# Relationships Between Genomic G+C Content, RNA Secondary Structures, and Optimal Growth Temperature in Prokaryotes

**Nicolas Galtier, J.R. Lobry**

CNRS UMR 5558-Laboratoire BGBP, Université Claude Bernard, 43 Bd. du 11-NOV-1918, F-69622 Villeurbanne cedex, France

**Abstract.** G:C pairs are more stable than A:T pairs because they have an additional hydrogen bond. This has led to many studies on the correlation between the guanine+cytosine (G+C) content of nucleic acids and temperature over the last 20 years. We collected the optimal growth temperatures ($T_{opt}$) and the G+C contents of genomic DNA; 23S, 16S, and 5S ribosomal RNAs; and transfer RNAs for 764 prokaryotic species. No correlation was found between genomic G+C content and $T_{opt}$, but there were striking correlations between the G+C content of ribosomal and transfer RNA stems and $T_{opt}$. Two explanations have been proposed—neutral evolution and selection pressure—for the approximate equalities of G and C (respectively, A and T) contents within each strand of DNA molecules. Our results do not support the notion that selection pressure induces complementary oligonucleotides in close proximity and therefore numerous secondary structures in prokaryotic DNA, as the genomic G+C content does not behave in the same way as that of folded RNA with respect to optimal growth temperature.

**Key words:** Prokaryotes — Optimal growth temperature — G+C content — Secondary structures — Parity rule — Directional mutation pressure

## Introduction

The Watson-Crick (1953) base-pairing rules imply exact equimolar amounts of adenine (A) and thymine (T) and exact equimolar amounts of cytosine (C) and guanine (G) in a double-stranded DNA molecule. Somewhat surprisingly, these relationships are also approximately true for single-stranded DNA molecules. These statistical relationships in single-stranded DNA, sometimes called Chargaff's rules (Chargaff 1979), were termed parity rule type 2 (PR2) by Sueoka (1995). They were first reported for the chemical composition of *Bacillus subtilis* DNA (Rudner et al. 1968; Karkas et al. 1968). Similar results were obtained for *Escherichia coli, B. megaterium, B. stearothermophilus, Proteus vulgaris, Salmonella typhimurium, Serratia marcescens,* and *E. coli* phage T4 (Rudner et al. 1969; Karkas et al. 1970; Rudner and LeDoux 1974). When long DNA sequences became available, PR2 was also found to hold for the eukaryotic SV40, polyoma, and Bk viruses (Nussinov 1982); for *Homo sapiens* (Fickett et al. 1992); and for many other species (Prabhu 1993; Lobry 1995).

Two distinct interpretations of PR2 have been suggested. The first is that PR2 is the consequence of directional mutation pressure (Sueoka 1962, 1988, 1992, 1993; Freese 1962). Under no-strand-bias conditions, when both DNA strands behave similarly for mutation and selection, the substitution pattern is such that PR2 should hold at equilibrium (Sueoka 1995; Lobry 1995). Any deviations from PR2 would therefore be the result of deviation from the ideal no-strand-bias conditions, and these deviations are indeed correlated with the replica-

*Correspondence to:* J.R. Lobry

tion origin in *E. coli, B. subtilis, Haemophilus influenzae* (Lobry 1996a), and *Mycoplasma genitalium* (Lobry 1996b). The second interpretation is that PR2 is the result of a ''selection pressure favoring mutations that generate complementary oligonucleotides in close proximity, thus creating a potential to form stem-loops'' (Forsdyke 1995a). According to this hypothesis, deviations from PR2 are deviations from the ideal case, where the whole genome is involved in secondary structures.

This paper examines these two hypothesis by analyzing the effect of temperature on the G+C content of bacterial genomes. Since bacteria do not regulate their temperature, their optimal growth temperature ($T_{opt}$) is a good indication of the temperature their DNA endures. Since there are three hydrogen bonds in G:C pairs and two in A:T pairs, G:C pairs should stabilize secondary structures at elevated temperatures (Wada and Suyama 1986). We have looked for this effect in molecules known to be involved in secondary structures (tRNAs, 5S rRNAs, and the stems of 16S and 23S rRNAs) and in the whole genome. If the second selectionist hypothesis is correct, a high proportion of the genomes should be involved in forming secondary structure, so genomic G+C content should follow the same pattern as that of the folded RNA G+C content.

## Materials and Methods

*Source of $T_{opt}$ and Genomic G+C Content Data.* The main source of data was *Bergey's Manual* (Staley et al. 1984), plus additional data for archaeal hyperthermophiles compiled by Dalgaard and Garrett (1993). The data set includes 764 prokaryotic (eubacterial and archaebacterial) species from 224 genera. The number of available species per genus varied greatly, from one to 64 (mean: 3.4 species per genus, distribution skewed to the right). The mean $T_{opt}$ and genomic G+C content for all the species in a given genus were computed to minimize this sampling imbalance.

*Optimal Growth Temperature ($T_{opt}$).* Care was taken to use data in which the reported temperatures unambiguously referred to the optimal growth temperature rather than to a permissive range. The temperatures are given for in situ conditions so that values above 100°C are possible for barophilic species. Points with a precision worse than ±5°C were discarded. The average accuracy is ±2.5°C in the data set, in good agreement with the confidence intervals for $T_{opt}$ estimators. Similar results are expected with $T_{min}$ or $T_{max}$ because of the strong correlation between cardinal temperatures (Rosso et al. 1993).

*Genomic G+C Content (224 Genera).* The buoyant density centrifugation (Bd) and the thermal denaturation midpoint determination (Tm) were the main methods used to measure the genomic G+C content. When both were reported for a given organism the results were compared and found to be very similar, so data were merged regardless of the method employed. Data with a precision worse than ±5% G+C were discarded. The average accuracy in the data set is ±2.7% G+C.

*16S rRNA (165 Genera) and 23S rRNA (38 Genera) G+C content.* The G+C contents of rRNA stems and rRNA loops were computed from the primary and secondary structures of 16S and 23S rRNA in the rRNA database (Van de Peer et al. 1994) available at URL http://rrna.uia.ac.be/.

*5S RNA G+C Content (71 Genera).* The 5S RNA sequences were extracted from the DDBJ/EMBL/GenBank database (Shin-I et al. 1994; Rodriguez-Tomé et al. 1996; Benson et al. 1996) structured under the entity-relationship model of ACNUC (Gouy et al. 1984, 1985a,b). Some sequences were removed from the raw data set to avoid duplicated data: all possible intraspecific sequence pairs were checked, and a single sequence—the longest—was used when the pair showed less than two differences, either in primary structure or in length.

*tRNA G+C Content (51 Genera).* The tRNA G+C content was computed from the database of tRNA sequences and sequences of tRNA genes (Sprinzl et al. 1996) available at URL ftp://ftp.ebi.ac.uk/pub/databases/trna. Some recent tRNA sequences from the DDBJ/EMBL/GenBank database were also included.

Single-stranded RNA positions were not removed from the 5S rRNA and tRNA data sets. About 60% of the nucleotides are paired in these molecules. The RNA G+C contents were computed from sequences for any species of a given genus, whether or not its particular $T_{opt}$ was known. Therefore, the species involved in $T_{opt}$ estimation did not exactly match species used to estimate the RNA G+C content for a given genus. We assumed here that intrageneric variation is small compared with intergeneric variation. This assumption is challenged by a few prokaryotic genera in which intrageneric variability greatly exceeded the usual level. The difference between the highest $T_{opt}$ value and the lowest $T_{opt}$ value was found to be over 20°C within the genera *Bacillus, Clostridium, Cytophaga, Methanobacterium, Methanococcus,* and *Pseudomonas.* Data from these genera were therefore analyzed separately.

*Availability.* The data set is available at URL ftp://biom3.univ-lyon1.fr/pub/dataset/JME97/.

## Results

### G+C Content of Known Secondary Structures as a Function of $T_{opt}$

Figure 1 shows a significant correlation between G+C content and temperature in tRNAs, 5S rRNAs, and the stems of 16S and 23S rRNAs ($P < 0.0001$ in all cases). These results are consistent with those from Dalgaard and Garret (1993) for the G+C content of 16S rRNAs in hyperthermophile archaebacteria. We also found that the C content increased faster than the G content when $T_{opt}$ increased ($P < 0.0001$ for the correlation between C content minus G content and $T_{opt}$, not shown), suggesting that conversion from G:U pairs to G:C pairs may be a mechanism of G+C enrichment in double-stranded RNA. Thus, the fraction of G:C pairs increased with temperature, as is required for stabilization of secondary structure. This relationship appeared to be universal. It is found in all the molecules unequivocally involved in secondary structures (ribosomal RNAs and transfer RNAs), and no thermophilic genus was identified that had a low or moderate G+C content in its secondary-structure-involved RNA. There was a weak correlation between the G+C content of 16S and 23S rRNA loops and $T_{opt}$ (results not shown). It may be due to the presence of residual stems within the regions quoted as loops in the database, or of a tertiary structure effect.
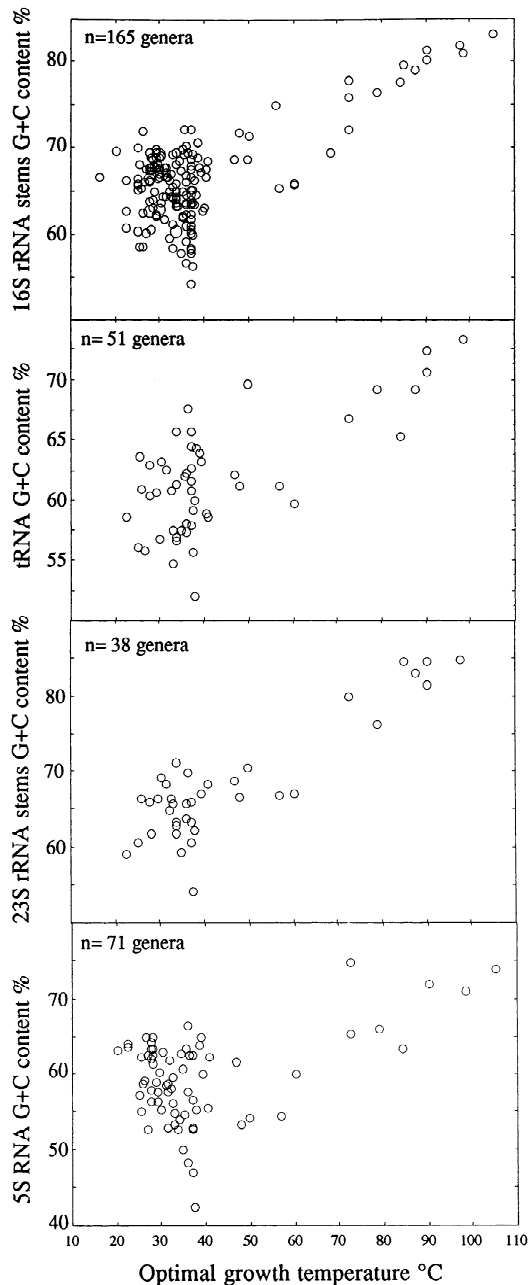
**Fig. 1.** G+C contents of 16S rRNA stems, 23S rRNA stems, 5S rRNAs and tRNAs plotted against optimal growth temperature.

*Genomic G+C Content as a Function of $T_{opt}$*

No correlation was found between the genomic G+C content and optimal growth temperature in prokaryotes (Fig. 2). The genomic G+C content was 25–77% in mesophilic genera (mean 54.2% GC) and 31–67% in thermophilic genera ($T_{opt} \geqslant 45°C$, mean 48.8% GC). The same pattern (a correlation between RNA G+C content and $T_{opt}$, but no correlation between genomic G+C content and $T_{opt}$) was found by intrageneric analyses of nine *Bacillus* species, 64 *Clostridium* species, 22 *Cytophaga* species, eight *Methanobacterium* species, seven *Methanococcus* species, and 31 *Pseudomonas* species (results not shown).

*G+C Content of 16S rRNA Stems as a Function of Genomic G+C Content*

The relationship between the genomic G+C content and the 16S rRNA stem G+C content for 165 prokaryotic genera is shown in Fig. 3. There was a strong correlation between them for the mesophilic genera ($T_{opt} < 45°C$, filled circles). This result is consistent with that of Muto and Osawa (1987) for a smaller (13 species) data set. But there was no correlation for thermophilic genera ($T_{opt} \geqslant 45°C$, open circles).

## Discussion

The G+C content of nucleic acids is known to be correlated with the stability of their double-helix (Marmur and Doty 1959; Wada and Suyama 1986) but the functional relevance of this G+C content is still debated. It has been suggested that thermal stability determines the nuclear G+C content in vertebrates (Bernardi and Bernardi 1986; Bernardi et al. 1988; Filipski 1990) and plants (Salinas et al. 1988). There have been fewer studies on this relationship in prokaryotes. An increased G+C content in third codon positions of genes in mildly thermophilic *Bacillus stearothermophilus* ($T_{opt}$: 62°C) and in the thermophile *Thermus thermophilus* ($T_{opt}$: 72°C) was interpreted as a selective advantage that helps to stabilize ''dynamic structures'' in mRNA during transcription and translation (Winter et al. 1983; Kagawa et al. 1984). However, there are also thermophilic species with low genomic G+C contents, such as *Pyrococcus furiosus* ($T_{opt}$: 97°C, G+C%: 38; Fiala and Stetter 1986), that indicate no selective advantage of a high genomic G+C content at high temperature. Moreover, as stressed by Muto and Osawa (1987), this relationship does not hold for mesophilic G+C–rich bacteria such as *Micrococcus* and *Streptomyces*. These objections were considered to be nonconclusive by Bernardi (1993), because the thermal stabilization of genomes might be due not to an increase in G+C but to other physiological adaptations.
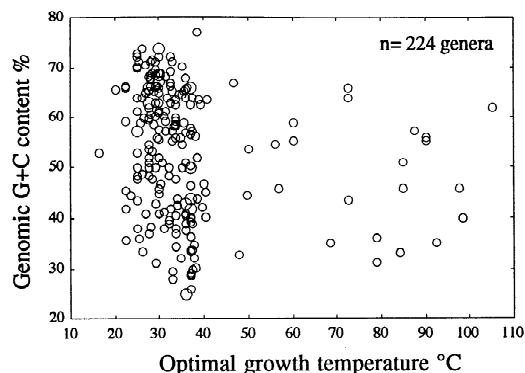


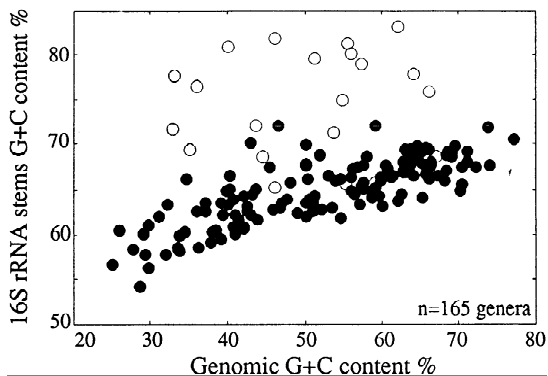**Fig. 2.** Genomic G+C content plotted against the optimal growth temperature for 224 prokaryotic genera.

**Fig. 3.** G+C content of 16S rRNA stems plotted against the optimal growth temperature for 165 prokaryotic genera. *Filled circles,* mesophilic genera, strong correlation. *Open circles,* thermophilic genera, no correlation.

The present analysis shows that the expected relationship between G+C content in RNA that has some secondary structure and optimal growth temperature is present in prokaryotes. This pattern was found in all the molecules and all the genera examined, suggesting that any secondary structure that must endure a high temperature requires a high G+C content. In contrast, no relationship was detected between genomic G+C content and $T_{opt}$ for prokaryotes.

Forsdyke (1995a) suggested that the rough equalities between C and G (respectively, A and T) contents in single-stranded DNA are the result of selection pressure favoring the evolution of numerous secondary structures in bacterial genomes. If such structures actually exist, the G+C content of genomic sequences should behave similarly to the G+C content of tRNA and rRNA with respect to temperature. Since there is no correlation, a high proportion of secondary structures in bacterial genomes is unlikely, and therefore PR2 is poorly explained by the above hypothesis. A similar conclusion was reached by Wada and Suyama (1986) based on a weaker argument: the great variations in the G+C contents of palindromic DNA sequences were found to be incompatible with actual secondary structures.

The work of Forsdyke is based on the comparison between the folding energy minimization values in actual genomic sequences and those in shuffled genomic sequences. Minimal energies are computed using the program FOLD (Zuker 1989). Genomic sequences generally have a better minimal energy than random sequences with the same base composition. This result is interpreted as being due to numerous intrastrand stem-loop structures in genomic DNA (Forsdyke 1995b), which contradicts results reported here. We have reproduced Forsdyke's results, but do not agree with their interpretation. We suggest that the results of Forsdyke (1995b) and the present above results can be interpreted as being due to alternative structures, such as codon context bias (Yarus and Folley 1985; Shpaer 1986; Gouy 1987), tetranucleotide bias due to very-short-patch repair (Gutiérrez et al.

1994, 1996), and dinucleotide bias due to transcription-coupled repair (Selby and Sancar 1993; Francino et al. 1996). These may all increase the frequency of palindromic patterns in DNA that does not correspond to secondary structures. There is experimental evidence for palindromic patterns that do not correspond to secondary structures in DNA in vivo (Sinden et al. 1983; Lyamichev et al. 1984). We therefore conclude that the intrastrand statistical equalities A = T and C = G are best explained as the result of neutral directional mutation pressure.

## References

Benson DA, Boguski M, Lipman DJ, Ostell J (1996) GenBank. Nucleic Acids Res 24:1–5

Bernardi G (1993) The vertebrate genome: isochores and evolution. Mol Biol Evol 10:186–204

Bernardi G, Bernardi G (1986) Compositional constraint and genome evolution. J Mol Evol 24:1–11

Bernardi G, Mouchiroud D, Gautier C, Bernardi G (1988) Compositional patterns in vertebrate genomes: conservation and change in evolution. J Mol Evol 28:7–18

Chargaff E (1979) How genetics got a chemical education. Ann NY Acad Sci 325:345–360

Dalgaard JZ, Garrett A (1993) Archaeal hyperthermophile genes. In: Kates M et al. (eds) The biochemistry of Archaea (Archaebacteria). Elsevier Science, Amsterdam, pp 535–562

Fiala G, Stetter KO (1986) *Pyrococcus furiosus* sp. nov. represents a new genus of marine heterotrophic archaebacteria growing optimally at 100°C. Arch Microbiol 145:56–61

Fickett JW, Torney DC, Wolf DR (1992) Base compositional structure of genomes. Genomics 13:1056–1064

Filipski J (1990) Evolution of DNA sequences. Contributions of mutational bias and selection to the origin of chromosomal compartments. In: Obe G (ed) Advances in mutagenesis research 2. Springer Verlag, Berlin, pp 1–54

Forsdyke DR (1995a) Relative roles of primary sequence and (G+C)% in determining the hierarchy of frequencies of complementary trinucleotide pairs in DNAs of different species. J Mol Evol 41:573–581

Forsdyke DR (1995b) A stem-loop ''kissing'' model for the initiation of recombination and the origin of introns. Mol Biol Evol 12:949–958

Francino MP, Chao L, Riley MA, Ochman H (1996) Asymmetries generated by transcription-coupled repair in enterobacterial genes. Science 272:107–109

Freese E (1962) On the evolution of the base composition of DNA. J Theor Biol 3:82–101

Gouy M (1987) Codon contexts in enterobacterial and coliphage genes. Mol Biol Evol 4:426–444

Gouy M, Milleret F, Mugnier C, Jacobzone M, Gautier C (1984) ACNUC: a nucleic acid sequence data base and analysis system. Nucleic Acids Res 12:121–127

Gouy M, Gautier C, Attimonelli M, Lanave C, di Paola G (1985a) ACNUC—a portable retrieval system for nucleic acid sequence databases: logical and physical designs and usage. Comput Appl Biosci 3:167–172

Gouy M, Gautier C, Milleret F (1985b) System analysis and nucleic acid sequence banks. Biochimie 67:433–436

Gutiérrez G, Casadesús J, Olivier JL, Marín A (1994) Compositional heterogeneity of the *Escherichia coli* genome: a role for VSP repair? J Mol Evol 39:340–346

Gutiérrez G, Casadesús J, Olivier JL, Marín A (1996) A possible re-

lationship between VSP mismatch repair and gene expression level. J Mol Evol 43:161–163

Kagawa Y, Nojima H, Nukiwa N, Ishizuka M, Nakajima T, Yasuhara T, Tanaka T, Oshima T (1984) High guanine plus cytosine content in the third letter of codons of an extreme thermophile: DNA sequence of the isopropylmalate dehydrogenase of *Thermus thermophilus.* J Biol Chem 259:2956–2960

Karkas JD, Rudner R, Chargaff E (1968) Separation of *B. subtilis* DNA into complementary strands. II. Template functions and composition as determined by transcription with RNA polymerase. Proc Natl Acad Sci USA 60:915–920

Karkas JD, Rudner R, Chargaff E (1970) Template properties of complementary fractions of denatured microbial deoxyribonucleic acids. Proc Natl Acad Sci USA 65:1049–1056

Lobry JR (1995) Properties of a general model of DNA evolution under no-strand-bias conditions. J Mol Evol 40:326–330, 41:680

Lobry JR (1996a) Asymmetric substitution patterns in the two DNA strands of bacteria. Mol Biol Evol 13:660–665

Lobry JR (1996b) Origin of replication of *Mycoplasma genitalium.* Science 272:745–746

Lyamichev V, Panyutin I, Frank-Kamenetskii MD (1984) The absence of cruciform structure from pAO3 plasmid DNA *in vivo.* J Biomol Struct Dyn 2:291–301

Marmur J, Doty P (1959) Heterogeneity in deoxyribonucleic acids. I. Dependence on composition of the configurational stability of deoxyribonucleic acids. Nature 183:1427–1429

Muto A, Osawa S (1987) The guanine and cytosine content of genomic DNA and bacterial evolution. Proc Natl Acad Sci USA 84:166–169

Nussinov R (1982) Some indications for inverse DNA duplication. J Theor Biol 95:783–791

Prabhu VV (1993) Symmetry observation in long nucleotide sequences. Nucleic Acids Res 21:2797–2800

Rodriguez-Tomé P, Stoehr PJ, Cameron GN, Flores TP (1996) The European Bioinformatics Institute (EBI) databases. Nucleic Acids Res 24:6–12

Rosso L, Lobry JR, Flandrois JP (1993) An unexpected correlation between cardinal temperatures of microbial growth highlighted by a new model. J Theor Biol 162:447–463

Rudner R, LeDoux M (1974) Distribution of pyrimidine oligonucleotides in complementary strand fractions of *Escherichia coli* deoxyribonucleic acid. Biochemistry 13:118–125

Rudner R, Karkas JD, Chargaff E (1968) Separation of B. subtilis DNA into complementary strands. III. Direct analysis. Proc Natl Acad Sci USA 60:921–922

Rudner R, Karkas JD, Chargaff E (1969) Separation of microbial deoxyribonucleic acids into complementary strands. Proc Natl Acad Sci USA 63:152–159

Salinas J, Matassi G, Montero LM, Bernardi G (1988) Compositional compartmentalization and compositional patterns in the nuclear genomes of plants. Nucleic Acids Res 16:4269–4285

Selby CP, Sancar A (1993) Molecular mechanism of transcription-repair coupling. Science 260:53–58

Shin-I T, Ikeo K, Tateno Y, Gojobori T (1994) The DNA database of Japan. Biochemist 16:18–21

Shpaer EG (1986) Constraints on codon context in *Escherichia coli* genes, their possible role in modulating the efficiency of translation. J Mol Biol 188:555–564

Sinden RR, Broyles SS, Pettijohn E (1983) Perfect palindromic *lac* operator DNA sequence exists as a stable cruciform structure in supercoiled DNA *in vitro* but not *in vivo.* Proc Natl Acad Sci USA 80:1797–1801

Sprinzl M, Steegborn C, Hübel F, Steinberg S (1996) Compilation of tRNA sequences and sequences of tRNA genes. Nucleic Acids Res 24:68–72

Staley JT, Bryant MP, Pfennig N, Holt JG (1984) Bergey's manual of systematic bacteriology. Williams and Wilkins, Baltimore

Sueoka N (1962) On the genetic basis of variation and heterogeneity of DNA base composition. Proc Natl Acad Sci USA 48:582–592

Sueoka N (1988) Directional mutation pressure and neutral molecular evolution. Proc Natl Acad Sci USA 85:2653–2657

Sueoka N (1992) Directional mutation pressure, selective constraints, and genetic equilibria. J Mol Evol 34:95–114

Sueoka N (1993) Directional mutation pressure, mutator mutations, and dynamics of molecular evolution. J Mol Evol 37:137–153

Sueoka N (1995) Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. J Mol Evol 40:318–325, 42:323

Van de Peer Y, Van den Broeck I, De Rijk P, De Wachter R (1994) Database on the structure of small ribosomal subunit RNA. Nucleic Acids Res 22:3488–3494

Wada A, Suyama A (1986) Local stability of DNA and RNA secondary structure and its relation to biological functions. Prog Biophys Mol Biol 47:113–157

Watson JD, Crick FHC (1953) A structure for deoxyribose nucleic acid. Nature 171:737–738

Winter G, Koch GLE, Hartley BS, Barker DG (1983) The amino acid sequence of the tyrosyl-tRNA synthetase from *Bacillus stearothermophilus.* Eur J Biochem 132:383–387

Yarus M, Folley LS (1985) Sense codons are found in specific contexts. J Mol Biol 182:529–540

Zuker M (1989) Computer prediction of RNA secondary structure. Methods Enzymol 180:262–289