Gene 534 (2014) 163–168

Contents lists available at ScienceDirect

Gene

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

Macroevolutionary trends of atomic composition and related functional group proportion in eukaryotic and prokaryotic proteins

Yu-Juan Zhang [a,b], Chun-Lin Yang [c], You-Jin Hao [b], Ying Li [b], Bin Chen [b,*], Jian-Fan Wen [a,**]

[a] State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan Province 650223, PR China
[b] Institute of Entomology and Molecular Biology, College of Life Sciences, Chongqing Normal University, Shapingba, Chongqing 401331, PR China
[c] Department of Biology, East Carolina University, Greenville, NC, USA

ARTICLE INFO

Article history:
Accepted 30 October 2013
Available online 18 November 2013

Keywords:
Amino acid
Prokaryotes
Eukaryotes
Macroevolution
Oxygen
Sulfur

ABSTRACT

To fully explore the trends of atomic composition during the macroevolution from prokaryote to eukaryote, five atoms (oxygen, sulfur, nitrogen, carbon, hydrogen) and related functional groups in prokaryotic and eukaryotic proteins were surveyed and compared. Genome-wide analysis showed that eukaryotic proteins have more oxygen, sulfur and nitrogen atoms than prokaryotes do. Clusters of Orthologous Groups (COG) analysis revealed that oxygen, sulfur, carbon and hydrogen frequencies are higher in eukaryotic proteins than in their prokaryotic orthologs. Furthermore, functional group analysis demonstrated that eukaryotic proteins tend to have higher proportions of sulfhydryl, hydroxyl and acylamino, but lower of sulfide and carboxyl. Taken together, an apparent trend of increase was observed for oxygen and sulfur atoms in the macroevolution; the variation of oxygen and sulfur compositions and their related functional groups in macroevolution made eukaryotic proteins carry more useful functional groups. These results will be helpful for better understanding the functional significances of atomic composition evolution.

1. Introduction

It is believed that eukaryotic cells evolved much later (only about 1–1.5 billion years ago) than prokaryotic organisms did, which existed as far back as 3.5–3.8 billion years ago (Schopf, 1978). During the evolution from prokaryotes to eukaryotes, mutations in DNA accumulated and resulted in the gains of new characteristics of the proteins coded by these DNA. A large number of studies have been carried out to investigate the difference between eukaryotic and prokaryotic proteins at amino acid (AA) level, such as homolog duplications, domain gains and losses (Tomasinsig and Zanetti, 2005; Zhang et al., 2009), variations in sequence length (Brocchieri and Karlin, 2005), and the evolution of higher structures (Thorne, 2000). In contrast, little attention has been paid on the evolution of atomic composition (Acquisti et al., 2007; Baudouin-Cornu et al., 2001).

All proteins are mainly composed of 20 kinds of AAs with different side-chains carrying varying amounts of oxygen, sulfur, nitrogen, carbon and hydrogen atoms. Previous studies showed that different proteins have different atomic composition biases during evolution, and the trend of atomic composition would be influenced by environment and nutrient availability (Acquisti et al., 2009; Baudouin-Cornu et al., 2001). For instance, Chen and Zhang (2008) reported that oxygenic photosynthetic proteins, such as the D1 peptide in photosynthetic reaction center type II complexes, contain more oxygen atoms than anoxygenic photosynthetic proteins. Similarly, sulfur-assimilatory proteins possess low sulfur frequency compared with other proteins in *Escherichia coli* and *Saccharomyces cerevisiae*. It was suggested that natural selection may have shaped the elemental composition of proteins with specific metabolic functions to respond to nutrient limitation (Baudouin-Cornu et al., 2001). Acquisti et al. (2009) studied the nitrogen frequency of the proteins involved in the anabolic and catabolic apparatus in four multicellular eukaryotic model organisms (*Arabidopsis thaliana*, *Drosophila melanogaster*, *Homo sapiens* and *Mus musculus*), and concluded that resource availability constrained the atomic composition and the molecular architecture of the biotic processes that enable cells to respond to the reduced nutrient availability.

The studies mentioned above mainly focused on a specific group of functional proteins, whereas the studies of Acquisti et al. (2007) and Sasidharan et al. (2008) focused on the oxygen composition of the transmembrane proteins in 19 taxa, and showed that eukaryotic transmembrane proteins have higher oxygen frequency than prokaryotic transmembrane proteins do, which might be correlated with atmospheric oxygen concentration over macroevolutionary time scales. But they did not distinguish homologous and non-homolog proteins in

---

Abbreviations: AA, amino acid; COG, Clusters of Orthologous Groups; COG1, 604 COGs distributed in both prokaryotes and eukaryotes; COG2, 456 COGs from 44 mesophilic species; K_e, the proportion of Group to Atom on the side-chains of AA in prokaryotic proteins; K_p, the proportion of Group to Atom on the side-chains of AA in eukaryotic proteins; $-COOH-$, carboxyl; $-H_2N-CO-$, acylamino; $-H_3C-S-H_3C-$, sulfide; $-OH-$, hydroxyl; $-SH$, sulfhydryl.

[*] Correspondence to: B. Chen, College of Life Sciences, Chongqing Normal University, Shapingba, Chongqing 401331, PR China. Tel./fax: +86 23 65910315.

[**] Corresponding author. Tel./fax: +86 871 85198682.

E-mail addresses: c_bin@hotmail.com (B. Chen), wenjf@mail.kiz.ac.cn (J.-F. Wen).

0378-1119/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.gene.2013.10.070

their analysis, which would make it complicated to find the trend of atomic composition change over macroevolutionary time scales. Moreover, they did not distinguish proteins from extreme-living and mesophilic organisms. Life style and growth temperatures may affect the charged AA composition of organisms, and thus affect their atomic compositions. That is, proteins in the same family but from extreme-living and mesophilic organisms may have different atomic composition bias, which should also be took into consideration in study.

The abovementioned studies of oxygen atom composition trigger us to think whether other atomic compositions also changed over the macroevolutionary process from prokaryotes to eukaryotes. Tracing the trend of atoms in macroevolutionary time scales is an interesting work, which can reflect environmental transition and nutrient availability, and is also helpful in clarifying the differentiation of function between prokaryotic and eukaryotic proteins. On the side-chain of AAs, specific groups of atoms constitute functional groups, which are responsible for the characteristic of biochemical reactions. There are two sulfur functional groups (sulfhydryl and sulfide) and three oxygen functional groups (hydroxyl, carboxyl and acylamino) on the side-chains of AAs. For example, disulfide bonds usually form from the oxidation of sulfhydryl ($-SH$) groups between cysteines (Trivedi et al., 2009), and hydroxyl ($-OH-$) can form hydrogen bonds in proteins (Mori et al., 2010). Based on the trend of atomic composition, analyzing the change pattern of the proportion of functional groups to their corresponding atoms would reveal the change pattern of the functional groups' proportion that leads to functional significance. However, until now, no studies have been carried out to study the evolutionary trends of the five kinds of atoms mentioned above and their corresponding functional groups over the macroevolutionary process.

The aim of this study is to analyze the variations of atomic composition and the related functional groups' proportion on side-chains of prokaryotic and eukaryotic proteomes. The five kinds of atomic compositions were investigated genome-widely. Atomic compositions were also calculated for those proteins in the same orthologous groups. The compositions of sulfur-containing functional groups, sulfhydryl and sulfide ($-H_3C-S-H_3C-$), and of oxygen-containing functional groups, hydroxyl, carboxyl ($-COOH-$) and acylamino ($-H_2N-CO-$), were also computed and analyzed.

2. Materials and methods

2.1. Data retrieving and ortholog clustering

All available prokaryotic (1051) and eukaryotic (66) genome data with annotated proteins were downloaded from GenBank (ftp://ftp.ncbi.nih.gov/genomes), in Dec. 2010. Then, 156 prokaryotes of these organisms were classified as mesophilic organisms (living under temperatures between 50 °C and 15 °C) according to PGTdb description (http://pgtdb.csie.ncu.edu.tw). The 156 mesophilic prokaryotes were further classified into 9 main groups as follows: actinobacteria, alphaproteobacteria, archaea, betaproteobacteria, CFB group bacteria, delta/epsilon proteobacteria, firmicutes, gammaproteobacteria and spirochetes. Each group contains at least 5 species (Tables 1–3 in Additional File 1).

All the proteins of the 4879 Clusters of Orthologous Groups (COG) (Tatusov et al., 1997) distributed in 66 species, were downloaded from NCBI, which is the same as the latest data. 1) 604 COGs (designated as COG1) distributed in both prokaryotes and eukaryotes were obtained (Table 4 in Additional File 1). 2) Sequences from hyperthermophilic, thermophilic and psychrophilic species were excluded according to PGTdb description. Finally, 456 COGs from 44 mesophilic species (designated as COG2) were obtained (Table 5 in Additional File 1).

2.2. Analysis of compositions of atoms and functional groups

Average frequency of the five kinds of atoms (carbon, hydrogen, oxygen, nitrogen and sulfur) and five related functional groups (sulfhydryl, sulfide, hydroxyl, carboxyl and acylamino) in a full protein sequence were estimated as follows:

$$[*\text{frequency}] = \sum wi \times \text{pi}/\text{L}$$

where wi is the number of * atom (or functional group) on the AA side-chain (ranging from 0 to 10), pi is the frequency of the i-th AA, and L is the sequence length. The frequency of an atom (or functional group) is the mean of all atoms (or functional group) in all sequences in one proteome or one orthologous group. All data were calculated by using Perl scripts.

2.3. Analysis of atom proportion of functional groups

2.3.1. Functional group proportion analysis

$$K_e = [\text{Group frequency}]_e/[\text{Atom frequency}]_e;$$

$$K_p = [\text{Group frequency}]_p/[\text{Atom frequency}]_p.$$

If 'Group' is sulfhydryl or sulfide, then 'Atom' is sulfur; if 'Group' is hydroxyl, carboxyl or acylamino, then 'Atom' is oxygen. K_e represents the proportion of Group to Atom on the side-chains of AAs in prokaryotic proteins; K_p represents the proportion of Group to Atom on the side-chains of AAs in eukaryotic proteins.

For example, if Group is sulfhydryl, K_e represents the proportion of sulfhydryl frequency to sulfur frequency on the side-chains of AAs in eukaryotic proteins; accordingly K_p represents the proportion of sulfhydryl frequency to sulfur frequency on the side-chains of AAs in prokaryotic proteins. K_e and K_p values were also calculated by using Perl scripts.

2.4. Statistical analysis

Mann–Whitney *U* test was performed using SPSS 13.0 statistics software (SPSS Inc., Chicago, IL). Using R 2.15.2 (Team, 2012), two independent sample *t*-test was performed for atomic compositions between 66 random samples from prokaryotic genomes (1051) and 66 eukaryotic genomes. Random sampling was repeated 1000 times. To improve robustness and consistency, $P < 0.001$ was considered to be statistically significant.

3. Results

3.1. Atomic composition at the whole genome level

To investigate whether the atom frequency varied over the macroevolutionary time scale, the five atom frequencies on the side-chains of proteins of 1051 prokaryotic genomes and 66 eukaryotic genomes were calculated and compared. The results showed that 1) eukaryotes have a higher frequency of oxygen than prokaryotes ($P < 1E-14$)—the average oxygen frequency on side-chains is 0.451 per AA for prokaryotes, and 0.487 per AA for eukaryotes; 2) a significant difference happened to sulfur frequency ($P < 1E-12$)—0.037 per AA for prokaryotes and 0.042 for eukaryotes; 3) nitrogen frequency also shows a significant difference ($P < 1E-15$) between prokaryotes (0.353 per AA) and eukaryotes (0.374 per AA); 4) carbon frequency shows no difference ($P > 0.001$) between prokaryotes and eukaryotes; and 5) a significant difference ($P < 0.001$) happened to the hydrogen frequency between prokaryotes (5.889 per AA) and eukaryotes (5.809 per AA) (Fig. 1 and Additional File 2). Similar significant differences of atom