THE EFFECT OF DNA STABILITY ON MUTATION AND SEQUENCE EVOLUTION

L. Medrano[1], G. Cocho[2], P. Miramontes[1] and J.L. Rius[2]
[1]Facultad de Ciencias, [2]Instituto de Física
Universidad Nacional Autónoma de México
Apartado postal 70-572. Cd. Universitaria
México 04510 D.F. MEXICO

Received 30 December 1992, 8 December 1993

ABSTRACT: The stacking interaction between consecutive base pairs along duplex nucleic acids, causes stability to depend on nucleotide sequence; this fact allows the inference of local energy distribution. We have analyzed the implications of DNA stability in different mutation events. It was found that spontaneous mutations such as some double-strand breakages and base substitutions occur more frequently in DNA regions with low stability. This trend seems not to be a process directly involving duplex DNA stability in mutation mechanisms. It is suggested that this bias could affect sequence evolution if selective constraints are absent or relaxed enough, as examined for the variability profile in one aligned set of cetacean D-loop sequences. The bias for the formation of stable duplexes could favor the accumulation of dinucleotides with G and C or the dimer ApA/TpT. The analysis of enthalpy distributions in coding sequences shows that their global stabilities strongly depend on the interaction between bases at the first and second codon positions. The hypermutable immunoglobulin regions D and J have remarkably low stability values in the interaction between the first and second codon positions and their energy distribution is different from the general pattern in coding sequences.

* * *

INTRODUCTION

Research on oligonucleotides has shown that the structure and physical properties of nucleic acids are heterogeneous along double strands. Base orientation is not constant; it greatly depends on the hindrance effect of neighboring opposite purines (Dickerson 1983). Since consecutive bases interact, duplex stability depends on nucleotide sequence. Breslauer *et al.* (1986) have reported 10 stability parameters given as the enthalpy (ΔH), free energy (ΔG) and entropy (ΔS) for the helix-to-coil transition in the different base dimers. DNA structure also varies in time as duplex DNA spontaneously bends with base pair opening (Ramstein and Lavery 1988). Therefore, we considered appropiate to ask whether structure fluctuations, as they depend on DNA heterogeneities, affect mutation events and sequence evolution.

Here, we have analyzed DNA sequences in terms of enthalpy values instead of free energy because the former are larger numbers and differences are more clearly seen. Also, we have considered DNA stability in the sense of the energy necessary to cause duplex fluctuations, and this energy is better reflected in enthalpy. Anyway, free energy and enthalpy track each other.

* * * *

Evolutionary Theory 10: 249-258 (August, 1994).
The editors thank two referees for help in evaluating this paper.

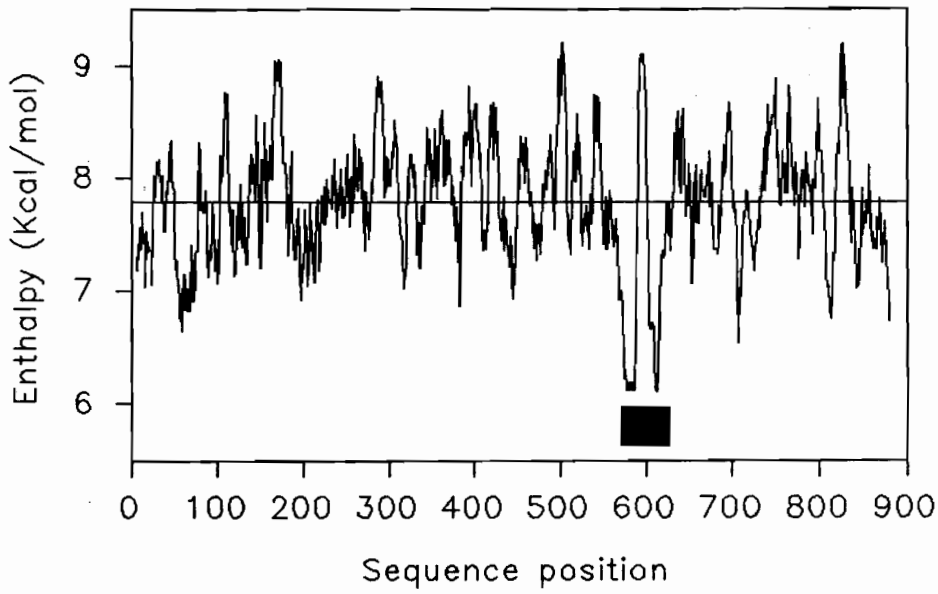

Figure 1. Enthalpy profile from the second intervening sequence of human $^{G}\gamma$-globin. The bar indicates the recombination hotspot and the horizontal line stands for the average sequence enthalpy.

DNA STABILITY AND MUTATION EVENTS

Different mutational phenomena exist and most probably there are diverse effects of DNA stability. In a general sense, it is expected that structure fluctuations occur more frequently in DNA sites having lower local duplex stability; thereby, it is also expected that such sites could be subject to a higher mutability independent of the mutation mechanisms.

Double-strand breakage is the most obvious event in which the effect of DNA stability on mutation could be expected. In order to see whether this could be true, we have analyzed the energy distribution in the second intervening sequence of the human $^{G}\gamma$-globin, which is known to bear a hotspot for recombination and related phenomena such as gene conversion (Scott *et al.* 1984, Slightom *et al.* 1980). In this hotspot region, large runs of the dimer TG are found and they clearly are of low stability. Accordingly, an energy profile for the whole intervening sequence shows that the hotspot is a region with remarkably low duplex stability (Figure 1).

Although the relationship of DNA stability to other mutational events such as base substitution is less clear in terms of mechanisms, since substitutions are mostly associated with replication, we decided to analyze whether our general hypothesis could be upheld. Our preliminary idea is that fluctuations in uncoiling DNA make the replication process prone to error. In this

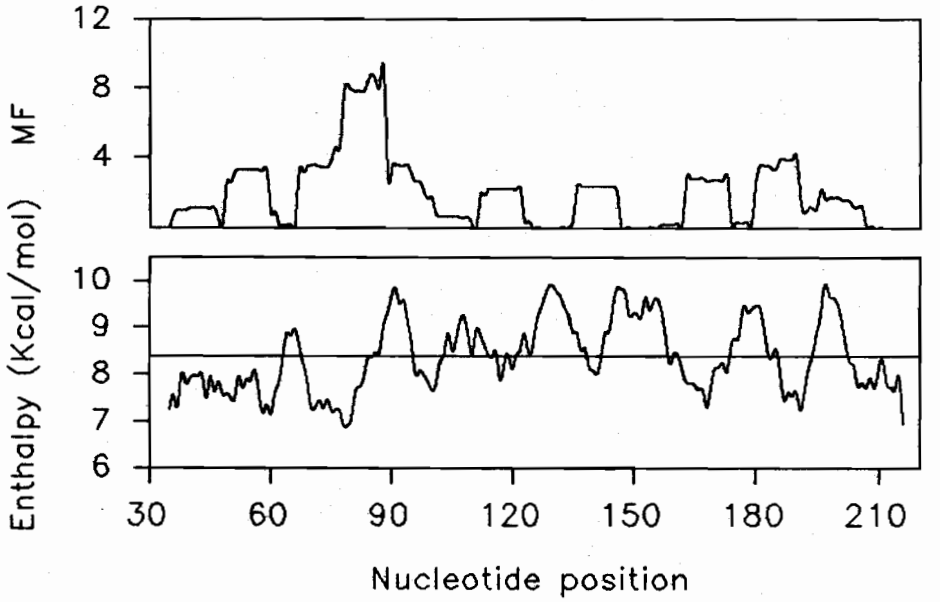

Figure 2. Top: Smoothed profile of spontaneous mutation frequency (MF) for a segment of the *lac*I gene according to Schaaper and Dunn (1987). Bottom: DNA-stability smoothed profile. The horizontal line shows the average enthalpy of the sequence.

context, Reanney and Pressing (1984) have proposed that the spontaneous substitution rate in replication depends on some energy factor of DNA structure and that substitution rate is affected by temperature. Schaaper and Dunn (1987) reported the spectra of spontaneous mutations in *E. coli*, obtaining sequences in a segment of the *lac*I gene from strains defective for mismatch correction. It was found that pairs with A and T are more frequently replaced than pairs with G and C. This increases global DNA stability in mutants, as expected from our hypothesis. In the present work, the ten parameters of ΔH from Breslauer *et al.* (1986) were assigned to the native sequence and the enthalpy profile was obtained by smoothing the series of values. For each position the number of mutants caused by base substitution were counted, and a smoothed profile of mutation frequency was also made. Figure 2 shows that spontaneous base substitutions are more frequent in the sites where the native sequence has lower energy values.

DNA STABILITY AND SELECTIVE CONSTRAINTS

Mutation, in conventional evolution theory, is a random event in that it is not causally related to the fitness of the organism. The evolution of nucleotide sequences derives from a mutation event followed by the fixation process. Amino acid sequence is thought to be the strongest constraint in protein coding sequences

(Kimura 1986, Kimura and Ohta 1974). Natural selection also accounts for additional factors such as codon-usage patterns (Ikemura 1985) and structural properties of DNA and RNA (Cocho *et al.* 1991). However, mutations are not randomly distributed throughout the genome, as shown by Benzer (1961) and some factors, like cytosine methylation, are known to favor base substitution (Coulondre *et al.* 1978). Mutation rate is also affected by temperature and, as a selective factor, it restrains base composition (Bernardi and Bernardi 1986, Reanney and Pressing 1984). The analysis of substitution patterns in pseudogenes has shown that cytosine methylation can bias sequence evolution in the absence of external constraints (Li *et al.* 1984). In that study, Li *et al.* (1984) found that base pairs with G and C are more frequently replaced than base pairs with A and T as a result of 5-methyl,cytosine deamination.

In order to study whether sequence evolution could be affected by the propensity of low-stability neighborhoods to mutation, sequence comparison is worthy of analysis. However, there is not, as a rule, a reliable assignment of which base is original and which is mutation-originated. On the other hand, the analysis of weakly constrained sequences is often difficult with respect to the alignment itself; multiple mutation events must be accounted for and gaps cannot be easily detected since sequences exhibit a high variability. Therefore, as a suitable example for correlating variability with DNA energy profile, a set of cetacean sequences from the mitochondrial D-loop has been analyzed. As usual in the mammalian mitochondrion genome, most of the sequence changes in the D-loop are substitutions. Since the D-loop does not code for any peptide, it should be selectively constrained by the DNA sequence itself but, having a high substitution rate in mammals (Wilson *et al.* 1985), it approximates to a non-restrained *locus*. Also, the cetacean genome evolves at a low rate compared to other mammals (Árnason 1974, Baker *et al.* 1993, Martin and Palumbi 1993, Schlöterer *et al.* 1991). All this implies that the sequence variability of the cetacean D-loop is close to reflecting single mutational events.

Figure 3 shows the variability and energy profiles from a set of three aligned sequences of the cetacean D-loop. As in the profile from spontaneous base substitutions, the regions where most of sequences have lower local stabilities exhibit higher variability. Accordingly, low-variability zones are observed to have regions of high enthalpy.

Substitution of base pairs composed of A and T could favor ΔH increase while replacement of base pairs composed of G and C could decrease ΔH. If substitution events are more frequent in sites having low stability, then an increase in GC content is expected, as dinucleotides with G and C are the most stable ones. However, this trend could favor the accumulation of dinucleotide ApA/TpT if dinucleotides with G and C are substituted, since ApA/TpT is the most stable dimer formed by A and T.

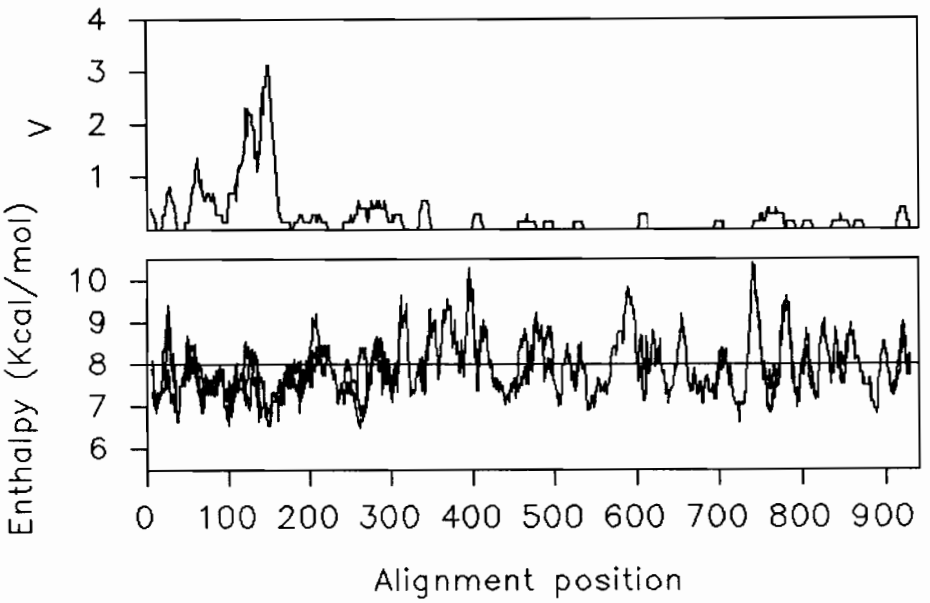

Figure 3. Top: Smoothed variability profile of three aligned cetacean sequences from the mitochondrial D-loop. Bottom: Smoothed profile of DNA stability for each sequence. The horizontal line stands for the average enthalpy of all sequences. The alignment includes the D-loop sequences from one blue whale, *Balaenoptera musculus*, one fin whale, *B. physalus*, and one humpback whale, *Megaptera novaeangliae* (Árnason 1991). Variability measurement at each alignment position (V) was modified from Alizon *et al.* (1986) as V = (D-1)/(A/N) where D is the number of different residues, A is the number of the more abundant residue and N is the number of aligned sequences.

In order to determine the relationship between two consecutive bases i and j in a sequence, we define the correlation index C_{ij}:

$$C_{ij} = (N_{ij}/(N-1) - (N_iN_j/N^2)$$

where N_{ij} is the number of the dimers formed by i and j, N_i and N_j are respectively the number of bases i and j and N is the total number of bases. There are 16 possible C_{ij} indexes but, of those only nine are independent. From these indices, we define the preference of ApA/TpT over ApT/TpA and TpA/ApT dimers as:

$$P_{aa/tt} = (C_{aa}+C_{tt}) - (C_{at}+C_{ta})$$

In Figure 4 the preference index of ApA/TpT for 280 eukaryotic sequences obtained from GenBank (Bilofsky *et al.* 1986) is plotted as a function of the proportion of dimers with G and C.

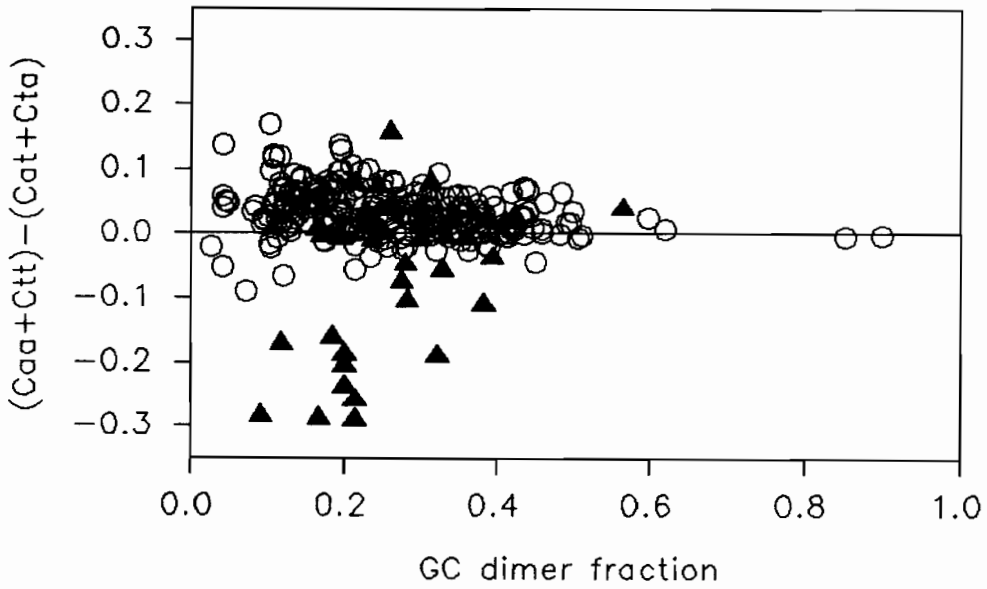

Figure 4. Preference of ApA and TpT over ApT and TpA measured with C_{ij} indexes for 280 eukaryotic sequences in relation to the content of GC dimers (GpG/CpC, GpC/CpG and CpG/GpC). Triangles stand for 42 hypermutable regions from immunoglobulins. Mean values of $(C_{aa}+C_{tt})-(C_{at}+C_{ta})$ are -0.046 ± 0.017 SE for hypermutable sequences and 0.032 ± 0.002 SE for the 238 remaining sequences in the analyzed set. The energy values for the AT dimers are, in Kcal/mol: TpA/ApT, 6.0; ApT/TpA, 8.6; ApA/TpT, 9.1.

As dinucleotides with G and C decrease in the sequences, preference for ApA/TpT shows a clear trend to increase from zero. Contrarily, most of the hypermutable regions from immunoglobulins have a preference for ApT/TpA and TpA/ApT dimers as seen by negative values of the $P_{aa/tt}$ index. If a muational bias towards the formation of stable sequences is recognized, it is questioned whether hypermutability arises from a particular DNA energy distribution in which unstable sites are prone to mutate.

DISTRIBUTION OF DNA ENTHALPY IN CODING SEQUENCES

For coding sequences, it is worthwhile to analyze the enthalpy distribution in relation to the codon positions because positions are differentially constrained. If xyz denotes any codon, xy, yz and zx are the interactions between the respective positions. Thus, by assigning the DNA enthalpy parameters in a sequence, the total average enthalpy (ΔH_w) and the average enthalpy by codon position (ΔH_{xy}, ΔH_{yz}, ΔH_{zx}) can be calculated. Figure 5 shows the distribution of ΔH_w by codon position for a set of coding sequences obtained from GenBank. ΔH is distributed heterogeneously among codon positions, leading to an enthalpy

periodicity of three bases characteristic of coding sequences (Cocho *et al.* 1990). It is observed that ΔH_{xy} closely follows the values of ΔH_w, as its corresponding slope is $s_{xy} \cong 1$. Since x and y positions basically define the amino acid identity, it is concluded that ΔH_w is determined by the peptide sequence. Position z then defines ΔH_{yz} and ΔH_{zx} with slopes $s_{yz}<1$ and $s_{zx}>1$ respectively. The slope values are interdependent because $s_{xy}+s_{yz}+s_{zx}=3$. Thus, $s_{zx}>1$ implies that the zx interaction varies more with a mutational event corresponding to z and x being the less restrained bases in the codon.

The relationship between ΔH_w and ΔH by codon position is variable for different sequences but they share, in general, the pattern previously described. As was observed in Figure 4, hypermutable sequences exhibit different C_{ij} values, suggesting that low stability in DNA could account for hypemutability. Therefore, the energy distribution in highly mutable sequences such as immunoglobulin D and J segments was analyzed for comparison with other coding sequences. The D and J set includes sequences from a wide variety of vertebrate immunoglobulins. Figure 5 (bottom) shows that D and J sequences have ΔH_{xy}, ΔH_{yz} and ΔH_{zx} distribution highly differentiated, with $s_{zx}>1$ as in the whole sequence set but having quite low ΔH_{xy} values with $s_{xy}<1$.

Figure 5 also shows that the particular energy distribution in D and J is not due to a frameshift in codon position but to a different energetic relationship between bases. In reference to the slopes of Figure 5, variable regions from immunoglobulins have an energy distribution similar to D and J sequences but enthalpy values are quite similar between the three positions while D and J show clearly low values at xy interaction and higher at zx. A t-test for the signed difference between ΔH_{xy} and ΔH_w shows that this value is significatively lower in hypermutable sequences (-0.41 Kcal/mol in average) compared to variable sequences (0.01 Kcal/mol in average) with $p<0.01$. This suggests that the particular low stability at xy interaction could be related to hypermutability. This hypothesis is reinforced by the fact that immunoglobulin mutability seems not to be related to antigen exposition or cell maturation (Wabl *et al.* 1985).

FINAL COMMENTS

When analyzing diferent problems of molecular evolution, selective schemes responsible for the non-randomness of DNA primary structure are generally invoked. Many of these explanations assume a random mutation pattern, which actually does not occur, as shown by research on mutagenesis. Thus, it is possible that nucleotide sequences are biased by mutation phenomena at weakly restrained sites. According to neutral theory, the evolutionary rate of nucleotide substitution approaches the spontaneous substitution rate as selective constraints become weaker (Kimura 1986). Thus, base substitution tends to reflect not only its spontaneous rate but the whole mutation event including its propensity at sites with low local DNA stability.

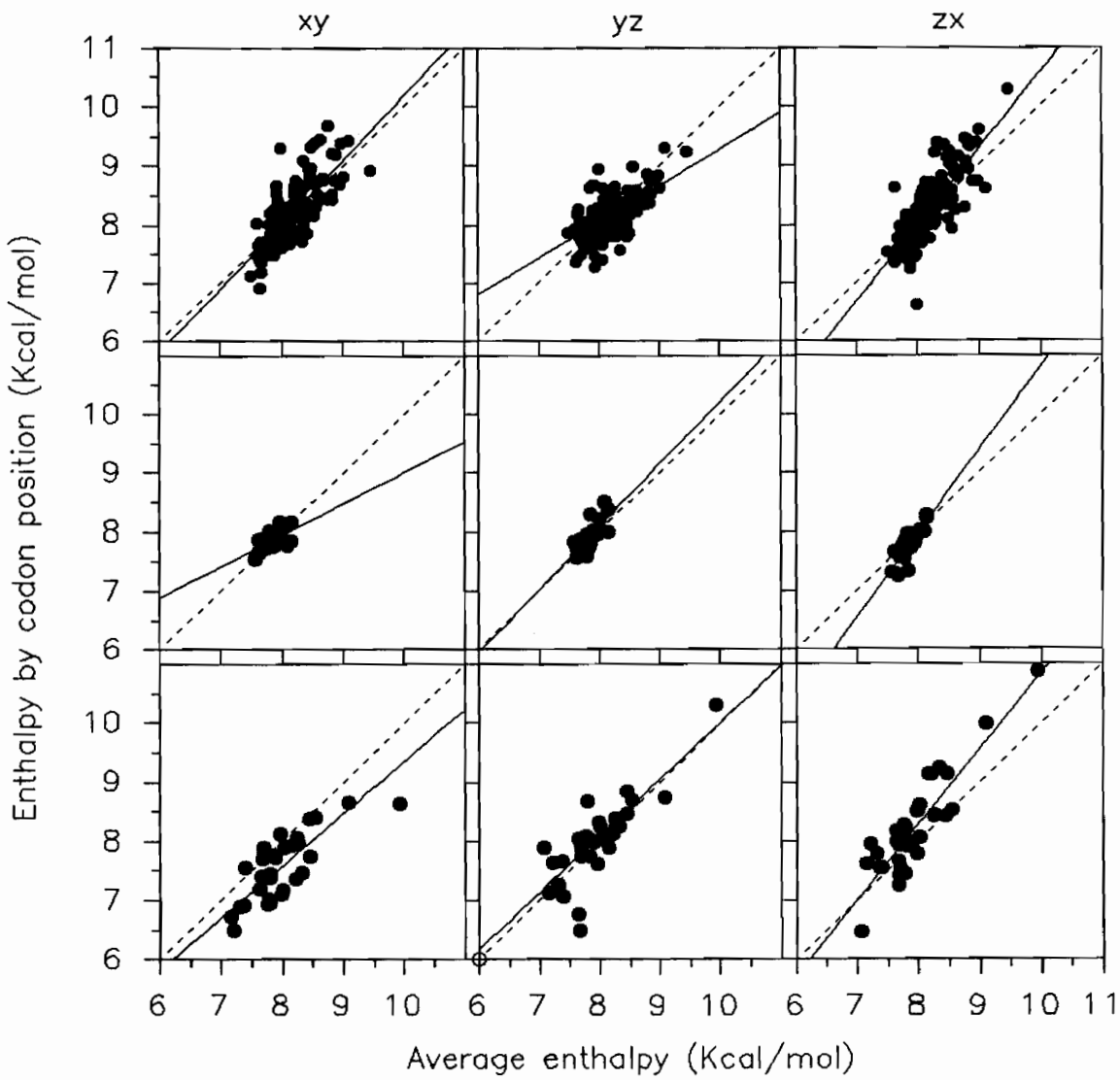

Figure 5. Energy distribution by codon position. Solid lines display linear regression. Top: Distribution for 173 coding sequences from different eukaryotes, prokaryotes and viruses. Center: Distribution for 21 variable regions of immunoglobulins. Bottom: Distribution for 37 hypermutable regions of immunoglobulins. Left: Energy between first and second codon position (xy). Center: second and third position (yz). Right: third and first position (zx).

ACKNOWLEDGEMENTS

We are grateful to A. Lazcano, S. Rodin, R. Austin, K. Breslauer and Ch.-I. Wu for critical comments, and to S.R. Palumbi and C.S. Baker from The Kewalo Marine Laboratory by kindly advicing and providing cetacean sequence data. We also are indebted to L. Van Valen by valuable help with editing. This work has been supported by Projects UNAM.IN 105289 from DGAPA UNAM and 0037 N-9106 from CONACyT. L. Medrano was also supported by SNI. This work is dedicated to M.E. Sandoval and R. Lara *in memoriam*.

REFERENCES

Alizon M, Wain-Hobson S, Montagnier L, Sonigo P (1986) Genetic variability of the AIDS virus: nucleotide sequence analysis of two isolates from african patients. Cell 46: 63-74

Árnason Ú (1974) Comparative chromosome studies in Cetacea. Hereditas 77: 1-36

Árnason Ú (1991) Sequence composition of the D-loop in mitochondrial DNA of the fin, blue and humpback whales. Report for the International Whaling Commission (SC/F91/F31)

Baker CS, Perry A, Bannister JL, Weinrich MT, Abernethy RB, Calambokidis J, Lien J, Lambertsen RH, Ramírez JU, Vasquez O, Clapham PJ, Alleng A, O'Brien SJ, Palumbi SR (1993) Abundant mitochondrial DNA variation and world-wide population structure in humpback whales. Proc Natl Acad Sci USA 90: 8239-8243

Benzer S (1961) On the topography of the genetic fine structure. Proc Natl Acad Sci USA 47: 403-415

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

Bilofsky HS, Burks C, Fickett JW, Goad WR, Lewitter FI, Rindone WP, Swindell CD, Tung CS (1986) The Genbank genetic sequence databank. Nucl Acid Res 14: 1-4

Breslauer KJ, Frank R, Blöcker H, Marky LA (1986) Predicting DNA duplex stability from the base sequence. Proc Natl Acad Sci USA 83: 3746-3750

Cocho G, Medrano L, Miramontes P, Rius JL (1991) Selective constraints over DNA sequence. In: Biologically inspired physics. Peliti L (ed) Plenum press: 63-69

Cocho G, Rius JL, Miramontes P, Medrano L (1990) Structural constraints, DNA periodicities and gene dynamics. In: Quasicrystals and incommensurate structures. Yacamán MJ, Romeu D, Castaño V, Gómez A (eds) World scientific press: 465-475

Coulondre G, Miller JH, Farabaugh PJ, Gilbert W (1978) Molecular basis of base substitution hotspots in *Escherichia coli*. Nature 274: 775-780

Dickerson RE (1983) Base sequence and helix structure variation in B and A DNA. J Mol Biol 166: 419-441

Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 2: 13-34

Kimura M (1986) DNA and the neutral theory. Phil Trans R Soc Lond B 312: 343-354

Kimura M Ohta T (1974) On some principles governing molecular evolution. Proc Natl Acad Sci USA 71: 2848-2852

Li WH, Wu CI, Luo CC (1984) Nonrandomness of point mutations as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J Mol Evol 21: 58-71

Martin AP, Palumbi SR (1993) Body size, metabolic rate, generation time and the molecular clock. Proc Natl Acad Sci USA 90: 4087-4091

Ramstein J, Lavery R (1988) Energetic coupling between DNA bending and base pair opening. Proc Natl Acad Sci USA 85: 7231-7235

Reanney DC, Pressing J (1984) Temperature as a determinative factor in the evolution of genetic systems. J Mol Evol 21: 72-75

Schaaper RM, Dunn RL (1987) Spectra of spontaneous mutations in *Escherichia coli* strains defective in mismatch correction: The nature of *in vivo* DNA replication errors. Proc Natl Acad Sci USA 84: 6220-6224

Schlöterer C, Amos B, Tautz D (1991) Conservation of polymorphic simple sequence *loci* in cetacean species. Nature 354: 63-65

Scott AF, Heath P, Trusko S, Boyer SH, Chang L-YE, Slightom J.L. (1984) The sequence of the gorilla fetal globin genes: Evidence for multiple gene conversions in human evolution. Mol Biol Evol 1: 371-389

Slightom JL, Blechl AE Smithies O (1980) Human fetal $^{G}\gamma$- and $^{A}\gamma$-globin genes: Complete nucleotide sequence suggests that the DNA can be exchanged between these duplicated genes. Cell 21: 627-638

Wabl M, Burrows PD, Von Gabain A, Steinberg Ch (1985) Hypermutation at the immunoglobulin heavy chain locus in a pre-B-cell line. Proc Natl Acad Sci USA 82: 479-482

Wilson AC, Cann RL, Carr SM, George M, Gyllenstein UB, Helm-Bychowski KM, Higuchi RG, Palumbi SR, Prager EM, Sage RD, Stoneking M (1985) Mitochondrial DNA and two perspectives on evolutionary genetics. Biol J Linn Soc 26: 375-400