent that such changes are not critical for survival, genomes are vulnerable to an on-going kaleidoscopic diversification – a constant shuffling – of the sequences they contain.

More than this, DNA molecules are promiscuous – meaning, literally, that DNA molecules are "pro-mixing." Place two duplex DNA molecules within a common cell wall and they will seek each other and attempt to recombine. We shall see that biological evolution became possible when DNA 'learned,' by adjusting sequence and structure, how to constrain and channel this tendency. Often the order of information in DNA is critical. Specific segments of DNA have specific 'addresses' in the chromosomes that contain them. The ability to accurately recombine specific segments of duplex DNA, while maintaining segment order and the integrity of functional units, is a fundamental property of living organisms. Indeed, US biologist George Williams, one of those responsible for our modern 'selfish gene' concept, thought it better to define genes in terms of their abilities to resist dismemberment by recombination, than in terms of their functions (see Chapter 11). The great evolutionary significance of recombination was pointed out by Crick in 1970 [13]:

> There is also a major problem to which I believe biologists have given insufficient attention. All biologists essentially believe that evolution is driven by natural selection, but … it has yet to be adequately established that the rate of evolution can be … explained by the processes which are familiar to us. It would not surprise me if nature has evolved rather special and ingenious mechanisms so that evolution can proceed at an extremely rapid rate – recombination is an obvious example.

A year later Crick presented his "unpairing postulate" to explain how the inward-looking bases in a DNA double-helix might look *outward* to recognize complementary bases in another helix (see Chapter 10).

Haploidy and Diploidy

So breath-taking was Watson and Crick's model that some potentially major criticisms were overlooked. If every line of the present book were repeated, after the fashion of the cat sentences at the beginning of this chapter, then the book would be twice as long as it now is. Not only does it make sense to minimize the duplication of information in books, but there are also circumstances where it would appear advantageous not to duplicate information in biological systems. Despite this, *duplication is the rule*. For example, one of the two forms of gamete, usually the male spermatozoon, has to be highly mobile and hence has a streamlined shape and, tadpole-like, is often equipped with a flagellum. There appears to have been a selection pressure to keep the quantity of contained information (i.e. DNA) to a minimum (Fig. 1.3). Virus genomes, which have to be packaged for transfer from organism to organism, are also very compact. Yet, the DNA of spermatozoa and viruses is always in duplex form (with a few special exceptions).

Many organisms alternate during their life cycle between haploidy (one copy of each chromosome, containing one DNA duplex, per cell) and diploidy (two copies of each chromosome per cell). Gametes are haploid and so contain only one copy of each DNA duplex. When male and female gametes unite, the product (zygote) is diploid with two copies of each DNA duplex-containing chromosome, one from the father and one from the mother. Some organisms, such as the malaria parasite *Plasmodium falciparum*, quickly switch back to the haploid state, so its adult form is haploid. But for many organisms, diploidy is the adult norm. Only when new gametes are formed is there a brief flirtation with haploidy.

Thus, there is redundancy of information not only because DNA molecules come as duplexes, but also because many organisms 'choose' for most of their life cycles to have two copies of each duplex. Since each duplex has *at least* two-fold redundancy, diploid organisms have *at least* four-fold redundancy in their content of DNA. Why "*at least*"? There is only at least four-fold redundancy because we have so far considered only *in-parallel* redundancy. The phenomenon of *in-series* redundancy was discovered when measurements were made of the rates at which duplexes would reform from single strands when in solution in test tubes.

In-Series Redundancy

In the 1950s it became possible to synthesize artificial single-stranded RNA sequences such as **UUUUUUUUUUUU** referred to as poly(r**U**), and **AAAAAAAAAAAA** referred to as poly(r**A**). The single strands when mixed together (e.g. poly(r**U**) + poly(r**A**)) formed a double-stranded hybrid, which had a helical structure similar to that of double-stranded DNA. Omitting the helix, this can be represented as:

$$\begin{matrix}\mathbf{UUUUUUUUUUUU}\\ \mathbf{AAAAAAAAAAAA}\end{matrix} \tag{2.4}$$

At the time it appeared amazing that hybridization could occur in a simple salt solution at room temperature in the absence of enzymes. The biologist Julian Huxley (grandson of Thomas Huxley) announced the discovery of "molecular sex" [14]. Whether said in jest, or from profound insight, the description fits perfectly (see Chapter 10).

What was going on in the privacy of the test-tube when millions of flexible, snake-like, poly(r**U**) molecules were mixed with millions of flexible, snake-like, poly(r**A**) molecules? Following the Watson-Crick base pairing rules, molecules of poly(r**U**) react only weakly with each other (since **U** pairs weakly with **U**). Furthermore, there is little inclination for the molecules to fold back on themselves, permitting internal pairing of **U** with **U**. The same applies for poly(r**A**). So there was nothing left but for **A**s to pair with **U**s (analogous to **A**-**T** pairing in DNA). Since the molecules had little internal secondary structure (no folding back on themselves), it

was easy for a writhing chain of **U**s to find a writhing chain of **A**s. Millions of relatively rigid, duplex molecules resulted. Their formation could be monitored by changes either in light absorption or in the viscosity of the solution.

Hybridization was studied similarly with natural nucleic acid duplexes that had been randomly fragmented into smaller duplexes (each about 1200 base pairs in length). From knowledge of the length of an original unfragmented DNA duplex and the number of such duplexes in a solution it was possible to calculate how rapidly the sheered duplex fragments should reform after their two strands had been separated from each other by heating. Like separated partners on a dance floor, to reform, each single-strand would have to find its complement. If there were just one original DNA duplex present, then each single strand would have no option but to find its *original* complementary partner. If two identical duplexes were present it would not matter if a strand found a partner from the other duplex (i.e. it would switch dancing partners). However, in this case there would be twice the chance of finding a partner in a given space and time, compared with when only one duplex was present. Thus, the more identical DNA duplexes present, the more rapidly should the strands reform duplexes (anneal) after heating.

When the experiment was carried out, it was found that for many DNA samples the rate of duplex reformation was far *greater* than anticipated [15]. This was particularly apparent in the case of species with very long DNA molecules. Further studies showed that within DNA there is a redundancy due to the presence of repetitive elements. There are many more copies of certain segments of DNA than the four expected from in-parallel considerations. Molecular 'dancing partners' may be found *in-series* as well as in-parallel. Why is there so much sequence redundancy? Could it all be beneficial (see Chapter 15 for discussion of "junk DNA")? We note below, that 'high flyers' long ago found they could manage quite nicely, thank you, without some of their repetitive elements.

Bits and Bats

There is a link with information theory. Since there are two main types of bases, purines (**R**) and pyrimidines (**Y**), then, disregarding the phosphate-ribose medium upon which the base message is written, a nucleic acid can be represented as a binary string such as:

YRYRRYRYYRRYRRYRYYRYR (2.5)

Electronic computers work with information in this form – represented as strings of 0s and 1s. If a **Y** and an **R** are equally likely alternatives in a sequence position, then each can be measured as one "bit" (binary digit) of information, corresponding to a simple yes/no answer.

Confronted with a generic base (often expressed as **N**) you could first ask if it was a purine (**R**). A negative reply would allow you to infer that **N** was a pyrimidine (**Y**).

You could then ask if it was cytosine (**C**). A positive reply would allow you to infer that **N** was not thymine (**T**). By this criterion, each position in a DNA sequence corresponds to two potential yes/no answers, or two "bits" of information. So the entire single-strand information content in the human haploid genome (3×10^9 bases) is 750 megabytes (since 8 bits make a byte). This is of the order of the amount of information in an audio compact disk.

This way of evaluating DNA information has been explored [16], but so far has not been particularly illuminating with respect to DNA function. One reason for this may be that DNA is not just a binary string. In the natural duplex form of DNA, a base in one string pairs with its complementary base in another string. Each base is 'worth' 2 bits, so that a base pair would correspond to 4 bits. However, even if not paired the two bases would still collectively correspond to 4 bits. Thus, the chemical pairing of bases increases their collective information content to some value greater than 4 bits. But does this come at a price?

At another level (literally and otherwise) consider flying organisms – bats, birds, insects. In every case we find duplex DNA in cells. Every cell of all multicellular organisms has duplex DNA, and flying organisms are no exceptions. Bats have 5.4 picograms of DNA per cell [17], whereas equivalent mammals (mice) have 7 picograms per cell. Bats appear to have shed some of the 'excess' DNA, but that which remains is still in *duplex* form. Birds have approximately 2.5 picograms of DNA/cell. A bird that could shed half its DNA and exist with single stranded DNA would seem to have a weight advantage compare with a bird that had duplex DNA. It should be able to fly faster and further than those with duplex DNA, a feature of particular importance for migratory birds. But again, the DNA is always in duplex form.

Nevertheless, relative to humans, birds seem to have shed (or to have not acquired) 'excess' DNA. Where did this excess originate? Humans and chickens have similar numbers of genes, but the average chicken chromosome is more crowded – 1 gene/40 kilobases compared with 1 gene/83 kilobases [18]. This suggests that birds have shed the DNA between genes. But the average chicken gene is half the size of the average human gene – 27 kilobases compared with 57 kilobases. So birds have also shed some DNA that is deemed "genic." Yet chicken proteins are the *same* size as human proteins. This suggests that birds have not shed the protein-encoding parts of genes (exons; see Chapter 13). Instead, they have shed non-protein-encoding parts of genes (introns).

In Figure 2.5 the lengths of exons and introns in some corresponding genes of chickens and humans are compared. In the case of exon lengths the slope is 1.0; so, on average, each chicken exon is the *same* size as the corresponding human exon. This is in keeping with bird proteins being the same size as human proteins. But in the case of intron lengths the slope is 0.4. The dashed lines indicate that, on average, a large 8 unit human intron would correspond to a large 6.5 unit chicken intron, and a small 4.5 unit human intron would correspond to a small 5.2 unit chicken intron. There is a big difference in the case of large introns, and a much smaller difference in the case of small introns (the scale is logarithmic). Large human introns are much bigger than the corresponding large chicken introns.

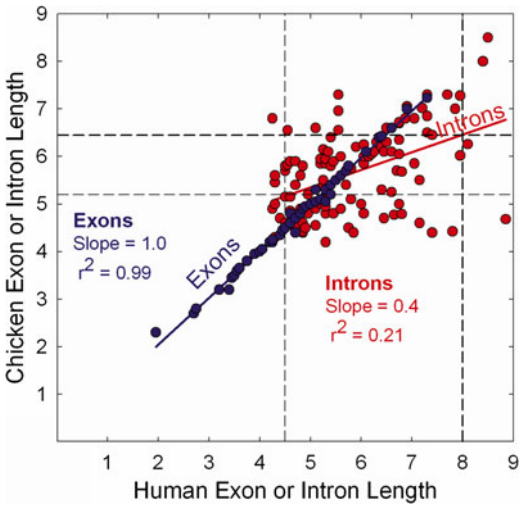

Fig. 2.5 Comparison of the lengths of exons and introns in corresponding genes of humans and chickens. The lines are drawn to best fit the points (see Appendix A). Both slopes are significantly different from zero. Note that, in general, exons (blue) are smaller than introns (red; see Chapter 13), and the points for exons fit more closely to their line (SEE = 0.11; $r^2 = 0.99$) than the points for introns fit to their line (SEE = 0.80; $r^2 = 0.21$). The lesser scattering of exon length values from their line, indicates that the exons in the set of genes studied have been under strong negative selection (i.e. individuals with exon mutations have tended not to survive). Thus, since the time when humans and chickens diverged from a common ancestor, exon sequences have been less successful at varying than intron sequences (i.e. individuals with intron mutations have survived more often than individuals with exon mutations; see Chapters 7 and 8). Dashed lines facilitate comparison between small and large introns (see text). This figure is adapted from reference [19]

When we examine the sequences we find that humans have an excess of repetitive elements (see Chapter 15). These lie both *between* genes (in intergenic DNA) and *within* genes (in introns).

It seems more likely that birds have *discarded* some repetitive elements, than that humans have acquired repetitive elements [19, 20]. When there is less DNA, the nucleus – the 'brain' of the cell – is smaller. So bird nuclei are smaller than human nuclei. And there are similar trade-offs at the level of the whole organism. We should not be surprised to find that migratory birds have smaller brains than non-migratory. Indeed, the further the distance migrated, the smaller the brain [21, 22]. Would the demands on cognition and memory be less in migratory birds? Or perhaps "bigger is not always better" where brains are concerned (see Part VII).

From all this it seems that there are compelling reasons for keeping DNA in duplex form at all stages of life. In biological systems there are conflicts and there have to be trade-offs. But abandoning the duplication of DNA information is seldom one of them.

Accidents and Non-Accidents

In our lives we encounter two classes of adverse events – random and non-random. The non-paranoid designate random adverse events as "accidents." There is conflict between the forces of disorder and the forces of order, the former usually being not deliberately hostile, but merely reflecting the tendency of things, if left alone, to become untidy rather than tidy. This tendency gets greater when things move faster, which usually means they get hotter, as will be considered at the molecular level in Chapter 15.

In this chapter we have considered error-generation as driven by random processes ("microchemical accidents"), and sequence redundancy as having arisen to permit error-detection, and so, possibly, error-correction. Redundancy means that the *qualitative* characteristics (e.g. sequence) of an organism's *own* DNA molecules can be compared, so allowing *quality control*. By mechanisms to be touched on in Chapter 6, the total quantity of DNA in a cell is maintained relatively constant. This is *quantity control*. Since the quantity of DNA determines the 'dose' of gene-products (e.g. proteins) that a cell contains (see Chapter 17), this implies that the quantity of cellular macromolecules can be regulated, directly or indirectly, by DNA quantity-control mechanisms.

Sometimes the forces of disorder have an appreciable non-random component, as when a virus (i.e. foreign DNA; V) deliberately enters a cell. The repertoire of 'self' macromolecules (M) then is supplemented by sets of 'not-self' macromolecules (VM). So the total quantity of cellular macromolecules (TM) can be written:

$$M + VM = TM \tag{2.6}$$

Under normal circumstances, the quantity of macromolecules of virus origin (VM) would be zero. As will be seen in Chapters 13, there are sophisticated host strategies for distinguishing 'self' from 'not-self,' and to be successful (i.e. to increase VM) a virus must outsmart them. The closer the virus, in its *qualitative* characteristics, can approach to self (i.e. become 'near-self' with respect to its host) the more likely it is to succeed. Host quality-control mechanisms, are then likely to be less effective. There is, however, the theoretical possibility of using *quantitative* characteristics of viruses (i.e. VM itself) as a basis for distinction by the host. The available strategies for organisms to respond *internally* to non-random adversities, in the forms of viruses, are somewhat similar to the available strategies for countries to respond *internally* to non-random adversities, in the forms of forgers of their currencies. The metaphor may be helpful.

The aim of a forger is to fool you with counterfeit currency. If successful the forger prospers, but if too successful there is the possibility that the entire monetary system would collapse. This would not serve the forger well and, so far as we know, no forger has gone to this extreme. Nevertheless, the counterfeit notes must be as like the real thing as the forger can contrive. At the qualitative level, your visual and tactile sensory mechanisms for distinguishing real notes from counterfeit notes

must be evaded. Accordingly, manufacturers of a country's true currency are engaged in an 'arms race' with the illegitimate manufacturers of false currency. As forgers get progressively better at counterfeiting currency that approaches progressively closer to the real thing, so the manufacturers of true currency must add embellishments that are difficult for forgers to imitate. This allows you to continue to make *qualitative* distinctions on a note-by-note basis.

At the level of the entire currency system, however, it should *in theory* be possible to detect that forged notes (FN) are present without looking at individual notes. Designating the quantity of real notes as N, we can write:

$$N + FN = TN \tag{2.7}$$

TN represents the total quantity of notes. Here is how it would work. Given knowledge of the initial quantity of real notes, and their rates of manufacture and of destruction when worn-out, then it should be possible to know how many real notes (N) exist. If there were a way of directly monitoring how many notes actually existed at a particular time-point (e.g. knowing the average 'concentration' of notes and the area over which they were distributed), then the actual number (TN) could be compared with the calculated number (N). If the actual number exceeded the calculated number, then the presence of forged notes (FN) would be inferred, alarm bells would ring, and appropriate corrective measures implemented. In principle, if the system were sufficiently sensitive, a small initial increase in forged notes would be immediately responded to. A forger would have difficulty opposing this form of monitoring. But, of course, in practice such monitoring is difficult for countries to implement.

Biological organisms are not so constrained. In general, 'self' molecules are manufactured at rates that have been fine-tuned over millions of years of evolution. Similarly, rates of destruction have been fine-tuned. Accordingly, the concentrations of many molecules, *especially proteins*, fluctuate between *relatively narrow limits*. Intrusive foreign 'not-self' macromolecules would tend to increase total macromolecule concentrations in ways that, in principle, should be detectable. This theme will be explored in Parts V and VI.

Summary

Most cell components undergo cycles of degradation and resynthesis ("turnover"), yet their concentrations fluctuate between only very narrow limits. The DNA of a cell provides information that specifies both the quality and quantity of these components. Accurate transmission of this information requires that errors in DNA be detected and corrected. If there is more than one copy of the information (redundancy) then one copy can be compared with another. For hereditary transmission of information, a 'message' is 'written' as a sequence of four base 'letters' – **A**, **C**, **G**, **T** – on a strand of phosphate and ribose (the 'medium'). In duplex DNA there is

two-fold redundancy – the 'top' strand is the complement of the 'bottom' strand. **A** on one strand matches **T** on the other, and **G** on one strand matches **C** on the other. A check for non-complementarity permits error-detection. Thus, Chargaff's first parity rule is that, for samples of duplex DNA, the quantity of **A** (adenine) equals the quantity of **T** (thymine), and the quantity of **G** (guanine) equals the quantity of **C** (cytosine). In diploid organisms there is four-fold *in parallel* sequence redundancy, due to the presence of a DNA duplex (chromosome) of maternal origin, and a DNA duplex (chromosome) of paternal origin. There is also some *in-series*, within-strand, redundancy. Trade-offs to optimize utilization of sequence space do not include abandonment of duplex DNA or diploidy. Birds lighten their DNA load by decreasing the number of repetitive sequences that would be located either between or within genes. DNA is promiscuous in readily acquiescing to a 'cutting-and-pasting' (recombination between and within strands) that shuffles the information it contains. Indeed, George Williams thought it better to define genes in terms of their abilities to resist dismemberment by recombination, than in terms of their functions. Furthermore, a codiscoverer of DNA structure, Francis Crick, questioned the potency of the natural selection of functional differences as an evolutionary force, and pointed to possible "ingenious mechanisms" involving recombination that might accelerate evolutionary processes.

References

1. Darwin C (1871) *The Descent of Man, and Selection in Relation to Sex*. Appleton, New York, pp 57–59
2. Watson JD, Crick FHC (1953) Molecular structure of nucleic acids. A structure for deoxyribose nucleic acid. *Nature* 171:737–738
3. Chargaff E (1951) Structure and function of nucleic acids as cell constituents. *Federation Proceedings* 10:654–659
4. Wyatt GR (1952) Specificity in the composition of nucleic acids. *Experimental Cell Research*, Supplement 2:201–217
5. Wyatt GR, Cohen SS (1953) The bases of the nucleic acids of some bacterial and animal viruses. *Biochemical Journal* 55:774–782
6. Israelachvili J, Wennerstrom H (1996) Role of hydration and water structure in biological and colloidal interactions. *Nature* 379:219–224
7. Muller HJ (1941) Resumé and perspectives of the symposium on genes and chromosomes. *Cold Spring Harbor Laboratory Symposium on Quantitative Biology* 9: 290–308
8. Sinden RR (1994) *DNA Structure and Function*. Academic Press, San Diego
9. Watson JD, Crick FHC (1953) Genetical implications of the structure of deoxyribonucleic acid. *Nature* 171:964–967
10. Kornberg A (1989) *For the Love of Enzymes: the Odyssey of a Biochemist*. Harvard University Press, Cambridge, MA
11. Muller HJ (1936) The needs of physics in the attack on the fundamental problems of genetics. *Scientific Monthly* 44:210–214
12. Forsdyke DR, Scott FW (1979) Exogenous purine deoxyribonucleosides do not prevent inhibition of DNA synthesis by hydroxyurea: evidence for nonconvergence of *de novo* and salvage pathways. In: Lynen F, Mothes K, Nover L (eds) *Cell Compartmentation and Metabolic Channelling*. Elsevier, Amsterdam, pp 177–184 [Viruses can sometimes escape dependence

on host *de novo* biosynthetic pathways by encoding their own enzymes. In turn, when alarmed, some hosts can activate enzymes that degrade the free nucleotides needed by viral polymerases; Yan N, Lieberman J (2012) SAMHD1 does it again, now in resting T cells. *Nature Medicine* 18:1611–1612]
13. Crick F (1970) Molecular biology in the year 2000. *Nature* 228:613–615
14. Forsdyke DR (2007) Molecular sex: the importance of base composition rather than homology when nucleic acids hybridize. *Journal of Theoretical Biology* 249:325–330
15. Waring M, Britten RJ (1966) Nucleotide sequence repetition: a rapidly reassociating fraction of mouse DNA. *Science* 154:791–794
16. Gatlin LL (1972) *Information Theory and Living Systems*. Columbia University Press, New York [In 'the classical period of molecular biology' the often misleading input of cybernetics and mathematicians tended to discourage a fuller exploration of informational aspects of biology.]
17. Burton DW, Bickham JW, Genoways HH (1989) Flow-cytometric analyses of nuclear DNA contents in four families of neotropical bats. *Evolution* 43:756–765
18. Bellott DW et al. (2010) Convergent evolution of chicken Z and human X chromosomes by expansion and gene acquisition. *Nature* 466:612–616
19. Hughes AL, Hughes MK (1995) Small genomes for better flyers. *Nature* 377:391 [In 2013 the Animal Genome Size Database maintained by Ryan Gregory had 475 avian records; the mean genome size was 1.39 gigabases (Gb) and 90% of the genomes fell between 1.0 and 1.6 Gb. The corresponding values for 657 mammalian species were 3.36 Gb with 90% between 2.2 and 4.9 Gb.]
20. Organ CL, Shedlock AM, Meade A, Pagel M, Edwards SV (2007) Origin of avian genome size and structure in non-avian dinosaurs. *Nature* 446: 180–184 [Mammalian red blood cells (Fig. 16.1) have no nuclei. It would seem advantageous, regarding body weight, that birds should be similar. Yet, avian red blood cells are nucleated.]
21. Safi K, Seid MA, Dechmann DKN (2005) Bigger is not always better: when brains get smaller. *Biology Letters* 1:283–286
22. Sol D, Garcia N, Iwaniuk A, Davis K, Meade A, Boyle WA, Szekely T (2010) Evolutionary divergence in brain size between migratory and resident birds. *PLOS One* 5:e9617

http://www.springer.com/978-3-319-28753-9

Evolutionary Bioinformatics

Forsdyke, D.R.

2016, XXXIV, 471 p. 116 illus., 24 illus. in color.,
Hardcover
ISBN: 978-3-319-28753-9