Article ID: 0253-2778(2010)07-0679-07

Nonparallel accumulation of synonymous and non-synonymous substitutions on human and chimpanzee chromosomes

SONG Yumei[1], CHEN Xi[2], CHEN Xueping[1]

(1. *Department of Chemistry, University of Science and Technology of China, Hefei* 230026, *China*;
2. *College of Life Science, Liaoning Normal University, Dalian* 116029, *China*)

Abstract: To reveal the evolutionary dynamics of primate chromosomes, using mice and dogs as backgrounds, we analyzed a total of 61949 pairs of genes by comparative genomics and bioinformatics methods, which included 16 427, 15 161, 15 802 and 14 559 ortholog pairs identified, from human-mouse, human-dog, chimpanzee-mouse and chimpanzee-dog phylogenies respectively. The results show that, in humans and chimpanzees, genes on chromosomes 16, 19, 21 and 22 have featured significantly higher synonymous substitution rates (d_S). Analysis of human-mouse-dog and chimpanzee-mouse-dog ortholog trios also indicates d_S and non-synonymous substitution rates (d_N) to be homogeneous across different phylogeny branches, suggesting that the relevant genes have been subjected to similar selection for base substitution rates. The analysis also suggests that local chromatin environment, such as GC content and gene density, may contribute to the accumulation of both types of substitutions on human chromosomes. Furthermore, recombination rates seem to have a significant influence on the d_S of human chromosomes.

Key words: chromatin environment; chromosome; non-synonymous substitution; phylogeny divergence; synonymous substitution

CLC number: Q349 **Document code**: A doi: 10.3969/j.issn.0253-2778.2010.07.004

人类和黑猩猩染色体同义和非同义替代的非平行累积

宋玉梅[1], 陈 曦[2], 陈学平[1]

(1. 中国科学技术大学化学系, 安徽合肥 230026; 2. 辽宁师范大学生命科学学院, 辽宁大连 116029)

摘要: 为了揭示灵长类染色体的进化动态, 以小鼠和狗作为外群, 使用比较基因组学和生物信息学的方法, 详细分析了人-小鼠、人-狗、黑猩猩-小鼠、黑猩猩-狗中的16 427, 15 161, 15 802和14 559同源基因. 结果表明, 人类和黑猩猩染色体16, 19, 21和22上的基因具有显著高的同义替代速率(d_S). 分析人类-小鼠-狗和黑猩猩-小鼠-狗的同源基因, 发现不同系谱基因碱基的同义和非同义替代速率(d_N)是相似的, 揭示了这些物种基

Received: 2010-04-26; **Revised**: 2010-06-05

Foundation item: Supported by National Natural Science Foundation of China (30970348).

Biography: SONG Yumei, female, born in 1985, master. Research field: bioinformatics. E-mail: sym11@mail.ustc.edu.cn

Corresponding author: CHEN Xueping, PhD/associate Prof. E-mail: chenxp08@ustc.edu.cn

因的碱基替代速率经历着相似的进化选择压力.此外,结果也表明局部染色质环境(GC 含量和基因密度)和染色体重组速率对人类染色体碱基替代的累积也有显著影响.

关键词:染色质环境;染色体;非同义替代;系统发育分化;同义替代

0 Introduction

The non-synonymous (d_N) and synonymous substitution rates (d_S) as well as their ratio have been widely used as an indicator of natural selection for gene sequences. Whilst excess of non-synonymous polymorphisms is thought to be indicative of diversifying selection, shortage of them has been construed as purifying selection imposed by constraints operating at the level of protein structure and function. It has become clear that nucleotide substitution rates at neutral sites may vary widely across mammalian genomes[1], and so may its correlations with GC content, CpG, local recombination rates, small insertions and deletions, insertions of transposable elements, single nucleotide polymorphisms (SNPs), and distance to telomeres[2-3]. Repeatability of substitutions, i.e., whether substitution rates are homogeneous or correlated across different phylogeny branches, has also been an interesting issue. Several studies have demonstrated that substitution rates for autosomes and regions of conserved synteny are similar across mammalian genomes[4-5]. Moreover, the fact that commonly observed clustering of essential genes, and that chromosomal distribution of genes are typically nonrandom[6-7] have implied that regional features may also impact on evolution of genes as well as chromosome organization.

With the completion of more and more eukaryotic, in particular, mammalian genome sequences, study of human genome evolution has become possible. Comparisons of human genome sequence with other mammalian genomes have facilitated characterization of nucleotide substitution patterns[8-9]. Nevertheless, study of the dynamics underlying the evolution of human chromosome organizations has been mainly focused on sex chromosomes[3-4,10-11].

In this study, we have re-addressed the issue of repeatability of synonymous and non-synonymous substitution rates by using genome-scale mouse, dog, human and chimpanzee datasets to explore the possibility of nonparallel base substitutions in protein-coding genes across human chromosomes. Considering the small divergence between human and chimpanzee, we also employed chimpanzee sequence to assess whether substitution rates are homogeneous or correlated across different phylogeny branches. With mouse and dog as outgroup, ortholog gene pairs were derived from selected human-mouse, human-dog, chimpanzee-mouse and chimpanzee-dog phylogenies. Subsequently, d_N and d_S were calculated and compared for each of the human and chimpanzee chromosomes. Repeatability of d_N and d_S across different branches were tested upon ortholog trios identified from human-mouse-dog and chimpanzee-mouse-dog phylogenies. Finally, we assessed the potential influence of gene density, GC content and local recombination rates on both types of substitutions.

1 Materials and methods

1.1 Data acquisition

Gene sequences were retrieved from the Ensembl database (http://www.ensembl.org) release 56 (Human build GRCh37, chimpanzee build CHIMP2.1, mouse build NCBI m37, dog build CanFam 2.0). The numbers of ortholog pairs identified from human-mouse, human-dog, chimpanzee-mouse and chimpanzee-dog phylogenies were 16 427, 15 161, 15 802 and 14 559, respectively. In addition, 14131 human-mouse-dog and 13636 chimpanzee-mouse-dog ortholog trios were identified. Genes were grouped according to their chromosomal locations.

1.2 Calculation of substitution rates

To calculate the substitution rates for the ortholog pairs, multiple alignments of coding sequence (cds) and corresponding proteins were obtained in FASTA format using a ClustalX program. Alignments were checked for correct reading frame[12-13]. Subsequently, a maximum likelihood method implemented in the PAML package was used for calculating the substitution rate for each pair of genes.

1.3 Calculation of gene density

Using the method described by Caron et al[14], gene density was defined as the inverse of the average distance between neighboring genes in a window of 40 adjacent transcripts and was calculated for all chromosomes. Correlation between gene density and other variables, such as synonymous substitution rate (SSR) and GC content, was computed for genes that have orthologs in corresponding species.

1.4 Recombination rates

Average recombination rates for human chromosomes were derived from the previous works by Kong et al[15].

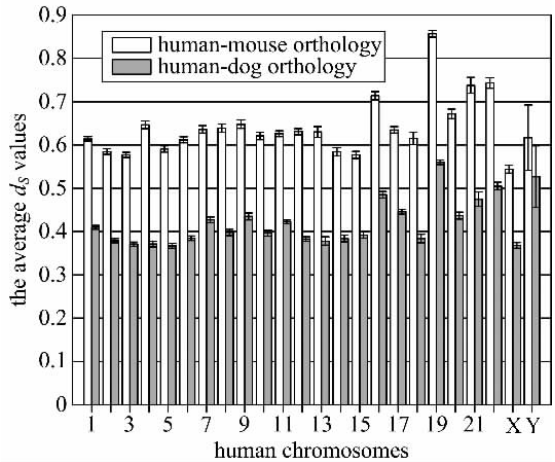
1.5 Statistical tests

All data were analyzed with the Statistic Analysis System (SAS). Values for each group were compared with a Student-Newman-Keuls Test.

2 Results and discussion

2.1 Distinct d_S values across various human and chimpanzee chromosomes

As synonymous sites are believed to be subject to little or no selection, d_S value has been frequently used as a rough measure for basal mutation rate. To compare the d_S of different human chromosomes, d_S values were first calculated from the human-mouse and human-dog phylogenies (Fig. 1). Based on human-mouse phylogeny, human chromosomes featured significantly different d_S values (ANOVA: $F = 64.45$, $P < 0.0001$), indicating that accumulation of synonymous substitutions on human chromosomes has been nonparallel. Particularly, chromosomes 19, 22, 21 and 16 have significantly higher average d_S values (0.857, 0.743, 0.738 and 0.714, respectively), whilst the X chromosome has the lowest average d_S (0.544).



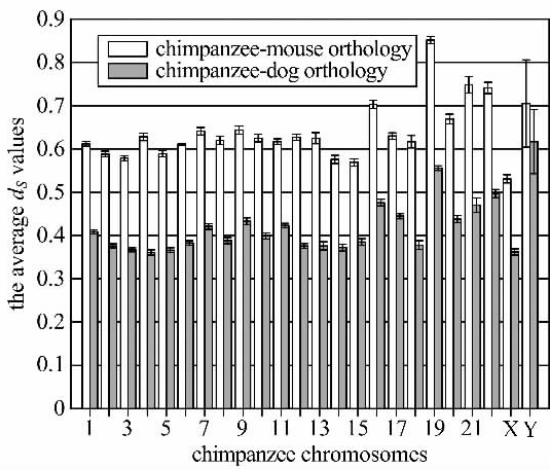
There were fewer homologous genes on the human Y (only 14 for human-mouse orthology and 12 for human-dog orthology), so the d_S on Y chromosome were omitted from the analysis.

Fig. 1 Substitution rates of ortholog pairs for human chromosomes

A very similar range of d_S for human chromosomes was derived from human-mouse and human-dog phylogenies, with genes from chromosomes 19, 22, 16 and 21 featuring significantly higher average d_S values (0.560, 0.505, 0.485 and 0.475, respectively) than others (ANOVA: $F = 57.22$, $P < 0.0001$), and those from X chromosome scoring the lowest d_S (0.368).

To explore the divergence between primate chromosomes, d_S values were also calculated from the chimpanzee-mouse and chimpanzee-dog phylogenies. The average d_S value on each chimpanzee chromosome is shown in Fig. 2. Analysis of chimpanzee-mouse ortholog pairs suggested that genes on chimpanzee chromosomes 19, 21, 22 and 16 have significantly higher average d_S values (0.852, 0.748, 0.741 and 0.703, respectively) than others (ANOVA: $F =$

60.46, $P<0.000\ 1$), with chimpanzee chromosome X also having the lowest average d_S value (0.531). From the chimpanzee-dog phylogeny, average d_S values for chimpanzee chromosomes 19, 22, 16 and 21 (0.556, 0.497, 0.476 and 0.470, respectively) also showed significant divergence from others (ANOVA: $F=55.73$, $P<0.000\ 1$). Again, genes from chimpanzee chromosome X scored the lowest average d_S value (0.362).



There were fewer homologous genes on the chimpanzee Y (only 11 for chimpanzee-mouse orthology and 12 for chimpanzee-dog orthology), so the d_S on Y chromosome were omitted from the analysis.

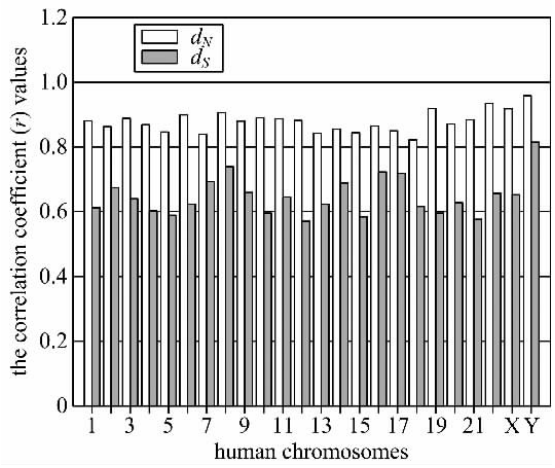
Fig. 2 Substitution rates of ortholog pairs for chimpanzee chromosomes

Pearson correlation analysis was carried out to assess the similarity in the range of d_S between human and chimpanzee chromosomes. Based on human-mouse and chimpanzee-mouse phylogenies, a strong correlation in average d_S value was found ($r=0.957\ 0$, $P<0.000\ 1$). Similar correlation was also derived from the human-dog and chimpanzee-dog phylogenies ($r=0.961\ 0$, $P<0.000\ 1$).

Although similar d_S across mammalian genomes were observed previously, nonparallel accumulation of synonymous substitutions on different chromosomes in diverse lineages has not been reported elsewhere. Large divergence and significant inter-chromosomal differences were only found with chromosome 19 based on the analysis of human-mouse orthologs[16]. Nonparallel accumulation of synonymous substitutions across both human and chimpanzee chromosomes, and average d_S values of human and chimpanzee chromosomes 16, 19, 21 and 22 diverged significantly from those of others, implying that inter-chromosomal differences in terms of d_S are rather complex.

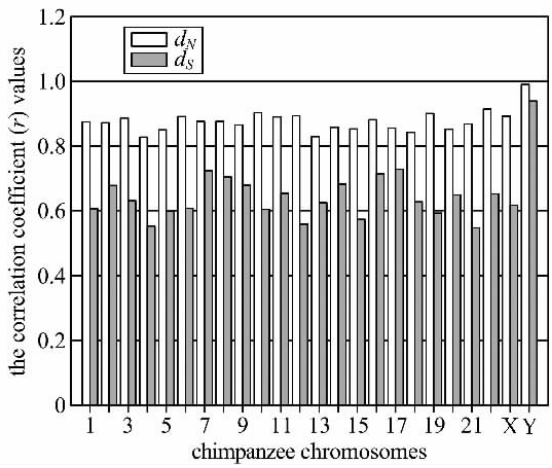
2.2 The d_S and d_N seem to be homogeneous among different phylogenic branches

The divergence in d_S across human and chimpanzee chromosomes also raised the question as to whether d_S or d_N are similar across different groups of orthologous sequences. To address this issue, ortholog trios of human-mouse-dog and chimpanzee-mouse-dog were identified and analyzed. Correlations between d_N and d_S for each of the phylogenies, namely human-mouse, human-dog, chimpanzee-mouse and chimpanzee-mouse were investigated. As shown in Fig. 3 and Fig. 4, a strong correlation between d_S and d_N were found in both of the phylogeny trios for all chromosomes except Y, suggesting that synonymous and non-synonymous substitution rates are homogeneous in different mammalian phylogenic branches. For lack of correlation between the d_S and d_N on Y chromosome, the explanation may be the



$P<0.000\ 1$ for all values, except Y chromosome

Fig. 3 Correlation coefficient (r) values in d_N and d_S for the human-mouse and human-dog phylogenies



$P<0.000\ 1$ for all values, except Y chromosome

Fig. 4　Correlation coefficient (r) values in d_N and d_S for the chimpanzee-mouse and chimpanzee-dog phylogenies

significantly fewer homologous genes identified (7 from the human-mouse-dog trio and 4 from the chimpanzee-mouse-dog trio).

Further analysis found the correlation coefficients between d_N to be significantly higher than between d_S in human-mouse-dog ($\chi^2=35.27$, $P < 0.000\ 1$) and chimpanzee-mouse-dog phylogenies ($\chi^2 = 29.86$, $P < 0.000\ 1$). Non-synonymous substitutions may change the codons, therefore, d_N may have been subjected to stronger purifying selection, which seems to be in keeping with the stronger correlation across different branches. Another possible factor is the presence of recombination. Without it, divergent and deleterious mutations may increase[17]. In addition, the selection increased by recombination may also result in the stronger correlation between d_N. Conversely, the stronger correlation of d_N between different phylogenies implied that d_N is less sensitive to divergence.

Similarity between synonymous substitutions was also found to be consistent with non-synonymous substitutions in human-mouse-dog and chimpanzee-mouse-dog phylogenies, though this tendency is relatively weaker than non-synonymous substitutions. A possible explanation for this may be purifying selection, which can keep the sites conserved across multiple species, has acted on the synonymous sites. Several recent studies have provided evidence for selection on synonymous sites[18-19]. Resch et al[20] also showed that positive selection in synonymous sites of mammalian genes is widespread. These seemed to suggest that the two aspects reported previously may be the reasons for the aforementioned discrepancies. On one hand, some synonymous sites may have been subjected to purifying selection owing to functional constraints. On the other hand, some sites are under positive selection, which may affect mRNA levels and translation through mRNA destabilization[20]. Additionally, the divergent correlation coefficient (r) values in d_S and d_N implied that synonymous sites are under stronger positive selection than non-synonymous ones. The two aspects may in part explain why synonymous substitutions are also consistent with selection and that this tendency is slightly weaker in comparison with that of the non-synonymous substitutions. Another possibility may be that certain neutral mutation patterns are not divergent among the studied ortholog trios due to relatively short-term evolution. Evolving by micro- and macro-rearrangements, mammalian genomes contain abundant heterogeneous neutral mutations. Contrary to the general belief that mutations are random events, Crombach et al[17] argued that certain types of mutations may occur preferentially. Therefore, the consistency between synonymous substitution and selection may have also resulted from the biased occurrence and preservation of mutations.

2.3 Potential influence of local chromatin environment on base substitutions

To further assess the impact of GC content on base substitution, the correlation between substitution rates and GC content of individual genes was calculated. Pearson correlations and P values are listed in Tab. 1. For all human chromosomes, the d_S calculated from human-mouse and human-dog phylogenies are strongly

correlated with GC content. In contrast, weak correlations between d_N and GC content were only found with several chromosomes (4, 5, 12, 19 and 20) (data not shown). Notably, chromosome 19 featured the weakest positive correlation for d_S and GC content and strongest negative correlation for d_N and GC content. The results seemed to have confirmed that the GC content may affect base substitution rates. Previous studies showed that regional GC content may influence the recombination frequency, and that the presence of recombination may alter the rate of substitution[15,21].

Tab. 1 Correlation between d_S and GC content of corresponding genes on different chromosomes

Human chromosome	Human-mouse orthology		Human-dog orthology	
	Pearson correlation (r)	P value	Pearson correlation (r)	P value
1	0.373	0.000 1	0.564	0.000 1
2	0.417	0.000 1	0.510	0.000 1
3	0.292	0.000 1	0.454	0.000 1
4	0.459	0.000 1	0.579	0.000 1
5	0.406	0.000 1	0.537	0.000 1
6	0.285	0.000 1	0.376	0.000 1
7	0.441	0.000 1	0.487	0.000 1
8	0.572	0.000 1	0.652	0.000 1
9	0.576	0.000 1	0.531	0.000 1
10	0.408	0.000 1	0.474	0.000 1
11	0.404	0.000 1	0.473	0.000 1
12	0.111	0.000 9	0.422	0.000 1
13	0.456	0.000 1	0.506	0.000 1
14	0.446	0.000 1	0.555	0.000 1
15	0.329	0.000 1	0.538	0.000 1
16	0.509	0.000 1	0.522	0.000 1
17	0.418	0.000 1	0.486	0.000 1
18	0.333	0.000 1	0.463	0.000 1
19	0.093	0.005 7	0.248	0.000 1
20	0.445	0.000 1	0.531	0.000 1
21	0.539	0.000 1	0.656	0.000 1
22	0.439	0.000 1	0.537	0.000 1
X	0.442	0.000 1	0.380	0.000 1

In the present work, we have further investigated the relationship between recombination rate and substitution rate in human chromosomes. The average recombination rates for human chromosomes were derived from the previous works by Kong et al[15]. The result showed that, in human-mouse phylogeny, strong correlation existed between recombination rate and d_S ($r=0.588\ 7$, $P=0.003\ 1<0.01$), whilst the correlation between recombination rate and d_N was not significant ($r=0.245\ 7$, $P=0.258\ 4>0.05$). For the human-dog phylogeny, significant correlation existed between recombination rate and d_S ($r=0.631\ 3$, $P=0.001\ 2<0.01$), whilst there was also significant correlation between recombination rate and d_N ($r=0.434\ 1$, $P=0.038\ 5<0.05$), indicating that recombination rates have an influence on substitution rate in human chromosomes. Schmegner et al[22] also noticed that the rate of base pair substitutions changes at a GC content transition, which may in part account for the aforementioned correlation. In the human genome, the number of synonymous codons (excluding AUG, UAG and UGG) that have C or G in the third position reaches 22 345 926 (55.0%), whilst other synonymous codons (excluding UAA and UGA) amount to 16 749 425 (41.2%) (original data derived from http://www.kazusa.or.jp/codon/). This suggests that the regions featuring higher GC content may harbor more synonymous codons, which comprise a substantial portion of all synonymous substitutions.

With human-mouse and human-dog phylogenies, we further explored the relationship between the substitution rates and density of corresponding genes (Tab. 2). The result indicated a significant correlation between d_S and gene density for most human chromosomes, whereas significant correlations between d_N and gene density were only found with chromosomes 4 and 8 (data not shown). A previous study argued that the regions featuring high gene density are correlated with open chromatin structures, whereas regions poor in genes tend to correlate with close or compact chromatin[23], which seems in keeping with our observation.

Tab. 2　Correlation between d_S and the density of corresponding genes on different chromosomes

Human chromosome	Human-mouse orthology		Human-dog orthology	
	Pearson correlation (r)	P value	Pearson correlation (r)	P value
1	0.047	0.051 6	0.216	0.000 1
2	0.031	0.299 4	0.061	0.051 6
3	0.030	0.349 0	0.182	0.000 1
4	0.343	0.000 1	0.342	0.000 1
5	0.136	0.000 2	0.184	0.000 1
6	0.111	0.001 2	0.203	0.000 1
7	0.207	0.000 1	0.280	0.000 1
8	0.435	0.000 1	0.448	0.000 1
9	0.279	0.000 1	0.244	0.000 1
10	−0.107	0.006 1	−0.073	0.077 8
11	0.118	0.000 1	0.137	0.000 1
12	−0.135	0.000 1	0.094	0.006 5
13	0.233	0.000 1	0.153	0.014 9
14	0.056	0.197 5	0.144	0.001 2
15	−0.153	0.000 5	−0.095	0.033 2
16	0.300	0.000 1	0.239	0.000 1
17	0.048	0.132 3	0.106	0.001 6
18	0.131	0.044 3	0.020	0.770 5
19	0.102	0.002 4	0.111	0.001 5
20	0.111	0.018 6	0.160	0.000 9
21	0.277	0.000 6	0.248	0.004 7
22	−0.140	0.007 8	−0.045	0.416 5
X	0.154	0.000 1	0.226	0.000 1

Open chromatin is indicative of transcriptional constancy, e.g., housekeeping genes, rather than transcriptional activity. Zhang et al[24] found d_S to be lower for housekeeping genes. In addition, in regions featuring higher gene density, nucleotide polymorphisms are fewer[25]. These observations may in part explain the correlation between gene density and d_S found in the present work. Researchers also found that genes in closed chromatin display the highest levels of selection at synonymous sites[26], and those with positive selection at synonymous sites have average higher d_S[20]. These also seem to be in keeping with our observation.

References

[1] Wolfe K H, Sharp P M, Li W H. Mutation rates differ among regions of the mammalian genome[J]. Nature, 1989, 337:283-285.

[2] McVicker G, Gordon D, Davis C, et al. Widespread genomic signatures of natural selection in hominid evolution[J]. PLoS Genetics, 2009, 5(5):e1000471.

[3] Duret L, Arndt P F. The impact of recombination on nucleotide substitutions in the human genome[J]. PLoS Genetics, 2008, 4(5):e1000071.

[4] Malcom C M, Wyckoff G J, Lahn B T. Genic mutation rates in mammals: local similarity, chromosomal heterogeneity, and X-versus-autosome disparity[J]. Molecular Biology and Evolution, 2003, 20(10):1 633-1 641.

[5] Chuang J H, Li H. Similarity of synonymous substitution rates across mammalian genomes[J]. Journal of Molecular Evolution, 2007, 65(3):236-248.

[6] Batada N N, Hurst L D. Evolution of chromosome organization driven by selection for reduced gene expression noise[J]. Nature Genetics, 2007, 39: 945-949.

[7] Hentges K E, Pollock D D, Liu B, et al. Regional variation in the density of essential genes in mice[J]. PLoS Genetics, 2007, 3(5):e72.

[8] Clark A G, Glanowski S, Nielsen R, et al. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios[J]. Science, 2003, 302:1 960-1 963.

[9] Yang S, Smit A F, Schwartz S, et al. Patterns of insertions and their covariation with substitutions in the rat, mouse, and human genomes[J]. Genome Research, 2004, 14:517-527.

[10] Zhang Rui, Peng Yi, Wang Wen, et al. Rapid evolution of an X-linked microRNA cluster in primates[J]. Genome Research, 2007, 17:612-617.

[11] Rozen S, Marszalek J D, Alagappan R K, et al. Remarkably little variation in proteins encoded by the Y chromosome's single-copy genes, implying effective purifying selection[J]. American Journal of Human Genetics, 2009, 85(6):923-928.

[12] Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood[J]. Comput Appl Biosci, 1997, 13(5):555-556.

[13] Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments[J]. Nucleic Acids Research, 2006, 34:W609-W612.

[14] Caron H, van Schaik B, van der Mee M, et al. The human transcriptome map: clustering of highly expressed genes in chromosomal domains[J]. Science, 2001, 291:1 289-1 292.

(下转第 705 页)