Article No. jmbi.1999.3275 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (1999) 294, 937–953

JMB

Tendency for Local Repetitiveness in Amino Acid Usages in Modern Proteins

Kazuhisa Nishizawa[1]*, Manami Nishizawa[1] and Ki Seok Kim[2]

[1]*Department of Biochemistry Teikyo University School of Medicine, Kaga, Itabashi Tokyo 173, Japan*

[2]*Laboratory of Immunobiology Dana-Farber Cancer Institute Boston, MA 02115, USA*

Systematic analyses of human proteins show that neural and immune system-specific, and therefore, relatively ''modern'' proteins have a tendency for repetitive use of amino acids at a local scale (~1-20 residues), while ancient proteins (human homologues of *Escherichia coli* proteins) do not. Those protein subsegments which are unique based on homology search account for the repetitiveness. Simulation shows that such repetitiveness can be maintained by frequent duplication on a very short scale (one to two codons) in the presence of substitutive point mutation, while the latter tends to mitigate the repetitiveness. DNA analyses also show the presence of cryptic (i.e. ''out of the codon frame'') repetitiveness, which cannot fully be explained by features in protein sequences. Simulative modification of the amino acid sequences of immune system-specific proteins estimate that 2.4 duplication events occur during the period equivalent to ten events of substitution mutation. It is also suggested that the repetitiveness leads to longitudinal unevenness within a given peptide domain. Those peptide motifs which contain similarly charged residues are likely to be generated more frequently in the presence of the tendency for repetitiveness than in its absence. Therefore, the neutral propensity of DNA for duplication, which can also tend to generate repetitiveness in amino acid sequences, seems to be manifested primarily when the constraints on amino acid sequences are relatively weak, and yet may be positively contributing to generation of unevenness in modern proteins.

*Corresponding author

Keywords: microsatellite; coding regions; peptide motif; triplet repeat

Introduction

Proteins have been assumed generally to evolve depending on their fitness and neutral drifts. On the other hand, it has been suggested that the propensities of genomes have significant influence on the ''neutral'' aspects and therefore on primary sequence of proteins (Nei, 1987; Bernardi, 1995). Even before the genetic codes were determined, Sueoka (1961) showed that there is a correlation between the nucleotide composition of genomic DNA and the amino acid composition of the proteins of the same organism. It has been demonstrated that eukaryote genomes have global scale unevenness which is referred to as isochore (Bernardi, 1995; Ikemura & Aota, 1988), and of note, it has been reported that genes belonging to different types of isochore encode proteins with different amino acid composition (D'Onofrio *et al.*, 1991) . We have shown that the occurrence of arginine (R) and lysine (K) residues has apparent arbitrariness based on the finding that the R *versus* K ratio correlates with local (~20 bp) G + C content of the corresponding gene (Nishizawa & Nishizawa, 1998). In fact, for the human gene, R/(R + K) equals 66 % on average when the corresponding genes are in the context of a DNA whose GC% = 70 ~ 80 %, (and 37 % when associated with the DNA of GC% = 30 ~ 40 %).

Although the origin of such global and local scale unevenness in a genome remains unclear, they may be generated, at least in part, by the duplication of DNA segments with various lengths. It is possible that duplication events of ~100 kb segments would affect the structure of an isochore, while duplications of 3-6 bp segments may cause local unevenness. The previous reports have suggested that the various scales of duplication is the fundamental process in evolution (Ohno, 1984, 1987; Doolittle, 1989). Tautz *et al.* (1986) also showed that a slippage-like mechanism

E-mail address of the corresponding author: kazunet@med.teikyo-u.ac.jp

0022-2836/99/490937–17 $30.00/0

is working for the generation of variation. Related to this study, microsatellite DNA has recently attracted much attention because of its potential usefulness for evolutionary and population genetics and medical diagnosis (for examples, see Weber & Wong, 1993; Goldstein *et al*., 1995; Rubinsztein *et al*., 1995; Kimmel *et al*., 1996). In general, such local repetitive motifs have been analyzed as a marker of evolution. It is not very clear how ubiquitously such phenomenon occurs over the whole genome. Our interest lies in how frequently the local scale tandem gene duplication occurs, particularly in the coding regions, and how it effects the protein sequences and structures.

Here, we perform the cumulative analysis of repetitiveness in amino acid occurrence in human proteins. We demonstrate that the iterative use of the same type of amino acid is a general feature for tissue-specific proteins. The frequency of gene duplication on scales of various lengths which can keep the repetitiveness at the poised state was estimated. Based on the simulative analysis, we show that such local scale duplication enhances the chance of occurrence of densely charged peptide segments, and, therefore, motifs such as heparin binding motifs and G-protein activating motifs, suggesting that repetitiveness enhances the chance for protein interaction occurrence.

Results and Discussion

Local repetitiveness in amino acid occurrence in human proteins

Our previous findings concerning the usage of amino acids R and K in correlation with the G + C content of the corresponding genome DNA, raised the possibility that local unevenness of genome DNA structure may be related to the tendency for repetitiveness of the genome (Nishizawa & Nishizawa, 1998). It has also been reported that eukaryotes, but not the prokaryote genome, have a tendency for repetitiveness (Tautz *et al*., 1986). We performed a cumulative analysis on yeast proteins and found that yeast amino acid sequences have repetitiveness (K.N., unpublished data). In the present study, we focus on human proteins, because of the relative ease in obtaining tissue-specific (thus, modern) and ancient protein sequences (Doolittle *et al*., 1986). Thus, human protein files were compiled from SwissProt, and cumulative analyses regarding the amino acid occurrence were performed. Figure 1 shows the frequency of each amino acid type at position +1 (left) and the frequency averaged over positions $+1, \ldots, +10$ from different types of residues (right). As will be defined in the Methods and Algorithms, F_X denotes the frequency of amino acid X over all proteins concerned. F_{XY} denotes the frequency of X in the proximity of amino acid Y. (X and Y are any of the 20 amino acid residues.) Figure 1 shows the percentage change from F_X, i.e. $100(F_{XY} - F_X)/F_X$. One primary feature is that for most of the 20 residues, there is a tendency of recurrence or ''self-clustering''. In fact, all the scores on the diagonal are positive, indicating that, near amino acid X, X itself tends to occur more frequently than average. Such tendency is especially strong for Q, E, S, H, R, K, A, P, G, Y and W residues. For C (cysteine), a more detailed analysis showed that $F_{CC}(i)$ (cysteine-cysteine) and $F_{HC}(i)$ (histidine-cysteine) profiles have clear peaks, likely due to the presence of many zinc finger proteins in the protein set (not shown). The tendency for recurrence appears to be modest for some amino acid residues, including T, M and V.

Figure 2(a) shows $(1/100)(\Sigma_X F_{XX}(i))$, which means the frequency of the amino acid recurrence, which were averaged over 20 amino acid types after weighting their frequency in the analyzed proteins, as described in Methods and Algorithms. As expected, the primary trend is that the same type amino acids tend to recur in close proximity at higher frequency, with the frequency gradually decreasing as the distance from the position increases. There are three clear peaks at 28, 56, and 84. Further analyses revealed that these peaks are created by zinc finger motifs: when those files whose description annotates on ''zinc finger'' were eliminated, the peaks were not observed (Figure 2(b)), while those files annotated as such gave rise to the peaks as shown in Figure 2(a).

However, it should be emphasized that even the protein set without zinc finger proteins shows a smooth curve, implying the tendency of amino acid recurrence at a local scale ($1 \sim 20$ residues) (Figure 2(b)). $F_{XX}(i)$ (recurrence profiles) of Q, E, S, H, R, K, A and P (Figure 3) and also of G, Y and L (not shown) for non-zinc finger proteins indicate the general tendency for repetitiveness. We conclude that human proteins have a general tendency for repetitive use of amino acids. Such tendency is also the case for other mammalian proteins and in *Saccharomyces cerevisiae* and *Caenorhabditis elegans* (not shown).

Strong repetitiveness in modern proteins

Given the previous findings on genomic instability mediated by gene duplication (Ohno, 1984, 1987; Weber & Wong, 1993; Goldstein *et al*., 1995; Rubinstein *et al*., 1995; Kimmel *et al*., 1996), it seems possible that the gene duplication on a very local scale is a factor generating the repetitiveness shown above. If it is so, then the tendency for repetitiveness may differ among proteins, because in general the modern proteins are under weak constraints, while the ancient proteins are under strong constraints so only limited patterns of mutation are accepted (Doolittle *et al*., 1986). From the original human protein sets, we chose those proteins which can be categorized into either of the following three subgroups, (A) neural system-specific, (B) immune system-specific, and (C) the proteins whose homologues are known in *Escherichia coli*. The averaged frequency of recurrence (1/

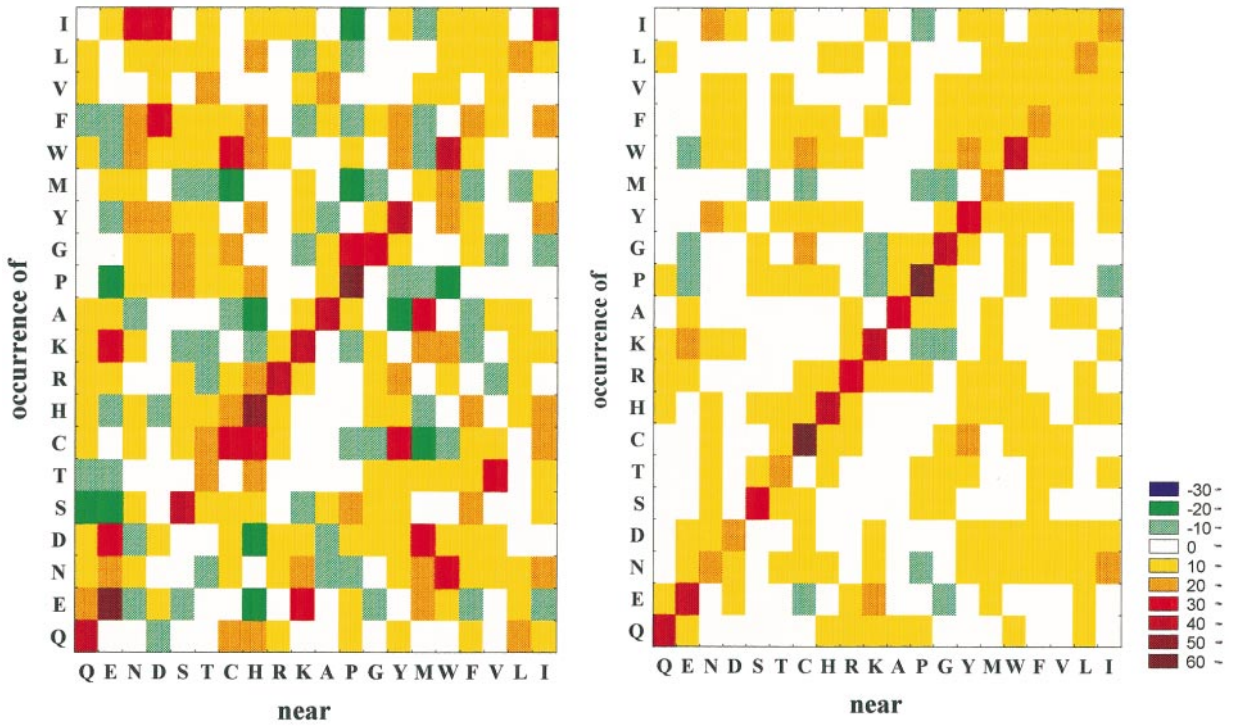

Figure 1. Amino acid occurrence near different types of amino acids: left, at the +1 position; right, over $+1, \ldots, +10$ positions (averaged). The percentage change from F_X, that is, $100(F_{XY} - F_X)/F_X$, is shown with the color indicated.

$100)(\Sigma_X\ F_{XX}(i))$ for each category is shown in Figure 4. Neural and immune system-specific proteins have higher degree of repetitiveness in amino acid sequences compared with ''ancient'' proteins. These observations seem consistent with our idea that the degree of repetitiveness reciprocally correlates with the strength of structural and functional constraints on protein sequences. One notable feature of the neural proteins, in particular, is that they have a repetitiveness score (~1.15 or above) significantly higher than 1.0 even at far distal locations. Although we have not investigated the cause as yet, one possibility may be that amino acid composition is different among the neural proteins, because some (e.g. ion channels) contain many hydrophobic residues, while some are known to be very acidic (e.g. GAP-43, neurofilaments etc.). While we believe this is an important issue, we would like to focus on more local repetitiveness in the current study. The genes for the neural proteins have a high G + C content (see below). However, it is presently unclear to what extent the amino acid repetitiveness at this ''intermediate distance (50-100 residues)'' is resulting from such a biased composition in DNA.

Mitigated repetitiveness in subsegments homologous with yeast proteins

These findings suggest the reciprocal relationship between the ''constraints'' on the protein sequences and the ''repetitiveness''. To further address this point, we also tested the human proteins for which yeast (*S. cerevisiae*) but not *E. coli* has the homologue. Because the overall ''homology'' between two proteins generally depends on the contribution from subsegments with different degrees of homology, we analyzed the repetitiveness of the two distinct populations of the subsegments derived from an identical protein set, with different degree of homology: those subsegments which can be aligned with the BLAST score (*e*-value) < *e*-40 (''homologous subsegments''), and those subsegments for which the BLAST analysis did not show the score smaller than *e*-40 (''non-homologous subsegments''). (The names of the files and the positions of the homologous subsegments are shown in Supplementary Material). As shown in Figure 5, while the non-homologous subsegments have a high degree of repetitiveness, the homologous subsegments have only a modest level of repetitiveness.

We also found that those yeast proteins for which *E. coli* has a homologue exhibit very weak repetitiveness, but those unique to yeast show strong repetitiveness (unpublished observations). Based on these results, we surmise that the tendency of repetitiveness of genome is the factor that causes the repetitive amino acid use, and that when the constraints on proteins are weak, such tendency becomes clear.

Repetitiveness in DNA sequences

Although the results presented above were obtained solely from the study on amino acid

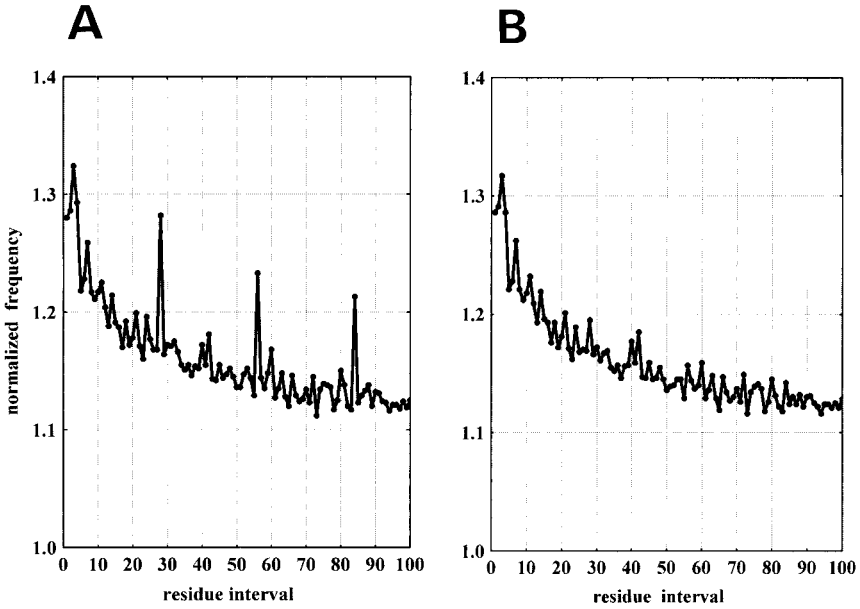

Figure 2. (a) Normalized frequency of recurrent use of identical amino acids in human proteins. To simplify the presentation, the profiles for 20 amino acid types obtained were combined as shown by $(1/100)(\Sigma_X F_{XX}(i))$. (b) Normalized frequency analyzed as in (a), but on the protein set from which zinc finger proteins were removed.

sequences, analyses of DNA sequences could also be informative. Here, let us classify the repetitiveness in DNA sequences into two categories, namely cryptic and non-cryptic (or ''3*n*'') types. For example, the nine nucleotide sequence *GA*T AAC *GA*C (encoding amino acid residues DND) have recurrent dinucleotide ''GA'' at the +6 interval, but this recurrence is not cryptic because it is

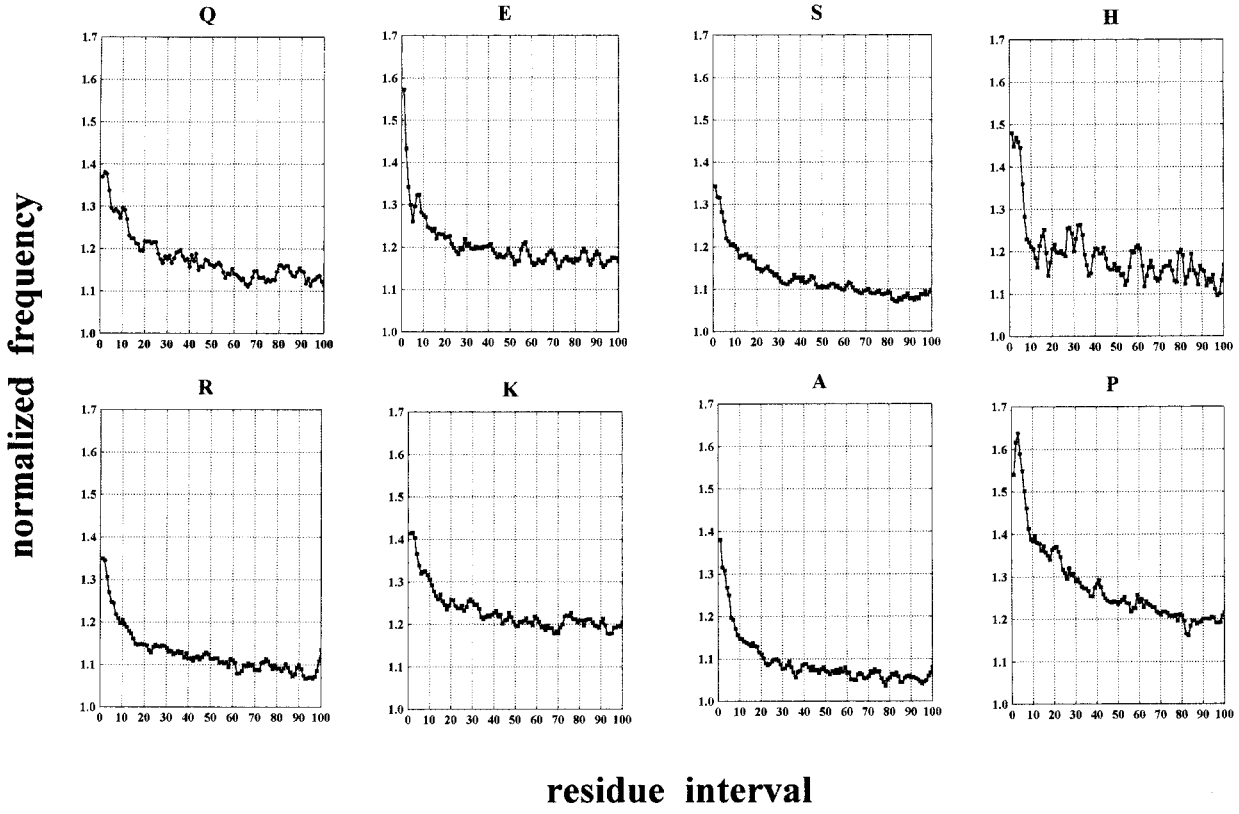

Figure 3. Recurrent use of individual amino acid types. $F_{XX}(i)/F_X$ is shown for indicated amino acid types. Analysis was performed on no-zinc finger human proteins.

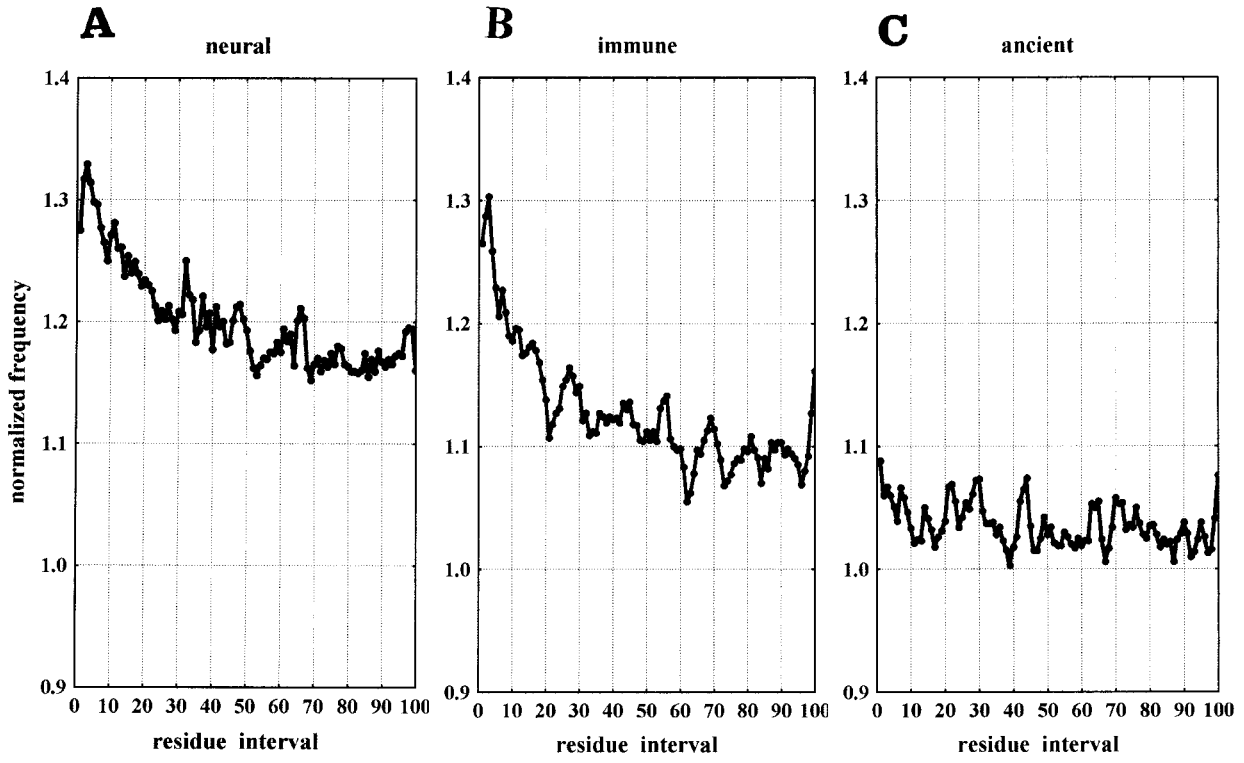

Figure 4. Recurrent use of identical amino acids in classified human proteins. The combined repetitiveness profiles $(1/100)(\Sigma_X F_{XX}(i))$ are shown, as in Figure 1. (a) Neural system specific protein (b) immune system-specific proteins; and (c) those human proteins for which *E. coli* has homologue(s).

directly linked to the repetitiveness in the amino acid sequence encoded. In fact, it has previously shown that coding DNA sequences tend to give rise to peaks at 3*n* (3, 6, 9 . . .) positions (Tsonis *et al.*, 1991). The other type is ''cryptic'' repetitiveness: for example, in the sequence *GA*T AA*G A*CC (encoding DKT), the dinucleotide GA recurs at +5 interval, so this repetitiveness is cryptic and not directly linked to that of amino acid. (Of course, there should be some constraints from protein sequences: in the above example, the second GA contains an A residue as the first nucleotide of the codon ACC and is likely to be influenced by the selection pressure on the encoded amino acid.) One reasonable approach here is to perform ''cumulative'' analyses to examine if such cryptic repetitiveness is present on a local scale as found in amino acid sequences.

Hence, we performed the analyses of repetitiveness of dinucleotide segments at 3*n* (3, 6, 9 . . .) and non-3*n* intervals (1, 2, 4, 5, 7 . . .), using the scoring method described in Methods and Algorithms. The repetitiveness at 3*n* intervals of the cDNA sequences for the modern (neural and immune) proteins was higher than that of the cDNA sequences for the ancient proteins (Figure 6(a), circles with a score above ~0.15). However, let us emphasize the presence of repetitiveness at two different scales; the local repetitiveness (at 1-50 nt intervals) and the more global correlation (affecting the height of the profile curve at 100-300 nt intervals), which we refer to as the intermediate scale repetitiveness. It is clear here that the cDNA sequences for the modern proteins have a higher level of intermediate scale repetitiveness, compared with the cDNAs for ancient proteins, in both terms of cryptic (<0.1) and non-cryptic (>0.15) repetitiveness. Regarding the local scale 3*n* repetitiveness, it appears that the difference between the modern and ancient sequences are small, compared with the difference in amino acid repetitiveness (cf. Figure 4). This may be due to the freedom allowed to the third position of codons, which may allow some local repetitiveness in the ancient genes as well.

The modern cDNA sequences also have cryptic (non-3*n*) repetitiveness (the filled circles below 0.1 in Figure 6(a)), which is stronger locally, with the scores for the proximal intervals being significantly higher than the average score for 100-300 intervals (see also the legend to Figure 6). Such tendency was weak but statistically significant for the cDNA sequences for the ancient proteins (see the open circles below 0.1). For the ancient proteins, the average of the scores at the intervals of 2-30 nt was −0.0114 and significantly higher than the −0.0254 values for the average at 100-300 nt intervals (with $P < 0.01$, t-test). We suggest that such cryptic repetitiveness in modern cDNA sequences is largely not linked to any features in amino acid sequences *per se*, because such repetitiveness was weak in the artificial DNA sequences whose codons were shuffled such that they encode the same proteins and have the same codon composition as that

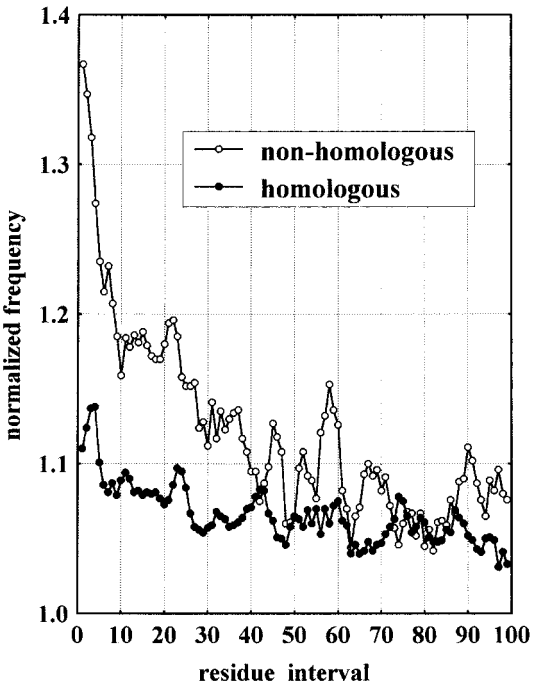

Figure 5. Recurrent use of identical amino acids in human protein subsegments classified based on the BLAST search against yeast proteins. The combined repetitiveness profiles $(1/100)(\Sigma_X\ F_{XX}(i))$ are shown. Filled circle, homologous subsegments; open circle, ''non-homologous subsegments''.

found in the original protein (cDNA) set. (Figure 6(b), triangles with scores below 0.1 for an example of codon shuffling, see the legend to Figure 6). Therefore, coding DNA sequences seem to have repetitiveness as their own propensity. This result is consistent with the idea that such propensity of genome DNA sequences is involved in and possibly contributing to repetitiveness in protein sequences.

One interesting feature in Figure 6(b) is that the intermediate scale $3n$ repetitiveness was considerably mitigated in the shuffled (artificial) sequences, as shown by the downward shift of the profile curve. A similar mitigation was observed for both the modern and the ancient genes (Figure 6(b) and data not shown). Intriguingly, when the codon shuffling was performed within each cDNA sequence, such mitigation was not clear (data not shown). This finding indicates that codon usage is similar ''within the gene'' but different ''between the genes''. This finding seems relevant to the previous findings regarding isochore (see Introduction).

Figure 6(a) shows that the cDNA sequences for the neural proteins have a higher level correlation (both $3n$ and cryptic) at the intermediate scale (~100-300 nt) than those for the ancient proteins. We surmise that the correlation at this scale is partly because of the biased nucleotide composition of the neural genes. In fact, the G + C content at the third position of codons are 63.2 % for the neural and 55.1 % for the ancient genes.

At present, genomic sequences for the proteins concerned are largely unavailable. One important thing we did not take into account in the above analyses was the presence of introns. While our preliminary simulation using the introns with representative lengths suggests that the presence of exon-intron boundaries do not considerably affect the local repetitiveness (< ~30 nt, with influence of <2 % changes in the score), we acknowledge that the intermediate scale repetitiveness should be regarded as an ''under-estimation'' of longitudinal correlation due to the disruption of correlation by introns. Although we believe that the consistency regarding the levels of repetitiveness (or correlation) for each cDNA set (Figure 6(a)) is noteworthy, more detailed analyses on genomic sequences are necessary in the future for more rigorous evaluation of the DNA repetitiveness.

Random artificial sequences and simulative duplications of segments

The precise cause of the local repetitiveness in tissue-specific proteins is unclear. Yet, it seems worthwhile to assume that gene duplication on a local scale is the major factor, and using artificial amino acid sequences, to estimate the frequency of duplications that could maintain the repetitiveness in the presence of point mutations. Note that random substitutive mutations tend to mitigate the repetitiveness. It seems also important to know how often very short scale duplications in contrast to long scale ones are likely to occur.

Hence, we generated artificial amino acid sequences that have realistic repetitiveness. As the first step, original artificial sequences, whose composition of amino acids is identical with that of real proteins (of the neural and immune set), were generated with the use of random number generator. Such sequences are naturally without any observable level of repetitiveness (Figure 7(a) and (b), broken lines). Repetitiveness was introduced into the original sequences using the duplication method to obtain the realistic repetitiveness (see Methods and Algorithms). The resultant repetitiveness is also shown (Figure 7(a) and (b)). To save space, the numbers of duplications introduced are partly shown in the legend to Figure 7. Although in Figure 7 repetitiveness of each protein set was mimicked without distinguishing 20 amino acid residues, we also found that individual simulation is possible, where repetitiveness with respect to individual amino acid types was mimicked (data not shown).

Simulation to maintain the repetitiveness in the presence of point mutations

Starting from the sequences in which repetitiveness had been introduced with the duplication method, we next estimated the relative frequency of

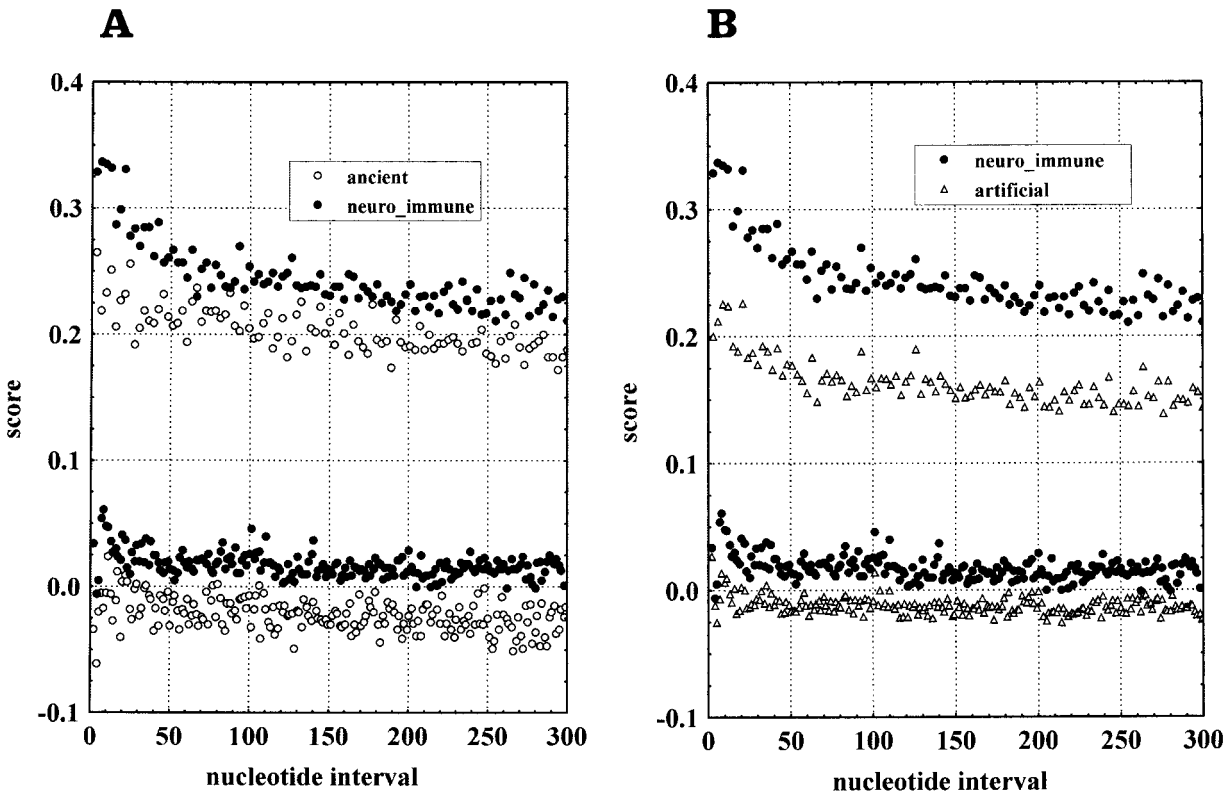

Figure 6. Repetitiveness in cDNA sequences. Repetitiveness at $\geqslant +2$ intervals is shown. (a) Repetitiveness at $3n$ (multiples of 3) nt intervals and non-$3n$ intervals in the cDNA sequences for the modern proteins (=neural + immune, filled circles) and the ancient proteins (open circles) as analyzed by dinucleotide-scanning method. The points above 0.15 represent the $3n$ (non-cryptic) intervals. (b) Repetitiveness ($3n$ and non-$3n$) in the cDNA in the codon-shuffled cDNA sequences (shown with triangles), which encode the same protein sequences and have the same codon frequency and nucletide compositions as those of the real cDNA sequences for modern proteins. For example, the ''codon shuffling'' in the set containing two cDNA sequences 1, CAG CAG CCG CCG (encoding QQPP) and 2, ATG CAA CAA CCA (encoding MQQP) may result in 1′, CAG CAA CCG CCA (still encoding QQPP) and 2′, ATG CAG CAA CCG (still MQQP), where the total codon composition was unchaged. Result for the cDNA for the modern proteins is also shown with filled circles. For the cryptic repetitiveness in the cDNA for the modern proteins, the average (±S.D.) over (+2)-(+30) positions was 0.0297(±0.017) and that over (+100)-(+300) was 0.0145(±0.0077). These two averages are significantly different (t-test, $P < 0.0005$), implying that the repetitiveness is high at the local scale. For the shuffled cDNA sequences, the corresponding values are −0.00434(±0.012) and −0.01325(±0.00548). The statistical difference between them is less significant than that for the modern cDNA (t-test, $0.001 < P < 0.01$).

duplication of various lengths that can maintain the repetitiveness in the presence of point mutations. First, we introduced point mutations according to the PAM (accepted point mutations) matrix (Dayhoff *et al.*, 1978). Note that ''one PAM'' represents one amino acid substitution per 100 residues. Random point mutations mitigate the repetitiveness and the repetitiveness profile tends to be flattened out (Figure 7(c)). Even the mutation of 10PAM causes a ∼15-20 % reduction of overall repetitiveness score. This seems noteworthy, because mutation rates in real proteins are fairly rapid: even cytochrome oxidase *c*2, an example of slowly evolving protein, has a 10 % difference in its amino acid sequence between human and pig (Nei, 1987).

We next estimated the frequency of duplication of peptide segment which can balance out the ''negative effect'' of point mutations. Table 1 shows, to maintain the repetitiveness seen in the neural system proteins, 1.7×10^3 events of duplications per 10^6 residue sequence are needed to balance against one PAM mutation. As expected, most of the duplications are those of short (one to two codons) fragments. Similarly, the repetitiveness of immune system proteins appears to be maintained by 2.4×10^3 duplication events per 10^6 residues during the time period of one PAM. It is suggested that, compared with the case of neural system proteins, more short scale duplications and less long scale duplications are occurring for immune system proteins. In theory, there are two types of duplications: (A) duplication accompanied by concomitant elongation as represented by x_1, x_2, x_2, x_3, x_4, generated from x_1, x_2, x_3, x_4, (where x_i ($i = 1, \ldots, 4$) is any of 20 amino acid residues, and x_2 is the duplicated residue), and (B) duplication accompanied by replacement of neighbor residue(s) represented by x_1, x_2, x_2, x_4, or x_2, x_2, x_3, x_4, generated from x_1, x_2, x_3, x_4. The results shown in Table 1 are based on type (A). When we assumed

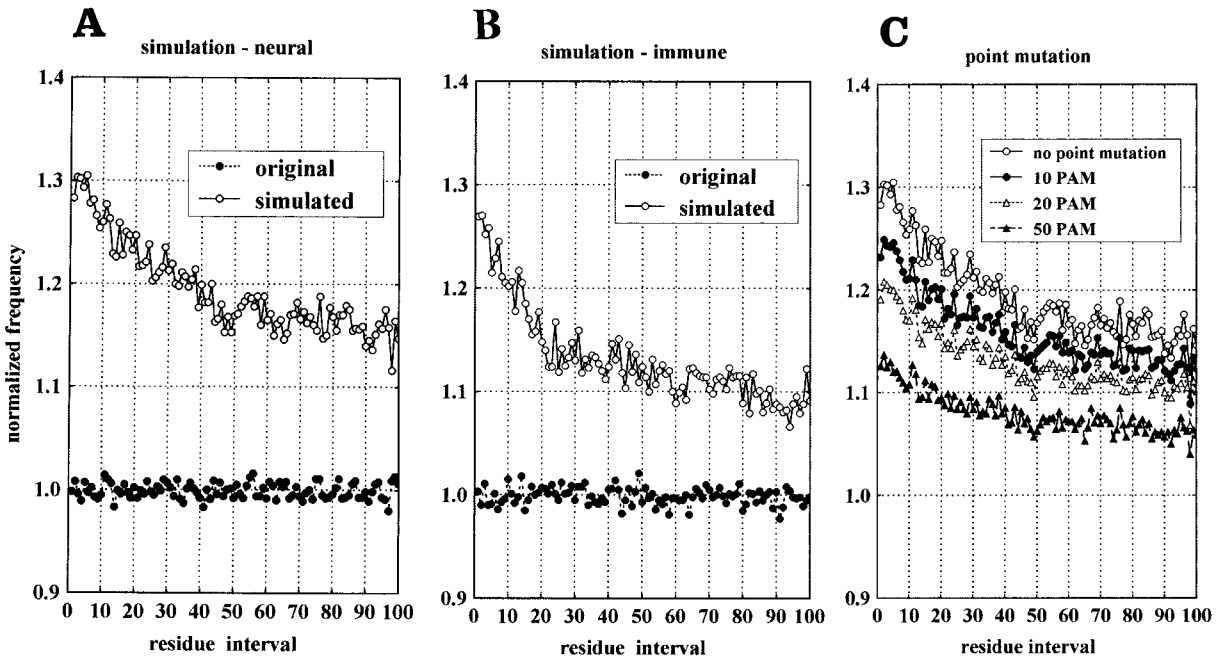

Figure 7. Simulation of repetitiveness using amino acid sequences. (a) Repetitiveness of the artificial sequences before (broken line) and after (continuous line) the simulation of repetitiveness mimicking the repetitiveness of neural proteins. Repetitiveness was introduced by the duplication method (see Methods and Algorithms). The number of duplication events needed (per 2.5×10^5 original residues) are partly shown below in the parenthesis following the number indicating segment length of duplication; 1 (6077); 2 (3282); 3 (2294); 4 (1673); 5 (1419); 10 (596); 20 (218); 30 (131); 40 (96); 50 (71); and 100 (32). (b) Combined repetitiveness profile of the artificial sequences before and after mimicking the immune system proteins. The number of duplication needed was: 1 (4961); 2 (2716); 3 (1487);,4 (1206); 5 (879); 10 (329); 20 (124); 30 (80); 40 (49); 50 (34); and 100 (15). (c) Effect of point mutations on the repetitiveness of artificial sequence mimicking that of neural proteins. Indicated numbers of mutations were introduced to the sequence, whose repetitiveness profile was shown as the top curve with open circles.

that all the duplication occurs like type (B), then 1.2×10^3 and 1.7×10^3 events were estimated for neural and immune proteins, respectively. In both cases, the frequency of estimated duplication appears substantially high ($1.2 \sim 2.4$ events per 1000 residues), given that one PAM indicates ten substitutive mutation events for segment of the same length. These results also show that duplications must occur more frequently on a very local scale like ~20 bp rather than $100 \sim 300$ bp.

It should be noted that we introduced several assumptions in this simulation using amino acid sequences. First, we assumed that tandem gene duplication is the sole mechanism for the enhancement of repetitiveness, disregarding other factors such as a local bias in point mutation probability. Second, we did not simulate the deletion and random insertion of the segments. Therefore, our duplication should be regarded as the ''net'' duplication, by which all the mutations (including deletion and random insertion) were taken into account. Third, we assumed that the overall repetitiveness is constant for an evolutionary period of at least ~20 PAM. Notwithstanding these assumptions, we believe that the figures are useful as an initial step for understanding the general feature of evolutionary dynamics of protein sequences.

Simulation using DNA sequences

The simulations presented above were concerned with amino acid level analyses. To estimate the effect of DNA mutation and to infer their relative strength, we also performed simulation by which DNA sequences were directly manipulated. First, the real cDNA sequences were modified by substitutive mutations according to the near-neutral mutation rates obtained by Li *et al.* (1984) using pseudogenes: this procedure mimics the evolution without any selection pressure. The substitutive mutations (of 20% of the amino acids and 10% of the nucleotides) under this condition resulted in quick mitigation of repetitiveness in both protein (Figure 8(a), left) and DNA sequences (right).

Next, more realistic conditions were examined, where mutations were partially ''accepted'' such that the resultant amino acid substitutions become similar to the Dayhoff matrix with the procedure described in Methods and Algorithm. Under the condition in which 20 PAM amino acid substitutions (which changes 17% of the nucleotides) were introduced, the repetitiveness in DNA sequences was weakened (Figure 8(b)). The cryptic repetitiveness in DNA sequences was also mitigated. We believe that the extent of mutations introduced here has biological relevance because

Table 1. Estimated frequency of duplication which balances with point mutation

Segment length	Neural proteins (times per 10^6 residues)	Immune proteins (times per 10^6 residues)
1 residue	377.6 ± 6.4	620.0 ± 8.4
2	201.3 ± 3.5	347.2 ± 7.2
3	139.8 ± 2.4	181.5 ± 2.9
4	111.0 ± 3.1	155.4 ± 2.7
5	89.1 ± 2.4	114.6 ± 3.0
10	37.6 ± 1.6	52.2 ± 1.3
20	18.2 ± 2.0	10.1 ± 1.7
30	10.7 ± 1.4	6.9 ± 1.2
40	5.6 ± 1.5	5.4 ± 1.3
50	4.3 ± 0.8	4.3 ± 0.9
Total of 51, 52 . . . 100	137.6 ± 3.8	111.8 ± 3.0
Total (for 1, . . . ,100 residues)	1.7×10^3	2.4×10^3

For each of the neural and immune protein sets, the average of five experiments is shown with S.D.

evolution of neural and immune proteins are fairly rapid: even for the homologous subsegments (that are the segments found with BLAST), only ~44% amino acid residues (mean) were identical between *Drosophila melanogaster* and human (our unpublished result). (Notably, both *D. melanogaster* and *C. elegans* homologues have comparable degree of repetitiveness with human homologues.)

Finally the evolution under the stronger constraints was mimicked, where DNA mutations were accepted only when the mutations were ''synonymous'' (thus not leading to amino acid alteration), with 10% of the nucleotides being changed. Even under this condition, the mitigating effect on repetitiveness was clear (Figure 8(c)). Taken together, biological levels of substitutive DNA mutations are likely to lead to the fairly rapid reduction in the repetitiveness in both DNA and protein sequences.

We also examined insertions and deletions (indels) using the frequency in pseudogenes as previously described (Ophir & Graur, 1997). However, indels which cause frameshift seem negligible for our purpose, because the frequency of insertions (respectively deletions) is equivalent to only 1% (respectively 2.5%) of that of substitutions (Ophir & Graur, 1997) and, moreover, indels appear to be very rarely accepted in human and rodents coding DNA sequences (Ophir & Graur, 1997; Gu & Li, 1995). Therefore, in the current study, we focused only on the indels of 3*n* (multiples of three) nucleotide long segments. Even under the condition where all the indels (with random component) were introduced at the near-neutral rate, indels of 3*n* nucleotide long segments did not effectively mitigate the repetitiveness of amino acid and DNA sequences: when five events of 3*n* indels per 1 kb sequences (at a typical probability distribution over different scales), which would occur during the period equivalent to the substitutive mutations of ~10% nucleotides (assuming the neutral frequency), were introduced, then only a slight decrease (<0.3%) in DNA repetitiveness was observed (our unpublished observation). Thus, the random indels at biological levels are not likely to have considerable effects on repetitiveness in sequences.

From the above analyses, it seemed likely that longitudinal duplications of DNA segment could generally have positive effects on the repetitiveness. To our knowledge, however, there has been no detailed analysis of the neutral frequency of duplications. Therefore, we implemented the system by which various rates of (both neighbor-replacing and elongating) duplications can be examined independently: we assumed, for example, that in the sequence GGA*AT*TCC, the neighbor-replacing duplication of AT generates G*ATAT*TCC or GGA*ATAT*C with an equal probability, while the elongating one results in GGA*ATAT*TCC. Regarding the elongating type, duplications of segments of non-3*n* lengths, which cause frameshift, are not likely to occur at a significant level for the proteins concerned (due to the same reason as considered in insertions and deletions). Thus, we simulated only the neighbor-replacing duplications and those elongating duplications of the 3*n* length segments. Regarding the neighbor-replacing type, we further discriminate the in-frame and out-of-frame type of neighbor-replacing duplications: for example, in a 12 nt sequence GAT C<u>TT A</u>GA GCA (encoding DLRA), the neighbor-replacing duplication of <u>TTA</u> results in GAT C<u>TT ATT A</u>CA or G<u>TT ATT A</u>GA GCA, changing two of the encoded amino acids (DLRA → DLIT or VIRA). Although this does not cause the frameshift downstream of this segment, we define that the <u>TTA</u> is not in-frame, but out-of-frame. (Note that the (neighbor-replacing) duplications of segments of non-3*n* nucleotide lengths always generate a new codon in the proximity, so can be regarded as out-of-frame type.)

Starting from the real protein and cDNA sequences, we introduced the substitutive mutations (resulting in 20 PAM amino acid substitutions) and obtained sequences with mitigated repetitiveness (as shown with filled circles in Figure 9(a)). We tried to find the type(s) of duplications which can bring the repetitiveness profiles back to the original level. When only the 3*n*, in-frame type was allowed, it rapidly augmented the repetitiveness in the amino acid and (rather slowly) that of DNA sequences at 3*n* intervals, but it did not considerably changed the cryptic DNA repetitiveness (Figure 9(a)). When the 3*n*, in-frame type duplications mixed with far less frequent non-3*n* and 3*n*, out-of-frame types duplications were introduced, the cryptic repetitiveness was slightly augmented (Figure 9(b)). When non-3*n* type and 3*n*, out-of-frame type duplications were introduced at a similar frequency to that of 3*n*, in-frame, they rather mitigated the amino acid and DNA repetitiveness at 3*n* intervals, while they augmented the cryptic repetitiveness (data not shown). These data suggest that 3*n*, in-frame type duplication is a

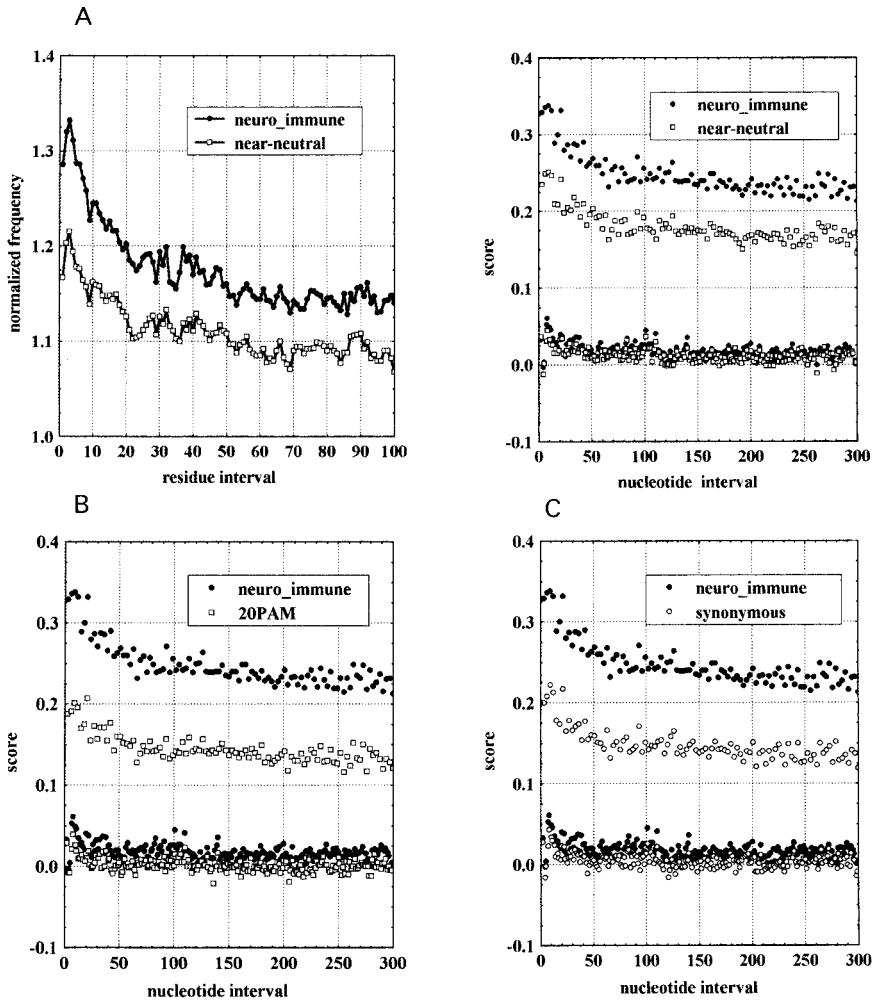

Figure 8. Simulation of substitutive mutations and effect of duplications using DNA sequences. (a) Repetitiveness profile in protein (left) and cDNA sequences (right) for the modern proteins after introducing virtual substitutions under the near-neutral condition, which changed 20% of the amino acids and 10% of the nucleotides. Profiles are shown as in Figure 2. (b) Profiles of DNA similar to (a), but after the DNA substitutions (17%) which result in 20PAM mutations were introduced as described in the text. (c) Repetitiveness in DNA similar to (a), but after synonymous substitutions were introduced until 10% of the nucleotides were changed.

major factor for maintaining the amino acid and DNA repetitiveness, while non-3*n* and 3*n*, out-of-frame types are also necessary for maintaining the cryptic repetitiveness. We also examined the elongating type duplications, but there was no considerable level of effect on repetitiveness even under the assumption that their frequency is equal to that of insertions measured in pseudogenes (data not shown).

At present, we still cannot find the condition which coordinately augments the repetitiveness in both the amino acid and DNA (at both 3*n* and non-3*n* intervals) toward the realistic levels within a biologically relevant number of events. The primary difficulty lies in augmenting the DNA repetitiveness at 3*n* intervals, although it is quite easy to augment amino acid repetitiveness (Figure 9 and data not shown). Although we cannot rule out the possibility that more extensive search could allow us to find an appropriate condition, we would rather surmise that this difficulty could be due to those factors which we did not consider in our system: nucleotide substiutions in real coding sequences should be influenced (or constrained) by the codon frequency, nucleotide compositions at the third position of codons and dinucleotide frequency. Such aspects seem related to the within gene correlation as shown by the downward shift of the profile by the between gene codon shuffling (Figure 6(b)). In any event, the real genes appear to be enhancing the intermediate scale repetitiveness in DNA ''without'' fully enhancing that of amino acid sequences.

Although it might thus not be easy to perform a quantitative study regarding the relative frequency of the 3*n* (mainly in-frame) *versus* non-3*n* duplications, these data imply that DNA duplications, in general, can augment the local repetitiveness in amino acid and DNA sequences. It is also suggested that 3*n*, in-frame type is a major factor, but also non-3*n* and 3*n*, out-of-frame types

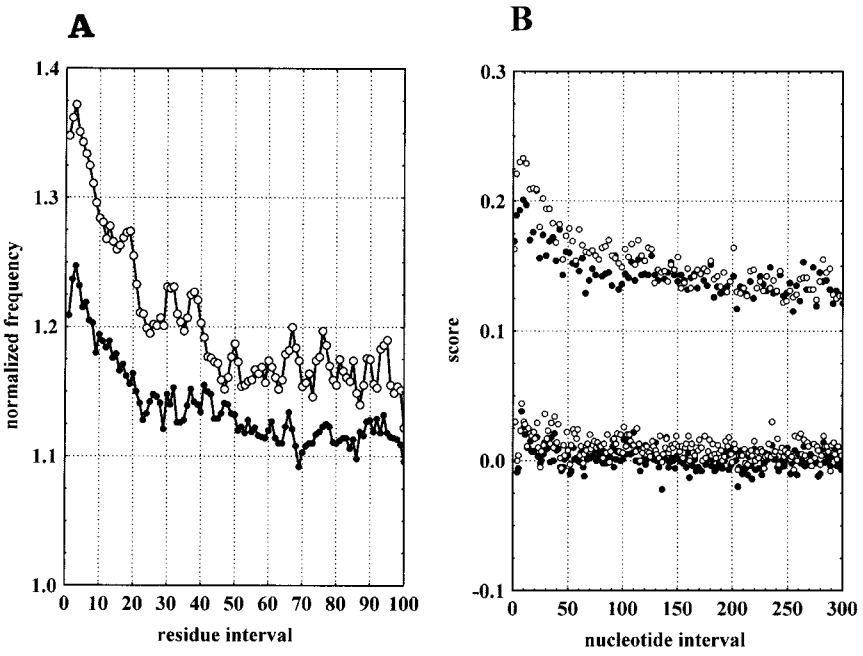

Figure 9. (a) Effect of the neighbor-replacing duplications of the 3*n* type on the repetitiveness in amino acid (left) and DNA sequences (right). Filled circles shows the repetitiveness profile of the amino acid and cDNA sequences (of the neural + immune set) whose repetitiveness was mitigated (with the method as used in Figure 8(b)). Open circles in both graphs indicate the profile after duplications of the 3*n* type were introduced to the sequences shown with the filled circles. The number of duplication events introduced (per 3.5×10^5 nucleotides) are partly shown below in the parenthesis following the number indicating segment length of duplication; 3 (520); 6 (269); 9 (190); 12 (135); 15 (109); 21 (82); 30 (60); 60 (19); 120 (12); and 300 (0). (b) Effect of the 3*n*, in-frame type mixed with less frequent non-3*n* and 3*n*, out-of-frame type duplications. Their relative frequencies are 10:1:1. The results are shown as in (a). The number of duplication events introduced (per 3.5×10^5 nucleotides) are, partly; 3 (688); 6 (396); 9 (288); 12 (198); 15 (187); 21 (90); 30 (68); 60 (35); 120 (14); and 300 (0).

are likely to be involved in maintaining the DNA repetitiveness.

Simulation of the effect of repetitiveness on properties of peptides

What is the biological or evolutionary relevance of the tendency for repetitiveness? It seems conceivable that such local dynamics could affect the variation in encoded peptide. Thus, it seems worthwhile to test how much effect such repetitiveness has on the local physicochemical nature of the local peptide. As an initial step, we examined the distribution of the pI of the 21-residue subsegments (Table 2). Notably, the occurrence of segments with extremely high (or low) pI is more frequent in the sequences with the realistic repetitiveness than in the randomly generated (but with identical composition) sequences. Estimation of the hydrophobicity using the procedure described by White (1994) suggested that, in terms of hydrophobicity as well, the repetitiveness enhances the regional diversity of the peptide segment (data not shown).

Some peptide motifs which have clusters of charged residues have been reported. We collected many motifs from PROSITE databases and the literature which employ at least one charged residue and tested their occurrence in the artificially generated sequences with and without the procedure mimicking the repetitiveness. As expected, the general trend is that the more identically charged residues the motif contains, the more frequent is the occurrence of the motif in the sequences with repetitiveness, as compared with the random sequences (Table 3). In fact, all the motifs whose occurrence have difference (>1.5-fold, marked with [b](see Table 3)) between rep^+ and rep^- sequences contain two or more identically charged amino acid (Table 3 and data not shown).

Our results suggest that compared with the sequences modified with only substitutive mutation, those modified by both substitution and duplication have higher likelihood to generate the segment of clustering similarly charged residues. It seems reasonable to surmise that duplication rate is also subject to adaptive evolution for the sake of the whole system's survival. While the intermediate and long-scale genetic rearrangements are believed to be important for evolution and help the generation of new combinations (shuffling) (Patthy, 1991; Iwabe *et al.*, 1996), local repetitiveness that we focus on may be convenient for generating a short motif consisting of similar residues. It may be helpful to imagine a situation where, for example, an occurrence of an SH3 ligand motif PPxP between pre-existing SH2 and SH3 domains would be helpful for better cellular function. In such a case, it may be easier to create a PPXP sequence by local scale duplication of the codons for P rather than ''copy (or cut) and paste'' PPxP from other genes without disrupting the SH2 and SH3. Many eukaryote motifs consist of the repetitive use of identical amino acid residues. It seems intriguing that such tendency is enhancing or regulating the chance of the occurrence of such motifs and thus, controlling the overall degree and frequency of protein-protein interactions, in eukaryotic cell.

Table 2. Number of 21-residue windows containing the peptide segment with indicated pI as estimated.

pI of 21 residues	Number of windows[a]			
	Neuro-rep$^+$	Neuro-rep$^-$	Immune-rep$^+$	Immune-rep$^-$
0.0-2.0	0	0	0	0
2.0-3.0	169	31	127	35
3.0-4.0	7980	6262	6554	5066
4.0-6.0	11,744	12,754	10,683	10,898
6.0-8.0	8042	8730	8059	8690
8.0-10.0	8751	9567	9954	11,064
10.0-11.0	8145	8175	8366	8586
11.0-12.0	3212	2894	3527	3608
12.0-13.0	1862	1577	2665	2045
13.0-	95	10	65	8

[a] A total of 50,000 independent 21-residue-long subsegments were analyzed for rep$^+$ and rep$^-$ generated sequences.

Based on the presented results, we surmise that the propensity of genome for repetitiveness tend to be clear in modern genes basically due to ''weak constraints'' on their product proteins. However, we cannot formally rule out the possibility that modern proteins ''favor'' the repetitive use of amino acids and that the tendency for repetitiveness of DNA is just ''employed'' to help the fitting process of such proteins.

In general, it is often not straightforward to discriminate positive selection from neutral drift. In our view, the possibility of positive selection of repetitive use of ''identical'' amino acids *per se* seems less plausible, because the BLAST-based classification generally show that those segments with high homology (even between human and rodent) tend to show weak repetitiveness (K.N., unpublished results). Yet, it still seems possible that the clustering of similar amino acids (in terms of physicochemical characteristics) may be beneficial to modern proteins. In this context, it still seems possible as well that local duplications of DNA (resulting in repetitive use of identical amino acids) may have been employed to achieve such clustering of similar amino acids, just because this is a rapid way. In other words, the recurrence of identical amino acids may be the byproducts produced during the evolution toward clustering of similar amino acids.

In the field of genetics, increasing number of studies are dealing with ''microsatellite'' structure, where the three to six nucleotides sequence motif expands rapidly, resulting in polymorphism. It seems possible that our findings are related to this phenomenon, and their basal origin could be the same. To build a basis for the approach toward such a question, more detailed analyses on fundamental parameters in genomic kinetics through

Table 3. The occurrence of motifs containing charged amino acid(s)

Motif	Rep$^+$ (each a.a.)	Rep$^+$ (average profile)	Rep$^-$	Reference
Heparin binding x-[RK]-[RK]-[RK]-x(2)-[RK]-x	14,749[b]	10,647[b]	5138	Fowlkes *et al.* (1997)
G-protein activating [RK]-[RK]-x(1,2)-[RK]	181,317[b]	146,418[b]	100,035	Okamoto & Nishimoto (1992)
Ubiquitous protease recognition site R-x-[KR]-R	29,983[b]	16,758	12,680	Nagahama *et al.* (1991)
Phospholipase C beta 1 activating [RK]-x-[RK]-x(3)-[RK]	87,004[b]	73,218	50,108	Piiper *et al.* (1997)
cAMP/cGMP prot. kinase phosphor. site [RK](2)-x-[ST]	94,552	78,471	73,812	PDOC00004[a]
Amidation site x-G-[RK]-[RK]	44,946	41,884	34,351	PDOC00009
N-myristoylation site G-{EDRKHPFYW}-x(2)-[STAGCN]-{P}	672,164	616,929	587,163	PDOC00008
ATP/GTP binding site motif A [AG]-x(4)-G-K-[ST]	3796	4049	3363	PDOC00017
N-glycosylation site N-{P}-[ST]-{P}	265,388	213,253	237,273	PDOC00001
Protein kinase C phosphorylation site [ST]-x-[RK]	782,888	662,283	725,937	PDOC00005
Casein kinaseII phosphorylation site [ST]-x(2)-[DE]	1368	886	1048	PDOC00006
Tyr kinase [RK]-x(2,3)-[DE]-x(2,3)-Y	32,766	28,572	31,336	PDOC00007
Protease recognition site 2 R-x(2)-R	188,345[b]	119,669	118,676	Molloy *et al.* (1992)
CDK2 cycline motif P-x-T-P-x-[RK]	2336	1493	1633	Higashi *et al.* (1995)
H(histone)1 H2b phosphorylation S-P-x-[RK]	29,993	25,138	31,649	Hill *et al.* (1990)
Guanine ring binding site N-K-x-D	4622	4957	4334	Zhong *et al.* (1995)
Sugar transporter motif K-x(2)-H-x(2)-D	2756	3282	2603	Poolman *et al.* (1995)
Neurofilament mutiphospho 1 K-S-P-x(2)	15,030	14,152	16,017	Bajaj & Miller (1997)
Sites by AMP-PK (total) [MVLI]-(R/K/H/x,x,)-x-[ST]-x(3)-[MVLI]	148,394	121,261	139,279	Weekes *et al.* (1993)
Phophate transport P-x-[DE]-x(2)-[RK]-x-[RK]	4469	3975	4242	Guerin *et al.* (1990)

[a] PROSITE entry number is shown.
[b] Indicates the difference greater than 1.5 fold as compared with rep$^-$.

homology and repetitiveness analyses seem to be necessary. As discussed above, it seems even possible that a significant part of ''apparently substitutive'' mutations may be cryptic neighbor-replacing duplications. It is hoped that our simple methods will be extended to more detailed investigation, for example, of homologous segments and pseudogenes of various organisms and eventually help the understanding of evolutionary changes in genes and proteins.

Methods and Algorithms

All the computer program source codes written in ANSI C-language used in this study are available from the authors upon request.

Analysis of correlation in amino acid occurrence

We define $F_X(\%)$ as the frequency of amino acid X over all the proteins in a given set. Our first goal is to calculate the frequency (%) of amino acid Y at a given distance i from amino acid X, which we refer to as $F_{YX}(i)$. Although the formal definition is given below, a simple example may help the understanding. Consider a 20 residue sequence MRKRTHSAVKNPTKCYRKSA (the single letter code is used). This sequence has two T (threonine) residues. Thus, $F_T = 100(2/20 = 10\,\%$. If we consider the +2 position from every K (lysine) residue, i.e. are the positions shown with italics in MRKR*T*HSAVKN*P*TK-C*Y*RKS*A*, we find T, P, Y and A. Thus, the frequency of T at +2 position from K, namely $F_{TK}(2)$, is $100(1/4) = 25\,\%$. Please also note that this sample sequence has only three +3 positions from the K residues because the K residue that is the closest to the C terminus has only two downstream neighboring residues.

Let us assume that the concerned protein set consists of N sequences and that the kth protein contains L_k residues. Let $S_{k,n}$ represent the nth residue of the kth protein.

For brevity, we represent the total number of X residues over the protein sets as $M(X)$. Thus:

$$M(X) = \sum_{k=1}^{N} \sum_{n=1}^{L_k} C_{k,n}(X)$$

where $C_{k,n}(X)$ is a ''counting function'' regarding X, therefore:

$$C_{k,n}(X) = \begin{cases} 1 & \text{if} \quad S_{k,n} \text{ is X,} \\ 0 & \text{if} \quad S_{k,n} \text{ is not X.} \end{cases}$$

Then:

$$F_X = 100M(X) / \left\{ \sum_{k=1}^{N} L_k \right\}$$

Similarly, when we have many sequences, $F_{YX}(i)$ is calculated cumulatively over all of the $M(X)$ X residues (or actually their neighbor residues). We assign a number p ($p = 1, \ldots, M(X)$) to each of these X residues, and if the pth X residue belongs to the $k(p)$th protein and corresponds to the $n(p)$th residue, our interest is in the residue at $(n(p) + i)$ th position of the $k(p)$th protein. In other words, we are interested in $S_{k(p),n(p)+i}$. Hence:

$F_{YX}(i) =$ {number of Y residues at the i position from each of the X residues}/{total number of residues at th i position from each of the X residues}

$$= 100 \left\{ \sum_{p=1}^{M(X)} C_{k(p),n(p)+i}(Y) \right\} \Big/ \left\{ \sum_{Z} \sum_{p=1}^{M(X)} C_{k(p),n(p)+i}(Z) \right\}$$

where Z = either of 20 amino acid types and:

$$C_{k(p),n(p)+i}(Z) = \begin{cases} 1 & \text{if} \quad S_{k(p),n(p)+i} \text{ is Z} \\ 0 & \text{if} \quad \{n(p)+i\} > L_{k(p)} \\ & \quad \text{or} \leqslant 0 \text{ (i.e. no such position!),} \\ & \quad \text{or if } S_{k(p),n(p)+i} \text{ is not Z} \end{cases}$$

In the denominator, Σ_Z indicates summing over 20 amino acid types.

Note that in both numerator and denominator, the position, which is the i residue apart from each X residue, is considered only when such a position really exists (i.e. $1 \leqslant \{n(p) + i\} \leqslant L_{k(p)}$). Therefore, when $|i|$ is large (~100), profile $F_{YX}(i)$ tends to be obtained from fewer data samples than in the case with small $|i|$.

Because $F_{YX}(i)$ gives the frequency (% among the 20 amino acid types) of Y, at position i, it is convenient to compare it with F_Y, the overall Y frequency of occurrence in the protein set. Thus, we use $\{F_{YX}(i)/F_Y\}$ as the normalized frequency of Y at position i of X. For example, if F_A, the frequency of alanine, is 8 % and if $F_{AQ}(2)$, the frequency of alanine at +2 position from each glutamine residue, is 10 %, the normalized frequency of alanine at the positions is $10/8 = 1.25$, meaning 25 % higher than the average frequency of alanine.

In this study, we were mainly concerned with $F_{XX}(i)$, which is just the special case of $F_{YX}(i)$, where X = Y. Thus:

$F_{XX}(i) =$ {number of X residues at the i position from each of the X residues}//{total number of residues at the i position from each of the X residues}

$$= 100 \left\{ \sum_{p=1}^{M(X)} C_{k(p),n(p)+i}(X) \right\} \Big/ \left\{ \sum_{Z} \sum_{p=1}^{M(X)} C_{k(p),n(p)+i}(Z) \right\}$$

Note that $F_{XX}(i)$ is obtained with the same denominator as the one used for $F_{YX}(i)$, and therefore, indicates the contribution (% among 20 amino acids) of X at position i.

In this study, we define the normalized repetitiveness profile of amino acid X as:

$$(F_{XX}(i)/F_X)$$

This profile needs to be determined for each amino acid type. For convenience, we also used the combined profile, for which profiles of individual amino acids were averaged using the formula:

$$(1/100) \left(\sum_{X} F_{XX}(i) \right)$$

where Σ_X indicates the sum over the 20 amino acid residues.

Note that this formula equals to $\Sigma_X[(F_{XX}(i)/F_X)(0.01F_X)]$, which indicates the normalized profiles $(F_{XX}(i)/F_X)$ for each amino acid are summed after being weighted by the overall frequency $(0.01F_X)$.

Analysis of repetitiveness in DNA sequences

Local repetitiveness in DNA sequences was analyzed with a procedure similar to the one used for amino acid sequences. We first measured the repetitiveness of DNA dinucleotide by comparing them with the neighbor dinucleotides. (Trinucletide or longer segments can also be used with the same scoring.) To score the similarity between two segments, each position was individually examined and given three points for an identical and -1 for different alignment: for example, the repetitiveness of the dinucleotide AG at a distance of i nt was scored as 6 if AG is found at the position i, while it was scored as 2 $(=3+(-1))$ against AX (X is A, C or T) or YG (Y = G, C or T), and as -2 if neither position had identity. Note that such details of scoring methods are trivial and do not affect the mathematical tractability (Karlin & Altschul, 1990). Next, we cumulatively calculated the average score with respect to individual dinucleotide and position, thus obtaining the profile. For the simplicity in presentation, we combined the profiles for the individual dinucleotide types after weighting the profiles by the frequency of each dinucleotide. (The same procedure was used for $\Sigma_X F_{XX}(i)$ in amino acids analyses.)

Simulation of amino acid sequence by repetitiveness generation and substitutive mutations

Artificial random protein sequences, whose composition of each type of amino acid is identical to that of real proteins of a given set, are generated with the use of a random number generator. For convenience, the lengths of the real proteins were not mimicked: typically, 5×10^6 residue long original sequences were modified. The generated sequences, or original sequences, are naturally without any significant level of repetitiveness (Figure 7(a) and (b)). These sequences were modified such that their repetitiveness was progressively augmented toward the level of the repetitiveness of real sequences. Both of the following methods were tested. (i) Insertion method: Of the 20 amino acid types, one type to be inserted is chosen for each cycle of procedure such that the relative abundance of each type of amino acid is largely constant. For example, if glutamine (Q) was chosen, the repetitiveness profile of Q was calculated and compared with the target repetitiveness profile obtained from real protein data. If, compared with the target profile, the repetitiveness profile of the simulated sequences shows the greatest discrepancy at the +3 (three residues downstream) position from the alanine residues, then one additional Q is inserted at the +3 position from an arbitrary Q, thereby doing feedback. These procedures are iterated for all amino acid types using the appropriate proportion, such that the relative abundance of each type of amino acid is unchanged. However, the insertion method did not produce sufficient repetitiveness such as the one seen in modern proteins. In fact, all of our trials (ten times) produced insufficient levels of repetitiveness (~1.15 in terms of $(1/100)(\Sigma_X F_{XX}(i))$) even after the extensive iterations of insertions, which lead to the elongation of the sequences by fivefold. Only when amino acid X was inserted into the ''X-rich peptide region'', was sufficient repetitiveness obtained.

(ii) Duplication method: One amino acid type is chosen, and its repetitiveness is compared with the real protein profile in a similar manner as in the insertion method. If the repetitiveness of Q needs to be augmented at the +3 position, one Q is arbitrarily chosen and one of the three residue segments containing the Q is duplicated. For example, if Q in the segment ''x_1, x_2, x_3, x_4, Q, y_1, y_2, y_3, y_4,'' (where x_i, y_j, $(i, j = 1, 2, 3, 4)$ are any type of amino acid) is chosen to be duplicated, one of the segments ''x_3, x_4, Q,'', ''x_4, Q, y_1'' or ''Q, y_1, y_2'', is randomly chosen and duplicated. All the duplication events performed during one simulation experiment were recorded for analysis. With the method (ii), procedures were iterated until the difference (between real and artificial proteins) in profile $[(1/100)(\Sigma_X F_{XX}(i))]$ became sufficiently small. That is:

$$(1/100)\left\{\sum_{i=1}^{100}[(\Sigma_X F_{simXX}(i)) - (\Sigma_X F_{realXX}(i))]^2\right\}^{1/2} < 0.001$$

where $F_{simXX}(i)$ and $F_{realXX}(i)$ denote the $F_{XX}(i)$ of artificial and real sequences, respectively. Note that with both methods (i) and (ii), the procedure to enhance the repetitiveness of Q tends to reduce the repetitiveness of the other types. Of note, since the (i) insertion method did not produce equivalent level of repetitiveness, we mainly used the (ii) as will be shown in the text.

Simulation of point mutations was performed by introducing point mutations into original sequences according to mutation probability scores obtained from the data of accepted point mutation (PAM). (Figure 8 by Dayhoff *et al.*, 1978). For point mutations for a longer time period, the score for 1 PAM was used in a iterative manner as described by Dayhoff *et al.* (1978). Every position of the sequence was equally subjected to the mutation probability score matrix.

Random point mutation should naturally tend to mitigate the repetitiveness of the sequence, as long as each point mutation event occurs independently from another event. If an appropriate frequency of duplication is introduced, the mitigating effect of point mutation on repetitiveness can be canceled out. Hence, the frequency of the duplication events that maintain the repetitiveness in the presence of substitutive point mutation can be inferred, by modifying the real human protein sequences as follows: Let $P(l)$ denote the frequency of duplication of the l residue segment per 10^6 residues during the period of 1 PAM. After each mutational procedure using the matrix for 1 PAM as previously mentioned, duplication is performed according to $P(l)$. Segments to be duplicated are randomly chosen. Our goal is to find the $P(l)$ that minimized the change in the repetitiveness profile produced by n iterations of the point *m*utation and *d*uplication procedure (or ''*m-d* loops'').

Let $[\Sigma_X F_{XX}(i)]_n$ indicate the repetitiveness profile (averaged over all amino acid types) after n rounds of m-d loops. We used the following function D_n to denote the difference between the profiles obtained before and after the n m-d loops:

$$D_n(P) = \{\Sigma_i([\Sigma_X F_{XX}(i)]_0 - [\Sigma_X F_{XX}(i)]_n)^2\}^{1/2}$$

In this study, we set $n = 20$. Our goal is to minimize $D_n(P)$ by seeking the appropriate $P(l)$. Because every trial failed with $P(l)$ greater than 10,000, $1 < P(l) < 10{,}000$ was tested. After testing 10,000 different $P(l)$ randomly defined, the best one was selected and subjected to a simulated annealing procedure (Müller *et al.*, 1995) for further improvement. In brief, an elementary move from $P(l)$ to $P'(l)$ was chosen at random, and the corresponding change in $D_n(P)$, i.e. $\Delta D_n = D_n(P) - D_n(P')$, was calculated. Depending on the sign and magnitude of ΔD_n, the move was accepted with a probability:

$$\mathrm{Prb}(P \rightarrow P') = \begin{cases} 1 & \text{for} \quad \Delta D_n < 0 \\ \exp(-\Delta D_n/T) & \text{for} \quad \Delta D_n > 0 \end{cases}$$

where T is the so-called temperature factor, which controls the efficacy of the minimization process. When T is high (large), the probability that $D_n(P)$ increases (i.e. temporarily gets worse) is high. When T is small, the movement of $D_n(P)$ is stable but is more likely to be trapped in a local minimum. In this study, as T, we used the minimum value of $D_n(P)$ that we reached until the time point. To improve the probability of finding the global optimum, the whole procedure was repeated ten times using new 10,000 sets. The procedure to keep the amino acid composition stable was employed as described in (ii) above.

Simulation using DNA sequences

We generated a simulation system where DNA sequences evolve under the influence both of the near-neutral substitution frequency (see below) and the virtual constraints on product protein. First, candidates of substitution were raised according to the frequency of substitution observed in pseudogenes (Li *et al.*, 1984) at positions randomly chosen. Such candidates were accepted only when they satisfied the conditions specified by user.

Here, three biologically important conditions were tested: (i) The condition in which all the candidates of substitutive mutation are always accepted without any discrimination. (ii) The condition that partially accepts the candidates such that alterations of encoded protein sequences mimic the Dayhoff matrix. Here, candidates (of DNA substitution) were recorded in the buffer and accepted only when their acceptance would either not change the amino acid sequences or bring the cumulative statistics (regarding the balance among different patterns of amino acid substitutions) closer to 1 PAM Dayhoff matrix. By this procedure, ~95% of 1 PAM mutations were carried out. Because some patterns of (amino acid) substitutions require two or three nucleotides to be changed in a codon and thus are not realized with this method, a complementary step for such substitutions (~5% of all the mutations equivalent to 1 PAM) were artificially performed to realize 1 PAM mutations. To mimic further evolutionary changes (such as 100 PAM), the procedure for 1 PAM was appropriately iterated. In both methods (i) and (ii), those substitutions which cause premature terminations were not accepted, because, for the protein species used, homology analyses over vertebrates suggested such cases are too rare to effect the overall repetitiveness (Nishizawa, unpublished result). (iii) The condition where the candidates are accepted only if they would lead to synonymous change (not leading to amino acid mutation). This condition corresponds to stringent constraints on the encoded proteins. Our system allows us to "blend" the procedure (ii) and (iii) and thus mimic various rates of evolution (and thus virtual constraints), although we did not extensively exploit this potential in this study.

One important factor that our current system does not take into account is positional bias in mutation rates: for each type of nucleotide, all the positions are equally subjected to near-neutral substitution events in our system.

Insertions and deletions mutations were also implemented according to the characteristics as previously measured (Ophir & Graur, 1997). As in the case of substitutive mutations, the candidates raised at near-neutral rate were partially accepted with the filter controllable by the user. Insertions and deletions causing premature terminations were not accepted in this study because of their rare occurrence in the real proteins and thus their weak effect on the overall repetitiveness. Our system also allows us to differentiate the parameters for the insertions and deletions depending on whether the mutation causes frameshift or not. We also implemented simulation of duplication mutations. Both types of duplication (the neighbor-replacing type and the elongating type) were independently considered. Estimation of frequency of duplications was performed in a similar manner to that used for the amino acid sequence analyses.

Estimation of pI of protein segments

The physicochemical nature of the protein segments was analysed with respect to hydrophobicity and the isoelectric point (pI). For, the pI of the segments, [H] which brings the total charge of the protein to zero was determined in the equation:

$$Z = -\Sigma_i\{a_i * K_i/([\mathrm{H}] + K_i)\} + \Sigma_j\{b_j * [\mathrm{H}]/([\mathrm{H}] + K_j)\}$$

where Z denotes the total charge of the protein, K_i is K_a (that is drawn from the p*Ka* value shown below) of negatively charged amino acids, a_i is the number of such residues, K_j is K_a of positively charged amino acids and b_j is the number of such residues (Manabe, 1990; Bjellqvist *et al.*, 1994). Σ_i and Σ_j indicate that the summing is performed over all the negatively (i) and positively (j) charged amino acid types shown below. For the calculations of K_i and K_j, the following values of pK_a, obtained for the side-chains of free amino acids were used: COOH group of the C terminus, 2.30; NH_2 group of the N terminus, 9.60; COOH group of Asp, 3.86; that of Glu, 4.25; OH group of Tyr, 10.07; imidazole group of His, 6.0; ε-NH_2 group of Lys, 10.53; Guanidil group of Arg, 12.48 and SH group Cys, 8.33. Typically, the subsegments within the window with the width of 21 residues were analyzed taking into account the N and C termini of the subsegments.

Effect of repetitiveness on the occurrence of peptide motifs

Artificial sequences generated with procedures enhancing repetitiveness (rep^+) and without such procedures (rep^-) were screened with respect to the motifs collected in PROSITE (Bairoch *et al.*, 1997). Because many PROSITE motifs very rarely occur, we first tested all the motifs against both rep^- and rep^+ sequences (2.5×10^6 residues for each), and those 95 motifs which appeared

at least once were chosen for the measurement on the 5×10^7 residues tested. In addition, we collected from the literature several motifs, because PROSITE does not contain all the motifs published, probably because definitive formalization of the consensus pattern is difficult especially when the number of examples are limited. Medline was screened with the query words ''motif'', ''basic (or acidic)'' and ''residue (or amino acid)'', and 92 papers whose summary includes motifs were collected. We then collected all of those motifs which are shorter than 10 residues and contain at least one residue (that is, R, K, [R or K], D, E, or [D or E]), Due to our interest in charged motifs in the present study. Because of redundancy, only 23 independent motifs were obtained.

Protein and DNA sequence files and classification analysis of repetitiveness

For humans, all human files in the SwissProt version 34 were compiled, and genes for HLA proteins, immunoglobulins, T cell receptors and other highly similar sequences were culled with the aid of the BLASTP algorithm (Altschul *et al.*, 1990) to remove redundancies, thus producing 3769 proteins (1.74×10^6 residues).

Ancient, neural-system-specific and immune-system-specific proteins were compiled from SwissProt/PIR. Using the appropriate keywords (''neuron'', ''neural'', ''nerve'', ''brain'' for neuro-specific proteins and ''lymphocyte'', ''T-cell'', ''B-cell'', ''immunoglobulin'', ''immune'', and ''immuno'' for immune-system-specific proteins), relevant files were chosen. For ancient proteins, all the files for which the *E. coli* homologues are known (with BLAST score (*e*-value) $< e - 50$) were compiled. The BLASTP algorithm was used (under the standard settings without filtering the monotonous sequences) for eliminating pairs of proteins mutually related to the degree of >25% identity over any of 200-residue segments. Even if the identity was <25% for any 200-residue segment, some proteins were discarded to avoid compilation of many entries from the same family. (e.g. T-cell receptors, HLA) The files of each category are shown in Supplementary Material.

For the analysis of those human proteins for which the yeast (*S. cerevisiae*) but not *E. coli* homologues are known (Figure 5), all the yeast protein files of SwissProt were searched, in the alphabetical order, using the standard BLASTP against SwissProt + PIR + translated GenBank sequences and against the *E. coli* subgroup. Those yeast files for which the human homologues are given in the search (with the BLAST score smaller than $e - 30$) were compiled and, to further select the files whose the *E. coli* homologues are not known (or BLAST score $> e - 10$), were subsequently tested against the *E.coli* proteins . The first two hundreds human files collected in this manner were used in the analysis shown in Figure 5. (see the Supplementary Material.)

The cDNA sequences were collected with the aid of SwissProt Web site (http://www. expasy.ch/cgi-bin/sprot-search-ful) and, when necessary, tblastn (of BLAST, at http://www. ncbi.nlm.nih.gov).

Acknowledgments

We thank the anonymous referees for valuable comments.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment tool. *J. Mol. Biol.* **215**, 403-410.

Bairoch, A. P., Bucher, P. & Hofmann, K. (1997). The PROSITE database, its status in 1997. *Nucl. Acids Res.* **25**, 217-221.

Bajaj, N. P. & Miller, C. C. (1997). Phosphorylation of neurofilament heavy-chain side-arm fragments by cyclin-dependent kinase-5 and glycogen synthase kinase-3alpha in transfected cells. *J. Neurochem.* **69**, 737-743.

Bernardi, G. D. (1995). The human genome: organization and evolutionary history. *Annu. Rev. Genet.* **29**, 445-476.

Bjellqvist, B., Basse, B., Olsen, E. & Celis, J. E. (1994). Reference points for comparisons of two-dimensional maps of proteins from different human cell types defined in a pH scale where isoelectric points correlate with polypeptide compositions. *Electrophoresis,* **15**, 529-539.

Dayhoff, M. O., Schwarz, R. M. & Orcutt, B. C. (1978). A model of evolutionary change in proteins. *Atlas Protein Sequence Struct.* **5**, **(Supple 3)**, 345-352.

D'Onofrio, G., Mouchirroud, D., Aissani, B., Gautier, C. & Bernardi, G. (1991). Correlation between the compositional properties of human genes. Codon usage and amino acid composition of proteins. *J. Mol. Evol.* **32**, 504-510.

Doolittle, R. F. (1989). Redundancies in protein sequences. In *Prediction of Protein Structure and the Principles of Protein Conformation* (Fasman, G. D., ed.), pp. 599-623, Plenum, New York.

Doolittle, R. F., Feng, D. F., Johnsom, M. S. & McClure, M. A. (1986). Relationships of human protein sequences to these of other organisms. *Cold Spring Harbor Symp. Quant. Biol.* **51**, 447-455.

Fowlkes, J. L., Thrailkill, K. M., George-Nascimento, C., Rosenberg, C. K. & Serra, D. M. (1997). Heparin-binding, highly basic regions within the thyroglobulin type-1 repeat of insulin-like growth factor (IGF)-binding proteins (IGFBPs) -3, -5, and -6 inhibit IGFBP-4 degradation. *Endocrinology,* **138**, 2280-2285.

Goldstein, D. B., Linares, A. R., Cavalli-Sforza, L. L. & Feldman, M. W. (1995). An evaluation of genetic distances for use with microsatellite loci. *Genetics,* **139**, 463-471.

Gu, X. & Li, W. H. (1995). The size distribution of insertions and deletions in human and rodent pseudogenes suggests the logarithmic gap penalty for sequence alignment. *J. Mol. Evol.* **40**, 464-473.

Guerin, B., Bukusoglu, C., Rakotomanana, F. & Wohlrab, H. (1990). Mitochondrial phosphate transport. N-ethylmaleimide insensitivity correlates with absence of beef heart-like Cys42 from the Saccharomyces cerevisiae phosphate transport protein. *J. Biol. Chem.* **265**, 19736-19741.

Higashi, H., Suzuki-Takahashi, I., Taya, Y., Segawa, K., Nishimura, S. & Kitagawa, M. (1995). Differences in substrate specificity between Cdk2-cyclin A and Cdk2-cyclin E *in vitro*. *Biochem. Biophys. Res. Commun.* **216**, 520-525.

Hill, C. S., Packman, L. C. & Thomas, J. O. (1990). Phosphorylation at clustered-Ser-Pro-X-Lys/Arg- motifs in sperm-specific histones H1 and H2B. *EMBO J.* **9**, 805-813.

Ikemura, T. & Aota, S. (1988). Global variation in G + C content along vertebrate genbome DNA. Possible

correlation with chromosome band structures. *J. Mol. Biol.* **203**, 1-3.
Iwabe, N., Kuma, K.-I. & Miyata, T. (1996). Evolution of gene families and relationship with organismal evolution: rapid divergence of tissue-specific genes in the early evolution of chordates. *Mol. Biol. Evol.* **13**, 483-493.
Karlin, S. & Altschul, S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl Acad. Sci. USA,* **87**, 2264-2268.
Kimmel, M., Chakraborty, R., Stivers, D. N. & Deka, R. (1996). Dynamics of repeat polymorphisms under a forward-backward mutation model: within- and between-population variability at microsatellite loci. *Genetics,* **143**, 549-555.
Li, W.-H., Wu, C.-I. & Luo, C.-C. (1984). Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. *J. Mol. Evol.* **21**, 58-71.
Manabe, T. (1990). Electrical properties of proteins. In *Proteins: Separation, Purification and Characterization*, vol. 1, pp. 336-339, Tokyo-Kagaku Dojin, Tokyo.
Molloy, S. S., Bresnahan, P. A., Leppla, S. H., Klimpel, K. R. & Thomas, G. (1992). Human furin is a calcium-dependent serine endoprotease that recognizes the sequence Arg-X-X-Arg and efficiently cleaves anthrax toxin protective antigen. *J. Biol. Chem.* **267**, 16396-16402.
Müller, B., Reinhardt, J. & Strickland, M. T. (1995). *Neural Networks*, 2nd edit., pp. 286-288, Springer-Verlag, Berlin.
Nagahama, M., Ikemizu, J., Misumi, Y., Ikehara, Y., Murakami, K. & Nakayama, K. (1991). Evidence that differentiates between precursor cleavages at dibasic and Arg-X-Lys/Arg-Arg sites. *J. Biochem.* **110**, 806-811.
Nei, M. (1987). Genes and mutation. In *Molecular Evolutionary Genetics*, pp. 19-110, Columbia University Press, New York.
Nishizawa, M. & Nishizawa, K. (1998). Biased usages of arginines and lysines in proteins are correlated with local-scale fluctuations of G + C content of DNA sequences. *J. Mol. Evol.* **47**, 385-393.
Ohno, S. (1984). Repeats of base oligomers as the primordial coding sequences of the primeval earth and their vestiges in modern genes. *J. Mol. Evol.* **20**, 313-321.
Ohno, S. (1987). Evolution from primordial oligomeric repeats to modern coding sequences. *J. Mol. Evol.* **25**, 325-329.
Okamoto, T. & Nishimoto, I. (1992). Detection of G protein-activator regions in M4 subtype muscarinic, cholinergic, and alpha 2-adrenergic receptors based upon characteristics in primary structure. *J. Biol. Chem.* **267**, 8342-8346.
Ophir, R. & Graur, D. (1997). Patterns and rates of indel evolution in processed pseudogenes from humans and murids. *Gene,* **205**, 191-202.
Patthy, L. (1991). Modular exchange principles in proteins. *Curr. Opin. Struct. Biol.* **1**, 351-361.
Piiper, A., Stryjek-Kaminska, D., Illenberger, D., Klengel, R., Schmidt, J. M., Gierschik, P. & Zeuzem, S. (1997). Synthetic peptides containing a BXBXXXB(B) motif activate phospholipase C-beta1. *Biochem. J.* **326**, 669-674.
Poolman, B., Knol, J. & Lolkema, J. S. (1995). Kinetic analysis of lactose and proton coupling in Glu379 mutants of the lactose transport protein of *Streptococcus thermophilus*. *J. Biol. Chem.* **270**, 12995-13003.
Rubenstein, D. C., Amos, W., Leggo, J., Goodburn, S. & Jain, S., Li, S. H., Margolis, R. L., Ross, C. A. & Ferguson-Smith, M. A. (1995). Microsatellite evolution-evidence for directionality and variation in rate between species. *Nature Genet.* **10**, 337-343.
Sueoka, N. (1961). Correlation between base composition of deoxyribonucleic acid and amino acid composition of proteins. *Proc. Natl Acad. Sci. USA,* **47**, 1141-1149.
Tautz, D., Trick, M. & Dover, G. A. (1986). Cryptic simplicity in DNA is a major source of genetic variation. *Nature,* **322**, 652-656.
Tsonis, A. A., Elsner, J. B. & Tsonis, P. A. (1991). Periodicity in DNA coding sequences: implications in gene evolution. *J. Theor. Biol.* **151**, 323-331.
Weber, J. L. & Wong, C. (1993). Mutation of human short tandem repeats. *Human Mol. Genet.* **2**, 1123-1128.
Weekes, J., Ball, K. L., Caudwell, F. B. & Hardie, D. G. (1993). Specificity determinants for the AMP-activated protein kinase and its plant homologue analysed using synthetic peptides. *FEBS Letters,* **334**, 335-339.
White, S. H. (1994). Global statistics of protein sequences: Implications of the origin, evolution and prediction of structure. *Annu. Rev. Biophys. Biomol. Struct.* **23**, 407-439.
Zhong, J. M., Chen-Hwang, M. C. & Hwang, Y. W. (1995). Switching nucleotide specificity of Ha-Ras p21 by a single amino acid substitution at aspartate 119. *J. Biol. Chem.* **270**, 10002-10007.

Edited by F. E. Cohen

(Received 10 June 1999; received in revised form 4 October 1999; accepted 5 October 1999)

http://www.academicpress.com/jmb

Supplementary material for this paper comprising a list of PDB files is available from JMB Online.