Evolutionary pattern of protein architecture in mammal and fruit fly genomes

Haiwang Yang¹, Ying Wu¹, Jing Feng, Sihai Yang⁎, Dacheng Tian *

State Key Laboratory of Pharmaceutical Biotechnology, Department of Biology, Nanjing University, Nanjing 210093, China

A R T I C L E   I N F O

Article history:
Received 9 May 2008
Accepted 13 September 2008
Available online 4 November 2008

Keywords:
Protein evolution
Divergence pattern
Mammal
Fruit fly
Protein domain

A B S T R A C T

Mutations, which can alter amino acid constitution, contribute greatly to protein evolution. However, little is reported of their pattern during protein structural evolution. We investigated the distribution of non-synonymous single nucleotide polymorphisms (nsSNPs) and insertions/deletions (indels) along mammal and fruit fly proteins. We found the nsSNPs (and dN) and indels increased in protein boundary regions, and this pattern is inversely correlated with the distribution of protein domain density. Additionally, synonymous substitutions (and dS) are reduced in 5′ and 3′ regions, indicating more variable protein boundaries, compared with central interior. All evidence suggests that the inner part of coding sequences (CDSs) is comparatively conserved, whereas the 5′ and 3′ regions, with higher evolution rates, are more variable. We assumed that due to greater frequencies of nsSNPs and indels in adaptive regions of CDSs it could be easier to ultimately alter, gain, or lose amino acids, thus becoming the front line of protein evolution.

© 2008 Elsevier Inc. All rights reserved.

In the long process of protein evolution, a relatively small set of ancestral domains gives rise to various kinds of proteins by means of divergent evolution [1,2]. Proteins evolve via a succession of mutations in the genome sequence, and among all these mutations, the Single Nucleotide Polymorphism (SNP) represents a major source of genetic heterogeneity [3]. A non-synonymous SNP (nsSNP) in coding regions could directly alter amino acid constitution of a protein, thus having a large effect on protein structure and stability. Insertion/deletion (indel), which could also lead to amino acid alternation, is another common type of polymorphism and thus could be a major driving force shaping genomes [4,5].

Both nsSNPs and indels play critical roles in protein evolution. The nsSNPs in coding regions could lead to loss of hydrogen bonds, breakage of disulfide bonds, introduction of proline into α-helix, and loss of salt bridges [6], and therefore have the potential to affect the function of the proteins. Moreover, nsSNPs could lead to increased biodiversity within one species, and over time, will enlarge the divergence between species, possibly resulting in new species [7,8]. Furthermore, it has been reported that a succession of insertions in protein variable regions could lead to the emergence of novel protein architectures [2,9,10]. Additionally, it was found that indels frequently occurred in the N- or C-terminus of a protein in bacteria, as a consequence of gene fusion/fission, which acts as a major contributor to evolution of multi-domain proteins [11].

Recent studies have shown that many properties of proteins, such as genome position, gene expression level, protein length, codon usage, and protein interactions [12–15], may contribute to differences in protein evolution. It has been reported that even in the same genome, the evolution rates of proteins may vary considerably [16]. For example, in yeast and vertebrates, the gene expression intensity relates inversely to the protein evolution rate on a genomic scale [12,14]. In Drosophila, proteins with higher rates of amino acid substitutions tend to be larger in size and expressed at lower mRNA abundances [13]. In addition, different rates of evolution in different parts of a protein have been observed. For instance, in coding regions, evolution rates were lowest near the intron–exon boundaries, at least partially owing to splicing enhancers [17]. Moreover, the buried amino acids in proteins evolve more slowly than exposed residues [18].

These studies indicate that there might be a general pattern of protein structural evolution to maintain a protein’s common role and to expand its diverse functions. However, little is reported on the general distribution pattern of nsSNPs/indels and the tendency of their positions in coding sequences (CDSs). Therefore, an overview of protein evolution pattern will provide a framework for understanding the mechanisms determining protein architecture. With the advent of a vast amount of genome-sequencing data, nsSNPs/indels identified by genome comparison provide a good opportunity for studying this problem. A statistical method was used to analyze the conserved orthologous CDSs shared by human (Homo sapiens) with three other mammals — chimpanzee (Pan troglodytes), rhesus (Macaca mulatta) and mouse (Mus musculus). The same analysis was used to compare D. melanogaster (Drosophila melanogaster) with three other fruit flies — D. simulans, D. yakuba (Drosophila yakuba) and D. ana (Drosophila ananassae). Our analysis indeed revealed a general pattern of protein structural evolution: a conserved central interior together with adaptable boundaries. This pattern may be an evolutionary

⁎ Corresponding authors. Fax: +86 25 83594720.
E-mail addresses: sibaiyang@nju.edu.cn (S. Yang), dtian@nju.edu.cn (D. Tian).
¹ These authors contributed equally to this work.

0888-7543/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2008.09.009
balance between the maintenance of common role and the expansion of diverse function in proteins. Some further analyses were also included in our study, such as GC-content along the CDS and features of exons in different locations.

Results

Distribution of non-synonymous substitutions in CDS

Six pairwise comparisons – human vs. chimpanzee, rhesus, and mouse and D. mel vs. D. sim, D. yak, and D. ana – were used to analyze the distribution of nsSNPs in a gene. The total amount of nsSNPs between two species is associated with the number of orthologous CDSs shared by the two species and the divergence of these CDS sequences (Table 1). To get an overall view of the nsSNP distribution in a CDS, we assumed each CDS was one unit and marked the relative position of each nsSNP. We then calculated the nsSNP number at each relative position (See Materials and methods).

If nsSNPs were unbiased to a certain position in the CDS, the overall distribution pattern would be a horizontal line. However, the nsSNPs were not evenly distributed along the CDS (Fig. 1). Instead, the 5′ and 3′ ends of the CDS accumulated much more nsSNPs than the inner part. In both mammals and fruit flies, for example, the nsSNP number at the 5′ end was almost twice as much as that in the middle of the CDS. The bias was even more distinct when a greater amount of nsSNP pairs.

To simplify this problem, we selected only the Human–Mouse and D. mel–D. yak pairs to represent mammals and fruit flies respectively, since both of them had the most nsSNPs and the clearest pattern in their groups (Details about the data range for the various features investigated are shown in Supplemental Table S1). Additionally, to get a clear view of nsSNPs distribution at the ends of CDS and also make it possible to fit the data into curves, we further analyzed the points before the relative position 0.2 and after the relative position 0.9 in Figs. 1d, e, in consideration that the decreasing pattern in 5′ regions is stronger than that in 3′ regions. A strong negative correlation between nsSNP number and distance to boundary is presented, and decay curves significantly fit the data (Supplemental Fig. S1).

Furthermore, we confirmed that the nsSNP distribution pattern in CDS was not determined by genes within a specific length or divergence (See Materials and methods; Supplemental Table S2 and S3), and the CDSs we used in our analysis had little protein bias compared with that in the entire genome (See Materials and methods; Supplemental Table S4 and Supplemental Fig. S2).

Table 1

<table>
<thead>
<tr>
<th></th>
<th>Total CDS number</th>
<th>Total nsSNP number</th>
<th>Total indel number</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CDSs without indels</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human–Chimpanzee</td>
<td>12,196</td>
<td>44,735</td>
<td>–</td>
</tr>
<tr>
<td>Human–Rhesus</td>
<td>9062</td>
<td>111,062</td>
<td>–</td>
</tr>
<tr>
<td>Human–Mouse</td>
<td>4194</td>
<td>138,643</td>
<td>–</td>
</tr>
<tr>
<td>D. mel–D. sim</td>
<td>5956</td>
<td>63,832</td>
<td>–</td>
</tr>
<tr>
<td>D. mel–D. yak</td>
<td>5451</td>
<td>317,307</td>
<td>–</td>
</tr>
<tr>
<td>D. mel–D. ana</td>
<td>982</td>
<td>17,040</td>
<td>–</td>
</tr>
<tr>
<td><strong>CDSs with indels</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human–Chimpanzee</td>
<td>2106</td>
<td>17,352</td>
<td>2885</td>
</tr>
<tr>
<td>Human–Rhesus</td>
<td>4335</td>
<td>129,318</td>
<td>8357</td>
</tr>
<tr>
<td>Human–Mouse</td>
<td>5659</td>
<td>534,106</td>
<td>17,570</td>
</tr>
<tr>
<td>D. mel–D. sim</td>
<td>4231</td>
<td>108,812</td>
<td>9701</td>
</tr>
<tr>
<td>D. mel–D. yak</td>
<td>5910</td>
<td>317,307</td>
<td>10,965</td>
</tr>
<tr>
<td>D. mel–D. ana</td>
<td>1961</td>
<td>119,016</td>
<td>13,342</td>
</tr>
</tbody>
</table>

Distribution of indels in CDS

Like nsSNPs, indels have a large impact on protein function and stability as well. To see the indel distribution pattern and how it related to nsSNPs, we used a similar approach to analyze the distribution of indels and nsSNPs in the indel-containing CDS alignments.

As expected, indels and nsSNPs showed a similar pattern in coding regions (Figs. 2a, b), and they were highly correlated with each other in both mammals (r = 0.964, P < 0.01, Fig. 2c) and fruit flies (r = 0.910, P < 0.01, Fig. 2d). Although the total amount of indels was ten times less than that of nsSNPs (Table 1), a similar pattern still existed, and was even clearer in mammals. This might suggest that indels have a greater effect than nsSNPs on protein function.

In addition, we examined the distributions of 3N (multiples of 3 bp) and Non-3N indel-groups (Supplemental Fig. S3). In both mammals and fruit flies, the 3N indel number outweighed that of the Non-3N. And while the patterns of 3N and Non-3N indels were similar, the Non-3N indels were more elevated in the 3′ region of CDS, compared to the 5′ region.

Study on distantly related genes

Our results above from the UCSC alignments only reveal the pattern of highly conserved orthologous CDSs. We could also use protein sequences to get quite accurate alignments for analyzing distantly related orthologs (See Materials and methods). However, since an intact boundary is the key to our analysis, the number of protein alignments was too sparse when we made sure all orthologous proteins were fully aligned (Supplemental Fig. S4). For instance, when we set protein similarity to 60% as a threshold, there are only 302 distantly related orthologous genes for Human–Mouse and 48 for D. mel–D. yak, which is insufficient for us to see a definite pattern. But it is notable that the average protein similarity for paralogs (0.77 and 0.69 for Human and D. mel respectively) is much lower than that of orthologs (0.87 and 0.93). Therefore, analyzing paralogs may enable us to have a glimpse of the pattern in highly divergent genes. Supplemental Fig. S5a, b show that the pattern still existed in paralogs, especially in the N-terminus.

Variation of GC-content in CDS

Different recombination rates and duplication times have been found between GC-rich and GC-poor regions in mammalian genomes [19]. For this reason, we were also intrigued by GC-content distribution along the CDS.

For the mammal species tested, the distribution curve of GC-content and nsSNPs for the 5′ end of CDSs could fit fairly well (Figs. 3a and 1e), since the correlation between GC-content and nsSNPs before the relative position 0.2 was significant (r = 0.99, P < 0.01; not shown). However, at the 3′ region, the nsSNPs number increased sharply, whereas the GC-content did not follow this trend. In fruit flies, the relationship between GC-content and nsSNPs was even more complicated as shown in Figs. 3b and 1d. The intricate relationship between GC-content and nsSNPs in fruit fly means that GC-content may not be responsible for the excess of nsSNPs at the end of CDS. Although GC-content had some correlation with mammal nsSNPs, the nsSNP distribution phenomenon cannot be solely explained by the level of the GC-content.

Distributions of $d_{A}$, $d_{G}$ and protein motifs in CDS

The ratio of non-synonymous to synonymous rates varies greatly among amino acid sites within a protein [20], presumably due to selection acting differently at different protein sites [21]. The $d_{A}$ and $d_{G}$ were defined as the numbers of non-synonymous and synonymous substitutions per site according to Nei and Gojobori (1986) [22], and
their ratio $\omega$ was frequently used to evaluate selection constraint and evolution rates in protein.

In both mammals and fruit flies, the ratio $\omega$ is far smaller than 1 along the CDS (Fig. 4), suggesting the role of purifying selection in protein coding regions. Additionally, it was believed that synonymous mutations, which are supposed to be functionally neutral, had a constant rate among different sites in protein [20,23]. However, in our results $d_S$ and synonymous SNPs (sSNPs) were not evenly distributed across the CDS, but rather decreased at the extremes, especially in the 5′ region (Fig. 4 and Supplemental Fig. S6). On the other hand, in both mammals and fruit flies, $d_N$ showed a similar pattern to that of nsSNP number, leading to a sharp increase of $\omega$ in 5′ region and a comparatively mild one in 3′ region. Therefore, although purifying selection governs the whole CDS, it is relaxed dramatically at the two extremes, especially in the 5′ region.

Protein domains are basic evolutionary units [24] and tend to have a highly conserved location in proteins [25]. Therefore, we also analyzed their density in the CDSs (Fig. 4), and in both mammals and fruit flies it was obvious that domains tended to avoid locating near the N- and C-terminal of protein, which is inversely correlated with the $\omega$ distribution ($r=-0.870$, $P<0.01$ and $r=-0.904$, $P<0.01$ for Figs. 4a, b, respectively).

Fig. 1. Patterns of nsSNP distribution across CDSs in mammals and fruit flies. The point 0.0 in X-axis represents the 5′ end, while 1.0 represents the 3′ end. Only non-indel CDS alignments were used.
Distribution of nsSNPs in individual exons

The excess of nsSNPs at the ends of CDSs is a phenomenon on the border between coding and non-coding sequences, suggesting that the excess of nsSNPs might be present at the ends of individual coding exons as well. Therefore, in intron-containing genes, we analyzed the nsSNPs distribution separately for end exons (5′ exon and 3′ exon) and inner exons (the separate exons between the end exons). In both mammals and fruit flies, the 5′ and 3′ exons showed a decreasing and an increasing pattern, which was consistent with the overall pattern respectively, whereas in the inner exons, nsSNP number decreased at the both ends (Fig. 5). These results suggest that the excess of nsSNPs is only present at the ends of the CDS not in all exon ends.

Alternative splicing in end and inner exons

In variable regions (5′ and 3′ regions of CDSs), higher possibilities to be alternatively spliced are expected. To test this assumption, multi-
transcript genes in human and D. mel were used to detect whether alternative splicing sites tend to be located in the end exons or the inner exons. Our calculations showed that in human multi-transcript genes, 35.5% of the end exons were alternatively spliced (Table 2), whereas in inner exons it was only 19.2%. The contrast was even more striking in the fruit fly, with 25.9% in end exons and 13.8% in inner exons.

Discussion

Our study had shown that nsSNPs and indels were not evenly distributed along the CDS, but elevated in 5′ and 3′ region. Specifically, the excess of nsSNPs in the 5′ region was generally more abundant than in the 3′ region. This pattern was confirmed to be present in both short and long, or conserved and non-conserved CDSs. In contrast, the sSNPs (and dS) and the motif density were inversely distributed from the nsSNPs/indels. Clearly, there is a conserved central interior, which is closely correlated with the density of protein motifs. In other words, the sequences and structures of protein motifs are more stably maintained.

The higher value of dS and the lower value of dN suggest the more variable 5′ or 3′ region is under a higher constraint for diversifying selection. The classic model assumes that synonymous substitutions are free from selection and their rates are constant across codon positions [20,23]. From our results, however, dN was reduced at the end of CDS, indicating the extremes, especially the 5′ part, might be generated at a relatively later stage. While our analysis didn’t sufficiently prove relative age, it is likely that the less variable sections have existed longer than more variable sections. Therefore, the domain scarcity observed at those regions may reflect the immaturity of motifs.

Additionally, the comparison of end and inner exons revealed that the 5′ and 3′ exons contribute greatly to the general pattern of nsSNP distribution in CDSs. On the other hand, fewer nsSNPs at the ends of inner exons indicates important functions exist in these regions. Previous studies have provided clues to such a phenomenon, that there existed some conserved exon splicing enhancer at the end of inner exon [17]. Moreover, our analysis showed that the end exons have a greater chance to be alternatively spliced, consistent with other results.

A conserved central interior together with adaptable boundaries could be evolutionarily advantageous. In the central interior of multi-domain protein, there exists an N to C-terminal series of domains, namely, the “domain architecture”; the sequential order of these domains is fairly conserved [26]. In the domain architecture, domain combinations have quite a limited repertoire compared with the huge number of possible combinations, and pairs of neighboring domains tend to have the same orientation in different proteins, indicating strong selection in this region [26–28]. The central interior
region is so important that multi-domain proteins mainly evolved by adding another domain at either terminus [29,30], since addition of domains in the middle would tend to disrupt the tertiary/quaternary structure [31].

The studies of generation and divergence of new genes in fruit flies may provide some hint for how adaptable boundaries could beneficially extend the function of genes. Yang et al. [32] reported that after duplication, genes could vary their peptide sequence at N- or C-terminals by accumulating deletions, shifting start and stop codons, or even by discarding some flanking exons and recruiting part of intron sequence into exons. The higher variability at N- or C-terminals is similar to the model “permutation by duplication”, which assumes that gene duplication and fusion occur first, followed by partial degradation of 5’ and 3’ coding regions, resulting in a new permuted gene [33–37]. Considering that 80% of eukaryotic proteins have more than one domain [24,38], our results were consistent with this model in that the fused multi-domain proteins tend to degrade or truncate their 5’ and 3’ regions.

More frequent insertions/deletions observed at the ends of CDSs could also contribute to the adaptive role of these regions, and ultimately to the evolution of protein structure. It has been reported that with a succession of insertion events and rapid evolution, the variable regions in a protein could generate novel structural architectures [39]. Several other studies have also confirmed insertion events could lead to incremental growth of variable regions, generating novel domains [29,10]. Recent study has revealed that indel polymorphism appears to act as mutational enhancers in sequences immediately surrounding the indel [40]. In addition, indels frequently occur in terminal regions of a protein, and may be relevant to the evolution of multi-domain proteins (at least in bacteria) [11]. It is clear that although protein structure evolves quite slowly [39,41], there exists some variable regions in proteins which could tolerate many more mutations. In these regions, indels could play an important role in generating new protein structures.

Considering both nsSNPs and indels accumulate at the variable boundaries, it is tempting to suggest the N- and C-terminals of a protein may have a greater chance of being the hotbed of novel domains and front line of protein evolution. The excess of both nsSNPs and indels at protein boundaries kept in accordance with this scenario. All evidence showed that the inner part of CDSs is conserved compared to the 5’ and 3’ regions, which have a higher evolution rate and thus are much more variable. The amino acids in the dynamic boundary regions of the protein could be altered, gained, or lost due to the high frequencies of nsSNPs/indels or the shifting of start/stop codons. This made these regions changeable, stretchable, and shrinkable in the course of adaptive evolution.

In conclusion, we found a general evolutionary pattern of protein architecture in mammals and fruit flies: adaptable boundaries and conserved inner parts in CDSs. This pattern is characterized by abundant nsSNPs/indels, a higher ω value, and a lower ds value in the boundaries. In contrast, a lower dw value and higher density of protein domains indicate the conserved central part of the protein. The existence of this pattern probably owes to stronger purifying selection in the interior region for the maintenance of conserved function. Additionally, adaptable boundaries are advantageous for the generation of new functions of genes. This scenario is consistent with recent reports, associated with novel gene generation and new protein structure, and also with our expectations. Therefore, the phenomenon discovered may provide the framework for building a better understanding of protein evolution.

Materials and methods

Sequences and alignments

For *H. sapiens* and *D. melanogaster*, the CDSs and their locations and lengths were retrieved from Ensembl (build 47) (http://www.ensembl.org/). All pairwise alignments were downloaded directly from UCSC Genome Browser (http://hgdownload.cse.ucsc.edu/). According to their locations, whole length CDS alignments were retrieved from the alignments above. Only protein coding genes were retained, and for genes encoding more than one protein isoform, only the first CDS provided by Ensembl database was used in order to make sure there was only one CDS for each gene. To confirm this manipulation has no bias, a random extraction was used, and nsSNPs/indels distribution was calculated in the same way. Then the nsSNP (and also indel) number at each relative position was plotted against that of the previous result, and we got *R = 0.98* and *P = 0.01* for all linear correlations, indicating that always using the first CDS did not affect the final result. We also excluded the genes that did not begin with the start codon (ATG) or did not end with the stop codon (TAA/TGA/TAG). To avoid misalignment, we defined orthologs as those CDSs with DNA similarity higher than 80%, and discarded the remainder. We also made sure the CDSs are not started or ended by gaps and have no base “N” inside. Then, the qualified CDS alignments were divided into two groups: those without indels and those with indels. The number of nsSNPs and the lengths of CDSs in the former group could be precisely calculated, thus becoming our main data source. The latter group was only used to calculate indel distribution and its relationship with nsSNPs. We also proved that in both human and *D. melanogaster* more than 97% of the genes we used in the former group have EST support (Supplemental Table S5).

Notable examples marking sSNPs, nsSNPs, indels, protein domain, and GC-content are provided in Supplemental Fig. S7.

Analysis of nsSNP/indel distribution and relevant statistics

Given each CDS one unit, we could determine relative positions of each nsSNP, e.g. a nsSNP in the middle of CDS having a relative position ‘0.5’ according to the formula: Relative Position = Real Position/CDS Length. To examine how these nsSNPs were distributed across the CDS, we equally divided a CDS into 100 parts (sub-regions), e.g. ‘0.00–0.01’, ‘0.01–0.02’, ‘0.02–0.49’, ‘0.50–0.99’, and then calculated the total number of nsSNPs in a particular position (sub-region). When graphically representing these data, the mid-point of each sub-region was used to represent the entire sub-region, e.g. 0.015 on X-axis representing the section ‘0.00–0.02’. The first and the last sub-regions, ‘0.00–0.01’ and ‘0.99–1.09’, were not shown in the graph, because the nsSNP number there could be affected by the start and stop codon. The distribution of GC-content, sSNPs and indels was calculated in the similar way. However, since indels themselves have a length, the indel length was neglected and the relative position of its 5’ flanking base was used as the position of each indel instead.

When using the intron-containing genes to analyze the nsSNPs distribution across individual exons, the CDSs were divided into 5’, 3’, and inner exons. Considering that sometimes codons at the end of exons are severed by introns, which prevent us from distinguishing nsSNPs from sSNPs, these codons were eliminated because of their incompleteness. We confirmed this “Cut-away” manipulation did not affect the ultimate result by another method, in which after removing the intron from the severed codon the lone base was joined to the remaining two bases, reforming the original codon. Analysis of nsSNP
distribution in exons was similar to that of the CDS, but we only depicted the distribution in ten sub-regions due to the shortness of individual exons.

The calculation of \( d_n \), \( d_s \), and \( \omega \) distribution was similar to that of the nsSNPs, except that it was based on codon pairs rather than bases. Nei and Gojobori method (1986) [22] was used to get \( d_n \), \( d_s \), and \( \omega \) for each sub-region of CDS.

Test with smaller groups of genes and protein bias analysis

To examine whether the excess of nsSNPs at the ends of CDSs was caused by a specific group of genes, we used two criteria – CDS length and divergence – to categorize our CDS alignments. The CDSs were first sorted by their lengths into 7 groups: five for the CDSs between 600 bp and 2100 bp (each with a 300 bp length range) and two groups for the CDSs shorter than 600 bp or longer than 2100 bp. Also, the divergence (measured applying the Jukes–Cantor correction to the number of substitutions per site) was used as a criterion to categorize the CDSs, which were equally divided into three groups, each with a low, medium or high value of divergence. In each group of genes, we analyzed the nsSNP distribution before the relative position 0.2 in 5′ ends and after 0.9 in 3′ ends. We characterized all correlations by decay curves, and calculated the significance in each group.

Additionally, to test whether the CDSs we retrieved from UCSC alignments had a protein bias compared with that in the entire genome, we used BioMart (Version 0.5, http://www.ensembl.org/biomart/) to classify the genes by GO (Gene Ontology) categories. In each category, we listed the number of genes in the genome and the dataset we used respectively, and plotted them against each other.

Analysis of the protein domain density

Protein sequences of Human and D. mel in the comparisons of Mouse, we used each human protein (the first protein was selected for multi-transcript genes) as a query in BLASTp search of all possible proteins in mouse genome. The threshold of expectation value (E-value) was set to \( 10^{-4} \). Then the hit protein with the smallest E-value was aligned with the query protein using ClustalW with default options. To make sure that proteins were completely aligned to one another, we retained only the alignments with no indels at the N- and C-terminals. Then we analyzed the distribution of amino acid alterations, which was highly correlated with the nsSNP distribution, and indel distribution. A similar method was used to obtain and analyze paralogous proteins.

Acknowledgment

This research was supported by NSFC (30570987) to D.T.

Appendix A. Supplementary data


References


