Virginia Commonwealth University
VCU Scholars Compass

Psychiatry Publications | Dept. of Psychiatry

2010

Features of Recent Codon Evolution: A Comparative Polymorphism-Fixation Study

Zhongming Zhao
Virginia Commonwealth University

Cizhong Jiang
Virginia Commonwealth University

Follow this and additional works at: http://scholarscompass.vcu.edu/psych_pubs

Copyright © 2010 Zhongming Zhao and Cizhong Jiang. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from
http://scholarscompass.vcu.edu/psych_pubs/11

This Article is brought to you for free and open access by the Dept. of Psychiatry at VCU Scholars Compass. It has been accepted for inclusion in Psychiatry Publications by an authorized administrator of VCU Scholars Compass. For more information, please contact libcompass@vcu.edu.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2010, Article ID 202918, 9 pages
doi:10.1155/2010/202918

Research Article

Features of Recent Codon Evolution: A Comparative Polymorphism-Fixation Study

Zhongming Zhao[1, 2, 3] and Cizhong Jiang[4]

[1] *Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, TN 37232, USA*
[2] *Department of Psychiatry, Vanderbilt University School of Medicine, Nashville, TN 37232, USA*
[3] *Bioinformatics Resource Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University, Nashville, TN 37203, USA*
[4] *Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, VA 23298, USA*

Correspondence should be addressed to Zhongming Zhao, zhongming.zhao@vanderbilt.edu

Received 14 March 2010; Accepted 31 March 2010

Academic Editor: Momiao Xiong

Copyright © 2010 Z. Zhao and C. Jiang. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Features of amino-acid and codon changes can provide us important insights on protein evolution. So far, investigators have often examined mutation patterns at either interspecies fixed substitution or intraspecies nucleotide polymorphism level, but not both. Here, we performed a unique analysis of a combined set of intra-species polymorphisms and inter-species substitutions in human codons. Strong difference in mutational pattern was found at codon positions 1, 2, and 3 between the polymorphism and fixation data. Fixation had strong bias towards increasing the rarest codons but decreasing the most frequently used codons, suggesting that codon equilibrium has not been reached yet. We detected strong CpG effect on CG-containing codons and subsequent suppression by fixation. Finally, we detected the signature of purifying selection against A|U dinucleotides at synonymous dicodon boundaries. Overall, fixation process could effectively and quickly correct the volatile changes introduced by polymorphisms so that codon changes could be gradual and directional and that codon composition could be kept relatively stable during evolution.

1. Introduction

Proteins are typically constructed from 20 different amino acids that are encoded by triplets of adjacent nucleotides, namely, codons. Investigation of the features and trends of amino-acid or codon changes among species can provide us important insights into protein or genome evolution. The order of introducing amino acids into the genetic code was recently inferred based on the change in the frequency of amino acids between the last universal ancestor of all extant species and today [1]. More recently, Trifonov [2] reconstructed the chronological order of entering amino acids and codons into the genetic code by applying 60 different criteria. Both studies revealed that the early amino acids were among those synthesized in imitation experiments [3], which suggested their abundance in the prebiotic environment, but the late amino acids were nonexistent or rare in the prebiotic environment.

The reconstructed amino acid chronology provided a useful index for further studying the evolutionary trend of amino acid changes. Brooks and Fresco [4] found that the frequencies of Cys, Try, and Phe have increased since the last universal ancestor. Jordan et al. [5] systematically compared sets of orthologous proteins in 15 taxa and revealed a universal trend of amino acid gain and loss, that is, amino acids under declining are those entered the genetic code earliest while those under increasing are generally late coming. This universal trend was later confirmed by simulations; however, the underlining mechanism has been under debate [6, 7].

So far, most studies have relied on the comparison of amino acid frequencies among different genomes. Such comparison may not have sufficient resolution for revealing mutational mechanisms because many genetic factors such as insertions, deletions, and duplications might have had dramatic changes of sequences, and recurrent and back

mutations might occur at the same site in the course of evolution. Single nucleotide polymorphisms (SNPs) and their recently fixed substitutions provide us an alternative and unique model for reliably estimating the mutational trend in protein sequences. A systematic comparison of inter- and intraspecific mutational changes observed at codons rather than amino acids should have a fine resolution of tracing the recent trend of mutations in protein sequences. Such a systematic comparison at the codon level has rarely been performed.

For this purpose, we performed a unique comparison of the mutation patterns at the polymorphic and fixed sites in a lineage in order to identify features and trends of mutations that led to codon gain and loss. Our comparative analysis indicated a prominent strand bias in the complementary transition pairs in amino-acid coding sequences. Both polymorphism and fixation had strong bias toward G/C → A/T relative to A/T → G/C changes, leading to a decrease of GC content in coding regions in genome evolution. We examined the correlation between the frequency change of codons and their attributes such as physiochemical properties, chronology, and codon usage. Several interesting features were observed, implying for the trend of codon gain and loss in the recent evolutionary history. Our results indicated that spontaneous mutations have volatile effects on shaping both the codon and amino acid pool but the fixation process could effectively keep composition relatively stable and minimize the structural disruption of the encoded proteins.

2. Materials and Methods

2.1. SNP Data. The human SNPs were downloaded from the NCBI dbSNP database (ftp://ftp.ncbi.nih.gov/snp/, dbSNP build 126). We extracted those SNPs that were biallelic, validated, and uniquely mapped in the human genome and excluded those SNPs in the repetitive sequences using the SNP process pipeline in our previous study [8]. We retrieved the gene annotation information from the ENSEMBL database (ftp://ftp.ensembl.org/pub/, version 32.35e). Among the SNPs selected above, we found 18,368 mapped in the exonic regions.

An exonic site may be annotated in more than one transcript because of alternative splicing or multiple genes at the same location; therefore, we removed such a site when it had more than one position in codon or it was encoded in different strands. We separated the retained exonic SNPs into three groups based on their positions (position 1, 2, or 3) in the corresponding codons. Each site was counted once to calculate GC content. The GC content at position 1, 2, or 3 was denoted by GC_1, GC_2, and GC_3, respectively. When the alleles and flanking sequence of a SNP had different orientation from the corresponding cDNA sequence, the alleles and flanking sequence were reversely complemented. These SNPs were used to infer the mutation direction at the polymorphic sites.

2.2. Inference of Mutation Direction at the Polymorphic Sites. We used chimpanzee as the outgroup to infer the ancestral allele of each human SNP. The chimpanzee genome was downloaded from the NCBI database (ftp://ftp.ncbi.nih.gov/genomes/, build 1, version 1). We ran MegaBLAST [9] to map human SNP sequences to the chimpanzee genome and then inferred the ancestral allele of each SNP using the stringent criteria as in our recent study [8]. The mutation direction was inferred by the maximum parsimony principle. For example, if a SNP A/G matches nucleotide "A" in chimpanzee sequence, the mutation direction would be A → G.

2.3. Substitutions at the Fixed Sites. For consistency, we used the alignments of 7552 triplets of human-chimpanzee-mouse orthologous genes (coding regions only) prepared in Jordan et al. [5]. To estimate the nucleotide substitutions for the human lineage, we identified the sites where the mouse and the chimpanzee carried the same nucleotide (e.g., A) but the human carried a different nucleotide (e.g., G). Because the rate of nucleotide change in mammalian genomes is low ($\sim 10^{-10}$) per site per year [10]), according to the maximum parsimony principle, we could infer the direction of nucleotide substitution (A → G in this example). We excluded the sites where any insertion or deletion occurred in a lineage. Similar to the polymorphisms, these sites were separated into three groups according to their positions (position 1, 2, or 3) in the corresponding codons. The GC content at position 1, 2, or 3, denoted by GC_1, GC_2, and GC_3, was calculated correspondingly.

2.4. Nucleotide and Codon Changes. The frequency of each nucleotide change was initially calculated by

$$f_{i->j} = \frac{n_{i->j}}{\sum_i \sum_{j \neq i} n_{i->j}} \times 100\%, \quad (1)$$

where $n_{i->j}$ is the counts of nucleotide changes from the ith type to the jth type (i, j = A, C, G or T). Because of unequal nucleotide composition in sequences, the frequency of each nucleotide change was then normalized by

$$f_{i->j} = \frac{n_{i->j}/N_i}{\sum_i \sum_{j \neq i} \left(n_{i->j}/N_i\right)} \times 100\%, \quad (2)$$

where N_i is the total counts of nucleotide i in the sequences.

The normalized difference (gain and loss) of a codon was defined as the difference of the number of mutations removing the codon from that creating the codon and divided by the total number of mutations in the codon. The equation is

$$\text{Codon change} = \frac{n_+ - n_-}{n_+ + n_-}, \quad (3)$$

where $n_+(n_-)$ denotes the mutations creating (removing) the codon. The gain or loss of an amino acid was calculated similarly.

2.5. Codon Usage Data. Frequencies of codon usage in the human were obtained from the Codon Usage Database

[11], which contained 38,691,091 codons from human genes (build: NCBI-GenBank Flat File Release 156.0, http://www.kazusa.or.jp/codon/).

3. Results and Discussion

3.1. Mutational Features at the Polymorphic and Fixed Sites in Coding Regions. After removing uncertain or low quality SNPs, we obtained a total of 11,253 exonic SNPs that were biallelic and uniquely mapped in the human genome and whose ancestral alleles could be inferred from their alignment with chimpanzee genomic sequences. Separately, we identified 14,007 substitutions that occurred in the human lineage based on the alignments of human-chimpanzee-mouse orthologous gene sequences. The direction of nucleotide changes was used to examine the mutational features at the polymorphic and fixed sites as well as gain and loss of amino acids and their codons.

We first examined the polymorphisms at each position (1, 2, or 3) of a codon. We observed more than half of the SNPs located at position 3 (2803 at position 1, 2466 at position 2, and 5984 at position 3). The frequencies of transitions (changes between a purine and another purine or between a pyrimidine and another pyrimidine) were much higher than those of transversions (changes between a purine and a pyrimidine) (Figure 1). Specifically, the frequencies of polymorphisms G → A (C → T) were higher than those of A → G (T → C) at each codon position. Importantly, the frequency was notably different between A → G (e.g., 15.9% at position 1) and T → C (e.g., 11.8% at position 1) at each position; such difference could not be detected in the whole genome or in noncoding regions in our previous study [8]. This feature might reflect the strand bias in the coding regions. Indeed, we observed 26.8% of A and 17.3% of T at position 1 and similar composition bias at other two positions in the coding regions (see Supplemental Table 1 in supplementary material available online at doi:10.1155/2010/202918). The strand bias might be resulted from different functional constraints. For example, the two DNA strands of a transcribed gene are under different selection pressure due to transcriptional process.

We obtained a total of 14,007 fixed sites (2765 at position 1, 2199 at position 2, and 9043 at position 3) based on the alignments of triplets of human-chimpanzee-mouse orthologous gene sequences. Similar to the polymorphism data, the frequencies of transitions were much higher than transversions (Figure 1), and the frequencies were not symmetric for the complementary substitution pairs (e.g., G → A versus A → G).

Opposite to the similar frequency of each transversion type, we observed strong difference in the frequency of each transition type at each position, especially at position 3, when compared polymorphism and fixation data (Figure 1). At positions 1 and 2, mutation G → A occurred most frequently based at the polymorphic and fixed sites. At position 3, mutation C → T occurred most frequently at the polymorphic sties whereas T → C occurred most frequently at the fixed sties.

There was a significant excess of G/C → A/T mutations at both polymorphic and fixed sites when compared to A/T → G/C mutations (Table 1). Here, G/C → A/T denotes changes of G or C to A or T. For example, the number of G/C → A/T and A/T → G/C changes was 1443 and 991 at position 1 at the fixed sites, respectively. The difference was significant ($P = 5.1 \times 10^{-20}$) assuming that they had the same chance in a random mutation model. This observation suggested a declining GC content in coding regions, which has higher GC content than noncoding regions. These results were consistent with our previous finding of more G/C → A/T than A/T → G/C mutations at polymorphic sites in each categorized human genomic region [8]. However, another study reported a weak fixation bias favoring mutations that increase GC content in noncoding regions despite of no significant difference at the fixed sites between G/C → A/T and A/T → G/C changes [12]. Of note, the difference between the frequencies of G/C → A/T and A/T → G/C changes was much smaller at the fixed sites than the polymorphic sites (Table 1). For example, for all mutation data in codons, the frequencies of G/C → A/T and A/T → G/C changes were 47.8% and 42.8% at the fixed sites, compared to 57.6% and 30.2% at the polymorphic sites. This indicated that fixation allows less variation in GC content than polymorphism in the course of genome evolution.

3.2. Trends of Amino Acid Gain and Loss. We next examined the trend of mutation at the amino acid level. We measured gain and loss of an amino acid by the difference of the number of mutations removing the amino acid from creating the amino acid divided by the total number of mutations in the amino acid (see Materials and Methods). Supplementary Figure 1 displays the gain and loss trend for the 20 amino acids in the human lineage. There was a great variation towards gain or loss among amino acids; however, the extent of variation by the polymorphisms was remarkably stronger than that by the fixed substitutions. For example, the normalized difference in Gly was −0.13 by the polymorphisms, compared to −0.016 by the fixed substitutions. Here, a minus value indicates the trend of loss in the recent history. This suggests that many new neutral mutations would have been eliminated by genetic factors such as genetic drift and natural selection.

We examined the trend of gain and loss of amino acids by their chronology [2]. Overall, the early-coming amino acids (Gly, Ala, and Asp) tended to be lost; an opposite trend (i.e., gain), though much weaker, was observed for the late coming ones (e.g., Trp, Phe, Cys, and His) (Supplementary Figure 1). Interestingly, polymorphisms resulted in an accumulation of almost all the latest 10 amino acids except Tyr which had a small loss. This implies that the ancient amino acids have been under loss whereas the late coming amino acids have been under accumulation. This observation supports the previous studies using substitution data among more evolutionarily distant species [4, 5]. Of note, Met, a late amino acid, tended to lose by fixation. This may reflect the functional constraint on this amino acid because it is

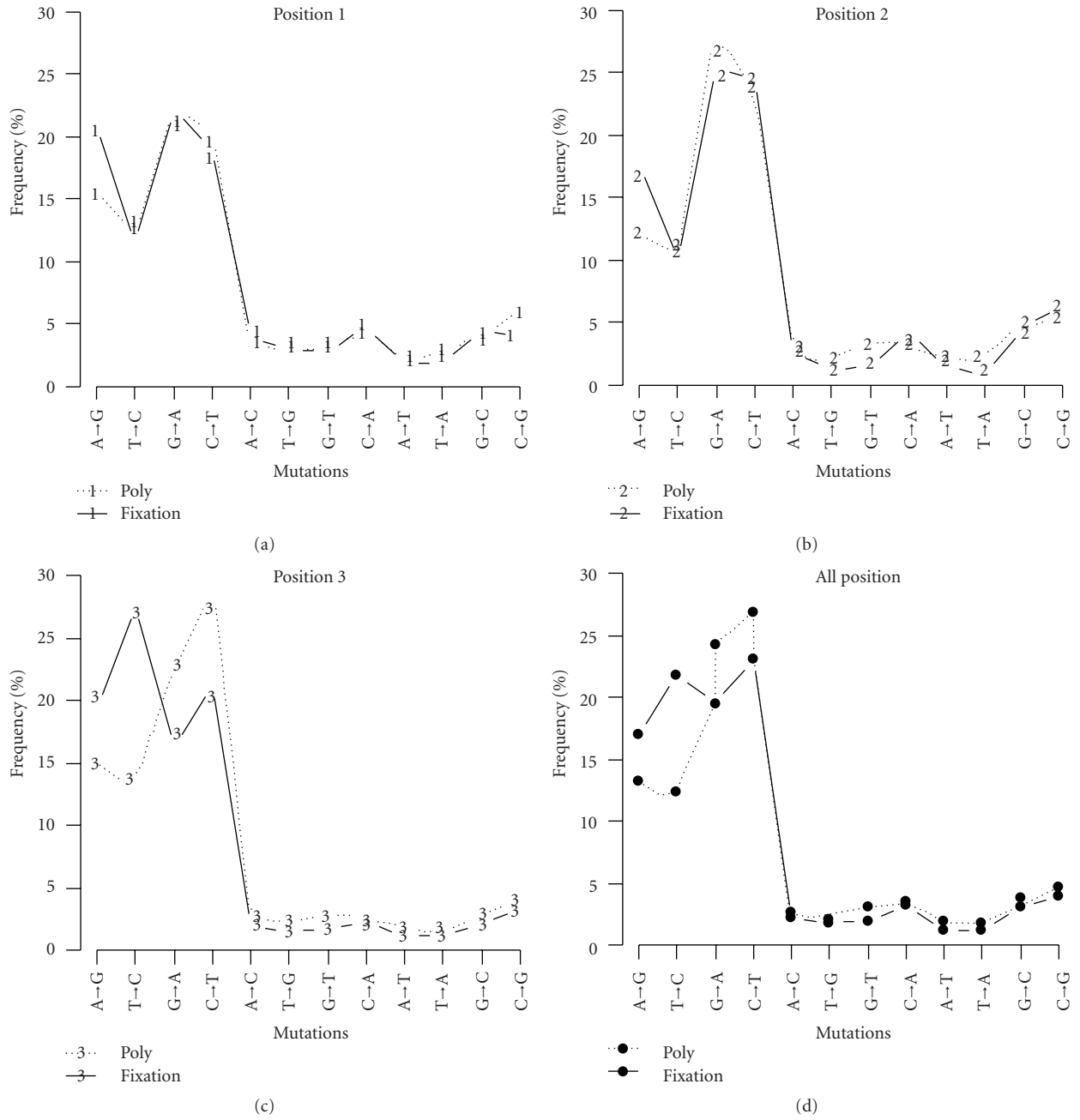


Figure 1: Normalized frequencies of nucleotide changes at the polymorphic and fixed sites in amino-acid coding regions. Figures 1(a)–1(c) show the frequencies of polymorphism (Poly) and fixation mutations at the first (a), second (b), and third (c) positions of a codon. (d) Frequency of all the mutation data. Dotted lines indicate the polymorphism and solid lines indicate the fixation.

Table 1: Comparison of mutations $G/C \rightarrow A/T$ and $A/T \rightarrow G/C$ at the polymorphic and fixed sites in coding regions.

	Position 1		Position 2		Position 3		Total	
	Poly (%)	Fixed (%)	Poly (%)	Fixed (%)	Poly (%)	Fixed (%)	Poly (%)	Fixed (%)
G/C → A/T	1547 (50.9)	1443 (48.7)	1324 (59.2)	1051 (55.4)	3791 (57.5)	4369 (41.7)	6662 (57.6)	6863 (47.8)
A/T → G/C	831 (33.7)	991 (39.3)	793 (26.6)	878 (31.4)	1593 (32.6)	3937 (50.6)	3217 (30.2)	5806 (42.8)
P-value	3.7×10^{-5}		2.0×10^{-7}		1.3×10^{-95}		1.8×10^{-90}	

The frequency of each mutation type was included in the parentheses. P-values were calculated by χ^2 test for 2×2 contingency tables at each position or for the total data. Poly: polymorphisms. Fixed: fixed substitutions.

the first amino acid in most eukaryotic proteins. Further investigation is warranted.

We performed linear regression analysis of amino acid changes with their physiochemical characters including charge, polarity, and polarity and volume using Zhang's classification [13]. No significant correlation was found, suggesting that the physiochemical attributes may not have a major effect on governing the recent amino acid composition changes.

3.3. Codon Gain or Loss with Chronology. Because most amino acids are encoded by more than one codon, a detailed examination of nucleotide substitution at the codon level should provide more insights on amino-acid compositional changes in protein evolution. We first examined the gain and loss of codons for the amino acids ordered by amino-acid chronology. There was no clear trend of loss of codons for early-coming amino acids or gain of codons for late coming amino acids; however, almost all amino acids that have multiple codons (synonymous codons) had a balancing effect by gaining some codons while losing the others (Supplemental Figure 2). For example, Gly has four codons. Both the polymorphisms and fixation resulted in two of them (GGC and GGG, early codons) to lose but the other two (GGA and GGT, late codons) to gain. This feature seems to be attributed to the different chronological orders of amino acids and codons entering the genetic code. Moreover, late coming amino acids might grab codon(s) encoding other amino acids through codon capture, which has been commonly proposed as a mechanism for adding amino acids into the code [14, 15]. For example, it was reported that AUA (Ile) also encodes methionine (Met) in yeast [16] and that AAA encodes asparagine (Asn) rather than lysine (Lys) in some animal mitochondria [17].

Therefore, we re-examined the gain and loss of codons in codon's temporal order constructed by Trifonov [2]. Remarkably, we observed an overall trend that the late coming codons were gainers whereas the early-coming codons were losers, with only a few exceptions (e.g., AAG and GTC) (Figure 2). Compared to the fixation, polymorphisms had a stronger trend. All the latest 10 codons were increasing while 10 of the earliest 11 codons were decreasing. This observation implies that, in the course of evolution, late coming codons were continuously added into the genetic code and encoded amino acids in the primitive proteins; meanwhile, the usage of early codons in proteins gradually declined. Finally, we did not find significant correlation between the codon changes and the physiochemical characters (charge, polarity, and polarity and volume) of amino acids encoded by the codons.

3.4. Codon Gain or Loss with Codon Usage. We further examined the relationship between codon usage and the trend of codon changes by the polymorphisms and fixation. We obtained the frequencies of codon usage in the human genome from the Codon Usage Database [11]. We defined a codon being rare when its frequency was ≤13 per 1,000 codons in the genome, as previously suggested in [18, 19]. Figure 3 shows the changes of codons ordered by the codon usage frequency. Interestingly, fixation had strong, but opposite, effect on rarely and frequently used codons (Figure 3). For rare codons, fixation increased 18 while decreased 6 codons. In contrast, for nonrare codons, fixation increased 13 but decreased 24 codons. The difference was statistically significant (2×2 contingency table, $P = .005$). Strikingly, fixation resulted in a strong increase of the frequencies (i.e., gain) of all the 10 rarest codons (Figure 3). The extent of increasing these codons by fixation was stronger than other codons. Conversely, for the 16 most frequent codons, we observed 12 under loss by fixation.

Polymorphisms did not show such a contrast pattern as observed in fixation. However, there was an interesting feature. There is a total of eight codons that contain CG dinucleotides; they could be termed NCG and CGN where "N" denotes any nucleotide (A, C, G, or T). When we examined the nucleotide attributes of rare codons, we found that all these CG containing codons belonged to rare codons. Interestingly, polymorphisms tended to decrease all these codons except CGT, which had a small increase (Figure 3). It is well known that the mutation rate of CG → TG/CA is much higher because of the hypermutability of methylated CpGs, that is, CpG effect [20, 21]. Our analysis of mutation direction using SNP data indicated much higher frequencies of CG → TG/CA than TG/CA → CG changes in the NCG and CGN codons (1,600 versus 560, 2.86 folds, Table 2), which confirmed the strong CpG effect on these codons. A more detailed examination revealed that this difference was largely attributed to the asymmetrical synonymous transition between NCG and NCA. For example, for CCG codon, we observed 206 synonymous transitions CCG → CCA compared to 80 CCA → CCG. Consequently, the frequencies of the NCG and CGN codons tended to decrease by polymorphisms. The only exception is CGT, a codon gainer. There were only 6 CGT → CGC compared to 53 CGC → CGT changes. This unique asymmetry of synonymous transition at the third position of CG containing codons resulted in an overall accumulation.

We further compared CG → TG/CA versus TG/CA → CG changes in the eight CG containing codons using the polymorphism and fixation data. In contrast to the strong CpG effect in the spontaneous nucleotide polymorphisms, we detected a strong suppression of CpG effect on NCG codons as well as a moderate suppression on CGN codons by fixation. For example, in polymorphisms, the ratios of CG → TG/CA over TG/CA → CG changes were at least 2; however, in fixation, most of CpG containing codons, especially the NCG codons, had the ratio not greater than 1 (Table 2). When we combined all eight codons, the ratio was 2.86 by polymorphisms but only 0.50 by fixation. Based on these results, we proposed that fixation process could fix the potential rapid loss of CpG containing codons caused by the CpG effect, therefore, avoid dramatic change of codons during short evolutionary period. It is worth noting that a similar suppression of polymorphisms at the CpG dinucleotides in CpG islands, which are often found in the promoter regions, was found as a mechanism to maintain high CpGs and GC content in the promoter regions [22, 23]. Purifying selection was suggested to play a role in

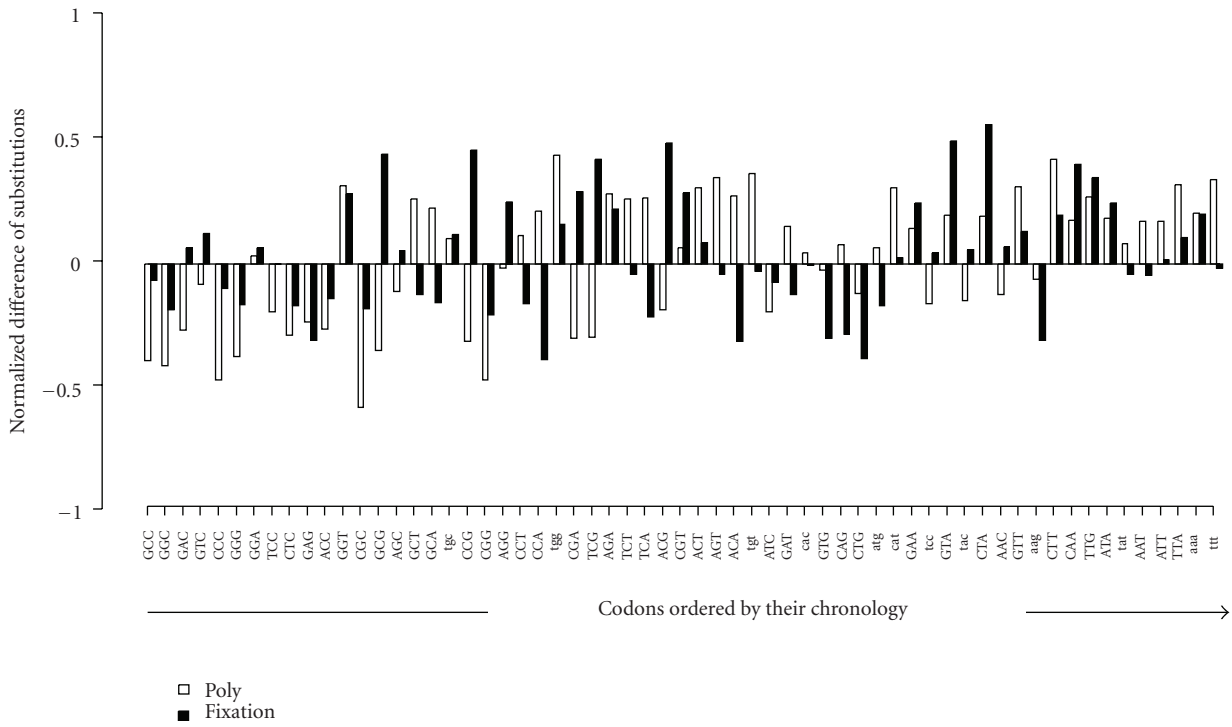


Figure 2: Normalized difference of codon changes (gain and loss) at the polymorphic and fixed sites in amino-acid coding regions. The normalized difference of a codon was defined as the difference of the number of mutations removing the codon from that creating the codon and divided by the total number of mutations in the codon. Codons are ordered (left: ancient; right: recent) in their temporal order in the genetic code according to Trifonov [2]. The codons associated with codon capture are in lower case.

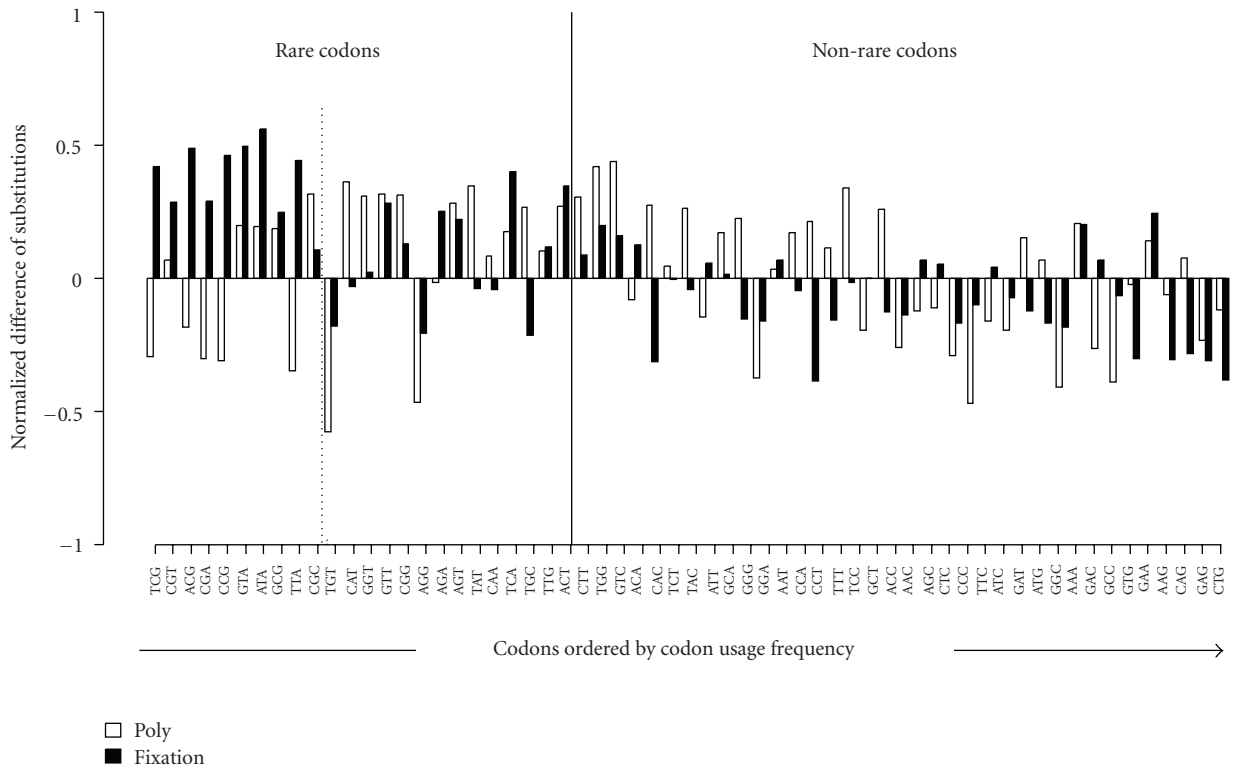


Figure 3: Normalized difference of codon changes (gain and loss) at polymorphic and fixes sites. Codons are ordered in the ascending order of frequencies in human codon usage based on the Codon Usage Database. The solid vertical line separates rare and nonrare codons. The 10 rarest codons are shown in the left side of the dotted line.

Table 2: Statistics of mutations *CG* ↔ *TG/CA* in eight CG-containing codons.

Codon	Polymorphism			Fixation		
	CG → TG/CA	TG/CA → CG	Ratio[a]	CG → TG/CA	TG/CA → CG	Ratio[a]
ACG	282	141	2.00	89	245	0.36
TCG	193	61	3.16	52	121	0.43
CCG	282	115	2.45	87	225	0.39
GCG	271	81	3.35	75	180	0.42
CGA	88	12	7.33	17	14	1.21
CGT	106	53	2.00	14	25	0.56
CGC	163	29	5.62	47	22	2.14
CGG	215	68	3.16	65	65	1.00
Total	1600	560	2.86	446	897	0.50

[a]Ratio of CG → TG/CA over TG/CA → CG changes.

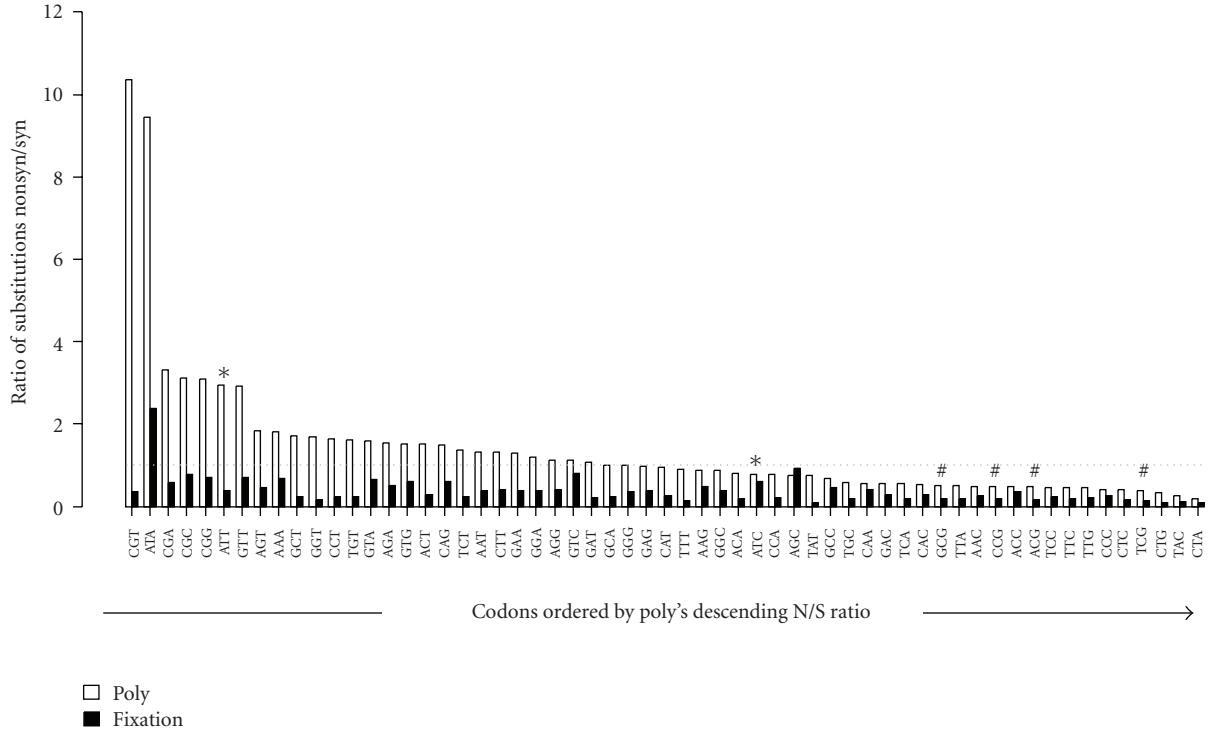


Figure 4: Nonsynonymous to synonymous substitutions (N/S) ratio calculated by the polymorphic and fixed nucleotide changes. Two codons (ATG and TGG) are not included because of their lack of synonymous substitutions. Symbol "#" indicates NCG codons. Codons are placed by the descending order of the polymorphism's N/S ratios.

the suppression of polymorphisms at the CpG dinucleotides in CpG islands [23]. Therefore, polymorphisms in the functional regions might have been under elimination due to functional constraints.

3.5. Selective Constraints on Codons. The different features of polymorphisms and fixation in codons, especially CpG containing codons, suggest that genetic factors such as selective constraints might have played important roles in codon evolution. We next calculated the ratio of nonsynonymous over synonymous substitution (N/S) for each codon, which is a typical method for detecting functional constraints on coding sequences [24]. Based on the fixation data, all codons except ATA had the N/S ratio smaller than 1, and most codons had the ratio much smaller than 1 (Figure 4). This suggests strong purifying selection on most codons during evolution. As expected, there were more synonymous mutations than nonsynonymous mutations because all codons, except ATG (coding Met), TGG (coding Trp), and ATA (coding Ile), could have transitions at their third position without changing amino acids and because the transition over transversion ratio has been observed to be approximate 2 in many vertebrate genomes [21, 25]. This type of synonymous transitional mutations was known to be "cheap" and occurred most frequently [26]. However, ATA lacks of such "cheap" mutations, which explains why the N/S ratio was greater than 1.

Table 3: Mutation pattern on $A|U$ dicodon boundaries.

	Polymorphism		Fixation	
	A|T	A|G	A|T	A|G
T/C/G → A	328	613	300	507
A → T/C/G	228	269	229	517
P-value[a]	4.49×10^{-5}		7.12×10^{-3}	

P-values were calculated by χ^2 test for 2×2 contingency tables. A|G dicodon boundaries were considered as control.
[a]Note the opposite pattern on A|U dicodon boundaries by polymorphism and fixation.

The N/S ratios using the polymorphism data were remarkably higher than the corresponding ones using the fixation data, with one exception (AGC, serine) (Figure 4). Nearly half of the codons had the N/S ratio greater than 1. Seven codons had ratio greater than 2, including four CGN codons. We found that spontaneous mutations favored transitions at the first and second positions (nonsynonymous) over transitions at the third position (mostly synonymous) in these seven codons, particularly in the CGN codons. For example, there were 106 CGT → TGT/CAT nonsynonymous changes but only 6 CGT → CGC synonymous changes. This resulted in a very high N/S ratio for codon CGT (Figure 4). However, these nonsynonymous changes were directly linked to the CpG effect, which conversely led to a high abundance of synonymous mutations at the second and third positions of NCG codons. Therefore, we observed small N/S ratios for NCG codons (Figure 4). As shown in Table 2, fixation had a much smaller ratio of CG → TG/CA over TG/CA → CG changes for NCG than for CGN codons. Taken together, CpG effect plays a dominant role in preferring mutations at the polymorphic sites. However, selective constraints during fixation process could largely correct it to balance the codon composition.

3.6. Selection Effect on Dicodon Boundary. An early study reported that AU dinucleotides was a cleavage site of RNase L, the 2′,5′-oligoadenylate-dependent ribonuclease [27]. Interestingly, a recent study extended this analysis in yeast and found that mRNAs containing A|U dinucleotides at synonymous dicodon boundaries had a short half-life due to more efficient 3′–5′ degradation by endonucleolytic cleavage [19]. If this was true in humans, we may detect the signature of purifying selection on A|U dinucleotides at synonymous dicodon boundaries. That is, mutations toward A at the third position of synonymous codon leading to A|U dicodon boundaries would not favor, but mutations from A at the third position of synonymous codon leading to non-A|U dicodon boundaries would favor. The mRNA half-life is irrelative of the use of synonymous A|G dicodon boundary [19]; therefore, we used synonymous A|G dicodon boundary as control. Our analysis of polymorphism data revealed a dramatic deficit of A|U dicodon boundary (Table 3). However, we observed an opposite effect of fixation on A|U dicodon boundary, that is, fixation would favor A|U dicodon boundaries. This opposite feature has not been reported in literature. The mechanism is unclear and needs further investigation.

4. Conclusions

Amino acid compositions and nucleotide substitutions have been extensively studied because they are fundamental in protein and genome evolution. So far, either interspecies nucleotide substitutions or intraspecies nucleotide polymorphisms, but not both, have been analyzed. In this study, we uniquely and systematically compared the inter- and intraspecies nucleotide changes in amino-acid coding sequences, especially at the codon level. Our results provided a detailed view on the spontaneous point mutations in codons and subsequent and rapid fixation in a lineage (e.g., human lineage) in recent genome evolution. We observed a trend of loss of the ancient codons while gain of the latest codons. Fixation had a strong bias towards increasing the rarest codons but decreasing the most frequent codons, suggesting that codon equilibrium has not been reached yet. Another major feature is that fixation could effectively suppress the strong CpG effect on the CG containing codons. Functional constraints such as purifying selection have likely played a major role in codon changes. Finally, we reported a unique and opposite mutation pattern on A|U dicodon boundaries. Although these findings are limited to the trends of codon gain and loss in the recent history, more specifically in the recent human history, they provide important insights on how nucleotide changes in the protein-coding regions have likely shaped the genome and protein sequence composition.

Acknowledgment

This project was partially supported by NIH Grants (LM009598 and AA017437) and the NARSAD young investigator award to Z. Zhao.

References

[1] D. J. Brooks, J. R. Fresco, A. M. Lesk, and M. Singh, "Evolution of amino acid frequencies in proteins over deep time: inferred order of introduction of amino acids into the genetic code," *Molecular Biology and Evolution*, vol. 19, no. 10, pp. 1645–1655, 2002.

[2] E. N. Trifonov, "The triplet code from first principles," *Journal of Biomolecular Structure and Dynamics*, vol. 22, no. 1, pp. 1–11, 2004.

[3] S. L. Miller, "A production of amino acids under possible primitive earth conditions," *Science*, vol. 117, no. 3046, pp. 528–529, 1953.

[4] D. J. Brooks and J. R. Fresco, "Increased frequency of cysteine, tyrosine, and phenylalanine residues since the last universal ancestor," *Molecular & Cellular Proteomics*, vol. 1, no. 2, pp. 125–131, 2002.

[5] I. K. Jordan, F. A. Kondrashov, I. A. Adzhubei, et al., "A universal trend of amino acid gain and loss in protein evolution," *Nature*, vol. 433, no. 7026, pp. 633–638, 2005.

[6] J. H. McDonald, "Apparent trends of amino acid gain and loss in protein evolution due to nearly neutral variation," *Molecular Biology and Evolution*, vol. 23, no. 2, pp. 240–244, 2006.

[7] L. D. Hurst, E. J. Feil, and E. P. C. Rocha, "Protein evolution: causes of trends in amino-acid gain and loss," *Nature*, vol. 442, no. 7105, pp. E11–E12, 2006.

[8] C. Jiang and Z. Zhao, "Mutational spectrum in the recent human genome inferred by single nucleotide polymorphisms," *Genomics*, vol. 88, no. 5, pp. 527–534, 2006.

[9] Z. Zhang, S. Schwartz, L. Wagner, and W. Miller, "A greedy algorithm for aligning DNA sequences," *Journal of Computational Biology*, vol. 7, no. 1-2, pp. 203–214, 2000.

[10] Z. Zhao, L. Jin, Y.-X. Fu, et al., "Worldwide DNA sequence variation in a 10-kilobase noncoding region on human chromosome 22," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 21, pp. 11354–11358, 2000.

[11] Y. Nakamura, T. Gojobori, and T. Ikemura, "Codon usage tabulated from international DNA sequence databases: status for the year 2000," *Nucleic Acids Research*, vol. 28, no. 1, p. 292, 2000.

[12] M. T. Webster, N. G. C. Smith, and H. Ellegren, "Compositional evolution of noncoding DNA in the human and chimpanzee genomes," *Molecular Biology and Evolution*, vol. 20, no. 2, pp. 278–286, 2003.

[13] J. Zhang, "Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes," *Journal of Molecular Evolution*, vol. 50, no. 1, pp. 56–68, 2000.

[14] J. T. Wong, "A co-evolution theory of the genetic code," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 72, no. 5, pp. 1909–1912, 1975.

[15] F. H. C. Crick, "The origin of the genetic code," *Journal of Molecular Biology*, vol. 38, no. 3, pp. 367–379, 1968.

[16] B. G. Barrell, S. Anderson, and A. T. Bankier, "Different pattern of codon recognition by mammalian mitochondrial tRNAs," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 77, no. 6, pp. 3164–3166, 1980.

[17] T. Ohama, S. Osawa, K. Watanabe, and T. H. Jukes, "Evolution of the mitochondrial genetic code IV. AAA as an asparagine codon in some animal mitochondria," *Journal of Molecular Evolution*, vol. 30, no. 4, pp. 329–332, 1990.

[18] G. Caponigro, D. Muhlrad, and R. Parker, "A small segment of the MATα 1 transcript promotes mRNA decay in Saccharomyces cerevisiae: a stimulatory role for rare codons," *Molecular and Cellular Biology*, vol. 13, no. 9, pp. 5141–5148, 1993.

[19] D. B. Carlini, "Context-dependent codon bias and messenger RNA longevity in the yeast transcriptome," *Molecular Biology and Evolution*, vol. 22, no. 6, pp. 1403–1411, 2005.

[20] B. K. Duncan and J. H. Miller, "Mutagenic deamination of cytosine residues in DNA," *Nature*, vol. 287, no. 5782, pp. 560–561, 1980.

[21] F. Zhang and Z. Zhao, "The influence of neighboring-nucleotide composition on single nucleotide polymorphisms (SNPs) in the mouse genome and its comparison with human SNPs," *Genomics*, vol. 84, no. 5, pp. 785–795, 2004.

[22] Z. Zhao and F. Zhang, "Sequence context analysis in the mouse genome: single nucleotide polymorphisms and CpG island sequences," *Genomics*, vol. 87, no. 1, pp. 68–74, 2006.

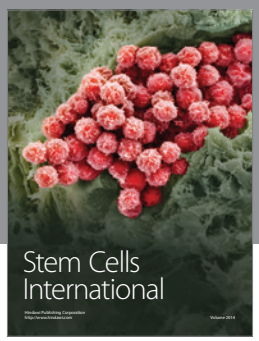
[23] Z. Zhao and F. Zhang, "Sequence context analysis of 8.2 million single nucleotide polymorphisms in the human genome," *Gene*, vol. 366, no. 2, pp. 316–324, 2006.

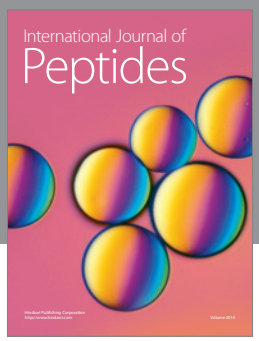
[24] A. Nekrutenko, K. D. Makova, and W.-H. Li, "The K_A/ K_S ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study," *Genome Research*, vol. 12, no. 1, pp. 198–202, 2002.

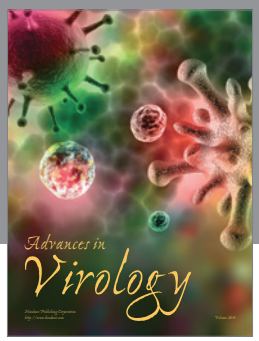
[25] Z. Zhao and E. Boerwinkle, "Neighboring-nucleotide effects on single nucleotide polymorphisms: a study of 2.6 million polymorphisms across the human genome," *Genome Research*, vol. 12, no. 11, pp. 1679–1686, 2002.

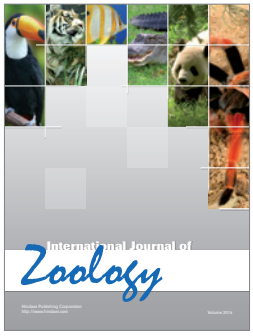
[26] A. D. Yoder, R. Vilgalys, and M. Ruvolo, "Molecular evolutionary dynamics of cytochrome b in strepsirrhine primates: the phylogenetic significance of third-position transversions," *Molecular Biology and Evolution*, vol. 13, no. 10, pp. 1339–1350, 1996.

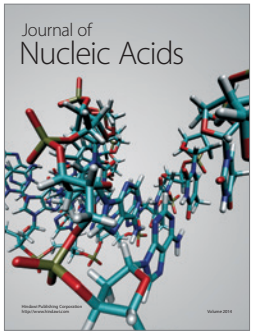
[27] S. S. Carroll, E. Chen, T. Viscount, et al., "Cleavage of oligoribonucleotides by the 2′,5′-oligoadenylate- dependent ribonuclease L," *The Journal of Biological Chemistry*, vol. 271, no. 9, pp. 4988–4992, 1996.

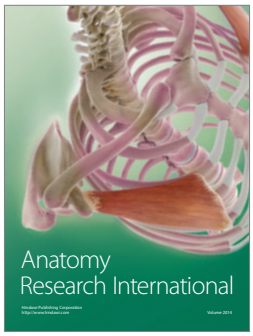


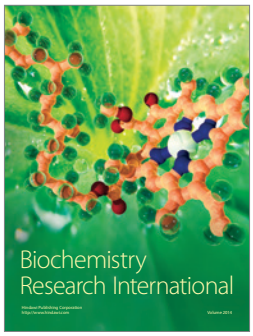


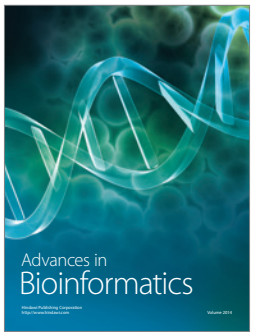


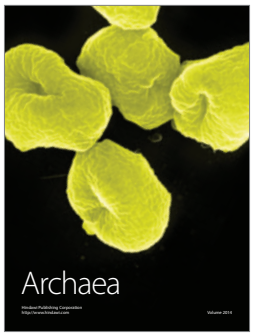


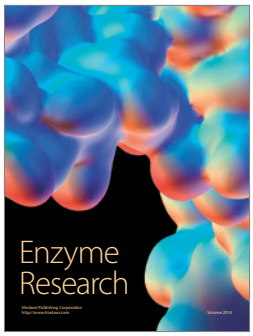


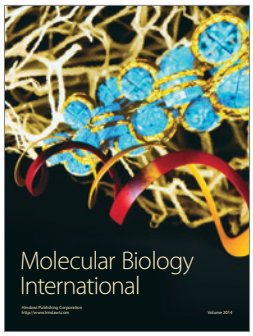












Hindawi

Submit your manuscripts at
http://www.hindawi.com