Single-base Mutation

Dan Graur, *Tel Aviv University, Tel Aviv, Israel*

DNA sequences are normally copied exactly during the process of chromosome replication;
however, new sequences are formed if errors in either DNA replication or repair occur.
These errors are called mutations.

Deoxyribonucleic acid (DNA) sequences are normally copied exactly during the process of chromosome replication. Sometimes, however, errors in either DNA replication or DNA repair occur, giving rise to new sequences. These errors are called mutations.

Mutations can occur in either somatic or germ-line cells. Somatic mutations are not inherited, so they can be disregarded in an evolutionary or genetic context. Germ-line mutations are the ultimate source of variation and novelty in evolution. Some organisms (e.g. plants) do not have a sequestered germ line, and therefore the distinction between somatic and germ-line mutations is not absolute.

Mutations may be classified by the length of the DNA sequence affected by the mutational event. If the mutational event affects two or more adjacent nucleotides, then we refer to the event as a *segmental mutation*. Here, we will only deal with mutations affecting a single nucleotide (*single-base* or *point mutations*). Specifically, we deal with *substitution mutations*, that is, the replacement of one nucleotide by another.

Substitution mutations are divided into *transitions* and *transversions*. Transitions are substitution mutations between A and G (purines) or between C and T (pyrimidines). Transversions are substitution mutations between a purine and a pyrimidine. There are four types of transitions (A→G, G→A, C→T and T→C) and eight types of transversions (A→C, A→T, C→A, C→G, T→A, T→C, G→C and G→T).

Substitution mutations occurring in protein-coding regions may also be classified according to their effect on the product of translation, the protein. Because of the degeneracy of all genetic codes, not all single-base mutations affect the sequence of the protein. A substitution mutation is defined as *synonymous* if it changes a codon into another that specifies the same amino acid as the original codon (**Figure 1**). Otherwise, it is *nonsynonymous*. A change in an amino acid due to a nonsynonymous mutation is called a *replacement*. The terms 'synonymous' and '*silent*' mutation are often used interchangeably because, in the great majority of cases, synonymous mutations do not alter the amino acid sequence of a protein and are therefore not detectable at the amino acid level. However, a synonymous mutation may not always be silent. A synonymous mutation may, for instance, create a new splicing site or obliterate an existing one, thus turning an exonic sequence into an intron or vice versa, and causing a different polypeptide to be produced. For example, a synonymous change from the glycine codon GGT to its synonymous codon GGA in codon 25 of the first exon of β globin has been shown to create a new splice junction, resulting in the

Figure 1 Types of substitution mutations in a coding region: (a) synonymous, (b) missense and (c) nonsense. (Reproduced from Graur and Li (2000).).
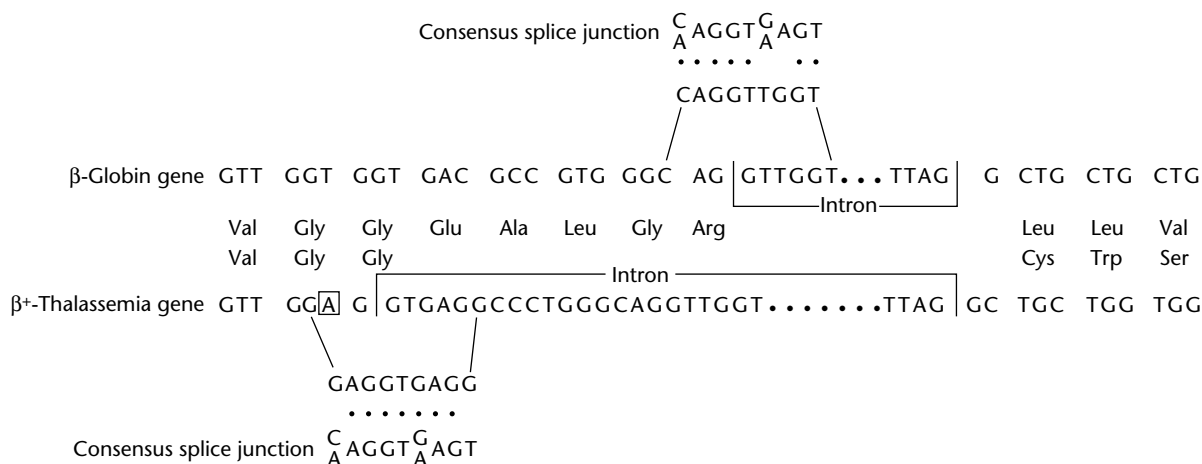
Figure 2 Nucleotide sequences at the borders between exon-1 and intron-I and exon-2 in the β-globin gene from a normal individual and a patient with β⁺-thalassemia, the mutated nucleotide boxed. The leader lines indicate the splicing sites. Each of the splice junctions is compared with the sequence of the consensus splice junction, and dots denote identity of nucleotides between the splice junction and the consensus sequence. Note that the nucleotide substitution in the β⁺-thalassemia gene is synonymous, because both GGT and GGA code for the amino acid glycine. It is not, however, silent, because the activation of the new splicing site in the β⁺-thalassemia gene results in the production of a frameshifted protein. (Reproduced from Graur and Li (2000) after Goldsmith *et al.* (1983).)

production of a frameshifted protein of abnormal length (**Figure 2**). Such a mutation is obviously not 'silent'. Therefore, it is advisable to distinguish between the two terms. Of course, all silent mutations in a protein-coding gene are synonymous.

Nonsynonymous or *amino acid-altering* mutations are further classified into *missense* and *nonsense* mutations (**Figure 1**). A missense mutation changes the affected codon into a codon that specifies a different amino acid from the one previously encoded. A nonsense mutation changes a sense codon into a termination codon, thus prematurely ending the translation process and ultimately resulting in the production of a truncated protein. Codons that can mutate to a termination codon by a single substitution mutation, for example, UGC (Tyr), which can mutate in one step into either UAG or UGA, are called *pretermination codons*.

Readthrough mutations, that is, mutations turning a stop codon into an amino acid-specifying condon, are also known to occur. Such mutations cause the translation to continue beyond the original termination codon until the translation apparatus encounters a new in-frame stop codon. If such a codon does not exist downstream of the mutated stop codon, the translation will continue up to the polyadenylation site.

Each of the sense codons can mutate to nine other codons by means of a single substitution mutation. For example, CCU (Pro) can experience six nonsynonymous changes, to UCU (Ser), ACU (Thr), GCU (Ala), CUU (Leu), CAU (His) or CGU (Arg), and three synonymous changes, to CCC, CCA or CCG. Since the universal genetic code consists of 61

Table 1 Relative frequencies of different types of substitution mutations in a random protein-coding sequence[a]

Mutation	Number	Percentage
Total in all codons	549	100
Synonymous	134	25
Nonsynonymous	415	75
Missense	392	71
Nonsense	23	4
Total in first codon	183	100
Synonymous	8	4
Nonsynonymous	175	96
Missense	166	91
Nonsense	9	5
Total in second codon	183	100
Synonymous	0	0
Nonsynonymous	183	100
Missense	176	96
Nonsense	7	4
Total in third codon	183	100
Synonymous	126	69
Nonsynonymous	57	31
Missense	50	27
Nonsense	7	4

[a]Reproduced from Graur and Li, 2000.

sense codons, there are $61 \times 9 = 549$ possible substitution mutations. If we assume that all possible substitution mutations occur with equal frequencies, and that all codons are equally frequent in coding regions, we can compute the expected proportion of the different types of substitution mutations from the genetic code. These are shown in **Table 1**. Of course, not all mutations occur with equal frequencies, nor is the frequency distribution of codons uniform. However, the results in **Table 1** give us an indication on the

buffering capacities of the genetic code against nonsynonymous and nonsense mutations. Because of the structure of the genetic code, synonymous mutations occur mainly at the third position of codons. Indeed, almost 70% of all the possible nucleotide changes at the third position are synonymous. In contrast, all the substitution mutations at the second position of codons are nonsynonymous, as are most nucleotide changes at the first position (96%).

In the vast majority of cases, the exchange of a codon by a synonym, that is, one that codes for the same amino acid, requires only one or at most two synonymous mutations. The only exception to this rule is the exchange of a serine codon belonging to the four-codon family (UCU, UCC, UCA and UCG) by one belonging to the two-codon family (AGU and AGC). Such an event requires two nonsynonymous mutations.

Nucleotide substitution mutations are thought to arise mainly from the mispairing of bases during DNA replication. As part of their formulation of DNA replication, Watson and Crick (1953) suggested that transitions might be due to the formation of purine–pyrimidine mispairs (e.g. A : C), in which one of the bases assumes an unfavored tautomeric form, that is, enol instead of keto in the case of guanine and thymine, or imino instead of amino in the case of adenine and cytosine (**Figure 3**). Topal and Fresco (1976) proposed that purine–purine mispairs can also occur, but pyrimidine–pyrimidine mispairs cannot. The purine–pyrimidine mispairs are $A^* : C$, $A : C^*$, $G^* : T$ and $G : T^*$ and the purine–purine mispairs are $A^* : A$, $A^* : G$, $G^* : A$ and $G^* : G$, in which the asterisk denotes an unfavored tautomeric form. The pathways via which substitution mutations arise are as follows: (1) transitions arise from purine–pyrimidine mispairing and can occur on either strand. For example, the transition A : T→G : C can arise from one of four possible mispairs: $A^* : C$, $A : C^*$, $G : T^*$ or $G^* : T$; (2) transversions arise from purine–purine mispairing but can only occur if the purine resides on the template strand. For instance, the transversion A : T↔T : A can arise only from $A^* : A$, where the unfavored tautomer A^* is on the template strand.

Because of the rarity of mutations, the rate of spontaneous mutation is very difficult to determine directly, and at present only a few such estimates exist at the DNA sequence level (e.g. Drake *et al.*, 1998). The rate of mutation, however, can be estimated indirectly by other means. Li *et al.* (1985) and Kondrashov and Crow (1993), for example, estimated the average rate of mutation in mammalian nuclear DNA to be $3-5 \times 10^{-9}$ substitution mutations per nucleotide site per year. The mutation rate, however, varies enormously with genomic region, and in microsatellites, for instance, the rate in humans was
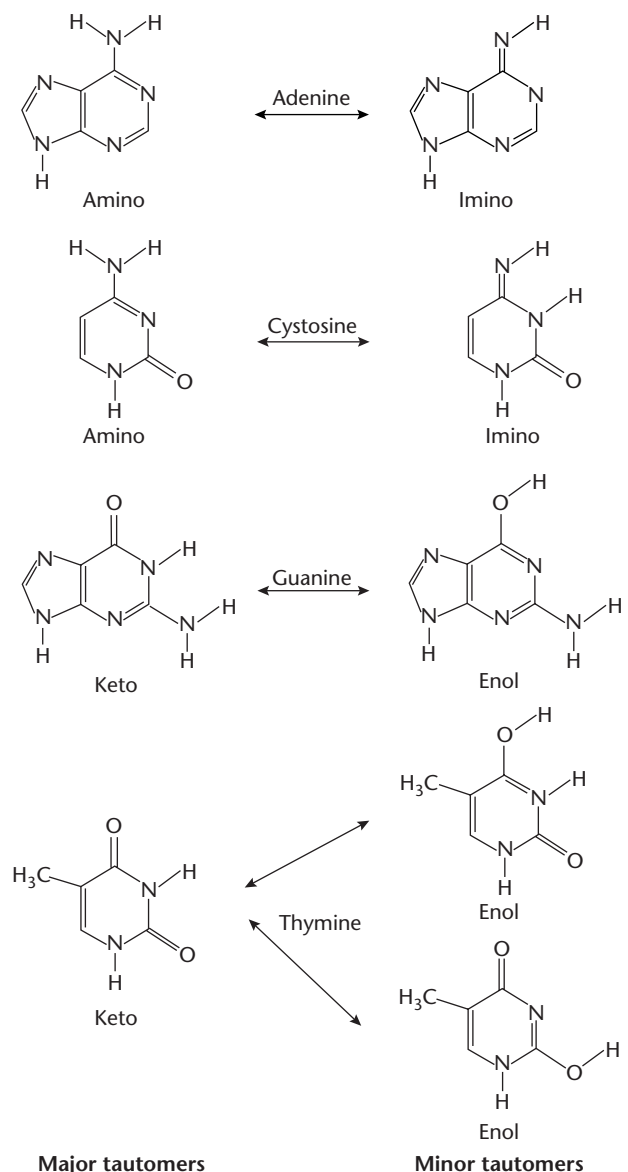

Figure 3 Amin↔imino and ket↔enol tautomerisms. Adenine and cytosine are usually found in the amino form, but rarely assume the imino configuration. Guanine and thymine are usually found in the keto form, but rarely form the enol configuration. Thymine has two enol tautomers. All minor tautomers can assume different rotational forms. (Reproduced from Graur and Li (2000).)

estimated to be over 10^{-3} substitution mutations per nucleotide site per year (Brinkmann *et al.*, 1998).

The rate of mutation in mammalian mitochondria has been estimated to be at least 10 times higher than the average nuclear rate (Brown *et al.*, 1982). Ribonucleic acid viruses have error-prone polymerases (i.e. ones lacking proofreading), and they lack an efficient postreplication repair mechanism for mutational damage. Thus, their rate of mutation is several orders

of magnitude higher. Gojobori and Yokoyama (1985), for example, estimated the rates of mutation in the influenza A virus and the Moloney murine sarcoma virus to be of the order of 10^{-2} substitution mutations per nucleotide site per year, that is, approximately 2 million times higher than the rate of mutation in the nuclear DNA of vertebrates. The rate of mutation in the Rous sarcoma virus may be even higher, as nine mutations were detected out of 65 250 replicated nucleotides, that is, a rate of 1.4×10^{-4} substitution mutations per nucleotide site per replication cycle (Leider et al., 1988).

Mutations do not occur randomly throughout the genome. Some regions are more prone to mutate than others, these being called *hot spots* of mutation. One such hot spot is the dinucleotide 5′-CG-3′ (often denoted as CpG), in which the cytosine is frequently methylated in many animal genomes, and is changed to 5′-TG-3′. The dinucleotide 5′-TT-3′ is a hot spot of mutation in prokaryotes but usually not in eukaryotes. In bacteria, regions within the DNA containing short *palindromes* (i.e. sequences that read the same on the complementary strand, such as 5′-GCCGGC-3′, 5′-GGCGCC-3′ and 5′-GGGCCC-3′) were found to be more prone to mutation than other regions.

The direction of mutation is nonrandom. In particular, transitions were found to occur more frequently than transversions. In animal nuclear DNA, transitions were found to account for about 60–70% of all mutations, whereas the proportion of transitions under random mutation is expected to be only 33%. Thus, in nuclear genomes, transitional mutations occur twice as frequently as transversions. In animal mitochondrial genomes, the ratio of transitions to transversions is about 20 transitions to 1 transversion. Some nucleotides are more mutable than others. For example, in the nuclear DNA of mammals, G and C tend to mutate more frequently than A and T.

Mutations are commonly said to occur 'randomly'. However, as we have seen previously, mutations do not occur at random with respect to genomic location, nor do all types of mutations occur with equal frequency. So, what aspect of mutation is random? Mutations are claimed to be random in respect to their effect on the fitness of the organism carrying them. That is, any given mutation is expected to occur with the same frequency under conditions in which this mutation confers an advantage on the organism carrying it, as under conditions in which this mutation confers no advantage or is deleterious. 'It may seem a deplorable imperfection of nature,' said Dobzhansky (1970), 'that mutability is not restricted to changes that enhance the adeptness of their carriers'. Indeed, the issue of whether mutations are random or not with respect to their effects on fitness is periodically debated in the literature, sometimes with fierce intensity (see e.g. Hall, 1990; Lenski and Mittler, 1993; Rosenberg et al., 1994; Sniegowski, 1995).

See also

Mutation Detection
Mutation Nomenclature
Mutation Rate
Mutational Change in Evolution
Mutations in Human Genetic Disease

References

Brinkmann B, Klintschar M, Neuhuber F, Huhne J and Rolf B (1998) Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *American Journal of Human Genetics* **62**: 1408–1415.

Brown WM, Prager EM, Wang A and Wilson AC (1982) Mitochondrial DNA sequences of primates: tempo and mode of evolution. *Journal of Molecular Evolution* **18**: 225–239.

Dobzhansky T (1970) *Genetics and the Evolutionary Process*. New York, NY: Columbia University Press.

Drake JW, Charlesworth B, Charlesworth D and Crow JF (1998) Rates of spontaneous mutation. *Genetics* **148**: 1667–1686.

Gojobori T and Yokoyama S (1985) Rates of evolution of the retroviral oncogene of *Moloney murine sarcoma virus* and of its cellular homologues. *Proceedings of the National Academy of Sciences of the United States of America* **82**: 4198–4201.

Goldsmith ME, Humphries RK, Ley T, et al. (1983) 'Silent' nucleotide substitution in a β^+-thalassemia globin gene activates splice site in coding sequence RNA. *Proceedings of the National Academy of Sciences of the United States of America* **80**: 2318–2322.

Graur D and Li W-H (2000) *Fundamentals of Molecular Evolution*, 2nd edn. Sunderland, MA: Sinauer Associates.

Hall BG (1990) Directed evolution of a bacterial operon. *BioEssays* **12**: 551–557.

Kondrashov AS and Crow JF (1993) A molecular approach to estimating human deleterious mutation rate. *Human Mutation* **2**: 229–234.

Leider JM, Palese P and Smith FI (1988) Determination of the mutation rate of a retrovirus. *Journal of Virology* **62**: 3084–3091.

Lenski RE and Mittler JE (1993) The directed mutation controversy and neo-Darwinism. *Science* **259**: 188–194.

Li W-H, Luo C-C and Wu C-I (1985) Evolution of DNA sequences. In: MacIntyre RJ (ed.) *Molecular Evolutionary Genetics*, pp. 1–94. New York, NY: Plenum.

Rosenberg SM, Longerich S, Gee P and Harris RS (1994) Adaptive mutation by deletions in small mononucleotide repeats. *Science* **265**: 405–409.

Sniegowski PD (1995) The origin of adaptive mutants: random or nonrandom? *Journal of Molecular Evolution* **40**: 94–101.

Topal MD and Fresco JR (1976) Complementary base pairing and the origin of substitution mutations. *Nature* **263**: 285–289.

Watson JD and Crick FHC (1953) Genetical implications of the structure of deoxyribonucleic acid. *Nature* **171**: 964–967.

Further Reading

Graur D and Li W-H (2000) *Fundamentals of Molecular Evolution*, 2nd edn. Sunderland, MA: Sinauer Associates.

Griffiths AJF, Miller JH and Suzuki DT (1996) *An Introduction to Genetic Analysis*, 6th edn. New York, NY: Freeman.

Portin P (1993) The concept of the gene: short history and present status. *Quarterly Review of Biology* **68**: 173–223.

Sinden RR, Pearson CE, Potaman VN and Ussery DW (1998) DNA: structure and function. *Advances in Genome Biology* **5A**: 1–141.