Gene 557 (2015) 182–187

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Relationship between amino acid usage and amino acid evolution in primates

Haoxuan Liu [1], Zhengqing Xie [1], Shengjun Tan, Xiaohui Zhang *, Sihai Yang *

State Key Laboratory of Pharmaceutical Biotechnology, School of Life Sciences, Nanjing University, Nanjing 210093, China

ARTICLE INFO

Article history:
Received 13 November 2013
Received in revised form 4 December 2014
Accepted 14 December 2014
Available online 16 December 2014

Keywords:
Amino acid gain/loss
Amino acid evolution
Primates
Neutral evolution

ABSTRACT

Amino acid usage varies from species to species. A previous study has found a universal trend in amino acid gain and loss in many taxa and a one-way model of amino acid evolution in which the number of new amino acids increases as the number of old amino acids decreases was proposed. Later studies showed that this pattern of amino acid gain and loss is likely to be compatible with the neutral theory. The present work aimed to further study this problem by investigating the evolutionary patterns of amino acids in 8 primates (the nucleotide and protein alignments are available online http://gattaca.nju.edu.cn/pub_data.html). First, the number of amino acids gained and lost was calculated and the evolution trend of each amino acid was inferred. These values were found to be closely related to the usage of each amino acid. Then we analyzed the mutational trend of amino acid substitution in human using SNPs, this trend is highly correlated with fixation trend only with greater variance. Finally, the trends in the evolution of 20 amino acids were evaluated in human on different time scales, and the increasing rate of 5 significantly increasing amino acids was found to decrease as a function of time elapsed since divergence, and the dS/dN ratio also found to increase as a function of time elapsed since divergence. These results suggested that the observed amino acid substitution pattern is influenced by mutation and purifying selection. In conclusion, the present study shows that usage of amino acids is an important factor capable of influencing the observed pattern of amino acid evolution, and also presented evidences suggesting that the observed universal trend of amino acid gain and loss is compatible with neutral evolution.

1. Introduction

In general, evolution consists of two processes: mutation and fixation. In protein evolution, most mutations result in amino acid substitution. Fixation depends on the fitness of the mutation (Kimura, 1983; Ohta, 1992; Gillespie, 1991). Multiple lines of evidence and systematic analysis of genomic data have demonstrated that various factors, including positive selection, purifying selection, and regional genomic base composition could affect the evolution of proteins (Koonin, 2005). One of the important aspects of protein evolution is amino acid usage. Amino acid (AA) usage can vary substantially in different organisms. For example, one of the factors that influence AA usage is GC-content. The GC-content of different organisms can vary from 25% to 75% (Lynch, 2007; Sueoka, 1962; Bentley and Parkhill, 2004). Proteins from high-GC organisms are encoded by more high GC-codons, and vice versa (Sueoka, 1961; Gu et al., 1998; Knight et al., 2001; Wang et al., 2004).

It has long been assumed that amino acid composition of proteins is in equilibrium with equal and reciprocal fluxes of amino acid substitutions and constant amino acid frequencies (Muller and Vingron, 2000; Goldman and Whelan, 2002; Veerassamy et al., 2003). In accordance with this, symmetrical substitution models have been used for modeling evolutionary processes (Henikoff and Henikoff, 2000). However, one study has shown that universal amino acid substitution trends exist in different taxa and that this process is one-way and irreversible (Jordan et al., 2005). This study inspired considerable debates, and several works exploring alternate explanations have been performed independently (McDonald, 2005; Goldstein and Pollock, 2006; Hurst et al., 2006). Richard Goldstein and David Pollock argue that the observations cited in the first study can be explained by statistical bias; and John McDonald and Laurence Hurst et al. separately asserted that they can be explained by nearly neutral variation. Jordan et al. responded that Hurst's model is invalid because too many substitutions were introduced to it.

We decided to study this interesting problem using complete genomes of 8 primate species that diverged from each other relatively recently (International human genome sequencing consortium, 2001; The chimpanzee sequencing and analysis consortium, 2005; Locke et al.,

Abbreviations: AA, amino acid; ND, normalized difference; Mya, million years ago; dS/dN, number of synonymous substitutions per synonymous site to the number of non-synonymous substitutions per non-synonymous site; CDS, coding sequence

* Corresponding authors.
E-mail addresses: xiaohuizhang@nju.edu.cn (X. Zhang), sihaiyang@nju.edu.cn (S. Yang).

[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.gene.2014.12.033
0378-1119/

2011; Scally et al., 2012; Prüfer et al., 2012). This allows not only a detailed study of primates' AA substitution but also the study of how AA substitution evolves on different time scales. The data collected here showed AA gain and loss to be closely associated with AA usage. Then 9 AAs (C; F; H; I; L; S; T; V; W) that constantly increased in frequency and 5 AAs (A; D; E; G; P) that constantly decreased in frequency in primates were identified. The effects of AA usage were also found to differ between domain and non-domain regions. Finally, we found evidences that the observed AA substitution patterns are subject to mutation and purifying selection and are compatible with nearly neutral evolution.

2. Methods

2.1. Data source, ortholog annotation and alignment

All the protein and CDS sequences of each species were downloaded from Ensembl (ftp.ensembl.prg/pub/release-76/) (Supplementary Fig. 1). For proteins with multiple transcripts, only the longest transcript was used.

9 species were employed in this study, 8 primates (*Callithrix jacchus*; *Gorilla gorilla*; *Homo sapiens*; *Macaca mulatta*; *Nomascus leucogenys*; *Otolemur garnettii*; *Pan troglodytes*; *Pongo abelii*) and mouse (*Mus musculus*). The 9 species were divided into 4 groups of triplet (species one-species two-outgroup): i) *H. sapiens–P. troglodytes–G. gorilla*; ii) *G. gorilla–P. abelii–N. leucogenys*; iii) *N. leucogenys–M. mulatta–C. jacchus*; and iv) *O. garnettii–C. jacchus–M. musculus*. BLASTP was performed between every two species in each triplet (E-value was set at 1e−30; match length ≥ 50%; identity ≥ 70%), and reciprocal best hits were annotated as orthologs and were aligned by ClustalW (Thompson et al., 1994), then aligned proteins were back translated into CDS. Both alignments are available online (http://gattaca.nju.edu.cn/pub_data.html).

The human SNP data were downloaded from dbSNP (ftp://ftp.ncbi.nih.gov:21/snp/organisms/human_9606_b141_GRCh38/VCF/All.vcf.gz). SNPs in CDS regions were extracted and aligned with *P. troglodytes* to annotate SNP polarization. For the list of SNPs and alignment among human CDS, human CDS with SNP and chimp CDS, please see http://gattaca.nju.edu.cn/pub_data.html.

To track the variation of AA substitution pattern in human over different divergent times, the 9 species were divided into 7 groups of triplet (human-other primate-mouse): i) *H. sapiens–P. troglodytes–M. musculus*; ii) *H. sapiens–G. gorilla–M. musculus*; iii) *H. sapiens–P. abelii–M. musculus*; iv) *H. sapiens–N. leucogenys–M. musculus*; v) *H. sapiens–M. mulatta–M. musculus*; vi) *H. sapiens–C. jacchus–M. musculus*; and vii) *H. sapiens–O. garnettii–M. musculus*. The orthologs were also annotated and aligned in each triplet. The alignments are available online (http://gattaca.nju.edu.cn/pub_data.html).

2.2. Amino acid gain and loss

To identify amino acid gain and loss in each primate, we analyzed the sites where the outgroup and one of the sisters carry the same AA but two sisters carry different AAs. For example, in human–chimp–gorilla alignments, two sisters are human and chimp, gorilla is the outgroup, and for a site where human and gorilla both carry the same AA, like M, but chimp carries a different AA, like H. Then an M is considered lost and an H is gained in chimp.

2.3. dS/dN calculation

The aligned protein sequences were back-translated to CDS sequences, and a Perl program recruiting the method of computing dS/dN in MEGA 5.0 was used to calculate the number of synonymous substitutions per synonymous site and the number of non-synonymous substitutions per non-synonymous site (dS/dN).

3. Results

3.1. Amino acid usage and amino acid gain and loss in primates

Protein sequences from 8 primate species (*C. jacchus*; *G. gorilla*; *H. sapiens*; *M. mulatta*; *N. leucogenys*; *O. garnettii*; *P. troglodytes*; *P. abelii*) were evaluated to assess variation among amino acid usage in primates (Supplementary Fig. 1). Then the 8 species were divided into 4 triplets of species (species one-species two-outgroup): i) *H. sapiens–P. troglodytes–G. gorilla*; ii) *G. gorilla–P. abelii–N. leucogenys*; iii) *N. leucogenys–M. mulatta–C. jacchus*; and iv) *O. garnettii–C. jacchus–M. musculus*. Totally, 8815–14,958 orthologous proteins were identified in each triplet. Then orthologous proteins were aligned for further analysis.

First, AA usage was calculated among these orthologs in the 8 primates. As expected, AA usage remained similar across different primate species (Mean r = 0.999 ± 0007, $P < 10^{-15}$, Supplementary Table 1). Then the numbers of AAs gained and lost were calculated for each species (Table 1). As with of the trends in AA usage, the AA gain and loss patterns of each primate were also found to be similar, and the mean r values were 0.969 (±019, $P < 10^{-5}$) and 0.968 (±0.026, $P < 10^{-5}$) for AA gain and AA loss, respectively. This suggests that AA gain and loss may be related to AA usage. Then linear regression was performed between the number of AAs gained and lost and AA usage. In all 8 species, the number of AAs gained or lost was found to be closely associated with AA usage. The mean r values were 0.766 (±0.038, $P < 10^{-3}$) and 0.850 (±0.054, $P < 10^{-4}$) for gain-usage and loss-usage, respectively.

The more AA used in one species, the greater the number of AAs gained and lost during the evolutionary process. However, it is not possible to infer whether the frequency of these AAs increases or decreases from this number alone. To address this question, normalized differences (ND) between gain and loss number for each AA (ND = (Gain − Loss) / (Gain + Loss)) were used (Table 1). All the primates were found to have similar AA evolutionary trends with mean r = 0.888 ± 0.052 ($P < 10^{-3}$). And ND values are negatively related with usage in all the 8 species. In 7 of the 8 primates the correlation is significant (mean r = −0.532 ± 0.085, $P < 0.05$). On average, ND value is significantly related with usage in the 8 primates (r = −0.602, $P < 0.005$). Thus we can infer that AA usage is an important factor that influences the evolution trend of AAs.

3.2. Evolutionary trends of AA usage in primates and comparison with previous study

To identify AAs that are constantly gained or lost in primates, significantly increased AA was here defined as ND > 0 in all 8 primates (sign-test, $P < 0.05$), and the same criteria were established for significantly decreased AA, ND < 0 in all 8 primates. As a result, 14 out of 20 AAs showed a clear evolving trend (significantly increased or decreased), of which 9 (C; F; H; I; L; S; T; V; W) showed significant increases and 5 (A; D; E; G; P) showed significant decreases. The rest 6 AAs are K, L, N, P, Q, R and Y.

The major difference between our study and Jordan's study (Jordan et al., 2005) is the choice of outgroups. For example, in *Homo–Pan* comparison, Jordan and co-workers used mouse as the outgroup (Jordan et al., 2005) and in this study we use gorilla instead, which is much closer related with *Homo–Pan* than mouse. A relatively distant outgroup may introduce more sites with multiple substitution. To test whether there is an outcome difference introduced by the two methods, we compare our results with Jordan's study. Jordan and co-workers analyzed the evolution trend of AAs in hominidae (*H. sapiens* and *P. troglodytes*, *M. musculus* as outgroup.) (Table 2). On the whole, these two sets of data are highly correlated with r = 0.869 ($P < 10^{-6}$), however, the variance is rather dramatic for each individual AA. For 19 of 20 AAs, the variance is more than 10%, and 12 of them have even more than 50% variance. And 4 AAs (K, N, T and Y) are identified with opposite trend